Bayesian Analysis for Longitudinal Semicontinuous Data

Pulak Ghosh[*] and Paul S. Albert [†]

April 10, 2007

Abstract

In many biomedical applications, researchers encounter semicontinuous data whereby data are either continuous or zero. When the data are collected over time the observations are correlated. Analysis of these kind of longitudinal semi-continuous data is challenging due to the presence of strong skewness in the data. In this paper, we develop a flexible class of zero-inflated models in a longitudinal setting. Improving on a likelihood-based approach (using Monte-Carlo EM) that was proposed by Albert and Shen (2005), we use a Bayesian approach to analyze longitudinal data from a acupuncture clinical trial in which we compare the effects of active acupuncture, sham acupuncture and standard medical care on chemotherapy-induced nausea in patients being treated for advanced stage breast cancer. A penalized spline model is introduced into the linear predictor of the model to explore the possibility of nonlinear treatment effect. The subject-specific effects involved in the model are assumed to follow a nonparametric Dirichlet process (DP) mixture. We also account

[*] Department of Mathematics and Statistics, Georgia State University, Atlanta, GA, 30303-3083, USA; Email: pghosh@mathstat.gsu.edu

[†] Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute

for possible serial correlation between successive observations using Brownian motion. Thus, the approach taken in this paper provides for a more flexible modeling framework and, with the use of WinBUGS, provides for a computationally simpler approach than direct maximum-likelihood. We illustrate the Bayesian methodology with the acupuncture clinical trial data analyzed by Albert and Shen (2005).

Key words: Acupunture; Dirichlet Process mixture; Emesis; Penalized-spline; Two-Part model; WinBUGS

1 Introduction

In many biomedical applications, a longitudinal response variable may take continuous distribution with a large number of values clustered at zero. These kind of longitudinal models with clumping at zero are referred to as two-part model by Lachenbruch (2002) and semicontinuous model by Olsen and Schafer (2001). Data structure in a two-part model is quite different from one that has been left-censored or truncated, because the zero's represent actual response values. For example, in the data set we consider, at some time points patients experience large amount of vomitting and at some other time points there is no vomitting at all (i.e., with zero vomitting). Treating these kind of data using a normal distribution is not suitable since ignoring the many zeros, especially when a sizeable proportion of the data is zero, implies that the underlying parametric distributional assumptions will not be met. This type of data may also be positively skewed for the nonzero values. So in a two-part model, the data can be considered as a mixture of 0's and highly skewed, continuously distributed, positive values (Robinson et al., 2006). Hence, in a two-part model, zeros should be analyzed separately from the nonzero continuous data.

The two-part model which originated in econometrics (Heckman, 1976; Duan et. al. 1983) is based on two equation. One equation (logistic model) is used to predict the probability of occurrence of a nonzero value, and a second equation (linear model) is used to predict the amount of nonzero values. Recently, Zhou and Tu (1999) and Tu and Zhou (1999) have proposed testing procedures for comparing different populations on the basis of a two-part model. More recently, Welsch and Zhou (2006) proposed methodology which allows for flexibly modeling the continuous part of the data. However, majority of the literature in the area addresses the cross-sectional case whereby only a single observation is measured on each individual. Excess zeros may also occur with longitudinal data and in that scenario the correlation among measurements on the same individual must be accounted for. Olsen et al. (2001) and Tooze et. al. (2002) have extended a two-part regression model to include random effects in both the logistic and linear stages of the model to capture unexplained heterogeneity among individuals in a longitudinal data. Recently Albert et al. (2005) developed a longitudinal two-part model with a Gaussian process to account for the serial correlation . Lu et al. (2004) and Albert (2005) discuss an estimating equations approach for two-part model with application to clustered data, and Li et al. (2005) introduced measurement error model in semicontinuous longitudinal data. While a majority of these approaches are based on maximum likelihood estimation, Zhang et al. (2006) developed a Bayesian two-part model to analyze health care data. Robinson et al. (2006) has developed a hierarchical Bayesian approach to analyze a multivariate two-part model. However, all these models are heavily dependent on the normal distribution and uses a parametric mixed model. Thus, the resulting estimates of the parameters may be compromised if the assumption of the parametric distribution is violated.

Shen et al (2000) presented the results of a clinical trial where daily emesis volume was collected longitudinally over a two week period for breast cancer patients being treated with a standard chemotherapy. 104 patients were randomized between three groups: (i) patients treated with standard chemotherapy only (34 patients), (ii) patients treated with sham acupuncture (33 patients), and (iii) patients treated with active acupuncture (37 patients). An important endpoint for this study was the daily measurement of the volume of emesis over a 14- day follow up period. Figure 1 presents a histogram of daily emesis volume of all patients combined, showing the presence of many 0s, and demonstrating that there are many days in which patients have no emesis. Interest primarily focused on comparing the longitudinal course of patients treated with sham acupuncture with those treated with active acupuncture. Albert et al (2005) used this dataset as motivation for a direct likelihood approach for fitting two-part models for longitudinal data. Their model assumed a standard linear model for incorporating time effects on both the probability of a positive volume and the mean volume given a positive volume. Further, they assumed that longitudinal dependence between measurements were described by a stochastic process which accounted for serial as well as between-subject dependence. Incorporating serial dependence (which was important to do in this application) made estimation difficult, requiring Monte-Carlo EM for parameter estimation. Further, standard errors were estimated using the bootstrap, making inference particularly computational.

In this article, we study the two-part model under both parametric and nonparametric Bayesian framework to analyze the data in Shen et al. (2000). Although we build our model based on Albert et al. (2005), our model offers an extension of the existing methods on several issues. *First,* a penalized spline (Ruppert et al., 2003) model is introduced in both the logistic and linear predictor of the model to flexibly model the interaction of treatment over time. Penalized spline models have a simple mixed model representation which allows the entire model to be cast within the mixed model framework. Figure 2 shows the nonlinearity of the longitudinal trajectory of the mean emesis volume for patients over the weeks for each treatment. The kind of semiparametric model developed is useful to explore the nonlinear effect of time and treatment on the responses. *Second,* we add more flexibility to the model by incorporating a Brownian motion to account for the serial correlation in the data. *Third,* we relax the distributional assumption of the random effect and propose a flexible random effect using Dirichlet Process (DP) mixture rather than assume the effect as Gaussian. The DP has the advantage of not assuming any parametric form of the distribution. Instead the data determine the shape of the random effects distribution. DP mixture models (Fergusion, 1973; Antoniak, 1974, Escober and West, 1995; Kleinman and Ibrahim, 1998), by now have an extensive literature in Bayesian analysis and have been established to provide a rather broad and flexible class of distributions. To our knowledge, none of the previous approaches considered DP mixture in the settings considered here. Another important contribution of our model is the easy and simple implementation in the freely available software `WinBUGS` (Spiegelhalter, 2005). This is incontrast to a likelihood approach where Monte-Carlo EM (see McCulloch (1997) for details) requires software development and is very computationally intensive.

In the next section we introduce the model and discuss the parametric Bayesian method. In section 3 we relax the normality assumption and develop a nonparametric method. This nonparametric method is particulary useful when the normality assumption is at stake. In Section 4 we illustrate the application of the proposed method in a real data set followed by discussion and further research in section 5.

2 Parametric Bayesian Model for Longitudinal Two-part Model

In this section a two-part model for semicontinuous longitudinal data is introduced. We develop our two-part model based on Albert et al., (2005).

Let y_{ij} be the volume of emesis for subject i $(i =, 2, \cdots, n)$ at week j $(j = 1, 2, \cdots, m)$, where n is the total number of subjects and m is the total number of follow-up times. For the acupuncture clinical trial, $n = 104$ and $m = 14$. Let R_{ij} be a random variable denoting the volume of daily emesis where,

$$R_{ij} = \begin{cases} 0, & \text{if } y_{ij} = 0 \\ 1, & \text{if } y_{ij} > 0 \end{cases} \tag{1}$$

with conditional probabilities

$$\Pr(R_{ij} = r_{ij} | \boldsymbol{\theta}) = \begin{cases} 1 - p_{ij}(\boldsymbol{\theta}_1), & \text{if } r_{ij} = 0 \\ p_{ij}(\boldsymbol{\theta}_1), & \text{if } r_{ij} = 1 \end{cases} \tag{2}$$

where $\boldsymbol{\theta}_1$ is a vector of parameters.

Let, $S_{ij} \equiv [y_{ij} | R_{ij} = 1]$, denote the mean transformed positive emesis volume for the ith subject at jth week with p.d.f. $f(s_{ij}|\theta_2)$ where $f(s_{ij})$ may be any distribution with $y_{ij} > 0$. We assume a continuous normal distribution for the log-transformed volumes ($\log(Y+1)$).

Stage One: We assume the following model for the two-part model:

$$\begin{aligned} \text{logit}(p_{ij}|\theta_1) &= X_{1ij}^T\beta_p + f^p(t_{ij}) + g_{c_i}^p(t_{ij}) + Z_{1ij}^T\mathbf{b}_{1i} + W_{ijp} && (3) \\ f(s_{ij}|\theta_2) &= X_{2ij}^T\beta_s + f^s(t_{ij}) + g_{c_i}^s(t_{ij}) + Z_{2ij}^T\mathbf{b}_{2i} + W_{ijs} + e_{ij} && (4) \end{aligned}$$

The logistic regression (3) models the probability of a positive volume and the linear mixed model (4) models the mean transformed positive emesis volume. Here, β_k $(k = p, s)$ are q_k vector of regression coefficients. The nonparametric function of time $f^k(t_{ij})$ is the spline model. In order to capture the nonlinear trajectory of the different treatments over time, we might be interested in fitting a separate mean curve for subjects receiving each treatment. To do that we use an arbitrary smooth function $g_{c_i}^k(t_{ij})$ to model interaction in which a categorical factor (treatment) interacts with a continuous predictor (Coull et al., 2001; Durban et al., 2004; Ruppert et al., 2003; Crainiceanu et al., 2006). Here, $c_i \in \{1, 2, \cdots, L\}$ represents the treatment group index corresponding to subject i, and $g_1^k, \cdots g_L^k$ are L different functions depending on the values of c_i. Note that $g_{c_i}^k$ are the deviations of the treatment group from the overall curve. The random effects $\mathbf{b}_i = (\mathbf{b}_{1i}, \mathbf{b}_{2i})^T$ accounts for the unobserved heterogeneity among the subjects. A random intercept in the logistic model allows some subjects to have consistently high or low probability of a positive volume, while a random intercept in the lognormal part allows individuals to have a tendency to high or low mean volume given they have a positive volume. To account for the serial correlation in the data (Albert et. al., 2005), we include a stochastic process apart from the usual random effects to flexible model the semicontinous longitudinal data. Thus, the process $W_{ij} = (W_{ijp}, W_{ijs})^T$ is similar to the bivariate stochastic process model involving Brownian motion of Sy et al. (1997). However, because the measurements in our data were taken on a fixed schedule, rather than irregularly as in the study data used by Sy et al. (1997), we instead use a bivariate random walk as the stochastic-process in our model. This

Brownian motion W_{ij} models the local variation and departure from the polynomial trend while the random effects $\mathbf{b}_i$ account for the variability of the trend across the subjects. Brownian motion also adds flexibility to the nonparametric function of time. The measurement error e_{ij} is assumed to have a $N(0, \sigma^2)$ distribution.

Stage two: The second stage of the the model (3-4) defines the distributional assumptions on the random subject effects vector $\boldsymbol{b}_i$ and the bivariate stochastic process W_{ij}. The random walk increments at time 0 are fixed at 0. Thus, we assume,

$$\mathbf{b}_i \sim N\left(\mathbf{0}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \sigma_{22} \end{bmatrix}\right) \tag{5}$$

$$\begin{bmatrix} W_{ijp} \\ W_{ijs} \end{bmatrix} | \begin{bmatrix} W_{i,j-1,p} \\ W_{i,j-1,s} \end{bmatrix} \sim N\left(\begin{bmatrix} W_{i,j-1,p} \\ W_{i,j-1,s} \end{bmatrix}, (t_j - t_{j-1})\Sigma_w\right); \; j = 2, \cdots, m; \tag{6}$$

For the first time point we assume the increments to be zero, i.e., $(W_{i1p} \equiv 0)$, and $(W_{i1s} \equiv 0)$. Although we assume a normal distribution for the random subject effect $\boldsymbol{b}_i$, we relax the distributional assumption in next section.

We estimate the smooth functions $\{f^k(t),\ g^k_{c_i}(t);\ k = p,\, s\}$ by penalized splines. Crainiceanu et al. (2006) outline a strategy for fitting penalized spline model in `WinBUGS`. Thus, following Ruppert et al. (2003) and Crainiceanu et al. (2006),we assume the linear spline estimator of the form

$$\begin{aligned} f^k(t) &= \alpha^k_1 + \alpha^k_2 t + \sum_{d=1}^{D_k} u^k_d (t - \kappa^k_d)_+; \;\; u^k_d \sim N(0, \sigma^2_{ku}) && (7) \\ g^k_{c_i}(t) &= \sum_{l=2}^{L} z^k_{il}(\gamma^k_{0l} + \gamma^k_{il} t) + \sum_{l=1}^{L} z^k_{il} \left\{ \sum_{d=1}^{D_k} v^{k^l}_d (t - \kappa^k_d)_+ \right\} && (8) \\ & \quad v^{k^l}_d \sim N(0, \sigma^2_{kv}); \;\; l = 2, 3, \cdots, L; \;\; k = p, s;. \end{aligned}$$

where $z_{il} = 1$ if $z_{il} = l$ and 0 otherwise for $l = 2, 3, \cdots, L$.

$$(t - \kappa_d^k)_+ = \begin{cases} 0, & t \leq \kappa_d^k \\ (t - \kappa_d^k), & t > \kappa_d^k \end{cases}$$

and $\kappa_1^k, \cdots, \kappa_D^k$ are knots for the $k = p, s$. The choice of the knots κ_d^k's will be described in the data analysis section. We have assumed same variance parameter for each curve, i.e., $N(0, \sigma_{kv}^2)$; $l = 2, 3, \cdots, L$, and the random effects are independent from function to function, i.e., the curves are different but with the same amount of smoothing. In order for the fixed effects to be identified we need to put constrains on α_1^k, γ_{0l}^k, γ_{1l}^k, we assume $\alpha_1^k = \gamma_{01}^k = \gamma_{11}^k = 0$. The specification of the treatment group curves is equivalent to $\sum_{d=1}^{D_k} v_d^{k^1}(t - \kappa_d^k)_+$ for $l = 1$ and $(\gamma_{0l}^k + \gamma_{il}^k t) + \left\{\sum_{d=1}^{D_k} v_d^{k^l}(t - \kappa_d^k)_+\right\}$ for treatment $l = 2, \cdots, L$. This avoids the nonidentifiability of the slope and intercept parameter. A higher degree splines may be used. However, the motivating application did not benefit from this extension.

3 Nonparametric Bayesian Approach

The parametric model described in Section 3 makes normal distribution as a specific distributional assumption for the random effects $\mathbf{b}_i$. The choice of the normal distribution for the random effects, however, is arbitrary. It may well happen that the normal distribution does not correctly fit the data at hand, for example if the data are skewed, contain outliers, or consist of diverse populations. Although, estimation for the fixed effects parameters is relatively robust to misspecification of the random effects, it is becoming widely recognized that inference for random effects can be misleading when normal distributional assumptions do not hold (Verbeke and Lesaffre, 1996). Thus a more general model that relaxes these parametric assumptions may be more desirable for making robust inferences. As pointed out by Gelfand (1999), in a situation where the parametric assumption may be too restrictive, a semiparametric model can be developed by a nonparametric specification of some portions of the model.

This section considers a model that generalizes the normality assumption of the random effects to include an entire class of distributions using a Dirichlet process mixture (DPM) prior (Antoniak, 1974). The aim of this generalization is to protect the inference from bias resulting from the incorrect specification of the random effect distribution. This prior can often be centered on a known distribution making it possible to both use the parametric form when appropriate and to move away from it when its fit is poor.

Thus, we model the distribution of $\boldsymbol{b}_i$ using a DPM prior.

$$\boldsymbol{b_i}|G \sim \mathrm{G} \text{ where } G|\nu, G_0 \sim \mathrm{DP}(\nu, G_0) \tag{9}$$

The basic idea of the above DPM model (9) is to model the uncertainty of the random effects $\boldsymbol{b}_i$ by specifying an unknown (random) distribution function G rather than a known parametric distribution. Since we assume that the form of G is unknown, the uncertainty about G needs to be modeled and thus we model the uncertainty about G by placing a prior distribution on G over all possible probability measures. Such a prior is called a DPM prior and we denote it by $G \sim DP(G_0, \nu)$ where G_0 is the known baseline prior and also the prior expectation of $G(.)$. The positive scalar parameter ν is the concentration parameter and the strength of belief parameter. A large value of ν suggests that G is likely to be close to G_0, and hence yield the results that are similar to those obtained from the parametric model with prior on G_0. On the other hand, a small value of ν implies that G is likely to place most of its probability mass on a few points and then, the distribution of the random effects will behave as a finite mixture of parametric distribution. This set up allows the unknown distribution function G to be nonparamtetrically estimated from the data. Another key feature is the almost sure discreteness of the random measure G that assigns positive probability to common values among all $\boldsymbol{b}_i$'s.

The above representation (9) provides a formal definition of the DPM prior. There are several ways to implement a DPM prior. Recent research has focussed on using the constructive definition of the DPM to produce MCMC algorithms (Sethuraman, 1994). Following Sethuraman's (1994) stick-breaking representation:

$$\begin{aligned} &G(\cdot) = \sum_{r=1}^{\infty} p_r \delta_{Z_r}(\cdot), \quad \text{where } Z_r \overset{\text{iid}}{\sim} G_0(\cdot|\kappa),\ r = 1, \ldots, \text{and} \\ &\text{with } p_1 = V_1, p_r = V_r \prod_{j=1}^{r-1}(1 - V_j),\ r = 2, \ldots,\ \text{ and } V_r \overset{\text{iid}}{\sim} \text{Beta}(1, \nu),\ r \geq 1 \end{aligned} \tag{10}$$

If we truncate the sum in (10) at a large integer R we obtain the models considered in Ishwaran and Zarepour (2002). This reduces $G(\cdot)$ into finite dimensional form as $G_R = \sum_{r=1}^{R} p_r \delta_{Z_r}(\cdot)$. Here Z_r are independent and identically distributed variables with some known distribution $G_0(.)$ and the distribution of $p = (p_1, \ldots, p_R)$ is specified by the stick-breaking construction. Ishwaran and Zarepour (2002) has shown that as $R \to \infty$, truncated DP, $G_R(.)$, converges to a DP as in equation (9). Another advantage here is that since the DP structure is reduced to a finite mixture model by this truncation and a non-conjugate structures can be more easily handled now. The advantage of the approximation is that the model reduces to a finite mixture model and can be fitted using the standard MCMC methods and implemented in `WinBugs` (Speigelhalter, 2005) software. We discuss the choice of R in detail in the Data analysis section.

4 Prior Specification

Let $\theta = (\beta_p, \beta_s, g^p_{c_i}, g^s_{c_i}, \Sigma, \Sigma_w, \sigma^2)$ be the set of parameters for model (3-4). In the Bayesian framework we assume independent priors for these parameters. We assume conditionally conjugate priors that leads to simpler updating schemes in the Markov chain sampling methodology. In particular We assume a normal distribution for the location parameters: $\beta_k \sim N(\beta_{0k}, D_k^{-1})$, $\alpha_2^k \sim N(m_k, \sigma_\alpha^2)$; $k = p, s$. The variance parameters are assumed an inverse gamma prior, i.e., $\sigma^2 \sim \text{IG}(a, b)$, $\sigma^2_{ku} \sim \text{IG}(a_{ku}, b_{ku})$, and $\sigma^2_{kv} \sim \text{IG}(a_{kv}, b_{kv})$; $k = p, s$. Here, $\text{IG}(a, b)$ denotes an inverse gamma distribution with density proportional to $\exp(-b/x)/x^{a+1}$. Note that small values of a, b corresponds to vague prior information. $\Sigma \sim IW(H, \eta_0)$, $\Sigma_W \sim IW(G, \eta_1)$ Here, Wishart(P^{-1}, ν_b) denotes the Wishart distribution with ν_b degrees of freedom and scale matrix P^{-1}.

The full Bayesian model in the present context is completed by assigning prior distribution on the DPM parameters ν and G_0. It is assumed that

$$\nu \sim \text{Gamma}(\nu_1, \nu_2), \quad G_0 \sim N_2(\mathbf{0}, \Delta), \quad \text{and } \Delta^{-1} \sim W_2(\zeta, \Psi) \tag{11}$$

The hyperparameters of all the prior distributions are assumed to be known.

5 Example

We analyze the clinical trial data on acupuncture for treating chemotherapy-induced vomiting in patients being treated for advanced stage breast cancer (Shen et al., 2000; Albert et al., 2005). 104 patients were randomized to one of three treatment groups (active acupuncture, sham acupuncture, and standard medical care) and were followed over a 14-day period. The only subject-specific covariate we consider is age. The data depict a large amount of serial correlation which diminishes and levels off with increasing distance between measurements (Albert and Shen, 2005). To account for this extra correlation we consider the random walk increments described in (5) and (6).

Thus, we consider the following model for analyzing the data:

$$\begin{aligned} \text{logit}(p_{ij}) &= \beta_1^p + \beta_2^p \text{age}_i + f^p(t_{ij}) + g_{c_i}^p(t_{ij}) + b_{i1} + W_{jp} && (12) \\ f(s_{ij}) &= \beta_1^s + \beta_2^s \text{age}_i + f^s(t_{ij}) + g_{c_i}^s(t_{ij}) + b_{i2} + W_{js} + e_{ij} && (13) \end{aligned}$$

There is no clear rule on how many knot points to include or where to locate them in the spline functions. More knots are needed in regions where the function is changing rapidly (Ruppert, 2002). Sometimes subject knowledge may be relevant in placing knots where a change in the shape of the curve is expected. Using too few knots or poorly sited knots means the approximation to the curve will be degraded. By contrast, a spline using too many knots will be imprecise. We select the knots among the existing values and they are equally spaced within the range $[\min(x), \max(x)]$. Thus we assume to have 6 knots and they are placed at time points 2, 4, 6, 8, 10, 12.

Our available data set is not large enough to allow part of it to be used for prior ellicitaion. Prior information based on expert opinion, even if available is user specific. Hence, for our data set we choose prior to be weakly informative, while making sure that the model remain identifiable. For each β in the model we take a $N(0, 1000)$ prior. Similarly for each α component we take $N(0, 1000)$ prior. For the variance parameters we take $IG(2.001, 1.001)$, giving rise to a prior mean of 1 and prior variance of $1,000$. Each of the variance covariance matrices Σ, Σ_W, Δ are assumed a IW (diag$(1, 1), 3)$prior. For the concentration parameter ν of the DPM prior we assume Gamma$(.1, .1)$ distribution. This choice of ν has a prior mean of 1. Note that $\nu = 1$ signifies that the probability of generating a new cluster is $\frac{1}{J+1}$ when we have a sample of size J. To assess the effect of this parameter on the inferences we also consider a Gamma$(2, 0.1)$ for the concentration parameter, and found the results to be very similar.

The posterior distributions are analytically intractable. We use Gibbs sampler (Gelfand and Smith, 1990) to obtain samples from the posterior distribution. Thus computations are done via Monte Carlo approximations with the help of the MCMC method. The methods are implemented in the freely available software packages R (R development Core Team, 2004), and WinBUGS (2005). Our code uses the R package R2WinBUGS (Sturtz et al., 2005) to execute WinBUGS while running a session in R. We ran a chain of $80,000$ iterations with the first $30,000$ discarded as burn-in. Convergence was assessed visually by monitoring the dynamic traces of Gibbs iterations and by computing the Gelman-Rubin (Brooks and Gelman, 1998) convergence statistic. We follow (Ghosh and Rosner, 2007) to check the sensitivity of the inference to the approximation of DPM by varying the upper limit of the number of components in the DPM model. We found that $R = 20$ works pretty well, in the sense that the MCMC puts negligible posterior probability on the number of unique components being larger than some number greater than R and thus increasing R beyond that value does not change the parameter estimates in our example. Figure 3 shows the histogram of the number of components. The initial values for the fixed parameters were selected by starting with the prior mean and covering ± 3 standard deviations. The initial values for the precision were arbitrarily selected.

Table 1 contains the posterior means, posterior standard deviations, and 95% credible intervals for the parameters of interest under the proposed Bayesian parametric and nonparametric methods as discussed before. The covariate age has no significant effect on both the logistic regression as well as on the linear regression. The large significant random intercept variance (Σ^{11}) for the logistic part shows that after accounting for covariate differences among the subjects, some subjects have a greater probability of positive emesis volume than others. The positive random intercept variance (Σ^{22}) shows that for the subjects who are vomiting tend to have a larger mean transformed emesis volume than others. The correlation between the random intercepts of the model (Σ^{12}) is positive implying that the probability of positive volume and mean emesis volume is correlated. However, this correlation increases substantially in the nonparametric model. Another aspect of the extra heterogeneity can be observed in the posterior estimate of ν which is 5.467. It indicates the presence of the sizeable number of distinct components in the distribution of the random subject effects. It also indicates the presence of mixands in the distribution of the random effect and hence evidence of possible multimodality on the distribution.

The estimated Σ_W suggests that there is significant autocorrelation in both the occurrance of a non-zero emesis and in the emesis volume given a positive volume. However, there was little cross-autocorrelation between the two components since 95% credible intervals for Σ_W^{12} included zero for both the parametric and non-parametric models.

Figure 4 presents the nonparametric mean profiles under the parametric and nonparametric model. The top two panels show the mean profiles for the logistic model, the lower two panels show the profiles for the log-normal part for both models. Both the parametric and nonparametric curves are similar in the overall shape, specially for the log-normal part. We focus on the logistic component (probability of a positive emesis volume) since this is where treatment differences appear. For both the parametric and non-parametric models, there is a sizable effect of treatment on the probability of a positive emesis during the first five days after randomization (comparing the dashed and dotted lines) with the maximal treatment effect appearing at 4 days post randomization. Further, the treatment effect appears to be diminished after day 5. Since, acupuncture was only provided over the first 5 days after randomization, the results suggest that active acupuncture may be beneficial over sham acupuncture over the time period in which acupuncture is given. The treatment effect appears to be stronger for the parametric model than the non-parametric model.

Figure 4 also shows the mean profile for the standard medical group (no acupuncture). A comparison of the sham acupuncture (dashed line) with the standard medical arm (solid) line is a measure of placebo effect. As with the treatment effect, placebo effect appears to be substantial over the first 5 days after randomization and to diminish after day 5. Also, the placebo effect over days 1-5 appears to be slightly larger for the nonparametric model than the parametric model.

To gain further insight into the differences between the treatments, specially between the sham acupuncture and acute acupuncture we find the absolute difference between the nonparametric curves in the logistic model and the corresponding posterior probability. Let $g_a(t)$ and $g_s(t)$ be the be the two function corresponding to acute acupuncture and sham acupuncture. Then, we define the difference between the function as,

$$\Lambda(t) = \int_{t_1}^{t_{14}} \frac{|g_a(t) - g_s(t)|}{t_{14} - t_1} dt$$

where, t_1, t_{14} are two boundary points (minimum and maximum follow up time). Then the posterior probability that $|\Lambda(t)| > 0$ can be approximated as:

$$P(|\Lambda(t)| > 0|\text{data}) \approx \frac{1}{L} \sum_{l=1}^{L} I(|\Lambda^l(t)| > 0)$$

where $\Lambda^l(t)$ is the lth iterate in the Gibbs sampler. We can then calculate the posterior probability based on MCMC output.

We estimated the absolute difference between the active and sham acupuncture groups separately over 1-5 days and 6-14 days for the parametric and nonparametric models. For the parametric model (normal random effect), the estimated absolute difference between the active and sham acupunture during days 1-5 is 1.832 with posterior probability 0.92, and during days 6-14 is 0.471 with posterior probability 0.32. For the nonparametric model (DPM random effect), the estimated absolute difference between the active and sham acupunture during days 1-5 is 1.136 with posterior probability 0.76, and during days 6-14 is 0.316 with posterior probability 0.29. Thus, there is some evidence (more for the parametric model than the more flexible nonparametric model) for an effect of treatment while acupuncture treatment is ongoing, but very little evidence for a treatment effect once acupuncture treatment stops.

We compare the performances of the parametric and nonparametric models using the model selection criteria, based on Deviance Information Criterion (DIC) proposed by Spiegelhalter et al. (2002) and defined as

$$\text{DIC} = \overline{D(\boldsymbol{\theta})} + p_D = -4E_{\boldsymbol{\theta}}[\log p(\boldsymbol{y}|\boldsymbol{\theta})|y] + 2\log p(\boldsymbol{y}|\overline{\boldsymbol{\theta}}).$$

Here $D(\boldsymbol{\theta}) = -2\log p(\boldsymbol{y}|\boldsymbol{\theta})$ is the deviance and $\overline{D(\boldsymbol{\theta})}$ is the average posterior deviance, $p_D = \overline{D(\boldsymbol{\theta})} - D(\overline{\boldsymbol{\theta}})$ is what Spiegelhalter et al. (2002) termed as the "effective dimension" and $\overline{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$ based on the data $\boldsymbol{y}$. Recently, Celeux et al. (2006) have pointed out that the "effective dimension" p_D can be negative in case of mixture of distributions. For models with mixtures or missing values, the recent Celeux et al. (2006) suggest eight different modifications of the DIC. The nonparametric model we propose utilize mixture structures through a DPM, and thus we choose DIC_3 (based on terminology used in Celeux et al. 2006), defined as

$$\text{DIC}_3 = -4E_{\boldsymbol{\theta}}[\log p(\boldsymbol{y}|\boldsymbol{\theta})|\boldsymbol{y}] + 2\log E_{\boldsymbol{\theta}}[p(\boldsymbol{y}|\boldsymbol{\theta})|\boldsymbol{y}]$$

as our model comparison criterion. Note that the second term is simply based on predictive distribution $p(\boldsymbol{y}|\boldsymbol{y}) = E_{\boldsymbol{\theta}}[p(\boldsymbol{y}|\boldsymbol{\theta})|\boldsymbol{y}]$. The model with smallest DIC is taken to be the best fitting model. Table 1 reports the DIC values. It can be seen that the DIC for the two-part models was calculated separately for each part of the model as well as for the overall model. Overall, we see the that the nonparametric DPM model fits the model better than the parametric random effect model.

6 Discussion

This paper presented a parametric and nonparametric Bayesian approach for modeling longitudinal semicontinuous data. The approach allowed for flexible inference on the treatment effect over time using a penalized spline model, the incorporation of serial correlation using a Brownian motion process, and more flexible distribution for the random effects using a Dirichlet process formulation. The approach is easily implemented in WinBugs (2005).

We used this methodology to analyze data from an acupuncture clinical trial. A goal of the trial was to compare daily emesis volume across a standard medical group, a sham acupuncture group, and an active acupuncture group, with the major focus on comparing the active and sham acupuncture group to assess treatment effect. The results show some evidence that active acupuncture reduced emesis relative to sham acupuncture over the period in which acupuncture was administered. The treatment effect quickly diminished after acupuncture was stopped. An assessment of the difference between the active and sham curves over days 1 to 5 was more substantial under the parametric model than the better fitting nonparametric model. Further, the posterior probability of the difference being greater than zero was higher for the parametric than the nonparametric model (0.92 and 0.76, respectively). For both models, the difference was substantially reduced for the period between days 6 and 14. These results were similar in spirit to those reported by Albert and Shen (2005) in their likelihood analysis.

Albert and Shen (2005) presented a likelihood-based approach to this problem. A Monte-Carlo EM procedure was used for parameter estimation to deal with the serial correlation. Unfortunately, this algorithm was very computationally intensive, requiring days of computation on a cluster of processes to make inference. Using WinBugs, the Bayesian approach was much more computationally feasible than the maximum-likelihood approach.

References

Albert, P. S. (2005). On the interpretation of marginal inference with a mixture model for clustered semi-continuous data. *Biometrics* **61**, 879-880.

Albert, P.S., and Shen, J. (2005). Modelling longitudinal semicontinuous emesis volume data with serial correlation in an acupuncture clinical trial. *Journal of the Royal Statistical Society: Series C* **54**, 707-720.

Antoniak, C. (1974). Mixture of Dirichlet Process with application to non parametric problems. *Annals of Statistics* **2**, 1152-1174.

Brooks, S. P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computaional and Graphical Statistics* **9**, 262-285.

Celeux, G., Forbes,F., Robert, C.P., and Titterington, D.M. (2006). Deviance Information Criteria for Missing Data Models. *Bayesian Analysis* **1**, 651–674.

Coull, B. A., Ruppert, D., Wand, M. P. (2001). Simple Incorporation of Interaction into Addtive Models. *Biometrics* **57**, 539-545.

Crainiceanu, C. (2006). Bayesian Analysis for Penalized Spline Regression Using WinBUGS. *Journal of Statistical Software* **14**, 1-24.

Duan, N., Manning, W, G., Morris, C. N., and Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. *Journal of Business and Economic Statistics* **1**, 115-126.

Durban, M., Harezlak, J., Wand, M. P., and Carroll, R. (2005). Simple fitting of subject-specific curves for longitudinal data. *statistics in medicine* **24**, 1153-1167.

Escober, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577-580.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209-230.

Gelfand, A. E. (1999). Approaches for Bayesian semi-parametric regression", in *Asymptotics, Nonparametrics, and Time Series: A Tribute to Madan Lal Puri*, eds. M. L. Puri and S. ghosh, New York: Marcel Dekker.

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**: 398-409.

Ghosh, P., and Rosner, G. (2007). A Semiparametric Bayesian Approach to Average Bioequivalence. *Statistics in Medicine*, **26**, 1224-1236.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator of such model. *Annals of Economic and Social Measurement* **5**, 475-492.

Ishwaran, H. and Zarepour, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica* **12**, 941-963.

Kleinman K, P., and Ibrahim J, G. ((1998). A Semiparametric Bayesian Approach to Generalized Linear Mixed Models. *Statistics in Medicine* **17**, 2579-2596.

Lachenbruch, P. A. (2002). Analysis of data with excess zeros. *Statistical Methods in Medical Research* **11**, 297302.

Lu, S-E., Lin, Y., and Shih, W., J. (2005). Analyzing Excessive No Changes in Clinical Trials with Clustered Data. *Biometrics* **60**, 257-267.

MacEachern, S. N. (1994). Estimating Normal Means with a Conjugate Style Dirichlet Process Prior. *Communications in Statistics: Simulation and Computation* **23**, 727-741.

McCulloch, C. E. (1997). Maximum-likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92**, 162-170.

Olsen, M, K., and Schafer, J, L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association* **96**, 730745.

R Development Core Team. (2004). A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Robinson, J, W., Zeger, S, L., and Forrest, C, B. (2006). A Hierarchical Multivariate Two-Part Model for Profiling Providers' Effects on Health Care Charges. *Journal of the American Statistical Association* **101**, 911-923.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11**, 735757.

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

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639-650.

Shen, J., Wenger, N., Glaspy, J., Hays, R., Albert, P., Choi, C. and Shekelle, P. (2000), Electroacupuncture for control of myeloablative chemotherapy-induced emesis: a randomized controlled trial. *Journal of the American medical Association* **284**, 27552761.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit (with discussion). *Journal of the Royal Statistical Society*, B, 64, 583-639.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2005). "WinBUGS User Manual, Version 1.4", MRC Biostatistics Unit, Institute of Public Health and Department of Epidemiology & Public Health, Imperial College School of Medicine, available at `http://www.mrc-bsu.cam.ac.uk/bugs`.

Sturtz, S., Liggers, U., and Gelman, A. (2005). R2WinBUGS: A Package for running WinBUGS from R. *Journal of Statistical Software.* **12**, 1-16.

Sy, J. P., Taylor, J. M. G., and Cumberland, W. G. (1997). A Stochastic Model for the Analysis of Bivariate Longitudinal AIDS Data. *Biometrics* **53**, 542-555.

Tooze, J. A., Grunwald, G. K., and Jones, R. H. (2002). Analysis of repeated measures with clumping at zero. *Statistical Methods in Medical Research* **11**, 341355.

Tu, W., and Zhou, X, H. (1999). A Wald test comparing medical costs based on log-normal distributions with zero value costs. *Statistics in Medicine* **18**, 2749-2762.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in random-effects population. *Journal of the American Statistical Association* **91**, 217-221.

Welsch, A., and Zhou, X, H. (2006). Estimating the retransformed mean in a heteroscedastic two-part model. *Journal of Statistical Planning and Inference* **36**, 860-880.

Zhou, X., and Tu, W. (1999). Comparison of several different population means when their samples contain log-normal and possibly zero observations. *Biometrics* **55**, 645-651.

Table 1: Parameter Estimates

	Parametric Model			Nonparametric Model		
$\text{DIC}_{\text{logistic}}$ $p_{D\text{logistic}}$	1747.81 68.382			1610.48 48.3		
$\text{DIC}_{\text{log-normal}}$ $p_{D\text{log-normal}}$	689.72 80.39			640.51 75.803		
Overall DIC p_{overall}	2437.53 148.772			2250.99 124.103		
Parameter	Mean	SD	95% CI	Mean	SD	95% CI
β_1^p	2.09	0.24	$(0.19, 2.31)$	1.452	0.52	$(0.1932, 2.579)$
β_1^s	2.279	0.14	$(2.01, 3.16)$	2.22	0.19	$(1.836, 2.591)$
β_2^p	-7.34E-04	0.0033	$(-0.0072, 0.0057)$	3.12E-04	0.0028	$(-0.0052, 0.0058)$
β_2^s	-0.0229	0.0135	$(-0.05, 0.001)$	-0.0228	0.012	$(-0.0477, 0.011)$
Σ^{11}	0.65	0.161	$(0.3843, 1.012)$	0.7605	0.249	$(0.2172, 1.777)$
Σ^{12}	0.17	0.028	$(-0.281, 0.661)$	0.669	0.035	$(-0.279, 1.47)$
Σ^{22}	0.0402	0.0074	$(0.0278, 0.0569)$	0.0896	0.0361	$(0.081, 0.1819)$
Σ_W^{11}	0.017	0.0085	$(0.0076, 0.0384)$	0.034	0.008	$(0.0091, 0.116)$
Σ_W^{12}	0.0024	0.003	(-0.0032,0.0103)	0.0058	0.003	$(-0.0016, 0.0206)$
Σ_W^{22}	0.0116	0.0034	$(0.0067, 0.02)$	0.013	0.0035	$(0.0073, 0.023)$
σ	0.1337	0.007	$(0.1197, 0.1493)$	0.173	0.005	$(0.129, 01535)$
ν	-	-	-	5.467	0.284	$(1.639, 9.643)$

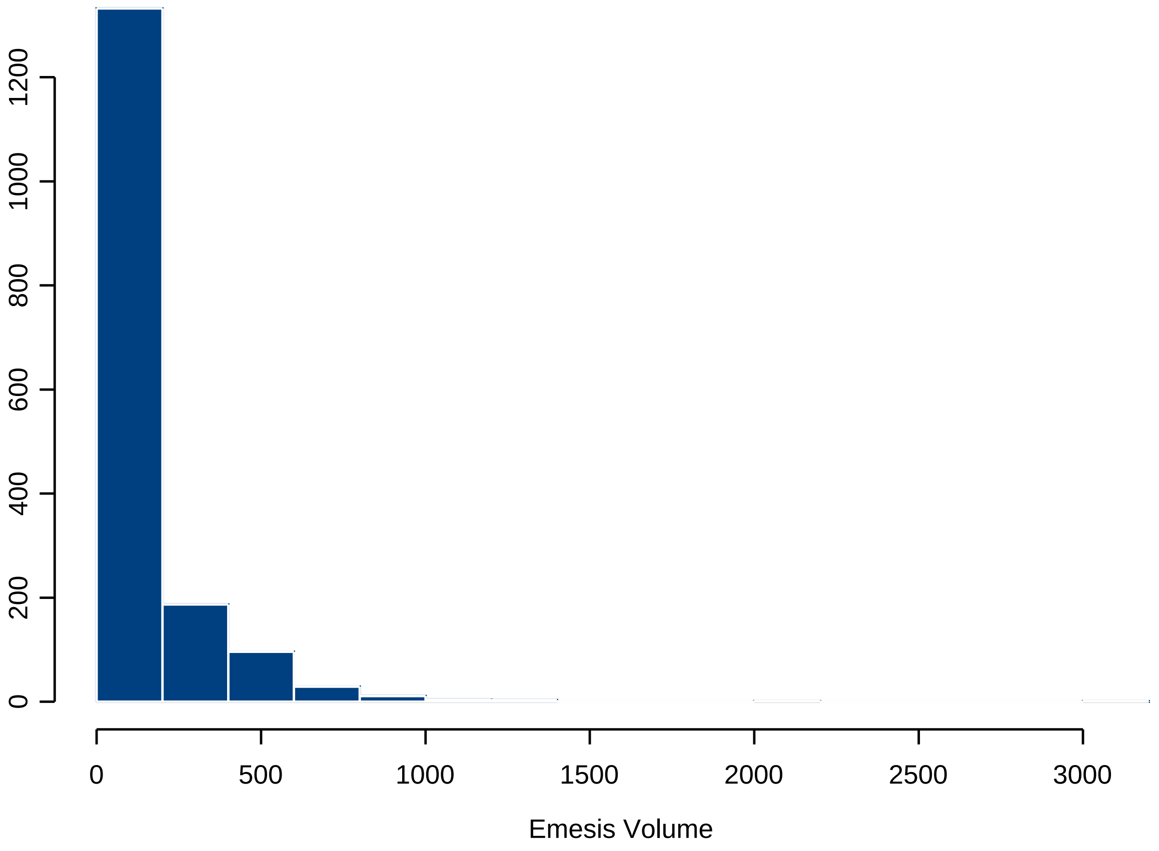

Figure 1: Histogram of the emesis volume

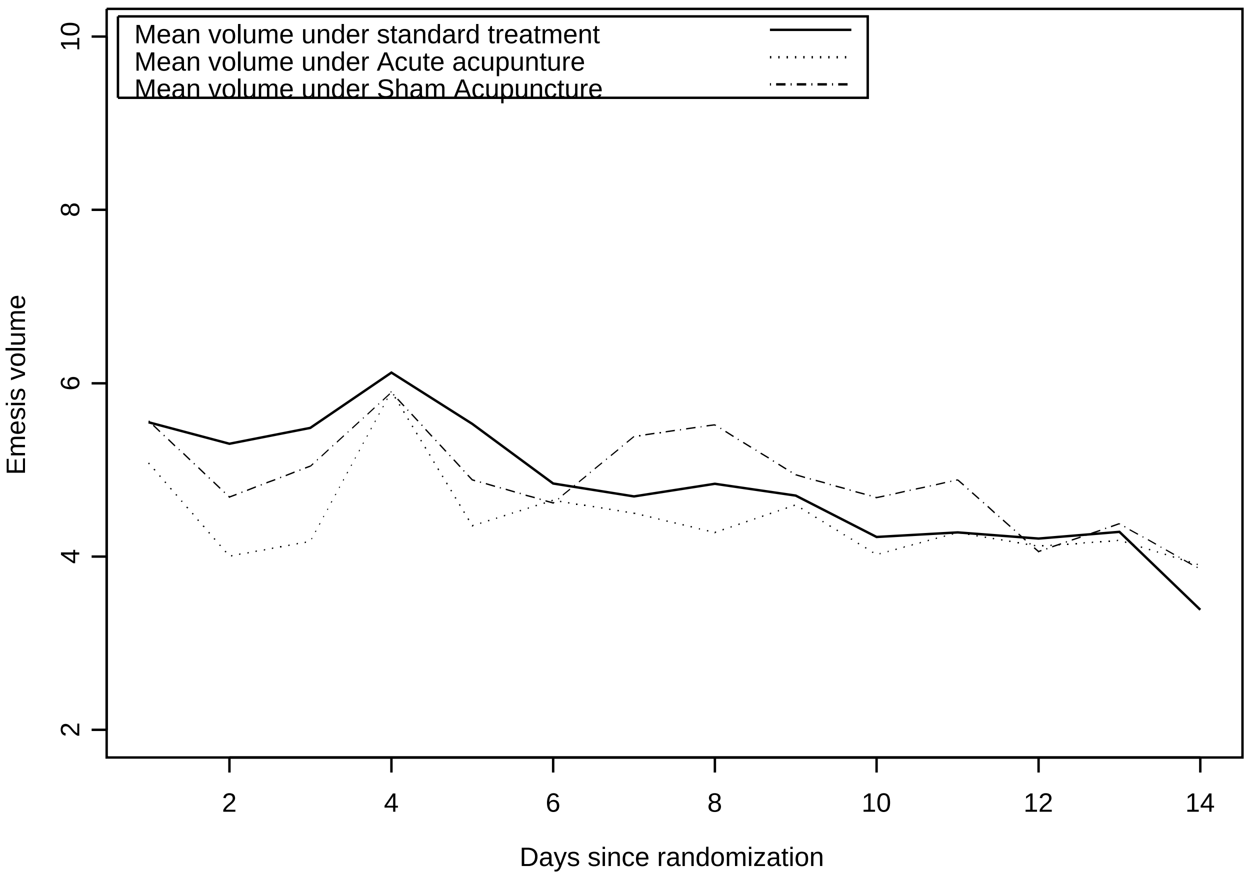

Figure 2: Mean of log(Emesis volume+1) for each treatment from day 1 to 14

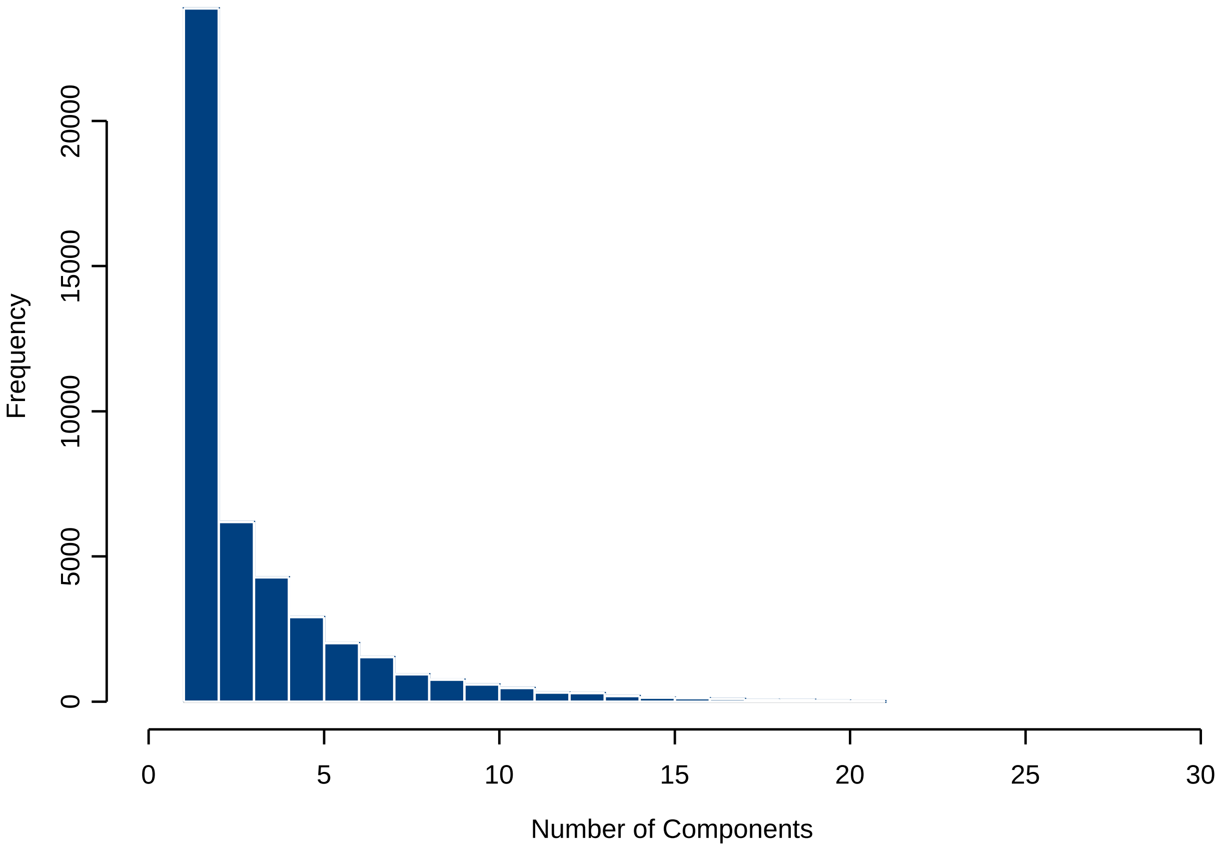

Figure 3: MCMC visits

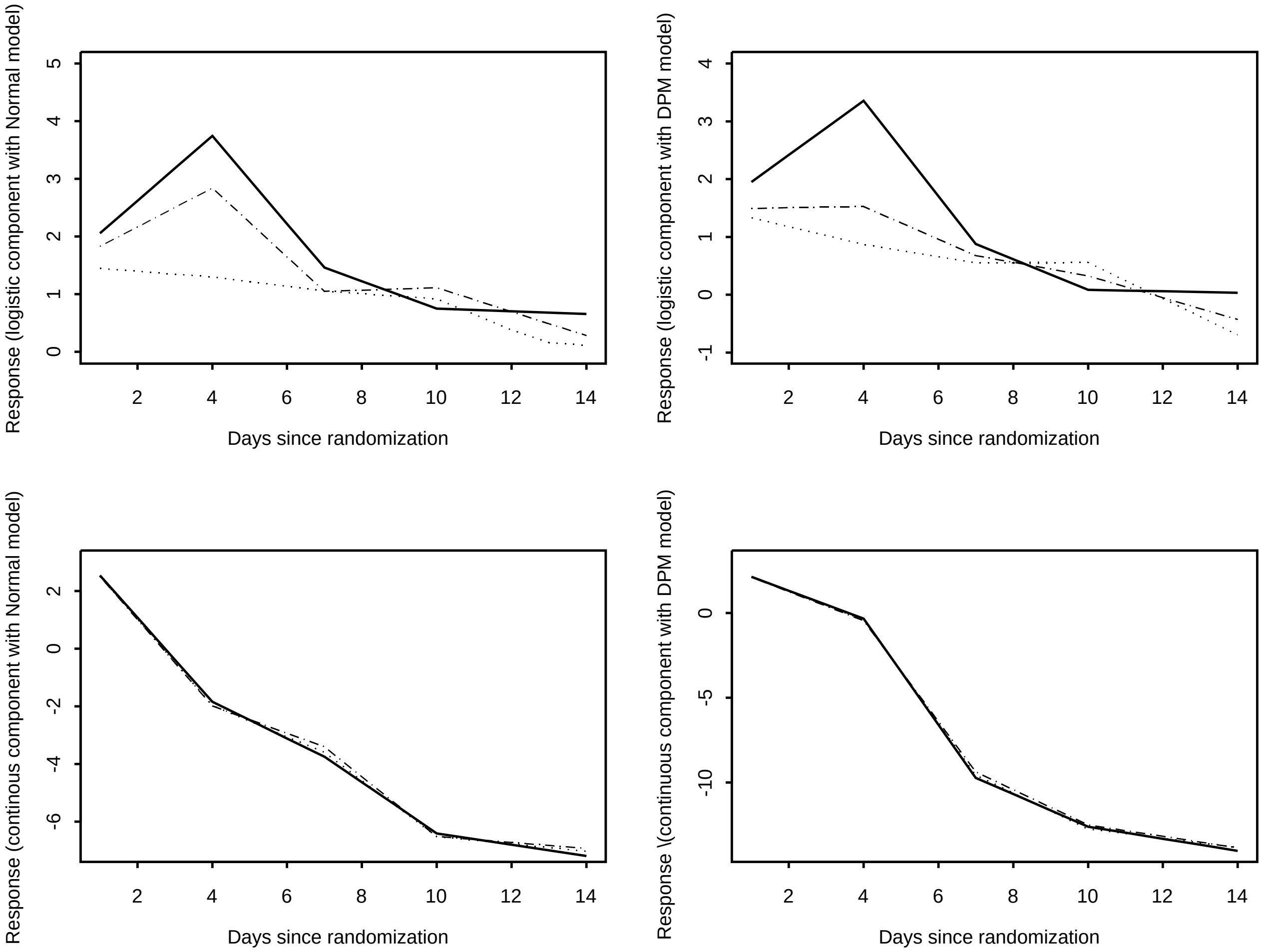

Figure 4: Treatment comparison. The top panel is the comparison of zero part and the lower panes is the same for the nonnegative longitudinal data; "–" denotes the standard medical group; ".-.-" denotes the sham acupuncture group; "..." denotes the active acupuncture group