




Nonparametric Estimation for Self-Exciting Point Processes—A Parsimonious Approach

Feng Chen & Peter Hall


To cite this article: Feng Chen & Peter Hall (2016) Nonparametric Estimation for Self-Exciting Point Processes—A Parsimonious Approach, Journal of Computational and Graphical Statistics, 25:1, 209-224, DOI: [10.1080/10618600.2014.1001491](https://doi.org/10.1080/10618600.2014.1001491)

To link to this article: <https://doi.org/10.1080/10618600.2014.1001491>

 View supplementary material 



 Accepted author version posted online: 13 Feb 2015.
Published online: 09 Mar 2016.

 Submit your article to this journal 

 Article views: 380

 View related articles 

 View Crossmark data 

 Citing articles: 3 View citing articles 

Nonparametric Estimation for Self-Exciting Point Processes—A Parsimonious Approach

Feng CHEN and Peter HALL[†]

There is ample evidence that in applications of self-exciting point-process models, the intensity of background events is often far from constant. If a constant background is imposed that assumption can reduce significantly the quality of statistical analysis, in problems as diverse as modeling the after-shocks of earthquakes and the study of ultra-high frequency financial data. Parametric models can be used to alleviate this problem, but they run the risk of distorting inference by misspecifying the nature of the background intensity function. On the other hand, a purely nonparametric approach to analysis leads to problems of identifiability; when a nonparametric approach is taken, not every aspect of the model can be identified from data recorded along a single observed sample path. In this article, we suggest overcoming this difficulty by using an approach based on the principle of parsimony, or Occam's razor. In particular, we suggest taking the point-process intensity to be either a constant or to have maximum differential entropy, in cases where there is not sufficient empirical evidence to suggest that the background intensity function is more complex than those models. This approach is seldom, if ever, used for nonparametric function estimation in other settings, not least because in those cases more data are typically available. However, our "ontological parsimony" argument is appropriate in the context of self-exciting point-process models. Supplementary materials are available online.

Key Words: Hawkes process; Identifiability; Nonparametric function estimation; Occam's razor; Temporal trend.

1. INTRODUCTION

The Hawkes self-exciting process (Hawkes 1971; Hawkes and Oakes 1974) is a point process model used widely in areas such as neuroscience, seismology, finance, social science, marketing research, and criminology (Chornoboy, Schramm, and Karr 1988; Ogata 1988; Chavez-Demoulin, Davison, and McNeil 2005; Crane and Sornette 2008; Kopperschmidt and Stute 2009; Errais, Giesecke, and Goldberg 2010; Mohler et al. 2011; Porter and White 2012). Statistical inference for the classical Hawkes process and its spatial-temporal variant has been studied widely; see, for example, Ogata (1978), Chornoboy,

Feng Chen, School of Mathematics & Statistics, University of New South Wales, Sydney, NSW 2052, Australia (E-mail: feng.chen@unsw.edu.au). Peter Hall, Department of Mathematics & Statistics, University of Melbourne, VIC 3010, Australia (E-mail: halpstat@ms.unimelb.edu.au). [†]Dr. Hall sadly passed away January 9, 2016.

© 2016 *American Statistical Association, Institute of Mathematical Statistics,*
and Interface Foundation of North America

Journal of Computational and Graphical Statistics, Volume 25, Number 1, Pages 209–224

DOI: [10.1080/10618600.2014.1001491](https://doi.org/10.1080/10618600.2014.1001491)

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jcgs.

Schramm, and Karr (1988), and Rathbun (1996). In the classical Hawkes process, the background event intensity is assumed to be a constant, not varying with time. This assumption is rather stringent and is likely to be violated in applications. For instance, in seismology, the frequency of after-shocks following a major earthquake is known to be decreasing over time; and in ultra-high frequency financial modeling, the arrival rate of trades of a stock is typically much higher near the points where the market opens or where it closes than at other times. This kind of temporal trend in the event rate renders inappropriate the constant background intensity assumption for the classical Hawkes self-exciting process, and motivates work to extend classical Hawkes process models to allow for time-varying background intensities, and to develop associated statistical methodology.

Indeed, there has been recent work (Chen and Hall 2013) on inference for parametric self-exciting processes with time-varying background intensities. However, parametric models are subject to misspecification, and so in this article we study the problem of nonparametric estimation of self-exciting processes with varying background intensities. Interestingly, it turns out that not every aspect of such a model can be identified from an observed sample path. We overcome this identifiability issue by appealing to Occam's razor.

The remainder of the article is organized as follows. In Section 2, we introduce the model, discuss identifiability, and introduce methodology for inference. Numerical performance of our methodology is assessed using simulations in Section 3. Applications are illustrated in Section 4, using real data from insurance. Section 5 briefly discusses numerical methods, and theoretical properties of our estimators are described in Section 6. All technical arguments are relegated to an online supplement.

2. MODEL AND METHODOLOGY

2.1 MODEL

The self-exciting point process is a mathematical model for the population size of a continuous time branching process, where the birth times of ancestors follow a Poisson process with a certain intensity function (the baseline intensity), and an offspring of any generation, once born, starts to give birth to its own offspring at times of a Poisson process with another intensity function (the fertility function, or excitation function).

The self-exciting point process (SEPP) model can be expressed as

$$N(t) = n \int_0^t v(u) du + \int_0^t dv \int_0^{v-} g(v-u | \beta_0) dN(u) + \epsilon(t), \quad (2.1)$$

where $n \asymp E\{N(1)\}$ is a normalizing constant. Here, $N(t)$ denotes the total number of points of the self-exciting point process in the interval $[0, t]$, $t \in \mathcal{I} = [0, 1]$, v is interpreted as a point-process intensity, β_0 is the true value of a parameter β (a p -vector), $g(\cdot - u | \beta)$ denotes the intensity of the superimposed point process “generated” by a point at u in the underlying (inhomogenous) process with intensity v , and the process ϵ has zero mean; it is a martingale if the component parts of the SEPP are Poisson processes. The function $g(\cdot | \beta)$ is nonnegative and continuous in β , and its form given β is assumed known. To ensure that the point process N has, with probability 1, a finite number of points on every finite interval, we ask that $\int g(x | \beta) dx < 1$.

The process N is only weakly dependent, and so

$$n^{-1} N(t) \rightarrow H(t) = \int_0^t h(u) du$$

in probability, uniformly in $t \in \mathcal{I}$ as $n \rightarrow \infty$, where the function h uniquely satisfies the equation (see Linz 1985, Theorem 3.5)

$$h(t) = v(t) + \int_0^t g(t-u | \beta_0) h(u) du, \quad (2.2)$$

for all $t \in \mathcal{I}$. Moreover, if we define $m(t) = E\{N(t)\}$, then, for each fixed n ,

$$m' = n h, \quad (2.3)$$

from which it follows that

$$n^{-1} m(t) = H(t).$$

To appreciate why (2.3) holds, observe that, taking expectation of both sides of (2.1),

$$m(t) = n \int_0^t v(u) du + \int_0^t dv \int_0^v g(v-u | \beta_0) m'(u) du,$$

which on differentiation gives $m'(t) = n v(t) + \int_0^t g(t-u | \beta_0) m'(u) du$. Comparing this with (2.2) gives (2.3).

2.2 IDENTIFIABILITY

As before, let h denote the solution of (2.2), and consider the following version of (2.2), where we replace the true value, β_0 , of β by a general value:

$$h(t) = v(t) + \int_0^t g(t-u | \beta) h(u) du. \quad (2.4)$$

We can interpret (2.4) as defining $v = v(\cdot | \beta)$ as a functional of β , for given h . Specifically,

$$v(t | \beta) = h(t) - \int_0^t g(t-u | \beta) h(u) du. \quad (2.5)$$

Of course, in order for v to have a valid practical interpretation as a point-process intensity, it should be nonnegative. We know that there is at least one solution (β_0, v) of (2.4), with $v \geq 0$ on \mathcal{I} , and it follows that generally there is an infinite number of solutions $(\beta, v(\cdot | \beta))$ for which $v(\cdot | \beta) \geq 0$. To appreciate why, it is necessary only to perturb the value of β in (2.5), and exploit the assumption that the functional $g(\cdot | \beta)$ is continuous in β . This argument shows that if the “initial” value v_0 of v , that is, the value in the pair (β_0, v) for which (2.2) and (2.4) coincide, is strictly positive on \mathcal{I} , then there is an infinity of solutions $(\beta, v(\cdot | \beta))$ of (2.4) for which $v(\cdot | \beta) \geq 0$. The same property also follows, after a little thought, in many cases where v_0 vanishes at one or more points in \mathcal{I} .

Therefore, the pair (v, β) is not, in general, identifiable from (2.4), or equivalently, from the mean alone. This is generally not a problem if we assume a parametric model for v , but if we are taking a nonparametric interpretation of the problem of estimating v , then the lack of identifiability is a challenge. In principle, we could make still further distributional

assumptions, and use the second-order properties (such as covariance) that they entail to characterize ν . However, it is difficult to be sure that those second-order aspects are valid; in particular, their validity is difficult to test empirically. Moreover, methodology for estimating ν in this setting is often particularly complex.

These issues motivate us to consider an alternative approach to removing the identifiability problem, using Occam's razor. That is, we choose the simplest candidate for ν , and we interpret simplicity in the sense of being closest to the constant intensity or, essentially equivalently, having maximum differential entropy.

2.3 METHODOLOGY

Let \hat{h} denote a nonparametric estimator of h ; examples are given in (2.10) and (2.11). Define the sets

$$\begin{aligned} B_0 &= \left\{ \beta \in B : h(t) - \int_0^t g(t-u|\beta) h(u) du \geq 0 \quad \text{for all } t \in \mathcal{I} \right\}, \\ \hat{B}_0 &= \left\{ \beta \in B : \hat{h}(t) - \frac{1}{n} \int_0^{t-} g(t-u|\beta) dN(u) \geq 0 \quad \text{for all } t \in \mathcal{I} \right\} \\ &= \left\{ \beta \in B : \hat{\nu}(t|\beta) \geq 0 \right\}, \end{aligned} \quad (2.6)$$

where

$$\hat{\nu}(t|\beta) = \hat{h}(t) - \frac{1}{n} \int_0^{t-} g(t-u|\beta) dN(u). \quad (2.7)$$

Note that the definition of \hat{B}_0 is free of n , in view of the definition of \hat{h} , such as that given in (2.10) and (2.11). We view \hat{B}_0 as an estimator of B_0 , and suggest the following approach to estimating β and ν .

If \hat{B}_0 is empty, then we define $\hat{\beta}$ to be the value of β that minimizes

$$\int_{\mathcal{I}} I\{\hat{\nu}(t|\beta) < 0\} dt,$$

where $I\{\cdot\}$ denotes the indicator function, or equivalently, which minimizes the proportion of \mathcal{I} for which $\hat{\nu}(t|\beta)$ is strictly negative. In this case, we take our estimator of ν to be $\hat{\nu}_1 \equiv \max\{\hat{\nu}(\cdot|\hat{\beta}), 0\}$. If \hat{B}_0 is nonempty, then we take $\hat{\beta}$ to be the value of β that minimizes, over $\beta \in \hat{B}_0$, a measure of the “disorder” of the function $\hat{\nu}(\cdot|\beta)$. In this instance, we estimate ν by $\hat{\nu}_1 \equiv \hat{\nu}(\cdot|\hat{\beta})$.

Disorder can be measured in a variety of ways, including the L_2 distance, $\hat{d}_1(\beta)$ say, of $\hat{\nu}(\cdot|\beta)$ from the nearest constant function:

$$\hat{d}_1(\beta) = \int_{\mathcal{I}} \{\hat{\nu}(t|\beta) - \hat{\nu}_{\text{av}}(\beta)\}^2 dt, \quad (2.8)$$

where $\hat{\nu}_{\text{av}}(\beta) = \int_{\mathcal{I}} \hat{\nu}(t|\beta) dt$. Note the disorder measure $\hat{d}_1(\beta)$ would be equal to 0 if $\hat{\nu}(\cdot|\beta)$ were a constant function. Bigger values of $\hat{d}_1(\beta)$ implies greater disorder in $\hat{\nu}(\cdot|\beta)$. Another measure of disorder is based on the differential entropy, $-\hat{d}_2(\beta)$ say, of the density $\hat{\nu}_{\text{den}}(\cdot|\beta) = \hat{\nu}(\cdot|\beta)/\hat{\nu}_{\text{av}}(\beta)$ that is proportional to $\hat{\nu}(\cdot|\beta)$:

$$-\hat{d}_2(\beta) = - \int_{\mathcal{I}} \hat{\nu}_{\text{den}}(t|\beta) \log \hat{\nu}_{\text{den}}(t|\beta) dt. \quad (2.9)$$

(Here it is assumed that $\beta \in \widehat{B}_0$, so that $\hat{v}_{\text{den}}(t | \beta) > 0$ for all $t \in \mathcal{I}$.) Note that $-\hat{d}_2(\beta)$ takes the maximum value zero when $\hat{v}_{\text{den}}(\cdot | \beta) \equiv 1$ or when $\hat{v}(\cdot | \beta)$ is a constant (Gokhale 1975). Thus, as a measure of disorder, $\hat{d}_2(\beta)$ will be minimal when $\hat{v}(\cdot | \beta)$ is a constant, and larger values of $\hat{d}_2(\beta)$ suggest further deviations of $\hat{v}(\cdot | \beta)$ from constancy. It is worth mentioning that we have used the word “disorder” to mean nonconstancy or variability of the baseline intensity function, while in some contexts, such as information theory, disorder refers to lack of information, which actually means the flatness, or constancy, of the density function in question.

Estimators of h include a standard kernel estimator,

$$\hat{h}(t) = \hat{h}_{\text{ker}}(t | b) = \frac{1}{nb} \sum_{i=1}^{N(1)} K\left(\frac{t - t_i}{b}\right), \quad (2.10)$$

where $t_1, \dots, t_{N(1)}$ is an enumeration of the points in the SEPP, b is a bandwidth, and K is a kernel function, which we take to be a bounded, symmetric probability density. Provided that $b \rightarrow 0$ and $nb \rightarrow \infty$ as $n \rightarrow \infty$, the estimator at (2.10) generally is consistent for $h(t)$ on $(0, 1)$, but is inconsistent at 0 and 1, and more generally, suffers from edge effects close to the boundaries. This problem can be alleviated by reflecting the data t_i in 0 and 1, in which case the estimator is consistent throughout \mathcal{I} , although it has bias of order h at 0 and 1, whereas it has bias of order h^2 in the interior of \mathcal{I} . Improved accuracy at the boundary can be achieved using so-called boundary kernels to estimate $h(t)$ when t is close to either end of \mathcal{I} . Local polynomial methods, based on interpreting the problem as one of nonparametric regression, are also possible.

An alternative approach, producing an estimator with properties similar to those of the kernel estimator based on reflected data, is to use orthogonal series methods founded on the cosine series:

$$\hat{h}(t) = \hat{h}_{\text{cos}}(t | r) = \sum_{j=1}^r \hat{\alpha}_j \phi_j(t), \quad (2.11)$$

where $\phi_0 \equiv 1$, $\phi_j(t) = 2^{1/2} \cos(j\pi t)$ for $j \geq 1$, and

$$\hat{\alpha}_j = \frac{1}{n} \sum_{i=1}^{N(1)} \phi_j(t_i)$$

is an estimator of the j th Fourier coefficient, $\alpha_j = \int_{\mathcal{I}} h(t) \phi_j(t) dt$, of h .

2.4 TARGETS OF INFERENCE

Analogously to the definitions of \hat{d}_1 and \hat{d}_2 at (2.8) and (2.9), put

$$d_1(\beta) = \int_{\mathcal{I}} \{v(t | \beta) - v_{\text{av}}(\beta)\}^2 dt, \quad d_2(\beta) = \int_{\mathcal{I}} v_{\text{den}}(t | \beta) \log v_{\text{den}}(t | \beta) dt, \quad (2.12)$$

where $v_{\text{av}}(\beta) = \int_{\mathcal{I}} v(t | \beta) dt$, $v_{\text{den}}(\cdot | \beta) = v(\cdot | \beta)/v_{\text{av}}(\beta)$, and $v(\cdot | \beta)$ is as defined at (2.5). Let β_1 denote the minimizer, over $\beta \in B_0$, of a chosen theoretical measure of disorder, such as d_1 and d_2 above, and put $v_1 = v(\cdot | \beta_1)$. We interpret $\hat{\beta}$ and \hat{v}_1 as estimators of β_1 and v_1 , respectively. It will be shown in Section 6 that the estimators are consistent.

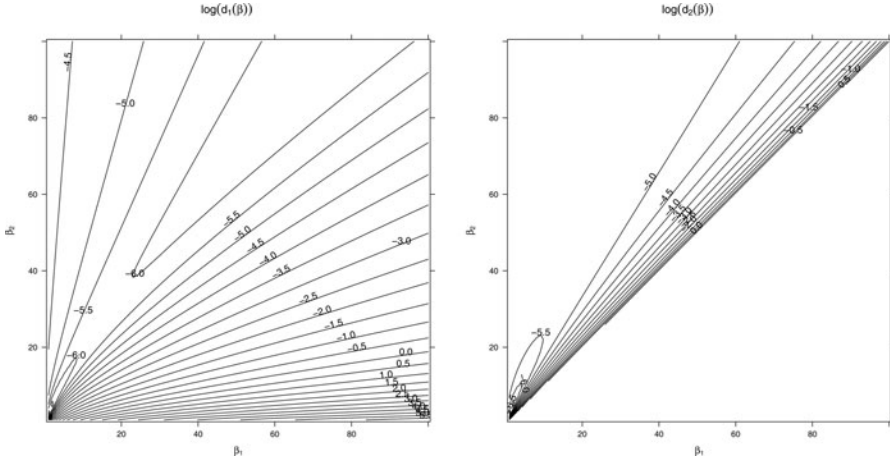


Figure 1. Contour plots of the disorder measures $d_1(\beta)$ and $d_2(\beta)$ on a logarithmic scale. The minimizer of $d_1(\beta)$ is $\beta_1 = (0.67, 1.27)$. The minimizer of $d_2(\beta)$ is $\beta_1 = (0.63, 1.13)$.

3. SIMULATIONS

3.1 PRELIMINARIES

To assess the numerical performance of our proposed estimators, we simulate sample paths of an SEPP, that is, $N(t)$ for $t \in [0, 1]$, with background intensity $n v_0$, where $v_0(t) = 20 + \cos(8t^2)$, and excitation function $g(t | \beta_0) = \beta_{01} \exp(-\beta_{02} t)$, for $n = 50, 100, 200, 400$ and $\beta_0 = (\beta_{01}, \beta_{02}) = (1, 2)$. In this model, the normalized mean intensity, $h_0 = n^{-1}m'$, is given by

$$h_0(t) = v_0(t) + \int_0^t v_0(u) \beta_{01} \exp\{-(\beta_{02} - \beta_{01})(t - u)\} du.$$

Since the disorder measures d_1 and d_2 , defined in (2.12), can be arbitrarily small if both components of β are allowed to be so large that $g(\cdot | \beta)$ can approach the Dirac delta function at 0, we also assume that the parameter space B_0 for β is bounded.

Specific to the simulation model considered here, we assume that $B_0 \subseteq \{b = (b_1, b_2) : b_1, b_2 \leq 100\}$. The first target of inference, β_1 , defined using the mean square disorder measure d_1 or the negative entropy disorder measure d_2 , is computed to be $\beta_1 = (0.67, 1.27)$ or $\beta_1 = (0.63, 1.13)$, respectively; see Figure 1 for contour plots of $d_1(\beta)$ and $d_2(\beta)$ on a logarithmic scale. The second target of inference,

$$v_1(t) = v(t | \beta_1) = h_0(t) - \int_0^t g(t - u | \beta_1) h_0(u) du,$$

is shown in Figure 2 together with v_0 .

For each value of $n = 50, 100, 200, 400$, we simulated 1000 sample paths of the SEPP, using the simulation methods of Brix and Kendall (2002) and Møller and Rasmussen (2005). Based on these simulated sample paths, the mean intensity was estimated using the kernel approach with data reflection. We employed the biweight kernel, $K(x) = (15/16)(1 - x^2)_+^2$, and the bandwidth was taken to be $b = N(1)^{-1/4}$. The Occam's razor estimator based

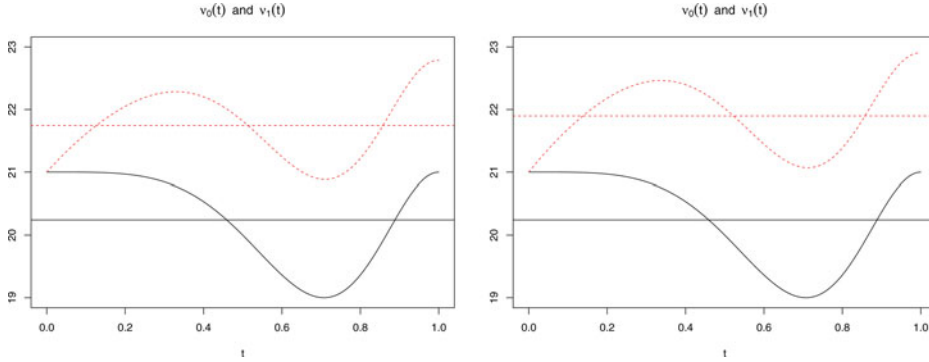


Figure 2. The baseline intensity function v_0 and its proxy v_1 (solid and dashed curves, respectively), with their respective average values (solid and dashed horizontal lines, respectively). Upper: with disorder measure d_1 ; lower: with disorder measure d_2 .

on the integrated square deviation d_1 or the negative entropy d_2 was then used to estimate β_1 and v_1 .

3.2 PERFORMANCE OF ESTIMATOR OF MEAN INTENSITY

In Figure 3, we graph, for each value of $n = 50, 100, 200, 400$, the 1000 estimated mean intensity functions \hat{h} (gray curves) and their pointwise average (dashed bold curve), together with the true mean intensity function h (bold curve in the center). It can be seen that the estimators perform satisfactorily, with biases and variances decreasing with n . In particular, the empirical values of mean integrated squared error (MISE) for $n = 50, 100, 200, 400$ equal 3.342, 1.896, 1.104, 0.665, respectively.

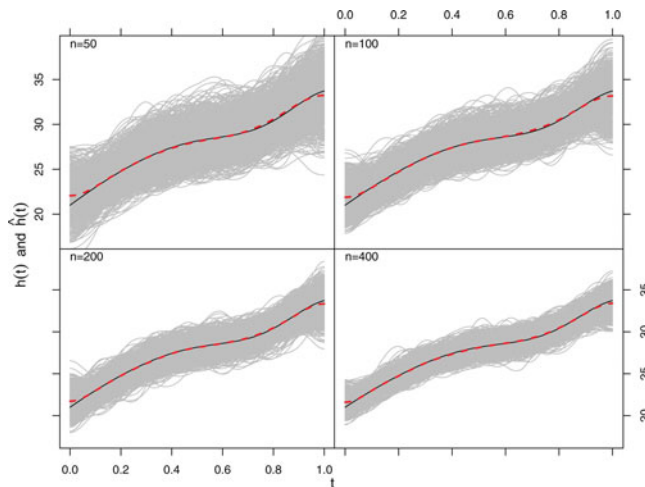


Figure 3. The true normalized mean intensity function h (central solid curve) and its kernel smoothed estimates (solid gray curves), based on simulated data with different scale parameter n , indicated in the panels. The averages of the estimated curves are shown as the broken curves in the center of each graph.

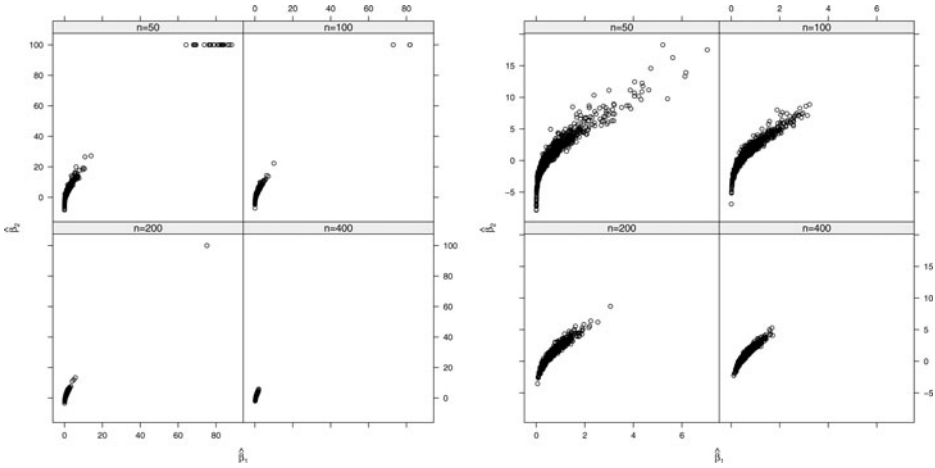


Figure 4. Estimated values of β_1 for $n = 50, 100, 200, 400$. As n increases, the distribution of $\hat{\beta}$ becomes increasingly concentrated in the neighborhood of the true value $\beta_1 = (0.67, 1.27)$, or $\beta_1 = (0.63, 1.13)$, according to disorder measure \hat{d}_1 and \hat{d}_2 , respectively. Left: based on disorder measure \hat{d}_1 ; right: based on disorder measure \hat{d}_2 .

The function $\log(\text{MISE}) = \gamma_1 + \gamma_2 \log n$ that fits best to the four values of (n, MISE) , in the sense of least squares, has $\hat{\gamma}_1 = 4.231$ and $\hat{\gamma}_2 = -0.777$, the latter with standard error 0.014. In particular, the estimate $\hat{\gamma}_2$ agrees well with the value $\gamma_2 = -0.75$ predicted by our asymptotic theory.

Figure 3 also reveals larger estimation biases near the boundaries of the estimation domain, reflecting the slower rate of convergence of the kernel estimator at the end-points discussed in Section 6.1. These results also illustrate the asymptotic theory in Section 6.

With either disorder measure, the proportion of well-behaved estimates of β_1 is high in all four cases, and increases with n . Figure 4 gives a scatterplot of $\hat{\beta}$ in all four cases, and from that figure we observe that the empirical distribution of $\hat{\beta}$ becomes more concentrated in the neighborhood of the true β_1 as n increases. The averages of the “well-behaved” estimates $\hat{\beta}$, when the disorder measure \hat{d}_1 is used, are $(1.02, 1.55)$, $(0.89, 1.53)$, $(0.77, 1.37)$, $(0.70, 1.28)$ for $n = 50, 100, 200, 400$, respectively. When the disorder measure \hat{d}_2 is used, the averages are $(0.74, 0.78)$, $(0.69, 0.94)$, $(0.64, 0.95)$, and $(0.62, 0.98)$, respectively. Reflecting our asymptotic theory, these values move successively closer to the true values, $\beta_1 = (0.67, 1.27)$ and $\beta_1 = (0.63, 1.13)$, with respective disorder measures \hat{d}_1 and \hat{d}_2 , as n increases. It seems that the estimates based on \hat{d}_1 converge faster than those based on \hat{d}_2 , although less stable.

We also computed estimates of the function v_1 ; they are shown in Figure 5. It can be seen from that figure that, as n increases, the curve estimates move closer to the true curves that represent the function v_1 , reflecting the uniform consistency of \hat{v}_1 for v_1 established in Section 6.2. There is also numerical evidence that \hat{v}_1 is asymptotically unbiased.

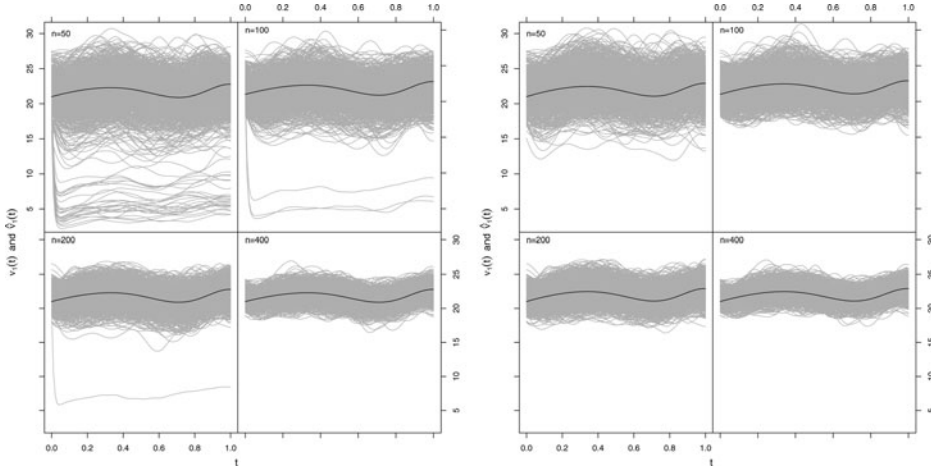


Figure 5. Estimated values of ν_1 for $n = 50, 100, 200, 400$. Central bold curve: truth; gray curves: estimates. Left: based on the disorder measure \hat{d}_1 ; right: based on the disorder measure \hat{d}_2 .

4. REAL DATA EXAMPLE

In this section, we apply the proposed methodology to a real dataset from insurance. The dataset `danish`, available from the R package `evir` Pfaff and McNeil (2012), contains 2167 large fire insurance losses/claims in Denmark (consisting of claims over one million Danish krone) from 1980 to 1990 inclusive. In each instance, the date and size of the claim were available. The data were used by McNeil (1997), among other articles, to illustrate the application of methodology for estimating loss distributions. Relatively less attention has been paid to modeling the claim-arrival process underlying this dataset, although in Burnecki and Weron (2005) it was argued that a nonhomogenous Poisson process (NHPP) with a linear time trend could fit a part of the data. However, in collective risk theory, the claim-arrival process is also an important component of the risk model. In many actuarial calculations, it is assumed to be a Poisson process, perhaps for purposes of mathematical tractability. The independence of claims over nonoverlapping time intervals implied by the Poisson process, however, often disagrees with the observed features of the claim-arrival process, for instance through features such as clustering of losses. Illustrating this point, the top panel of Figure 6, which shows the claim times in the Danish fire insurance claim dataset, reveals that the claims tend to arrive in clusters.

Since the self-exciting process is designed specifically to model the event clustering phenomenon, we modeled the claim-arrival process for Danish large fire insurance claims, during the study period, by a self-exciting process $N(t)$, for $0 \leq t \leq T = 11$ (years). The excitation function was assumed to be exponential, $g(t) = \beta_{01} \exp(-\beta_{02}x)$, with true parameter β_0 , and the baseline intensity function, ν_0 , was assumed to be a smooth function of an unknown form. We estimated the mean claim intensity, h , using the kernel method with data reflection, the biweight kernel, and bandwidth equal to $N(T)^{-1/4} T$. The parameters β_1 and ν_1 , defined as in Section 2.4 with mean square disorder measure d_1 , were then estimated using the proposed procedure. Here we have used d_1 instead of d_2 , since

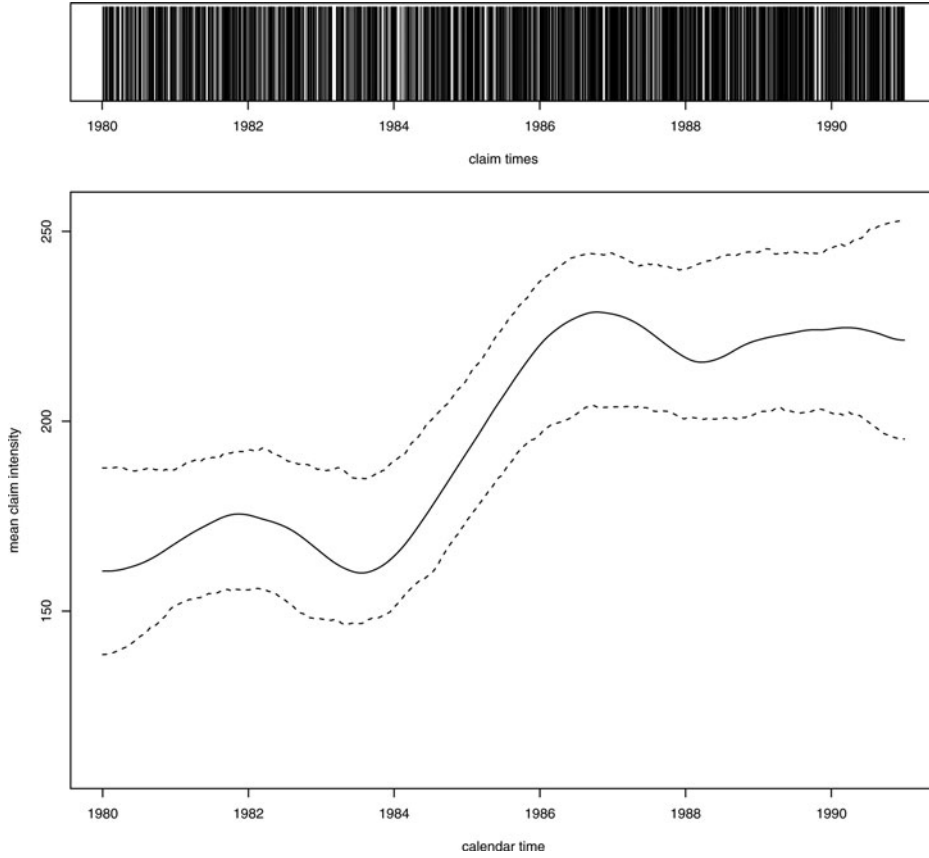


Figure 6. Danish large fire insurance claim times (upper panel) and estimated mean claim intensity in claims per year (lower panel). Each vertical line in the upper panel indicates the occurrence time of a claim. Broken curves in the lower panel indicate pointwise 95% bootstrap confidence limits.

we have seen earlier that the results with both disorder measures are quite similar but the estimator with \hat{d}_1 seems more efficient. Confidence intervals were constructed using a percentile bootstrap method, where we first simulated $B = 500$ sample paths of the SEPP with baseline intensity \hat{v}_1 and exponential excitation function with parameter $\beta = \hat{\beta}$, and then applied the above estimation procedure to each simulated sample path, to obtain the bootstrap samples $\mathcal{H}^* = \{\hat{h}_1^*, \dots, \hat{h}_B^*\}$, $\mathcal{B}^* = \{\hat{\beta}_{11}^*, \dots, \hat{\beta}_{1B}^*\}$, and $\mathcal{N}^* = \{\hat{v}_{11}^*, \dots, \hat{v}_{1B}^*\}$ for \hat{h} , $\hat{\beta}$, and \hat{v}_1 , respectively.

The estimated mean intensity curve, with pointwise 95% bootstrap confidence limits, is shown in the lower panel of Figure 6, and seems to suggest that the claim-arrival process is nonstationary, with mean claim intensity during 1986–1990 significantly higher than that during 1980–1983. The point estimate of β_1 is $\hat{\beta} = (0.068, 0.130)$, with 95% bootstrap confidence intervals $(0.016, 0.211)$ and $(-0.111, 0.521)$ for the two components, respectively. The estimated curve \hat{v}_1 , together with the pointwise 95% bootstrap confidence limits, is shown in Figure 7. To assess goodness of fit, we computed the point process

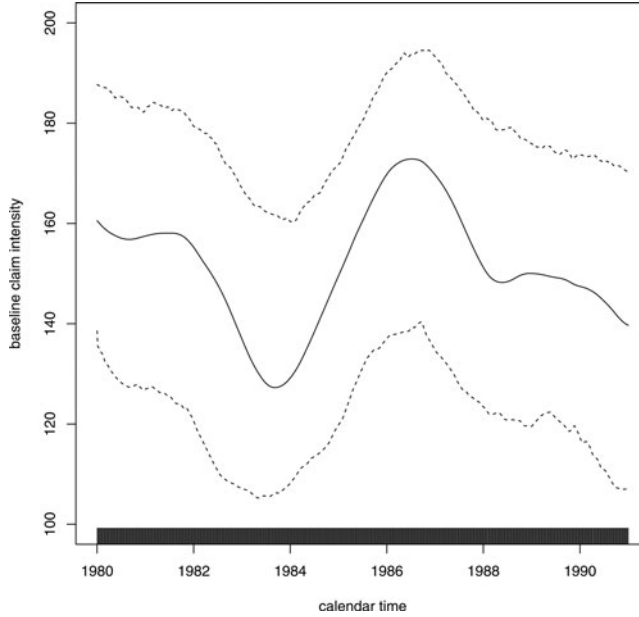


Figure 7. Occam's razor estimate of the baseline claim intensity in claims per year, with mean square disorder measure; broken curves indicate pointwise 95% bootstrap confidence limits.

residuals, defined as $\{\hat{\Lambda}(t_i), | i = 1, \dots, N(T)\}$ with

$$\hat{\Lambda}(t) \equiv \int_0^t \hat{\lambda}(u) du = \int_0^t n \hat{v}_1(u) du + \int_0^t \int_0^u g(u-v | \hat{\beta}) dv du .$$

Sufficient SEPP models, including NHPPs as special cases, should have their point process residuals, thus defined, uniformly distributed over $[0, \hat{\Lambda}(T)]$. A histogram plot of the residuals is shown in the top left panel of Figure 8, which suggests uniformly distributed residuals. A χ^2 -test of equality of histogram bin probabilities returns a large p -value of 0.69. We therefore conclude that an SEPP with baseline intensity \hat{v}_1 and excitation function $g(\cdot | \hat{\beta})$ fits well to the claim times.

Since Figure 7 reveals a rather wide confidence band, through which a horizontal straight line can fit easily, we also fitted an SEPP with a constant baseline intensity $v(t) \equiv a$ to the data. For comparison, we also fitted two other parametric models, a linear intensity NHPP model with intensity function $\lambda(t) = a + bt$, as considered in Burnecki and Weron (2005), and a linear baseline intensity SEPP with baseline intensity $v(t) = a + bt$. In the SEPP models, the excitation function was exponential with parameter β , as before. The maximum likelihood estimates of the parameters of these models are shown in Table 1. The histograms of their point process residuals are shown in the bottom left, top right, and bottom right panels, respectively, of Figure 8. The p -values of the χ^2 -tests of uniformity of the histogram bin probabilities are 0.02, 0.02, and 0.14, respectively. At the significance level 0.05, only the SEPP model with linear baseline intensity gives an acceptable fit to the data on claim times. The estimated value of the vector parameter in the excitation function is $\hat{\beta} = (13.48, 63.19)$, with standard errors 3.86 and 19.05, respectively, for the

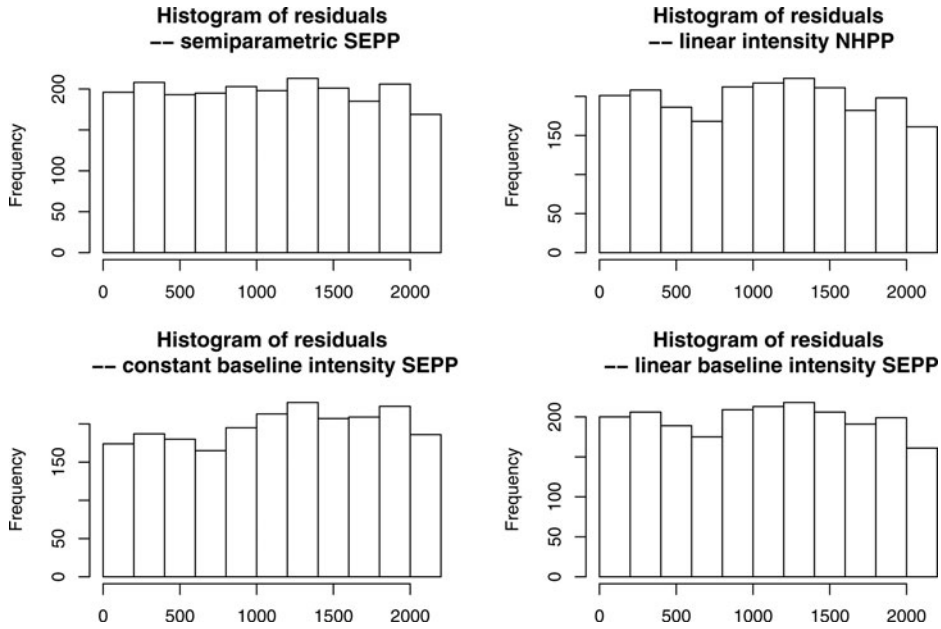


Figure 8. Histograms of the point process residuals of the four models fitted to the claim times data. The p -values of the χ^2 -tests of equality of the histogram bin probabilities are, respectively, 0.69, 0.02, 0.14, and 0.02, clockwise from the top left.

two components. Therefore, the self-exciting effect is statistically significant. Since this model fitted the data reasonably well, and is much simpler than the semiparametric model, we attempt to interpret the data using this model.

To appreciate why the self-excitation effect exists, note that a large fire insurance claim implies that the damage to the insured property was likely to have been relatively extensive, for example, because of failure to control the fire in a timely manner, which in turns implies that the fire might spread and cause damage to nearby properties, rendering other large claims more likely to arise. This is a possible mechanism for self-excitation. The slope of the assumed linear baseline intensity was estimated to be 63.77, with a standard error of 15.26, and therefore is highly statistically significant, with a normal approximation p -value of less than 10^{-4} . This increasing trend in the baseline claim intensity might be due to an

Table 1. Estimated parameters of the three parametric models

	SEPP with $v(t) \equiv a$			NHPP with $\lambda(t) = a + bt$			SEPP with $v(t) = a + bt$		
	Est.	S.E.	p -Value	Est.	S.E.	p -Value	Est.	S.E.	p -Value
a	137.9	12.87	9.1e-27	155.4	7.81	4.7e-88	123.2	11.1	1.3e-28
b	—	—	—	83.13	14.22	5.0e-9	63.77	15.26	2.9e-5
β_1	14.84	4.03	2.3e-4	—	—	—	13.48	3.86	4.8e-4
β_2	49.24	16.52	2.9e-3	—	—	—	63.19	19.05	9.1e-4

increasing trend in the frequency of outbreaks of fire over the period, which itself could be due to changes in environmental factors or growth in overall population.

5. DISCUSSION OF NUMERICAL PROPERTIES

In unreported simulation experiments, we found that cosine series-based estimators perform similarly to those constructed using kernel methods with data reflection. We also found that using either the differential entropy- or L_2 distance-based disorder measure, along with our Occam's razor approach, produces very similar estimates of the excitation parameter and baseline intensity function.

As noted in Section 2.3, local linear or, more generally, local polynomial methods can also be used to estimate h . However, while those approaches enjoy good technical performance in the vicinity of support endpoints, in finite samples they are often inferior at those places because of substantial variance inflation.

To implement our methodology, we used $\hat{v}(t | \beta) = \hat{h}(t) - \int_0^t g(t - u | \beta) \hat{h}(u) du$, instead of the estimator $\hat{v}(t | \beta)$ defined in (2.7), for reasons of computational efficiency. This substitution has negligible effect on estimated values of β_1 and v_1 in the numerical examples we considered. However, evaluation of the disorder measures d_1 and d_2 becomes much faster, since with the new definition the integral involved in the definitions of d_1 and d_2 has a smooth, rather than the previously ragged, integrand, and therefore converges relatively quickly.

The final estimators $\hat{\beta}_1$ and \hat{v}_1 clearly depend on choice of the parametric form of the excitation function $g(\cdot | \beta)$. Taking an exponential form leads to an explicit formula for the mean intensity function, as well as to other simplified calculations, and therefore is quite popular in the literature. However, from a practical viewpoint, there is little reason to stick with the exponential form. In the Danish fire insurance example, we also considered the following polynomial excitation function: $g(t | \beta_0) = \beta_{01} (1 + t)^{-\beta_{02}-1}$. With this polynomial form for g , the resulting estimates $\hat{\beta}$ and \hat{v}_1 are quite similar to those obtained using the exponential function.

The smoothing parameter, for example, the bandwidth b in the kernel case, used to estimate h influences the performance of our estimators. Data-driven smoothing parameter selection methods that are more refined than ours are worthy of investigation.

6. THEORETICAL PROPERTIES

6.1 CONSISTENCY OF MEAN INTENSITY ESTIMATORS

Theorem 1 asserts specific convergence rates for $\hat{h}_{\text{ker}}(\cdot | b)$ and $\hat{h}_{\text{cos}}(\cdot | b)$, in terms of integrated mean squared error, if the following condition holds:

- (a) for all square-integrable functions ϕ on \mathcal{I} , if T is chosen randomly from

$$t_1, \dots, t_{N(1)} \text{ then } E\{\phi(T)\} = \int_{\mathcal{I}} \phi(t) h(t) dt, \text{ and } E\left(\left[\int_{\mathcal{I}} \phi(t) d\{N(t) - \right.\right.$$

$$\left. m(t)\right\} \Big]^2 \Big) \leq C n \int_{\mathcal{I}} \phi^2, \text{ where the constant } C \text{ does not depend on } \phi \text{ or } n;$$

and (b) h has two bounded derivatives on \mathcal{I} . (6.1)

Theorem 1. Assume that (6.1) holds. Then, if $\hat{h} = \hat{h}_{\text{ker}}$ is the kernel estimator at (2.10), using the reflection method to reduce the impact of boundary bias; and if $b = b(n) \rightarrow 0$ and $nb \rightarrow \infty$ as $n \rightarrow \infty$, and K is a bounded, symmetric probability density satisfying $\int u^2 K(u) du < \infty$; then

$$\int_{\mathcal{I}} \{\hat{h}_{\text{ker}}(t | b) - h(t)\}^2 dt = O_p\{(nb)^{-1} + b^3\}. \quad (6.2)$$

If $\hat{h} = \hat{h}_{\text{cos}}$ is the orthogonal series estimator at (2.11), where ϕ_j , for $j \geq 0$, is the cosine basis on \mathcal{I} ; and if $r = r(n) \rightarrow \infty$ and $r/n \rightarrow 0$; then

$$\int_{\mathcal{I}} \{\hat{h}_{\text{cos}}(t | r) - h(t)\}^2 dt = O_p(n^{-1}r + r^{-3}). \quad (6.3)$$

Clearly, the rate at which the right-hands side of (6.2) and (6.3) converge to zero is optimized by taking $b \asymp n^{-1/4}$ and $r \asymp n^{1/4}$, in which case \hat{h} converges to h at rate $n^{-3/4}$ in terms of mean squared error. However, if in (6.2) and (6.3) we replace \mathcal{I} by $[\epsilon, 1 - \epsilon]$ where $0 < \epsilon < \frac{1}{2}$, then it can be proved that (6.2) and (6.3) continue to hold if, in (6.2), we replace b^3 on the right-hand side by b^4 , and, on the right-hand side of (6.3), we replace r^{-3} by r^{-4} . In particular, the best achievable convergence rate of either form of \hat{h} , taken over any closed subset of \mathcal{I} that excludes 0 and 1, is $n^{-4/5}$ rather than $n^{-3/4}$. The rate $n^{-4/5}$ is optimal for functions h with two uniformly bounded derivatives.

6.2 CONSISTENCY OF ESTIMATORS OF β_1 AND ν_1

As a prelude to showing that $\hat{\beta}$ and $\hat{\nu}_1$ converge to their respective targets β_1 and ν_1 , the assumptions in (6.4) are imposed on the counting process N , the functions g and h , the estimator \hat{h} of h , and the disorder measure. In particular, (6.4)(a) is identical to (6.1)(a). Conditions (6.5) are imposed on the sets B and B_0 of p -vectors β :

- (a) for all square-integrable functions ϕ on \mathcal{I} , if T is chosen randomly from

$$t_1, \dots, t_{N(1)}, \text{ then } E\{\phi(T)\} = \int_{\mathcal{I}} \phi(t) h(t) dt, \text{ and } E\left(\left[\int_{\mathcal{I}} \phi(t) d\{N(t) - m(t)\}\right]^2\right) \leq C n \int_{\mathcal{I}} \phi^2, \text{ where the constant } C \text{ does not depend on } \phi \text{ or } n;$$

- (b) the function $g(t | \beta)$ is bounded and continuous in $(t, \beta) \in \mathcal{I} \times B$;

- (c) h is continuous on \mathcal{I} ; (d) $\hat{h} = \hat{h}_{\text{ker}}$, the kernel estimator at (2.10),

and uses a bandwidth $b = b(n)$ satisfying $b \rightarrow 0$ and $nb^3 \rightarrow \infty$, and a kernel K that is a symmetric, differentiable probability density satisfying $\sup |K'| < \infty$; and (e) the disorder measure is taken to be either d_1 or d_2 , at (2.12), with corresponding empirical form \hat{d}_1 or \hat{d}_2 , at (2.8) or (2.9),

respectively; (6.4)

- (a) the set B is a closed, convex, bounded subset of \mathbb{R}^p , where $p \geq 1$;
- (b) $B_0 \subseteq B$ and the boundary ∂B_0 of B_0 , defined at (2.6), is a $(p - 1) -$ dimensional surface in B having a continuously turning tangent; and
- (c) the value β_0 of β that minimizes $d_1(\beta)$ (respectively, $d_2(\beta)$) over B_0 is uniquely defined. (6.5)

Note that the conditions in (6.4)(d) on b hold if b is of the size, either $n^{-1/4}$ or $n^{-1/5}$, discussed immediately below Theorem 1.

Recall the definitions of $\hat{\beta}$ and $\hat{\nu}_1$ given in the paragraph below (2.7), and the definitions of β_1 and ν_1 given in Section 2.4. Theorem 2 asserts that $\hat{\beta}$ and $\hat{\nu}_1$ are consistent for β_1 and ν_1 , respectively.

Theorem 2. If (6.4) and (6.5) hold, then $\hat{\beta} \rightarrow \beta_1$ and $\sup_{t \in \mathcal{T}} |\hat{\nu}_1(t) - \nu_1(t)| \rightarrow 0$, where both convergences are in probability.

SUPPLEMENTARY MATERIALS

Proofs: The technical proofs of the theoretical results in Section 6 are available in the supplementary PDF file. (ChenHall-npSEPP-tech.pdf)

R-code: The R code for simulation of the self-exciting process and estimating it using the proposed methods (with the kernel smoothing estimator) is available in the files `simHawkes.R`, `HawkesRaMEKer.R`, and `HawkesRaSSKer.R`, respectively, in the online supplementary tar archive file. The file `readme.txt` in the archive describes the contents of all the other files in the archive. (ChenHall-npSEPP-Rcode.tar.gz)

ACKNOWLEDGMENTS

The authors gratefully acknowledge the valuable comments of the reviewers, the associate editor, and the editor. The work was partly supported by UNSW Australia SFRG grants (for F.C.).

[Received April 2014. Revised August 2014.]

REFERENCES

- Brix, A., and Kendall, W. S. (2002), "Simulation of Cluster Point Processes Without Edge Effects," *Advances in Applied Probability*, 34, 267–280. [214]
- Burnecki, K., and Weron, R. (2005), "Modelling of the Risk Process," in *Statistical Tools for Finance and Insurance*, eds. P. Čížek, W. K. Härdle, and R. Weron, Berlin: Springer, pp. 319–339. [217,219]
- Chavez-Demoulin, V., Davison, A. C., and McNeil, A. J. (2005), "Estimating Value-at-Risk: A Point Process Approach," *Quantitative Finance*, 5, 227–234. [209]
- Chen, F., and Hall, P. (2013), "Inference for a Non-stationary Self-Exciting Point Process With an Application in Ultra-High Frequency Financial Data Modeling," *Journal of Applied Probability*, 50, 1006–1024. [210]
- Chornoboy, E., Schramm, L., and Karr, A. (1988), "Maximum Likelihood Identification of Neural Point Process Systems," *Biological Cybernetics*, 59, 265–275. [209]
- Crane, R., and Sornette, D. (2008), "Robust Dynamic Classes Revealed by Measuring the Response Function of a Social System," *Proceedings of the National Academy of Sciences*, 105, 15649–15653. [209]
- Errais, E., Giesecke, K., and Goldberg, L. R. (2010), "Affine Point Processes and Portfolio Credit Risk," *SIAM Journal on Financial Mathematics*, 1, 642–665. [209]
- Gokhale, D. V. (1975), "Maximum Entropy Characterizations of Some Distributions," in *A Modern Course on Statistical Distributions in Scientific Work* (Vol. 3, chapter 7), eds. G. P. Patil, S. Kotz, and J. K. Ord, Boston MA: Reidel, pp. 299–304. [213]
- Hawkes, A. G. (1971), "Spectra of Some Self-Exciting and Mutually Exciting Point Processes," *Biometrika*, 58, 83–90. [209]
- Hawkes, A. G., and Oakes, D. (1974), "A Cluster Process Representation of a Self-Exciting Process," *Journal of Applied Probability*, 11, 493–503. [209]
- Kopperschmidt, K., and Stute, W. (2009), "Purchase Timing Models in Marketing: A Review," *ASTA Advances in Statistical Analysis*, 93, 123–149. [209]
- Linz, P. (1985), *Analytical and Numerical Methods for Volterra Equations*, Philadelphia, PA: SIAM. [211]
- McNeil, A. J. (1997), "Estimating the Tails of Loss Severity Distributions Using Extreme Value Theory," *ASTIN Bulletin*, 27, 117–237. [217]
- Mohler, G. O., Short, M. B., Brantingham, P. J., Tita, G. E., and Schoenberg, F. P. (2011), "Self-Exciting Point Process Modeling of Crime," *Journal of the American Statistical Association*, 106, 100–108. [209]
- Møller, J., and Rasmussen, J. G. (2005), "Perfect Simulation of Hawkes Processes," *Advances in Applied Probability*, 37, 629–646. [214]
- Ogata, Y. (1978), "The Asymptotic Behaviour of Maximum Likelihood Estimators for Stationary Point Processes," *Annals of the Institute of Statistical Mathematics*, 30, 243–261. [209]
- (1988), "Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes," *Journal of the American Statistical Association*, 83, 9–27. [209]
- Pfaff, B., and McNeil, A. (2012), *evir: Extreme Values in R. R Package Version 1.7–3*, Available at <http://CRAN.R-project.org/package=evir>. [217]
- Porter, M. D., and White, G. (2012), "Self-Exciting Hurdle Models for Terrorist Activity," *The Annals of Applied Statistics*, 6, 106–124. [209]
- Rathbun, S. L. (1996), "Asymptotic Properties of the Maximum Likelihood Estimator for Spatio-Temporal Point Processes," *Journal of Statistical Planning and Inference*, 51, 55–74. [210]