Smoothing Spline Analysis of Variance for Correlated Non-Gaussian Data

P. Karcher and Y. Wang

University of California
Dep. of Statistics and Applied Probability
Santa Barabara, CA 93117

Abstract

Generalized Linear Mixed Effects Models (GLMM) provide useful tools for correlated and overdispersed non-Gaussian data. In this paper we consider Generalized Nonparametric Mixed Effects Models (GNMM) which relax the rigid linear assumption on the conditional predictor in a GLMM. We use smoothing splines to model fixed effects. The random effects are general and may also contain stochastic processes corresponding to smoothing splines. We show how to construct smoothing spline ANOVA (SS ANOVA) decompositions for the predictor function. Components in a SS ANOVA decomposition have nice interpretations as main effects and interactions. The experiment design decides naturally which components are fixed or random. We estimate all parameters and spline functions using stochastic approximation with Markov Chain Monte-Carlo. We illustrate our method using a longitudinal data with binary response.

Keywords: Exponential family; Generalized Linear Mixed Effects Models; Smoothing Spline ANOVA; Nonparametric Mixed Effects Models; Markov Chain Monte-Carlo; Stochastic approximation; Fixed and random effects; Penalized likelihood.

1 Introduction

Linear Models (LM) with normal random errors are some of the lasting and popular tools in statistics. They assume a linear relationship between mean response and parameters. These models have been generalized in three directions. One extension is to Linear Mixed Effects Models (LMM) which permit fixed and random effects. LMM are used to model both the mean and covariance structure (Robinson 1991). The second extension is to Generalized Linear Models (GLM) which provide a unified regression methodology for independent non-Gaussian responses (McCullagh & Nelder 1989). The third extension is to Nonparametric Regression Models (NRM) which allow more flexibility for the mean function (Wahba 1990). Combinations of any two of these three directions lead to Generalized Linear Mixed Models (GLMM) (Breslow & Clayton 1993), Nonparametric Mixed Effects Models (NMM) (Wang 1998), and Generalized Nonparametric Regression Models (GNRM) (Wahba, Wang, Gu, Klein & Klein 1995). GLMMs model correlation and/or overdispersion for non-Gaussian data. NMMs model both the mean and covariance structure nonparametrically for Gaussian data. GNRMs model the prediction function nonparametrically for independent data from the exponential family. There has been little research that generalizes along all three directions. One exception is Zhang & Lin (1999). In this paper we propose Generalized Nonparametric Mixed Effects Models (GNMM), which extend all models discussed above. Figure 1 shows the relationships between various models.

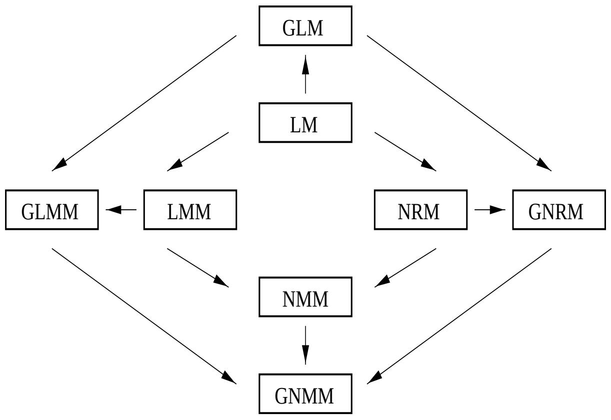

Figure 1: The overview of different models. An arrow represents an extension to a more general model.

Our model is different from that of Zhang & Lin

(1999) in terms of both modeling and estimation. Zhang & Lin (1999) used an additive model for the fixed effects. We will build our models using Smoothing Spline ANOVA decompositions which are defined on general domains, allow for arbitrary interactions, specify fixed and random effects in a natural way and allow spline structures in the random effects. Zhang & Lin (1999) used Laplace approximation to estimate parameters which may have large bias (Lin & Breslow 1996, Jiang 1998). We will use a stochastic approximation with a Markov Chain Monte-Carlo (MCMC) method which leads to asymptotically consistent estimates.

Many applications collect correlated non-Gaussian data, for example, in terms of counts or proportions over time from several experimental units. Observations from the same experimental unit and/or observations which are close in time are oftentimes correlated. GLMMs are commonly used to analyze such data. As with all parametric models, the linear relationship between the predictor and the parameters may be too restrictive. Preliminary and residual plots are less informative when the response is discrete. Nonparametric models let the data suggest the the relationship. Therefore they provide powerful tools for modeling and diagnostics.

In Section 2 we define the GNMM and describe the Smoothing Spline ANOVA decompositions. We develop the estimation method using stochastic approximation with Markov Chain Monte-Carlo in Section 3. A longitudinal data set is analyzed in Section 4 to demonstrate our method. In Section 5 we state our conclusions.

2 Generalized Nonparametric Mixed Effects Model

2.1 The Model

Suppose we observe $\{(y_i, \mathbf{t}_i), i = 1, \ldots, n\}$ where y_i is the response with a d-dimensional covariate $\mathbf{t}_i = (t_{i1}, \ldots t_{id})$. Let $\mathbf{U}$ be a q dimensional vector of random effects with mean zero and density $f_{\mathbf{U}}(\mathbf{u}; \psi)$, where ψ is a parameter vector. Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ be a realization of the random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$. Assume that conditional on $\mathbf{U}$, Y_i's are independent samples from the exponential family

$$f_{Y_i|\mathbf{U}}(y_i) = exp\left\{(\theta_i y_i - b(\theta_i))/a_i(\alpha) + K_i(\alpha, y_i)\right\}. \quad (1)$$

Denote the link function by g. We will model the conditional predictor as

$$g(E(\mathbf{Y}|\mathbf{U})) = \mathbf{f} + \mathbf{Z}\mathbf{U}, \quad (2)$$

where $\mathbf{f} = (f(\mathbf{t}_1), f(\mathbf{t}_2), \ldots, f(\mathbf{t}_n))^T$, and $\mathbf{Z}_{n\times q}$ is the design matrix of the random effects. The function f is used to model the fixed effects and $f \in \mathcal{M}$, where $\mathcal{M}$ is a chosen functional space discussed in Section 2.2. The conditional density (1), the link function g, and the conditional predictor (2) constitute the *Generalized Nonparametric Mixed Effects Model (GNMM)* .

2.2 Smoothing Spline ANOVA Decomposition

Suppose that each covariate takes values in some fairly arbitrary index set $\mathcal{T}_k$. Interesting examples are $\{1, \ldots, K\}$ for a discrete, $[0, 1]$ for a continuous, and $\mathbb{R}^2$ for a spatial covariate. f is a real-valued function of $\mathbf{t} = (t_1, \ldots, t_d) \in \mathcal{T} = \mathcal{T}_1 \otimes \mathcal{T}_2 \otimes \ldots \otimes \mathcal{T}_d$. In the following we show how to construct the model space $\mathcal{M}$ using *Reproducing Kernel Hilbert Spaces (RKHS)*. See Aronszajn (1950) and Wahba (1990) for details on RKHS.

Suppose we want to model the effect of the kth covariate t_k using space $\{1\} \oplus \mathcal{H}^k$, where $\{1\}$ is a one dimensional space of constant functions on $\mathcal{T}_k$ and $\mathcal{H}^k$ is a RKHS of functions $\mathcal{T}_k$. Let $\oplus$ be the direct sum and $\otimes$ be the tensor product of RKHSs. Consider the tensor product space

$$\begin{aligned} \mathcal{H} &= \prod_{k=1}^{d} [\{1\} \oplus \mathcal{H}^k] = \mathcal{H}_\emptyset \oplus \sum_{k=1}^{d} \mathcal{H}_{\{k\}} \\ &\quad \oplus \sum_{k<l} \mathcal{H}_{\{k,l\}} \oplus \cdots \oplus \mathcal{H}_{\{1,\ldots,d\}}, \end{aligned} \quad (3)$$

where, for example, $\mathcal{H}_{\{1,3\}}$ represents $\mathcal{H}^1 \otimes \{1\} \otimes \mathcal{H}^3 \otimes \{1\} \otimes \ldots \otimes \{1\}$. The decomposition in (3) is equivalent to decomposing f as

$$\begin{aligned} f(t_1, \ldots, t_d) &= \mu + \sum_{k=1}^{d} f_k(t_k) + \sum_{k<j} f_{kj}(t_k, t_j) \\ &\quad + \cdots + f_{1,\ldots,d}(t_1, \ldots, t_d). \end{aligned} \quad (4)$$

Decompositions (3) and (4) are called Smoothing Spline ANOVA (SS ANOVA) decompositions. As in classical ANOVA models, all components in (4) have nice interpretations: μ as the overall mean, $f_k(t_k)$ as the main effect of t_k, $f_{kj}(t_k, t_j)$ as the interaction between the kth and the jth covariate, and so forth. The SS ANOVA decomposition can also be derived by defining appropiate averaging operators. See Wahba (1990) and Gu & Wahba (1993) and Section 4 for details. The design of the experiment decides which terms in (4) are fixed effects and which terms are random effects. See Wang (1998), Wahba & Wang (1998) and the example in Section 4. The random effects terms can easily be written into the random effects form $\mathbf{Z}\mathbf{U}$ in (2).

To overcome the curse of dimensionality only some terms in (3), usually lower order terms, are included in the model space. After a model is chosen, we can rearrange the terms and write the model space for the fixed effects as

$$\mathcal{M} = \mathcal{M}^0 \oplus \sum_{k=1}^{p} \mathcal{M}^k \tag{5}$$

where $\mathcal{M}^0$ is a finite dimensional space containing terms which are not going to be penalized. $\mathcal{M}^k$ are orthogonal subspaces containing terms that we want to penalize. See Wahba & Wang (1998) for discussion on model selection.

3 Estimation

3.1 Penalized Likelihood

The log-likelihood of $\mathbf{y} = (y_1, \ldots, y_n)^T$ is

$$\begin{aligned} & l_{\mathbf{Y}}(f, \alpha, \psi) = \\ & \quad \mathit{log}\left(\int \prod_{i=1}^{n} f_{Y_i|\mathbf{U}}(y_i; f, \alpha) f_{\mathbf{U}}(\mathbf{U}; \psi) d\mathbf{U} \right). \end{aligned} \tag{6}$$

The log-likelihood of $\mathbf{y}$ conditional on $\mathbf{U}$ is

$$\begin{aligned} l_{\mathbf{Y}|\mathbf{U}}(f, \alpha) &= \mathit{log}\left(\prod_{i=1}^{n} f_{Y_i|\mathbf{U}}(y_i; f, \alpha) \right) \\ &= \sum_{i=1}^{n} \mathit{log} f_{Y_i|\mathbf{U}}(y_i|\mathbf{U}; f, \alpha). \end{aligned} \tag{7}$$

Denote $|| \quad ||_k$ as the norm of $\mathcal{M}^k$ and P^k as the orthogonal projection in $\mathcal{M}$ onto $\mathcal{M}^k$. We estimate f by minimizing the the penalized likelihood

$$\mathit{pl}(f, \alpha, \psi) = -l_{\mathbf{Y}}(f, \alpha, \psi) + \lambda \sum_{k=1}^{p} \tau_k^{-1} ||P^k f||_k^2, \tag{8}$$

where $-l_{\mathbf{Y}}(f, \alpha, \psi)$ measures the goodness-of-fit, and the second term applies a penalty to functions f. The smoothing parameters λ/τ_k control the balance between goodness-of-fit and 'roughness'. Let $\phi_1(\mathbf{t}), \ldots, \phi_m(\mathbf{t})$ be a basis of $\mathcal{M}^0$, R_k be the *Reproducing Kernel* (RK) of $\mathcal{M}^k$, and $R(\mathbf{s}, \mathbf{t}) = \sum_{k=1}^{p} \tau_k R_k(\mathbf{s}, \mathbf{t})$ be the RK of $\sum_{k=1}^{p} \mathcal{M}^k$ under a norm that depends on the τ_k's.

It can be shown that the function that minimizes (8) has the form

$$\begin{aligned} f(\mathbf{t}) &= (\phi_1(\mathbf{t}), \ldots, \phi_m(\mathbf{t}))\mathbf{d} + \\ & \quad (R(\mathbf{t}_1, \mathbf{t}), R(\mathbf{t}_2, \mathbf{t}), \ldots, R(\mathbf{t}_n, \mathbf{t}))\mathbf{c}. \end{aligned} \tag{9}$$

Thus estimation of f can reduced to estimation of two vectors $\mathbf{c}$ and $\mathbf{d}$. From now on we will consider the log-likelihood and the penalized likelihood functions as functions of $\mathbf{c}$ and $\mathbf{d}$ instead of f. Furthermore, we will drop the dependence on the parameters and write pl for $\mathit{pl}(\mathbf{d}, \mathbf{c}, \alpha, \psi)$, and $l_{\mathbf{Y}|\mathbf{U}}$ for $l_{\mathbf{Y}|\mathbf{U}}(\mathbf{d}, \mathbf{c}, \alpha)$ whenever the meaning is clear.

3.2 Estimation of c and d

Since pl is complicated in general, $\mathbf{c}$ and $\mathbf{d}$ can not be solved directly. For fixed values of α, ψ, and λ, we use the Fisher scoring procedure to calculate $\mathbf{c}$ and $\mathbf{d}$. Let $\mathbf{T} = \{\phi_j(\mathbf{t}_i)\}_{i=1}^{n}{}_{j=1}^{m}$, $\mathbf{\Sigma} = \{R(\mathbf{t}_j, \mathbf{t}_i)\}_{i=1}^{n}{}_{j=1}^{n}$, $\mu_i = E[Y_i|\mathbf{U}]$, $w_i = (b''(\theta_i)g'(\mu_i)^2)^{-1}$, and $\mathbf{W} = \mathit{diag}(w_1/a_1(\alpha), \ldots, w_n/a_n(\alpha))$. Note that μ_i and w_i are random since they depend on $\mathbf{U}$. It is not difficult to check that

$$\begin{aligned} \partial \mathit{pl}/\partial \mathbf{d} &= -E_{\mathbf{U}|\mathbf{Y}}\left[\frac{\partial l_{\mathbf{Y}|\mathbf{U}}}{\partial \mathbf{d}} \right], \\ \partial \mathit{pl}/\partial \mathbf{c} &= -E_{\mathbf{U}|\mathbf{Y}}\left[\frac{\partial l_{\mathbf{Y}|\mathbf{U}}}{\partial \mathbf{c}} \right] + n\lambda \mathbf{\Sigma} \mathbf{c}, \\ E_{\mathbf{Y}|\mathbf{U}}\left[\partial^2 l_{\mathbf{Y}|\mathbf{U}}/\partial \mathbf{d} \partial \mathbf{d}^T \right] &= -\mathbf{T}^T \mathbf{W} \mathbf{T}, \\ E_{\mathbf{Y}|\mathbf{U}}\left[\partial^2 l_{\mathbf{Y}|\mathbf{U}}/\partial \mathbf{d} \partial \mathbf{c}^T \right] &= -\mathbf{T}^T \mathbf{W} \mathbf{\Sigma}, \\ E_{\mathbf{Y}|\mathbf{U}}\left[\partial^2 l_{\mathbf{Y}|\mathbf{U}}/\partial \mathbf{c} \partial \mathbf{d}^T \right] &= -\mathbf{\Sigma} \mathbf{W} \mathbf{T}, \\ E_{\mathbf{Y}|\mathbf{U}}\left[\partial^2 l_{\mathbf{Y}|\mathbf{U}}/\partial \mathbf{c} \partial \mathbf{c}^T \right] &= -\mathbf{\Sigma} \mathbf{W} \mathbf{\Sigma}. \end{aligned}$$

Define

$$\tilde{\mathbf{W}} = E_{\mathbf{U}|\mathbf{Y}}\left[\mathbf{W} \right], \tag{10}$$

and $\tilde{\mathbf{y}}$ equals

$$E_{\mathbf{U}|\mathbf{Y}}\left[\mathbf{W}((y_1 - \mu_1)g'(\mu_1), \ldots, (y_n - \mu_n)g'(\mu_n))^T \right]. \tag{11}$$

The Fisher scoring equation at iteration $k+1$ is

$$\begin{aligned} & \begin{bmatrix} \mathbf{\Sigma} \tilde{\mathbf{W}} \mathbf{\Sigma} + n\lambda \mathbf{\Sigma} & \mathbf{\Sigma} \tilde{\mathbf{W}} \mathbf{T} \\ \mathbf{T}^T \tilde{\mathbf{W}} \mathbf{\Sigma} & \mathbf{T}^T \tilde{\mathbf{W}} \mathbf{T} \end{bmatrix} \begin{bmatrix} \mathbf{c}^{(\mathbf{k+1})} - \mathbf{c}^{(\mathbf{k})} \\ \mathbf{d}^{(\mathbf{k+1})} - \mathbf{d}^{(\mathbf{k})} \end{bmatrix} = \\ & \qquad \begin{bmatrix} \mathbf{\Sigma} \tilde{\mathbf{y}} - n\lambda \mathbf{\Sigma} \mathbf{c}^{(\mathbf{k})} \\ \mathbf{T}^T \tilde{\mathbf{y}} \end{bmatrix}, \end{aligned} \tag{12}$$

where $\mathbf{c}^{(\mathbf{k})}$ and $\mathbf{d}^{(\mathbf{k})}$ refer to the solutions at the kth iteration. $\mathbf{Td} + \mathbf{\Sigma c}$ is unique if $\mathbf{T}$ is of full column rank. It can be shown that a solution to (13) is also a solution to (12):

$$\begin{aligned} \tilde{\mathbf{W}} \mathbf{T} \mathbf{d}^{(\mathbf{k+1})} + (\tilde{\mathbf{W}} \mathbf{\Sigma} + n\lambda \mathbf{I}) \mathbf{c}^{(\mathbf{k+1})} & \\ &= \tilde{\mathbf{y}} + \tilde{\mathbf{W}}(\mathbf{T}^T \mathbf{d}^{(\mathbf{k})} + \mathbf{\Sigma} \mathbf{c}^{(\mathbf{k})}) \\ \mathbf{T}^T \mathbf{c}^{(\mathbf{k+1})} &= 0. \end{aligned} \tag{13}$$

With the transformations $\tilde{\mathbf{Q}} = \tilde{\mathbf{W}}^{1/2}\boldsymbol{\Sigma}\tilde{\mathbf{W}}^{1/2}$, $\tilde{\mathbf{T}} = \tilde{\mathbf{W}}^{1/2}\mathbf{T}$, $\tilde{\mathbf{z}} = \tilde{\mathbf{W}}^{-1/2}(\tilde{\mathbf{W}}\mathbf{T}^T\mathbf{d}^{(\mathbf{k})} + \tilde{\mathbf{W}}\boldsymbol{\Sigma}\mathbf{c}^{(\mathbf{k})}) + \tilde{\mathbf{W}}^{-1/2}\tilde{\mathbf{y}}$, and $\tilde{\mathbf{c}}^{(\mathbf{k+1})} = \tilde{\mathbf{W}}^{-1/2}\mathbf{c}^{(\mathbf{k+1})}$, (13) becomes

$$\begin{aligned} (\tilde{\mathbf{Q}} + n\lambda\mathbf{I})\tilde{\mathbf{c}}^{(\mathbf{k+1})} + \tilde{\mathbf{T}}\mathbf{d}^{(\mathbf{k+1})} &= \tilde{\mathbf{z}}, \\ \tilde{\mathbf{T}}^T\tilde{\mathbf{c}}^{(\mathbf{k+1})} &= 0. \end{aligned} \tag{14}$$

RKPACK (Gu 1989) can be used to solve equations (14) and estimate the smoothing parameters.

3.3 Estimation of α and ψ

To estimate α we set the first derivative of pl with respect to α equal to zero,

$$\begin{aligned} \frac{\partial pl}{\partial \alpha} &= -E_{\mathbf{U}|\mathbf{Y}}\left[\frac{\partial log f_{\mathbf{Y}|\mathbf{U}}(\mathbf{U};\alpha)}{\partial \alpha}\right] \\ &= \sum_{i=1}^{n} \frac{\partial a_i(\alpha)}{\partial \alpha} \frac{y_i E_{\mathbf{U}|\mathbf{Y}}[\theta_i] - E_{\mathbf{U}|\mathbf{Y}}[b(\theta_i)]}{[a_i(\alpha)]^2} \\ &+ \frac{\partial K_i(\alpha, y_i)}{\partial \alpha} = 0. \end{aligned} \tag{15}$$

To estimate ψ we set the first derivative of pl with respect ψ equal to zero,

$$\begin{aligned} \frac{\partial pl}{\partial \psi} &= \frac{\partial}{\partial \psi}\left\{-log f_{\mathbf{Y}}(\mathbf{Y};\psi) + n\lambda\mathbf{c}^T\boldsymbol{\Sigma}\mathbf{c}\right\} \\ &= -E_{\mathbf{U}|\mathbf{Y}}\left[\frac{\partial log f_{\mathbf{U}}(\mathbf{U};\psi)}{\partial \psi}\right] = 0. \end{aligned} \tag{16}$$

Any routine for solving a system of non-linear equations can be used to solve (15) and (16).

3.4 Markov Chain Monte Carlo (MCMC) Method

In order to solve the system (14) the matrix $\tilde{\mathbf{W}}$ and the vector $\tilde{\mathbf{y}}$ have to be computed. From definitions of $\tilde{\mathbf{W}}$ and $\tilde{\mathbf{y}}$ in (10) and (11) we need to calculate q-dimensional integrals. This is not feasible in most cases since q is usually large in our models. Therefore, we calculate $\tilde{\mathbf{W}}$ and $\tilde{\mathbf{y}}$ using MCMC as follows: first, we produce random draws, $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(R)}$, from the conditional distribution of $\mathbf{U}|\mathbf{Y}$ using a Metropolis algorithm (Gilks et al. 1996), then we replace $\tilde{\mathbf{W}}$ and $\tilde{\mathbf{y}}$ by their Monte Carlo approximations. See the algorithm in Section 3.5 for details. Since the dimension q of the random effects can be fairly large, we choose a single component Metropolis-Hastings method. This avoids small acceptance rates of the sampler. Other sampling methods are possible and may work better in some cases. See Gilks et al. (1996) for advantages and disadvantages of the various sampling methods.

To produce sample points from $f_{\mathbf{U}|\mathbf{Y}}$ we do the following:

> After some burn-in iterations let $\mathbf{u}^{(t)}$ be the sampled vector of our Markov Chain at iteration t. Each of the q components in $\mathbf{u}^{(t)}$ has to be updated. Denote $\mathbf{u}^* = (u_{t+1}^1, \ldots, u_{t+1}^{i-1}, x, u_t^{i+1}, \ldots, u_t^q)$, where u_{t+1}^j for $j = 1, \ldots, i-1$ are the values that are already updated, and u_t^j for $j = i+1, \ldots, q$ are the corresponding components of $\mathbf{u}^{(t)}$. x is the proposed value for ith component sampled from the proposal distribution $q_i(\cdot|\tilde{\mathbf{u}})$ where $\tilde{\mathbf{u}} = (u_{t+1}^1, \ldots, u_{t+1}^{i-1}, u_t^i, u_t^{i+1}, \ldots, u_t^q)$. Then the ith component is updated by either accepting or rejecting x. If x is rejected, the ith component of $\mathbf{u}^*$, u_{t+1}^i, will be set to the old value u_t^i. Otherwise u_{t+1}^i will be set to x. The probability to accept x is
>
> $$\min\left\{\frac{f_{\mathbf{U}|\mathbf{Y}}(\mathbf{u}^*)q_i(u_i^t|\mathbf{u}^*)}{f_{\mathbf{U}|\mathbf{Y}}(\tilde{\mathbf{u}})q_i(x|\tilde{\mathbf{u}})}\right\}.$$
>
> After all components have been updated in $\mathbf{u}^*$, we set $\mathbf{u}^{(t+1)}$ to $\mathbf{u}^*$.

The choice of the proposal distribution could be fairly arbitrary in theory. In practice, however, it should be close to the marginal distribution of $U_i|\mathbf{Y}$. In our implementations we choose $q_i(\cdot|\cdot)$ to be the marginal of the random effects, $f_{U_i}(u)$.

3.5 The Algorithm

i) Choose initial values for $\mathbf{c}^{(0)}$,$\mathbf{d}^{(0)}$, $\alpha^{(0)}$, $\psi^{(0)}$. Set $k = 0$.

ii) Generate R values $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(R)}$ from the distribution $f_{\mathbf{U}|\mathbf{Y}}(\mathbf{u}; \mathbf{c}^{(k)}, \mathbf{d}^{(k)}, \alpha^{(k)}, \psi^{(k)})$ using a Metropolis-Hastings algorithm as described in section 3.4. Let $\mu^{(l)} = g^{-1}(\mathbf{Td}^{(k)} + \Sigma\mathbf{c}^{(k)} + \mathbf{Zu}^{(l)})$ and $\mu^{(l)} = (\mu_1^{(l)}, \ldots, \mu_n^{(l)})$ for $l = 1, \ldots, R$.

iii) Solve the estimating equations (14) for $\mathbf{c}^{(k+1)}$ and $\mathbf{d}^{(k+1)}$ where the iith element of the matrix $\tilde{\mathbf{W}}$ is replaced by its estimate $\frac{1}{R}\sum_{l=1}^{R} 1/\left[b''(\theta_i(\mu_i^{(l)}))g'(\mu_i^{(l)})^2 a_i(\alpha)\right]$ and the ith element of vector $\tilde{\mathbf{y}}$ is replaced by its estimate $\frac{1}{R}\sum_{l=1}^{R}(y_i - \mu_i^{(l)})/(b''(\theta_i(\mu_i^{(l)}))g'(\mu_i^{(l)}))$.

iv) Derive $\alpha^{(k+1)}$ to be the solution to (15) with $E_{\mathbf{U}|\mathbf{Y}}[\theta_i] \approx \frac{1}{R}\sum_{l=1}^{R}\theta_i(\mu_i^{(l)})$and $E_{\mathbf{U}|\mathbf{Y}}[b(\theta_i)] \approx \frac{1}{R}\sum_{l=1}^{R} b(\theta_i(\mu_i^{(l)}))$.

v) Derive $\psi^{(k+1)}$ to be the solution to (16).

vi) If estimates have not converged yet then set $k := k+1$ and return to ii).

4 Application

To demonstrate our method, data on the feeding behavior of hummingbirds from Graham & Petkau (1994) are considered. The data were down-loaded from StatLib. In this experiment 18 hummingbirds were randomly assigned to one of the following three treatment groups: full tape (FT), partial tape (PT), and no tape (NT). The birds in the FT group were put in a feeding room in which a light clue was connected with a corresponding feeder by a continuous fluorescent tape. For the PT group the tape was discontinuous, and for the NT group there was no tape at all. Success or failure of feeding was recorded. Each bird repeated the experiment 180 times under the same treatment.

Our main purpose here is to demonstrate our method rather than providing a thorough analysis of the data.

As mention in section 2, in this section we will demonstrate how to decompose the conditional predictor using averaging operators. We have three covariates: treatment group, birds, and time. We consider time and treatment group as fixed effects. Since birds were randomly sampled from a population we consider bird effects as random. Notice that birds are nested within treatment groups. Denote by $\mathcal{B}_g$ the population of birds receiving treatment g and by P_g the sampling distribution. Define the conditional predictor as

$$\eta(g,b,t) = logit(P(Y=1|g,b,t)).$$

The domain for the function $\eta(g,b,t_j)$ is

$$\{\{1\}\otimes\mathcal{B}_1, \{2\}\otimes\mathcal{B}_2, \{3\}\otimes\mathcal{B}_3\}\otimes[0,1].$$

We want to model the time effect using a cubic spline using Sobolev space W_2

$$W_2 = \{\eta : \eta^{(\nu)} \text{ absolutely continuous }, \nu = 0,1, \int_0^1 (\eta^{(2)}(t))^2 dt < \infty\}. \quad (17)$$

We want to use one way ANOVA models for treatment and bird. To derive the SS ANOVA decompositions, we define the following averaging operators as

$$\begin{aligned} A_2\eta(g,t_j) &= \int_{\mathcal{B}_g} \eta(t,g,b)dP_g(b), &(18)\\ A_1\eta(t) &= \sum_{g=1}^{3} A_2\eta(t,g)/3, &(19)\\ A_3\eta(g,b) &= \int_0^1 \eta(t,g,b)dt, &(20)\\ A_4\eta(b,t) &= [\int_0^1 \eta'(t,g,b)dt](t-0.5). &(21)\end{aligned}$$

We can decompose η into

$$\begin{aligned} \eta \;=\;& [A_1 + (A_2 - A_1) + (I - A_2)]\\ & [A_3 + A_4 + (I - A_3 - A_4)]\eta\\ =\;& A_1A_3\eta + A_1A_4\eta + A_1(I - A_3 - A_4)\eta\\ +\;& (A_2 - A_1)A_3\eta + (A_2 - A_1)A_4\eta\\ +\;& (A_2 - A_1)(I - A_3 - A_4)\eta\\ +\;& (I - A_2)A_3\eta + (I - A_2)A_4\eta\\ +\;& (I - A_2)(I - A_3 - A_4)\eta\\ =\;& \mu + \alpha(t - .5) + s_1(t)\\ +\;& \beta_g + \gamma_g(t - .5) + s_2(g,t)\\ +\;& \delta_b + \zeta_b(t - .5) + s_3(b,g,t). \end{aligned} \quad (22)$$

Since birds are random samples, all terms involving b are random. Therefore the first six terms in (22) are fixed and the last three are random. The first three terms represent the population curve. The next three terms measure the departure of a treatment group from the population curve. The last three terms measure the departure of a bird from the population curve for a specific treatment group. We may fit several nested models based on The decomposition (22) comprises several nested models. We consider the following here:

Model A contains the first eight terms in (22).

$$\begin{aligned} Y_{gbj}|\mathbf{U} &\sim \text{Bernoulli}(p_{gbj})\\ logit(p_{gbj}) &= \mu + \alpha(t_j - .5) + s_1(t_j) + \beta_g + \gamma_g(t_j - .5)\\ &+ s_2(g,t_j) + \delta_b + \zeta_b(t_j - .5), \end{aligned} \quad (23)$$

where $\mathbf{U}^T = (\delta_b, \zeta_b) \sim N((0,0), diag(\sigma_1^2, \sigma_2^2))$. In model A each group has its own conditional predictor curve $\mu + \alpha(t_j - .5) + s_1(t_j) + \beta_g + \gamma_g(t_j - .5) + s_2(g,t_j) \in W_2$, and for each bird the intercept and slope deviate from the population mean values by δ_b and ζ_b, respectively.

To estimate the parameters we implemented Algorithm 3.5 using Fortran. In Figure 2 the components of the SS ANOVA decomposition are shown.

The left plot is the estimate for $s_1(t_j)$, the middle plot is the estimate for $s_2(g,t_j)$, and right plot is the estimate for $\gamma_g(t_j - .5)$. The solid line corresponds to treatment NT, the broken line corresponds to treatment PT, and the dotted line corresponds to treatment FT. The scales of the right plot is ten times the scale of the left two plots, indicating that both smooth terms do not have much impact. A neglectable $s_2(g,t_j)$ indicates small interaction between time and treatment. In Figure 3, we have the overall estimate of the conditional predictor which is the sum of all components.

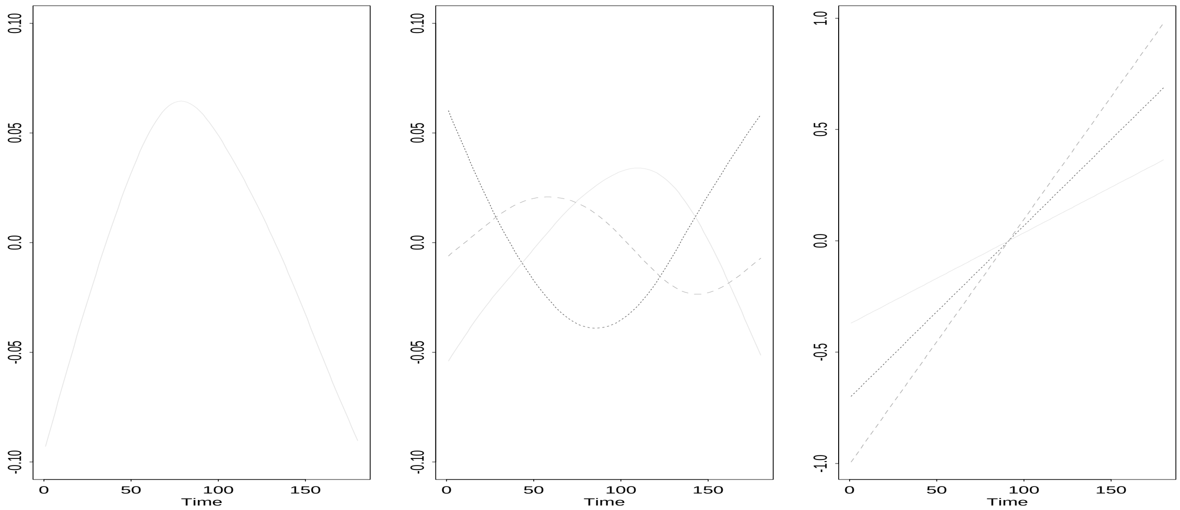

Figure 2: Plots of several estimated components in our model. Left: $s_1(t)$. Middle: $s_2(g,t)$. Right: $\gamma_g(t-.5)$. For the right two plots the solid line is treatment NT, the broken lines are treatment PT, the dotted lines are treatment FT.

We see all mean responses increase with time. The FT group does better than the PT group. The taped groups do worse than the NT group at the beginning, but catch up at the end.

The conclusion is that the linear model with without $s_1(t_j)$ and $s_2(g,t_j)$ appears to be appropiate. This example shows that SS ANOVA decomposition provides a tool to check linear assumptions. Next we want to show that our model can catch nonlinear trend when it does exist.

To illustrate this we generate data based on the experimental design of the hummingbird data. We have 18 birds nested within three treatments. For each bird we generate 30 observations with the following conditional predictor,

$$\begin{aligned}\eta(1,b,t) &= 2+log(10t+.2)+\delta_b+\zeta_b(t-.5),\\ \eta(2,b,t) &= -3+4t^4+\delta_b+\zeta_b(t-.5),\\ \eta(3,b,t) &= -1.2+3.5t+3sin(6.28t)+\delta_b+\zeta_b(t-.5),\end{aligned}$$

where $(\delta_b,\zeta_b) \sim N((0,0)^T, diag(.05^2,.01^2))$. Figure 4 shows the plots of the true curves and their estimates. This plot shows that our model catch more complicated structures.

5 Conclusion

We have proposed a very general class of models GNMM which generalize many existing models such as GLMM, NMR, and GNM. The Smoothing Spline ANOVA decomposition of the conditional predictor has the usual interpretation of main effects and interactions. The model can be applied to wide a range of data such as growth curves, repeated measures, spatial-temporal and longitudinal data. The algorithm for the estimation of the parameters has been implemented using Fortran. Open is how to construct confidence intervals for the parameters and confidence bands for nonparametric components.

Acknowledgements

This research was supported by NIH grant R01 GM58533.

References

Aronszajn, N. (1950), 'The theory of reproducing kernel', *Trans. Amer. Math. Soc.* **68**, 337–404.

Breslow, N. E. & Clayton, D. G. (1993), 'Approximate inference in generalized linear mixed models', *Journal of the American Statistical Association* **88**, 9–25.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. E. (1996), *Markov Chain Monte Carlo in Practice*, Chapman Hall:London.

Graham, J. & Petkau, A. J. (1994), Spatial association learning in hummingbirds, *in* 'Case Studies Biometry', pp. 159– 187.

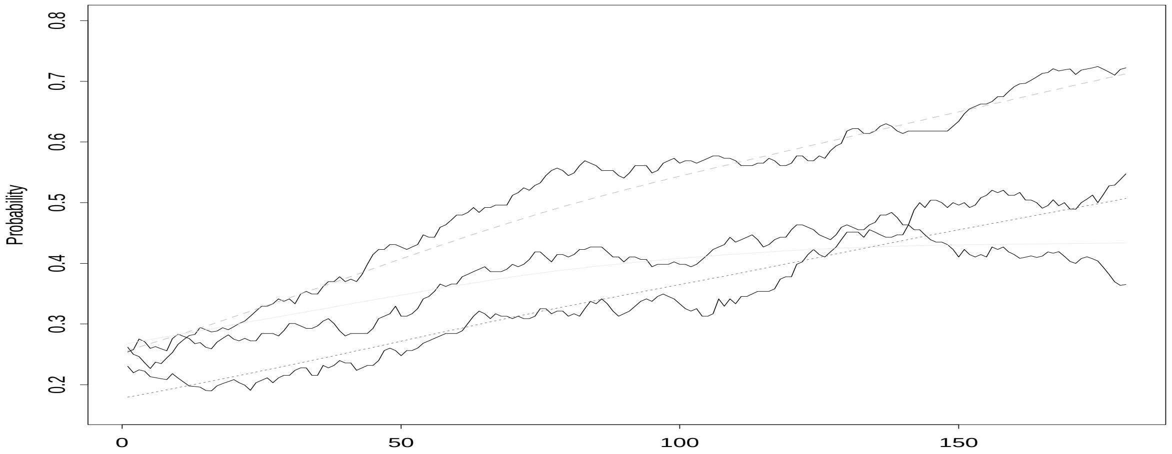

Figure 3: Estimated conditional predictor for three treatment groups. Solid line is treatment NT, broken line is treatment PT, dotted line is treatment FT.

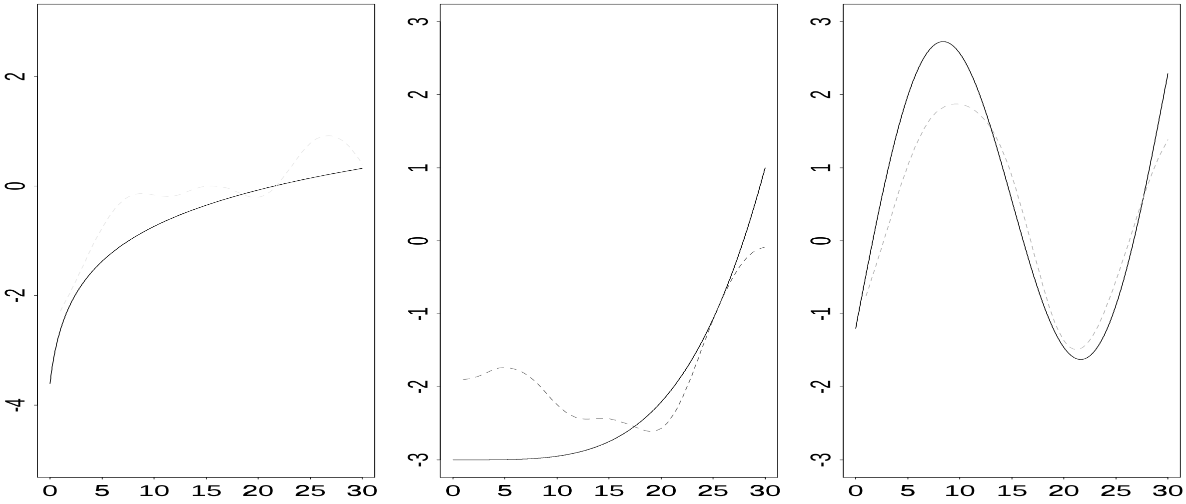

Figure 4: Comparison of true curve and estimated curves Solid lines are the true curves. Left plot shows treatment 1, middle plot shows treatment 2, and right plot shows treatment 3.

Gu, C. (1989), Rkpack and its applications: Fitting smoothing spline models, *in* 'ASAProStCp', pp. 42–51.

Gu, C. & Wahba, G. (1993), 'Semiparametric analysis of variance with tensor product thin plate splines', *Journal of the Royal Statistical Society B* **55**, 353–368.

Jiang, J. (1998), 'Consistent estimators in generalized linear mixed models', *Journal of the American Statistical Association* **93**, 720–729.

Lin, X. & Breslow, N. E. (1996), 'Bias correction in generalized linear mixed models with multiple components of dispersion', *Journal of the American Statistical Association* **91**, 1007–1016.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models (Second Edition)*, ChapmnHall:Lnd.

Robinson, G. K. (1991), 'That Blup is a good thing: The estimation of random effects (disc: P32-51)', *StatSci* **6**, 15–32.

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM:PA.

Wahba, G. & Wang, Y. (1998), 'Discussion of "smoothing spline models for the analysis of nested and crossed samples of curves" by brumback and rice', *Journal of the American Statistical Association* **93**, 976–980.

Wahba, G., Wang, Y., Gu, C., Klein, R. & Klein, B. (1995), 'Smoothing spline Anova for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy', *Annals of Statistics* **23**, 1865–1895.

Wang, Y. (1998), 'Mixed-effects smoothing spline anova', *Journal of the Royal Statistical Society B,* **60**, 159–174.

Zhang & Lin (1999), 'Inference in generalized additive mixed models by using smoothing splines', *Journal of the Royal Statistical Society B* **61**, 381–400.