Gamma frailty non-homogeneous Poisson process model for panel count data with nonparametric modeling of the baseline mean function

Lu Wang, Xiaoyan Lin*, and Lianming Wang

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

**Email: lin9@mailbox.sc.edu*

Abstract: A novel Bayesian approach is introduced for analyzing panel count data under the gamma frailty non-homogeneous Poisson process model. The baseline mean function of the Poisson process is an undefined nondecreasing function which is approximated non-parametrically. Specifically, the idea of innermost interval is adopted to define a sequence of tiny time intervals based on the data and the increments in those intervals follows a gamma process. When units or individuals are observed over time it is often apparent that the recurrent event happens at different rates, even though no differences in treatment or environment are present. Thus, gamma frailty has been introduced in the model to explain the individual level differences. Meanwhile, introducing the gamma frailty variable can account for the intra-correlation between the panel counts of the counting process. An efficient and easy-to-implement Gibbs sampler is developed for the posterior calculation. An extensive simulation study is conducted to illustrate the proposed approach.

Keywords: Gamma frailty; Poisson process; Gamma process; Panel count data.

1 Introduction

Panel count data frequently occur in clinical and observational studies especially in long-term studies. In those studies each subject may experience multiple recurrences of the event of interest but they are monitored or observed only at finite discrete time points instead of continuously. In this situation, the observed data include only the numbers of the occurrences of the event of interest between observation time points, and the exact occurrence times of the event are unknown (Zhu et al. 2018). Usually, the primary interest is the occurrence rate or the mean function of the counting process. In 1995, Sun & Kalbfleisch first studied nonparametric estimation of the mean function of panel count data using the monotonicity of the mean function of a counting process. They applied the Isotonic regression technique to estimate the mean function nonparametrically. Wellner & Zhang (2000) assumed that the counting process is a nonhomogeneous Poisson process, established a pseudo-likelihood of panel count data by ignoring the intracorrelation between the counts within a subject. Subsequent researches(Wellner & Zhang(2000,2007)) showed that the underlying conditional Poisson process assumption is robust with respect to the actual distribution of the underling counting process. To account for the within-subject correlation, Zhang & Jamshidian (2003) and Yao et al. (2016) introduced gamma-frailty variable and established EM algorithms to fit the model. Some researches (Hua et al. (2014),Xu et al. (2018)) showed that the gamma-frailty assumption is robust to the misspecification of the distribution of the latent variables.

Based on the most widely used proportional mean model, we adopted the non-homogeneous Poisson process with the gamma-frailty variable to estimate the mean function of the count-

ing process. Specifically, we assume that given the within subject frailty variable, ϕ, the counting process $N(t)$ is a Poisson process with the conditional mean function

$$E(N(t)|\phi) = \mu_t = \mu_0(t)\exp(\mathbf{x}'\boldsymbol{\beta})\phi.$$

Usually $\mu_0(t)$ is assumed to be a fixed nondecreasing function and approximated by a linear combination of spline functions (Hua & Zhang (2012),Hua et al. (2014),Yao et al. (2016)) or by estimating jump sizes of a step-function at every observation time points(Zhu et al. (2018)). Different from the previous researches, we propose to treat $\mu_0(t)$ as a nondecreasing stochastic process with independent increments. Gamma process is used as a prior distribution, that is $\mu_0(t) \sim GP(\eta H_0)$. A Bayesian estimation of the baseline mean function is established and a Gibbs sampler has been developed to fit the model.

Gamma process is a Lévy process with gamma distributed increments. It has been extensively studied and applied in various studies since been introduced by Doksum (1974) as one of the processes neutral to the right. It is commonly used to estimate nondecreasing function. Kalbfleisch (1978) used gamma process to model the cumulative hazard function and built up a Bayesian analysis of the semi-parametric regression and life model of Cox(1972). Ferguson & Phadia (1979) used the processes neutral to the right as prior distributions for the unknown distribution function F and derived general theory that can be used in the estimation of F given some right censored data. Using gamma process as prior distribution is interpretable. Consideration of the partition $(0, t]$, $[t, \infty)$ shows immediately that $\mu_0(t) \sim Ga(\eta H_0(t), \eta)$ and $E\{\mu_0(t)\} = H_0(t)$, $\text{var}\{\mu_0(t)\} = H_0(t)/\eta$. The specified $H_0(t)$ can be viewed as an initial guess at $\mu_0(t)$ and η is a specification of the weight attached to that guess. Some also think it is helpful to reveal the intrinsic property of the event of

interest. For example, Nozer (1995) use gamma process to model the hazard rate process in a dynamic environment. Because of those properties, gamma process is also widely used in models that describe the process of deterioration or degradation in units or systems. Since for certain types of degradation processes a model involving independent nonnegative increments is appropriate. For such kind of researches please refer to Lawless & Crowder (2004), Wang (2009), Sinha et al. (2015), etc.

The organization of the rest of this paper is as follows. In Section 2, we includes the setting of the model; an introduction of the notations; a brief discuss of using the idea of Turnbull interval to build successive intervals; a detail introduction of the application of gamma process; and a step of data augmentation which is important for the computation of posterior distributions. In section 3, the proposed Gibbs sampler is exhibited. In Section 4, the performance of proposed Bayesian approach is assessed by a group of simulation study. In Section 5, the proposed method is applied to real data. Section 6 is summary discussion.

2 The Poisson model with frailties

2.1 Notations

Suppose n independent subjects are observed periodically in a study. We assume that the observational process and the recurrent event process are conditionally independent given the time-independent covariates $\mathbf{x}_i$, a $p \times 1$ vector. Let $\{t_{ij}, j = 1, ..., J_i\}$ denote the actual observation times for subject i, where J_i is the number of observations and t_{iJ_i} is the last observation time. Then $N_i(t)$ is the counting process for subject i, which is only observed at t_{ij}'s. In order to account for the within-subject correlation, we adopt a gamma frailty

non-homogeneous Poisson process model for the recurrent event process $N_i(t)$. 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 $\mu_0(0) = 0$ and ϕ_i's are independently and identically distributed from $Gamma(v, v)$ with mean 1 and variance v^{-1}.

By the properties of non-homogeneous Poisson process, define $Z_{ij} = N_i(t_{ij}) - N_i(t_{ij-1})$, the count of recurrent events within time interval $(t_{ij-1}, t_{ij}]$. All Z_{ij}'s are conditionally independent given ϕ_i. And we have

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

So the observed data likelihood has the form:

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

where $g(.|v)$ is the pdf of $Gamma(v, v)$, $\mathcal{P}(.|\mu_{ij})$ is the pdf of Poisson distribution with rate $\{\mu_0(t_{ij}) - \mu_0(t_{ij-1})\}\exp(\mathbf{x}_i'\boldsymbol{\beta})\phi_i$.

A similar stochastic model was proposed by Sinha (2004), which deal with the counting process and terminal event at the same time. This method require all the subjects inspected over time follow the same scheduled time points. The proposed method improved their method by adopting the idea of Turnbull interval (Turnbull 1976) so that each subject can have different inspection times.

2.2 Turnbull intervals

Turnbull intervals also called "innermost intervals", is most commonly used in survival analysis for interval-censored data. Basically, they are gaps over the real line upon which Turn-

bull's F_t, a non-parametric maximum likelihood estimator (NPMLE), is defined. It is also known as the regions of the maximal cliques for the univariate data in an intersection graph (Gentleman & Vandal 2002). Here we use Li et al. (1997)'s notation, and define a Turnbull's interval $\{s_m, s_{m-1}\}$ to be the non-empty intersection of observed intervals $\{t_{ij}, t_{ij-1}\}$ such that $\{s_m, s_{m-1}\} \cap \{t_{ij}, t_{ij-1}\}$ is either an empty set or $\{s_m, s_{m-1}\}$. Based on this step, the time between each inspection for subject i can be expressed as a union of certain Turnbull's intervals. That is $\{t_{ij} - t_{ij-1}\} = \bigcup_{l=1}^{k_{ij}-k_{ij-1}} \{s_{k_{ij}+l}, s_{k_{ij}+l-1}\}$, where k_{ij} is an index of the position of t_{ij} on the scale of $\mathbf{s} = \{s_0, ..., s_M\}$. Specifically, we can write the relation between observation times and the boundaries of innermost intervals as $s_{k_ij} = t_{ij}$.

2.3 Gamma process

The baseline mean function of panel count data, $\mu_0(t)$, is an unspecified nondecreasing function with $\mu_0(0) = 0$. It is natural and appropriate to model it with a process with independent nonnegative increments. One such is the gamma process, which can be thought of as arising from a compound Poisson process of gamma-distributed increments in which the Poisson rate tends to infinity while the sizes of the increments tend to zero in proportion (Lawless & Crowder 2004).

Notationally, we treat the baseline mean function as a non-negative-valued processes $\{\mu_0(t), t \geq 0\}$ with the property that its increments $\Delta\mu_0(t) = \mu_0(t+\Delta t) - \mu_0(t)$ are independent with $Gamma(\eta H_0(\Delta t), \eta)$ distributions, where $H_0(.)$ is a given, monotone increasing function. Because these distributions depend only on the length of the interval, and not its location, such process are said to have stationary independent increments. We denote it as $\mu_0(t) \sim GP(\eta H_0), t \geq 0$ for short.

Furthermore, by combining the independent increment property of gamma process with the Turnbull intervals we can define $\lambda_m = \mu_0(s_m) - \mu_0(s_{m-1})$, so that $\lambda_m \sim Gamma(\{H_0(s_m) - H_0(s_{m-1}\}\eta, \eta)$. Then the increment of baseline mean function for subject i in the $j-th$ time interval has the form $\mu_0(t_{ij}) - \mu_0(t_{ij-1}) = \sum_{l=1}^{k_{ij}-k_{ij-1}} \lambda_{K_{ij-1}+l}$. Since gamma process is robust to the choose of H_0 so we simply let $H_0(t) = at$.

In this way, the observational data likelihood can be written as:

$$\begin{aligned} L_{obs} \propto \Big(\prod_{i=1}^{n} [\prod_{j=1}^{J_i} (\sum_{l=1}^{k_{ij}-k_{ij-1}} \lambda_{k_{ij-1}+l})^{z_{ij}} e^{\mathbf{x}_i' \boldsymbol{\beta} z_{ij}} \phi_i^{z_{ij}} \exp\big\{ - (\sum_{l=1}^{k_{ij}-k_{ij-1}} \lambda_{k_{ij-1}+l}) e^{\mathbf{x}_i' \boldsymbol{\beta}} \phi_i \big\}] \\ \times g(\phi_i | v) \Big) \prod_{m=1}^{M} g\{ \lambda_m | a(s_m - s_{m-1})\eta, \eta \} \end{aligned} \tag{1}$$

2.4 Data Augmentation

Basing on the fact that for $(u_1, ..., u_K) \sim Multinomial(1, (\frac{1}{K}, ..., \frac{1}{K}))$, integrate $\prod_{l=1}^{K} [\lambda_l]^{u_l}$ with respect to $(u_1, ..., u_K)$ equals to $\sum_{l=1}^{K} \lambda_l / K$. We introduce $\mathbf{u}_{ij} \sim Multi(1, (\frac{1}{n_{ij}}, ..., \frac{1}{n_{ij}}))$, where $n_{ij} = k_{ij} - k_{ij-1}$, for $i = 1, ..., n$ and $j = 1, ..., J_i$. The likelihood function obtain the form of a simple production as following:

$$\begin{aligned} L_{aug} \propto \Big(\prod_{i=1}^{n} [\prod_{j=1}^{J_i} (\prod_{l=1}^{k_{ij}-k_{ij-1}} \lambda_{K_{ij-1}+l}^{u_{ijl}})^{z_{ij}} e^{\mathbf{x}_i' \boldsymbol{\beta} z_{ij}} \phi_i^{z_{ij}} \exp\big\{ - (\sum_{l=1}^{k_{ij}-k_{ij-1}} \lambda_{k_{ij-1}+l}) e^{\mathbf{x}_i' \boldsymbol{\beta}} \phi_i \big\}] \\ \times ga(\phi_i | v) \Big) \prod_{m=1}^{M} g\{ \lambda_m | a(s_m - s_{m-1})\eta, \eta \}, \end{aligned} \tag{2}$$

which is crucial for the development of the Gibbs sampler described below.

3 Gibbs sampler

For the purpose of providing flexible modeling while also allowing for efficient posterior computation, we assign conventional vague priors for all of the parameters in the Bayesian approach. Specifically, we assign a multivariate normal $\mathcal{N}(\mu_0, \Sigma_0)$ prior for the regression coefficients $\boldsymbol{\beta}$, with mean vector zero and large independent variances such as 10 or 10^6. In practical, the very noninformative prior can balance both the skeptical and the enthusiastic views about the effects of covariates such as assigned treatmentsSinha (2004). We adopt independent Gamma$(1, 1)$ priors for v, a and η. By giving all the parameters fixed initial values 1, our Gibbs sampler iterates through the following steps:

1: Sample λ_m , for $m = 1, ..., M$, from

$$\lambda_m \sim \mathcal{G}amma(\sum_i \sum_j \sum_l u_{ijl} Z_{ij} + a(s_m - s_{m-1})\eta, \sum_i \exp(\mathbf{x}_i'\boldsymbol{\beta})\phi_i + \eta),$$

for all i, j, l such that $k_{ij-1} + l = m$.

2: Sample ϕ_i **from Gamma(**a_i**,**b_i**) with**

$$a_i = Z_i + v$$

$$b_i = \{\sum_{j=1}^{J_i} \sum_{l=1}^{k_{ij}-k_{ij-1}} \lambda_{k_{ij-1}+l} \exp(\mathbf{x}_i'\boldsymbol{\beta})\} + v,$$

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

3: Sample $U_{ij1}, ..., U_{ijn_{ij}} \sim Multinomial(1, (p_{ij1}, ..., p_{ijn_{ij}}))$, where

$$p_{ijl} = \frac{\lambda_{K_{ij-1}+l}}{\sum_{l=1}^{k_{ij}-k_{ij-1}} \lambda_{K_{ij-1}+l}} \qquad l = 1, ..., L$$

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

$$L(\boldsymbol{\beta}|\boldsymbol{.}) \propto \exp\Bigg\{\sum_{i=1}^{n}\Big\{Z_i\mathbf{x}_i'\boldsymbol{\beta} - (\sum_{j=1}^{J_i}\sum_{l=1}^{k_{ij}-k_{ij-1}}\lambda_{K_{ij-1}+l})\phi_i\exp(\mathbf{x}_i'\boldsymbol{\beta})\Big\} - (\boldsymbol{\beta}-\boldsymbol{\mu}_0)'\Sigma_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)/2\Bigg\}$$

6: Sample v, by using ARMS.

$$L(v|\boldsymbol{.}) \propto \exp(-v)\{\frac{v^v}{\Gamma(v)}\}^n(\prod_{i=1}^{n}\phi_i)^{v-1}\exp(-v\sum_{i=1}^{n}\phi_i)$$

7: Sample a by using ARMS.

$$L(a|\boldsymbol{.}) \propto \frac{\eta^{a\eta s_M}}{\prod_{m=1}^{M}\Gamma\{a(s_m-s_{m-1})\eta\}}\prod_{m=1}^{M}\lambda_m^{a(s_m-s_{m-1})\eta-1}\exp(-a)$$

8: Sample η

$$L(\eta|\boldsymbol{.}) \propto \frac{\eta^{a\eta s_M}}{\prod_{m=1}^{M}\Gamma\{a(s_m-s_{m-1})\eta\}}\prod_{m=1}^{M}\lambda_m^{a(s_m-s_{m-1})\eta-1}\exp(-\eta\sum_{m=1}^{M}\lambda_m)\exp(-\eta)$$

4 Simulation study

Comprehensive simulations are conducted to evaluate the proposed approach. To generate simulated data for each subject, we set 50 evenly allocated examination time points on time interval $(0, 10]$ to imitate the pre-decided research time span and observation scheme. Then we randomly remove 20% of examination time points for each of the subject. In this way, all the subject have different total observation times and gap times. 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 x_{i1} is continuous variable that follows a normal distribution, $N(0, 0.5^2)$ and x_{i2} is a binary variable that follows the Bernoulli distribution, $Bernoulli(0.5)$. The true regression coefficients are $\beta_1 = \{-1, 1\}$, $\beta_2 = \{-1, 1\}$. The distribution of ϕ_i is $Gamma(1, 1)$. We assessed the proposed approach with two different baseline mean function: $\mu_0(t) = log(1 + t) + t^{1.5}$ and $\mu_0(t) = t + sin(t)$. For each setting, 100 data sets with sample size $n = 100$ are generated.

The proposed Gibbs sampler in Section 3 is implemented to fit the Gamma frailty proportional mean model for each of the simulated data set. Table 1 present the frequentist operating characteristics of the estimates of the regression parameters. Bias is the difference between the average of 100 posterior means and the true parameter value; ESE is the average of the estimated posterior standard errors; SSD is the sample standard deviation of the 100 posterior means; and CP95 is the empirical coverage probability based on the 100 95% credible intervals.The results from our proposed method indicate that the proposed method performs well in terms of the estimation of the regression parameters, as the estimates show no bias, ESD and SSD are close to each other, and the coverage probabilities are all close to 0.95.

Table 1: Estimation of regression parameters based on 100 simulated data sets from the proposed Bayesian method. Empirical bias (BIAS), the average of the estimated standard errors (ESD) and standard deviation (ESD) of β, and the empirical coverage probabilities associated with 95% confidence probability (CP95).

		$\mu_0(t) = log(1+t) + t^{1.5}$				$\mu_0(t) = t + sin(t)$			
Method	(β_1,β_2)	BIAS	ESD	SSD	CP95	BIAS	ESD	SSD	CP95
GP	(-1,-1)	0.0217	0.2072	0.1715	0.98	0.0014	0.2228	0.2526	0.94
		0.0066	0.2128	0.2186	0.97	0.0145	0.2343	0.2567	0.94
SP		0.0427	0.2056	0.1697	0.97	-0.0185	0.2246	0.2501	0.95
		0.0030	0.2133	0.2218	0.93	0.0142	0.2341	0.2568	0.94
GP	(-1,1)	0.0250	0.2092	0.2119	0.93	-0.0366	0.2249	0.2403	0.93
		-0.0216	0.2126	0.2283	0.95	-0.0099	0.2323	0.2567	0.96
SP		0.0370	0.2111	0.2167	0.92	-0.0607	0.2277	0.2394	0.93
		-0.0168	0.2135	0.2301	0.92	-0.0089	0.2347	0.2545	0.96
GP	(1,-1)	-0.0038	0.1934	0.2096	0.93	0.0074	0.2101	0.2329	0.91
		0.0017	0.2032	0.2076	0.92	-0.0164	0.2194	0.2196	0.95
SP		0.0039	0.1933	0.2026	0.93	-0.0148	0.2106	0.2250	0.90
		0.0111	0.1994	0.2116	0.94	-0.0158	0.2211	0.2272	0.95
GP	(1,1)	0.0160	0.1997	0.2009	0.95	0.0248	0.2112	0.2284	0.93
		0.0355	0.2091	0.2005	0.94	-0.0016	0.2150	0.1986	0.97
SP		0.0170	0.1947	0.2013	0.91	-0.0034	0.2103	0.2245	0.94
		0.0295	0.2107	0.2075	0.91	-0.0050	0.2175	0.1981	0.97

5 Real-life data application

5.1 The patent study

We applied the proposed method to analysis of an industrial economics data set from the r package 'pglm'. The current data set is an extract from a larger data set that is collected by Hall et al. for their study of the relationship between patenting and research and development activity at the firm level by the U.S. manufacturing sector during the 1970's. This dataset

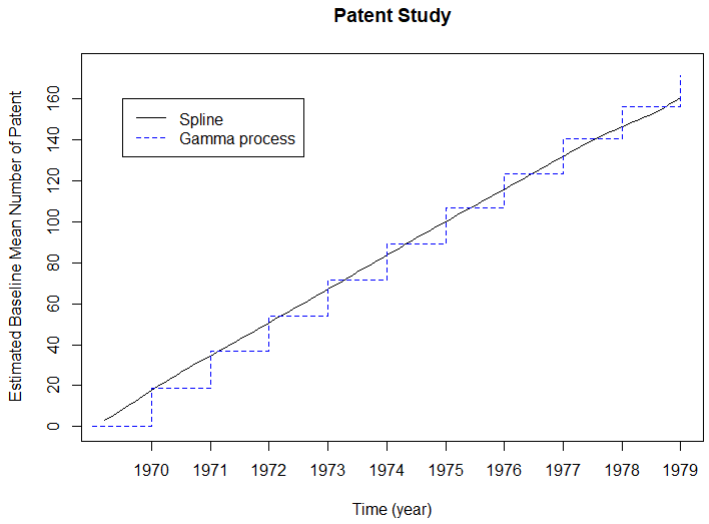
contains 346 firms in the United States. Among them, 147 firms are in the scientific sector. During 1970 to 1979, the number of patents applied for in each year that were eventually granted is recorded for every firm. The data set also includes all the firms' book value of capital in 1972 and their annual research and development (R & D) spending.

In this section, the primary objective of our analysis is to assess the relationship between the mean number of patents and the characteristics of the firm. x_{i1} is binary variable that indicate if the firm i is in the scientific sector. x_{i2} and x_{i3} are the book value of capital in 1972 and the average annual research and development (R & D) spending for the firm i, respectively. To mitigate the problem of collinearity, we standardized x_2 and x_3 before fitting the model. For the purpose of comparison, we also analyzed this data set with GFNPM (Yao et al. 2016).

As shown in Table 2, the estimation of regression coefficients from both methods are accordance with each other. The result indicates that the mean number of patent applied by firms in the scientific sector is 0.7 times higher than firms that not in the scientific sector. At the same time, a firm's book value and its R & D spending have significant positive effect to the patenting development. In Figure 5.1, we superimposed the estimated baseline mean functions of patent counts between 1970 and 1979 obtained by both methods. The two lines are very close to each other, which implies the proposed method provides a similar estimation of the baseline mean function to GFNPM.

Table 2: Patent data analysis from the proposed approach and GFNPM. Summarized results are the point estimates (Point), the standard errors (SE), and the 95% credible interval for all the regression parameters and the frailty variance parameter v.

	GFGP			GFNPM		
	Point	SE	CI95	Point	SE	CI95
$\hat{\beta}_1$	0.537	0.146	(0.240,0.834)	0.561	0.127	(0.311, 0.798)
$\hat{\beta}_2$	1.032	0.148	(0.812,1.427)	1.130	0.122	(0.818, 1.303)
$\hat{\beta}_3$	0.617	0.144	(0.335,0.801)	0.795	0.144	(0.483, 1.005)
$\hat{v}$	0.549	0.037	(0.480,0.624)	0.555	0.037	(0.485, 0.630)



5.2 The bladder tumor study

We also 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 & 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 & Zhang (2007), 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.

Table 3: 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 95% credible interval for all the regression parameters and the frailty variance parameter v.

	GFGP			GFNPM			WZ		
	Point	SE	CI95	Point	SE	CI95	Point	SE	CI95
$\hat{\beta}_1$	0.333	0.107	(0.131,0.550)	0.336	0.106	(0.128,0.544)	0.207	0.078	(0.054,0.360)
$\hat{\beta}_2$	0.001	0.122	(-0.224,0.244)	0.012	0.120	(-0.223,0.247)	0.036	0.086	(-0.133,0.205)
$\hat{\beta}_3$	-0.021	0.427	(-0.851,0.833)	-0.033	0.409	(-0.835,0.769)	0.066	0.431	(-0.779,0.911)
$\hat{\beta}_4$	-1.152	0.427	(-2.051,-0.261)	-1.140	0.435	(-1.993,-0.287)	0.797	0.360	(0.091,1.503)
$\hat{v}$	0.326	0.058	(0.225, 0.453)	0.351	0.062	(0.229,0.473)	-	-	-

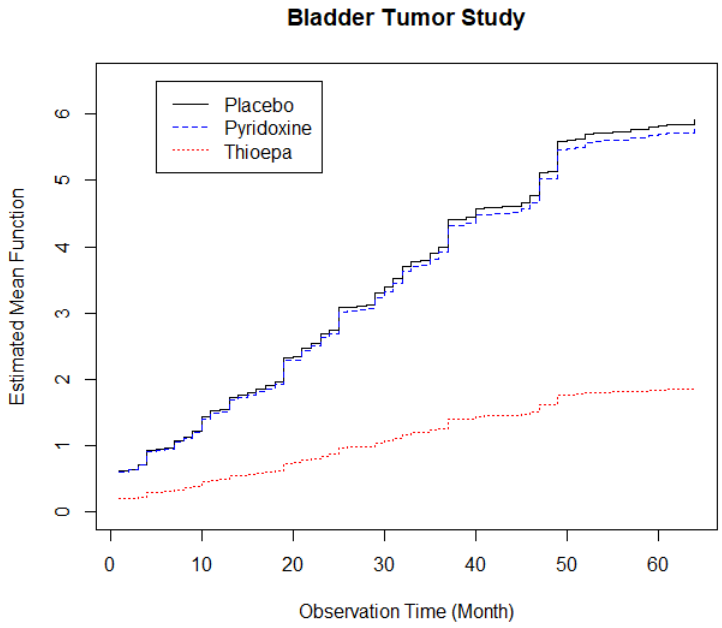


Table 3 shows the results from the proposed approach and two other competitive approaches, i.e. Yao et al. (2016) and Wellner & Zhang (2007). The results from these two competitors are directly drawn from their papers. Both of these two competitive approaches are

likelihood-based approaches under the non-homogeneous Poisson model. Yao et al. (2016)'s method considered the within-subject correlation while Wellner & Zhang (2007)'s method did not consider the within-subject correlation.

As seen in Table 3, 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 5.2 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 & 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.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 & Zhang 2012).

6 Discussion

This article introduces a Bayesian approach which can fit the gamma frailty non-homogeneous Poisson process model for panel count data. The approach decomposed each observation time interval using the idea of innermost interval, which results in the ability of processing con-

tinuous observation and missing data. The method estimates the baseline mean function nonparametrically by adopting gamma process prior, which allows us to develop a straightforward and easy to implement Gibbs sampler. This approach have appealing numerical performance in terms of providing efficient, accurate and reliable estimation of regression coefficients and baseline mean function. Additionally, because of the intrinsic connection between panel count data and interval censored data this framework can be further extended to fit interval censored data without introducing any complexity of the optimization problem. Because of using gamma process as prior, one limitation of our approach is that the realization $\Lambda(t)$ is discrete with probability one. We hope to tackle this potential issue in the future research.

References

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.

Cox, D. R. (1972), 'Regression models and life tables', *Journal of the Royal Statistical Society* **34**, 187–220.

Doksum, K. (1974), 'Tailfree and neutral random probabilities and their posterior distributions', *The Annals of Probability* **2**, 183–201.

Ferguson, S. T. & Phadia, G. E. (1979), 'Bayesian nonparametric estimation based on censored data', *The Annals of Statistics* **1**, 163–186.

Gentleman, R. & Vandal, A. (2002), 'Nonparametric estimation of the bivariate cdf for arbitrarily censored data', *The CanadianJournal of Starisrics* **30**, 557–571.

Hall, B., Zvi, G. & Jerry, H. (1986), 'Patents and r and d: Is there a lag?', *International Economic Review* **27**, 265–283.

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', *Canadian Journal of Statistics* **42**, 217–245.

Kalbfleisch, D. J. (1978), 'Non-parametric bayesian analysis of survival time data', *Journal of the Royal Statistical Society* **40**, 214–221.

Lawless, J. & Crowder, M. (2004), 'Covariates and random effects in a gamma process model with application to degradation and failure', *Lifetime Data Analysis* **10**, 213–227.

Li, L., Watkins, T. & Yu, Q. (1997), 'An em algorithm for smoothing the selfconsistent estimator of survival functions with interval-censored data', *Scandinavian Journal of Statistics* **24**, 531–542.

Nozer, D. S. (1995), 'Survival in dynamic envirnoments', *Statistical Science* **10**(1), 86–103.

Sinha, A., Chi, Z. & Chen, M. (2015), 'Bayesian inference of hidden gamma wear process model for survival data with ties', *Statistical Sinica* **25**, 1613–1635.

Sinha, D. (2004), 'A bayesian approach for the analysis of panel-count data with dependent termination', *Biometrics* **60**, 34–40.

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

Turnbull, B. (1976), 'The empirical distribution function with arbitrarily grouped, censored and truncated data', *Journal of the Royal Statistical Society* **38**, 290–295.

Wang, X. (2009), 'Nonparametric estimation of the shape function in a gamma process for degradation data', *The Canadian Journal of Statistics* **37**(1), 102–118.

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, A. & Zhang, Y. (2007), 'Two likelihood-based semiparametric estimation methods for panel count data with covariates', *The Annals of Statistics* **35**(5), 2105–2142.

Xu, D., Zhao, H. & Sun, J. (2018), 'Joint analysis of interval-censored failure time data and panel count data', *Lifetime Data Analysis* **24**, 94–109.

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.

Zhu, L., Zhang, Y., Li, Y., Sun, J. & Robison, L. L. (2018), 'A semiparametric likelihood-based method for regression analysis of mixed panel-count data', *Biometrics* **74**, 488–497.