Ann. Inst. Statist. Math.
Vol. 51, No. 4, 657–678 (1999)

ROBUSTNESS FOR INHOMOGENEOUS POISSON POINT PROCESSES

RENATO ASSUNÇÃO[1] AND PETER GUTTORP[2]

[1] *Departamento de Estatística, UFMG, Caixa Postal 702, Belo Horizonte MG Brazil 30161-970, Brazil*
[2] *Department of Statistics, University of Washington, Seattle, WA 98195-1720, U.S.A.*

(Received March 17, 1995; revised May 25, 1998)

Abstract. We consider robustness for estimation of parametric inhomogeneous Poisson point processes. We propose an influence functional to measure the effect of contamination on estimates. We also propose an M-estimator as an alternative to maximum likelihood estimator, show its consistency and asymptotic normality.

Key words and phrases: Robustness, influence function, M-estimator, point process, Poisson process.

1. Introduction

Only recently has the statistical literature started to consider more frequently problems of statistical inference for stochastic processes. However, robust methods are almost completely absent except for time series data (Künsch (1984), Martin and Yohai (1986, 1991)). The reason seems to be the great technical difficulties in dealing with dependent data as, for example, the need to consider a large variety of possible deviations from a given model, and the difficulty of obtaining limit theorems or distributions of functionals (Martin and Yohai (1986)). Unfortunately, for two reasons it is not possible to generalize directly most of the methods proposed for time series. First, the methods are mainly for ARMA parameters which do not have an immediate analogue for many stochastic processes, in particular, point processes. Second, the methods and estimators use strongly the time direction, which does not have an equivalency for spatial processes where the dimension is higher than one. To our knowledge, the only other work concerning robust methods in a specific case of scalar parameter estimation of point processes in $\boldsymbol{R}$ is Yoshida and Hayashi (1990), to be discussed later on.

We consider the inhomogeneous Poisson point process (IHPP) N in a finite convex polygon $A \subset \boldsymbol{R}^k$ and the same notation, N, is used for a point process and a realization of it. Let $\lambda(\boldsymbol{x}, \boldsymbol{\theta}) > 0$ be the (first-order) intensity of N assumed to depend on covariates through the parameter $\boldsymbol{\theta} \in \boldsymbol{R}^q$. The maximum likelihood estimator (MLE) maximizes the log likelihood and, if the loglikelihood is differentiable, it is a solution of the likelihood equation:

$$(1.1) \qquad \mathbf{0} = \int_A \boldsymbol{l}(\boldsymbol{x},\boldsymbol{\theta})(N(d\boldsymbol{x}) - \lambda(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}) = N(\boldsymbol{l}(\boldsymbol{x},\boldsymbol{\theta})) - E_{\boldsymbol{\theta}}N(\boldsymbol{l}(\boldsymbol{x},\boldsymbol{\theta}))$$

where $\boldsymbol{l}(\boldsymbol{x},\boldsymbol{\theta}) = \partial \log \lambda(\boldsymbol{x},\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. Note that the last equality is of the method-of-moments type and, if $q > 1$, (1.1) is a set of estimating equations and $\boldsymbol{l}(\boldsymbol{x},\boldsymbol{\theta})$ is a vector.

Earlier works have considered the asymptotic properties of the maximum likelihood estimator (MLE). Kutoyants (1984) worked with an inhomogeneous Poisson process on $[0,T] \subset \boldsymbol{R}$. Krickeberg (1982) discussed the extension of Kutoyants' results to a scalar parameter θ and to compact sets A in a locally compact Hausdorff space, which includes $\boldsymbol{R}^k$. Rathbun and Cressie (1994) presented a proof of Krickeberg's results for a vector parameter $\boldsymbol{\theta}$ and a compact $A \subset \boldsymbol{R}^k$. They showed that under regularity conditions the MLE is consistent, asymptotically Gaussian, and asymptotically efficient as the observation region increases. In distribution, $J^{1/2}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges to $N_q(\mathbf{0}, I)$, where I is the identity matrix and $J(\boldsymbol{\theta}) = \int_A \boldsymbol{l}(\boldsymbol{x},\boldsymbol{\theta})\boldsymbol{l}'(\boldsymbol{x},\boldsymbol{\theta})\lambda(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}$. Rathbun (1996) considers the asymptotic behavior of estimators when covariates are only observed for a sample of sites rather than at all locations in the region.

Robust methods can be a valuable addition to the analysis of point processes patterns as the following simple example shows. Consider the occurrence of plants, viewed as discrete points, along a transect. Soil variation coupled with ecological mechanisms creates environmental heterogeneity expressed in terms of different plant intensities along the transect. Suppose the heterogeneity depends on a single variable, soil moisture, which is known along the transect (in practice, it would be measured at a fine grid and interpolated). We assume that the plant locations can be described by an inhomogeneous Poisson process with intensity function $\lambda(x;\theta) = e^{\theta g(x)}$, where $g(x)$ is the soil moisture, measured on some scale, which creates heterogeneity in the soil. This inhomogeneous Poisson process is simulated with $\theta = 1$ and the 185 realized plants are represented by dots along the horizontal axis in Fig. 1. Given these data, we want to estimate the parameter θ.

The MLE is $\hat{\theta} = 1.107$ with estimated standard deviation $s(\hat{\theta}) = 0.079$ and approximate 95% confidence interval of $(0.953, 1.262)$. Three plants, corresponding to 1.5% of the 185 initially present, were erroneously recorded near a low-intensity region (at $x = 81.26$, $x = 81.27$, $x = 81.28$). After this spurious addition, the MLE for θ is $\hat{\theta}^* = .747$. The difference $\hat{\theta} - \hat{\theta}^*$ is equal to $4.56s(\hat{\theta})$ and it is 25% smaller than the true value of θ. Increasing the contamination makes the estimate worse: adding 9 plants (5% of the total) in the region $(79, 82)$ changes the estimator to $\hat{\theta} = .514$, implying a much smaller estimated effect of moisture.

As a second example, consider the periodic intensity $\lambda(t,\theta) = \exp(3\sin(\theta t))$ with $\theta = 5$. Figure 2(A) shows this intensity as a solid line curve and the events of a realization as dots along the horizontal line at height 0. Figure 2(B) shows the multimodal loglikelihood of θ as a solid line curve together with a loglikelihood of a contaminated process as a dashed line curve. The latter is composed by five spurious events around $t = 9.4$ added to the 52 original ones shown in Fig. 2(A). It is clear that there is only a small change on the maximum likelihood estimate. In

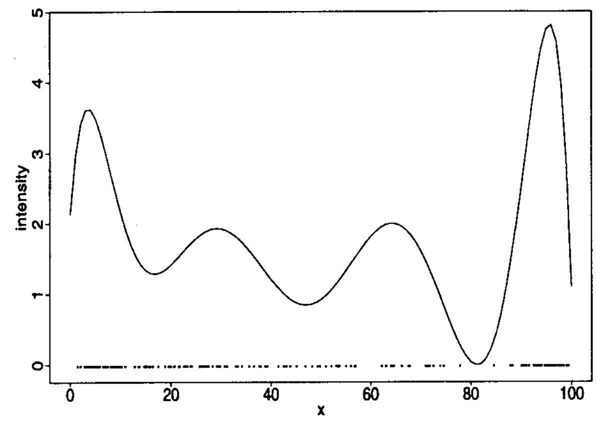

Fig. 1. Plants along the horizontal axis were generated by an inhomogeneous Poisson process with intensity given by $\lambda(x, \theta) = e^{\theta g(x)}$ with $\theta = 1$ and displayed as a curve in the plot. The continuous function $g(x)$ represents some covariate (e.g., moisture) which creates heterogeneity in the soil.

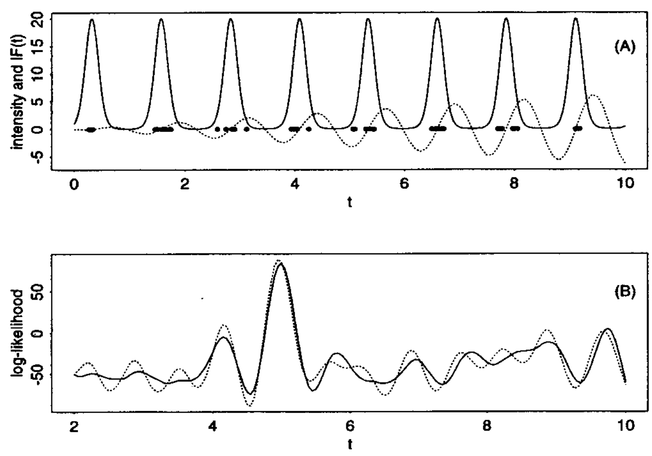

Fig. 2. Events along the horizontal line $y = 0$ in plot (A) were generated by Poisson process with intensity $\lambda(t, \theta) = \exp(3\sin(\theta t))$ with $\theta = 5$. The solid line curve is the intensity and the dashed line curve is the influence function $IF(t)$. The plot (B) shows the log-likelihood function of the original data (solid line curve) and after addition of spurious events (dashed line curve).

fact, this is an expected result according to our definition of an influence function, presented on Subsection 2.1. The spurious events were added at the locations they would influence most the estimates and even so the effect is negligible, as predicted by our results. This influence function is plotted in Fig. 2(A) as a dashed line curve. This shows that robust methods can be useful to point out when some contamination is not likely to affect usual estimators like the MLE.

These simple examples are artificial and the heterogeneity in Example 1 as expressed by $\sup_x \lambda(x,\theta)/\inf_x \lambda(x,\theta) = 122$ is extreme. However, they motivate the search for robust estimation methods in point processes and a general treatment of the problem can be found in Assunção (1994). Previously, Yoshida and Hayashi (1990) proposed an M-estimator for θ in the one-dimensional case and, as far as we know, this is the only robustness approach in the point process literature. By generalizing (1.1), they defined an M-estimator for θ as a solution of $\int_A h(x,\theta)N(dx) - \int_A H(x,\theta)dt = 0$ where h and H are to be chosen. Their approach differs from ours in several important aspects. First, they consider only the special case of a Poisson process on the line and $\theta \in \boldsymbol{R}$ while we intend to generalize to processes in higher dimensions and vector parameters rather than scalars. Second, they suppose that the true intensity is a combination of periodic functions given by $(1-\epsilon)f(t-\theta) + \epsilon c(t)$ with $\lambda(t,\theta) = f(t-\theta)$, $\epsilon \in (0,1]$, $\theta \in (-.5,.5)$ and that f and c are both periodic functions. This is a restrictive set of assumptions and the model considers only contamination by addition of spurious points. Their main results are proofs of consistency and minimax asymptotic variance properties of their estimator. Hence, there is no local robustness as expressed by our definition of influence functional. Third, we choose a robustifying function in a different way. Their robustifying functions h and H in the definition of the M-estimator are related through their assumption (3), page 491, while we adopt a single robustifying function to substitute $\boldsymbol{l}(\boldsymbol{x};\boldsymbol{\theta})$ in (1.1). We find justification for our approach on the link with logistic regression as explained in Subsection 3.1.

In this paper, we concentrate on parametric inhomogeneous Poisson point processes (IHPP). However, in Section 2 we outline a proposal, to be developed elsewhere, of a contamination model and an influence functional for general point processes. We treat IHPP models in Section 3 introducing an M-estimator and its influence curve. We also point out the relationship of our proposal with bounded leverage estimators in generalized linear models (GLM) and we state the results concerning consistency and asymptotic distribution of our M-estimator. Section 4 deals with the choice of a robustifying function and diagnostic tools derived from our proposal. We give an example in Section 5 and present the proofs of the asymptotic results in Section 6.

2. Contamination model

Let $A \subseteq \boldsymbol{R}^k$, $k \geq 1$, be the space where the points of the process lie and $\mathbf{M}(A)$ be the set of all counting measures on A. Given a probability space with sample space Ω, a point process N on A is a random element of $\mathbf{M}(A)$. We assume its first moment measure μ is given by $\mu(B) = \int_B \lambda(\boldsymbol{x})d\boldsymbol{x}$, where $\lambda(\boldsymbol{x})$ is the first order intensity function and $B \in \mathcal{B}$, the Borel sets of A.

An extension of the influence curve to the case of point processes must define a contamination model and the estimators appropriately. A contamination model specifies the character and amount of error in observing N. A simple example of a contaminated process is a thinned process where we observe the original process with some of its points deleted. For instance, suppose that interest is in the spatial configuration of objects identified as mines at the bottom of the ocean. Some of these mines could be missed after being covered by sand. In another context, an epidemic could rip plants of a small region. A particle-counting device may switch off randomly if intensity of incoming particles is above some threshold. Another contaminating mechanism corresponds to the addition of points of another process to the underlying process. For example, a particle-counting device could have a constant level of noise in the form of spurious events being recorded, or plants of a different species could be mislabeled as the species of interest in an ecological study. Measurement errors, such as rounding errors, in the locations of the events provide a third contaminating mechanism. Another common error in forestry is the swap of x and y coordinates: the position of a tree in a region is taken with respect to two coordinate axes and occasionally the x coordinate is recorded in the y column and the y coordinate in the x column of the sheet.

A fairly general model for contaminated counting processes C^γ can be defined as a combination of these mechanisms. The number of events of the contaminated process in $B \in \mathcal{B}$ is given by

$$C^\gamma(B) = \int_B Z^{\gamma_d}(\boldsymbol{x}) N^*(d\boldsymbol{x}) + M^{\gamma_a}(B).$$

Here, Z^{γ_d} is a 0-1 process corresponding to the deletion of points. In general, Z^{γ_d} may depend on the realization N. For example, an event could be deleted with high probability if there are too many neighboring events within a distance r. The parameter γ_d represents the amount of contamination obtained through the thinning of N. When $EN(A) < \infty$, γ_d is the fraction of expected deleted points

$$\gamma_d = \frac{1}{E(N(A))} E\left(N(A) - \int_A Z^{\gamma_d}(\boldsymbol{x}) N(d\boldsymbol{x})\right).$$

For the case when $EN(A) = \infty$, we give an asymptotic interpretation for this fraction by assuming that there exists a sequence of bounded Borel sets B_n such that $B_n \uparrow A$, $EN(B_n) < \infty$, and defining γ_d as the limit of $E^{-1}(N(B_n))E(N(B_n) - \int_{B_n} Z^{\gamma_d}(\boldsymbol{x})N(d\boldsymbol{x}))$ as n tends to infinity.

The process M^{γ_a} is a point process yielding spurious points which are mistaken as points of N. The parameter γ_a represents the amount of contamination introduced by M^{γ_a}. When $\mu(A)$ is finite, γ_a is the fraction of expected added points: $\gamma_a = EM^{\gamma_a}(A)/EN(A)$. When $\mu(A) = \infty$, we proceed as before giving this interpretation an asymptotic character.

Finally, N^* is N subjected to measurement error at each event τ_i, so that $\tau_i^* = (1-\zeta_i)\tau_i + \zeta_i \alpha_i$. The ζ_i's are Bernoulli random variables, possibly depending on N, with probability of success p_i, and α_i's are the τ_i's shifted by the measurement errors. If $E(N(A)) < \infty$, the amount of contamination is measured by the proportion of expected number of points measured with error $\gamma_m = \sum_i p_i/\mu(A)$. For the $E(N(A)) = \infty$ case, we work as before.

The total amount of contamination γ is the sum of all contamination types, $\gamma = \gamma_a + \gamma_d + \gamma_m$. We want to study the properties of estimators when the contamination is small and for that we will be considering situations as $\gamma \to 0$. We assume that the contamination amounts due to addition, deletion and measurement error decrease at a certain relative ratio: $\gamma_a/\gamma \to f_a$, $\gamma_d/\gamma \to f_d$, $\gamma_m/\gamma \to f_m$, which implies $f_a + f_d + f_m = 1$ since $\gamma = \gamma_a + \gamma_d + \gamma_m$.

2.1 *Influence functional for point processes*

In the point process setting, it is usually possible to represent the asymptotic value of a parameter estimate as a functional $T = T(P)$ evaluated at the underlying probability distribution P. Denoting by $P_{\boldsymbol{\theta}}$ the distribution of a parametric family, the asymptotic value should be the parameter of interest in order to satisfy the Fisher-consistency criterium $T(P_{\boldsymbol{\theta}}) = \boldsymbol{\theta}$. The estimators appear as those functionals evaluated at the probability distribution which puts mass one at the observed $N(\omega) \in \mathbf{M}(A)$ where $\omega \in \Omega$. Many of the usual estimators of non-parametric descriptors of point processes can be represented as functionals (Assunção (1994)).

Let T be the functional associated with a certain estimator and P_N and P^γ be the probability distributions of N and C^γ, respectively. Then, following Martin and Yohai (1986), the *influence functional of* T is defined as

$$IF(\{P^\gamma\}, T, P_N) = \lim_{\gamma \to 0} \frac{T(P^\gamma) - T(P_N)}{\gamma},$$

when this limit exists. Note that the influence functional depends on the particular path $\{P^\gamma, \gamma \geq 0\}$ followed by the contaminated distribution as γ goes to zero.

To compute explicit expressions of the influence functional IF, we need to step down from the general definition given earlier. We consider a simpler situation, still general enough to contain some interest. We let $N^* = N$, ignoring measurement errors. Additionally, we impose some independence properties on the deletion and addition processes. From now on, when we refer to the contamination model, we are considering the model that follows.

We let the process Z^{γ_d} be independent of N. We take $Z^{\gamma_d}(\boldsymbol{x})$ to be a Bernoulli random variable with parameter $(1 - \delta p(\boldsymbol{x}))$, where p is a fixed function and $\delta > 0$. That is, an observed N event at $\boldsymbol{x}$ is deleted with probability $\delta p(\boldsymbol{x})$ and we assume that, given N, the vector $(Z^{\gamma_d}(\boldsymbol{x}_1), \ldots, Z^{\gamma_d}(\boldsymbol{x}_{N(B)}))$ is composed by independent random variables. Note that $p(\boldsymbol{x})$ could depend, for example, on the non-random first-order intensity $\lambda(\boldsymbol{x})$ of N.

Let μ be the mean measure of the underlying process. Then we find the following relationship between δ and γ_d:

$$\gamma_d = \frac{1}{E(N(A))} E\left(\int_A (1 - Z^{\gamma_d}(\boldsymbol{x}))N(d\boldsymbol{x})\right) = \frac{\delta}{\mu(A)} \int_A p(\boldsymbol{x})\mu(d\boldsymbol{x}) = \delta \frac{\mu(p)}{\mu(A)}.$$

In particular, δ goes to zero if γ_d goes to zero and vice-versa. For convenience, we take $\mu(p)/\mu(A) = 1$ which implies $\gamma_d = \delta$.

Let the process M^{γ_a} be a point process independent of N and Z^{γ_d}. The first-order intensity of M^{γ_a} is $\epsilon\eta(\boldsymbol{x})$, where $\epsilon > 0$ and η is a fixed non-negative function satisfying $\int_A \eta(\boldsymbol{x})/\mu(A)d\boldsymbol{x} = 1$ when $\mu(A) < \infty$. The function $\eta(\boldsymbol{x})$ could be a function of non-random functions of the distribution of N, such as $g(\lambda(\boldsymbol{x}))$. For example, $g(\lambda(\boldsymbol{x}))$ could be inversely proportional to $\lambda(\boldsymbol{x})$ indicating that we tend to add spurious events on the voids of the underlying N.

The influence functional IF can be computed explicitly for many estimators of point process parameters (Assunção (1994)). In this paper, we concentrate on IHPP and the M-estimator suggested in the next section.

3. An M-estimator for parametric IHPP

To simplify notation, we write $\{N(d\boldsymbol{x})-\lambda(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}\}$ as $dM(\boldsymbol{x},\boldsymbol{\theta})$. If the process occurs on the real line, $dM(\boldsymbol{x},\boldsymbol{\theta})$ can be usefully viewed as a martingale increment. In higher dimensions, it is possible to establish a filtration $\mathcal{F}_t$ such that the differential dM is a zero-mean martingale with respect to it. However, there will be many important events not $\mathcal{F}_t$-measurable rendering this approach useless for $k > 1$ (Assunção (1994)). We write integrals of the form $\int_A f(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}$, $\int_A f(\boldsymbol{x},\boldsymbol{\theta})N(d\boldsymbol{x})$, and $\int_A f(\boldsymbol{x},\boldsymbol{\theta})dM(\boldsymbol{x},\boldsymbol{\theta})$ either as $\int f_{\boldsymbol{\theta}}$, $\int f_{\boldsymbol{\theta}}N$, and $\int(fdM)_{\boldsymbol{\theta}}$, or just as $\int f$, $\int fN$, and $\int fdM$, respectively, with an implicit integration region. The transpose of a matrix C is denoted by C' and the derivative of a function f by $\dot{f}$.

In order to introduce our definition of an M-estimator, we cast (1.1) in a more abstract framework. Let $\mu_{\boldsymbol{\theta}}(B) = E_{\boldsymbol{\theta}}N(B) = \int_B \lambda(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}$ be the boundedly finite mean measure and define the functional T by $T(P_{\boldsymbol{\theta}}) = \boldsymbol{\theta}$. Take expectations with respect to the true $\boldsymbol{\theta}^*$ in (1.1) to obtain

$$\mathbf{0} = \int_A \boldsymbol{l}(\boldsymbol{x},\boldsymbol{\theta})\lambda(\boldsymbol{x},\boldsymbol{\theta}^*)d\boldsymbol{x} - \int_A \boldsymbol{l}(\boldsymbol{x},\boldsymbol{\theta})\lambda(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}. \tag{3.1}$$

We see that taking $\boldsymbol{\theta}$ equal to $\boldsymbol{\theta}^*$ trivially satisfies (3.1). That is, the functional $T(P_{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ is a solution to the functional equation

$$\mathbf{0} = E_{P_{\boldsymbol{\theta}}} \int_A \boldsymbol{l}(\boldsymbol{x},T(P_{\boldsymbol{\theta}}))dM(\boldsymbol{x},T(P_{\boldsymbol{\theta}})).$$

The robust approach assumes that the true distribution P of N is in the neighborhood of the parametric $P_{\boldsymbol{\theta}}$. For N in this neighborhood and extending the definition of T conveniently to this neighborhood, we can write the functional equation on T

$$\mathbf{0} = E_P \int_A \boldsymbol{l}(\boldsymbol{x},T(P))dM(\boldsymbol{x},T(P)).$$

We now introduce a robustifying function $\boldsymbol{\psi}$ instead of $\boldsymbol{l}$ and define an *M-estimator of* $\boldsymbol{\theta}$ by the functional $T(P)$ which is a solution of the estimating equation

$$\begin{aligned}\mathbf{0} &= E_P \int_A \boldsymbol{\psi}(\boldsymbol{x},T(P))dM(\boldsymbol{x},T(P)) \\ &= \int_A \boldsymbol{\psi}(\boldsymbol{x},T(P))(\mu(d\boldsymbol{x}) - \lambda(\boldsymbol{x},T(P)))d\boldsymbol{x}\end{aligned} \tag{3.2}$$

where $\mu(B) = E_P(N(B))$ is the mean measure function of N. We discuss conditions on $\boldsymbol{\psi}$ later on. Our M-estimator is trivially Fisher-consistent because, if $P_{\boldsymbol{\theta}}$ is the true distribution of the point process N, then (3.2) becomes $\mathbf{0} = \int_A \boldsymbol{\psi}(\boldsymbol{x},T(P_{\boldsymbol{\theta}}))\{\lambda(\boldsymbol{x},\boldsymbol{\theta}) - \lambda(\boldsymbol{x},T(P_{\boldsymbol{\theta}}))\}d\boldsymbol{x}$, which obviously has $T(P_{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ as a solution.

When the point process is observed on A, the observed locations give origin to the observed counting measure $N(\omega,\cdot)$, where ω is a random element of the sample space Ω. This is used to formulate the empirical version of (3.2). We take a point process distribution $P_{N(\omega,\cdot)}$ with its probability mass concentrated on the observed counting measure $N(\omega,\cdot)$ and use the estimate $\hat{\boldsymbol{\theta}} = T(P_{N(\omega,\cdot)})$, written as $T(\omega)$ to simplify notation. Then, the M-estimating equation is

$$0 = \int_A \boldsymbol{\psi}(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) dM(\boldsymbol{x}, \hat{\boldsymbol{\theta}}). \tag{3.3}$$

3.1 *Relationship with GLM*

Suppose that the intensity $\lambda(\boldsymbol{x},\boldsymbol{\theta})$ is a smooth function depending on a q-dimensional covariate $\boldsymbol{z}(\boldsymbol{x})$ measured at location x. We divide a finite region $A \subset \boldsymbol{R}^k$ into m small k-dimensional rectangles of volume $\delta > 0$, denoted by A_i. The number of events Y_i in A_i are independent random variables and, if δ is small enough, they have approximately a Bernoulli distribution with parameter $p_i = \lambda(\boldsymbol{x}_i, \boldsymbol{\theta})\delta$, where $\boldsymbol{x}_i$ is the center of A_i. Taking derivatives with respect to $\boldsymbol{\theta}$, the likelihood score of Y_i is approximately $(y_i - p_i)\dot{p}_i/p_i = (y_i - p_i)\partial \log \lambda(\boldsymbol{x}_i, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and hence the likelihood equation is, approximately,

$$\begin{aligned} 0 &= \sum_{i=1}^{m} (y_i - p_i) \frac{\partial \log \lambda(\boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \to \int_A \boldsymbol{l}(\boldsymbol{x}, \boldsymbol{\theta})(N(d\boldsymbol{x}) - \lambda(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x}) \\ &= \int_A \boldsymbol{l}(\boldsymbol{x}, \boldsymbol{\theta}) dM(\boldsymbol{x}, \boldsymbol{\theta}) \end{aligned}$$

as $m \to \infty$. Therefore, we can interpret $\boldsymbol{l}(\boldsymbol{x},\boldsymbol{\theta})dM(\boldsymbol{x},\boldsymbol{\theta})$ as a generalized likelihood score.

A direct extension from the M-estimator in the i.i.d. case would allow for $\boldsymbol{\psi}$ functions involving the whole score in a non-linear way. However, the special meaning of the increments $dM(\boldsymbol{x},\boldsymbol{\theta})$ in the likelihood equation of the point process creates technical difficulties for such a general definition. One alternative is to define robustifying $\boldsymbol{\psi}$ functions involving only the term $\boldsymbol{l}(\boldsymbol{x},\boldsymbol{\theta})$, with $\boldsymbol{\psi}(\boldsymbol{x},\boldsymbol{\theta})dM(\boldsymbol{x},\boldsymbol{\theta})$ as the generalized robust score. An estimator based on $\boldsymbol{\psi}dM$ would have a close relationship with the *bounded leverage estimator*, a Schweppe-type estimator, defined by Stefanski *et al.* (1986) for generalized linear models. For logistic regression models, their estimator is the solution of $\sum_i (y_i - F(\boldsymbol{\theta}'\boldsymbol{z}_i))\boldsymbol{\psi}(\boldsymbol{z}_i,\boldsymbol{\theta})$ where F is the logistic function and $\boldsymbol{z}$ is a vector with the covariates. The function $\boldsymbol{\psi}(\boldsymbol{x},\boldsymbol{\theta})$ depends on the covariates $\boldsymbol{z}$ and on $\boldsymbol{\theta}$ but not on the observation y_i which makes their estimator unbiased, and hence Fisher-consistent. The analogy with our situation is clear: the function $\boldsymbol{\psi}$ assumes the same role, and the "residual" $dM(\boldsymbol{x},\boldsymbol{\theta}) = N(d\boldsymbol{x}) - \lambda(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}$ substitutes for $y_i - F(\boldsymbol{\theta}'\boldsymbol{z}_i)$.

If $\lambda(\boldsymbol{x},\boldsymbol{\theta}) = \exp(\boldsymbol{\theta}'\boldsymbol{z}(\boldsymbol{x}))$ then we have the generalized likelihood score equal to $z(x)dM(\boldsymbol{x},\boldsymbol{\theta})$, which is of the same form as the logistic score. The use of the log-link function make the intensity log-linear in the covariates and can be seen in two ways. First, as we did above, if we approximate the number of events in a small region by a Bernoulli distribution, then the log link is an approximation for a logistic link function since the probability of success is small. On the other hand, we can consider the true distribution of points in small regions Δ, which is a Poisson distribution with parameter $\int_\Delta \lambda(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x} \approx \lambda(\boldsymbol{x}_0,\boldsymbol{\theta})\Delta$. Then we can adopt the logarithm on the grounds that it is the canonical link function for Poisson distributions.

3.2 *Influence function*

Considering our contamination model C^γ from the beginning of Section 3, it is not difficult to obtain an explicit expression for the influence functional of the M-estimator. Consider a one-dimensional parameter θ. Recall that the M-estimator is the solution $T(P)$ of (3.2). The contaminated process C^γ has intensity $\lambda^\gamma(\boldsymbol{x}) = \lambda(\boldsymbol{x},\theta)(1-\gamma_d p(\boldsymbol{x})) + \gamma_a \eta(\boldsymbol{x})$. Letting P^γ denote the distribution of C^γ (3.2) becomes

$$\int_A \psi(\boldsymbol{x},T(P^\gamma))(\lambda(\boldsymbol{x},\theta)(1-\gamma_d p(\boldsymbol{x})) + \gamma_a\eta(\boldsymbol{x}) - \lambda(\boldsymbol{x},T(P^\gamma)))d\boldsymbol{x} = 0. \tag{3.4}$$

Taking derivatives on both sides of (3.4) with respect to γ and evaluating at $\gamma = 0$ yields

$$IF(\{P^\gamma\},T,P_\theta) = \dot{T}(P_\theta) = \int_A \frac{\psi(\boldsymbol{x},\theta)(f_a\eta(\boldsymbol{x}) - f_d p(\boldsymbol{x})\lambda(\boldsymbol{x},\theta))d\boldsymbol{x}}{E_\theta \int_A \psi(\boldsymbol{x},\theta)l(\boldsymbol{x},\theta)N(d\boldsymbol{x})}. \tag{3.5}$$

For q-dimensional $\boldsymbol{\theta}$, we obtain a matrix version of (3.5)

$$\dot{T}(P_{\boldsymbol{\theta}}) = \left(\int_A \boldsymbol{\psi}(\boldsymbol{x},\boldsymbol{\theta})\boldsymbol{l}'(\boldsymbol{x},\boldsymbol{\theta})\lambda(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}\right)^{-1}\int_A \boldsymbol{\psi}(\boldsymbol{x},\boldsymbol{\theta})(f_a\eta(\boldsymbol{x}) - f_d p(\boldsymbol{x})\lambda(\boldsymbol{x},\boldsymbol{\theta}))d\boldsymbol{x}.$$

As a simple example of (3.5), consider the case $f_a = 1$ and $f_d = 0$ where contamination consists only of addition of points. Assume also that, for $c > 0$, the function $\eta(\boldsymbol{x}) = c1_V(\boldsymbol{x})$ is concentrated and constant in a region V of finite k-dimensional volume $|V|$. Since we set $\int \eta(\boldsymbol{x})d\boldsymbol{x} = \int \lambda(\boldsymbol{x},\theta)d\boldsymbol{x}$, then $\eta(\boldsymbol{x}) = |V|^{-1}\int \lambda(\boldsymbol{y},\theta)d\boldsymbol{y}1_V(\boldsymbol{x})$. Denote the function $\lambda/\int\lambda$ by λ^*. Then the IF converges to

$$\psi(\boldsymbol{x}_0,\theta)\int\lambda \bigg/ \int\psi l\lambda = \psi(\boldsymbol{x}_0,\theta) \bigg/ \int \psi l \lambda^*$$

if $|V|$ shrinks to zero around the point x_0 and under the assumption that the denominator is nonzero.

The MLE of one-dimensional θ has IF given by

$$l(\boldsymbol{x}_0,\theta) \bigg/ \int_A l^2\lambda^*. \tag{3.6}$$

Note that $l(\boldsymbol{x},\theta)$ measures the relative change on the intensity $\lambda(\boldsymbol{x},\theta)$ induced by infitesimal changes on θ. If spurious points are added around locations where this percent change is high and $\lambda(\boldsymbol{x},\theta)$ is close to zero, we could expect large effects on the MLE.

In Fig. 2, we plot the function $IF(t) = l(t,5)/\int l^2\lambda^*$ where $\lambda(t,\theta) = \exp(3\sin(\theta t))$ and $\theta = 5$. This is the influence function (3.6) under the above periodic model and local contamination. The effect of spurious events added at t on the MLE is periodic with increasing amplitude. Note that the peaks on $IF(t)$ are not in phase with those of $\lambda(t,5)$ which shows that the maximum effect on the MLE is not obtained by events added at points of minimum intensity. Since the maximum $IF(t)$ is around 5, we can expect the change $T(P^{\gamma_a}) - T(P)$ on the estimator approximately equal to $5\gamma_a$ when adding a fraction γ_a of spurious events. Adding 5 events around $t = 9.4$ to the 52 original ones in Fig. 2(A) gives a change of approximately 0.5 which is 10% of 5, the true value of θ. Therefore, we should not expect large changes on the MLE due to addition of events in this model.

Similarly, a deletion-only contamination with deletion function $p(x)$ increasingly concentrating around a point x_0 produces a limiting IF equal to $-\psi(x_0,\theta)/\int \psi l\lambda^*$. In both situations described above, bounded IF for $A = \boldsymbol{R}^k$ occurs if and only if ψ is bounded.

3.3 *Asymptotic properties*

Results assuring the consistency and asymptotic normality of the M-estimator are given below and proved in Section 6. The main practical use of these asymptotic results is to obtain approximate confidence intervals and hypothesis tests.

Previous work in the asymptotic properties of the MLE used the general approach of Ibragimov and Has'minskii (1981) which consists of proving the uniform local asymptotic normality of a parametric family of the probability measures. In our work, we adopt Crámer's approach to prove consistency. The main reason for that is that the M-estimator defined by (3.3) can not be written in an useful way as the gradient of a function $\rho(N,\boldsymbol{\theta})$. This constrains us to work directly with (3.3). That is, our estimator is not viewed as a minimizer of some criterion function. The proof of the asymptotic normality assumes a Lindeberg-type of condition. Less restrictive results seem to require the more powerful locally asymptotic normality theory of Ibragimov and Has'minskii (1981).

Let $A \subset \boldsymbol{R}^k$ be a compact set with positive Lebesgue measure. Define $A_t = \{t\boldsymbol{x}, \boldsymbol{x} \in A\}$ and assume that $A_t \uparrow \boldsymbol{R}^k$ as $t \to \infty$. Suppose N is an IHPP on $\boldsymbol{R}^k$ with mean measure $\mu_{\boldsymbol{\theta}}$ absolutely continuous with respect to (w.r.t.) the Lebesgue measure and parametrized by $\{\boldsymbol{\theta} \in \Theta \subset \boldsymbol{R}^q\}$. The intensity function, denoted by $\lambda(x,\boldsymbol{\theta})$, $x \in \boldsymbol{R}^n$, $\boldsymbol{\theta} \in \Theta$, is the Radon-Nikodym derivative of $\mu_{\boldsymbol{\theta}}$ w.r.t. the Lebesgue measure.

We consider the q-dimensional M-estimator $\hat{\boldsymbol{\theta}}_t$ as the solution of the estimating equation

$$\mathbf{0} = \boldsymbol{g}_t(\boldsymbol{\theta}) \equiv \int_{A_t} \boldsymbol{\psi}(x,\boldsymbol{\theta})\{N(dx) - \lambda(x,\boldsymbol{\theta})dx\}, \tag{3.7}$$

which is a set of simultaneous equations on $\boldsymbol{\theta}$. We want to consider the asymptotic properties of $\hat{\boldsymbol{\theta}}_t$ as $t \to \infty$.

THEOREM 1. *Assume that conditions* (A)–(F) *from Section* 6 *hold. Then, with probability tending to* 1 *as* $t \to \infty$, *the M-equation* $\boldsymbol{g}_t(\boldsymbol{\theta}) = \mathbf{0}$ *has a* $q \times 1$ *root* $\hat{\boldsymbol{\theta}}_t$ *such that* $\hat{\boldsymbol{\theta}}_t \to_p \boldsymbol{\theta}_0$.

THEOREM 2. *Assume that conditions* (A)–(G) *from Section* 6 *hold. If* $\hat{\boldsymbol{\theta}}_t$ *is a consistent solution of* $g_t(\boldsymbol{\theta}) = 0$, *then* $\sqrt{\Lambda_t(\boldsymbol{\theta}_0)}(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0) \rightarrow_d N_q(\mathbf{0}, C^{-1}(\boldsymbol{\theta}_0)F(\boldsymbol{\theta}_0)C'^{-1}(\boldsymbol{\theta}_0))$ *as* $t \to \infty$.

We estimate the covariance matrix by substituting $\boldsymbol{\theta}$ by the consistent estimator $\hat{\boldsymbol{\theta}}_t$ on the formulae above and evaluating the expressions at A rather than at the limit when $t \to \infty$.

4. Choosing a ψ function

In any application we have to choose a specific function ψ. Ideally, this choice should be guided by what specific contamination patterns we want to protect from. In this respect, the IF defined earlier is valuable in showing how the different contamination patterns will affect the estimator. For example, if addition or deletion is suspected in certain regions, the calculations at the end of 3.2 show how to bound that influence. However, we also suggest some members of the ψ class that can be used in general.

Strong results concerning optimal M-estimators in regression contexts have been achieved by Hampel (1978), Krasker (1980), and Krasker and Welsch (1982) and, in GLM models, by Stefanski *et al.* (1986) and Künsch *et al.* (1989). Stefanski *et al.* (1986) studied the GLM model with covariate vector $\boldsymbol{z}$ and response Y considering M-estimators $\hat{\boldsymbol{\theta}}_\psi$ satisfying

$$\sum_{i=1}^{n} \psi(y_i, \boldsymbol{z}_i, \hat{\boldsymbol{\theta}}_\psi) = 0.$$

They require ψ to be Fisher-consistent, i.e.

$$E_\theta \psi(Y, \boldsymbol{z}, \theta) = 0.$$

They impose a bound b^2 on the *self-standardized sensitivity* of the estimator $\hat{\boldsymbol{\theta}}_\psi$ defined as

$$s(\psi) = \sup_{y,\boldsymbol{z}} \sup_{\lambda \neq 0} \frac{\lambda^T IC_\psi(y, \boldsymbol{z})}{(\lambda^T V_\psi \lambda)^{.5}}, \tag{4.1}$$

where $IC_\psi(y, \boldsymbol{z})$ is the influence curve of $\hat{\theta}_\psi$ and V_ψ is its covariance matrix.

A ψ^* estimator satisfying these conditions is called *efficient* if $V_\psi - V_{\psi^*}$ is positive definite for any other ψ in the class. They prove that, if there is such an efficient estimator, it must be

$$\psi(y, \boldsymbol{z}, \boldsymbol{\theta}) = \begin{cases} s(y, \boldsymbol{z}, \boldsymbol{\theta}) - C(\boldsymbol{\theta}), \\ \qquad \text{if } \; b^2 \geq (s - C)' B^{-1} (s - C) \\ (s(y, \boldsymbol{z}, \boldsymbol{\theta}) - C(\boldsymbol{\theta})) b / \sqrt{(s - C)' B^{-1} (s - C)}, \\ \qquad \text{otherwise}, \end{cases} \tag{4.2}$$

with $B(\boldsymbol{\theta})$ and $C(\boldsymbol{\theta})$ satisfying the implicit equations $B(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(\psi_{\boldsymbol{\theta}}\psi'_{\boldsymbol{\theta}})$ and the Fisher-consistency $E_{\boldsymbol{\theta}}(\psi) = 0$ condition.

However, the problem is far from settled. For example, it is known that bounded influence estimators can have unbounded maximal bias curve for small amounts of contamination (Martin *et al.* (1989)). In addition, it is difficult to extend optimality results to point process data since it is not clear how to define an appropriate measure of influence to bound. A direct translation of the sensitivity is unsatisfactory because our IF includes the contamination path through the deletion function $p(\boldsymbol{x})$ and the addition function $\eta(\boldsymbol{x})$ making the supremum difficult to handle. Resorting to the GLM approximation of 3.1 does not work either, since as the number of observations Y_i increases (by taking finer partitions of A), the parameters $P(Y_i = 1)$ go to zero. However, we can use the approximation and results from Stefanski *et al.* (1986) to suggest an element of the ψ class. Although not proven to be optimal in the point process situation, this suggested estimator should have good robustness properties. Another approach, a Huber-type estimator, is discussed in Assunção (1994).

Consider the M-estimator given by

$$\psi(\boldsymbol{x},\boldsymbol{\theta}) = \begin{cases} \boldsymbol{l}(\boldsymbol{x},\boldsymbol{\theta}), & \text{if } b^2 \geq l^{*\prime}B^{-1}l^* \\ \boldsymbol{l}(\boldsymbol{x},\boldsymbol{\theta})b/\sqrt{l^{*\prime}B^{-1}l^*}, & \text{otherwise,} \end{cases} \tag{4.3}$$

where

$$\boldsymbol{l}^*(\boldsymbol{x},\boldsymbol{\theta}) = \boldsymbol{l}(\boldsymbol{x},\boldsymbol{\theta}) - \int_A \boldsymbol{l}(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x},$$

$$\psi^*(x,\theta) = l^*(x,\theta)\min\left\{1, \frac{b^2}{l^{*\prime}(x,\theta)B^{-1}(\theta)l^*(x,\theta)}\right\},$$

and $B(\theta) = \int_A \psi^*\psi^{*\prime}\lambda$.

Given the observed point process N, we obtain this M-estimator as the solution of the simultaneous equations

$$0 = \int_A l(\boldsymbol{x},\boldsymbol{\theta})\sqrt{\min\left\{1, \frac{b^2}{\boldsymbol{l}^{*\prime}(\boldsymbol{x},\boldsymbol{\theta})B^{-1}(\boldsymbol{\theta})\boldsymbol{l}^*(\boldsymbol{x},\boldsymbol{\theta})}\right\}}dM(\boldsymbol{x},\boldsymbol{\theta})$$

and

$$B(\boldsymbol{\theta}) = \int_A \boldsymbol{l}^*(\boldsymbol{x},\boldsymbol{\theta})\boldsymbol{l}^{*\prime}(\boldsymbol{x},\boldsymbol{\theta})\min\left\{1, \frac{b^2}{\boldsymbol{l}^{*\prime}(\boldsymbol{x},\boldsymbol{\theta})B^{-1}(\boldsymbol{\theta})\boldsymbol{l}^*(\boldsymbol{x},\boldsymbol{\theta})}\right\}\lambda(\boldsymbol{x},\boldsymbol{\theta})dx.$$

Any ψ leads to Fisher-consistency and centering is not required. However, there is another reason why centering could be useful and its motivation comes from the relationship with the GLM (see Subsection 3.1). The value of $\boldsymbol{l}(x,\theta)$ is weighted by $\boldsymbol{l}'(x,\theta)B^{-1}\boldsymbol{l}(x,\theta)$, which is a norm of the vector $\boldsymbol{l}(x,\theta)$ defined by the matrix B. Hence, the weights depend on the B-norm of $\boldsymbol{l}(x,\theta)$ rather than on its distance from a convenient average value. The value of $\boldsymbol{l}'B^{-1}\boldsymbol{l}$ on our definition is just a point process version of the diagonal elements h_{ii} of the hat matrix in regression and this is already a centered measure of distance in the covariate space. Hence, we can take $\boldsymbol{l}^* = \boldsymbol{l}$ when the intensity is a log-linear function of the covariates with a constant term as one of the predictors. However, in general, it is useful to take the centered $\boldsymbol{l}^*$ defined previously.

The $\boldsymbol{\psi}$ we propose have two desirable characteristics. First, it is relatively close to the term $\boldsymbol{l}(\boldsymbol{x},\boldsymbol{\theta})$ hence retaining part of the MLE efficiency and rendering valid the asymptotic results (see Assumption C in Section 6). The second is that this $\boldsymbol{\psi}$ estimator bounds $\boldsymbol{l}(\boldsymbol{x},\boldsymbol{\theta})$ outside a certain region and downweighs the information conveyed there (events or no events). The region is $\{x \in A : l^{*\prime}B^{-1}l^* > b^2\}$ and it depends on ψ itself. It can be seen as a distance measure for the term $l(x,\theta)$, standardized by the covariance matrix of the point process score ψdM.

To fully specify the estimator one must determine the value of b. The choice of b involves a trade-off between efficiency and robustness. A small value of b will downweigh more regions in the estimation procedure increasing protection against deviation from the model. However, decreasing b also decreases efficiency by making ψ different from the efficient $l(x,\theta)$ in larger regions. In principle, $b(\theta)$ should not be too small and this is found through the use of the information inequality. Krasker and Welsch (1982) and Stefanski *et al.* (1986) have found minimum bounds for b. Unfortunately, these contributions can not be used in our problem since there is no meaningful sensitivity measure. We leave for future work a more detailed study of this tuning constant. In Section 5, we calculate our estimator using some possible values for b in order to build some knowledge of its effect on simple situations.

An iterative method is required in order to estimate θ and we can use a direct analogue of that proposed by Stefanski *et al.* (1986), or a modification of it which improves speed (Assunção (1994)).

4.1 *Residuals and diagnostics*

Diagnostics are the dual aspect of robust estimation. Based on a robust estimator, parts of the data departing from the model are more easily diagnosed. Since this anomalous part is downweighed during the estimation procedure, it does affect the estimate and hence it does not mask its presence.

A number of residual techniques are available in one dimension, $(0,T)$. Let τ_i be the observed events. Lewis (1972) proposed to look at the transformed process with events $v_i = \Lambda(\tau_i,\boldsymbol{\theta})$ where $\Lambda(t,\boldsymbol{\theta})$ is the accumulated intensity $\int_0^t \lambda(x,\boldsymbol{\theta})dx$. This new process is a Poisson process with rate 1. To use it as a diagnostic, the estimated value of $\boldsymbol{\theta}$ is used to calculate v_i, that is, $v_i = \Lambda(\tau_i,\hat{\boldsymbol{\theta}})$. Ogata (1988) suggested to plot the cumulative number of v-events as a function of the transformed time $\Lambda(t,\hat{\boldsymbol{\theta}})$, and applied a Kolmogorov-Smirnov test to assess the fit. Berman (1983) suggested to test whether the interevent distances are i.i.d. with distribution exp(1). Some other techniques have been discussed in Ogata (1988). They strongly use the ordered structure of the line, and hence they are not easily generalized to higher dimensions. For point processes observed in $[0,a] \times [0,b]$, Cressie ((1991), p. 656) proposed to project the events onto either $[0,a]$ or $[0,b]$, and analyze them as a one-dimensional point process.

Lawson (1993) defined a deviance residual for spatial Poisson processes. Let $\{x_i, i = 1,\ldots,n\}$ be the location of n events observed in the finite region A. The

Poisson process has intensity $\lambda(x,\boldsymbol{\theta})$. Let T_i be the Dirichlet tile area corresponding to the i-th event and $\hat{\theta}$ be the MLE of $\boldsymbol{\theta}$. Lawson obtains a saturated model estimate of $\lambda(x_i,\boldsymbol{\theta})$ as T_i^{-1} and, using $\hat{\lambda}_i$ to denote $\lambda(x_i,\hat{\boldsymbol{\theta}})$ he defines the deviance contribution of the i-th event

$$d_i = \log\left(\sum_{i=1}^{n} \hat{\lambda}_i T_i/n\right) - \log(\hat{\lambda}_i T_i) - 1 + n\frac{\hat{\lambda}_i T_i}{\sum_{i=1}^{n} \hat{\lambda}_i T_i}$$

and the residual

$$r_i = \mathrm{sgn}(T_i^{-1} - \hat{\rho}\hat{\lambda}_i)\sqrt{d_i}, \qquad \text{where} \quad \hat{\rho} = \frac{n}{\sum_{i=1}^{n} \hat{\lambda}_i T_i}.$$

Our robust estimation procedure gives a natural measure of influence of small regions on the fit. We can plot the values of the weights $\omega(\boldsymbol{x},\hat{\boldsymbol{\theta}}_\psi,\hat{B})$ evaluated in a fine grid over the region A. Hence, we can check which areas are being severely downweighed. These are the areas with potential to alter the MLE fit. Then, a scatterplot of the weights against the residuals $\Delta M_i = N_i - (\lambda(x_i,\hat{\boldsymbol{\theta}}_\psi)) \times$ the area of cell i can show whether the model is well fitted to this cell or not. If suspicious cells happen to be neighboring areas, then the robust diagnostic will have identified a subregion where the fit is not good *and* with great influence on a non-robust fit.

5. Example

We return to the analysis of the simulated plant data of Section 1. Let $\theta_0 = 0$ and

$$B_0 = \int_0^{100} (g^*(x))^2 e^{\theta_0 g(x)} dx = \int_0^{100} g^2(x)dx = 215,$$

where $g^*(x) = g(x) - \int g(x)dx$, as initial values for the M-estimator of $\theta = 1$. Table 1 shows these results. The row with $b = \infty$ corresponds to the maximum likelihood estimate. The other lines correspond to the robust estimates for several values of b. We can see that the ψ estimates and its estimated standard deviation are almost invariant for all choices of the tuning constant b and they are equal to the MLE results except by $b = .1$ when we see a slight difference.

The 1.5% contaminated data set adds 3 plants in the region of low intensity around 80, at $x = 81.26$, $x = 81.27$, $x = 81.28$. The new MLE is $\hat{\theta}_l = .747$ with approximated asymptotic standard deviation 0.096 and nominal 95% confidence interval $(0.559, 0.935)$. From Table 1 we note that the robust estimates are very accurate for $b \leq .5$ when they start to degrade and to get close to the maximum likelihood estimate. If we let the contamination increase to 5% by adding 9 plants in the region $(79, 82)$, for $b \leq .2$ we get very good results for the M-estimator.

Figure 3 shows Ogata's diagnostic plot for the transformed process $v_i = \Lambda(\tau_i, \hat{\theta}_l)$ with 1.5% of contamination. It is obvious that the model is not fitting the data well. However, this plot does not give any indication of what is causing the problem. This is a situation where the model could be discarded because of a

Table 1. Maximum likelihood ($b = \infty$) and robust estimates.

	Non-contaminated Data			1.5% of Contamination			5% of Contamination		
b	$\hat{\theta}_\psi$	$\hat{\sigma}$	B	$\hat{\theta}_\psi$	$\hat{\sigma}$	B	$\hat{\theta}_\psi$	$\hat{\sigma}$	B
.1	1.097	.080	112.48	1.073	.082	104.18	1.024	.084	89.64
.2	1.107	.079	161.03	1.060	.081	152.44	.965	.085	136.80
.5	1.107	.079	161.29	.989	.084	140.97	.758	.095	110.33
1	1.107	.079	161.29	.871	.090	123.94	.569	.104	93.17
2	1.107	.079	161.29	.747	.096	109.17	.533	.109	84.07
∞	1.107	.079		.747	.096		.514	.106	

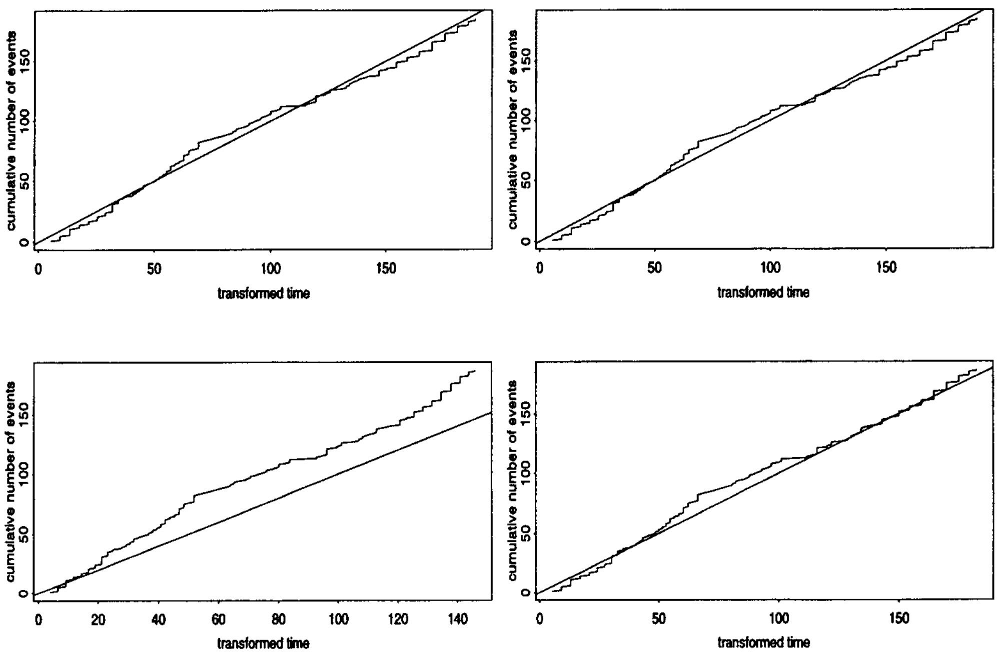

Fig. 3. Cumulative numbers of the transformed process $v_i = \Lambda(\tau_i, \hat{\theta}_l)$ for the non-contaminated process (top plots) and the 1.5% contaminated process (bottom plots). The first column of plots correspond to the MLE and the second, to the robust estimator with $b = 0.2$.

small portion of the data not fitting and washing away the estimate. The robust estimator, being insensitive to the spurious events, will not change when we pass from the non-contaminated process to the contaminated one (see Fig. 3). The plot for the contaminated process shows that the model is fitting well but it is not able to detect the spurious events. We could finish the analysis unable to detect those anomalous events even after adopting the robust estimator.

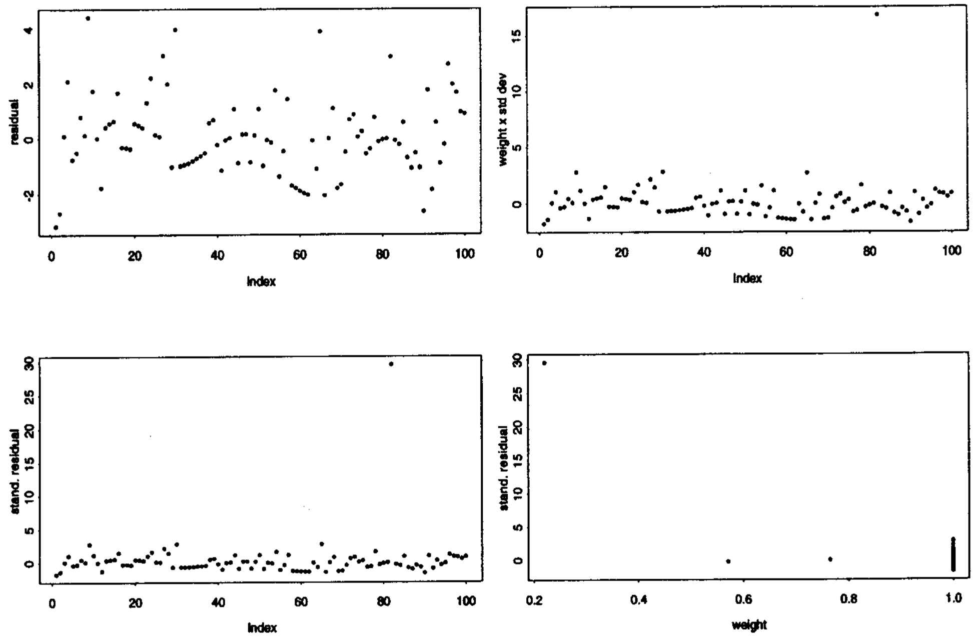

Fig. 4. Robust residuals using $b = .2$. The top plot on the left shows the raw residuals r_i with the index i labeling $A_i = (i-1, i)$ (see text for definitions). The other plots show the standardized residuals r_i^* and the weights.

We also use some robust diagnostic tools, in addition to Ogata's plot. We divide the segment $[0, 100]$ into 100 equally sized subintervals $A_1, \ldots, A_{100}$. For each A_i, we calculate the number of events $N(A_i)$ occurring in A_i, and the approximate expected number $\lambda(i - .5, \hat{\theta}_\psi)|A_i|$ at A_i. We then define the robust residual

$$r_i = \Delta M(A_i, \hat{\theta}_\psi) = N(A_i) - \lambda(i - .5, \hat{\theta}_\psi)$$

and the standardized robust residual

$$r_i^* = \frac{r_i}{\sqrt{\lambda(i - .5, \hat{\theta}_\psi)}}.$$

The values $\omega_i = \omega(i - .5, \hat{\theta}_\psi)$ are, approximately, the weights assigned to $x \in A_i$. A plot of ω_i against r_i^* can show what points are being downweighed and with what standardized residual. Another way to use these weights is to obtain $v_i = \omega_i r_i^*$, a combined indicative of misfit and downweighed influence.

The first row of plots in Fig. 4 shows the robust residuals for the contaminated data using $b = .2$. The plot on the left shows the raw residuals r_i and the plot on the right shows the standardized residuals r_i^* with the index i labeling $A_i = (i-1, i)$. The raw residuals do not indicate any problem but the standardized residuals show one obvious region with large residuals.

The plots in the second row of Fig. 4 use the weights ω_i with $b = .2$. The plot on the left shows $\omega_i r_i^*$ versus the index i of $A_i = (i-1, i]$ and the plot on the right shows ω_i against r_i^*. Again, there is one subinterval with large combined residual and influence. The bottom plot is the most interesting one. It shows that there are 3 subregions being downweighed. These subregions are (80, 81], (81, 82], (83, 84] with (81, 82] receiving the smallest weight. This is the region where the spurious events lie. Although the regions (80, 81] and (83, 84] are moderately downweighed, their residuals are small hence the model is well-fitted there, as expected. This useful information is made available as a by-product of our robust estimator.

6. Proofs of asymptotic results

With $\boldsymbol{g}_t(\boldsymbol{\theta})$ defined in (3.7) we have $E_{\boldsymbol{\theta}} = \boldsymbol{g}_t(\boldsymbol{\theta}) = \mathbf{0}$ when $\int(\boldsymbol{\psi}\lambda)_{\boldsymbol{\theta}}$ is finite. Then $\text{Cov}_{\boldsymbol{\theta}}\, \boldsymbol{g}_t(\boldsymbol{\theta}) = \int_{A_t} \boldsymbol{\psi}(\boldsymbol{x},\boldsymbol{\theta})\boldsymbol{\psi}'(\boldsymbol{x},\boldsymbol{\theta})\lambda(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}$. Aitchison and Silvey (1958), who can be consulted for the proof, gave the following lemma.

LEMMA 1. *If $\boldsymbol{g}_t$ is a continuous function mapping of R^q into itself with the property that $\boldsymbol{\theta}'\boldsymbol{g}_t(\boldsymbol{\theta}) > 0$ for every $\boldsymbol{\theta}$ such that $\boldsymbol{\theta}'\boldsymbol{\theta} = 1$, then there exists a point $\hat{\boldsymbol{\theta}}$ such that $\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}} < 1$ and $\boldsymbol{g}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$.*

REGULARITY CONDITIONS.

(S) $\lambda(\boldsymbol{x},\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \Theta \subset R^q$ and $\boldsymbol{x} \in R^n$. There exist a neighborhood Θ_0 of $\boldsymbol{\theta}_0$ such that, for all $t > 0$, $\boldsymbol{\theta} \in \Theta_0$, and $\boldsymbol{x} \in R^n$, there exist the derivatives with respect to $\boldsymbol{\theta}$ of $\lambda(\boldsymbol{x},\boldsymbol{\theta})$ up to third-order and of $\boldsymbol{\psi}(\boldsymbol{x},\boldsymbol{\theta})$ up to second-order. With probability 1, the random integral $\boldsymbol{g}_t(\boldsymbol{\theta})$ may be differentiated twice with respect to $\boldsymbol{\theta} \in \Theta_0$ by interchanging the order of integration and differentiation.

(A) $\Lambda_t(\boldsymbol{\theta}_0) \equiv \int_{A_t} \lambda(\boldsymbol{x},\boldsymbol{\theta}_0)d\boldsymbol{x} \to \infty$ as $t \to \infty$.

(B) Let $0 \le \alpha_{1t} \le \cdots \le \alpha_{qt}$ be the eigenvalues of the symmetric and positive semi-definite $q \times q$ matrix $W_t(\boldsymbol{\theta}_0) \equiv \int_{A_t} \boldsymbol{\psi}(\boldsymbol{x},\boldsymbol{\theta}_0)\boldsymbol{\psi}'(\boldsymbol{x},\boldsymbol{\theta}_0)\lambda(\boldsymbol{x},\boldsymbol{\theta}_0)d\boldsymbol{x}$. We assume that $0 < \alpha_{1t}$ for all t and that $\Lambda_t^{-1}(\boldsymbol{\theta}_0)W_t(\boldsymbol{\theta}_0) \to F(\boldsymbol{\theta}_0)$, as $t \to \infty$, where $F(\boldsymbol{\theta}_0)$ is a symmetric positive definite matrix.

(C) Let $\boldsymbol{l}(\boldsymbol{x},\boldsymbol{\theta}_0)$ denote $\partial \log \lambda(\boldsymbol{x},\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and

$$C_t(\boldsymbol{\theta}_0) = \int_{A_t} \boldsymbol{\psi}(\boldsymbol{x},\boldsymbol{\theta}_0)\boldsymbol{l}'(\boldsymbol{x},\boldsymbol{\theta}_0)\lambda(\boldsymbol{x},\boldsymbol{\theta}_0)d\boldsymbol{x}.$$

We assume that $\Lambda_t^{-1}(\boldsymbol{\theta}_0)C_t(\boldsymbol{\theta}_0) \to C(\boldsymbol{\theta}_0)$, which is a positive definite matrix.

(D) Let $\boldsymbol{u}$ be a $q \times 1$ unitary vector and let $\partial\boldsymbol{\psi}_i(\boldsymbol{x},\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}$ denote $\partial\boldsymbol{\psi}_i(\boldsymbol{x},\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Furthermore, let $T_t(\boldsymbol{u})$ be the $q \times q$ symmetric semi-definite positive matrix with ij-element

$$\int_{A_t} \boldsymbol{u}'\left(\frac{\partial\boldsymbol{\psi}_i(\boldsymbol{x},\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial\boldsymbol{\psi}_j(\boldsymbol{x},\boldsymbol{\theta}_0)'}{\partial\boldsymbol{\theta}}\right)'\boldsymbol{u}\lambda(\boldsymbol{x},\boldsymbol{\theta}_0)d\boldsymbol{x}$$

and eigenvalues $\{\beta_{it}(\boldsymbol{u}); i = 1,\ldots,q\}$. If $\beta_t = \sup_{\boldsymbol{u}}\{\max_{1\le i\le q}\beta_{it}(\boldsymbol{u})\}$, then we assume that $\Lambda_t^{-2}(\boldsymbol{\theta}_0)\beta_t \to 0$ as $t \to \infty$.

(E) For all $\boldsymbol{\theta} \in \Theta_0$, we have

$$\max_{1\le i,j,k\le q} |\partial^2 \psi_k(\boldsymbol{x},\boldsymbol{\theta})/\partial\theta_i\theta_j| \le G(\boldsymbol{x})$$

with $G(\boldsymbol{x})$ satisfying

$$\lim_{t\to\infty} \int_{A_t} \Lambda_t^{-2}(\boldsymbol{\theta}_0) G^2(\boldsymbol{x})\lambda(\boldsymbol{x},\boldsymbol{\theta}_0) = 0$$

and

$$\limsup_t \int_{A_t} \Lambda_t^{-1}(\boldsymbol{\theta}_0) G(\boldsymbol{x})\lambda(\boldsymbol{x},\boldsymbol{\theta}_0) d\boldsymbol{x} = M_1 < \infty.$$

(F) For all $\boldsymbol{\theta} \in \Theta_0$ we have

$$\left| \frac{\partial^2 \psi_k \lambda}{\partial\theta_i\theta_j} \bigg|_{(\boldsymbol{x},\boldsymbol{\theta}_0)} \right| \le H(\boldsymbol{x})$$

with $H(\boldsymbol{x})$ satisfying $\limsup_t \int_{A_t} H(\boldsymbol{x})/\Lambda_t(\boldsymbol{\theta}_0) dx = M_2 < \infty$.

(G) For t sufficiently large, we can find partition $A_t = \sum_{i=1}^{n_t} A_{it}$ such that, for any fixed $\boldsymbol{v} \in R^q$ and any $\epsilon > 0$, the sequence of independent random variables $Y_{it} = \int_{A_{it}} \boldsymbol{v}'\boldsymbol{\psi}(\boldsymbol{x},\boldsymbol{\theta}_0) dM(\boldsymbol{x},\boldsymbol{\theta}_0)$ with mean zero and variance $\mathrm{Var}_{\boldsymbol{\theta}_0} Y_{it} = \int_{A_{it}} \boldsymbol{v}'\boldsymbol{\psi}(\boldsymbol{x},\boldsymbol{\theta}_0)\boldsymbol{\psi}'(\boldsymbol{x},\boldsymbol{\theta}_0)\boldsymbol{v}\lambda(\boldsymbol{x},\boldsymbol{\theta}_0) d\boldsymbol{x}$ satisfies

$$\lim_{t\to\infty} \Lambda_t^{-1}(\theta_0) \sum_{i=1}^{n_{it}} E_{\theta_0}(Y_{it}^2 I_{[Y_{it}^2 > \epsilon\Lambda_t(\theta_0)]}) \to 0.$$

Condition (S) refers to the smoothness of λ and $\boldsymbol{\psi}$. Condition (A) requires the increase of the number of events. The unbounded growth of $\Lambda_t(\boldsymbol{\theta}_0)$ guarantees that we do not stop observing events after some finite t. $W_t(\boldsymbol{\theta}_0)$ can be interpreted as the expected information about $\boldsymbol{\theta}_0$ in A_t as we will see in the study of the asymptotic distribution of $\hat{\boldsymbol{\theta}}$. Then condition (B) says that information about $\boldsymbol{\theta}$ accumulates and goes to infinity as $t \to \infty$, and that the asymptotic growth occurs at a constant rate with the increase of the expected number of events. This is a minimum requirement if asymptotic results are to be valid. Condition (C) requires $\boldsymbol{\psi}$ and $\boldsymbol{l}$ to be somewhat similar. This is also reasonable and we recall that the Hampel-optimum M-estimator of a scalar θ in an i.i.d. context is given by the score l centered and bounded below and above by constants $-b$ and b (see Hampel *et al.* (1986), p. 119). Condition (D) imposes a constraint on the accumulation rate of the derivative of $\boldsymbol{\psi}$ with respect to that of $\Lambda(\boldsymbol{\theta}_0)$. For example, in the one-dimensional case, it is satisfied if the growth of $\int \dot{\psi}^2\lambda$ is of the same order as that of $\Lambda_t(\theta_0)$, i.e., if $\int \dot{\psi}^2\lambda = O(\Lambda_t(\theta_0))$. Finally, conditions (E) and (F) impose bounds on second derivatives and condition (G) is a Lindeberg-type condition. Note that it is always possible to obtain Y_{it} with constant variance. The fact that $T_t(\boldsymbol{u})$ in condition (D) is positive semi-definite will become clear during the consistency proof.

In the MLE case, when $\boldsymbol{\psi} = \boldsymbol{l}$, we have $C(\boldsymbol{\theta}_0) = F(\boldsymbol{\theta}_0)$. Many classical models can be easily checked with respect to these conditions. For example, for one-dimensional θ, the log-linear intensity $\lambda(\boldsymbol{x}, \theta) = \exp(\boldsymbol{z}(\boldsymbol{x})\theta)$ have $\boldsymbol{l}(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{z}(\boldsymbol{x})$ and hence (D), (E), and (F) are trivially true. As with (C), integrating on A_t, we have $\int \boldsymbol{z}(\boldsymbol{x})\boldsymbol{z}(\boldsymbol{x})'\Lambda_t^{-1}(\theta_0)\lambda(\boldsymbol{x}, \theta_0)d\boldsymbol{x}$ equal to the second moment of a random variable $\boldsymbol{z}(\boldsymbol{X})$ where $\boldsymbol{X}$ is a random point chosen in A_t with density proportional to $\lambda(\boldsymbol{x}, \theta_0)$. Therefore, (C) is equivalent to require convergence of this second moment.

PROOF OF CONSISTENCY. Let $S(\boldsymbol{\theta}_0, \delta)$ be the surface of a q-dimensional sphere centered at $\boldsymbol{\theta}_0$ and with radius $\delta > 0$ arbitrarily small. Suppose we prove that

$$P(\Lambda_t^{-1}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'g_t(\boldsymbol{\theta}) < 0 \text{ for all } \boldsymbol{\theta} \in S(\boldsymbol{\theta}_0, \delta)) \to 1 \quad \text{as} \quad t \to \infty.$$

Then, for t sufficiently large, by Lemma 1, there exists $\hat{\boldsymbol{\theta}}_t$ in the interior of the sphere such that $\Lambda_t^{-1}(\boldsymbol{\theta}_0)\boldsymbol{g}_t(\hat{\boldsymbol{\theta}}_t) = \mathbf{0}$ which implies $\boldsymbol{g}_t(\hat{\boldsymbol{\theta}}_t) = \mathbf{0}$. Hence, it suffices to prove the inequality.

Define $H(\boldsymbol{\theta}) \equiv \Lambda_t^{-1}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\boldsymbol{g}_t(\boldsymbol{\theta})$. Using (A) and Taylor's expansion, we expand $\boldsymbol{g}_t(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_0$ for $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$ to obtain

$$\begin{aligned} H(\boldsymbol{\theta}) &= \Lambda_t^{-1}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\boldsymbol{g}_t(\boldsymbol{\theta}_0) + \Lambda_t^{-1}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'D\boldsymbol{g}_t(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &\quad + 0.5\Lambda_t^{-1}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'D^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\theta}^*) \\ &\equiv R_1 + R_2 + 0.5R_3. \end{aligned}$$

The $q \times q$ matrix $D\boldsymbol{g}_t(\boldsymbol{\theta}_0)$ has ij-th element given by $\partial g_{it}(\boldsymbol{\theta})/\partial\theta_j$ evaluated at $\boldsymbol{\theta}_0$. The $q \times 1$ vector $D^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ has k-th element given by $(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'D_k^2(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ where $D_k^2(\boldsymbol{\theta}^*)$ is the $q \times q$ matrix with ij-th element given by $\partial^2 g_{kt}/\partial\theta_i\partial\theta_j$ evaluated at $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^*$ is in the line connecting $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. The rest of the proof consists of controlling each of the terms R_1, R_2, and R_3, and then showing that $R_2 < 0$ dominates $R_1 + R_2 + R_3$.

For $\boldsymbol{\theta}$ on $S(\boldsymbol{\theta}_0, \delta)$, we have $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \delta\boldsymbol{u}$ where $\boldsymbol{u}$ is an unitary vector. Therefore,

$$\begin{aligned} \text{Var}_{\boldsymbol{\theta}_0}(R_1) &= \Lambda_t^{-2}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \operatorname{Cov}_{\boldsymbol{\theta}_0}(g_t(\boldsymbol{\theta}_0))(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &= \delta^2\Lambda_t^{-2}(\boldsymbol{\theta}_0)\boldsymbol{u}'W_t(\boldsymbol{\theta}_0)\boldsymbol{u} \le \delta^2\Lambda_t^{-1}(\boldsymbol{\theta}_0)\alpha_{qt} \end{aligned}$$

which goes to zero uniformily in $S(\boldsymbol{\theta}_0, \delta)$ as $t \to \infty$ by (A) and (B). Since $E_{\boldsymbol{\theta}_0}R_1 = 0$, Chebyshev's inequality gives that $R_1 \to 0$ in probability.

For the term R_2, we have

$$R_2 = \Lambda_t^{-1}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \left(\int_{A_t} \frac{\partial\boldsymbol{\psi}(x, \boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} dM(x, \boldsymbol{\theta}_0) - C_t(\boldsymbol{\theta}_0) \right) (\boldsymbol{\theta} - \boldsymbol{\theta}_0). \tag{6.1}$$

Using (C), the second term in the right hand side converges to $-(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'C(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) < 0$ if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. For $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \delta\boldsymbol{u} \in S(\boldsymbol{\theta}_0, \delta)$ we have $-(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'C(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \le -\gamma_1\delta^2 < 0$ where γ_1 is the smallest eigenvalue of $C(\boldsymbol{\theta}_0)$. The $q \times 1$ vector

$$\int_{A_t} \frac{\partial\boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)dM(\boldsymbol{x}, \boldsymbol{\theta}_0)$$

has mean zero and covariance matrix $\delta^2 T_t(\boldsymbol{u})$ with ij-th element given by

$$\int_{A_t} (\boldsymbol{\theta}-\boldsymbol{\theta}_0)' \left(\frac{\partial \boldsymbol{\psi}_i(\boldsymbol{x},\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right) \left(\frac{\partial \boldsymbol{\psi}_j(\boldsymbol{x},\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right)' (\boldsymbol{\theta}-\boldsymbol{\theta}_0)\lambda(\boldsymbol{x},\boldsymbol{\theta}_0)d\boldsymbol{x}.$$

Incidentally, this proves that $T_t(\boldsymbol{u})$ in condition (D) is a symmetric semi-definite positive matrix. Therefore, the random variable $\Lambda_t^{-1}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)' \int (\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}} dM)_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$ has mean zero and variance

$$\delta^2 \Lambda_t^{-2}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'T_t(\boldsymbol{u})(\boldsymbol{\theta}-\boldsymbol{\theta}_0) = \delta^4 \Lambda_t^{-2}(\boldsymbol{\theta}_0)\boldsymbol{u}'T_t(\boldsymbol{u})\boldsymbol{u} \le \delta^4 \beta_t \Lambda_t^{-2}(\boldsymbol{\theta}_0) \to 0$$

by condition (D). We use Chebyshev's inequality once again to obtain $R_2 = o_p(1) - \gamma_1\delta^2 < -\gamma_1\delta^2/2 < 0$ for all $\boldsymbol{\theta} \in S(\boldsymbol{\theta}_0, \delta)$ if t is large enough.

Considering R_3, we have

$$|[D_k^2(\boldsymbol{\theta}^*)]_{ij}| \le \int G(\boldsymbol{x})N(d\boldsymbol{x}) + \int H(\boldsymbol{x})d\boldsymbol{x}$$

by conditions (E) and (F). Therefore,

$$|D^2(\boldsymbol{\theta}^*)| \le \delta^2 \sum_i |\boldsymbol{u}'D_i^2(\boldsymbol{\theta}^*)\boldsymbol{u}| \le \delta^2 \left\{\int G(\boldsymbol{x})N(d\boldsymbol{x}) + \int H(\boldsymbol{x})d\boldsymbol{x}\right\}.$$

This implies that, for $\boldsymbol{\theta} \in S(\boldsymbol{\theta}_0, \delta)$,

$$\begin{aligned} |R_3| &= \frac{|(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'D^2(\boldsymbol{\theta}^*)|}{\Lambda_t(\boldsymbol{\theta}_0)} = \frac{\delta|\boldsymbol{u}||D^2(\boldsymbol{\theta}^*)|}{\Lambda_t(\boldsymbol{\theta}_0)} \\ &\le \frac{\delta^3 q}{\Lambda_t(\boldsymbol{\theta}_0)} \left\{\int G(\boldsymbol{x})N(d\boldsymbol{x}) + \int H(\boldsymbol{x})d\boldsymbol{x}\right\}. \end{aligned}$$

Using condition (E) we obtain

$$\int \frac{G(\boldsymbol{x})}{\Lambda_t(\boldsymbol{\theta}_0)} N(d\boldsymbol{x}) = \int \frac{G(\boldsymbol{x})}{\Lambda_t(\boldsymbol{\theta}_0)} dM(\boldsymbol{x},\boldsymbol{\theta}_0) + \int \frac{G(\boldsymbol{x})}{\Lambda_t(\boldsymbol{\theta}_0)} \lambda(\boldsymbol{x},\boldsymbol{\theta}_0)d\boldsymbol{x} \le o_p(1) + M_1$$

for large t since

$$\mathrm{Var}_{\boldsymbol{\theta}_0}\left(\int_{A_t} \frac{G(\boldsymbol{x})}{\Lambda_t(\boldsymbol{\theta}_0)} dM(\boldsymbol{x},\boldsymbol{\theta}_0)\right) = \int_{A_t} \frac{G^2(\boldsymbol{x})}{\Lambda_t^2(\boldsymbol{\theta}_0)} \lambda(\boldsymbol{x},\boldsymbol{\theta}_0)d\boldsymbol{x} \to 0.$$

Condition (F) gives $|\Lambda_t^{-1}(\boldsymbol{\theta}_0)\int H(\boldsymbol{x})d\boldsymbol{x}| < M_2$ for large t.

Let S_t be the event that all the three inequalities, $|R_1| < \delta^3$, $R_2 < -\alpha_1/2\delta^2$, and $|R_3| < q(M_1+M_2+1)\delta^3$, are satisfied uniformly in $S(\boldsymbol{\theta}_0, \delta)$. Then $P(S_t) > 1-\epsilon$ if t is large enough. For such t, for $\omega \in S_t$, and δ sufficiently small, $H(\boldsymbol{\theta}) < 0$ for $\boldsymbol{\theta} \in S(\boldsymbol{\theta}_0, \delta)$ and we have the desired result.

PROOF OF ASYMPTOTIC NORMALITY. Since

$$\mathbf{0} = g_t(\hat{\boldsymbol{\theta}}_t) = g_t(\boldsymbol{\theta}_0) + Dg_t(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0) + 0.5D^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\theta}^*),$$

we obtain

$$W_t^{-1/2}(\boldsymbol{\theta}_0)g_t(\boldsymbol{\theta}_0) = -W_t^{-1/2}(\boldsymbol{\theta}_0)Dg_t(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - 0.5W_t^{-1/2}(\boldsymbol{\theta}_0)D^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\theta}^*). \tag{6.2}$$

By conditions (B) and (G) and the Crámer-Wold device, the left-hand side of (6.2) converges in distribution to a $N_q(\mathbf{0}, I)$. From the control of the R_2 term on the consistency proof, we have

$$\Lambda_t^{-1}(\boldsymbol{\theta}_0)Dg_t(\boldsymbol{\theta}_0) = -C(\boldsymbol{\theta}_0) + o_p(1)$$

and hence

$$-\sqrt{\Lambda_t(\boldsymbol{\theta}_0)}W_t^{-1/2}(\boldsymbol{\theta}_0)\frac{Dg_t(\boldsymbol{\theta}_0)}{\Lambda_t(\boldsymbol{\theta}_0)} \to_p F^{-1/2}(\boldsymbol{\theta}_0)C(\boldsymbol{\theta}_0),$$

which implies that

$$-W_t^{-1/2}(\boldsymbol{\theta}_0)Dg_t(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = (F^{-1/2}(\boldsymbol{\theta}_0)C(\boldsymbol{\theta}_0) + o_p(1))\sqrt{\Lambda_t(\boldsymbol{\theta}_0)}(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0).$$

The k-th entry of $D^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\theta}^*)/\sqrt{\Lambda_t(\boldsymbol{\theta}_0)}$ can be written as

$$(\Lambda_t^{-1}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0)'D_k^2(\boldsymbol{\theta}^*))\sqrt{\Lambda_t(\boldsymbol{\theta}_0)}(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0) = o_p(1)\sqrt{\Lambda_t(\boldsymbol{\theta}_0)}(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0)$$

and hence the right-hand side of (6.2) is

$$[(F^{-1/2}(\boldsymbol{\theta}_0)C(\boldsymbol{\theta}_0) + o_p(1)) + (0.5(F^{-1/2}(\boldsymbol{\theta}_0) + o_p(1))o_p(1))]\sqrt{\Lambda_t(\boldsymbol{\theta}_0)}(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0)$$

which is

$$[F^{-1/2}(\boldsymbol{\theta}_0)C(\boldsymbol{\theta}_0) + o_p(1)]\sqrt{\Lambda_t(\boldsymbol{\theta}_0)}(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0).$$

Then, Lemma 6.4.1 from Lehmann (1983) gives

$$\sqrt{\Lambda_t(\boldsymbol{\theta}_0)}(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0) \to_d N_q(\mathbf{0}, C^{-1}(\boldsymbol{\theta}_0)F(\boldsymbol{\theta}_0)C'^{-1}(\boldsymbol{\theta}_0)). \tag{6.3}$$

Acknowledgements

This paper is based on part of the first author's Ph.D. dissertation from the Statistics Department of the University of Washington and it was completed during a visit to that University. The authors want to thank Doug Martin, Adrian Raftery and Victor Yohai for helpful discussions and the associate editor and the referees for their careful reading which substantially improved this paper. The first author was supported by CAPES-Brazil and FAPEMIG-Brazil and the second author was partially supported by EPA-Corvallis grant #CR821799-01-2. This document has not been subjected to internal EPA review and does in no way reflect official EPA policy.

REFERENCES

Aitchison, J. and Silvey, S. D. (1958). Maximum-likelihood estimation of parameters subject to restraints, *Ann. Math. Statist.*, **29**, 813–828.

Assunção, R. M. (1994). Robust estimation in point processes, Ph.D. Thesis, University of Washington, Seattle.

Berman, M. (1983). Comment on "Likelihood analysis of point processes and its applications to seismological data" by Y. Ogata, *Bulletin of the International Statistical Institute*, **50**, Book 3, 412–418.

Cressie, N. A. C. (1991). *Statistics for Spatial Data*, Wiley, New York.

Hampel, F. R. (1978). Optimally bounding the gross-error-sensitivity and the influence of position in factor space, *ASA 1978 Proceedings of Statistical Computing Section*, 59–64, ASA, Washington, D.C.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York.

Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotical Theory*, Springer, New York.

Krasker, W. S. (1980). Estimation in linear regression models with disparate data points, *Econometrica*, **48**, 1333–1346.

Krasker, W. S. and Welsch, R. E. (1982). Efficient bounded-influence regression estimation, *J. Amer. Statist. Assoc.*, **77**, 595–604.

Krickeberg, K. (1982). Processus ponctuels en statistique, *Ecole d'Ete de Probabilities de Saint-Flour X-1980* (ed. P. L. Hennequin), 205–313, Springer, Berlin.

Künsch, H. R. (1984). Infinitesimal robustness for autoregressive processes, *Ann. Statist.*, **12**, 843–863.

Künsch, H. R., Stefanski, L. A. and Carroll, R. J. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models, *J. Amer. Statist. Assoc.*, **84**, 460–466.

Kutoyants, Y. A. (1984). *Parameter Estimation for Stochastic Processes* (translated by B. L. S. Prakasa Rao), Heldermann Verlag, Berlin.

Lawson, A. (1993). A deviance residual for heterogeneous spatial Poisson process, *Biometrics*, **49**, 889–897.

Lehmann, E. L. (1983). *Theory of Point Estimation*, Wiley, New York.

Lewis, P. A. W. (1972). Recent results in the statistical analysis of univariate point processes, *Stochastic Point Processes* (ed. P. A. W. Lewis), 1–54, Wiley, New York.

Martin, R. D. and Yohai, V. (1986). Influence functionals for time series, *Ann. Statist.*, **14**, 781–818.

Martin, R. D. and Yohai, V. J. (1991). Bias robust estimation of autoregression parameters, *Directions in Robust Statistics and Diagnostics* (ed. W. Stahel and S. Weisberg), Part 1, 233–246, Springer, New York.

Martin, R. D., Yohai, V. and Zamar, R. H. (1989). Min-max bias robust regression, *Ann. Statist.*, **17**, 1608–1630.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes, *J. Amer. Statist. Assoc.*, **83**, 9–27.

Rathbun, S. L. (1996). Estimation of Poisson intensity using partially observed concomitant variables, *Biometrics*, **52**, 226–242.

Rathbun, S. L. and Cressie, N. (1994). Asymptotic properties of estimators for the parameters of a spatial inhomogeneous Poisson point process, *Advances in Applied Probability*, **26**, 122–154.

Stefanski, L. A., Carroll, R. and Ruppert, D. (1986). Optimally bounded score functions for generalized linear models with applications to logistic regression, *Biometrika*, **73**, 413–424.

Yoshida, N. and Hayashi, T. (1990). On the robust estimation in Poisson processes with periodic intensities, *Ann. Inst. Statist. Math.*, **42**, 489–507.