Electronic Journal of Statistics
Vol. 0 (2023) 1–17
ISSN: 1935-7524

Bayesian regression analysis of panel count data under frailty nonhomogeneous Poisson process model with an unknown frailty distribution

Lu Wang[*]

Department of Mathematics, Western New England University, Springfield, U.S.A
e-mail: wanglustat@gmail.com

Chunling Wang

Wells Fargo Bank, San Francisco, U.S.A
e-mail: pennywang527@gmail.com

Xiaoyan Lin and Lianming Wang

Department of Statistics, University of South Carolina, Columbia, U.S.A
e-mail: lin@stat.sc.edu*;* wangl@stat.sc.edu

Abstract: Panel count data arise when recurrent events are observed periodically in a study. The response variable of interest is the number of recurrent events within different time windows instead of the exact onset times of the events. The gamma frailty Poisson process model has been proposed to accommodate the within-subject correlation and overdispersion in panel count data. Although the existing methods based on the gamma frailty Poisson process model have shown some robustness against frailty distribution misspecifications, they are also found to produce biased estimates in some other cases when the gamma frailty assumption is violated. In this paper, we generalize the gamma frailty Poisson process model to allow an unknown frailty distribution for analyzing panel count data. Specifically the frailty distribution is modeled nonparametrically by assigning a Dirichlet Process Gamma Mixture prior. An efficient Gibbs sampler is developed to facilitate the Bayesian computation. Extensive simulation results suggest that the proposed Bayesian approach has an excellent performance in estimating the regression parameters and the baseline mean function and outperforms the corresponding Bayesian method based on the gamma frailty Poisson model when the gamma frailty distribution is misspecified. The proposed method is applied to a skin cancer dataset for an illustration.

Keywords and phrases: Bayesian nonparametric, Dirichlet process mixture, frailty, panel count data, Poisson process.

1. Introduction

Panel count data commonly arise when recurrent events are observed periodically in a study. The response variable of interest is the number of recurrent

[*] Corresponding author: Lu Wang

events within different time windows instead of the exact onset times of the events. Panel count data often have the following characteristics: (1) subjects may be observed or examined at different time points during the study period; (2) the number of observations varies from subject to subject. When covariates are not considered, the focus of panel count data analysis is to estimate the mean function, and existing work include [27] adopting isotonic regression techniques, [34] based on nonhomogeneous Poisson process models, and [18] approximating the mean function with monotone splines among others.

When covariates are considered, the major research goals in the analysis of panel count data are to estimate the mean function for the recurrent event over time and to identify significant covariates and assess their effects on the response. Such goals are completed by regression analysis, and many methods have been developed in this category. For example, [28] proposed an estimation approach based on the proportional mean model when both observation and censoring times may depend on covariates; [39] studied a semiparametric pseudolikelihood estimation method based on the nonhomogeneous Poisson process proportional mean model; [35] considered both pseudo-likelihood estimator and maximum likelihood estimator under the same model; [19] studied the spline-based sieve version of the MLE by approximating the baseline mean function using monotone B-spline functions; and [10] developed estimating equations for a class of marginal mean models which leave the dependence structures for related types of recurrent events completely unspecified.

Despite the popularity of the non-homogeneous Poisson process model for analyzing panel count data, this model fails to accommodate the overdispersion [12, 40] and the within-subject correlation [36] for panel count data. In general, failing to address the overdispersion will cause underestimation of the standard errors of the regression parameter estimates and thus lose estimation efficiency [4, 12]. The gamma frailty Poisson model has been popular to address these problems in the analysis of panel count data. Among others, [40] proposed an EM algorithm and [11] developed an estimation approach based on estimating equations when there are no covariates, [36] studied the within-subject correlation in panel count data and developed an efficient EM algorithm, and [13] developed a sieve maximum likelihood method adopting monotone B-splines for the baseline mean function and established the asymptotic properties of their spline-based estimator, all using the gamma frailty Poisson process model. [12] developed a spline-based generalized estimating equation (GEE) method and showed that the GEE method is actually equivalent to the likelihood method based on the gamma-frailty Poisson process model.

Both [13] and [36] have shown some robustness of their methods under the gamma frailty Poisson process models in some cases where the gamma frailty distribution is misspecified. However, both of their simulation studies also reveal that their performances are less desirable in some other cases of misspecification of the frailty distribution. To tackle this problem, we propose a more general frailty Poisson process model with unknown frailty distribution and develop an efficient Bayesian estimation approach. Specifically, the unknown frailty distribution is modeled by the Dirichlet process Gamma mixture. Under this model, the within-subject correlation can be quantified in terms of Pearson's correlation with an explicit form. An efficient Gibbs sampler is developed for Bayesian posterior computation. The proposed approach has an excellent performance in estimating both the regression coefficients and the baseline mean function under a variety of frailty distributions in the simulation study. It outperforms the corresponding Bayesian competitor under the Gamma frailty Poisson process model.

The rest of the paper is organized as follows. Section 2 presents the frailty non-homogeneous Poisson process model, monotone splines for modeling the baseline mean function, and the Dirichlet process Gamma mixture for the frailty distribution. Section 3 provides the details of our exact block Gibbs sampler for the posterior computation. In Section 4, extensive simulation studies are conducted to evaluate the performance of our approach in a variety of misspecifications of the frailty distribution. Section 5 provides an illustration of our proposed method and its comparison with existing methods via an analysis of the skin cancer data. Finally, some concluding remarks are given in Section 6.

2. The proposed model

2.1. A nonparametric frailty Poisson process model

Consider a study that consists of n independent subjects. It is assumed in this paper that the observational process and the recurrent event process are conditionally independent given covariates. Let $N_i(\cdot)$ denote the recurrent event counting process for subject i, and $\{t_{ij}, j = 1, \ldots, J_i\}$ the set of examination times for subject i, and $\mathbf{x}_i$ a vector of $q \times 1$ time-independent covariates for subject i. The observed data are $\{\mathbf{x}_i, N(t_{ij}), j = 1, \ldots, J_i, i = 1, \ldots, n\}$.

We propose a general frailty non-homogeneous Poisson process model as follows. Conditioning on frailty ϕ_i, $N_i(\cdot)$ is a non-homogeneous Poisson process with mean function $\mu_0(\cdot)\exp(\mathbf{x}_i'\boldsymbol{\beta})\phi_i$, where $\mu_0(\cdot)$ is an unspecified non-decreasing baseline mean function with $\mu_0(0) = 0$, and the frailty ϕ_i follows a continuous distribution with density function f. We allow f to be unknown and model it nonparametrically in this paper. The details are to be given in the next subsection. It is worth noting that taking f to be a gamma density with a common shape and rate parameter leads to the popular gamma frailty Poisson process model (GFPM). The GFPM was introduced to deal with overdispersion [12, 13] and to account for the within-subject correlation among the counts within the same subject [36] in the regression analysis of panel count data.

Define $Z_{ij} = N_i(t_{ij}) - N_i(t_{ij-1})$ as the count of recurrent events within time interval $(t_{ij-1}, t_{ij}]$ for $j = 1, \ldots, J_i$ and $i = 1, \ldots, n$, with $t_{i0} = 0$ for each i for notational convenience. By the properties of non-homogeneous Poisson process, all Z_{ij}'s are conditionally independent given ϕ_i and

$$Z_{ij}|\phi_i \sim \text{Poisson}\left[\{\mu_0(t_{ij}) - \mu_0(t_{ij-1})\}\exp(\mathbf{x}_i'\boldsymbol{\beta})\phi_i\right]$$

for $j = 1, \ldots, J_i$ and $i = 1, \ldots, n$. Thus, the observed likelihood is given by

$$L_{obs} = \prod_{i=1}^{n} \int_{0}^{\infty} \left\{ \prod_{j=1}^{J_i} \mathcal{P}(z_{ij}|\phi_i) \right\} f(\phi_i) d\phi_i, \tag{2.1}$$

where $\mathcal{P}(\cdot|\phi_i)$ is the conditional Poisson probability mass function with mean $\{\mu_0(t_{ij}) - \mu_0(t_{ij-1})\} \exp(\mathbf{x}_i'\boldsymbol{\beta})\phi_i$ for each i.

2.2. Dirichlet process Gamma mixture for the frailty density

There is extensive work of modeling frailty distribution nonparametrically in the Bayesian literature. Potential methods include Dirichlet process (DP) [5], Dirichlet process mixture [1], Polya tree [14, 15], mixtures of Polya tree models [9], and Bernstein polynomials [24], etc. Among these methods, Dirichlet process priors are particularly useful when modeling unknown distributions with certain level of clustering characteristics, and such studies include [7], [20], and [22], among many others. However, a DP is not suitable to model unknown densities due to the almost sure discreteness of the random measure generated by the DP [3]. Instead, the Dirichlet process mixture (DPM) prior is naturally adopted in density estimation using a DP prior for the mixing distribution. The DPM is popular since it provides a smooth estimation of the density, and existing works based on the DPM include [6], [8], and [31], among many others.

In many frailty or random effect models, it is usually necessary to set a certain constraint to ensure the identifiability of the models. For instance, a common constraint is to set the mean or median of the frailty to be 0 or 1. In our study, the mean of the frailty is set to be 1 for the purpose of identifiability. We propose the following Dirichlet process gamma mixture (DPGM) for the frailty density f,

$$f(\phi_i) = \int g(\phi_i|\tau_i) d\pi(\tau_i), \quad \tau_i|\pi \sim \pi, \quad \pi \sim DP(\alpha G_0),$$

where $g(\cdot|\tau_i)$ is a gamma density function with common shape and rate parameter τ_i, and $DP(\alpha G_0)$ is a Dirichlet process with a precision parameter α and a base measure G_0 that has support on $(0, \infty)$. The use of common shape and rate parameters in the gamma kernel guarantees that the nonparametric density f has a mean 1 constraint.

Based on the stick-breaking construction of DP [26], the random distribution π can be expressed as

$$\pi = \sum_{h=1}^{\infty} p_h \delta_{\theta_h}, \quad p_h = V_h \prod_{k<h} (1 - V_k),$$

where δ_{θ_h} is a Dirac probability measure concentrated at θ_h, θ_h's are i.i.d. from G_0, and V_h's are i.i.d. random variables from $Beta(1, \alpha)$. Under this stick-breaking representation of DP, the unknown frailty density $f(\phi_i)$ can be written as a mixture of gamma with an infinite number of components as follows,

$$f(\phi_i) = \sum_{h=1}^{\infty} p_h g(\phi_i|\theta_h). \tag{2.2}$$

The inclusion of frailty ϕ_i induces correlations among the counts Z_{ij}'s for subject i, which is called the within-subject correlation. This is one way to handle the overdispersion problems appearing in the corresponding non-frailty models. Additionally, the within-subject correlation can be quantified in a simple closed form. Specifically, we consider the counts Z_1 and Z_2 of recurrent events within two non-overlapping intervals $(t_1, t_2]$ and $(t_3, t_4]$ from the same subject with covariates $\mathbf{x}$. It can be shown that Pearson's correlation coefficient between Z_1 and Z_2 takes the following form

$$\rho(Z_1, Z_2) = \frac{1}{\sqrt{\{1+\lambda_1^{-1}\text{var}(\phi)^{-1}\}\{1+\lambda_2^{-1}\text{var}(\phi)^{-1}\}}}, \tag{2.3}$$

where $\lambda_1 = \{\mu_0(t_2) - \mu_0(t_1)\}\exp(\mathbf{x}'\boldsymbol{\beta})$ and $\lambda_2 = \{\mu_0(t_4) - \mu_0(t_3)\}\exp(\mathbf{x}'\boldsymbol{\beta})$ are the mean numbers of the recurrent event within the time intervals $(t_1, t_2]$ and $(t_3, t_4]$, respectively, and $\text{var}(\phi) = \sum_{h=1} p_h \theta_h^{-1}$ is the variance of the frailty with the proposed DPGM prior (2.2). This expression (2.3) suggests that both the mean numbers of the recurrent events and the variance of the frailty term affect the within-subject correlation. This finding agrees with the conclusion in [36] for the gamma frailty Poisson process model.

3. The proposed method

3.1. Monotone splines

Since $\mu_0(\cdot)$ is an unspecified positive non-decreasing function over the positive real line, estimating $\mu_0(\cdot)$ can be challenging because it is infinitely dimensional. For a finite sample size, the number of parameters involved in $\mu_0(\cdot)$ is on the order of sample size when the observation times vary from subject to subject. To overcome this difficulty, we model the baseline mean function $\mu_0(\cdot)$ with the monotone spline of [25] in the following manner,

$$\mu_0(t) = \sum_{l=1}^{L} \gamma_l b_l(t), \tag{3.1}$$

where $b_l(\cdot)$'s are integrated spline (I-spline) basis functions and γ_l's are nonnegative spline coefficients to ensure that $\mu_0(\cdot)$ is nondecreasing. Each I-spline basis function is a piecewise polynomial that starts from 0 in the initial region, increases in the mid-region, and plateaus at 1 in the final region [25]. Here the nonnegative constraints of γ_l's are a simple but sufficient condition to keep the monotonicity of $\mu_0(\cdot)$.

The use of monotone spline (3.1) is very appealing because it provides great flexibility for modeling monotone functions with only a finite number of parameters [29, 36]. The spline basis functions are totally determined by the specification of the knots and degree. In general, the degree determines the overall smoothness, and the placement of the knots determines the shape of these basis functions. Setting the degree to 2 and 3 results in piecewise linear and quadratic polynomials, respectively, and both provide adequate smoothness in practice. The placement of knots can be either equally spaced or quantile based. The number of interior knots affects the modeling flexibility and thus potentially affects the estimation performance although many spline-based approaches have shown a robust performance over such spline specifications [21, 33]. [25] recommended using only a few knots, say, at the median or at the three quartiles. [16], [17], and others have advocated that using 10 $\sim$ 30 knots provides adequate modeling flexibility for various types of survival data with a sample size up to hundreds of thousands [16, 17, 32] in Bayesian survival literature. They are using a shrinkage prior for the spline coefficients, which functions to shrink the coefficients of those unnecessary basis functions towards zero and thus prevents the over-fitting problem that is potentially caused by using too many knots. We adopt the same strategy in this paper.

3.2. Data augmentation

The observed likelihood (2.1) is difficult to work with since it involves multiple integrals that do not have closed forms. Instead, we consider the following conditional likelihood by treating all the frailties ϕ_i's as missing data,

$$L_{con} = \prod_{i=1}^{n} f(\phi_i) \prod_{j=1}^{J_i} \quad (Z_{ij}!)^{-1} [\{\mu_0(t_{ij}) - \mu_0(t_{ij-1})\} \exp(\mathbf{x}_i'\boldsymbol{\beta})\phi_i]^{Z_{ij}} \\ \times \exp[-\{\mu_0(t_{ij}) - \mu_0(t_{ij-1})\} \exp(\mathbf{x}_i'\boldsymbol{\beta})\phi_i].$$

Ignoring the multiplicative constants and plugging in the monotone spline representation of $\mu_0(\cdot)$, the conditional likelihood takes the form

$$L_{con} \propto \prod_{i=1}^{n} f(\phi_i) \prod_{j=1}^{J_i} \left\{ \sum_{l=1}^{L} \gamma_l B_{lij} \exp(\mathbf{x}_i'\boldsymbol{\beta})\phi_i \right\}^{Z_{ij}} \exp\left\{ -\sum_{l=1}^{L} \gamma_l B_{lij} \exp(\mathbf{x}_i'\boldsymbol{\beta})\phi_i \right\}, \tag{3.2}$$

where $B_{lij} = b_l(t_{ij}) - b_l(t_{ij-1})$ for $l = 1, \ldots, L$, $j = 1, \ldots, J_i$ and $i = 1, \ldots, n$.

The above conditional likelihood involves summations in the product, which will cause much trouble in the posterior computation. To solve this problem, we decompose each z_{ij} as a sum of conditionally independent Poisson random variables $Z_{ij} = \sum_{l=1}^{L} Z_{ijl}$ with

$$Z_{ijl} | \phi_i \sim \text{Poisson}\{\gamma_l B_{lij} \exp(\mathbf{x}_i'\boldsymbol{\beta})\phi_i\},$$

for $l = 1, \ldots, L$, $j = 1, \ldots, J_i$ and $i = 1, \ldots, n$. Treating all Z_{ijl}'s as latent variables, the augmented data likelihood has the following form

$$L_{com} \propto \prod_{i=1}^{n} f(\phi_i) \prod_{j=1}^{J_i} \prod_{l=1}^{L} \Big\{ \gamma_l B_{lij} \exp(\mathbf{x}_i' \boldsymbol{\beta}) \phi_i \Big\}^{Z_{ijl}} \exp \Big\{ -\gamma_l B_{lij} \exp(\mathbf{x}_i' \boldsymbol{\beta}) \phi_i \Big\}. \tag{3.3}$$

This likelihood is appealing with only multiplicative terms and will be treated as the complete data likelihood for our posterior computation.

3.3. Prior specifications

We need to specify the prior distributions for all of the unknown parameters. Following the idea of [16] and [32] among others, we adopt shrinkage priors for the spline coefficients γ_l's. Specifically we assign independent exponential priors $\mathcal{E}(\lambda)$ for γ_l's and assign a Gamma prior $\mathcal{G}a(a_\lambda, b_\lambda)$ for the hyperparameter λ. This prior specification is closely related to Bayesian Lasso [23] and is equivalent to the penalized likelihood approach with L_1 penalty imposed on the spline coefficients, where λ serves as a tuning parameter. These shrinkage priors function to penalize large values of the spline coefficients and shrink those coefficients of the unnecessary spline basis functions towards 0, thus preventing over-fitting.

To complete the prior specifications, we assign a multivariate normal prior $\mathcal{N}(\mu_0, \Sigma_0)$ for $\boldsymbol{\beta}$ and a gamma prior $\mathcal{G}a(a_\alpha, b_\alpha)$ for the DP precision parameter α. The base distribution G_0 in the DP is specified as $\mathcal{G}a(a_0, b_0)$ in our paper.

3.4. Gibbs sampler

The posterior distribution is proportional to the product of the complete likelihood (3.3) and all of the prior densities including the DPM prior for the frailty density f in Section 2.2 and the priors in Section 3.3. For the posterior computation, we propose an efficient block Gibbs sampler by considering the full conditional distributions for each parameter and latent variable.

To develop an efficient posterior computation algorithm, we first rewrite the proposed DPGM into the following hierarchical form,

$$\phi_i \overset{ind}{\sim} \mathcal{G}(\tau_i, \tau_i), \quad \tau_i | \pi \overset{iid}{\sim} \pi, \quad \pi = \sum_{h=1}^{\infty} p_h \delta_{\theta_h}$$

for $i = 1, \cdots, n$. To avoid sampling from the infinite components of the DP, we propose an exact block Gibbs sampler [37] that only requires us to sample from a mixture of a finite number N of components for the frailty distribution without the need of truncating the DP [31, 37]. Let $\{\theta_1, \cdots, \theta_N\}$ denote the set of unique values in $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$ and $\mathcal{K} = \{K_1, ..., K_n\}$ denote the vector of configuration indicators so that $\tau_i = \theta_{K_i}$. Let $m_h = \sum_{i=1}^{n} I(K_i = h)$ be the number of subjects that share the same gamma frailty distribution with parameter θ_h for $h = 1, \ldots, N$. Our algorithm allows N to increase when an additional component $\theta_h \sim G_0$ is needed for the DP.

In the proposed Gibbs sampler, the initial values of all unknown parameters and latent variables are generated from their prior distributions. Specifically for the parameters involved in the DPGM, we initialize N to be 2 and sample θ_h's independently from $G_0 = \mathcal{G}a(a_0, b_0)$ for $h = 1, \ldots, N$. Sample α from $\mathcal{G}a(a_\alpha, b_\alpha)$ and V_h's independently from $Beta(1, \alpha)$ for $h = 1, \ldots, N$ and then calculate p_h's according to their relationship with V_h's in section 2.2. The initial value of K_i takes the position of 1 in the vector generated from Multinomial$(1, \mathbf{p})$ for each i, where $p = (p_1, \ldots, p_N)$. Then the frailty ϕ_i is generated from $\mathcal{G}a(\theta_{K_i}, \theta_{K_i})$ for each i. After generating the initial values for all unknowns, the proposed exact Gibbs sampler cycles through the following steps:

1. For each i and j, sample $(Z_{ij1}, ..., Z_{ijL})$ from Multinomial$(z_{ij}, (q_{ij1}, ..., q_{ijL}))$, where
$$q_{ijl} = \frac{\gamma_l\{b_l(t_{ij}) - b_l(t_{ij-1})\}}{\sum_{j=1}^{L} \gamma_j\{b_j(t_{ij}) - b_j(t_{ij-1})\}}, \qquad l = 1, \ldots, L.$$
2. For each $l = 1, \ldots, L$, sample γ_l from
$$\mathcal{G}a\left(\sum_{i=1}^{n}\sum_{j=1}^{J_i} Z_{ijl} + 1, \sum_{i=1}^{n}\{b_l(t_{iJ_i}) - b_l(t_{i0})\}\exp(\mathbf{x}_i'\boldsymbol{\beta})\phi_i + \lambda\right).$$
3. Sample λ from $\mathcal{G}a(a_\lambda + L, b_\lambda + \sum_{l=1}^{L}\gamma_l)$.
4. Sample $\boldsymbol{\beta}$ by using adaptive rejection metropolis sampling (ARMS) from the following full conditional distribution
$$\begin{aligned} L(\boldsymbol{\beta}|\cdot) \propto \quad & \exp\left\{\sum_{i=1}^{n}\sum_{j=1}^{J_i} Z_{ij}\mathbf{x}_i'\boldsymbol{\beta} - \sum_{i=1}^{n}\{\mu_0(t_{iJ_i}) - \mu_0(t_{i0})\}\phi_i\exp(\mathbf{x}_i'\boldsymbol{\beta})\right. \\ & \left. -\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)'\Sigma_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)\right\}. \end{aligned}$$
5. Sample ϕ_i from $\mathcal{G}a(Z_i + \tau_i, \{\mu_0(t_{iJ_i}) - \mu_0(t_{i0})\}\exp(\mathbf{x}_i'\boldsymbol{\beta}) + \tau_i)$ for each i.
6. Sample θ_h from the following full conditional distribution
$$\theta_h \propto \theta_h^{a_0 - 1}\exp(-b_0\theta_h)\prod_{\{i: K_i = h\}}\frac{\theta_h^{\theta_h}}{\Gamma(\theta_h)}\phi_i^{\theta_h - 1}\exp(-\theta_h\phi_i)$$
by using ARMS for $h = 1, ..., N$.
7. Sample V_h from $Beta(1 + m_h, \alpha + \sum_{s=h+1}^{N} m_s)$, for $h = 1, ..., N$. Then calculate $p_1 = V_1$ and $p_h = V_h(1 - V_{h-1})\cdots(1 - V_1)$ for $h = 2, ..., N$.
8. Sample $U_i \sim$ Uniform$(0, p_{K_i})$, for $i = 1, \ldots, n$.
9. Sample K_i for $i = 1, \ldots, n$ as follows. Let $U^* = \min(U_1, \cdots, U_n)$.

(a) If $\sum_{h=1}^{N} p_h > 1 - U^*$, sample K_i from Multinomial$(1, q_i)$, where $q_i = (q_{i1}, ..., q_{iN})$ and

$$q_{ih} = \frac{p_h g(\phi_i|\theta_h)}{\sum_{l=1}^{N} p_l g(\phi_i|\theta_l)}$$

for $h = 1, \ldots, N$ and $i = 1, \ldots, n$. Then update $\tau_i = \theta_{K_i}$ for $i = 1, \ldots, n$.

(b) Otherwise, keep updating $N = N + 1$ and sampling $V_N \sim Beta(1, \alpha)$ and $\theta_N \sim \mathcal{G}a(a_0, b_0)$ until $\sum_{h=1}^{N} p_h > 1 - U^*$. Then sample K_i as in (a).

10. Sample α from $\mathcal{G}a(a_\alpha + N, b_\alpha - \sum_{h=1}^{N} \log(1 - V_h))$.

This Gibbs sampler allows us to sample from the DPGM without any approximation with only a finite number N of components. It allows N to increase in the algorithm when a new component is needed. A truncated version of DPGM with a fixed N can be also used but is less desirable here because in general using an overly large value of N will result in a loss of computational efficiency while using a too small value will restrict the flexibility of the frailty distribution and affect the overall estimation performance.

Steps 6-10 in the Gibbs sampler are used to update the parameters involved in the DPGM prior for the nonparametric frailty. The proposed algorithm can be simplified to analyze panel count data under the gamma frailty Poisson process model by deleting steps 6-10. This simplified version was reported in [30] and is a Bayesian version of [36] under the gamma frailty Poisson process model.

4. Simulation study

Extensive simulation studies were conducted to assess the performance of the proposed approach. We considered the following scenarios for the frailty distribution,

I. $\phi_i \sim \mathcal{G}a(0.5, 0.5)$;
II. $\phi_i \sim 0.5\mathcal{G}a(1, 1) + 0.5\mathcal{G}a(15, 15)$;
III. $\phi_i \sim \mathcal{LN}(-0.55, 1.05)$;
IV. $\phi_i \sim \mathcal{LL}(\pi, \sin(1))$;
V. $\phi_i \sim 0.5\mathcal{LN}(-0.11, 0.47) + 0.5\mathcal{LN}(-0.35, 0.83)$;
VI. $\phi_i \sim 0.5\mathcal{G}a(10, 10) + 0.5\mathcal{LN}(-0.55, 1.05)$,

where $\mathcal{LN}$ and $\mathcal{LL}$ represent log-normal distribution and log-logistic distribution, respectively. All these frailty distributions have the mean 1 constraint.

For each scenario, we simulated 500 data sets with sample size $n = 100$ for each. To generate the observational process for subject i, we first generated J_i, the total number of observation times, from 1 plus a Poisson distribution with a mean of 6. This mechanism guarantees a minimum of one observation time per subject, while also allowing for varying numbers of observation times among subjects. Then we generated J_i gap times independently from an exponential distribution with a rate parameter 2 to form up the observation times $\{t_{ij}, j =$

$1, \ldots, J_i\}$ for subject i. We generated ϕ_i from a frailty distribution in Scenarios I∼VI, and conditioning on ϕ_i, the counting process associated with subject i was generated from the following model,

$$Z_{ij} = N_i(t_{ij}) - N_i(t_{ij-1}) \sim \text{Poisson}\left[\{\mu_0(t_{ij}) - \mu_0(t_{ij-1})\}\exp(x_{i1}\beta_1 + x_{i2}\beta_2)\phi_i\right],$$

where $\mu_0(t) = \log(1+t) + t$, $x_{i1} \sim Bernoulli(0.5)$, $x_{i2} \sim N(0, 0.5^2)$, and the true values of regression coefficients (β_1, β_2) were taken to be $(1, -1)$ or $(-1, 1)$.

As reported in the literature, methods based on the monotone splines are relatively robust with respect to the degree and number of knots when the knots are adequate [29, 33, 36]. Here we fixed the order of monotone spline as 3 for adequate smoothing and used 18 equally-spaced interior knots in all the simulations for illustration. In the simulation, we also implemented the Bayesian approach under the gamma frailty Poisson process model mentioned in Section 3.4 as a benchmark for comparison with the proposed approach. It was observed that all the chains mix well and converge fast for both approaches. For each method, we ran MCMC with a total of 5000 iterations with the first 1000 iterations as a burn-in.

Table 1 summarizes the simulation results from the proposed method (termed as DPGM-PM) using the Dirichlet process gamma mixture prior for the frailty distribution and the Bayesian method (termed as GFPM) under the gamma frailty Poisson model in the six different scenarios of frailty distributions. The presented results include the estimation of (β_1, β_2) from the two methods in terms of bias, the difference between the average of 500 posterior means and the true parameter value; ASD, the average of the estimated posterior standard deviations; 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, both the proposed method, DPGM-PM, and GFPM exhibit commendable performance in Scenario I, where the true frailty distribution is gamma. Both methods demonstrate minimal bias in the point estimate, the ASD in close agreement with SSD, and 95% coverage probability close to the nominal value of 0.95. In this Scenario, DPGM-PM yields results comparable to GFPM regarding bias, estimated standard deviations, and coverage probability.

However, in Scenarios II - VI where the true frailty distributions are non-gamma, GFPM's performance is unsatisfactory, with a large discrepancy between ASD and SSD and a poor coverage probability. It is clear that DPGM-PM has a better estimation performance than GFPM across these scenarios, presenting smaller differences between ASD and SSD, and a 95% coverage probability that aligns more closely with the nominal value of 0.95. Also, DPGM-PM produces smaller ASDs and SSDs than GFPM in all settings, indicating that DPGM-PM is more efficient than GFPM in estimating the regression parameters. These results demonstrate the superior performance of DPGM-PM over GFPM when the gamma frailty assumption is violated.

Regarding the baseline mean function estimation, Figure 1 shows the true baseline mean function and the average of the baseline mean function estimates

Table 1

Simulation results from the proposed method (DPGM-PM) using the Dirichlet process gamma mixture prior for the frailty distribution and the Bayesian method (GFPM) under the gamma frailty Poisson model in six different scenarios (I $\sim$ VI) of frailty distributions. Summarized results include the bias (Bias), the average of the estimated posterior standard deviations(ASD), the sample standard deviation of the 500 posterior means (SSD), and the 95% coverage rate (CP95).

			DPGM-PM				GFPM			
	(β_1, β_2)	Est	Bias	ASD	SSD	CP95	Bias	ASD	SSD	CP95
I	$(1, -1)$	$\hat{\beta}_1$	-0.0110	0.3024	0.3085	0.958	-0.0105	0.3004	0.2982	0.956
		$\hat{\beta}_2$	-0.0083	0.3182	0.3344	0.942	-0.0154	0.3174	0.3283	0.932
	$(-1, 1)$	$\hat{\beta}_1$	-0.0247	0.3282	0.3184	0.952	-0.0315	0.3250	0.3147	0.952
		$\hat{\beta}_2$	0.0149	0.3472	0.3355	0.964	0.0173	0.3456	0.3270	0.958
II	$(1, -1)$	$\hat{\beta}_1$	-0.0114	0.1425	0.1440	0.952	-0.0204	0.1608	0.1637	0.942
		$\hat{\beta}_2$	0.0068	0.1458	0.1340	0.950	0.0077	0.1662	0.1588	0.952
	$(-1, 1)$	$\hat{\beta}_1$	-0.0088	0.1870	0.1991	0.928	-0.0149	0.1912	0.2113	0.916
		$\hat{\beta}_2$	0.0061	0.1878	0.1943	0.932	0.0043	0.1950	0.2041	0.930
III	$(1, -1)$	$\hat{\beta}_1$	-0.0355	0.2287	0.2529	0.906	-0.0119	0.2269	0.2914	0.872
		$\hat{\beta}_2$	0.0054	0.2382	0.2550	0.932	-0.0110	0.2372	0.2709	0.900
	$(-1, 1)$	$\hat{\beta}_1$	-0.0138	0.2632	0.2893	0.904	-0.0334	0.2623	0.3187	0.882
		$\hat{\beta}_2$	0.0249	0.2700	0.2899	0.930	0.0346	0.2721	0.3074	0.914
IV	$(1, -1)$	$\hat{\beta}_1$	-0.0171	0.1412	0.1520	0.934	-0.0101	0.1444	0.1672	0.909
		$\hat{\beta}_2$	-0.0103	0.1453	0.1507	0.926	-0.0050	0.1491	0.1642	0.925
	$(-1, 1)$	$\hat{\beta}_1$	-0.0218	0.1776	0.1783	0.952	-0.0317	0.1797	0.1880	0.940
		$\hat{\beta}_2$	-0.0127	0.1790	0.1826	0.946	0.0006	0.1823	0.1923	0.932
V	$(1, -1)$	$\hat{\beta}_1$	0.0008	0.1581	0.1624	0.954	0.0093	0.1593	0.1777	0.916
		$\hat{\beta}_2$	-0.0113	0.1626	0.1733	0.928	-0.0221	0.1653	0.1859	0.912
	$(-1, 1)$	$\hat{\beta}_1$	-0.0361	0.1919	0.1890	0.930	-0.0466	0.1927	0.2029	0.920
		$\hat{\beta}_2$	0.0043	0.1947	0.1975	0.948	0.0156	0.1969	0.2014	0.940
VI	$(1, -1)$	$\hat{\beta}_1$	-0.0006	0.1621	0.1661	0.952	0.0088	0.1726	0.2089	0.900
		$\hat{\beta}_2$	0.0049	0.1677	0.1684	0.950	0.0011	0.1802	0.2080	0.918
	$(-1, 1)$	$\hat{\beta}_1$	-0.0188	0.1970	0.2003	0.942	-0.0003	0.2034	0.2382	0.906
		$\hat{\beta}_2$	-0.0031	0.2004	0.2062	0.944	0.0189	0.2099	0.2365	0.912

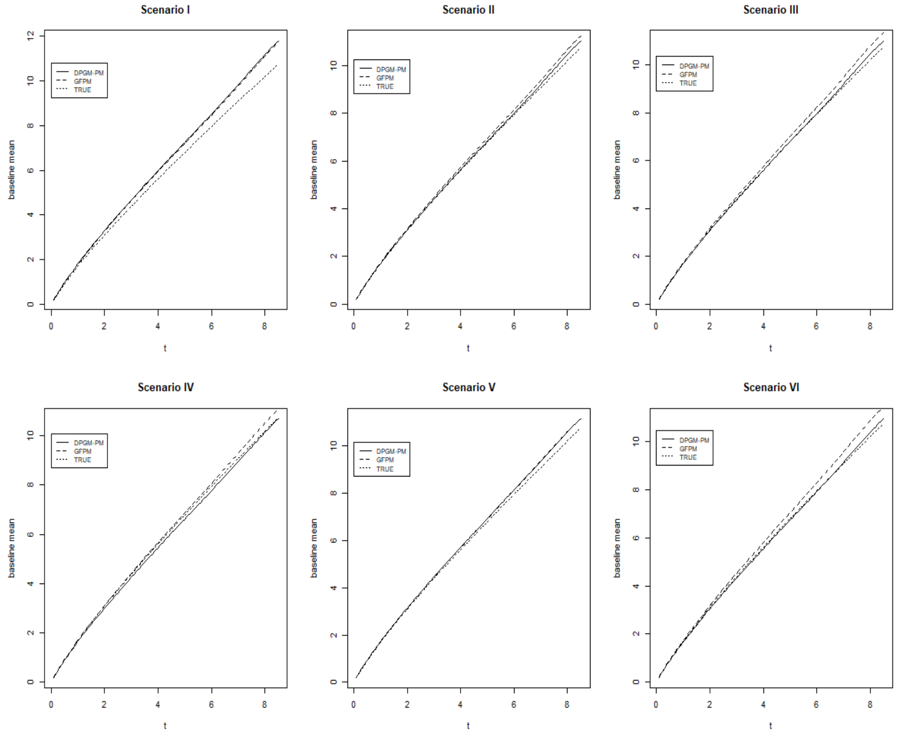

Fig 1. The true baseline mean function (dotted) and the average of the estimated baseline mean curves under DPGM-PM (solid) and GFPM (dashed) in the six simulation scenarios.

from DPGM-PM and GFPM over 500 data sets when $(\beta_1, \beta_2) = (1, -1)$ in the 6 simulation Scenarios. Examining the plots in Figure 1, we can conclude that the two estimated curves are overlapping in Scenario I but the mean estimate from DPGM-PM seems to be closer to the true mean baseline function than that from GFPM in all other scenarios.

In summary, these two methods have comparable and good performance when frailty follows a gamma distribution. However, DPGM-PM performs much better in terms of parameter estimation, inferential characteristics, and baseline mean function estimation when the true frailty distribution is non-gamma. This is not surprising for the following two reasons: (1) the multiple observations from the same subjects in panel count data provide much information about the underlying frailty distribution, which is not necessarily gamma; and (2) our proposed method is developed to accommodate the uncertainty in the frailty distribution and can provide an accurate estimation of the covariate effects and the baseline mean function.

5. Real data applications

The skin cancer study was conducted by the University of Wisconsin Comprehensive Cancer Center in Madison, Wisconsin. The primary goal of the study was to determine whether daily dose of $500mg/m^2$ of difluoromethylornithine (DFMO) would reduce new skin cancers in patients with a history of non-melanoma skin cancers including basal cell carcinoma and squamous cell carcinoma. In this randomized, double-blind, placebo-controlled phase 3 clinical trial, two hundred and ninety-one participants with a history of prior non-melanoma skin cancer were randomized to receive DFMO or placebo for 3 to 5 years, depending on when they entered the study. These patients were scheduled to be assessed every six months and the number of recurrences of new skin cancers between the observation times suppose to be recorded. However, the actual observational times varied from subject to subject. For more details about this study, please refer to [2].

Our analysis is focused on a total of 290 patients (147 in the placebo group and 143 in the DFMO group) after deleting one patient with missing cancer information. Four covariates were included in the analysis: x_1 (age at enrollment), x_2 (gender with 1 for male and 0 otherwise), x_3 (treatment indicator with 1 for DFMO group and 0 for placebo group), and x_4 (the number of prior skin tumors). The two quantitative covariates, age and the number of prior skin tumors, were standardized before analysis. The proposed nonparametric frailty Poisson process model was fitted with 21 equally-spaced knots and degree of 3. For comparison purposes, we also analyze the data under the gamma frailty Poisson model using the Bayesian method (GFPM-B) developed by [30] and the EM algorithm (GFPM-EM) developed by [36]. The length of the MCMC chains for both the proposed method and GFPM-B was fixed at 30000. It took 33.65 and 29.88 minutes for the proposed method and GFPM-B to complete the estimation, respectively. For both methods, the trace plots indicated that all the chains mix well and converge fast. The autocorrelation function plots showed that the autocorrelations of these key parameters go to zero quickly. To play conservatively, we summarized the estimation results based on the thinned samples by taking every 30th element of each chain after discarding the first 3000 iterations.

Table 2 presents the summarized estimation results in terms of the point estimate (MLE or posterior mean), the estimated standard error (deviation), and the 95% confidence (credible) interval for each regression parameter from the three methods. As seen in Table 2, the DFMO treatment did not have a significant effect on the recurrence of the new skin tumors, indicating that the DFMO treatment was not effective in reducing the number of new cancers. Also, the age and gender of patients did not seem to be significantly related to the tumor recurrence. In contrast, the number of prior skin tumors had a significantly positive effect on the occurrence of new skin tumors. These conclusions are consistent with prior studies in [2], [36], and [38].

We also compared the estimation of frailty distribution from the three methods. Under the gamma frailty Poisson model, the estimates of parameter for

Table 2

Analysis results of the skin cancer data analysis from the DPGM-PM and two competitive methods: GFPM [30] and GFPM-EM [36]. Summarized results are the point estimates (Point), the standard errors (SE) or deviations, and the corresponding 95% credible (or confidence) intervals for all the regression parameters ($\beta_1 \sim \beta_4$).

	DPGM-PM			GFPM			GFPM-EM		
	Point	SE	CI95	Point	SE	CI95	Point	SE	CI95
β_1	0.01	0.08	(-0.13, 0.16)	-0.00	0.07	(-0.15, 0.14)	-0.01	0.07	(-0.15, 0.13)
β_2	0.17	0.15	(-0.14, 0.45)	0.17	0.15	(-0.13, 0.49)	0.24	0.15	(-0.05, 0.53)
β_3	-0.10	0.15	(-0.38, 0.18)	-0.10	0.15	(-0.39, 0.21)	-0.04	0.15	(-0.32, 0.25)
β_4	0.63	0.09	(0.47, 0.80)	0.63	0.09	(0.47, 0.81)	0.65	0.08	(0.49, 0.81)

gamma frailty from the Bayesian method and EM algorithm were 1.24 and 1.26, respectively. Figure 2 presents the estimated frailty density curves from the nonparametric frailty Poisson process model and gamma frailty Poisson model. As seen in Figure 2, there is some noticeable difference between the estimated frailty densities although their overall shapes are close. These results suggest that the true frailty distribution is not seriously deviated from gamma if it is not a gamma distribution. Using the proposed approach does not need to worry about this gamma assumption since the approach does not assume any specific frailty distribution.

To illustrate the estimation of within-subject correlation, the Pearson correlation coefficient between the numbers of recurrences of new skin cancers within the first and second year was calculated using the formula proposed in Section 2.2. For example, for females who were at mean age (60.9 years old) in the placebo group and had mean number (4.3) of skin tumors before the study, this correlation is 0.71, while the same correlation is 0.62 from the corresponding Bayesian approach under the Gamma frailty Poisson model. These results suggest that there is a high within-subject correlation in the data.

6. Concluding remarks

In this paper, we propose a frailty non-homogeneous Poisson process model with an unknown frailty distribution for analyzing panel count data. Specifically, a Dirichlet process gamma mixture prior is assigned to the unknown frailty density. The adoption of monotone splines for the baseline mean function provides a simple form with only a finite number of parameters but maintains great modeling flexibility. An efficient Bayesian approach is then developed to carry out the estimation based on a well-calibrated data augmentation. The proposed exact block Gibbs sampler allows us to sample all the parameters and latent variables without the need of truncating the Dirichlet process. Our simulation results suggest that the proposed method works well in estimating both the regression parameters and the baseline mean function when frailty follows a variety of different distributions. In contrast, the corresponding Bayesian method under the gamma frailty Poisson process model produces unsatisfactory estimation results when the gamma frailty assumption is violated.

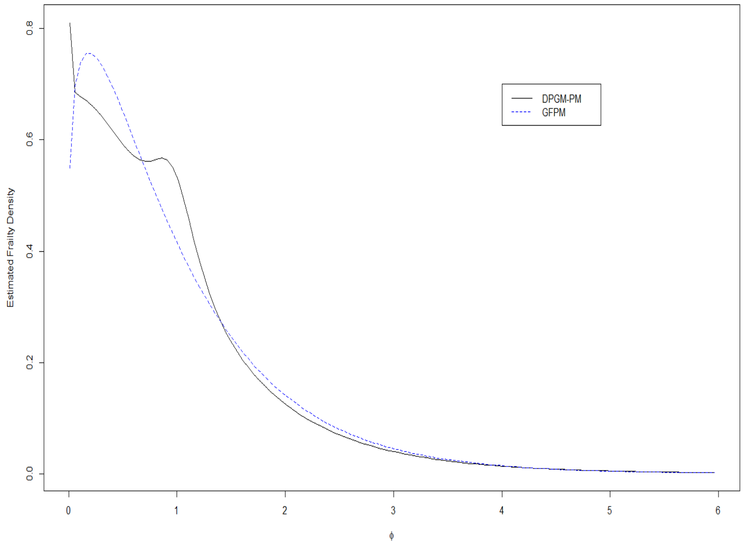

Fig 2. Plot of the estimated frailty density curves from nonparametric frailty Poisson process model (DPGM-PM) and gamma frailty Poisson model (GFPM) with $v = 1.24$.

For identification purpose, the mean of the frailty distribution is required to be one, and this is accomplished by restricting the shape and rate parameters in the gamma kernel to be the same in our proposed model. Although this strategy is simple and sufficient to meet the mean one constraint for the unknown frailty distribution, the theoretic properties of the proposed frailty distribution are unclear and are worth studying for future research.

References

[1] Antoniak, C. E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics* **2** 1152-1174.

[2] Bailey, H., Kim, K., Verma, A., Sielaff, K., Larson, P., Snow, S., Lenaghan, T., Viner, J., Douglass, J., Dreckschmidt, N., Hamielec, M., Pomplun, M., Sharata, H., Puchalsky, D., Berg, E., Havighurst, T. and Carbone, P. (2010). A randomized, double-blind, placebo-controlled phase 3 skin cancer prevention study of DFMO in subjects with previous history of skin cancer. *Cancer Prev Res (Phila)* **3** 35-47.

[3] Blackwell, D. (1973). The discreteness of Ferguson selections. *Annals of Statistics* **1** 356-358.

[4] Cox, D. R. (1983). Some Remarks on Overdispersion. *Biometrika* **70** 269-274.

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

[6] Ferguson, T. S. (1983). Bayesian Density Estimation by Mixtures of Normal Distribution. *Recent Advances in Statistic* 287-302.

[7] Gasperoni, F., Ieva, F. and Paganoni, A. M. (2020). Non-parametric frailty Cox models for hierarchical time-to-event data. *Biometrics* **21** 531-544.

[8] Gelfand, A. E. and Kottas, A. (2002). A Computational Approach for Full Nonparametric Bayesian Inference Under Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics* **11** 289-305.

[9] Hanson, T. E. (2006). Inference for Mixtures of Finite Polya Tree Models. *Journal of the American Statistical Association* **101** 1548-1565.

[10] He, X., Tong, X., Sun, J. and Cook, J. R. (2008). Regression analysis of multivariate panel count data. *Biostatistics* **9** 234-248.

[11] Hu, X. J., Lagakos, S. W. and Lockhart, R. A. (2009). Marginal analysis of panel counts through estimating functions. *Biometrika* **96** 445-456.

[12] Hua, L. and Zhang, Y. (2012). Spline-based semiparametric projected generalized estimating equation method for panel count data. *Biostatistics* **13** 440-454.

[13] Hua, L., Zhang, Y. and Tu, W. (2014). A Spline-based semiparametric sieve likelihood method for over-dispersed panel count data. *The Canadian Journal of Statistics* **42** 217-245.

[14] Lavine, M. (1990). Some Aspects of Polya Tree Distributions for Statistical Modeling. *The Annals of Statistics* **20** 1222-1235.

[15] Lavine, M. (1994). More Aspects of Polya Tree Distributions for Statistical Modeling. *The Annals of Statistics* **22** 1161-1176.

[16] Lin, X., Cai, B., Wang, L. and Zhang, Z. (2015). A Bayesian proportional hazards model for general interval-censored data. *Lifetime Data Analysis* **21** 470-490.

[17] Lin, X. and Wang, L. (2010). A Semiparametric Probit Model for Case 2 Interval-censored Failure Time Data. *Statistics in Medicine* **29** 972-981.

[18] Lu, M., Zhang, Y. and Huang, J. (2007). Estimation of the mean function with panel count data using monotone polynomial splines. *Biometrika* **94** 1060-1070.

[19] Lu, M., Zhang, Y. and Huang, J. (2009). Semiparametric estimation methods for panel count data using monotone B-splines. *Journal of the American Statistical Association* **104** 1060-1070.

[20] Manda, O. M. S. (2011). A Nonparametric Frailty Model for Clustered Survival Data. *Communications in Statistics-Theory and Methods* **40** 863-875.

[21] McMahan, C. S., Wang, L. and Tebbs, J. M. (2013). Regression analysis for current status data using the EM algorithm. *Statistics in Medicine* **32** 4452-4466.

[22] Naskar, M. (2008). Semiparametric analysis of clustered survival data under nonparametric frailty. *Statistica Neerlandica* **62** 155-172.

[23] Park, T. and Casella, G. (2008). The Bayesian Lasso. *Royal Statistical Society* **103** 681-686.

[24] Petrone, S. (1999). Bayesian density estimation using Bernstein polynomials. *The Canadian Journal of Statistics* **27** 105-126.

[25] Ramsay, J. O. (1988). Monotone Regression Splines in Action. *Statistical Science* **3** 425-461.

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

[27] Sun, J. and Kalbfleisch, J. D. (1995). Estimation of the mean function of point processes based on panel count data. *Statistica Sinica* **5** 279-290.

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

[29] Wang, C. and Lin, X. (2020). A Bayesian approach for semiparametric regression analysis of panel count data. *Lifetime Data Analysis* **26** 402-420.

[30] Wang, J. (2018). Bayesian Semiparametric Methods for Analyzing Panel Count Data, PhD thesis, University of South Carolina Retrieved from https://scholarcommons.sc.edu/etd/4778.

[31] Wang, L. and Dunson, D. (2011). Bayesian isotonic density regression. *Biometrika* **98** 537-551.

[32] Wang, L. and Dunson, D. (2011). Semiparametric Bayes' proportional odds models for current status data with under reporting. *Biometrics* **67** 1111-1118.

[33] Wang, L., McMahan, C. S., Hudgens, M. G. and Qureshi, Z. P. (2016). A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics* **72** 222-231.

[34] Wellner, A. J. and Zhang, Y. (2000). Two Estimators of the Mean of a Counting Process with Panel Count Data. *The Annals of Statistics* **28** 779-814.

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

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

[37] Yau, C., Papaspiliopoulos, O., Roberts, G. O. and Holmes, C. (2011). Bayesian nonparametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society: Series B* **73** 37-57.

[38] Zhang, H., Zhao, H., Sun, J., Wang, D. and Kim, K. (2013). Regression analysis of multivariate panel count data with an informative observation process. *Journal of Multivariate Analysis* **119** 71-80.

[39] Zhang, Y. (2002). A semiparametric pseudolikelihood estimation method for panel count data. *Biometrika* **89** 39-48.

[40] Zhang, Y. and Jamshidian, M. (2003). The Gamma-Frailty Poisson Model for the Nonparametric Estimation of Panel Count Data. *Biometrics* **59** 1099-1106.