Published in final edited form as:

Biometrics. 2011 September ; 67(3): 730–739. doi:10.1111/j.1541-0420.2011.01557.x.

Second-Order Analysis of Semiparametric Recurrent Event Processes

Yongtao Guan*

Division of Biostatistics, Yale University, CT 06520, USA

Summary

A typical recurrent event dataset consists of an often large number of recurrent event processes, each of which contains multiple event times observed from an individual during a followup period. Such data have become increasingly available in medical and epidemiological studies. In this paper, we introduce novel procedures to conduct second-order analysis for a flexible class of semiparametric recurrent event processes. Such an analysis can provide useful information regarding the dependence structure within each recurrent event process. Specifically, we will use the proposed procedures to test whether the individual recurrent event processes are all Poisson processes and to suggest sensible alternative models for them if they are not. We apply these procedures to a well-known recurrent event dataset on chronic granulomatous disease and an epidemiological dataset on Meningococcal disease cases in Merseyside, UK to illustrate their practical value.

Keywords

Pair correlation function; recurrent event process; second-order analysis

1. Introduction

A typical recurrent event dataset consists of an often large number of recurrent event processes, each of which contains multiple event times observed from an individual during a followup period. Such data have become increasingly available in many fields especially in medical and epidemiological studies. For example, the chronic granulomatous disease (CGD) data given in Section 6.1 consist of recurrence time of pyogenic infections of 128 CGD patients, and the childhood Meningococcal disease (CMD) data given in Section 6.2 contain time of diagnosis for CMD cases from properly defined geographical units in Merseyside, UK.

In this paper we introduce the concept of second-order analysis for recurrent event data. A major interest when conducting such an analysis for recurrent event processes is to test whether the individual recurrent event processes can be viewed as Poisson processes. If the Poisson assumption is reasonable, then it would justify the validity of an analysis that is based on this assumption (e.g., Ng and Cook, 1999). A departure from the Poisson assumption often implies two alternatives, namely a clustered or a regular point pattern. For the CGD data, these alternatives in turn correspond to a more or less clustered infection pattern than what would be expected under complete randomness, and therefore are of relevance to how the disease progresses over time.

*yongtao.guan@yale.edu.

Upon a rejection of the Poisson assumption, it is often necessary to study in detail the dependence structure within the individual recurrent event processes. For the CMD data, the diagnosis time for the CMD cases is expected to be clustered due to the contagious nature of the disease. From a disease-controlling point of view, it is of interest to know the time window when an elevated disease risk should be expected following a confirmed CMD case, and how long that period will last. As we will illustrate in Section 6.2, our proposed second-order analysis procedures can be used to help answer these questions.

Second-order analysis has been widely used for spatial point pattern analysis (e.g., Diggle, 2003). We stress that there are important differences between second-order analysis of spatial point patterns and recurrent event data. For the former, it is often necessary to fully specify the mean structure of the process. Furthermore, both the size of the observation window and the number of events observed on it need to be large. For the latter, it is common to assume a semiparametric regression model for the mean structure of the recurrent event processes, and the number of events per process can be extremely small. Specifically, for the CGD data, the total number of events is 76 for 128 CGD patients; for the CMD data, the total number of CMD cases is 864 from 182 reporting units. It is challenging to extract dependence information from individually sparse and also not fully parametric recurrent event processes.

Overcoming the above difficulties, we develop novel statistical methods to conduct second-order analysis for recurrent event processes. Specifically, we propose consistent estimators for two second-order integrals, based on which useful second-order functions can be derived. Our main theoretical results are based on a flexible asymptotic framework that combines features from both spatial point processes and (commonly studied) recurrent event processes. Consequently, the proposed methods built upon these results can be applied to real data arising from a broad range of settings. As an important application, we develop a nonparametric graphical procedure to test whether the individual recurrent event processes are Poisson processes.

We organize the rest of the paper as follows. In Section 2 we give necessary background and set up the notation. In Section 3 we develop the proposed estimators for the second-order integrals and study their theoretical properties. In Section 4 we introduce the graphical procedure to test the Poisson assumption. We conduct a simulation study in Section 5 to assess the performance of the proposed estimators and apply them to two real examples in Section 6. We conclude this article with a discussion in Section 7. Proofs for our theoretical results and some additional simulation results are given in the web Supplementary Materials.

2. Background and Notation

2.1 Notation

Consider n independent recurrent event processes $N_1, \dots, N_n$. Let $dN_i(t)$ denote the number of events from the i th process observed over the time interval $[t, t+dt)$, where $dt > 0$. Define the first- and second-order rate functions of N_i as

$$\lambda(t; N_i) = \lim_{dt \rightarrow 0} \frac{E[dN_i(t)]}{dt},$$

$$\lambda_2(t_1, t_2; N_i) = \lim_{dt_1, dt_2 \rightarrow 0} \frac{E[dN_i(t_1)dN_i(t_2)]}{dt_1 dt_2}.$$

Intuitively, $\lambda(t; N_i)dt$ and $\lambda_2(t_1, t_2; N_i)dt_1 dt_2$ are the approximate probabilities to have one event in $[t, t + dt)$ and in each of $[t_1, t_1 + dt_1)$ and $[t_2, t_2 + dt_2)$, respectively. We will refer to

$\lambda(t; N_i)$ as the rate function (Pepe and Cai, 1993; Lawless and Nadeau, 1995) and assume that it takes the following semiparametric form:

$$\lambda(t; N_i) = \lambda_0(t) \rho(\beta, \mathbf{X}_i), \quad (1)$$

where $\lambda_0(t)$ is an unspecified baseline rate function and $\rho(\beta, \mathbf{X}_i)$ is a parametric function depending on both some predictors $\mathbf{X}_i$ and unknown parameters β . A consistent estimator for β can be obtained by applying the estimating equation based approaches given in Lawless and Nadeau (1995) and Lin et al. (2000) without having to specify $\lambda_0(t)$; these estimating equations have the same form as the score equations of Andersen and Gill (1982) except that they are expressed in terms of the rate function. We assume that $\mathbf{X}_i$ does not change with t for ease of theoretical derivation. However, our methods presented in Section 3 can be easily generalized to the case when $\mathbf{X}_i$ is time varying.

Given the first- and second-order rate functions, we define the pair correlation function

$$g(t_1, t_2; N_i) = \frac{\lambda_2(t_1, t_2; N_i)}{\lambda(t_1; N_i) \lambda(t_2; N_i)}.$$

The pair correlation function is very useful for second-order analysis of point processes (Møller and Waagepetersen, 2004). If the point process N_i is Poisson (conditional on the covariates $\mathbf{X}_i$), then $g(t_1, t_2; N_i) = 1$. Otherwise, $g(t_1, t_2; N_i)$ is larger than 1 if N_i shows attraction at the time locations (t_1, t_2) and is less than 1 if it exhibits inhibition at those locations.

2.2 Pair correlation function for Cox processes

Cox (1955) introduced the so-called doubly stochastic Poisson process, which generalizes a Poisson process by defining its rate function as a realization of a random field. This type of process is commonly referred to as Cox process. Henderson et al. (2000) discussed a very general class of Cox processes in the context of modeling recurrent event data. Here we consider one such example.

Specifically, let u_i be an unobserved normal random variable with mean zero and variance σ^2 , and let $Z_i(t)$ be a stationary Gaussian process that is independent of u_i and has mean zero and an isotropic covariance function $R(|t_1 - t_2|) = \text{cov}\{Z_i(t_1), Z_i(t_2)\}$. Assume that conditional on u_i and $Z_i(t)$, N_i is a Poisson process with rate function

$$\lambda\{t; N_i | u_i, Z_i(t)\} = \lambda_0(t) \exp\{u_i + \mathbf{X}_i^T \beta + \eta Z_i(t)\}, \quad (2)$$

where β and η are some real constants and $\mathbf{X}_i$ is a vector of known covariates. By definition, N_i is a Cox process. For the CGD data given in Section 6.1, u_i in (2) accounts for the variability across subjects due to non-treatment factors, e.g., one's overall health conditions, $\mathbf{X}_i$ indicates the treatment assignment, and $Z_i(t)$ is a latent process to explain the local temporal variations. In the special case that $\eta = 0$, (2) becomes a random effect (frailty) model, which has been widely used in literature, e.g., see Ng and Cook (1999) and the references therein.

Conditional on u_i , N_i is a log-Gaussian Cox process (Müller et al., 1998) and its pair correlation function is given by

$$g(t_1, t_2; N_i | u_i) = \exp\{\eta^2 R(|t_1 - t_2|)\}.$$

Because u_i is not observed, it is more sensible to consider the pair correlation function without conditioning on u_i . This is given by

$$g(t_1, t_2; N_i) = \exp\{\sigma^2 + \eta^2 R(|t_1 - t_2|)\}.$$

Interestingly, $g(t_1, t_2; N_i)$ is simply $g(t_1, t_2; N_i | u_i)$ multiplied by a constant and $g(t_1, t_2; N_i) = g(t_1, t_2; N_i | u_i)$ if and only if $\sigma^2 = 0$, i.e., if the random effect term u_i does not apply. Such a relationship is generally true even if u_i and $Z_i(t)$ are nonnormal. Note also that $g(t_1, t_2; N_i | u_i) = 1$, which implies $g(t_1, t_2; N_i)$ being a constant, for all $t_1 \neq t_2$ if $\eta = 0$. In this case (2) leads to a Poisson process when conditional on u_i . Moreover, $g(t_1, t_2; N_i | u_i)$ converges to one and $g(t_1, t_2; N_i)$ converges to a constant as $R(|t_1 - t_2|)$ goes to zero.

In practice, it is often difficult to estimate $g(t_1, t_2; N_i | u_i)$ consistently due to the unobserved random effect u_i . We will introduce in Section 2.3 two second-order integrals based on which a consistent nonparametric estimator for $g(t_1, t_2; N_i)$ can be derived.

Our proposed second-order analysis is useful for model building and model diagnostics for Cox processes. Specifically, the proposed test for the Poisson assumption can be used to check whether η in (2) is equal to zero, i.e., whether (2) can be reduced to a simpler random effect model, whereas our proposed estimator for $g(t_1, t_2; N_i)$ (see Section 2.3) can be used to assess the goodness-of-fit of a fitted model such as (2) by comparing it against its theoretical counterpart under the fitted model.

2.3 The proposed second-order integrals

Let $f(t_1, t_2)$ be a known, nonnegative function satisfying that $f(t_1, t_2) = f(t_2, t_1)$. Our main interest is to estimate the following two quantities:

$$A = \sum_{i=1}^n \int_0^{\tau_i} \int_0^{\tau_i} \lambda_0(t_1) \lambda_0(t_2) f(t_1, t_2) g(t_1, t_2; N_i) dt_1 dt_2, \quad (3)$$

$$B = \sum_{i=1}^n \int_0^{\tau_i} \int_0^{\tau_i} \lambda_0(t_1) \lambda_0(t_2) f(t_1, t_2) dt_1 dt_2, \quad (4)$$

where $[0, \tau_i]$: $i = 1, \dots, n$, are the observation intervals for the n recurrent event processes and τ_i 's are assumed to be independent of N_i 's.

Integrals of the forms (3) and (4) play an essential role in second-order analysis of recurrent event processes. To see this point, consider

$$f(t_1, t_2) = \frac{1}{h} \cdot K\left(\frac{|t_1 - t_2| - r}{h}\right), \quad (5)$$

where $K(\cdot)$ is a bounded kernel function with a compact support over $[-1, 1]$. Assume that $g(t_1, t_2; N_i) = g(|t_1 - t_2|)$ for $i = 1, \dots, n$, i.e., the pair correlation functions of the n recurrent event processes are identical and also isotropic. If $g(\cdot)$ is continuous and h is small, then

$$\begin{aligned} A &= \sum_{i=1}^n \int_0^{\tau_i} \int_0^{\tau_i} \lambda_0(t_1) \lambda_0(t_2) \frac{1}{h} \cdot K\left(\frac{|t_1 - t_2| - r}{h}\right) g(|t_1 - t_2|) dt_1 dt_2 \\ &\approx \sum_{i=1}^n \int_0^{\tau_i} \int_0^{\tau_i} \lambda_0(t_1) \lambda_0(t_2) \frac{1}{h} \cdot K\left(\frac{|t_1 - t_2| - r}{h}\right) g(r) dt_1 dt_2 \\ &= g(r)B. \end{aligned}$$

Thus, A/B forms an approximately unbiased estimator for the pair correlation function.

The two quantities A and B defined through (5) depend on a user-specified bandwidth. This can be avoided by considering

$$f(t_1, t_2) = I(|t_1 - t_2| \leq r), \quad (6)$$

which leads to the cumulative quantities

$$\begin{aligned} A &= \sum_{i=1}^n \int_0^{\tau_i} \int_0^{\tau_i} \lambda_0(t_1) \lambda_0(t_2) I(|t_1 - t_2| \leq r) g(|t_1 - t_2|) dt_1 dt_2, \\ B &= \sum_{i=1}^n \int_0^{\tau_i} \int_0^{\tau_i} \lambda_0(t_1) \lambda_0(t_2) I(|t_1 - t_2| \leq r) dt_1 dt_2. \end{aligned}$$

If N_i is Poisson conditional on the covariates $\mathbf{X}_i$ and some random effect u_i , then $g(r) = c$ for some $c > 0$ and for all $r > 0$, where $c = 1$ if N_i depends only on $\mathbf{X}_i$. Then,

$$A = c \sum_{i=1}^n \int_0^{\tau_i} \int_0^{\tau_i} \lambda_0(t_1) \lambda_0(t_2) I(|t_1 - t_2| \leq r) dt_1 dt_2 = cB.$$

Thus, A/B is a constant for all $r > 0$, under the null hypothesis that all individual recurrent event processes are Poisson processes. If an estimate for A/B can be obtained, we can then plot it against r so as to check for departures from the null hypothesis.

3. Estimation of Second-Order Integrals

In this section we develop consistent estimators for the two second-order integrals defined in (3) and (4). Because the function $f(\cdot)$ used to define these integrals may take very different forms, we wish to make our estimators as general as possible so they can be applied in a broad range of settings. For ease of presentation but without losing generality, we assume that the ending times τ_i 's are all distinct and more specifically, $0 = \tau_0 < \tau_1 < \tau_2 < \dots < \tau_n$. Our proposed estimators can be easily extended to the situation with tied ending times and we will give the detailed expressions in Appendix A.

3.1 The proposed estimators

Estimator for A—For any given process N_i , define

$$A_i = \int_0^{\tau_i} \int_0^{\tau_i} \lambda_0(t_1) \lambda_0(t_2) f(t_1, t_2) g(t_1, t_2; N_i) dt_1 dt_2.$$

We can then rewrite A as

$$A = \sum_{i=1}^n A_i.$$

Assume that a consistent estimator $\hat{\beta}$ is available for β . It then follows from Campbell's Theorem (Daley and Vere-Jones, 2008, Section 13.1) that

$$\int_0^{\tau_i} \int_0^{\tau_i} \frac{f(t_1, t_2) I(t_1 \neq t_2)}{\rho(\hat{\beta}, \mathbf{X}_i)^2} dN_i(t_1) dN_i(t_2)$$

is an approximately unbiased estimator for A_i . Note that we consider only one recurrent event process at a time, because distinct recurrent event processes are assumed to be independent and consequently contain no information regarding the pair correlation functions in A_i ; $i = 1, \dots, n$. This then leads to the following estimator for A :

$$\hat{A} = \sum_{i=1}^n \int_0^{\tau_i} \int_0^{\tau_i} \frac{f(t_1, t_2) I(t_1 \neq t_2)}{\rho(\hat{\beta}, \mathbf{X}_i)^2} dN_i(t_1) dN_i(t_2). \quad (7)$$

Estimator for B—For any two positive integers $l, m \leq n$, define

$$B_{lm} = \int_{\tau_{l-1}}^{\tau_l} \int_{\tau_{m-1}}^{\tau_m} \lambda_0(t_1) \lambda_0(t_2) f(t_1, t_2) dt_1 dt_2.$$

Because the ending times are assumed to be distinct, we can rewrite B as

$$B = \sum_{i=1}^n \sum_{l=1}^i \sum_{m=1}^i B_{lm} = \sum_{l=1}^n \sum_{m=1}^n \{n - \max(l, m) + 1\} B_{lm}.$$

Then, for any two distinct recurrent event processes N_j and N_k satisfying that $j \geq l$ and $k \geq m$, i.e., $\tau_j \geq \tau_l$ and $\tau_k \geq \tau_m$,

$$\int_{\tau_{l-1}}^{\tau_l} \int_{\tau_{m-1}}^{\tau_m} \frac{f(t_1, t_2)}{\rho(\hat{\beta}, \mathbf{X}_j) \rho(\hat{\beta}, \mathbf{X}_k)} dN_j(t_1) dN_k(t_2)$$

is an approximately unbiased estimator for B_{lm} . Let $C_{lm} = \{n - \max(l, m) + 1\} \{n - \min(l, m)\}$ be the total number of such distinct pairs of recurrent event processes. Define

$$\widehat{B}_{lm} = \frac{1}{C_{lm}} \sum_{j=l}^n \sum_{\substack{k=m \\ k \neq j}}^n \int_{\tau_{l-1}}^{\tau_l} \int_{\tau_{m-1}}^{\tau_m} \frac{f(t_1, t_2)}{\rho(\widehat{\beta}, \mathbf{X}_j) \rho(\widehat{\beta}, \mathbf{X}_k)} dN_j(t_1) dN_k(t_2).$$

The above leads to the following estimator for B :

$$\widehat{B} = \sum_{l=1}^n \sum_{m=1}^n \frac{1}{n - \min(l, m)} \sum_{j=l}^n \sum_{\substack{k=m \\ k \neq j}}^n \int_{\tau_{l-1}}^{\tau_l} \int_{\tau_{m-1}}^{\tau_m} \frac{f(t_1, t_2)}{\rho(\widehat{\beta}, \mathbf{X}_j) \rho(\widehat{\beta}, \mathbf{X}_k)} dN_j(t_1) dN_k(t_2). \quad (8)$$

Note that unlike in the case of estimating A , we estimate B based on all possible pairs of distinct recurrent event processes.

A major novelty of this paper is that the proposed estimators do not require estimating the unknown baseline rate function $\lambda_0(\cdot)$. This is desirable because the estimation of $\lambda_0(\cdot)$ can be problematic especially when the number of processes is small.

3.2 Theoretical justification

Let $\alpha < 1$ be a positive real value. Given the conditions in web Appendix A, the following two theorems give the mean squared errors of $\widehat{A}$ and $\widehat{B}$, respectively.

Theorem 1—Assume that conditions (A.1)–(A.4), (A.6)–(A.9) in web Appendix A are true. Then

$$(\widehat{A} - A)^2 = O_p \left\{ \left(\sum_{i=1}^n \tau_i \right)^{1+\alpha} \right\} + o_p \left\{ \left(\sum_{i=1}^n \tau_i \right)^2 \right\}.$$

Proof: See web Appendix B.

Theorem 2—Assume that conditions (A.1)–(A.8) in web Appendix A are true. Then

$$(\widehat{B} - B)^2 = O_p \left\{ (n \log n) \tau_n + \tau_n \left(\sum_{i=1}^n \tau_i \right)^\alpha \right\} + o_p \left\{ \left(\sum_{i=1}^n \tau_i \right)^2 \right\}.$$

Proof: See web Appendix C.

Assume that both A and B are of order $\sum_{i=1}^n \tau_i$. Then Theorems 1 and 2 suggest that both $\widehat{A}/A$ and $\widehat{B}/B$ converge to one in probability, if

$$\sum_{i=1}^n \tau_i \rightarrow \infty \text{ and } \frac{(n \log n) \tau_n}{\left(\sum_{i=1}^n \tau_i \right)^2} \rightarrow 0.$$

The above result has important practical implications. Specifically, it says that if the lengths of the observation intervals (i.e., τ_i 's) are small but the number of independent recurrent event processes (i.e., n) is large, then the two integrals A and B can still be estimated accurately. This is also true if τ_i 's are large, even though n may be small. In the most extreme case, consistent estimators for A and B can still be obtained with $n = 2$. Thus, our asymptotic framework is a hybrid of the commonly used asymptotic framework in spatial statistics including spatial point processes, which requires the size of the observation window to increase to infinity (Cressie, 1993), and that in recurrent event processes, which assumes that the number of independent processes n increases to infinity (Lin et al., 2000). Such an asymptotic framework is rarely considered in practice.

4. Graphical Assessment of Poisson Assumption

4.1 The null hypothesis

Let $\gamma_i > 0$ be an unobserved random effect for N_i . We are interested in testing the null hypothesis that conditional on γ_i and $\mathbf{X}_i$, N_i is a Poisson process and has the rate function

$$\lambda(t; N_i | u_i) = \lambda_0(t) \gamma_i \rho(\beta, \mathbf{X}_i) \equiv \lambda_0(t) \theta_i. \quad (9)$$

In general, θ_i can be any positive function that involves γ_i and $\mathbf{X}_i$. We in fact do not need to know the exact form of this function in order to conduct the test.

In terms of model (2), $\gamma_i = \exp(u_i)$ and $\rho(\beta, \mathbf{X}_i) = \exp(\mathbf{X}_i^T \beta)$, and testing (9) is equivalent to testing whether $\eta = 0$. It is worth noting that when $\eta \neq 0$, model estimation in particular maximum likelihood estimation can be quite challenging, because $Z_i(t)$ is an infinite-dimension process (Møller and Waagepetersen, 2004). In contrast, model (2) when $\eta = 0$ is simply a random frailty model and can therefore be estimated with relative ease. This example shows the potential benefits for conducting the Poisson test.

We will perform the test based on the second-order statistics defined in the last section. To be specific, let $\hat{G} = \hat{A}/\hat{B}$, where $\hat{A}$ and $\hat{B}$ are defined in (7) and (8) with $f(t_1, t_2) = I(|t_1 - t_2| \leq r)$. We rewrite $\hat{G}$ as $\hat{G}(r)$ to make explicit its dependence on r . In light of the consistency results given in Theorems 1 and 2, we expect $\hat{G}(r) \approx c$ for some unknown constant c under the null hypothesis. However, if N_i 's are clustered (regular), we then expect $\hat{G}(r)$ to take larger (smaller) values at some small lag values r .

4.2 The proposed procedure

We propose to generate permuted replicates, say $\hat{G}_l(r)$: $l = 1, \dots, L$, for some positive integer L . The original statistics $\hat{G}(r)$ can then be plotted together with permutation envelope obtained from the replicates $\hat{G}_l(r)$ in order to assess evidence of departure from the null hypothesis. Specifically, we will conclude that the processes are clustered if $\hat{G}(r)$ falls above the permutation envelope and are regular if it falls below.

To that end, we first note that the union of all recurrent event processes, which we refer to as the combined process, is also a Poisson process but has a different rate function

$\lambda(t) = \lambda_0(t) \sum_{i=1}^n \theta_i$. The original n recurrent event processes can be viewed as a result of randomly labeling events in the combined process into n processes. To better illustrate this point, assume that the ending times satisfy $0 = \tau_0 < \tau_1 < \tau_2 < \dots < \tau_n$. Because of (9), an event $t \in [\tau_{j-1}, \tau_j)$ is assigned to N_i with probability $\theta_i / \sum_{k=j}^n \theta_k$ if $i \geq j$ and with probability zero if $i < j$. Note that these probabilities are the same for all events in $[\tau_{j-1}, \tau_j)$.

Let m_{ij} denote the number of events in $[\tau_{j-1}, \tau_j)$ from N_i and define $m_i = \sum_{j=1}^n m_{ij}$. An obvious approach to generate permuted replicates for the original processes is to randomly select m_{ij} events from all events in $[\tau_{j-1}, \tau_j)$ with equal probability for all possible i and j . This would then generate n new processes that have the same distribution as the original ones, given that they share the same m_{ij} values. However, if the subintervals $[\tau_{j-1}, \tau_j)$ are all small, e.g., $\tau_j = j/n$, then the resulting permuted replicates may follow the micro-structure of the original processes too closely so that $\hat{G}_l(r)$: $l = 1, \dots, L$, are all very similar to $\hat{G}(r)$; this may lead to a test with little power.

We instead propose to generate permuted replicates conditional on the total number of events (i.e., m_i) in each process. The rationale is as follows. First, note that for an arbitrary event $t \in [0, \tau_1)$ in the combined process, the probability that it comes from N_1 is $\theta_1 / \sum_{i=1}^n \theta_i$. In other words, all events observed in $[0, \tau_1)$ have the same probability for being from N_1 . We may thus select a simple random sample of m_1 events from all events in $[0, \tau_1)$ as a replicate for N_1 . The remaining $\sum_{i=2}^n m_i$ events will then form a replicate for $\cup_{i=2}^n N_i$. We next generate a replicate for N_2 . This can be done by selecting m_2 random samples from the remaining events in $[0, \tau_2)$, because every such event has the same probability $\theta_2 / \sum_{i=2}^n \theta_i$ to be from N_2 . The process can be repeated until the last m_n events are selected as a replicate for N_n . Under the null hypothesis and conditional on the total number of events in each process, the resulting n new recurrent event processes have the same distribution as the original ones.

It is worth noting that the proposed permutation procedure is independent of the assumed parametric model in terms of $\mathbf{X}_i$. As such, it is valid even if the fitted parametric part of the model is wrong. Moreover, because $\mathbf{X}_i$'s are fixed, the estimate for β typically depends only on the total number of events in each process; the same β will therefore be used given that m_i 's are fixed. Our proposed testing procedure can be summarized in three steps:

- Step 1** Fit the multiplicative rate function model (1) to data and calculate $\hat{G}(r)$.
- Step 2** Set $m_0 = 0$. Randomly select m_j events from all available events in $[0, \tau_j)$, assign the selected m_j events as a permuted sample for the j th process, and remove them from the available events. Repeat this for $j = 1, \dots, n$ to obtain a permuted sample for all processes; calculate $\hat{G}_l(r)$ based on the l th permuted sample, for $l = 1, \dots, L$.
- Step 3** Plot $\hat{G}(r)$ and the permutation envelope based on $\hat{G}_l(r)$: $l = 1, \dots, L$.

4.3 Comparison with existing methods

There is a rich literature available on testing for the Poisson assumption, often given under the names of testing for overdispersion or homogeneity; for example, see Dean (1992), Lambert and Roeder (1995), and Ng and Cook (1999). For most of the existing methods, a rejection of the Poisson assumption is interpreted as the presence of unobserved heterogeneity between subjects, i.e., a larger-than-zero variance for the random effect term γ_i given in (9). However, this can be overly simplistic, because the presence of within-process correlation, as is the case with $\eta \neq 0$ in model (2), could also lead to overdispersion in the total counts of events from each process and hence to a rejection of the Poisson assumption. Our proposed testing procedure can be a valuable supplementary to the existing approaches by testing the within-process correlation. Moreover, most available tests are expected to be sensitive to the misspecification of the parametric part of model (1). However, our test can maintain the correct test size as long as the more general form (9) is valid.

5. Simulation

We have conducted a simulation study to evaluate the finite sample performance of the proposed estimators. Specifically, we have generated n independent recurrent event processes from either Poisson or Poisson cluster processes over the time intervals $[0; a\tau_i^*]$: $i = 1, \dots, n$, where $(n, a) = (128, 1), (128, 4)$ or $(512, 1)$ and τ_i^* 's are the ending times in the CGD data. Specifically, the minimum and maximum of the τ_i^* 's are equal to 91 and 439, respectively.

For the Poisson processes, we use the rate function $\lambda_i(t) = 2.5\kappa \exp(\text{trt}_i\beta)$ for the i th process, where $\kappa = 0.001$, trt_i is a binary treatment variable for the i th process ($=1$ for treatment and $=0$ for control) and $\beta = -1.0971$ as in the CGD data (Lin et al., 2000). For the Poisson cluster processes, we first simulate a Poisson process with intensity $\kappa = 0.001$ as the parent process, and then generate a Poisson random variable with mean equal to $2.5 \exp(\text{trt}_i\beta)$ for the number of offspring that each parent generates. The position of each offspring relative to its parent is then determined by a Gaussian random variable with standard deviation $\omega = 10$ or 20 . Note that a smaller ω value leads to a more clustered process. The resulting rate function for the i th process is the same as that in the Poisson case. For $n = 128$ and $a = 1$, the expected numbers of events in all cases are 16 and 46 for the treatment and control groups, respectively. Note that the total number of simulated events in each simulation is slightly smaller than, but nevertheless comparable to, that for the CGD data given in Section 6.1 ($=76$). For $n = 512$ and $a = 4$, the expected numbers of events become 64 and 184.

For each simulated set of n realizations, we calculate $\hat{A}$, $\hat{B}$ and $\hat{G}$ with $f(t_1, t_2) = I(|t_1 - t_2| \leq r)$, where $r = 20, 40, 80$. Table 1 shows the empirical bias and standard deviation in each case from 1000 simulations. Both the bias and standard deviation have been divided by the target parameter for a fair comparison. The biases for all three estimators are typically small especially when $n = 512$ and $a = 4$, but appear to be larger for $\hat{G}$ and when $r = 20$. The larger bias for $\hat{G}$ is expected given that $\hat{G}$ is defined as the ratio of $\hat{A}$ and $\hat{B}$; the larger bias when $r = 20$ is probably due to the fact that fewer events are separated by a distance less than or equal to 20. The standard deviations are much smaller when $n = 512$ and/or $a = 4$ than when $(n, a) = (128, 1)$. Combining this result with the fact that the biases are also smaller in these cases, we can conclude that the accuracy of the proposed estimators improves as n and a increase. This provides support for our theoretical results given in Theorems 1 and 2. In general, the standard deviations become smaller when r increases, especially when $(n, a) = (128, 1)$. Moreover, the standard deviations typically decrease for $\hat{A}$ and $\hat{G}$ but increase for $\hat{B}$ when the clustering becomes stronger.

We have also evaluated the performance of our proposed test procedure given in Section 4.2. The detailed results are presented in we Appendix D. Briefly speaking, our test procedure has high power in detecting clustering with a remarkably small sample size. The power increases as the sample size increases especially as a increases, but decreases as ω increases, i.e., when the clustering strength decreases. The maximum power is typically achieved at $r \approx 3\omega$. When the lag is set to be too large or too small, the power can be greatly reduced. For the Poisson cluster processes, the dependence beyond a lag distance 4ω is often ignorable; hence it is desirable to base the test on a lag value that is slightly smaller than the dependence range ($\approx 4\omega$ in this case). In practice, we can roughly estimate the dependence range based on a plot of the empirical pair correlation function.

The size of our test is not affected by whether the parametric part of the multiplicative model (1) is correctly specified or not, and does not appear to be very sensitive to modest departures from the overall multiplicative structure either. In terms of power, we have surprisingly found that an incorrect choice of $\beta = 0$ has led to a better power in our

simulation; hence, it may be sufficient to directly test for clustering without first modeling the potential heterogeneity across subjects. Finally, we note that although our test is effective in detecting clustering, it generally has a low power in detecting regularity especially when the sample size is relatively small. This is because the lower permutation envelope is bounded by zero.

6. Applications to Real Data

6.1 Application to chronic granulomatous disease data

Chronic granulomatous disease is a group of inherited rare disorders of the immune function that are characterized by recurrent pyogenic infections. The data being considered here came from a placebo controlled trial of gamma interferon in CGD and consist of recurring pyogenic infection time for 128 patients. We define the rate function as $\lambda_i(t) = \lambda_0(t) \exp(\text{trt}_i \beta)$, where $\hat{\beta} = -1.0971$ is obtained by the estimating equation approach given in Lin et al. (2000).

We apply our proposed graphical procedures in Section 4 to assess whether an individual's recurring pyogenic infection time is Poisson. To that end, we calculate the function $\hat{G}(r)$ and its permuted replicates $\hat{G}_l(r)$: $l = 1, \dots, 49$, for $0 < r \leq 40$. Figure 1 plots $\hat{G}(r)$ and the associated permutation envelope defined as the maximum and minimum of $\hat{G}_l(r)$: $l = 1, \dots, 49$ for each given r value. Under the Poisson assumption, the use of 49 permuted replicates implies that $\hat{G}(r)$ has a 2/50 chance to fall outside of the permutation envelope for each given r . Because $\hat{G}(r)$ is completely included in the permutation envelope, we conclude no evidence against the null hypothesis. Note that the permutation envelope is quite wide especially at the beginning. This is a result of the very small number ($=76$) of events being observed. Nevertheless, $\hat{G}(r)$ appears to be quite flat, indicating a great compatibility with the Poisson assumption. This result justifies the use of a random effect model (Ng and Cook, 1999), as opposed to a more complex model.

To further illustrate the power of our test procedure, we apply it to 128 realizations from the Poisson cluster processes considered in Section 5 with $\kappa = 0.001$ and $\omega = 10$. The total number of simulated events is comparable to that in the real data example (see the two histograms in Figure 1). Different from the real data example, however, $\hat{G}(r)$ now falls completely above the permutation envelope, indicating a strong evidence for clustering (see Figure 1). This is quite remarkable given such a small sample size. Note also that the shapes of the permutation envelopes for the real data and the simulated data are very similar, in that both are very wide at the beginning.

6.2 Application to childhood meningococcal disease data

Meningococcal disease is a severe bacterial infection of the bloodstream or meninges that is caused by the meningococcus germ. Infants and children are particularly vulnerable to being infected. The data being considered here consist of time of diagnosis for 864 meningococcal disease patients of age 0–15 living in Merseyside, UK during the period January 1, 1981 to December 31, 2007. For each case, the full unit post-code of residence is also available. Each post-code is converted to a grid reference using the online tool GeoConvert, developed by the Census Dissemination Unit at the University of Manchester, UK (<http://cdu.mimas.ac.uk>).

There are 182 Middle Level Super Output Areas (MSOAs) in Merseyside, belonging to five local authorities, namely Knowsley, Liverpool, St Helens, Sefton and Wirral, that make up the region. Each MSA contains at least 5000 residents and 2000 households and will be treated as a reporting unit in our analysis. This in turn leads to 182 recurrent event processes.

For comparison, we also use the five local authorities as the reporting units; this corresponds to the case of n being small in our asymptotic framework given in Section 3.

Our main interests of the analysis are to assess whether the reported time of diagnosis was random (i.e., Poisson) in each reporting unit and to understand the dependence if it was not. Clustering is suspected given the contagious nature of the disease. To formally verify this, we calculate the function $\hat{G}(r)$ and its permuted replicates $\hat{G}_l(r)$: $l = 1, \dots, 49$, for $0 < r \leq 40$ days. The first plot of Figure 2 shows the resulting $\hat{G}(r)$ and the associated permutation envelope. Because $\hat{G}(r)$ exceeds the upper permutation envelope at multiple locations for $r < 20$, we conclude that there is a strong evidence of clustering in the reported time of diagnosis. Our analysis is novel in the sense that we require only (9) but do not need to assume any specific model for the reporting patterns in order to perform the test.

Figure 3 plots the estimated pair correlation function, $\hat{g}(r)$ for $0 < r \leq 60$ days using bandwidth $h = 19$. Note that $\hat{g}(r)$ decreases with r for $r \leq 25$ and becomes relatively flat for $r > 25$. The former again suggests clustering in the time of diagnosis within a given reporting unit, whereas the latter suggests that the correlation diminishes roughly after $r = 25$. Combining these facts together, we conclude that a meningococcal disease case will more likely lead to another case in the same reporting unit within a short period (say ≤ 10 days) after its diagnosis, but is expected to have no significant effect beyond 25 days. Such a conclusion can not be reached from the plot of $\hat{G}(r)$. Note also that $\hat{g}(r)$ is consistently larger than one, which indicates a strong between-unit variation. As a result, variables specific to a reporting unit such as socioeconomic variables should be considered so as to explain this variability.

The second plot of Figure 2 shows the function $\hat{G}(r)$ and its associated permutation envelope when the local authorities are used as the reporting units. The plot suggests no evidence of clustering but a strong evidence for between-unit variation. The former is due to the facts that temporal clustering is expected to be significant only at relatively small spatial scales and that the use of larger reporting units tends to dilute the evidence of clustering. The latter is a result of the high heterogeneity in the social composition and degree of urbanization among the five local authorities, which could affect the risk of developing the meningococcal disease (Diggle et al., 2010). The purpose of this analysis is simply to show what one should expect when the between-process variation dominates over the within-process clustering.

7. Discussion

We have introduced the new concept of second-order analysis for semiparametric recurrent event processes. We have applied the proposed methods to both simulated and two real data examples, and have shown through these applications that they can provide valuable information on the dependence structure of individual recurrent event process. Specifically, for the CGD example, we have concluded that an individual's recurrent pyogenic infection time is compatible with a Poisson process. This justifies the use of the popular random effect models for this dataset. For the CMD example, we have formally confirmed the existence of clustering in the diagnosis time of meningococcal disease cases, and revealed that the clustering is highly significant with a lag distance less than or equal to 10 days but diminishes after a lag distance of 25 days. We do not need to model the potentially heterogeneous temporal trend for either example, which is a great strength of our proposed methods.

Our main theoretical results are derived based on a flexible asymptotic framework that combines main features of those commonly used in both spatial statistics and recurrent event

processes. Consequently, the proposed methods based on these theoretical results can be used in many different settings, including processes that are either individually sparse but have many replicates or individually rich of events but have few replicates.

Our theoretical results are currently limited to consistency of the proposed estimators for the two second-order integrals defined in Section 2.3. We conjecture that the joint distribution of these estimators will be asymptotically normal under suitable conditions. However, a proof of this result will require advanced statistical theories on either empirical processes (e.g., Lin et al., 2000) or mixing conditions (e.g., Guan and Sherman, 2007) or both, depending on the exact type of asymptotics to be considered. Our preliminary results also show that the asymptotic covariance involves a large number of complex integrals in terms of the rate functions up to the forth order. It is nontrivial to develop a consistent yet computationally realistic estimator for the asymptotic covariance. These problems are important but are beyond the scope of this paper. To test the Poisson assumption, we prefer the proposed permutation test over a test based on asymptotic theories, because it is exact and also is not affected the validity of the fitted rate function model.

8. Supplementary Materials

Theoretical results and additional simulation results, referenced in Sections 1, 3.2 and 5, are available under the Paper Information link at the *Biometrics* website <http://www.biometrics.tibs.org>.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by NSF grant DMS-0845368 and by NIH grant 1R01DA029081. The CMD data were collected by Alistair Thomson, Omnia Marzouk, Andrew Riordan, Paul Baines, Enitan Carrol, Fauzia Paize, Scott Hackett and Niten Makwana, Alder Hey Children's Hospital, Merseyside. The authors thank the reviewers for their constructive comments and suggestions, Peter Diggle for helpful discussions, and Michelle Stanton for organizing the CMD data.

References

- Andersen PK, Gill RD. Cox's regression model for counting processes: A large sample study (Com: P1121–1124). *The Annals of Statistics*. 1982; 10:1100–1120.
- Cox DR. Some statistical models related with series of events. *Journal of the Royal Statistical Society, Series B*. 1955; 17:129–164.
- Cressie, NAC. *Statistics for Spatial Data*. 2. New York: John Wiley & Sons; 1993.
- Daley, DJ.; Vere-Jones, D. *General Theory and Structure*. 2. Vol. II. New York: Springer-Verlag; 2008. *An Introduction to the Theory of Point Processes*.
- Dean CB. Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association*. 1992; 87:451–457.
- Diggle, PJ. *Statistical Analysis of Spatial Point Patterns*. 2. London: Edward Arnold; 2003.
- Diggle PJ, Guan Y, Hart AC, Paize F, Stanton M. Estimating individual-level risk in spatial epidemiology using spatially aggregated information on the population at risk. *Journal of the American Statistical Association*. 2010 To appear.
- Guan Y, Sherman M. On least squares fitting for stationary spatial point processes. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*. 2007; 69(1):31–49.
- Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostatistics (Oxford)*. 2000; 1(4):465–480.

- Lambert D, Roeder K. Overdispersion diagnostics for generalized linear models. *Journal of the American Statistical Association*. 1995; 90:1225–1236.
- Lawless JF, Nadeau C. Some simple robust methods for the analysis of recurrent events. *Technometrics*. 1995; 37:158–168.
- Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, Series B, Methodological*. 2000; 62(4): 711–730.
- Møller J, Syversveen AR, Waagepetersen RP. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*. 1998; 25(3):451–482.
- Møller, J.; Waagepetersen, RP. *Statistical Inference and Simulation for Spatial Point Processes*. New York: Chapman & Hall; 2004.
- Ng ETM, Cook RJ. Adjusted score tests of homogeneity for Poisson processes. *Journal of the American Statistical Association*. 1999; 94:308–319.
- Pepe MS, Cai J. Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association*. 1993; 88:811–820.

Appendix A: Expressions with Tied Ending Time

The expression of $\hat{A}$ remains unchanged so we consider only $\hat{B}$. For ease of presentation, write $\rho_i(\beta)$ for $\rho(\beta, \mathbf{X}_i)$. Assume that $0 = \tau_{(0)} < \tau_{(1)} < \dots < \tau_{(n^*)}$ are the distinct ending time, where $n^* < n$. Let $n_{(i)}$ denote the number of point processes with ending time equal to $\tau_{(i)}$.

Define $R_{(l)} = \sum_{i=l}^{n^*} n_{(i)}$, $L_{(l)} = n - R_{(l)}$ and

$$B_{(l)(m)} = \int_{\tau_{(l-1)}}^{\tau_{(l)}} \int_{\tau_{(m-1)}}^{\tau_{(m)}} \lambda_0(t_1) \lambda_0(t_2) f(t_1, t_2) dt_1 dt_2.$$

Then,

$$B = \sum_{i=1}^{n^*} n_{(i)} \sum_{l=1}^i \sum_{m=1}^i B_{(l)(m)} = 2 \sum_{l=1}^{n^*} \sum_{m < l} R_{(l)} B_{(l)(m)} + \sum_{l=1}^{n^*} R_{(l)} B_{(l)(l)}.$$

For $m \leq l$, we estimate $B_{(l)(m)}$ by

$$\widehat{B}_{(l)(m)} = \frac{1}{R_{(l)} \{R_{(m)} - 1\}} \sum_{j=L_{(l)}+1}^n \sum_{\substack{k=L_{(m)}+1 \\ k \neq j}}^n \int_{\tau_{(l-1)}}^{\tau_{(l)}} \int_{\tau_{(m-1)}}^{\tau_{(m)}} \frac{f(t_1, t_2)}{\rho_j(\widehat{\beta}) \rho_k(\widehat{\beta})} N_j(dt_1) N_k(dt_2).$$

Figure 1.

Histograms of counts of events and plots of $\hat{G}(r)$ for the CGD data and the simulated Poisson cluster process data, where the counts are the total number of events for each recurrent event process. In plots (b) and (d), the solid lines are for $\hat{G}(r)$, the dashed lines are the permutation envelopes, and the dotted lines are the line of constant one.

Figure 2. Plots of $\hat{G}(r)$ (solid lines) and the associated permutation envelopes (dashed lines) for the CMD data. The reporting units are MSOAs for plot (a) and local authorities for plot (b). The dotted line in each plot is the line of constant one.

Figure 3.
Plot of the empirical pair correlation $\hat{g}(r)$ for the CMD data with $h = 19$.

Table 1

Empirical biases and standard deviations (STDs) of $\hat{A}$, $\hat{B}$ and $\hat{G} = \hat{A}/\hat{B}$ from 1,000 simulations, with $f(t_1, t_2) = I(|t_1 - t_2| \leq r)$ for $r = 20, 40, 80$. PCP₁ and PCP₂ stand for the Poisson cluster process with $\omega = 20$ and 10, respectively.

	Process	(n, a)	$\hat{A}$			$\hat{B}$			$\hat{G}$		
			r = 20	40	80	20	40	80	20	40	80
BIAS	Poisson	(128,1)	.031	.000	.001	.005	.004	.006	.040	-.006	.005
		(128,4)	.005	.006	-.003	.004	.005	.004	.004	.001	-.006
		(512,1)	-.009	-.016	-.012	.004	.003	.003	-.009	-.017	-.012
	PCP ₁	(128,1)	.027	.010	.002	.018	.018	.017	.082	.050	.037
		(128,4)	.015	.017	.016	.012	.012	.013	.022	.022	.017
		(512,1)	-.004	-.001	-.002	.000	.001	.000	.011	.012	.011
	PCP ₂	(128,1)	.014	.008	.007	.025	.024	.022	.074	.052	.043
		(128,4)	.016	.017	.016	.007	.011	.013	.030	.024	.019
		(512,1)	.005	.004	.001	.002	.002	.002	.018	.016	.013
STD	Poisson	(128,1)	1.322	.906	.668	.302	.293	.293	1.274	.819	.617
		(128,4)	.569	.410	.321	.174	.171	.169	.541	.377	.276
		(512,1)	.559	.408	.306	.151	.151	.152	.540	.380	.276
	PCP ₁	(128,1)	.676	.615	.598	.571	.565	.560	.527	.394	.362
		(128,4)	.329	.315	.312	.313	.311	.308	.210	.179	.163
		(512,1)	.277	.265	.260	.271	.271	.273	.188	.158	.145
	PCP ₂	(128,1)	.620	.603	.600	.608	.588	.575	.433	.377	.361
		(128,4)	.314	.311	.310	.314	.313	.311	.200	.175	.164
		(512,1)	.271	.265	.266	.279	.278	.278	.162	.151	.148