Variational Inference for Gaussian Processes with Panel Count Data

Hongyi Ding[13], **Young Lee**[2], **Issei Sato**[13], **Masashi Sugiyama**[31]
[1]The University of Tokyo, Japan
[2]National Univeristy of Singapore, Singapore
[3]The RIKEN Center for AIP, Tokyo, Japan

Abstract

We present the first framework for Gaussian-process-modulated Poisson processes when the temporal data appear in the form of panel counts. Panel count data frequently arise when experimental subjects are observed only at discrete time points and only the numbers of occurrences of the events between subsequent observation times are available. The exact occurrence timestamps of the events are unknown. The method of conducting the efficient variational inference is presented, based on the assumption of a Gaussian-process-modulated intensity function. We derive a tractable lower bound to alleviate the problems of the intractable evidence lower bound inherent in the variational inference framework. Our algorithm outperforms classical methods on both synthetic and three real panel count sets.

1 INTRODUCTION

Background and issues. Temporal data frequently arise as outcomes of an underlying *temporal point process* (Kingman, 1993) in continuous time. Temporal data can generally be classified into two types. One is from experiments that monitor subjects in a continuous fashion; and thereby the exact timestamps of all occurrences of the events are fully observable. These data are usually referred to as *recurrent event data* (Cook and Lawless, 2007). On the other hand, we have the so-called *panel count data* (Sun and Zhao, 2016), which is the focus of our paper. Under this framework, subjects are examined or observed only at discrete time-points and thus give only the numbers of occurrences of the events between subsequent observation times.

Characteristics of panel count data. A common characteristic of the panel count data is that we only have the numbers of occurrences between subsequent observation times. In particular, the exact occurrence times of the events are unknown. Hence, panel counts are non-negative integers and they represent the number of occurrences of events within a fixed period. Classical examples often arise in the clinical trials (Thall and Lachin, 1988) where patients are required to go back to the hospital after a certain treatment and only the numbers of symptoms between subsequent visits are recorded, such as the number of vomits or new tumors. Figure 1 gives an example of panel count data.

Objective of this study. The purpose of this paper is to present the variational Bayesian inference on **G**aussian-**p**rocess-modulated **P**oisson **p**rocesses (GP3) that permits panel data observations.

There have been extensive studies on GP3 models and various inference algorithms are introduced for *recurrent event data* when timestamps of the events are fully observable, e.g., Monte Carlo sampling (Diggle et al., 2013; Adams et al., 2009), Laplace approximation (Flaxman et al., 2015) and variational inference (Lloyd et al., 2015). Among these approaches, the variational inference method (Lloyd et al., 2015) provides a computationally efficient estimate of the intensity function and does not require a careful discretization of the underlying space.

To the best of our knowledge, however, there has not been any study carried out on the variational inference of the GP3 model when the data come in the form of panel counts. Our ultimate goal is to infer the underlying intensity function in the panel count data.

Related statistical works. Based on the maximum likelihood criterion, several non-parametric estimators have been proposed to infer the underlying intensity function (Sun and Zhao, 2016), e.g., a

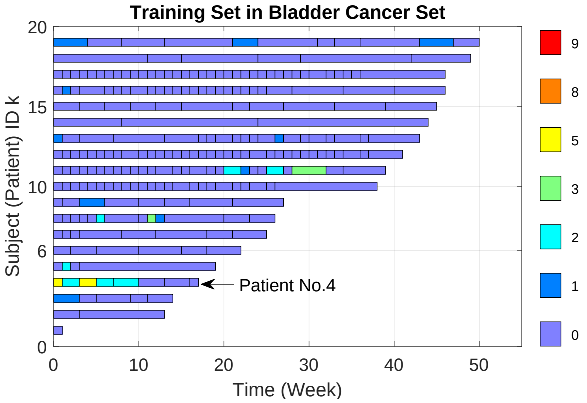

Figure 1: **Bladder Cancer Data Set**. This figure illustrates the panel count data from the patients. For the kth subject (or the kth patient), his/her observation window $\mathcal{X}^{(k)}$ is divided into disjoint intervals. The ith interval is denoted as $\mathcal{X}_i^{(k)}$. For example, patient No. 4 ($k = 4$) has an observation window which is divided into 8 disjoint intervals, i.e., $\bigcup_{i=1}^{8} \mathcal{X}_i^{(4)} = \mathcal{X}^{(4)}$ and $X_i \cap X_j = \emptyset$ for $i \neq j$. Patients may drop out from the study at any time and therefore their observation windows are different. An interval is shown by a rectangle. We use different colors to indicate the different numbers of new bladder tumors observed in this interval. Note that we only have access to *the number of events* in each interval.

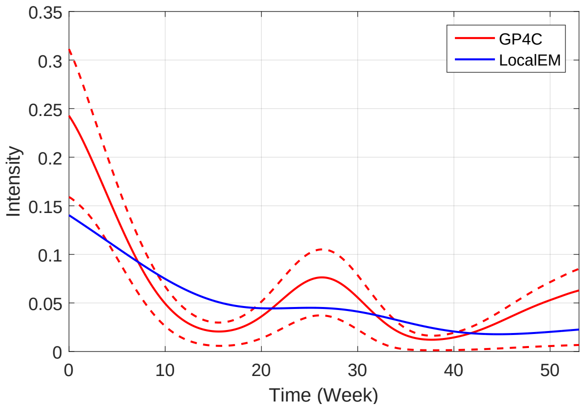

Figure 2: **Bladder Cancer Data Set**. Inferred intensity function by the LocalEM and GP4C methods. For GP4C, a 75% credible interval is given by dotted lines. Our estimator GP4C provides the additional uncertainty in the estimated intensity function compared with LocalEM. See Section 5 for details.

non-parametric maximum pseudo-likelihood estimator (NPMPLE) (Wellner and Zhang, 2000), a non-parametric maximum pseudo-likelihood estimator with gamma frailty (NPMPLGF) (Zhang and Jamshidian, 2003) and the local Expectation-Maximization (LocalEM) estimator (Fan et al., 2011). Unlike NPMPLE and NPMPLGF, which only estimate the cumulative intensity function at a set of points, LocalEM provides a smooth estimate of the underlying intensity function due to the use of an exponential quadratic kernel (Fan et al., 2011).

Besides the computational cost in selecting the bandwidth of the exponential quadratic kernel, the estimators obtained by the LocalEM algorithm and other similar algorithms are point-estimates in the sense that the estimated intensity function is a point in the functional space. These point-estimates fail to capture the uncertainty in the data set. We show an example of the estimated intensity function by LocalEM in Figure 2. The uncertainty of the intensity function helps us understand the difficulty of the prediction at a given time.

Contributions. The contributions of our work are twofold. 1) In the first place it undertakes to construct a variational inference procedure for the **G**aussian-**P**rocess-modulated **P**oisson **P**rocess model for **P**anel **C**ount data (GP4C). 2) To carry out a variational inference in this setting, we derive a simple and tractable lower bound of the intractable evidence lower bound and demonstrate through empirical evidence that with this lower bound, GP4C outperforms a non-Bayesian method.

2 BACKGROUND

Throughout this paper, we denote the set of panel count data from $K \in \mathbb{N}^+$ independent subjects as $\mathcal{D}$. Each subject will generate a sequence of events in the continuous space $\mathcal{X}$. We only consider the temporal point processes where the continuous space $\mathcal{X}$ is a subset of $\mathbb{R}$. In the *recurrent event data*, the timestamps of the events are fully observable. We denote the timestamps from the kth subject as $\{x_j^{(k)} \in \mathcal{X}\}$.

In the *panel count data*, the kth subject is assessed in N_k disjoint intervals $\{\mathcal{X}_i^{(k)}\}_{i=1}^{N_k}$, where $\cup_i \mathcal{X}_i^{(k)} = \mathcal{X}^{(k)} \subset \mathcal{X}$. We have access to each interval $\mathcal{X}_i^{(k)}$ and the number of events observed in this interval $m_i^{(k)} = |\{x_j^{(k)} \in \mathcal{X}_i^{(k)}\}|$. Let $\boldsymbol{d}_k = \{(\mathcal{X}_i^{(k)}, m_i^{(k)})\}_{i=1}^{N_k}$ and $\mathcal{D} = \{\boldsymbol{d}_k\}$. Figure 1 illustrates an example of the panel count data.

2.1 LIKELIHOOD OF PANEL COUNT DATA

In the *recurrent event data*, one approach to modeling the events $\{x_j^{(k)} \in \mathcal{X}\}$ from each subject is to use the inhomogeneous Poisson processes (IPP) (Kingman, 1993) and assume that there is a fixed underlying intensity function $\lambda(x) : \mathcal{X} \to \mathbb{R}^+$. Given the intensity function $\lambda(x)$, the likelihood for the observed events is

$$p(\{x_j^{(k)}\}|\lambda(x)) = \exp\Big(-\int_{\mathcal{X}} \lambda(x)dx\Big)\prod_j \lambda(x_j^{(k)}).$$

To derive the likelihood of the *panel count data* $\mathcal{D}$, we use two important features of an IPP (Kingman, 1993). The first is that given the intensity function $\lambda(x)$, the probability that we observe $m_i^{(k)}$ events in the interval $\mathcal{X}_i^{(k)}$ is given as follows:

$$p(m_i^{(k)}|\lambda(x); \mathcal{X}_i^{(k)}) = \frac{r_{ik}^{m_i^{(k)}}}{m_i^{(k)}!}\exp(-r_{ik}), \tag{1}$$

where $r_{ik} \triangleq \int_{\mathcal{X}_i^{(k)}} \lambda(x)dx$ is the rate parameter of the Poisson distribution. Hereafter, we omit the dependency on $\mathcal{X}_i^{(k)}$ for simplicity. However, the likelihood depends on the intervals and even for the same sequence, after censored with different intervals, the likelihood of the sequence will vary. See Appendix E.1 for a brief discussion.

The second feature is that on two disjoint intervals $\mathcal{X}_i^{(k)}$ and $\mathcal{X}_j^{(k)}$ ($\mathcal{X}_i^{(k)} \bigcap \mathcal{X}_j^{(k)} = \emptyset$), the numbers of events on these intervals are independent random variables.

$$p(m_j^{(k)}, m_i^{(k)}|\lambda(x)) = p(m_j^{(k)}|\lambda(x))p(m_i^{(k)}|\lambda(x)). \tag{2}$$

Based on these two features, the likelihood of the panel count data $\mathcal{D}$ can be derived. We assume that all subjects share the same intensity function $\lambda(x)$. Since K subjects are independent of each other and for the kth subject, the N_k intervals $\{\mathcal{X}_i^{(k)}\}_{i=1}^{N_k}$ are disjoint, we obtain the following likelihood:

$$p(\mathcal{D}|\lambda(x)) = \prod_{k=1}^{K} p(\boldsymbol{d}_k|\lambda(x)) = \prod_{k=1}^{K}\prod_{i=1}^{N_k} p(m_i^{(k)}|\lambda(x)). \tag{3}$$

Several maximum likelihood estimators have been proposed on the basis of this likelihood or its variants, e.g., NPMPLE (Wellner and Zhang, 2000; Wellner et al., 2007), NPMPLGF (Zhang and Jamshidian, 2003) and the LocalEM estimator (Fan et al., 2011). An estimate from LocalEM on the data set in Figure 1 is given in Figure 2. As we discussed, these estimators fail to model the uncertainty in the intensity function.

2.2 GP3 MODEL

In order to model the uncertainty of the intensity function $\lambda(x)$ via a kernel, the traditional approach is to use the Cox process (Kingman, 1993). A Cox process is defined via a stochastic intensity function $\lambda(x)$. The stochastic process to generate the intensity function is usually chosen to be a Gaussian process (GP) (Adams et al., 2009) and the model using a GP is called a GP3 model.

For the *recurrent event data*, GP3 models have been studied extensively (Adams et al., 2009; Gunter et al., 2014; Lloyd et al., 2015). The following model is an example of GP3 models (Lloyd et al., 2015),

$$\lambda(x) = f^2(x),\ f \sim \mathcal{GP}(g(x), \kappa(x, x')), \tag{4}$$

where $\mathcal{GP}(g(x), \kappa(x, x'))$ denotes the Gaussian process with mean function $g(x)$ and covariance function $\kappa(x, x')$. The function $f(x)$ drawn from a GP prior is squared to ensure the non-negativity of the intensity function. The GP3 model in Equation (4) admits a complete variational inference framework. Moreover, this intensity model can be enhanced with an independent variable for each subject or a mixture structure (Lloyd et al., 2016) to flexibly model the heterogeneity of the intensity functions across several subjects.

3 OUR MODEL GP4C : GP3 MODEL FOR PANEL COUNT DATA

In order to retain the scalability and efficiency of the variational inference approach (Lloyd et al., 2015) and add the uncertainty on the intensity function when we only observe the panel count data, we use the GP3 model defined in Equation (4) as the underlying intensity model.

The joint distribution $p(\mathcal{D}, f)$ can be obtained by combining the likelihood model in Equation (3) and the intensity model in Equation (4).

$$p(\mathcal{D}, f) = \Big[\prod_{k=1}^{K} p(\boldsymbol{d}_k|\lambda(x))\Big]p(f; g, \kappa). \tag{5}$$

We call this model the **GP**-modulated **P**oisson **P**rocess model for **P**anel **C**ount data (GP4C).

4 INFERENCE

In this section, we will discuss the problems when applying variational inference techniques on the GP4C model.

4.1 VARIATIONAL INFERENCE

We use sparse GPs to reduce the computational complexity with the set of pseudo inputs $\{x_r\}_{r=1}^{R}$ on $\mathcal{X}$ (Titsias, 2009). Let $\boldsymbol{f}_R \triangleq [f(x_1), \ldots, f(x_R)]^\top$. The joint model with additional pseudo inputs is $p(\mathcal{D}, f, \boldsymbol{f}_R) = p(\mathcal{D}|f)p(f|\boldsymbol{f}_R)p(\boldsymbol{f}_R)$ and the variational distribution is defined as follows:

$$q(f, \boldsymbol{f}_R) = p(f|\boldsymbol{f}_R)q(\boldsymbol{f}_R), \tag{6}$$

where $q(\boldsymbol{f}_R)=\mathcal{N}(\boldsymbol{\mu},\Sigma)$ and $\mathcal{N}(\boldsymbol{\mu},\Sigma)$ denotes the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. The evidence lower bound (ELBO) $\mathcal{L}$ can be obtained by using Jensen's inequality.

$$\begin{aligned}\ln p(\mathcal{D}) &\geq \iint q(f,\boldsymbol{f}_R)\ln\frac{p(\mathcal{D},f,\boldsymbol{f}_R)}{q(f,\boldsymbol{f}_R)}df\,d\boldsymbol{f}_R\\ &=\sum_{k=1}^{K}\sum_{i=1}^{N_k}\Big(m_i^{(k)}\mathbb{E}_q\Big[\ln\int_{\mathcal{X}_i^{(k)}}f^2(x)dx\Big]-\ln(m_i^{(k)}!)\Big)\\ &-\sum_{k=1}^{K}\mathbb{E}_q\Big[\int_{\mathcal{X}^{(k)}}f^2(x)dx\Big]+\mathbb{E}_q\Big[\ln\frac{p(\boldsymbol{f}_R)}{q(\boldsymbol{f}_R)}\Big]\triangleq\mathcal{L}.\end{aligned}\tag{7}$$

In ELBO, when assuming that the covariance function $\kappa(x,x')$ is the automatic relevance determination (ARD) function $\kappa(x,x')=\gamma\exp\Big(-\frac{(x-x')^2}{2a^2}\Big)$, $x,x'\in\mathcal{X}$, the second term in the ELBO can be analytically calculated (Lloyd et al., 2015) as follows:

$$\begin{aligned}\mathbb{E}_q\Big[\int_{\mathcal{X}^{(k)}}f^2(x)dx\Big]&=\gamma|\mathcal{X}^{(k)}|-\mathrm{tr}(K_{RR}^{-1}\Phi)\\ &+\mathrm{tr}(K_{RR}^{-1}\Phi K_{RR}^{-1}(\boldsymbol{\mu}\boldsymbol{\mu}^\top+\Sigma)),\end{aligned}\tag{8}$$

where Φ is an $R\times R$ matrix related to the pseudo inputs with its (i,j)-th entry equal to $\int_{\mathcal{X}^{(k)}}\kappa(x_i,x)\kappa(x,x_j)dx$ and K_{RR} is the covariance matrix computed at the pseudo inputs. However, the ELBO $\mathcal{L}$ is still intractable, since we can not analytically compute the expected integral $\mathbb{E}_q\Big[\ln\int_{\mathcal{X}_i^{(k)}}f^2(x)dx\Big]$ in the first term.

4.2 A TRACTABLE LOWER BOUND

We tackle the intractable expectation by deriving a tractable lower bound. First we introduce a relevant lemma on the expectation of the logarithm of the square of a normal-distributed random variable.

Lemma 1. *Let $y\sim\mathcal{N}(\mu,\sigma^2)$ and $\varphi=(\mu/\sigma)^2$. Then*

$$\mathbb{E}_y[\ln y^2]=\ln(2\sigma^2)+\sum_{j=0}^{\infty}\frac{(\varphi/2)^j\exp(-\varphi/2)}{j!}\psi(j+1/2),\tag{9}$$

where $\psi(\cdot)$ is the digamma function.

The proof of Lemma 1 can be found in Appendix A. Let

$$g_m(y)=\sum_{j=0}^{\infty}\frac{y^j\exp(-y)}{j!}\psi(j+m).\tag{10}$$

Then $\mathbb{E}_y[\ln y^2]=\ln(2\sigma^2)+g_{0.5}(\varphi/2)$. The function $g_m(y)$, where y is a positive real number and m is a positive integer, has been studied in the analysis of mobile and wireless communication systems (Moser, 2007). For $m=1/2$, $g_{0.5}(\varphi/2)$ can be computed using a confluent hyper-geometric function $G(\cdot)$ (Lloyd et al., 2015), which is stored in a pre-computed look-up table.

$$g_{0.5}(\varphi/2)=-G(-\varphi/2)-2\ln 2-C,\tag{11}$$

where C is Euler's constant and $C\approx 0.5772$. However, to the best of our knowledge, it is still not clear how to calculate the integral of the function $G(-\varphi/2)$ when using a GP. To derive a tractable lower bound of the intractable expectation, we introduce the following lemma to give a lower bound of the function $g_m(y)$ and the proof can be found in Appendix B.

Lemma 2. *Let $y\sim\mathcal{N}(\mu,\sigma^2)$ and C be Euler's constant.*

$$\mathbb{E}_y[\ln y^2]\geq\ln(\mu^2+b\sigma^2)-C-\ln 2,\ \forall b\in[0,1].\tag{12}$$

Based on Lemma 2, we propose the following lower bound for the intractable expectation in the ELBO.

Theorem 1. *Let f be a GP as defined in Equation (4). For $b\in[0,1]$, the following bound holds:*

$$\begin{aligned}\mathbb{E}_q\Big[\ln\int_{\mathcal{X}_i^{(k)}}f^2(x)dx\Big]&\geq -C-\ln 2\\ &+\ln\Big(\int_{\mathcal{X}_i^{(k)}}\Big(\mathbb{E}_q^2f(x)+b\mathrm{Var}_qf(x)\Big)dx\Big),\end{aligned}\tag{13}$$

where the distribution q is given in Equation (6).

Proof. We first use Jensen's inequality on the logarithm function and then interchange the order of integration and expectation.

$$\begin{aligned}\mathbb{E}_q\Big[\ln\int_{\mathcal{X}_i^{(k)}}f^2(x)dx\Big]&=\mathbb{E}_q\Big[\ln\int_{\mathcal{X}_i^{(k)}}\tilde{p}(x)\frac{f^2(x)}{\tilde{p}(x)}dx\Big]\\ &\geq\int_{\mathcal{X}_i^{(k)}}\tilde{p}(x)\mathbb{E}_q\Big[\ln\frac{f^2(x)}{\tilde{p}(x)}\Big]dx,\end{aligned}\tag{14}$$

where $\tilde{p}(x)$ is a probability distribution on $\mathcal{X}_i^{(k)}$. Furthermore, maximizing this lower bound with respect to $\tilde{p}(x)$ yields the optimal distribution:

$$\tilde{p}_{\mathrm{opt}}(x)\propto\exp\Big(\mathbb{E}_q\ln f^2(x)\Big).\tag{15}$$

We remark that this result is analogous to that of the discrete version presented in Paisley (2010). Substituting Equation (15) into the right-hand side of Equation (14) yields

$$\mathbb{E}_q\left[\ln\int_{\mathcal{X}_i^{(k)}} f^2(x)dx\right] \geq \ln\left(\int_{\mathcal{X}_i^{(k)}} e^{\mathbb{E}_q \ln f^2(x)}dx\right)$$

$$\overset{(13)}{\geq} \ln\left(\int_{\mathcal{X}_i^{(k)}} e^{\ln(\mathbb{E}_q^2 f(x)+b\mathrm{Var}_q f(x))-C-\ln 2}dx\right)$$

$$= \ln\left(\int_{\mathcal{X}_i^{(k)}} \left(\mathbb{E}_q^2 f(x)+b\mathrm{Var}_q f(x)\right)dx\right) - C - \ln 2,$$

where we have invoked Lemma 2 in the penultimate line whilst defining $y := f(x)$. □

It should be emphasized that we are making no further assumptions on the dimensionality of x in the proof of Theorem 1. Hence we may augment the dimensionality of x in Theorem 1 such that it can also be applied to problems in spatial point processes. In summary, the ELBO in Equation (7) inherits an analytical bound. We present the following:

Theorem 2. *A tractable lower bound of the ELBO $\mathcal{L}$ in the GP4C model is given as follows:*

$$\mathcal{L} \geq \tilde{\mathcal{L}} \triangleq -\sum_{k=1}^{K}\mathbb{E}_q\left[\int_{\mathcal{X}^{(k)}} f^2(x)dx\right] + \mathbb{E}_q\left[\ln\frac{p(\boldsymbol{f}_R)}{q(\boldsymbol{f}_R)}\right]$$

$$+\sum_{k=1}^{K}\sum_{i=1}^{N_k} m_i^{(k)}\ln\left(\int_{\mathcal{X}_i^{(k)}}\left(\mathbb{E}_q^2 f(x)+b\mathrm{Var}_q f(x)\right)dx\right)$$

$$-\sum_{k=1}^{K}\sum_{i=1}^{N_k}\left(m_i^{(k)}(C+\ln 2)+\ln(m_i^{(k)}!)\right). \qquad (16)$$

The details of the proof are deferred to Appendix C. The derivations of $\mathbb{E}_q^2 f(x)$ and $\mathrm{Var}_q f(x)$ follow similar lines to the derivation of Equation (8). The third part of $\tilde{\mathcal{L}}$ is a constant and thus can be omitted when maximizing the lower bound. Let $\Psi = \{\boldsymbol{\mu}, \Sigma\}$ and $\Phi = \{\gamma, a\}$ be the variational parameters and hyper-parameters in the covariance function of a GP, respectively. We use the variational Expectation-Maximization (vEM) algorithm (Dempster et al., 1977) to update the parameters Ψ and Φ iteratively on the modified ELBO $\tilde{\mathcal{L}}$.

4.3 THE VALUE OF PARAMETER b

A natural question is, how do we select the parameter b in Theorem 1? Recall that two inequalities were used in the proof. It is cumbersome to evaluate Inequality (14) since it is an integral over $\mathcal{X}_i^{(k)}$. We first examine different choices of b in Lemma 2.

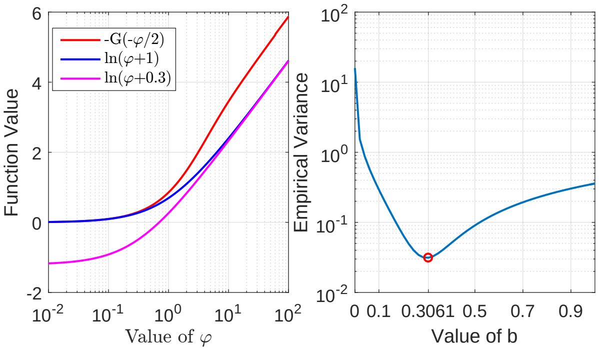

Figure 3: **Influences of b in Lemma 2**. (Left) The true value of $-G(-\varphi/2)$ by a look-up table and two simple lower bounds. The bound $\ln(\phi + b)$ with $b = 0.3$ correlates with the curve of the true value better. (Right). The variance $\mathrm{Var}[h(\varphi; b)]$ when varying the choices of b and the best b is shown with a red circle.

In Paisley et al. (2012), a more correlated lower bound of the ELBO serves as a better control variate in reducing the variance of a stochastic gradient. Inspired by this study, we introduce a heuristic method and conduct the following experiment to evaluate the correlation for different choices of b. In Lemma 2, the difference between the lower bound and the true value is

$$\ln(\mu^2 + b\sigma^2) - C - \ln 2 - \mathbb{E}_y[\ln y^2]$$

$$= \ln(\varphi + b) + G(-\varphi/2) \triangleq h(\varphi; b). \qquad (17)$$

For each choice of b, we vary $\varphi = (\mu/\sigma)^2$ on a vector of 5000 logarithmically spaced points between 10^{-6} and 10^6 and evaluate the correlation between the lower bound and the true value by the variance $\mathrm{Var}[h(\varphi; b)]$. We calculate $\mathrm{Var}[h(\varphi; b)]$ on a vector of 50 evenly spaced choices of b between 0 and 1 and the result is shown in Figure 3. We see that the optimal choice of b is 0.3061 if φ ranges from 10^{-6} to 10^6. In the actual situation, this optimal value of b depends on the range of φ in the data and the influence of Inequality (14), we evaluate several choices of b on synthetic data sets in Section 5.

4.4 COMPUTATIONAL COMPLEXITY

Let each interval in temporal point processes be $\mathcal{X}_i^{(k)} = [x_{ai}^{(k)}, x_{bi}^{(k)}]$ with two end points $x_{ai}^{(k)}$ and $x_{bi}^{(k)}$. Two intervals are different if at least one end point is different. We denote the number of different intervals in the data set as N and the number of pseudo inputs as M. For each interval, the computation complexity of GP4C is $\mathcal{O}(M^3)$ which is determined by the matrix-matrix calculation when evaluating $\mathrm{Var}_q f(x)$ in Equation (16). The computational complexity during one iteration of the vEM algorithm is $\mathcal{O}(NM^3)$ since in our implementation, we calculate the integral of all N different intervals.

We analyze the computational complexity of the LocalEM (Fan et al., 2011) algorithm for comparison. In LocalEM, $\{x_{ai}^{(k)}\}$ and $\{x_{bi}^{(k)}\}$ are first merged into a single ordered set X where duplicated values are removed. We denote the size of the merged set X as $\bar{N}$ and generally $\bar{N} \leq N$. Then the Gaussian quadratic rule with $\bar{M}$ points is used to calculate the integral of the intensity function between subsequent values in the set X and the computational complexity during one iteration is $\mathcal{O}(\bar{N}^2\bar{M}^2)$. If the size of the merged set $\bar{N}$ is significantly smaller than N, LocalEM may be computationally more efficient than GP4C. However, if $\bar{N} \approx N$, LocalEM may suffer from the term $\bar{N}^2$ in the computational complexity. We provide additional experiments on the influence of the number $\bar{N}$ in Appendix E.3.

5 EXPERIMENTS

We evaluate our proposed GP4C model and compare it with the benchmark methods on both synthetic and real-world data sets. The algorithms are programmed in Matlab R2015b and run on an Intel Xeon E5-2667 CPU with a memory of 64GB. Our code is available at `github.com/Dinghy/GP4C`.

5.1 EXPERIMENT SETTINGS

For each data set $\mathcal{D}$, we randomly partition the subjects into training and testing sets, which we denote as $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, respectively. We repeat each setting for $S = 40$ times. In the sth trial, the training and testing sets are denoted as $\mathcal{D}_{\text{train}}^{(s)}$ and $\mathcal{D}_{\text{test}}^{(s)}$.

Benchmark. Two benchmark algorithms are used.

a) **GP3** (Lloyd et al., 2015). This benchmark reflects the best performance that can be obtained if we obtain the recurrent event data set where we have the exact timestamps.

b) **LocalEM** (Fan et al., 2011). Both LocalEM and GP4C are nonparametric estimators based on the maximum likelihood criterion. To fairly compare the computation time, we implemented the LocalEM algorithm in MATLAB based on the R code provided in Fan et al. (2011). This method produces a smooth estimate of the intensity function due to the use of an exponential quadratic kernel. We use a 5-fold cross-validation on the training set to select the bandwidth of the exponential quadratic kernel.

Evaluation Metric. We evaluate the performance of the algorithms in terms of three metrics.

a) Mean of the integrated squared error (MISE). In synthetic data sets, we have the ground truth of the intensity function λ_{true} and the integrated squared error can be calculated using our estimated intensity function $\lambda_{\text{est}}^{(s)}$ during the sth trial. To measure the bias of each estimator, we calculate the mean of the integrated squared error as follows:

$$\text{MISE}(s) \triangleq \int_{\mathcal{X}} (\lambda_{\text{est}}^{(s)}(x) - \lambda_{\text{true}}(x))^2 dx. \quad (18)$$

For GP4C, to measure its bias, we omit the variance of the estimator and use the expectation of the intensity function $\mathbb{E}_{q^{(s)}}[f^2(x)]$ as $\lambda_{\text{est}}^{(s)}(x)$.

b) Test log likelihood $\mathcal{L}_{\text{test}}$. During the sth trial, the logarithm of the test likelihood can be written as follows:

$$\mathcal{L}_{\text{test}}(s) \triangleq \ln \int p(\mathcal{D}_{\text{test}}^{(s)}|f)p(f|\mathcal{D}_{\text{train}}^{(s)})df. \quad (19)$$

For LocalEM, since this estimator provides a point-estimate and we directly use the estimated function $f^{(s)}$ to calculate $\mathcal{L}_{\text{test}}(s)$. For GP4C and GP3, we need to sample the function $f^{(s)}$ from the variational distribution and the detailed calculation can be found in Appendix D.

c) Computation time T. We record the training time measured in seconds for each setting. For GP3 and GP4C, we record the computation time of the training process. For LocalEM, it includes the time of 5-fold cross-validation on the training set to select the bandwidth of the exponential quadratic kernel and the time of a training process over the whole training set.

Optimization Settings. For GP3 and GP4C, following Lian et al. (2015), we use the re-parametrization trick $\Sigma = LL^\top$ by Cholesky decomposition and add positivity constraints to the diagonal elements in L. Due to this constraint on L, we use the limited-memory projected quasi-Newton algorithm (Schmidt et al., 2009) to optimize the variational parameters $\Psi = \{\boldsymbol{\mu}, \Sigma\}$. We add a jitter term ϵI where $\epsilon = 10^{-6}$ to the covariance matrix K_{RR} to avoid numerical instability (Titsias, 2009).

5.2 SYNTHETIC DATA SETS

We test three synthetic data sets which we denote as the Synthetic A, B and C data sets, respectively.

On the Synthetic A data set, the intensity function is a square wave function $h_1(x)$ as follows. See Figure 4 for an illustration of $h_1(x)$.

$$h_1(x) = \begin{cases} 7 \text{ if mod}\left(\left[\dfrac{x}{10}\right], 2\right) = 0, \\ 2 \text{ otherwise.} \end{cases}$$

Table 1: **Synthetic data sets**. Mean and standard deviation of statistics about different choices of b over 40 runs. GP3 uses the *recurrent event data* while LocalEM and GP4C use the *panel count data*. For GP4C, $b = 0.3$ and $b = 0$ perform better than $b = 1$ in terms of MISE and $\mathcal{L}_{\text{test}}$.

Method	**MISE**	$\mathcal{L}_{\text{test}}$	$T[s]$
(Synthetic A)			
GP3	29.5±1.0	-1366.5±17.4	16±4
GP4C(1)	41.8±6.2	-3236.9±542.3	25±5
GP4C(0)	40.8±3.3	-1378.1±16.9	19±4
GP4C(0.3)	40.2±3.2	-1377.8±17.5	20±3
LocalEM	44.6±3.1	-1383.5±17.0	33±2
(Synthetic B)			
GP3	0.5±0.2	-783.1±20.7	8±1
GP4C(1)	1.9±2.1	-1005.8±81.5	55±44
GP4C(0)	2.7±0.8	-794.5±20.1	17±3
GP4C(0.3)	2.4±0.7	-794.2±20.2	17±4
LocalEM	3.5±0.7	-800.3±19.6	33±2
(Synthetic C)			
GP3	1.2±0.4	-864.1±14.9	8±3
GP4C(1)	2.3±1.5	-1194.6±100.5	52±53
GP4C(0)	2.1±0.6	-871.2±15.9	17±2
GP4C(0.3)	2.0±0.7	-872.0±15.7	18±3
LocalEM	5.2±1.1	-882.7±16.5	34±2

On the Synthetic B and C data set, the underlying intensity functions are drawn according to Equation (4). We first draw a function from a GP on a vector of 3001 evenly-spaced points in $\mathcal{X} = [0, T]$, where $T = 60$. We approximate the value of the function at an arbitrary position with linear interpolation. The function is then squared to guarantee the positiveness of the intensity function. See Figure 5 for an illustration.

During the sth trial, we first generate a *recurrent event data set* with 100 subjects on the same observation window $\mathcal{X}^{(k)} = \mathcal{X}$. Then we generate the corresponding *panel count data set* $\mathcal{D}^{(s)}$ by censoring each subject with 10 intervals. We generate the censored intervals by a draw from a Dirichlet distribution $\boldsymbol{w}^{(k)} \sim \text{Dir}(\boldsymbol{\theta})$ and $\boldsymbol{\theta}$ is a 10-dimensional vector with all elements equal to 1. The ith interval of the kth subject can be computed as $\mathcal{X}_i^{(k)} = [\sum_{j=1}^{i-1} w_j^{(k)} T, \sum_{j=1}^{i} w_j^{(k)} T]$. We randomly partition $\mathcal{D}^{(s)}$ into two parts, where 50 subjects are used for training and 50 for testing.

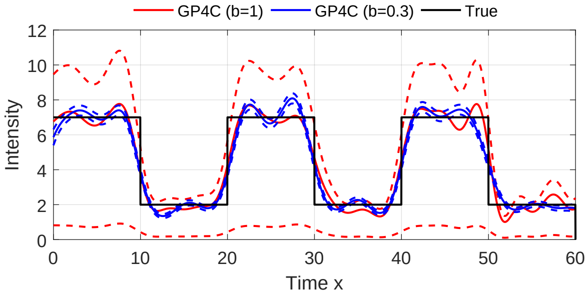

Figure 4: **Synthetic A Data Set**. The estimated intensity functions from GP4C ($b = 1$) and GP4C ($b = 0.3$) are shown with 75% credible intervals. True intensity function $h_1(x)$ is given for comparison. We see that GP4C ($b = 1$) over-estimates the variance of the intensity function.

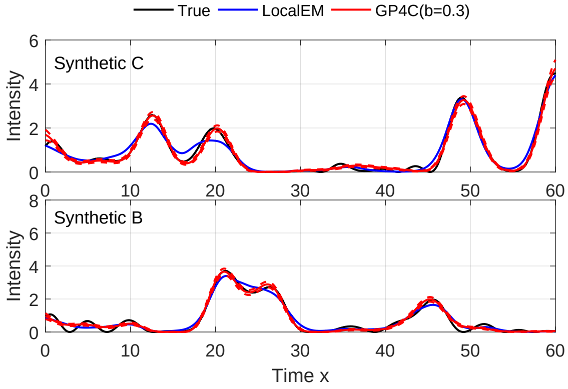

Figure 5: **Synthetic B & C Data Sets**. An illustration of the underlying intensity functions and inferred intensity functions by the LocalEM and GP4C methods. The underlying intensity function is drawn from a Gaussian process. For GP4C, a 75% credible interval is given by dotted lines.

Different choices of the hyper-parameter b. On all three synthetic data sets, we test three different choices of b in $\{0, 0.3, 1\}$. We choose the number of pseudo inputs to be 30. We calculate the MISE and $\mathcal{L}_{\text{test}}$ and the results are provided in Table 1. We see that $b = 0, 0.3$ generally outperform $b = 1$ on these simple synthetic data sets. However, the difference between $b = 0$ and $b = 0.3$ is not significant. The reason is that Inequality (14) and the range of φ on $\mathcal{X}$ are also relevant to the actual performance of different b, as we discussed in Section 4.3.

To investigate the reason behind the bad performance of $\mathcal{L}_{\text{test}}$ when $b = 1$, we plot the best result in terms of MISE during 40 trials in Figure 4. We see that GP4C ($b = 1$) over-estimates the variance of the intensity function and the over-estimated variance leads to the poor performance in $\mathcal{L}_{\text{test}}$. We fix $b = 0.3$ during the remaining experiments for simplicity.

Number of the pseudo inputs. We vary the number of pseudo inputs in GP3 and GP4C since this number de-

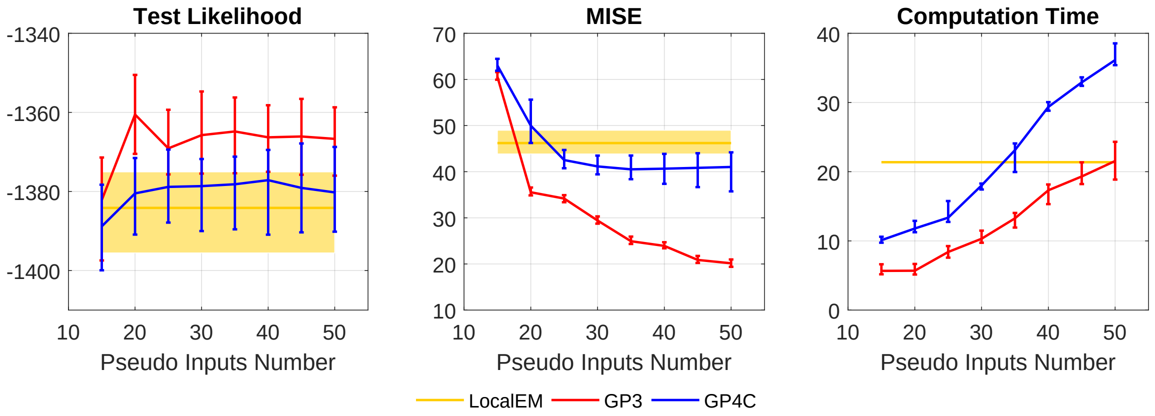

Figure 6: **Synthetic Data Set**. Comparison of performance of GP3, GP4C and LocalEM in terms of $\mathcal{L}_{\text{test}}$, MISE and T when varying the number of pseudo inputs for sparse GPs. For the test likelihood, MISE and the computation time, the median, the 0.25 and 0.75 quantiles of the statistics in 40 experiments are shown with error bars or shaded area. For GP3 and GP4C, MISE and $\mathcal{L}_{\text{test}}$ stay relatively stable with the increase of the number of pseudo inputs.

termines the accuracy of approximation in a sparse GP. We expect that for GP-based methods the test likelihood will be relatively stable when increasing the number of pseudo inputs according to previous studies on sparse GPs (Titsias, 2009).

The result for the Synthetic A data set is given in Figures 6. In Figure 6, we see that for GP3 and GP4C, MISE and $\mathcal{L}_{\text{test}}$ stay relatively stable with the increase of the number of pseudo inputs. The computation time of GP3 and GP4C will grow with the increase of the number of pseudo inputs.

In both Table 1 and Figure 6, we see that GP4C outperforms LocalEM on these three datasets. However, we also notice that there is still a gap between GP3 and GP4C in terms of $\mathcal{L}_{\text{test}}$ and MISE in Table 1. Two reasons may account for this fact. The first one is that the data are provided in the form of panel counts rather than exact timestamps. The second reason is that we use a lower bound of the true ELBO to perform the variational inference, which may lead to a bias. This bias can be alleviated with the stochastic variational inference (Paisley et al., 2012), where our lower bound can serve as a control variate. We leave this as a future study.

An additional experiment in which we increase the number of training subjects to evaluate the gain in performance on the Synthetic A data set is given in Appendix E.2.

5.3 REAL WORLD DATA SETS

Sun and Zhao (2016) provided three panel count data sets. Some statistics can be found in Table 2. A brief description about the these data sets can be found in Appendix F.

Table 2: Statistics about the three data sets, where K, $\mathcal{X}$, $\bar{N}$ and N denote the number of subjects in each data set, the underlying continuous space, the number of different end points and the number of different intervals $\mathcal{X}_i^{(k)}$, respectively.

Data Set	$\mathcal{X}$	K	$\bar{N}$	N
Na-A	$[0, 55]$	65	45	109
Na-B	$[0, 55]$	48	38	84
Bl-A	$[0, 53]$	38	52	176
Bl-B	$[0, 53]$	47	52	201
Sk-A & Sk-B	$[0, 61.57]$	143	751	816
Sk-C & Sk-D	$[0, 62.63]$	147	808	887

We use 18 pseudo inputs for all real world experiments. In each trial, we randomly split each data set into two parts, which are $\mathcal{D}_{\text{train}}^{(s)}$ (50%) and $\mathcal{D}_{\text{test}}^{(s)}$ (50%). On these three data sets, since the original data are in the form of panel counts, GP3 is not tested. We compare GP4C with LocalEM in terms of $\mathcal{L}_{\text{test}}$ and the computation time T.

The results are given in Table 3. The standard deviation of the likelihood is large since the likelihood depends on the censored intervals of the subjects, which vary greatly in different train/test split. We conduct an experiment to reduce the standard deviation in Appendix H. In Table 3, LocalEM performs better on the Nausea and Bladder data sets in terms of the computation time T. GP4C outperforms LocalEM in terms of test likelihood $\mathcal{L}_{\text{test}}$ in all data sets.

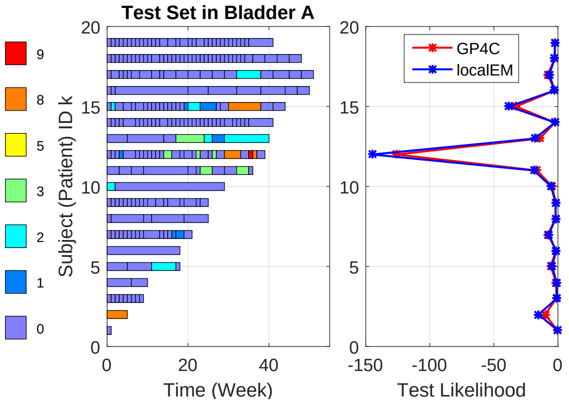

Figure 7: **Bladder A Data Set**. An illustration of the panel count data in the test set (Left) and the test likelihood from GP4C and LocalEM of each subject (Right). GP4C mainly outperforms LocalEM on two subjects whose numbers of newly-occurred cancers are large (No. 12 and 15).

To see the difference between GP4C and LocalEM, we show the result of inferred intensities by two algorithms during one trial on the Bladder A data set in Figure 2. We see that GP4C provides the additional uncertainty which helps improve $\mathcal{L}_{\text{test}}$ compared with LocalEM. Since the Bladder A set is small, we plot the panel count data in the training set in Figure 1. The test set and the test likelihood of all its subjects are given in Figure 7. From the test likelihood of each subject, we see that GP4C outperforms LocalEM on two subjects whose counts of newly-occurred tumors are large (No. 12 and No. 15). The count 8 never occurs in the training set and a point-estimate will fail to model this uncertainty while a GP-modulated method will take the uncertainty into consideration.

Another observation about this data set is that there is a heterogeneity across all subjects. The traditional approach to modeling heterogeneity is to add an additional variable on the intensity function for each subject (Cook and Lawless, 2007). We briefly discuss how to add this change to GP4C in Appendix G.

6 CONCLUSION

We presented the first framework for GP-modulated Poisson processes when data appear in the form of panel counts. We derived a tractable lower bound for the intractable evidence lower bound when modeling the panel count data using the GP-modulated intensity function. Our model, GP4C, outperforms a non-Bayesian method using the maximum likelihood criterion in terms of test likelihood and achieves comparable results in terms of computational time.

Table 3: Mean and standard deviation of the test likelihood ($\mathcal{L}_{\text{test}}$) and the computation time T measured in seconds on the three panel count data sets over 40 runs. LocalEM performs better on the Nausea and Bladder data sets in terms of computation time. In all data sets, GP4C performs better on the test likelihood and outperforms LocalEM on computation time in the Skin data sets.

Data Set	**METHOD**	$\mathcal{L}_{\text{test}}$	$T[s]$
Na-A	LocalEM	-492.1±306.1	1±0
	GP4C	-484.9±201.8	10±10
Na-B	LocalEM	-473.2±212.2	1±0
	GP4C	-411.0±184.3	10±7
Bl-A	LocalEM	-201.8±46.9	1±0
	GP4C	-182.2±47.3	25±9
Bl-B	LocalEM	-313.1±54.2	1±0
	GP4C	-310.4±54.9	26±21
Sk-A	LocalEM	-259.1±27.3	39±3
	GP4C	-258.7±26.7	33±6
Sk-B	LocalEM	-198.1±47.1	39±3
	GP4C	-191.2±42.5	24±4
Sk-C	LocalEM	-358.0±35.8	47±4
	GP4C	-355.7±36.0	21±12
Sk-D	LocalEM	-200.9±31.9	46±3
	GP4C	-198.9±30.6	27±4

In the future, we plan to implement the stochastic variational inference algorithm to evaluate the bias in the tractable lower bound. We are also considering to find an applicable two-dimensional data set where we can extend our algorithm to spatial point processes.

Acknowledgements

We thank the anonymous reviewers for their helpful suggestions. MS was supported by KAKENHI 17H00757.

References

Adams, R. P., Murray, I., and MacKay, D. J. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 9–16. ACM.

Cook, R. J. and Lawless, J. (2007). *The Statistical Analysis of Recurrent Events*. Springer Science & Business Media.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science*, pages 542–563.

Fan, C.-P. S., Stafford, J., and Brown, P. E. (2011). Local-EM and the EMS algorithm. *Journal of Computational and Graphical Statistics*, 20(3):750–766.

Flaxman, S., Wilson, A., Neill, D., Nickisch, H., and Smola, A. (2015). Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods. In *International Conference on Machine Learning*, pages 607–616.

Gunter, T., Lloyd, C., Osborne, M. A., and Roberts, S. J. (2014). Efficient Bayesian nonparametric modeling of structured point processes. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 310–319. AUAI Press.

Kingman, J. F. C. (1993). *Poisson Processes*. Wiley Online Library.

Lian, W., Henao, R., Rao, V., Lucas, J., and Carin, L. (2015). A multitask point process predictive model. In *International Conference on Machine Learning*, pages 2030–2038.

Lloyd, C., Gunter, T., Osborne, M., and Roberts, S. (2015). Variational inference for Gaussian process modulated Poisson processes. In *International Conference on Machine Learning*, pages 1814–1822.

Lloyd, C., Gunter, T., Osborne, M., Roberts, S., and Nickson, T. (2016). Latent point process allocation. In *Artificial Intelligence and Statistics*, pages 389–397.

Moser, S. M. (2007). Some expectations of a non-central chi-square distribution with an even number of degrees of freedom. In *TENCON 2007-2007 IEEE Region 10 Conference*, pages 1–4. IEEE.

Paisley, J. (2010). Two useful bounds for variational inference. Technical report, Technical report, Department of Computer Science, Princeton University, Princeton, NJ.

Paisley, J., Blei, D. M., and Jordan, M. I. (2012). Variational Bayesian inference with stochastic search. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1363–1370. Omnipress.

Schmidt, M., Berg, E., Friedlander, M., and Murphy, K. (2009). Optimizing costly functions with simple constraints: A limited-memory projected quasi-Newton algorithm. In *Artificial Intelligence and Statistics*, pages 456–463.

Sun, J. and Zhao, X. (2016). *Statistical Analysis of Panel Count Data*. Springer.

Thall, P. F. and Lachin, J. M. (1988). Analysis of recurrent events: Nonparametric methods for random-interval count data. *Journal of the American Statistical Association*, 83(402):339–347.

Titsias, M. K. (2009). Variational model selection for sparse Gaussian process regression. *Report, University of Manchester, UK*.

Wellner, J. A. and Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Annals of Statistics*, pages 779–814.

Wellner, J. A., Zhang, Y., et al. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *The Annals of Statistics*, 35(5):2106–2142.

Zhang, Y. and Jamshidian, M. (2003). The gamma-frailty Poisson model for the nonparametric estimation of panel count data. *Biometrics*, 59(4):1099–1106.