Regression Analysis of Panel Count Data Accounting for Within-Subject Correlation with Nonparametric Frailty Distribution

Lu Wang*, Chunling Wang, Xiaoyan Lin and Lianming Wang

Department of Statistics, University of South Carolina, 1523 Greene Street, Columbia, South Carolina, 29208, U.S.A.

**Email: lw3@email.sc.edu*

Abstract: In this article, a novel Bayesian approach is proposed to analyze panel count data under the proportional mean model. The underlying counting process is assumed to be a nonhomogeneous Poisson process. The within-subject correlation is taken into consideration by introducing a frailty term, whose distribution is modeled nonparametrically via Dirichlet process mixture (DPM) technique. The unspecified nondecreasing baseline mean function is approximated by a linear combination of monotone spline functions. This proposed approach aims to provide a robust inference under no assumption about the frailty distribution. The commonly recognized overdispersion problem is avoided and estimation efficiency has improved. Further, an easy-to-implement Gibbs sampler is developed to estimate the regression parameters and the spline coefficients simultaneously. Simulation studies are conducted under different scenarios to investigate the performance of the proposed model, to compare the estimation efficiency with other existed models and to explore the performance of variance estimates. Lastly, we illustrate the proposed method by fitting the famous bladder cancer dataset.

Keywords: Panel count data; Dirichlet process mixture; Gibbs sampler; Monotone spline;

Poisson process.

1 Introduction

Panel count data is longitudinal count data that have the following characteristics: subjects are observed at several discrete time points during the study period; the number of observations varies from subject to subject; the observation times are treated as continuous random variables (Zhang and Jamshidian 2003). Panel count data often occur in a long-term clinical, industrial or animal study where the primary end point is the time to a specific event and each subject may experience several such events over time (Sun and Wei 2000). Usually an estimation of the mean function for the recurrent event over time is researcher's main interest.

Various methods have been established to analyze the panel count data. Among all the methods, two are most studied. One is the likelihood-based methods and the other is generalized estimating equation methods. When Sun and Kalbfleisch (1995) first studied the estimation of the mean function of panel count data, they constructed a nonparametric estimator based on the isotonic regression technique. Based on their study, Wellner and Zhang (2000) proposed a nonparametric pseudo-likelihood estimator which ignores the dependence between counts in the counting process as successive observation times, treating these successive counts as if they were independent random variables to form a "pseudo likelihood ". They also proposed a full nonparametric maximum likelihood Estimator (NPMLE) of the mean function in the same paper. This NPMLE adopted the assumption that the counting process is a (nonhomogeneous) Poisson process, so that this method can take account of

the dependence of the successive counts via independence of the increments of the process. The corresponding method for panel count data with covariate (Wellner and Zhang 2007) considered the proportional mean regression model under the nonhomogeneous Poisson process. Poisson process then became a commonly used tool for analyzing panel count data (Hua et al. 2014). For example, Lu et al.(2007,2009) studied the spline-based sieve version of MPLE and MLE by approximating the baseline mean function using monotone B-spline functions.

Although Poisson likelihood-based estimation methods are consistent and robust against the underlying Poisson process assumption (Wellner and Zhang 2000), the Poisson process-based likelihood does not take into account the overdispersion problem that often occurs in various applications of longitudinal count data (Hua and Zhang 2012). According to Cox (1983), overdispersion in general has two effects: underestimation of standard errors of the estimated regression parameters and loss of estimation efficiency. Both effects have been observed in the analysis of panel count data when overdispersion is neglected. Similar to introduce latent variable in GLM (Nielsen et al. (1992), Murphy (1995), Pan (2001),etc), adding a multiplicative or additive frailty term in Poisson likelihood has been widely used in the analysis of panel count data. Zhang and Jamshidian (2003) proposed an EM algorithm based on the gamma frailty Poisson model without incorporating covariates. Huang et al. (2006) introduced a latent frailty to account for informative observation times and avoided specifying its distribution through a conditional maximum likelihood approach. Yao et al. (2016) studied semiparametric regression analysis of panel count data under the gamma frailty Poisson model and derived an estimator of the within-subject correlations. Besides likelihood based methods, Hu et al. (2009) discussed an alternative method based on quasi-

score equations with additional quadratic estimation equations to account for the overdispersion. Hua and Zhang (2012) developed a spline-based semiparametric projected generalized estimating equation method and showed that the semiparametric GEE method is actually equivalent to a semiparametric likelihood method based on a gamma-frailty Poisson process model. Later, Hua et al. (2014) established the asymptotic properties of this spline based estimators and claimed that the gamma-frailty Poisson process model is robust to frailty distribution misspecification. On the other hand, Yao et al. (2016) showed in the simulation studies that the estimation on the regression parameters may be biased when the gamma frailty assumption does not hold; thus robustness of almost all the likelihood-based methods in terms of addressing the overdispersion problem is not particularly addressed. In this paper, the primary objective is to model the distribution of frailty nonparametrically so that it can properly account for the within-subject correlation and solve the overdispersion problem in all circumstances. Specifically, Dirichlet process (Ferguson 1973) is adopted here to model the frailty distribution and we call this proposed approach nonparametric frailty Possion model (NPFPM).

However, because of the almost sure discreteness of the random measure generated by Dirichlet process, modeling the frailty distribution with DP is usually used when the frailty can only get limited values, especially in analysis of clustered survival data. Naskar and Das (2006) propose a semiparametric bivariate binary model in which the subject-specific effects involved in the bivariate log odds ratio and the univariate logit components are assumed to follow a nonparametric Dirichlet process. Naskar (2008) and Manda (2011) demonstrate the use of Dirichlet process prior for the frailty under Cox proportional hazard model, where the cluster-specific shared frailty is modeled nonparametrically with DP. Additionally, Pennell

and David B. Dunson (2006) use Dirichlet process priors for modeling subject-specific shared frailty and for modeling multiplicative innovations on this frailty over time intervals. To remove the restriction of discreteness and accommodate DP to modeling subject level frailty, we applied DP mixture by adding an extra level to the hierarchy of DP. Then by adopting an approximate DP representation, we developed an easy-to-implement Bayesian estimate based on the block Gibbs sampler of Ishwaran and James (2001). This approach has the ability to take into account the positive correlation of the panel counts using the frailty variable, while preserving the simplicity of computation.

The rest of the paper is organized as follows. Section 2 introduces the proportional mean model with nonparametric frailty. Section 3 shows modeling the baseline mean function with monotone splines, modeling the frailty distribution with the Dirichlet process mixture and provides the details of an easy-to-implement Gibbs sampler for the posterior computation. Section 4 provides an extensive simulation study to evaluate the performance of our approach. Section 5 provides a real-life data application which involves the analysis of the bladder tumor data. Finally, we give some concluding remarks and discuss some further issues in Section 6.

2 The proposed model

2.1 Notation, model, and the likelihood

Consider a study that consists of n independent subjects. The observational process and the recurrent event process are assumed conditionally independent given covariates. For subject i, let x_i denote a vector of $p \times 1$ time-independent covariates and $\{t_{ij}, j = 1, ..., J_i\}$ denote the actual observation times, where J_i is the number of observations and t_{iJ_i} is the last

observation time. Let $N_i(t)$ denote the counting process for subject i, and this process is only observed at time points t_{ij}'s. We assume the proportional mean model for the counting process. In order to account for the within-subject correlation, we introduce a frailty ϕ_i for each $Ni(t)$. Specifically, conditional on ϕ_i, the frailty associated with subject i, $N_i(t)$ is a non-homogeneous Poisson process with mean function $\mu_0(t)\exp(\mathbf{x}'\boldsymbol{\beta})\phi_i$, where $\mu_0(t)$ is an unspecified nondecreasing baseline mean function with $\mu_0(0) = 0$. We adopt monotone splines to model $\mu_0(t)$ so that it only involves a finite number of parameters.

Then, by the property of a non-homogeneous Poisson process, define $Z_{ij} = N_i(t_{ij}) - N_i(t_{ij-1})$, the count of recurrent events for subject i within the time interval $(t_{ij-1}, t_{ij}]$. All Z_{ij}'s are conditionally independent given ϕ_i, and

$$Z_{ij}|\phi_i \sim Poi\Big[\{\mu_0(t_{ij}) - \mu_0(t_{ij-1})\}\exp(\mathbf{x}'\boldsymbol{\beta})\phi_i\Big].$$

Therefore, the observed data likelihood has the form:

$$L_{obs} = \prod_{i=1}^{n}\int h(\phi_i)\prod_{j=1}^{J_i}\mathcal{P}(z_{ij}|\mu_{ij})d\phi_i,$$

where $h(\phi_i)$ is the pdf of the frailty ϕ_i, whose form is unspecified; $\mathcal{P}(\centerdot|\mu_{ij})$ is the pmf of the Poisson distribution with mean $\{\mu_0(t_{ij}) - \mu_0(t_{ij-1})\}\exp(\mathbf{x}_i'\boldsymbol{\beta})\phi_i$.

Modeling the distribution of ϕ_i nonparametrically by using Dirichlet process directly incurs the problem of unidentifiability and the discreteness of the posterior distribution. In order to guarantee the identifiability of the parameters, the underling distribution needs to have mean equal to 1. Meanwhile, since DP prior ensure that the posterior distributions of the frailties are almost surely discrete (Blackwell 1973), we need to circumvent the discrete constrain of DP as well. To accomplish these two goals, the Dirichlet process mixing of frailty distribution has been utilized instead of simple DP. More specifically, we assume the

frailty to follow an unspecified distribution $h(\phi_i) = \int g(\phi_i|v_i)d\pi(v_i)$. This is equivalent to a Dirichlet process mixture model, in which the frailty follows a conditional Gamma distribution with mean 1 and variance $1/v_i$, denoted by $\phi_i|v_i \sim Ga(v_i, v_i)$, and the distribution of v_i is generated from Dirichlet process, i.e. $v_i \sim \pi(\centerdot)$ and $\pi(\centerdot) \sim DP(\alpha G_0)$. In this expression, $DP(\alpha G_0)$ refers to $\pi(\centerdot)$ being a random distribution generated by a Dirichlet process with base measure G_0 and total mass parameter α.

2.2 Within-subject correlation

Under the proposed model, the panel counts for different subjects are independent, while the panel counts within subject i shares common frailty ϕ_i which induces within-subject correlation. Without losing generality, we can quantify the within-subject correlation by considering two non-overlapping intervals $(t_1, t_2]$ and $(t_3, t_4]$. Let Z_1 and Z_2 denote the counts of recurrent events within these two intervals, respectively, from the same subject with covariates $\mathbf{x}$. As shown in Appendix A, the corresponding Pearsons correlation coefficient between Z_1 and Z_2 takes the following form,

$$\rho(Z_1, Z_2) = \frac{1}{\sqrt{\{1 + \lambda_1^{-1}\text{var}(\phi)^{-1}\}\{1 + \lambda_2^{-1}\text{var}(\phi)^{-1}\}}},$$

where $\text{var}(\phi) = \int v^{-1}d\pi(v)$, $\lambda_1 = \{\mu_0(t_2) - \mu_0(t_1)\}\exp(\mathbf{x}'\boldsymbol{\beta})$ and $\lambda_2 = \{\mu_0(t_4) - \mu_0(t_3)\}\exp(\mathbf{x}'\boldsymbol{\beta})$ are the mean numbers of the recurrent events occurring within $(t_1, t_2]$ and $(t_3, t_4]$, respectively. It is clear that ρ is decided by $\text{var}(\phi)$ and the mean numbers of recurrent events within the two considered time spans. In general, the smaller the variance of the frailties is, the smaller the within-subject association is.

3 The proposed Bayesian method

3.1 Monotone splines

Estimating the baseline mean function $\mu_0(.)$ is important as it is an indispensable part of the mean function but is also challenging because it is infinitely dimensional. In a nonparametric estimation, the number of parameters involved in $\mu_0(.)$ is on the order of sample size when the observation times differ from subject to subject. To handle this situation, we follow Yao et al. (2016)'s tactic and approximate the baseline mean function $\mu_0(.)$ with monotone splines (Ramsay 1988) in the following manner,

$$\mu_0(t) = \sum_{l=1}^{L} \gamma_l b_l(t), \tag{1}$$

where $b_l(.)$'s are integrated spline basis functions and γ_l's are unknown spline coefficients, for $l = 1...L$. Each of the I-spline basis function is a piecewise polynomials of specified degree $d-1$. Each starts from 0 in an initial flat region, increases in a mid region, and then plateaus at 1 at higher values (Wang and Dunson 2011). In such a way, by constraining the basis coefficients to be nonnegative, the monotonicity of $\mu_0(.)$ can be guaranteed.

One can determine the form of integrated spline basis functions by specifying knots and degree. The placement of the knots determines the shape and the degree determines the smoothness of the monotone splines. The detailed formulation of the basis function $b_l(.)$ can be found in Ramsay (1988) or Lin et al. (2015). According to previous studies, 2 or 3 degree can provide adequate smoothness. As for the placement of knots, Cai et al. (2011) have shown that using $10-30$ knots (equally-spaced or based on quantiles) provide adequate modeling flexibility for data sets containing up to thousands of observations (Cai et al. 2011,

Wang and Dunson 2011). Additionally, we adopt a shrinkage prior for the spline coefficients in the Gibbs sampler to prevent over-fitting problems.

3.2 Data augmentation

In order to exploit the monotone spline representation of $\mu_0(\cdot)$ in (1), a data augmentation is considered. By taking advantage of the Poisson likelihood and the additive form of the spline expression, we decompose Z_{ij} into the sum of L conditionally independent Poisson latent variables $\{Z_{ijl}\}_{l=1}^{L}$, for subject i and the time interval $(t_{i,j-1}, t_{ij}]$, such that

$$Z_{ij} = \sum_{l=1}^{L} Z_{ijl}$$

$$Z_{ijl}|\phi_i \sim Poi[\gamma_l\{b_l(t_{ij}) - b_l(t_{ij-1})\}\exp(\mathbf{x}_i'\boldsymbol{\beta})\phi_i].$$

Under this setting, the augmented data likelihood obtains a simple multiplication form:

$$L_{aug} = \prod_{i=1}^{n} h(\phi_i) \prod_{j=1}^{J_i} \prod_{l=1}^{L} \mathcal{P}(z_{ijl}|\gamma_l\{b_l(t_{ij}) - b_l(t_{ij-1})\}\exp(\mathbf{x}_i'\boldsymbol{\beta})\phi_i), \tag{2}$$

where $\mathcal{P}(\cdot|\gamma_l\{b_l(t_{ij}) - b_l(t_{ij-1})\}\exp(\mathbf{x}_i'\boldsymbol{\beta})\phi_i)$ is the pmf of Poisson distribution with mean $\gamma_l\{b_l(t_{ij}) - b_l(t_{ij-1})\}\exp(\mathbf{x}_i'\boldsymbol{\beta})\phi_i$.

3.3 Dirichlet process mixture

To fit the Dirichlet process model, Ishwaran and James (2001) introduced the blocked Gibbs sampler which is based on an approximated DP. In this approximation, the prior is assumed to be a finite dimensional measure whose random weight is generated by the stick-breaking construction:

$$p_1 = V_1 \quad \text{and} \quad p_k = (1 - V_1)(1 - V_2)(1 - V_{k-1})V_k, \quad k = 2, ..., N - 1,$$

where $V_1, V_2, ..., V_{N-1}$ are independent and identically distributed from Beta(1, α). V_N is fixed to 1 so that $\sum_{k=1}^{N} p_k = 1$. Sethuraman (1994) showed that this truncated DP converges almost surely to DP(αG_0) as $N \to \infty$.

However, because the distribution generated by DP is discrete, Dirichlet process mixture is preferred while modeling the density. In particular, under the proposed model, define $\boldsymbol{\theta} = \{\theta_1, ..., \theta_N\}$ as the set of distinct v_i's, where $N \leq n$ is the number of distinct elements in $\mathbf{v} = \{v_1, ..., v_n\}$. Let $\mathcal{K} = \{K_1, ..., K_i\}$ denote the vector of configuration indicators such that $v_i = \theta_{K_i}$. Then denote the size of the h-th cluster as n_h, that is $n_h = \sum_{i=1}^{n} I(K_i = h)$. The random distribution which generated by a truncated DP has the from $\pi_N(.) = \sum_{h=1}^{N} p_h \delta_{\theta_h}$.

The finite dimensionality of such priors is a key to the success of block sampler because it allows us to express our model in terms of a finite number of random variables (Ishwaran and James 2001). Then the frailty part in the Bayesian semiparametric model can be written hierarchically as a truncated DP mixture:

$$\begin{aligned} (\phi_i|\boldsymbol{\theta}, \mathcal{K}) &\overset{ind}{\sim} Ga(\theta_{K_i}, \theta_{K_i}), \quad i = 1, ..., n \\ (K_i|\mathbf{p}) &\overset{iid}{\sim} \sum_{h=1}^{N} p_h \delta_h \\ (\mathbf{p}, \boldsymbol{\theta}) &\sim \pi_N(\mathbf{p}) \times G_0^N(\boldsymbol{\theta}). \end{aligned} \tag{3}$$

This expression allows the blocked Gibbs sampler to update blocks of parameters. Because of the nature of the prior, those parameters are drawn from simple multivariate distributions. Note that usually Gibbs sampling schemes in mixture of Dirichlet process models are restricted to using conjugate base measures which allow analytic evaluation of the transition probabilities or alternatively need to rely on approximate numeric evaluations of some transition probabilities (Maceachern and Muller 1998). For convinience of computation, we take

G_0 to be $Ga(1, 1)$.

3.4 Prior specification and posterior computation

Since the posterior distribution is intractable for exact inference, we use Gibbs sampler (Geman and Geman 1984) for our posterior computation. Gibbs sampler is one of the most popular Markov chain Monte Carlo (MCMC) (Robert and Casella 2004) algorithms for Bayesian computation. It repeatedly and sequentially generates all unknown parameters and latent variables from their full conditional distributions. The MCMC theory guarantees that the limiting distribution of the samples from a Gibbs sampler is the same as the joint posterior distribution under certain regularity conditions. The Gibbs sampler we developed is a combination of non-homogeneous Poisson process and Dirichlet process. The former part is just a standard derivation of full conditional distribution while the latter part is an application of blocked Gibbs sampler.

We need to specify the prior distributions for the unknown parameters $\boldsymbol{\beta}$ and γ_l's, then combining with the complete likelihood (Equation 2) we can obtain full conditional distribution of the parameters. We simply adopt conventional vague priors and assign independent exponential priors $\mathcal{E}xp(\lambda)$ to γ_l's, Gamma(a_α, b_α) prior to α and prior distribution $\mathcal{N}(\mu_0, \Sigma_0)$ to $\boldsymbol{\beta}$, with mean vector zero and large independent variance such as 10. As mentioned before, we also assign a $Gamma(a_\lambda, b_\lambda)$ hyper prior for λ. Theoretically, such a prior specification is closely related to Bayesian Lasso (Park and Casella 2008) and is equivalent to the penalized likelihood approach with L1 penalty imposed on those spline coefficients, where λ serves as a tuning parameter. Lin et al. (2015) showed this shrinkage priors for the spline coefficients naturally prevents over-fitting and allow for automatic tuning with much less computational

efforts.

3.5 Gibbs sampler

The initial values of θ_h's are sampled from Ga$(1,1)$ independently. Their corresponding p_h is generated from Dirichlet$(\alpha/N, ..., \alpha/N)$, where α is generated from Ga(a_α, b_α). Then the initial value of K_i's are generated independently from $Multinomial(1, \mathbf{p})$. With $\mathcal{K}$ we identify v_i for each observation and generate frailty ϕ_i independently. The Gibbs sampler iterates through the following steps:

1: Sample $Z_{ij1}, ..., Z_{ijL}|z_{ij} \sim Multinomial(z_{ij}, (q_{ij1}, ..., q_{ijL}))$, where

$$q_{ijl} = \frac{\gamma_l\{b_l(t_{ij}) - b_l(t_{ij-1})\}}{\sum_{j=1}^{L}\gamma_j\{b_j(t_{ij}) - b_j(t_{ij-1})\}} \qquad l = 1, ..., L$$

2: Sample γ_l from Gamma(A_l,B_l) with

$$A_l = \sum_{i=1}^{n}\sum_{j=1}^{J_i} Z_{ijl} + 1,$$

$$B_l = \sum_{i=1}^{n}\{b_l(t_{iJ_i}) - b_l(t_{i0})\}\exp(\mathbf{x}_i'\boldsymbol{\beta})\phi_i + \lambda.$$

3: Update λ Sample

$$\lambda \sim Ga(\alpha_\lambda + L, \beta_\lambda + \sum_{l=1}^{L}\gamma_l)$$

4: Sample $\boldsymbol{\beta}$ by ARMS as

$$L(\boldsymbol{\beta}|\boldsymbol{.}) \propto \exp\left\{\sum_{i=1}^{n} Z_i\mathbf{x}_i'\boldsymbol{\beta} - \sum_{i=1}^{n}\{\mu_0(t_{iJ_i}) - \mu_0(t_{i0})\}\phi_i\exp(\mathbf{x}_i'\boldsymbol{\beta}) - (\boldsymbol{\beta} - \boldsymbol{\mu}_0)'\Sigma_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)/2\right\}$$

where $Z_i = \sum_{j=1}^{J_i} Z_{ij}$.

5: Sample ϕ_i from Gamma(a_i,b_i) with

$$a_i = Z_i + v_i$$

$$b_i = \{\mu_0(t_{iJ_i}) - \mu_0(t_{i0})\} \exp(\mathbf{x}_i'\boldsymbol{\beta}) + v_i.$$

6: Sample θ_h, for $h = 1, ..., N$, by using ARMS.

$$\theta_h \propto \exp(-\theta_h) \prod_{\{i:K_i=h\}} \frac{\theta_h^{\theta_h}}{\Gamma(\theta_h)} \phi_i^{\theta_h - 1} \exp(-\theta_h \phi_i)$$

7: Sample $K_i \sim Multinomial(1, \mathbf{p_i})$, where

$$p_{ih} = \frac{p_h g(\phi_i|\theta_h)}{\sum_{r=1}^{N} p_r g(\phi_i|\theta_r)},$$

for $h = 1, ..., N$.

8: Sample $V_h \overset{ind}{\sim} Beta(1 + n_h, \alpha + \sum_{s=h+1}^{N} n_s)$, for $h = 1, ..., N-1$ and let $V_N = 1$. Then let $p_1 = V_1$ and $p_h = V_h(1 - V_{h-1})...(1 - V_1)$, for $h = 2, ..., N$.

9: Sample α,

$$\alpha \sim Ga(a_\alpha + N - 1, b_\alpha - \sum_{h=1}^{N-1} log(1 - V_h))$$

4 Simulation study

An extensive simulation study is conducted to assess the performance of the proposed approach thoroughly. To demonstrate the robustness and the other advantages of the proposed nonparametric frailty Poisson model (NPFPM), we compared it with gamma frailty Poisson

model (GFPM) under six different frailty distributions: (1) the simplest case where data were generated from the gamma frailty Poisson model where the frailty ϕ_i's were generated from $Ga(0.5, 0.5)$; (2) the data were generated from the mixture gamma frailty Poisson model, in which the frailty follows a gamma mixture distribution $0.5Ga(1,1)+0.5Ga(50,50)$; (3) To further explore the performance of our approach with more complicated frailty form, we generate data with the frailty follows a mixture of four gamma distributions, i.e., $\phi_i \sim 0.25Ga(0.5,0.5) + 0.25Ga(1,1) + 0.25Ga(10,10) + 0.25Ga(50,50)$; some other distributions are also considered, such as (4) ϕ_i's follow a lognormal distribution with mean 1 and variance 2; (5) ϕ_i's follow a log-logistic distribution with shape π and scale $\sin(1)$, so that it has mean 1 and variance around 0.56; and (6) ϕ_i's follow a lognormal mixture distribution $0.5\mathcal{LN}(1,2) + 0.5\mathcal{LN}(1,0.02)$, which has mean 1 and variance 1.01. Note that the frailty distribution becomes more complicated in a progressive manner. For the purpose of illustration, we fix the degree of monotone spline to 2 and use 18 equally-spaced interior knots in all the situations. For the part involving Dirichlet process in the NPFPM, we fixed N, the number of distinct values of v_i's to 20.

For each setting, we simulated 500 data sets and there are $n = 100$ subjects in each data set. To generate the observation process for subject i, we first generate the total number of observation times from 1 plus a Poisson random variable with mean 6, then generate gap times independently from an exponential distribution with a rate parameter 2. The counting process associated with subject i was generated from the following model,

$$Z_{ij}|\phi_i = N_i(t_{ij}) - N_i(t_{ij-1}) \sim Poi\Big[\{\mu_0(t_{ij}) - \mu_0(t_{ij-1})\}\exp(x_{i1}\beta_1 + x_{i2}\beta_2)\phi_i\Big],$$

where $\mu_0(t) = log(1+t) + t$, $x_{i1} \sim Bernoulli(0.5)$, $x_{i2} \sim N(0, 0.5^2)$, the true values of

regression coefficients (β_1, β_2) have values $(1, -1)$ or $(-1, 1)$, and the frailty ϕ_i comes from different distributions shown previously.

Table 1 summarizes the simulation results on the estimation of (β_1, β_2) from NPFPM and GFPM in terms of bias, the difference between the average of 500 posterior means and the true parameter value; ESE, the average of the estimated posterior standard errors; SSD, the sample standard deviation of the 500 posterior means; and CP95, the coverage rate based on the 500 95% credible intervals.

As shown in Table 1, NPFPM performs well and steady with the frailty following different distributions. The small values of BIAS suggest that the point estimate of regression parameters obtained by NPFPM are close to their corresponding true values hence the estimation is unbiased. Additionally, ESE is close to SSD in each situation, indicating the proposed approach effectively solves the problem of overdispersion. This advantage is more obvious when compared to GFPM, whose ESE and SSD have larger differences when frailty does not follow gamma distribution. Meanwhile, the magnitudes of ESE and SSD are smaller in NPFPM than that in GFPM, so that NPFPM can offer a shorter confidence interval and provide more precise inferences. The empirical coverage probabilities (CP95) for the confidence intervals for the regression parameters are close to the nominal level 0.95, which means the Bayesian inference based on NPFPM is reliable.

Table 1: Simulation results from NPFPM and GFNPM when the data were generated from Poisson model in which the frailty generated 6 different distributions. Summarized results include the bias (Bias), the average of the estimated posterior standard errors(ESE), the sample standard deviation of the 500 posterior means (SSD), and the 95% coverage rate (CP95). The true frailty distributions are: GM: gamma distribution; MTGM: mixture of two gamma distributions; MFGM: mixture of four gamma distributions; $\mathcal{LN}$, lognormal distribution; $\mathcal{LL}$, log-logistic distribution; and mix-$\mathcal{LN}$, mixture lognormal distribution.

Dist	(β_1, β_2)	Est	NPFPM				GFNPM			
			Bias	ESE	SSD	CP95	Bias	ESE	SSD	CP95
GM	$(1,-1)$	$\hat{\beta}_1$	-0.0535	0.3001	0.2854	0.966	-0.0539	0.3000	0.2914	0.950
		$\hat{\beta}_2$	0.0164	0.3169	0.3282	0.944	0.0175	0.3173	0.3206	0.942
	$(-1,1)$	$\hat{\beta}_1$	-0.0484	0.3276	0.3317	0.924	-0.0493	0.3260	0.3241	0.928
		$\hat{\beta}_2$	-0.0027	0.3468	0.3518	0.952	-0.0035	0.3436	0.3436	0.952
MTGM	$(1,-1)$	$\hat{\beta}_1$	-0.0014	0.1423	0.1373	0.957	-0.0041	0.1605	0.1634	0.945
		$\hat{\beta}_2$	-0.0103	0.1438	0.1429	0.955	-0.0120	0.1658	0.1659	0.945
	$(-1,1)$	$\hat{\beta}_1$	0.0040	0.1857	0.1791	0.954	0.0092	0.1893	0.1890	0.948
		$\hat{\beta}_2$	0.0054	0.1880	0.1881	0.958	0.0070	0.1957	0.1934	0.952
MFGM	$(1,-1)$	$\hat{\beta}_1$	-0.0071	0.1592	0.1602	0.958	-0.0139	0.1862	0.2010	0.912
		$\hat{\beta}_2$	-0.0002	0.1618	0.1617	0.948	-0.0013	0.1919	0.2009	0.946
	$(-1,1)$	$\hat{\beta}_1$	-0.0095	0.2042	0.2049	0.954	-0.0099	0.2118	0.2220	0.938
		$\hat{\beta}_2$	0.0074	0.2065	0.2144	0.948	0.0164	0.2177	0.2312	0.934
$\mathcal{LN}$	$(1,-1)$	$\hat{\beta}_1$	-0.0315	0.2267	0.2548	0.918	-0.0155	0.2269	0.2878	0.884
		$\hat{\beta}_2$	0.0076	0.2380	0.2545	0.926	-0.0138	0.2368	0.2717	0.904
	$(-1,1)$	$\hat{\beta}_1$	-0.0168	0.2636	0.2900	0.916	-0.0338	0.2622	0.3176	0.884
		$\hat{\beta}_2$	0.0255	0.2698	0.2903	0.932	0.0337	0.2717	0.3062	0.914
$\mathcal{LL}$	$(1,-1)$	$\hat{\beta}_1$	-0.0171	0.1412	0.1491	0.930	-0.0119	0.1444	0.1691	0.916
		$\hat{\beta}_2$	0.0025	0.1456	0.1604	0.922	-0.0033	0.1497	0.1757	0.900
	$(-1,1)$	$\hat{\beta}_1$	-0.0009	0.1782	0.1904	0.922	-0.0120	0.1807	0.2010	0.914
		$\hat{\beta}_2$	0.0069	0.1781	0.1783	0.954	0.0230	0.1825	0.1936	0.938
M$\mathcal{LN}$	$(1,-1)$	$\hat{\beta}_1$	-0.0128	0.1344	0.1303	0.956	-0.0047	0.1446	0.1558	0.942
		$\hat{\beta}_2$	-0.0001	0.1355	0.1298	0.962	-0.0092	0.1475	0.1591	0.934
	$(-1,1)$	$\hat{\beta}_1$	-0.0018	0.1727	0.1734	0.946	-0.0103	0.1770	0.1958	0.926
		$\hat{\beta}_2$	-0.0082	0.1715	0.1756	0.946	0.0015	0.1795	0.1859	0.934

Regarding baseline mean function, Figure 1 shows the true baseline mean function and the average of the baseline mean function estimates from NPFPM and GFPM when $(\beta_1, \beta_2) = (1, 1)$ and frailty follows $Gamma(0.5, 0.5)$, mixture of four gamma distribution specified above, and $log - logistic(\pi, sin(1))$. As seen in Figure 1, NPFPM and GFPM have similar performance in general because the averaged baseline mean estimates from NPFPM and GFPM essentially overlaps with the true curve. However, after close inspection, we can see GFPM gives better estimation when frailty follows gamma distribution while NPFPM gives better estimation in the other two settings.

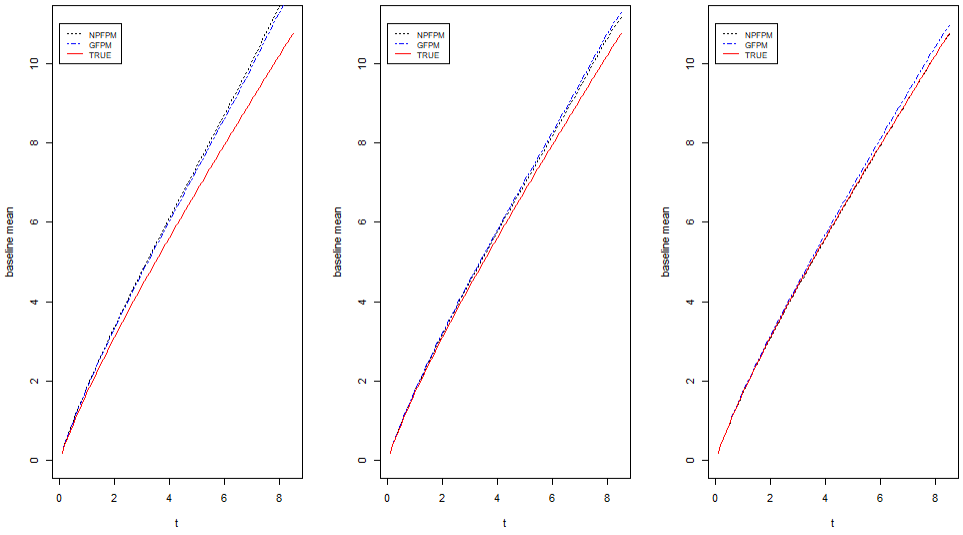

Figure 1: The true baseline mean function and the average of the estimated baseline mean curves under NPFPM and GFPM, when frailty follows gamma distribution, mixture of four gamma distribution and log-logistic distribution.

In summary, these two methods have comparable performance when frailty follows gamma distribution. However, NPFPM performs better in terms of parameter estimation, inferen-

tial characteristics, and baseline mean function estimation when the frailty distribution is different from gamma. Thus, we conclude that estimating the frailty nonparametrically with Dirichlet process mixture is robust and it solves the problem of underestimating variances and increases the coverage probabilities. This makes sense because there are multiple counts from the same subjects which can provide adequate information to estimate the frailty distribution accurately (Yao et al. 2016).

5 Real-life data application

In this section, we apply the proposed method to the most widely used panel count data example in the literature, which arose from a bladder cancer study conducted by the Veterans Administration Cooperative Urological Research Group (Byar and Blackard 1977). In this randomized clinical trial study, all the 118 patients had experienced superficial bladder tumors when they entered the trial. They were randomized into one of three treatment groups: placebo, thiotepa, and pyridoxine. During the study at each follow-up visit, new tumors since the last visit were counted, measured and then removed transurethrally. The number of follow-up clinical visits and follow-up times vary noticeably from patient to patient. The primary objective of the study is to determine if any treatment could significantly reduce the recurrence of bladder tumor.

This data set has been analyzed extensively using many different approaches in the literature. Following Wellner and Zhang (2007) and Lu et al. (2009), we focused on 116 patients in the study, who had at least one follow-up observation after the study enrollment. Let $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})'$ denote the covariate vector for patient i, where x_{i1} and x_{i2} represent

the number of bladder tumors and the size of the largest bladder tumors for patient i at the beginning of the trial, and x_{i3} and x_{i4} are the binary variables indicate whether patient i was assigned to the treatment of pyridoxine pills or thiotepa installation, respectively. When applying the proposed method, we use 20 equally-spaced knots within the data range $0-64$ months for the monotone spline specification. In addition, we fix the number of distinct values of v_i's to 20 as in the simulation.

Table 2: Bladder tumor data analysis from the proposed approach, the GFNPM approach and the WZ approach in Wellner and Zhang (2007). Summarized results are the point estimates (Point), the standard errors (SE), and the p-values for all the regression parameters.

	NPFPM			GFNPM			WZ		
	Point	SE	CI95	Point	SE	CI95	Point	SE	CI95
$\hat{\beta}_1$	0.317	0.101	(0.109,0.502)	0.336	0.106	(0.128,0.544)	0.207	0.078	(0.054,0.360)
$\hat{\beta}_2$	-0.026	0.129	(-0.260,0.255)	0.012	0.120	(-0.223,0.247)	-0.036	0.086	(-0.133,0.205)
$\hat{\beta}_3$	-0.107	0.391	(-0.838,0.716)	-0.033	0.409	(-0.835,0.769)	0.066	0.431	(-0.779,0.911)
$\hat{\beta}_4$	-1.219	0.442	(-2.043,-0.330)	-1.140	0.435	(-1.993,-0.287)	-0.797	0.360	(0.091,1.503)

Table 2 shows the results from the proposed approach and two other competitive approaches, i.e. Yao et al. (2016) and Wellner and Zhang (2007). The results from these two competitors are directly drawn from their papers. Both of these two competitive approaches are likelihood-based approaches under the non-homogeneous Poisson model. Yao et al. (2016)'s method considered the within-subject correlation while Wellner and Zhang (2007)'s method did not consider the within-subject correlation.

As seen in Table 2, the results from our method indicates that the number of initial bladder tumors was positively related to the recurrence of the tumor while the size of the largest tumor at the enrollment did not have a significant effect. It also reveals that the

thiotepa instillation treatment significantly reduced the recurrence rate of bladder tumors, while the treatment of pyridoxine pills did not have a significant effect. Figure 2 plots the estimated mean functions of bladder tumor counts for the control and the other two treatment groups. It is clear that the estimated mean functions for the control and the pyridoxine treatment groups are close to each other and they are higher than the one for the thiotepa treatment group. These conclusions are consistent with those made in Wellner and Zhang (2007) and more close to Yao et al. (2016) in terms of regression coefficients estimation and baseline mean function estimation. This is because Yao et al.'s method also accounts for the within-subject correlation. In addition, the estimated density function of the frailty from the proposed method and Yao et al.'s method highly matches with each other as shown in Figure 3. For this data, the within-subject correlation is not ignorable because the tumor number at baseline is positively related to the recurrence of bladder tumor (Hua and Zhang 2012).

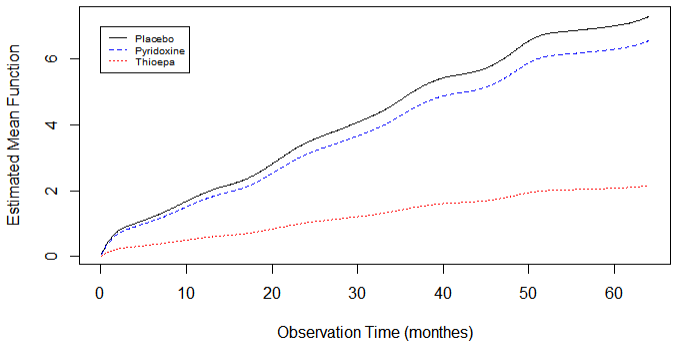

Figure 2: The estimated mean functions for different groups for the bladder tumor data.

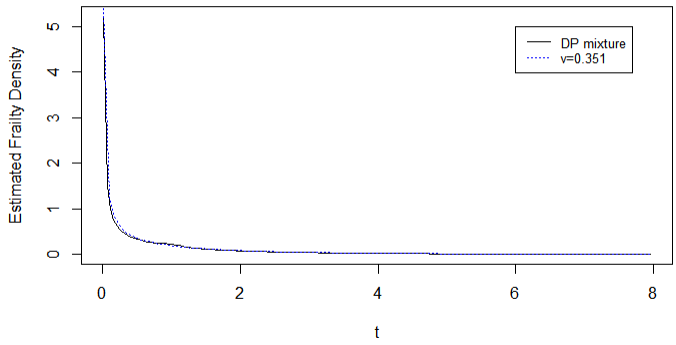

Figure 3: The estimated mean functions for different groups for the bladder tumor data.

To quantify such within-subject correlation and compare the result with previous research, we follow Yao et al. (2016)'s setting: define Z_1 and Z_2 to be the random counts of bladder tumors within the first six months and within the next six months for a patient with the median number ($x_1 = 1$) of bladder tumors and the median size of the largest tumors ($x_2 = 1$) at the enrollment. The Pearsons correlation coefficients between Z_1 and Z_2 estimated by the proposed NPFPM and GFNPM are shown in Table 3. If the patient was in the placebo group, the Pearsons correlation coefficient between Z_1 and Z_2 is estimated to be 0.8435 using NPFPM and 0.6993 using GFNPM; if the patient was in the thiotepa treatment group, the corresponding Pearson's correlation estimated by NPFPM and GFNPM are 0.6285 and 0.4325, respectively. As for the pyridoxine treatment group, Yao et al. (2016) didn't show their result, but the present method shows the Pearson's correlation is 0.8209 which is very close to the placebo group. Both methods suggest a medium to large within-subject correlation and this conclusion is consistent with Yao et al. (2016)'s research.

Table 3: Pearson correlations of counts of bladder tumors within the first six months and the next six months for a patient with the median number of bladder tumors and the median size of the largest tumors at the enrollment.

Method	Placebo	Pyridoxine	Thiotepa
NPFPM	0.8435	0.8209	0.6285
GFNPM	0.6993	*	0.4325

6 Discussion

In this paper, we proposed a Bayesian estimation approach to analyze panel count data under the proportional mean model assuming a non-homogeneous Poisson process. By introducing multiplicative frailty term for the proportional mean model, this approach is able to account for within-subject correlation. The frailty distribution is estimated nonparametrically with a Dirichlet process mixture which solves the problem of overdispersion in likelihood based methods. The baseline mean function is approximated by monotone splines which leads to a finite number of parameters to estimate and thus save the computation effort. An easy-to-implement Gibbs sampler is established upon Poisson data augmentation and the blocked Gibbs sampler of Ishwaran and James (2001). The proposed method shows an excellent performance of estimating the regression parameters and the baseline mean function when frailty follows different distributions as shown in our simulation studies and the real data application. Our future effort will be devoted to developing more robust methods and extend this strategy to multivariate penal count data.

References

Blackwell, D. (1973), 'The discreteness of ferguson selections', *Annals of Statistics* **1**, 356–358.

Byar, D. & Blackard, C. (1977), 'Comparisons of placebo, pyridoxine, and topical thiotepa in preventing recurrence of stage i bladder cancer', *The Veterans Administration Cooperative Urological Research Group* **10**, 556–561.

Cai, B., Lin, X. & Wang, L. (2011), 'Bayesian proportional hazards model for current status data with monotone splines', *Computational Statistical Data Analysis* **55**, 26442651.

Cox, D. (1983), 'Some remarks on overdispersion', *Biometrika* **70**(1), 269–274.

Ferguson, T. S. (1973), 'A bayesian analysis of some nonparametric problems', *Annals of Statistics* **1**, 209–230.

Geman, S. & Geman, D. (1984), 'Stochastic relaxiation, gibbs distributions, and the bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Hu, X. J., Lagakos, S. W. & Lockhart, R. A. (2009), 'Marginal analysis of panel counts through estimating functions', *Biometrika* **96**, 445–456.

Hua, L. & Zhang, Y. (2012), 'Spline-based semiparametric projected generalized estimating equation method for panel count data', *Biostatistics* **13**, 440–454.

Hua, L., Zhang, Y. & Tu, W. (2014), 'A spline-based semiparametric sieve likelihood method for over-dispersed panel count data', *The Canadian Journal of Statistics* **42**, 217–245.

Huang, C., Wang, M. & Zhang, Y. (2006), 'Analysing panel count data with informative observation times', *Biometrika* **93**, 763–775.

Ishwaran, H. & James, F. L. (2001), 'Gibbs sampling methods for stick-breaking priors', *Journal of the American Statistical Association* **96**, 161–173.

Lin, X., Cai, B., Wang, L. & Zhang, Z. (2015), 'A bayesian proportional hazards model for general interval-censored data', *Lifetime Data Analysis* **21**, 470–490.

Lu, M., Zhang, Y. & Huang, J. (2007), 'Estimation of the mean function with panel count data using monotone polynomial splines', *Biometrika* **94**, 1060–1070.

Lu, M., Zhang, Y. & Huang, J. (2009), 'Semiparametric estimation methods for panel count data using monotone b-splines', *Journal of the American Statistical Association* **104**, 1060–1070.

Maceachern, N. & Muller, P. (1998), 'Estimating mixture of dirichlet process models', *Journal of Computational and Graphical Statistics* **7**(2), 223–238.

Manda, O. M. S. (2011), 'A nonparametric frailty model for clustered survival data', *Communications in Statistics-Theory and Methods* **40**, 863–875.

Murphy, S. A. (1995), 'Asymptotic theory for the frailty model', *Annals of Statistics* **23**, 182–198.

Naskar, M. (2008), 'Semiparametric analysis of clustered survival data under nonparametric frailty', *Statistica Neerlandica* **62**, 155–172.

Naskar, M. & Das, K. (2006), 'Semiparametric analysis of two-level bivariate binary data', *Biometrics* **62**, 1004–1013.

Nielsen, G. G., Gill, R. D., Andersen, P. K. & Sorensen, T. I. A. (1992), 'A counting process approach to maximum likelihood estimation in frailty models', *Scandinavian Journal of Statistics* **19**, 25–44.

Pan, W. (2001), 'Using frailties in the accelerated failure time model', *Lifetime Data Analysis* **7**, 55–64.

Park, T. & Casella, G. (2008), 'The bayesian lasso', *Royal Statistical Society* **103**, 681–686.

Pennell, L. M. & David B. Dunson, B. D. (2006), 'Bayesian semiparametric dynamic frailty models for multiple event time data', *Biometrics* **62**, 1044–1052.

Ramsay, J. O. (1988), 'Monotone regression splines in action', *Statistical Science* **3**, 425–461.

Robert, C. P. & Casella, G. (2004), *Monte Carlo statistical methods*, Springer, New York.

Sethuraman, J. (1994), 'A constructive definition of dirichlet priors', *Statistica Sinica* **4**, 639–650.

Sun, J. & Kalbfleisch, J. D. (1995), 'Estimation of the mean function of point processes based on panel count data', *Statistica Sinica* **5**, 279–290.

Sun, J. & Wei, L. J. (2000), 'Regression analysis of panel count data with covariate-dependent observation and censoring times', *Royal Statistical Society* **62**, 293–302.

Wang, L. & Dunson, D. (2011), 'Semiparametric bayes proportional odds models for current status data with underreporting', *Biometrics* **67**, 1111–1118.

Wellner, A. & Zhang, Y. (2000), 'Two estimators of the mean of a counting process with panel count data', *The Annals of Statistics* **28**(3), 779–814.

Wellner, J. & Zhang, Y. (2007), 'Two likelihood-based semiparametric estimation methods for panel count data with covariates', *The Annals of Statistics* **35**, 2106–2142.

Yao, B., Wang, L. & He, X. (2016), 'Semiparametric regression analysis of panel count data allowing for within-subject correlation', *Computational Statistics and Data Analysis* **97**, 47–59.

Zhang, Y. & Jamshidian, M. (2003), 'The gamma-frailty poisson model for the nonparametric estimation of panel count data', *Biometrics* **59**, 1099–1106.

Appendix A. Derivation of the within-subject correlation

Consider two non-overlapping intervals $(t_1, t_2]$ and $(t_3, t_4]$. Let Z_1 and Z_2 denote the counts of recurrent events within these two intervals, respectively, from the same subject with covariates $\mathbf{x}$. The derivation of Pearsons correlation coefficient between Z_1 and Z_2 is shown below under the proposed nonparametric frailty Poisson model. Define $\lambda_1 = \{\mu_0(t_2) - \mu_0(t_1)\}\exp(\mathbf{x}'\boldsymbol{\beta})$ and $\lambda_2 = \{\mu_0(t_4) - \mu_0(t_3)\}\exp(\mathbf{x}'\boldsymbol{\beta})$, it is clear that $Z_1|\phi \sim Poi(\lambda_1\phi)$ and $Z_2|\phi \sim Poi(\lambda_2\phi)$ under the proposed model. Meanwhile, the frailty ϕ follows the distribution $h(\cdot)$ which is generated by the DP mixture. First, using the law of iterated conditional

expectations, one obtains

$$\begin{aligned} \text{cov}(Z_1, Z_2) &= \text{cov}\{E(Z_1|\phi), E(Z_1|\phi)\} + E\{\text{cov}(Z_1, Z_2|\phi)\} \\ &= \text{cov}\{\lambda_1\phi, \lambda_2\phi\} + 0 \\ &= \lambda_1\lambda_2\text{var}(\phi) \end{aligned} \tag{4}$$

$$\begin{aligned} \text{var}(Z_1) &= E\{\text{var}(Z_1|\phi)\} + \text{var}\{E(Z_1|\phi)\} \\ &= E\{\lambda_1\phi\} + \text{var}\{\lambda_1\phi\} \\ &= \lambda_1 + \lambda_1^2\text{var}(\phi) \end{aligned} \tag{5}$$

Similarly, $\text{var}(Z_2) = \lambda_2 + \lambda_2^2\text{var}(\phi)$. Then the correlation between Z_1 and Z_2 can be written as:

$$\rho(Z_1, Z_2) = \frac{1}{\sqrt{\{1 + \lambda_1^{-1}\text{var}(\phi)^{-1}\}\{1 + \lambda_2^{-1}\text{var}(\phi)^{-1}\}}}.$$

As for $\text{var}(\phi)$ we can use iterated rule again based on the fact that $\phi|v \sim Ga(v, v)$. So we can get:

$$\text{var}(\phi) = E\{\text{var}(\phi|v)\} + \text{var}\{E(\phi|v)\} = E(v^{-1}) + \text{var}(1) = \int v^{-1} d\pi(v). \tag{6}$$

Using the notation in Section 3.3 it can be further written as $\text{var}(\phi) = \sum_{h=1} \theta_h p_h$.