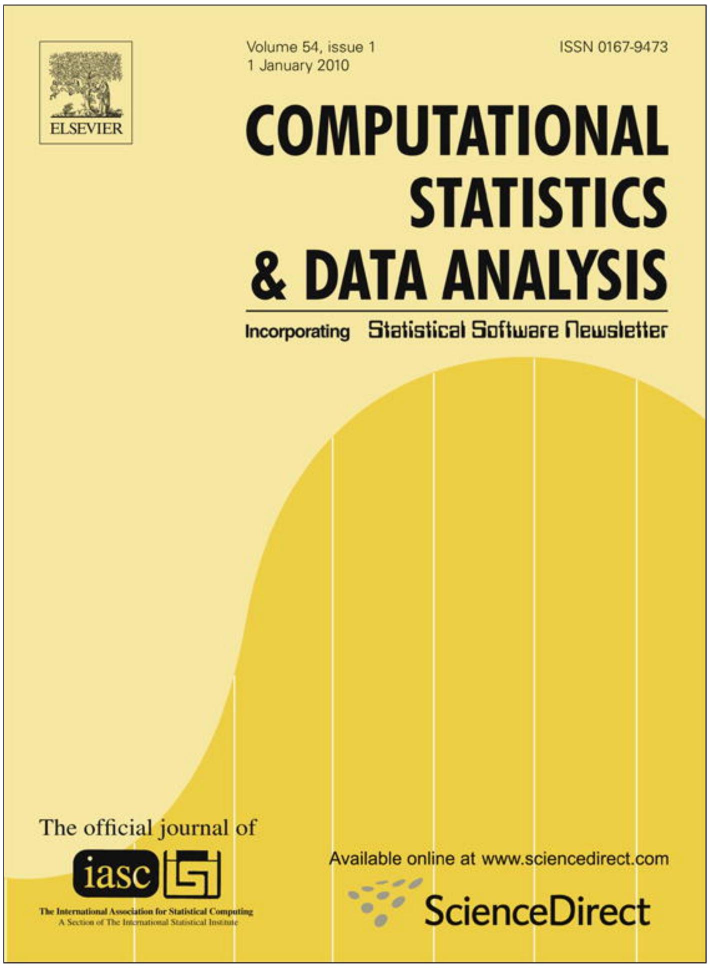
Volume 54, issue 1
1 January 2010
ISSN 0167-9473
ELSEVIER
COMPUTATIONAL STATISTICS & DATA ANALYSIS
Incorporating Statistical Software Newsletter
The official journal of
iasc
Available online at www.sciencedirect.com
ScienceDirect

Computational Statistics and Data Analysis 54 (2010) 135–142

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A nonparametric test for the equality of counting processes with panel count data

N. Balakrishnan [a], Xingqiu Zhao [b,*]

[a] Department of Mathematics & Statistics, McMaster University, Hamilton, Canada

[b] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 5 November 2008
Received in revised form 13 July 2009
Accepted 15 July 2009
Available online 21 July 2009

ABSTRACT

This paper considers the problem of nonparametric comparison of counting processes with panel count data, which arise naturally when recurrent events are considered. For the problem considered, we construct a new nonparametric test statistic based on the nonparametric maximum likelihood estimator of the mean function of the counting processes over observation times. The asymptotic distribution of the proposed statistic is derived and its finite-sample property is examined through Monte Carlo simulations. The simulation results show that the proposed method is good for practical use and also more powerful than the existing nonparametric tests based on the nonparametric maximum pseudo-likelihood estimator. A set of panel count data from a floating gallstone study is analyzed and presented as an illustrative example.

1. Introduction

Consider a study that concerns some recurrent event and suppose that each subject in the study gives rise to a counting process $N(t)$, denoting the total number of occurrences of the event of interest up to time t. Also suppose that for each subject, observations include only the values of $N(t)$ at discrete observation times or the numbers of occurrences of the event between the observation times. Such data are usually referred to as *panel count data* (Sun and Kalbfleisch, 1995; Wellner and Zhang, 2000), which frequently occur in medical follow-up studies and reliability experiments, for example. Our focus here will be on the situation when such a study involves $k\,(\geq 2)$ groups. Let $\Lambda_l(t)$ denote the mean function of $N(t)$ corresponding to the lth group for $l = 1, \ldots, k$. The problem of interest is then to test the hypothesis $H_0 : \Lambda_1(t) = \cdots = \Lambda_k(t), t \in (0, \tau]$, where τ is the maximum observation time.

For the analysis of panel count data, Sun and Kalbfleisch (1995) and Wellner and Zhang (2000) studied estimation of the mean function of $N(t)$. Sun and Wei (2000), Zhang (2002), Hu et al. (2003) and Sun et al. (2007) discussed regression analysis for such data. The model-based methods depend on the expression of the model used and model checking is needed but difficult in practice. To test the hypothesis H_0, Thall and Lachin (1988) suggested to transform the problem to a multivariate comparison problem and then apply a multivariate Wilcoxon-type rank test. Sun and Fang (2003) proposed a nonparametric procedure for this problem under the assumption that treatment indicators can be regarded as independent and identically distributed random variables. Park et al. (2007) proposed a class of nonparametric tests for the two-sample comparison based on the isotonic regression estimator of the mean function of counting process. Zhang (2006) also presented nonparametric tests for the problem based on the nonparametric maximum pseudo-likelihood estimator that is equivalent to the isotonic regression estimator (Wellner and Zhang, 2000). Also, Wellner and Zhang (2000) showed

* Corresponding author. Tel.: +852 27666921; fax: +852 23629045.
E-mail address: xingqiu.zhao@polyu.edu.hk (X. Zhao).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.07.015

through Monte Carlo simulations that the nonparametric maximum likelihood estimator (NPMLE) of the mean function is more efficient than the nonparametric maximum pseudo-likelihood estimator (NPMPLE). However, no nonparametric tests have been discussed in the literature for panel count data based on the NPMLE since the NPMLE is more complicated both theoretically and computationally. It is, therefore, particularly important to develop nonparametric tests based on the NPMLE for panel count data. However, unlike the isotonic regression estimate, the maximum likelihood estimate has no closed-form expression and its computation requires an iterative convex minorant algorithm. In this paper, for simplicity, we focus on the situation considered by Sun and Fang (2003) and propose a nonparametric test using the maximum likelihood estimator and then compare its power with those of existing tests for the problem of two-sample nonparametric comparison of counting processes with simulated panel count data.

The rest of this paper is organized as follows. Section 2 discusses estimation of the mean function and the existing nonparametric tests for the hypothesis H_0 when only panel count data are available. Section 3 presents a new nonparametric test statistic motivated by the property of the NPMLE and the idea used by Sun and Fang (2003). Also, the asymptotic normality of the test statistic is established. In Section 4, finite-sample property of the proposed test statistic is examined through Monte Carlo simulations. In Section 5, we apply the proposed method to a data from a floating gallstone study. Finally, some concluding remarks are made in Section 6.

2. Nonparametric maximum likelihood estimation of mean function

Wellner and Zhang (2000) studied two estimators of the mean of a counting process with panel count data: the nonparametric maximum pseudo-likelihood estimator and the nonparametric maximum likelihood estimator. To describe the test statistics, we first introduce the NPMLE. Suppose that $N = \{N(t) : t \geq 0\}$ is a nonhomogeneous Poisson process with the mean function $E(N(t)) = \Lambda_0(t)$. Also suppose that for each subject, observations include only the values of $N(t)$ at discrete observation times $0 < T_{K,1} < T_{K,2} < \cdots < T_{K,K}$, where the total number of observations K is an integer-valued random variable. The observed data from the counting process consist of $X = (K, T, N)$, where $T = (T_{K,1}, T_{K,2}, \ldots, T_{K,K})$ and $N = (N(T_{K,1}), N(T_{K,2}), \ldots, N(T_{K,K}))$. We assume that (K, T) are independent of N. Let $X = (K, T, N)$. Then, $X_i = (K_i, T_i, N_i)$, $i = 1, 2, \ldots, n$, with

$$T_i = (T_{K_i,1}, T_{K_i,2}, \ldots, T_{K_i,K_i}) \quad \text{and} \quad N_i = (N_i(T_{K_i,1}), N_i(T_{K_i,2}), \ldots, N_i(T_{K_i,K_i})),$$

is a random sample of size n from the distribution of X. Let $\mathbf{X} = (X_1, \ldots, X_n)$. Then the log likelihood function for the mean function Λ is

$$l_n(\Lambda|\mathbf{X}) = \sum_{i=1}^{n}\sum_{j=1}^{K_i}\left(N_i(T_{K_i,j}) - N_i(T_{K_i,j-1})\right)\log\left(\Lambda(T_{K_i,j}) - \Lambda(T_{K_i,j-1})\right) - \sum_{i=1}^{n}\Lambda(T_{K_i,K_i})$$

after omitting the parts independent of Λ.

Let $t_1 < \cdots < t_m$ denote the ordered distinct observation time points in the set of all observation time points $\{T_{K_i,j}, j = 1, \ldots, K_i, i = 1, \ldots, n\}$. For $\ell = 1, \ldots, m$, define $\Lambda_\ell = \Lambda(t_\ell)$ and write $\underline{\Lambda} = (\Lambda_1, \ldots, \Lambda_m)$. For $1 \leq r < \ell \leq m$, set

$$A_{r,\ell} = \sum_{i=1}^{n}\sum_{j=1}^{K_i}\left(N_i(T_{K_i,j}) - N_i(T_{K_i,j-1})\right) 1_{\{T_{K_i,j}=t_\ell, T_{K_i,j-1}=t_r\}},$$

and for $\ell = 1, \ldots, m$,

$$B_\ell = \sum_{i=1}^{n} 1_{\{T_{K_i,K_i}=t_\ell\}}.$$

Then $l_n(\Lambda|\mathbf{X})$ can be rewritten, with a slight abuse of notation, as

$$l_n(\Lambda|\mathbf{X}) = l_n(\underline{\Lambda}|\mathbf{X}) = \sum_{r=1}^{m}\sum_{\ell=r+1}^{m} A_{\ell,r}\log(\Lambda_\ell - \Lambda_r) - \sum_{\ell=1}^{m} B_\ell\Lambda_\ell$$

and the NPMLE of Λ_0, $\hat{\Lambda}_n$, is defined to be the nondecreasing, nonnegative step function with possible jumps only occurring at t_ℓ, $\ell = 1, \ldots, m$, that maximizes $l_n(\Lambda|\mathbf{X})$. Here, only $\Lambda_1, \ldots, \Lambda_m$ are identifiable. Wellner and Zhang (2000) gave the characteristic and the algorithm for computing this estimator, and studied its asymptotic properties.

The existing nonparametric tests (Park et al., 2007; Zhang, 2006) are based on the asymptotic normality of a smooth functional of the nonparametric maximum pseudo-likelihood estimator $\tilde{\Lambda}_n$ (the isotonic regression estimator). However, it is unknown if the asymptotic normality of the functional of the nonparametric maximum likelihood estimator still holds because of the complexity of the NPMLE. We observe that the test presented by Sun and Fang (2003) is related to the characteristic of the $\tilde{\Lambda}_n$ given by

$$\sum_{i=1}^{n}\sum_{j=1}^{K_i}(\tilde{\Lambda}_n(T_{K_i,j}) - N_i(T_{K_i,j})) = 0. \tag{1}$$

However, from Eq. (2.13) of Wellner and Zhang (2000), the corresponding characteristic of the NPMLE can be written as

$$\sum_{i=1}^{n}\left[\sum_{j=1}^{K_i-1}\hat{\Lambda}_n(T_{K_i,j})\left\{\frac{\Delta N_i(T_{K_i,j+1})}{\Delta\hat{\Lambda}_n(T_{K_i,j+1})}-\frac{\Delta N_i(T_{K_i,j})}{\Delta\hat{\Lambda}_n(T_{K_i,j})}\right\}+\hat{\Lambda}_n(T_{K_i,K_i})\left\{1-\frac{\Delta N_i(T_{K_i,K_i})}{\Delta\hat{\Lambda}_n(T_{K_i,K_i})}\right\}\right]=0, \tag{2}$$

where $\Delta\Lambda(T_{K,j}) = \Lambda(T_{K,j}) - \Lambda(T_{K,j-1})$ and $\Delta N(T_{K,j}) = N(T_{K,j}) - N(T_{K,j-1})$. Clearly, the structure of (2) is different from that of (1) and considerably more complicated. Therefore, we need to develop a new form of test statistic when the NPMLE is used to estimate the mean function of counting process with panel count data.

3. A nonparametric test with panel count data

Consider a longitudinal study that is concerned with some recurrent event and involves n independent subjects from k different groups. Let Z_i denote the group indicator of subject i ($i = 1, \ldots, n$) and assume that group indicator is a scalar variable. For the two-sample comparison problem, $k = 2$ and $Z_i = 0$ or 1. For the dose-effects problem, k is the number of doses tested in the experiment and Z_i denotes the dose given to subject i ($i = 1, \ldots, n$). Let $N_i(t)$ denote the counting process arising from subject i and $\Lambda_l(t)(l = 1, \ldots, k)$ be defined as before, for $i = 1, \ldots, n$. Suppose that each subject is observed only at discrete time points $0 < T_{K_i,1} < \cdots < T_{K_i,K_i}$ and that no information is available about $N_i(t)$ between observation times; that is, only panel count data are available. Also assume that N_i and $(K_i, T_{K_i,1}, \ldots, T_{K_i,K_i})$ are independent of Z_i. For simplicity, assume that H_0 is true, and let $\Lambda_0(t)$ denote the common mean function of the $N_i(t)$'s.

Let $\hat{\Lambda}_n$ be the nonparametric maximum likelihood estimate of Λ_0 based on the combined data. To test the hypothesis H_0, motivated by the characteristic of the NPMLE and an idea used in Sun and Fang (2003), we propose the statistic

$$U_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i\left[\sum_{j=1}^{K_i-1}\hat{\Lambda}_n(T_{K_i,j})\left\{\frac{\Delta N_i(T_{K_i,j+1})}{\Delta\hat{\Lambda}_n(T_{K_i,j+1})}-\frac{\Delta N_i(T_{K_i,j})}{\Delta\hat{\Lambda}_n(T_{K_i,j})}\right\}+\hat{\Lambda}_n(T_{K_i,K_i})\left\{1-\frac{\Delta N_i(T_{K_i,K_i})}{\Delta\hat{\Lambda}_n(T_{K_i,K_i})}\right\}\right].$$

Let $\mathcal{B}$ denote the collection of Borel sets in $\mathcal{R}$, and let $\mathcal{B}_{[0,\tau]} = \{B \cap [0, \tau], B \in \mathcal{B}\}$. On $([0, \tau], \mathcal{B}_{[0,\tau]})$, set

$$\mu(B) = \sum_{k=1}^{\infty}\Pr(K = k)\sum_{j=1}^{k}\Pr(T_{k,j} \in B|K = k).$$

For establishing asymptotic result on U_n, we need the following regularity conditions:

C.1. There exists a constant K_0 such that $\Pr\{K \le K_0\} = 1$ and that the random variables $T_{k,j}$'s take values in a bounded set $[\tau_0, \tau]$, where $\tau_0, \tau \in (0, \infty)$.
C.2. The mean function Λ_0 is continuous such that $\Lambda_0(\tau_0) > 0$ and $\Lambda_0(\tau) \le M$ for some constant $M \in (0, \infty)$.
C.3. There exists a constant L_0 such that

$$\Pr\left\{\min_{1\le j\le K}(\Lambda_0(T_{K,j+1}) - \Lambda_0(T_{K,j})) \ge L_0\right\} = 1.$$

C.4. $E\{N(t)\}^2 \le M_1$ for all $t \le \tau$ where M_1 is a constant.
C.5. $\mu(\{\tau_0\}) > 0$ and for all $\tau_0 < \tau_1 < \tau_2 < \tau$, $\Lambda_0(\tau_1) < \Lambda_0(\tau_2) < \Lambda_0(\tau)$ implies $\mu((\tau_1, \tau_2)) > 0$.

Condition C.3 holds if Λ_0 is differentiable, Λ_0' has a positive lower bound in $[\tau_0, \tau]$, and $\Pr\{\min_{1\le j\le K}(T_{K,j} - T_{K,j-1}) \ge s_0\}$ for some fixed time s_0, where s_0 can be considered as the smallest length of consecutive observation times. Condition C.5 holds if Λ_0 is strictly increasing, $\Pr\{T_{K,1} = \tau_0\} > 0$ and $\mu'(t) > 0$ for $t \in (\tau_0, \tau)$. The asymptotic distribution of U_n is as presented in the following theorem.

Theorem 1. *Suppose that Conditions* C.1–C.5 *hold. Also suppose that Z_i's are independent and identically distributed random variables. Then as $n \to \infty$,*

$$U_n \longrightarrow U$$

in distribution, where U has a normal distribution with mean zero and variance σ^2 with

$$\sigma^2 = E\left[(Z - E(Z))\left\{\sum_{j=1}^{K-1}\Lambda_0(T_{K,j})\left(\frac{\Delta N(T_{K,j+1})}{\Delta\Lambda_0(T_{K,j+1})}-\frac{\Delta N(T_{K,j})}{\Delta\Lambda_0(T_{K,j})}\right)+\Lambda_0(T_{K,K})\left(1-\frac{\Delta N(T_{K,K})}{\Delta\Lambda_0(T_{K,K})}\right)\right\}\right]^2$$

which can be consistently estimated by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(Z_i - \bar{Z})\left\{\sum_{j=1}^{K_i-1}\hat{\Lambda}_n(T_{K_i,j})\left(\frac{\Delta N_i(T_{K_i,j+1})}{\Delta\hat{\Lambda}_n(T_{K_i,j+1})}-\frac{\Delta N_i(T_{K_i,j})}{\Delta\hat{\Lambda}_n(T_{K_i,j})}\right)+\hat{\Lambda}_n(T_{K_i,K_i})\left(1-\frac{\Delta N_i(T_{K_i,K_i})}{\Delta\hat{\Lambda}_n(T_{K_i,K_i})}\right)\right\}\right]^2,$$

where $\bar{Z} = \sum_{i=1}^{n} Z_i/n$.

The proof is given in the Appendix. Note that the asymptotic result requires Z_i's as independent and identically distributed random variables. For example, this assumption is satisfied in randomized clinical trials, where all patients in the study are randomly assigned to one of the treatments. Form the proof of the theorem, this assumption can be replaced by the assumption that the Z_i's are uncorrelated random variables with common means and variances. By Theorem 1, if n is large, the hypothesis H_0 can be tested by using the statistic $T = U_n/\hat{\sigma}$, which has an asymptotic standard normal distribution.

Remark. In Theorem 1, for convenience, we assume that Z_i's are scalars. The results are easily extended to vectors Z_i's. For the k-sample problem ($k > 2$), let Z_i be the k-dimensional vector of treatment indicators associated with subject i whose lth element is equal to one if it is from population l and zero otherwise. Then U_n has an asymptotic normal distribution with mean vector $\mathbf{0}$ and covariance matrix

$$\Sigma = E\left[(Z - E(Z))(Z - E(Z))^{\mathrm{T}}\left\{\sum_{j=1}^{K-1}\Lambda_0(T_{K,j})\left(\frac{\Delta N(T_{K,j+1})}{\Delta\Lambda_0(T_{K,j+1})} - \frac{\Delta N(T_{K,j})}{\Delta\Lambda_0(T_{K,j})}\right) + \Lambda_0(T_{K,K})\left(1 - \frac{\Delta N(T_{K,K})}{\Delta\Lambda_0(T_{K,K})}\right)\right\}^2\right]$$

which can be consistently estimated by

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\left[(Z_i - \bar{Z})(Z_i - \bar{Z})^{\mathrm{T}}\left\{\sum_{j=1}^{K_i-1}\hat{\Lambda}_n(T_{K_i,j})\left(\frac{\Delta N_i(T_{K_i,j+1})}{\Delta\hat{\Lambda}_n(T_{K_i,j+1})} - \frac{\Delta N_i(T_{K_i,j})}{\Delta\hat{\Lambda}_n(T_{K_i,j})}\right) + \hat{\Lambda}_n(T_{K_i,K_i})\left(1 - \frac{\Delta N_i(T_{K_i,K_i})}{\Delta\hat{\Lambda}_n(T_{K_i,K_i})}\right)\right\}^2\right].$$

Note that the sum of all components of U_n equals 0. Let $\mathbf{U}_0$ denote the first $(k-1)$ components of $\mathbf{U}_n$ and $\hat{\mathbf{\Sigma}}_0$ the matrix obtained by deleting the last row and column of $\hat{\mathbf{\Sigma}}$. Then, using Theorem 1, the hypothesis H_0 can be tested by means of the statistic $\chi_0^2 = \mathbf{U}_0^{\mathrm{T}}\hat{\mathbf{\Sigma}}_0^{-1}\mathbf{U}_0$, which have asymptotically a central χ^2-distribution with $(k-1)$ degrees of freedom.

4. Simulation study

To examine the finite-sample property of the proposed test statistic and compare it with those of the tests presented by Sun and Fang (2003), Park et al. (2007) and Zhang (2006), we carry out a simulation study for the two-sample comparison problem. Let T be defined as in Section 3. Let T_{SF} denote the test proposed by Sun and Fang (2003), and let T_i $(i = 1, 2, 3)$ denote the tests presented by Park et al. (2007) and Zhang (2006) with three different weight processes: $W_n^{(1)}(t) = 1$, $W_n^{(2)}(t) = Y_n(t) = \sum_{i=1}^{n} I(t \le T_{K_i,K_i})/n$, and $W_n^{(3)}(t) = \{Y_{n_1}(t)Y_{n_2}(t)\}/Y_n(t)$, where $Y_{n_l}(t) = \sum_{i\in S_l} I(t \le T_{K_i,K_i})/n_l$, S_l denotes the set of indices for subjects in group l and n_l is the number of subjects in group l, $l = 1, 2$. To generate panel count data $\{K_i, T_{K_i,j}, N_i(T_{K_i,j}), j = 1, \ldots, K_i, i = 1, \ldots, n\}$, we mimic medical follow-up studies similar to the example discussed in the next section. We first generate the number of observation times K_i from the uniform distribution $U\{1, \ldots, 10\}$, and then, given K_i, we generate observation times $T_{K_i,j}$'s from $U\{1, \ldots, 10\}$, for simplicity. To generate $N_i(T_{K_i,j})$'s, we assume that N_i's are nonhomogeneous Poisson or mixed Poisson processes. In particular, let $\{\nu_i, i = 1, \ldots, n\}$ be independent and identically distributed random variables, and given ν_i, let $N_i(t)$ be a Poisson process with mean function $\Lambda_i(t|\nu_i) = E(N_i(t)|\nu_i)$. Here, it is assumed that $Z_i = 0$ for $i \in S_1$ and $Z_i = 1$ for $i \in S_2$. To assess the performance of the proposed test, we consider two cases as follows:

Case 1. $\Lambda_i(t|\nu_i) = \nu_i t$ for $i \in S_1$, $\Lambda_i(t|\nu_i) = \nu_i t \exp(\beta)$ for $i \in S_2$.

Case 2. $\Lambda_i(t|\nu_i) = \nu_i t$ for $i \in S_1$, $\Lambda_i(t|\nu_i) = \nu_i\sqrt{\beta t}$ for $i \in S_2$.

Figs. 1 and 2 display the graphs of the true mean functions for two cases with $\nu = 1$ and different values of β. It can be seen that the two mean functions do not overlap in Case 1 and they cross over in Case 2.

For each case, we consider $\nu_i = 1$ and $\nu_i \sim \text{Gamma}(2, 1/2)$ corresponding to Poisson and mixed Poisson processes, respectively. For each setting, we consider two sample sizes, $n_1 = n_2 = 50$ and 100, respectively. The NPMLE $\hat{\Lambda}_n$ is computed by using the modified iterative convex minorant algorithm (MICM); see Wellner and Zhang (2000). We first compute the NPMPLE $\tilde{\Lambda}_n$ (the isotonic regression estimate) by using the max–min formula given in (2.6) of Wellner and Zhang (2000). Then, following Wellner and Zhang (2000), we choose the linearly interpolated $\tilde{\Lambda}_n$ as the starting point for the MICM algorithm. Based on this initial set-up, the MICM algorithm preforms well as Wellner and Zhang (2000) pointed out. Here, we use the R code for the MICM algorithm provided by Zhang and Liu. All the results reported are based on 1000 Monte Carlo replications using R software.

Tables 1 and 2 present the estimated sizes and powers of the proposed test statistic T, the test T_{SF} (Sun and Fang, 2003) and the test statistics T_i's (Park et al., 2007; Zhang, 2006) at significance level $\alpha = 0.05$ for different values of β for the two cases, respectively. When $\nu_i = 1$, the $N_i(t)$'s are Poisson processes; when $\nu_i \sim \text{Gamma}(2, 1/2)$, the $N_i(t)$'s are mixed Poisson processes. The first part of the table is for the situation with the total sample size of 100 and the second part is for the situation with the total sample size of 200. For the situations considered here, the proposed test displays the highest power. As expected, the power increases when the sample size increases, and the power decreases in the presence of more variability. In particular, when the mean functions overlap, it can be seen from Table 2 that the proposed test has a good power while the powers of the tests based on the NPMPLE are very poor for $\beta = 5$ in Case 2 and worse for the case of the mixed Poisson process. In this case, the proposed test based on the NPMLE is much more powerful than the existing tests based on NPMPLE.

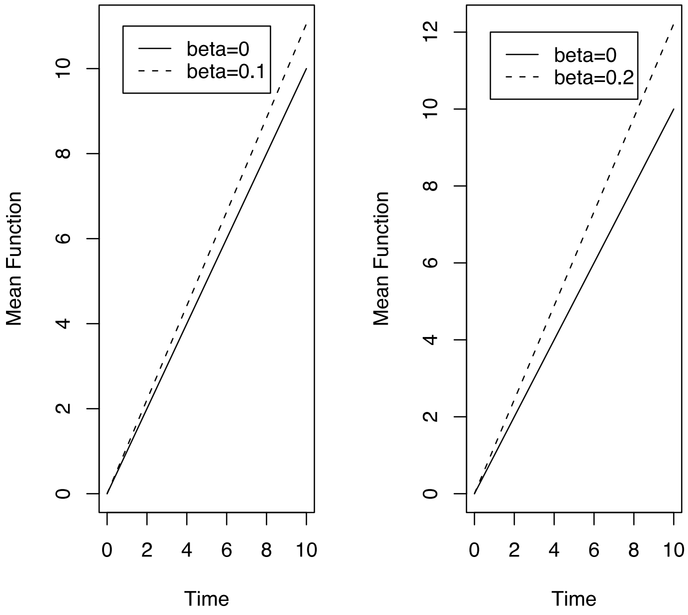

Fig. 1. True mean functions for Case 1 with $\nu = 1$ and $\beta = 0.1, 0.2$.

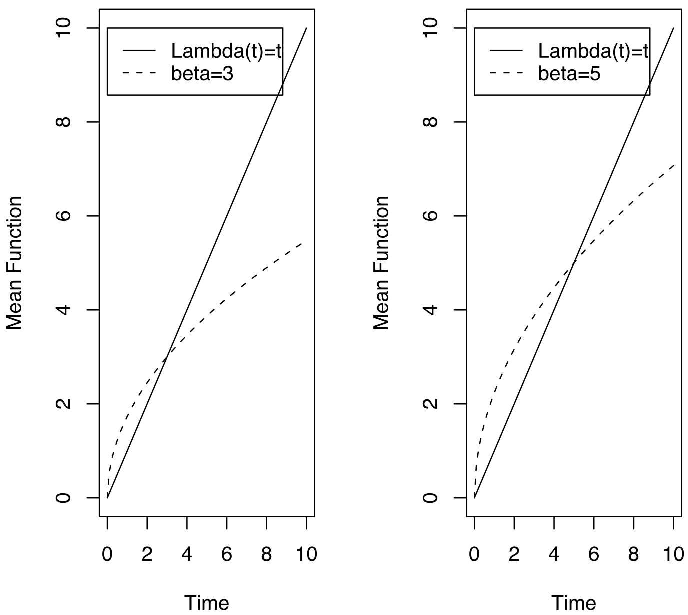

Fig. 2. True mean functions for Case 2 with $\nu = 1$ and $\beta = 3, 5$.

Table 1
Estimated sizes and powers of the proposed test T and other tests T_{SF}, T_1, T_2, T_3 in Case 1.

$n_1 = n_2$	β	N_i's are Poisson processes					N_i's are mixed Poisson processes				
		T	T_{SF}	T_1	T_2	T_3	T	T_{SF}	T_1	T_2	T_3
50	0.0	0.051	0.047	0.053	0.055	0.055	0.046	0.044	0.045	0.047	0.047
	0.1	0.298	0.207	0.214	0.200	0.200	0.098	0.083	0.084	0.085	0.085
	0.2	0.855	0.693	0.697	0.667	0.665	0.223	0.183	0.185	0.184	0.184
	0.3	1.000	0.979	0.981	0.974	0.974	0.450	0.370	0.380	0.375	0.375
100	0.0	0.049	0.041	0.043	0.047	0.047	0.043	0.046	0.048	0.045	0.045
	0.1	0.553	0.422	0.423	0.405	0.405	0.141	0.111	0.114	0.111	0.111
	0.2	0.990	0.957	0.958	0.948	0.947	0.411	0.316	0.317	0.307	0.307
	0.3	1.000	1.000	1.000	1.000	1.000	0.710	0.590	0.596	0.592	0.592

5. Illustrative example

To illustrate the proposed method, we consider a floating gallstone study presented by Thall and Lachin (1988). The data comprise the first year follow-up of the patients in two study groups, placebo (48) and high-dose chenodiol (65), from the

Table 2
Estimated sizes and powers of the proposed test T and other tests T_{SF}, T_1, T_2, T_3 in Case 2.

$n_1 = n_2$	β	N_i's are Poisson processes					N_i's are mixed Poisson processes				
		T	T_{SF}	T_1	T_2	T_3	T	T_{SF}	T_1	T_2	T_3
50	3	1.000	0.955	0.956	0.900	0.899	0.864	0.380	0.386	0.321	0.318
	4	0.998	0.592	0.600	0.439	0.437	0.635	0.185	0.188	0.141	0.141
	5	0.972	0.188	0.189	0.113	0.111	0.403	0.086	0.089	0.071	0.071
100	3	1.000	0.999	0.999	0.993	0.993	0.994	0.691	0.695	0.594	0.594
	4	1.000	0.894	0.896	0.726	0.725	0.894	0.309	0.313	0.238	0.237
	5	1.000	0.284	0.290	0.140	0.139	0.667	0.095	0.096	0.066	0.066

National Cooperative Gallstone Study. The data include the successive visit times in study weeks and the associated counts of episodes of nausea. The whole study consists of 916 patients who were randomized to placebo, low dose, or high dose group and followed for up to two years and one of the objectives of the study is to test the difference of the two treatments with respect to the incidence rate of nausea.

During the study, patients were scheduled to return for clinical visits at 1, 2, 3, 6, 9, and 12 months. In reality, most of the patients visited about six times within the first year, but actual visit times differed from patient to patient. Some patients had only one visit and some had 9 visits. As pointed out by Thall and Lachin (1988), there is no evidence that the number of observations and actual observation times are related to the incidence of nausea, and so it seems reasonable to assume that conditions required for the asymptotic result are satisfied.

Define $Z_i = 1$ for patients in the placebo group and $Z_i = 0$ otherwise. To test the difference between the two groups, we apply the proposed method to the data from 113 gallstone patients in the two groups. To compute the NPMLE Λ_n of the common mean function under the null hypothesis, we choose the linearly interpolated NPMPLE $\tilde{\Lambda}_n$ as the starting point for the MICM algorithm again. Here, the MICM works well as in simulation. We then obtain $T = 0.264$ which yields a p-value of 0.792 for testing H_0 based on the asymptotic result in Theorem 1. This result suggests that the incidence rates of nausea were not significantly different for the patients in the two groups, which agrees with the findings of Schoenfield et al. (1981), Sun and Fang (2003), and Park et al. (2007).

6. Concluding remarks

This paper discusses the problem of the nonparametric comparison of counting processes when only panel count data are available. The nonparametric maximum likelihood estimators are used to estimate the mean functions of counting processes. A new nonparametric test is proposed for the problem and the asymptotic property of the test statistic is derived. Simulation studies are carried out which suggest that the proposed method works well for practical situations, and is more powerful than the existing tests based on the nonparametric maximum pseudo-likelihood estimators of the mean functions.

Note that the proposed procedure depends on the assumption that treatment indicators are uncorrelated random variables with common means and variances and distributed independently of panel counts. If the treatment indicators Z_i's and the event processes N_i's are correlated, then the asymptotic result of U_n no longer holds. More discussions about it can be found in Sun and Kalbfleisch (1993). Our test also relies on the assumption that the observation scheme with respect to the total number of observations K and panel observation times T is the same across different samples. If the observation scheme is different for different treatment groups, then the asymptotic result of U_n no longer holds since the asymptotic properties of the NPMLE $\hat{\Lambda}_n$ based on combined sample are not available anymore. Further research will be to develop a class of tests applicable to general situations by using the nonparametric maximum likelihood estimates instead of the nonparametric maximum pseudo-likelihood estimates for the mean functions.

Acknowledgements

The authors are very grateful to the editor and two referees for their helpful comments and suggestions that greatly improved the paper. They thank Drs. Ying Zhang and Minggen Lu for supplying the R code for the modified iterative convex minorant algorithm presented by Wellner and Zhang (2000). The research of Xingqiu Zhao is supported in part by a grant from the Hong Kong Polytechnic University.

Appendix

Proof of Theorem 1. Let

$$h(X,\Lambda) = \sum_{j=1}^{K-1} \Lambda(T_{K,j}) \left\{ \frac{\Delta N(T_{K,j+1})}{\Delta\Lambda(T_{K,j+1})} - \frac{\Delta N(T_{K,j})}{\Delta\Lambda(T_{K_i,j})} \right\} + \Lambda(T_{K,K}) \left\{ 1 - \frac{\Delta N(T_{K,K})}{\Delta\Lambda(T_{K,K})} \right\}.$$

From Eq. (2), we have

$$\sum_{i=1}^{n} h(X_i, \hat{\Lambda}_n) = 0.$$

Now U_n can be expressed as

$$U_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Z_i - E(Z_i)\} h(X_i, \hat{\Lambda}_n) = V_n + \Delta_n,$$

where

$$V_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Z_i - E(Z_i)\} h(X_i, \Lambda_0)$$

and

$$\Delta_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Z_i - E(Z_i)\}\{h(X_i, \hat{\Lambda}_n) - h(X_i, \Lambda_0)\}.$$

It is easy to see that V_n is a U-statistic and has an asymptotic normal distribution with mean zero and variance σ^2 that can be consistently estimated by $\hat{\sigma}^2$ presented earlier in Theorem 1. Hence, it is sufficient to show that Δ_n converges in probability to zero. Let

$$\mathcal{F} = \{\Lambda : [0, \tau] \to [0, \infty) | \Lambda \text{ be nondecreasing}, \ \Lambda(0) = 0\},$$

and let d be the $L_2(\mu)$ metric on $\mathcal{F}$. Then for $\Lambda_1, \Lambda_2 \in \mathcal{F}$,

$$\begin{aligned} d^2(\Lambda_1, \Lambda_2) &= \int |\Lambda_1(t) - \Lambda_2(t)|^2 \mathrm{d}\mu(t) \\ &= E\left[\sum_{j=1}^{K} \{\Lambda_1(T_{K,j}) - \Lambda_2(T_{K,j})\}^2\right]. \end{aligned}$$

Wellner and Zhang (2000) showed that

$$d(\hat{\Lambda}_n, \Lambda_0) \xrightarrow{a.s.} 0$$

and hence the uniform consistency of $\hat{\Lambda}_n$ can be shown by using the similar arguments to Proposition 5 of Schick and Yu (2000); that is, conditions C.1, C.2, C.4 and C.5 imply that

$$\sup_{t \in [\tau_0, \tau]} |\hat{\Lambda}_n(t) - \Lambda_0(t)| \xrightarrow{a.s.} 0.$$

Note that the uniform consistency of $\hat{\Lambda}_n$ implies for every $0 < \delta_0 < \min(L_0/2, \Lambda_0(\tau_0))$ and any $\varepsilon > 0$, there exists a positive integer N_ε such that

$$\sup_{n > N_\varepsilon} \Pr\left\{\sup_{t \in [\tau_0, \tau]} |\hat{\Lambda}_n(t) - \Lambda_0(t)| > \delta_0\right\} < \varepsilon.$$

Here, we fix δ_0. Let

$$\mathcal{F}_0 = \left\{\Lambda : \Lambda \in \mathcal{F}, \sup_{t \in [\tau_0, \tau]} |\Lambda(t) - \Lambda_0(t)| \le \delta_0\right\}.$$

Define $\hat{\Lambda}_n^*$ as

$$\hat{\Lambda}_n^* = \underset{\Lambda \in \Omega \cap \mathcal{F}_0}{\operatorname{argmax}} \left\{\sum_{i=1}^{n} \sum_{j=1}^{K_i} \left(\Delta N_i(T_{K_i,j}) \log(\Delta\Lambda(T_{K_i,j})) - \Delta\Lambda(T_{K_i,j})\right)\right\},$$

where Ω is the class of nondecreasing step functions with possible jumps only at the observation time points $\{T_{K_i,j}, j = 1, \ldots, K_i, i = 1, \ldots, n\}$. Clearly, we have

$$\sup_{n > N_\varepsilon} \Pr(\hat{\Lambda}_n \neq \hat{\Lambda}_n^*) \le \sup_{n > N_\varepsilon} \Pr\left\{\sup_{t \in [\tau_0, \tau]} |\hat{\Lambda}_n(t) - \Lambda_0(t)| > \delta_0\right\} < \varepsilon.$$

Let Δ_n^* denote the version of Δ_n obtained by replacing $\hat{\Lambda}_n$ with $\hat{\Lambda}_n^*$. Then, to prove that Δ_n converges to zero in probability, it is sufficient to show that $\Delta_n^* = o_p(1)$ since $P\{\hat{\Lambda}_n \neq \hat{\Lambda}_n^*\} < \varepsilon$. From the assumption that Z is independent of $X = (K, T, N)$, we have

$$E\left\{\left(\Delta_n^*\right)^2 | X_1, \ldots, X_n\right\} = \frac{\sigma_z^2}{n} \sum_{i=1}^{n} \left\{h(X_i, \hat{\Lambda}_n^*) - h(X_i, \Lambda_0)\right\}^2$$

where $\sigma_z^2 = E\{Z - E(Z)\}^2 < \infty$. Also from the definition of $\hat{\Lambda}_n^*$ and conditions C.1–C.3, we have

$$0 < \Lambda_0(\tau_0) - \delta_0 \le \Lambda_0(t) - \delta_0 \le \hat{\Lambda}_n^*(t) \le \Lambda_0(t) + \delta_0 \le \Lambda_0(\tau) + \delta_0 \le M + \delta_0$$

for $t \in [\tau_0, \tau]$ and

$$0 < L_0 - 2\delta_0 \le \Delta\Lambda_0(T_{K_i,j}) - 2\delta_0 \le \Delta\hat{\Lambda}_n^*(T_{K_i,j}) \le \Delta\Lambda_0(T_{K_i,j}) + 2\delta_0 \le 2M + 2\delta_0$$

for $j = 1, \ldots, K_i$, $i = 1, \ldots, n$ with probability 1. Hence, we have

$$\begin{aligned}
\left|h(X_i, \hat{\Lambda}_n^*) - h(X_i, \Lambda_0)\right| &\le c_1 \sum_{j=1}^{K_i-1} \Delta N_i(T_{K_i,j+1}) \left\{|\hat{\Lambda}_n^*(T_{K_i,j}) - \Lambda_0(T_{K_i,j})| + |\hat{\Lambda}_n^*(T_{K_i,j+1}) - \Lambda_0(T_{K_i,j+1})|\right\} \\
&\quad + c_2 \sum_{j=1}^{K_i-1} \Delta N_i(T_{K_i,j}) \left\{|\hat{\Lambda}_n^*(T_{K_i,j}) - \Lambda_0(T_{K_i,j})| + |\hat{\Lambda}_n^*(T_{K_i,j-1}) - \Lambda_0(T_{K_i,j-1})|\right\} \\
&\quad + |\hat{\Lambda}_n^*(T_{K_i,K_i}) - \Lambda_0(T_{K_i,K_i})| \\
&\quad + c_3 \sum_{j=1}^{K_i-1} \Delta N_i(T_{K_i,K_i}) \left\{|\hat{\Lambda}_n^*(T_{K_i,K_i}) - \Lambda_0(T_{K_i,K_i})| + |\hat{\Lambda}_n^*(T_{K_i,K_i-1}) - \Lambda_0(T_{K_i,K_i-1})|\right\} \\
&\le c_4 \left\{1 + \sum_{j=1}^{K_i} \Delta N_i(T_{K_i,j})\right\} \sup_{t \in [\tau_0, \tau]} |\hat{\Lambda}_n^*(t) - \lambda_0(t)| \\
&= c_4 \left\{1 + N_i(T_{K_i,K_i})\right\} \sup_{t \in [\tau_0, \tau]} |\hat{\Lambda}_n^*(t) - \lambda_0(t)|,
\end{aligned}$$

for some constants c_1, c_2, c_3 and c_4, and so

$$\frac{1}{n} \sum_{i=1}^{n} \left\{h(X_i, \hat{\Lambda}_n^*) - h(X_i, \Lambda_0)\right\}^2 \le c_4 \sup_{t \in [\tau_0, \tau]} |\hat{\Lambda}_n^*(t) - \Lambda_0(t)|^2 \frac{1}{n} \sum_{i=1}^{n} \left\{1 + N_i(T_{K_i,K_i})\right\}^2.$$

Thus, $\Delta_n^* = o_p(1)$. This completes the proof of the theorem. □

References

Hu, X.J., Sun, J., Wei, L.J., 2003. Regression parameter estimation from panel counts. Scand. J. Statist. 30, 25–43.

Park, D.-H., Sun, J., Zhao, X., 2007. A class of two-sample nonparametric tests for panel count data. Commun. Stat. - Theory Methods 36, 1611–1625.

Schick, A., Yu, Q., 2000. Consistency of the GMLE with mixed case interval-censored data. Scand. J. Statist. 27, 45–55.

Schoenfield, L.J., Lachin, J.M., et al., 1981. National cooperative gallstone study: A controlled trial of the efficiency and safety of chenodeoxycholic acid for dissolution of gallstones. Ann. Inter. Med. 95, 257–292.

Sun, J., Fang, H.B., 2003. A nonparametric test for panel count data. Biometrika 90, 199–208.

Sun, J., Kalbfleisch, J.D., 1993. The analysis of current status data on point processes. J. Amer. Statist. Assoc. 88, 1449–1454.

Sun, J., Kalbfleisch, J.D., 1995. Estimation of the mean function of counting processes based on panel data. Statist. Sinica 5, 279–289.

Sun, J., Tong, X., He, X., 2007. Regression analysis of panel count data with dependent observation times. Biometrics 63, 1053–1059.

Sun, J., Wei, L.J., 2000. Regression analysis of panel count data with covariate-dependent observation and censoring times. J. Roy. Statist. Soc., Ser. B 62, 293–302.

Thall, P.F., Lachin, J.M., 1988. Analysis of recurrent events: Nonparametric methods for random-interval count data. J. Amer. Statist. Assoc. 83, 339–347.

Wellner, J.A., Zhang, Y., 2000. Two estimators of the mean of a counting process with panel count data. Ann. Statist. 28, 779–814.

Zhang, Y., 2002. A semiparametric pseudolikelihood estimation method for panel count data. Biometrika 89, 39–48.

Zhang, Y., 2006. Nonparametric k-sample tests for panel count data. Biometrika 93, 777–790.