

**A note on the separability of multidimensional
point processes with covariates**

Frederic Paik Schoenberg

Department of Statistics, University of California, Los Angeles, CA 90095–1554, USA.

phone: 310-794-5193

fax: 310-206-5658

email: frederic@stat.ucla.edu

Postal address: UCLA Dept. of Statistics

8142 Math-Science Building

Los Angeles, CA 90095–1554, USA.

Abstract

For models used to describe multi-dimensional marked point processes with covariates, the high number of parameters typically involved and the high dimensionality of the process can make model evaluation, construction, and estimation using maximum likelihood quite difficult. Conditions are explored here under which parameters governing one set of coordinates or covariates affecting a multi-dimensional marked point process may be estimated separately. The resulting estimates are, under the given conditions, similar to maximum likelihood estimates.

Key words: maximum likelihood estimation, intensity function, weighted least squares estimation, consistency, Poisson process, conditional intensity.

1 Introduction.

Given a multi-dimensional point process, it is well-known that if the coordinates in one dimension are independently drawn from some fixed density, then the parameters governing this density may be estimated separately from the parameters governing the rest of the process. In point process models for earthquake occurrences, for instance, the distribution of earthquake magnitudes is typically modeled as constant over time, and under this assumption the estimation of the earthquake size distribution is especially straightforward. When one or more dimensions of a point process have coordinates whose entries are i.i.d. draws from a fixed distribution, the process is called *separable*; see e.g. Rathbun (1996) or Schoenberg (2004) for examples. This paper investigates more general conditions under which components of a point processes can be estimated separably.

The separability of a component in a point process model is very important in that if a parameter or collection of parameters may be estimated individually, this greatly facilitates model building, fitting, and assessment. For multi-dimensional models, each separable coordinate may be plotted individually to suggest functional forms for the model, and the fit of the model is also much more readily inspected due to the reduction in the number of dimensions. Further, while maximum likelihood estimates have well-understood properties such as consistency and asymptotic efficiency under rather general conditions, in practice maximum likelihood typically requires an iterative optimization procedure which, when many parameters are being estimated, can fail to converge to a global maximum and which often relies heavily on starting values, the choice of which can be very problematic. Estimation is greatly facilitated when only a few parameters are estimated at a time. Hence it is worth exploring situations in which point process models can be decomposed so that the certain parameters can be estimated separably, i.e. without optimizing over all values of the other parameters.

Another motivation for the study of separable estimation of point process parameters is in the case of missing covariates or dimensions. Even in point process applications where many coordinates are recorded with each point and where many covariates are analyzed, there are typically some variables that are excluded from the analysis yet which could potentially affect the estimation of the conditional intensity of the point process. It is therefore important to investigate conditions under which the omission of these variables will not substantially affect the maximum likelihood estimates of parameters in a point process model, and this is precisely what is meant by separability in the current paper.

Tests for separability of point process models have been proposed by Schoenberg (2004). Here, we focus on the estimation of separable point process models, including processes with covariates, and address the question of what types of models have components that may be estimated separately. Rathbun (1996) noted that models that are multiplicative in all dimensions may be estimated separably, and methods for estimating such models are detailed by Baddeley and Turner (2000). The present paper extends this to a much wider class of models. Our main results may roughly be summarized as follows: for models that are multiplicative in the dimensions of the point process, and either multiplicative or additive in the covariates, the individual components of the model may, under general conditions, be estimated separately. The resulting estimates will be equivalent, or in the case of Theorem 4.1 below will converge in probability, to the ordinary maximum likelihood estimates.

2 Preliminaries

Suppose N is a point process whose domain D is a measurable product space, $D = D_0 \times D_1 \times \dots \times D_k$, equipped with Lebesgue measure μ . For instance, in the case of earthquake occurrences, D might be the product of a portion of space-time and a mark space. Suppose that each of the domains D_i is measurable and is equipped with a Lebesgue measure μ_i , and that in particular $D_0 = [0, T]$ is a portion of the real (time) line.

For any point $x = (t, m_1, m_2, \dots, m_k)$ in D , let $\lambda(t, m_1, m_2, \dots, m_k)$ denote the conditional intensity of the point process. That is, following Brown et al. (1986) or Merzbach and Nualart (1986), beginning with a filtration \mathcal{F}_x on D , we define \mathcal{F}_x^1 as the filtration generated by the \mathcal{F} -adapted, left-continuous processes, and say a process is *predictable* if

it is \mathcal{F}^1 -adapted. Then the *conditional intensity* (or *1-intensity*) λ is any non-negative, \mathcal{F}^∞ -predictable process such that for any measurable subset S of $D_1 \times D_2 \times \dots \times D_k$, $N([0, t] \times S) - \int_0^t \int_S \lambda(u, m_1, m_2, \dots, m_k) d\mu_0 d\mu_1 \dots d\mu_k$ is an \mathcal{F} -martingale. Note that each coordinate m_i may be a multi-dimensional vector, or a point in the arbitrary measurable space D_i .

Suppose that λ is governed by a parameter vector θ from some compact parameter space Θ , and that Θ is a product of compact parameter spaces $\Theta_0, \Theta_1, \dots, \Theta_k, \Theta_{k+1}$. We assume in what follows that Θ_{k+1} is a compact subset of \mathbf{R}^+ , but each of the other spaces Θ_i may be multi-dimensional.

To ease notation, it will be useful to introduce the following conventions. For any integer i in $\{0, 1, \dots, k\}$, let D_{-i} represent the product space $D_0 \times D_1 \dots \times D_{i-1} \times D_{i+1} \times D_{i+2} \times \dots \times D_k$, and let μ_{-i} denote Lebesgue measure on D_{-i} . Similarly, let $m_{-i} = (m_1, m_2, \dots, m_{i-1}, m_{i+1}, m_{i+2}, \dots, m_k)$, and let θ_{-i} denote the parameter vector $\{\theta_0, \theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \theta_{i+2}, \dots, \theta_k\}$.

We say λ is *completely separable* if

$$\lambda(t, m_1, \dots, m_k; \theta) = \theta_{k+1} \lambda_0(t; \theta_0) \lambda_1(t, m_1; \theta_1) \dots \lambda_k(t, m_k; \theta_k), \quad (1)$$

where $\theta_i \in \Theta_i$, and each λ_i is \mathcal{F}^1 -predictable. θ_{k+1} simply represents a multiplicative constant; if this is not desired, Θ_{k+1} may simply be taken to be the point set 1. In some applications, it may be unreasonable to suppose that the process is completely separable. However, more generally one might suppose that a given component is separable, as in the following definition.

We say λ (or equivalently, the point process N) is *separable in mark m_i* if the 1-intensity

may be written

$$\lambda(t, m_1, \dots, m_k; \theta) = \theta_{k+1} \lambda_i(t, m_i; \theta_i) \lambda_{-i}(t, m_{-i}; \theta_{-i}). \quad (2)$$

Note that mark m_i may be multiplicative and yet may influence the conditional rates λ_i and λ_{-i} at future times and that the distribution of mark m_i may vary with t and may depend on any facets of the history of the process. The key feature in (??) is that the parameter θ_i only influences the process λ_i .

For point processes in general, the loglikelihood for the full parameter vector θ may in general be written (Daley and Vere-Jones 1988):

$$L(\theta) = \int_D \log \lambda(x; \theta) dN - \int_D \lambda(x; \theta) d\mu. \quad (3)$$

The parameter vector $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k, \hat{\theta}_{k+1})$ is called the maximum likelihood estimate (MLE) of θ .

For a point process thought to be separable in mark m_i , one may instead consider maximizing the *partial loglikelihood*

$$\tilde{L}_i(\theta_i, \theta_{k+1}) = \int_D \log[\theta_{k+1} \lambda_i(t, m_i; \theta_i)] dN - \theta_{k+1} \int_{D_0} \int_{D_i} \lambda_i(t, m_i; \theta_i) d\mu_i dt. \quad (4)$$

The parameters $\tilde{\theta}_i, \tilde{\theta}_{k+1}$ maximizing \tilde{L} may be called *partial maximum likelihood estimates* (PMLEs).

3 PMLEs for Multiplicative Models.

For processes that are separable in a certain mark, when estimating the parameters governing the component of the rate related to this mark, the MLE and PMLE are often quite

similar. Some conditions under which the two estimates are exactly the same are given in the following elementary result.

Lemma 3.1. Let N be a point process whose 1-intensity λ is separable in mark m_i as in (??). Suppose that both $L(\theta)$ and $\tilde{L}_i(\theta_i)$ are differentiable with respect to θ_i , and that $\tilde{\theta}_i$ is the unique value of θ_i satisfying $\frac{\partial \tilde{L}_i}{\partial \theta_i} = 0$. Suppose also that (at least) one of the following three conditions holds, for some scalar γ :

$$\int_{D_{-i}} \lambda_{-i}(t, m_{-i}; \theta_{-i}) d\mu_{-i} = \gamma, \forall \theta_{-i} \quad (5)$$

$$\int_{D_i} \lambda_i(t, m_i; \theta_i) d\mu_i = \gamma, \forall \theta_i \quad (6)$$

$$\int_D \lambda(t, m_1, \dots, m_k; \theta) d\mu = \theta_{k+1} \int_{D_0} \int_{D_i} \tilde{\lambda}(t, m_i; \theta_i) d\mu_i dt = \gamma, \forall \theta. \quad (7)$$

Then $\tilde{\theta}_i = \hat{\theta}_i$.

Proof.

When λ is separable in mark m_i , (??) becomes

$$\begin{aligned} \tilde{L}_i(\theta_i, \theta_{k+1}) &= N(D) \log(\theta_{k+1}) + \left[\int_D \log \lambda_i(t, m_i; \theta_i) + \log \lambda_{-i}(t, m_{-i}; \theta_{-i}) \right] dN \\ &\quad - \theta_{k+1} \int_{D_i} \lambda_i(t, m_i; \theta_i) d\mu_i \int_{D_{-i}} \lambda(t, m_{-i}; \theta_{-i}) d\mu_{-i}. \end{aligned}$$

Hence

$$\begin{aligned} 0 &= \frac{\partial L(\theta)}{\partial \theta_i} \\ &= \frac{\partial}{\partial \theta_i} \int_D \log \lambda_i(t, m_i; \theta_i) dN - \theta_{k+1} \frac{\partial}{\partial \theta_i} \int_{D_i} \lambda_i(t, m_i; \theta_i) \left[\int_{D_{-i}} \lambda_{-i}(t, m_{-i}; \theta_{-i}) d\mu_{-i} \right] d\mu_i. \end{aligned}$$

By assumption, $(\tilde{\theta}_{k+1}, \tilde{\theta}_i)$ is the unique solution to the equation

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta_i} \tilde{L}(\theta) \\ &= \frac{\partial}{\partial \theta_i} \int_D \log \lambda_i(t, m_i; \theta_i) dN - \theta_{k+1} \int_{D_0} \int_{D_i} \frac{\partial \lambda_i(t, m_i; \theta_i)}{\partial \theta_i} d\mu_i dt. \end{aligned} \quad (8)$$

Therefore under condition (??), $(\hat{\theta}_{k+1}, \hat{\theta}_i)$ is the unique solution to (??). If (??) or (??) holds, then neither $\int_{D_0} \int_{D_i} \lambda_i(t, m_i; \theta_i) d\mu_i dt$ nor $\int_{D_0} \int_{D_i} \lambda_i(t, m_i; \theta_i) \left[\int_{D_{-i}} \lambda_{-i}(t, m_{-i}; \theta_{-i}) d\mu_{-i} \right] d\mu_i dt$ depends on θ_i , so both $\tilde{\theta}_i$ and the MLE $\hat{\theta}_i$ must uniquely satisfy $\frac{\partial}{\partial \theta_i} \int_D \log \lambda_i(t, m_i; \theta_i) dN = 0$. \square

Equations (??)-(??) are not impossibly restrictive. The following three examples illustrate conditions under which these assumptions may be met.

Example 3.1. The Epidemic-Type Aftershock Sequence (ETAS) model of Ogata (1988; 1998) is a type of branching model that is widely used in seismology. The marks include the magnitudes of the earthquakes and may also include the spatial locations of the events. According to the ETAS model, the conditional rate λ is separable with respect to magnitude, and can be written $\lambda(t, m, \mathbf{x}) = \lambda_1(t, \mathbf{x})f(m)$, where $f(m)$ is the magnitude density, which is posited not to change over time. Thus the LHS of (??) becomes $\int f(m; \theta_m) dm = 1$, since f is a density. As noted in Schoenberg (2004), it is important to clarify that the magnitudes of prior events may influence the conditional intensity subsequently, but the process may nevertheless be separable in magnitude provided (??) holds, i.e. if the parameters governing the magnitude distribution do not influence the other marginal distributions of the process.

Example 3.2. In the analysis of wildfires, one important mark is the amount of area

burned, and it has often been noted that the density of area burned may change from year to year. This density (assuming it exists) may depend on the fuel age distribution and other dynamic conditions. It nevertheless must always integrate to unity, and models have been proposed (see e.g. Peng et al. 2005) which posit that the parameters governing this density do not interact with the other parameters governing the other distributions of the process in violation of (??). Hence (??) is satisfied for such models with m_i the burn area of a fire (or equivalently (??) is satisfied where m_{-i} is the burn area, and m_i contains information on all other marks).

Example 3.3. When implementing maximum likelihood estimation algorithms in practice, one must verify that the optimization routine does not converge to a local maximum. A common way of checking whether the routine's output is reasonable is by ensuring that the integral term in (??) is approximately equal to the number $N(D)$ of observed points, since $E \int_D \lambda(x; \theta) d\mu = E \int_D dN = EN(D)$. Similarly, in maximizing the partial likelihood, one would typically ensure that $\theta_{k+1} \int_{D_0} \int_{D_i} \lambda_i(t, m_i; \theta_i) d\mu_i dt$ is approximately equal to $N(D)$. If one imposes the constraint that each of these integrals must equal $N(D)$, then (??) is satisfied with $\gamma = N(D)$.

Example 3.4. In some models, spatial background rates are fitted by kernel smoothing of a certain fixed subset of n points (e.g. Ogata 1988, Schoenberg 2003), and the bandwidth of the kernel density may be estimated by maximum likelihood or by maximizing the partial likelihood governing only the spatial coordinates. In such situations, if the spatial domain

has no boundary, or if boundary effects are negligible, or if a correction is used in the fitting so that each of the n points identically contributes a value of one to the total background rate, then as in the previous example, (??) holds with $\gamma = n$.

Recall that in the parameterization of each component $\lambda_i(t, m_i; \theta_i)$, the parameter θ_i need not be a scalar, but may instead be a vector in \mathbf{R}^d . (Similarly, m_i may also be vector-valued.) Recall also that although $\lambda_i(t, m_i)$ must be \mathcal{F} -predictable, it may depend on covariates, including external observations and/or functionals of the history of the point process. We turn now to the estimation of the parameters governing the effect of these covariates on λ . The next results indicate conditions under which the parameters governing each covariate may be estimated separately.

Suppose that the parameterization of one particular component $\lambda_i(t, m_i; \theta_i)$ of the 1-intensity can be decomposed into a product of terms

$$\lambda_i(t, m_i; \theta_i) = f_1(X(t, m_i); \beta_1) f_2(Y(t, m_i); \beta_2), \quad (9)$$

where $\theta_i = (\beta_1, \beta_2)$, and X and Y are predictable processes. Such a model may arise for example when f_1 represents the effect on the rate caused by one collection of covariates, and f_2 represents the effect of another group of covariates. Note that X and Y need not be scalars, but may be vector-valued or may take values in an arbitrary measurable space.

Let $H_1(x)$, $H_2(y)$, and $H(x, y)$ denote the empirical cumulative distribution functions on D of $X(t, m_i)$, $Y(t, m_i)$, and of the pair (X, Y) , respectively. Of particular interest is the special case where H has the multiplicative form

$$H(x, y) = H_1(x)H_2(y). \quad (10)$$

Let $\check{\beta}_1$ denote the maximum likelihood estimate when the parameter (vector) β_1 is estimated separately, i.e. the value of β_1 maximizing

$$\check{L}(\beta_1) := \int_{D_0} \int_{D_i} \log[\theta_0 f_1(X(t, m_i); \beta_1)] dN(t, m_i) - \theta_0 \int_{D_0} \int_{D_i} f_1(X(t, m_i); \beta_1) d\mu_i dt. \quad (11)$$

Theorem 3.2. Suppose that the conditions of Lemma 3.1 hold and that λ_i is multiplicative as in (??). Suppose that \check{L} is differentiable with respect to β_1 , and that there exists a unique solution $(\check{\theta}_0, \check{\beta}_1)$ satisfying $\frac{d\check{L}}{d\beta_1} = 0$. If H has the multiplicative form (??), then $\check{\beta}_1$ is the MLE of β_1 .

Proof. Reparameterizing the second term in \check{L}_i , one may write

$$\begin{aligned} \theta_0 \int_{D_0} \int_{D_i} \lambda_i(t, m_i; \theta_i) d\mu_i dt &= \theta_0 \int_x \int_y f_1(x; \beta_1) f_2(y; \beta_2) dH(x, y) \\ &= \theta_0 \int_x f_1(x; \beta_1) dH_1(x) \int_y f_2(y; \beta_2) dH_2(y). \end{aligned}$$

Hence $\check{\beta}_1$ satisfies

$$\begin{aligned} 0 &= \frac{d}{d\beta_1} L(\check{\theta}_i) \\ &= \frac{d}{d\beta_1} \int_{D_0} \int_{D_i} \log f_1(X(t, m_i); \check{\beta}_1) dN(t, m_i) - \theta_0 \int_y f_2(y; \theta_2) dH_2(y) \frac{d}{d\beta_1} \int_x f_1(x; \theta_1) dH_1(x). \end{aligned}$$

One may similarly reparameterize $\check{L}(\beta_1)$ to obtain

$$\frac{d}{d\beta_1} \check{L}(\beta_1) = \frac{d}{d\beta_1} \int_{D_0} \int_{D_i} \log f_1(X(t, m_i); \beta_1) dN(t, m_i) - \theta_0 \frac{d}{d\beta_1} \int_x f_1(x; \beta_1) dH_1(x).$$

Thus $(\theta_0 \int_y f_2(y; \theta_2) dH_2(y), \check{\beta}_1)$ is the unique solution to the equation $\frac{d}{d\beta_1} \check{L}(\beta) = 0$. Therefore, using Lemma 2.1, $\hat{\beta}_1 = \check{\beta}_1 = \check{\beta}_1$. \square

Example 3.5. In the *log-linear* or *exponential* family of models considered by Baddeley and Turner (2000; 2005), the conditional rate is purely multiplicative with respect to all

marks and covariates, and thus satisfies conditions (??) and (??). According to Theorem 3.2, if two of the covariates X and Y satisfy (??), then the parameters governing their components in the conditional rate λ may equivalently be estimated separately.

4 Additive Models

The result in Theorem 3.2 may seem intuitively obvious given condition (??), but note that this condition does not necessarily imply that the effects of X and Y may be estimated separately. For additive models, for instance, this result of Theorem 3.2 does not generally hold. For a simple example, suppose that N is a 1-dimensional point process whose conditional intensity has the form $\lambda(t) = \alpha X(t) + \beta Y(t)$, and suppose that $X(t) = 1$ and $Y(t) = t$, for all t . Then (??) holds, but the estimate $\check{\beta}$ obtained by separately estimating the coordinate $f_2(Y(t)) = \beta Y(t)$ is simply the MLE of β for the model $\lambda(t) = \beta t$, which is obviously different from the MLE of β for the model $\lambda(t) = \alpha + \beta t$.

This Section explores conditions under which parameters may be estimated separately for the case of components of λ that are additive rather than multiplicative. As an alternative to the product form in (??), suppose instead that λ_i is parameterized as a sum of functions of the covariates X and Y . That is,

$$\lambda_i(t, m_i; \theta_i) = f_1(X(t, m_i); \beta_1) + f_2(Y(t, m_i); \beta_2), \quad (12)$$

where $\theta_i = (\beta_1, \beta_2)$, and X, Y are predictable processes.

Consider the maximum likelihood estimate $\hat{\beta}_1(T)$ when the parameter (vector) β_1 is estimated individually, using observations on $[0, T] \times D_1 \times \dots \times D_k$. That is, $\hat{\beta}_1(T)$ is the

value of β_1 maximizing

$$\hat{L}^{(T)}(\beta_1) := \int_0^T \int_{D_i} \log[f_1(X(t, m_i); \beta_1)] dN(t, m_i) - \int_0^T \int_{D_i} f_1(X(t, m_i); \beta_1) d\mu_i dt. \quad (13)$$

General conditions for the convergence in probability of the MLE $\hat{\theta}$ to the true parameter vector θ^* have been given by a variety of authors; see for instance Theorem 2 of Ogata (1978) for stationary one-dimensional processes, or Theorem 1 of Rathbun (1996) for more general multi-dimensional point processes. In the following result, it is assumed that N satisfies such conditions. Further conditions are provided under which the estimate $\hat{\beta}_1$ is consistent as well, as T approaches infinity.

Theorem 4.1. Suppose that N satisfies the conditions for Theorem 2 of Ogata (1978). Suppose also that N satisfies the conditions of Lemma 2.1, and that λ_i has the additive form (??), where f_1 and f_2 are continuous in β_1 and β_2 , respectively. Suppose also that $E \int \int |\lambda(t, m_i; \theta_i^*) \log \lambda(t, m_i; \theta_i)| d\mu_i dt < \infty$ and $E \int \int |\lambda(t, m_i; \theta_i^*) \log f_1(X(t, m_i); \beta_1)| d\mu_i dt < \infty$, and that there exists an open neighborhood $U = U_1 \times U_2 \times \dots \times U_k$ of the true parameter vector θ^* , where U_1 is a neighborhood of the true parameter β_1^* , such that for θ in U , $\frac{1}{T} \int_0^T \int_{D_i} f_2(Y(t, m_i); \beta_2) d\mu_i dt$ and $\frac{1}{T} \int_0^T \int_{D_i} \frac{\lambda(t, m_i; \theta_i^*) f_2(Y(t, m_i); \beta_2)}{f_1(X(t, m_i); \beta_1)} d\mu_i dt$ converge to zero in probability for θ in U . Then $\hat{\beta}_1(T)$ is a consistent estimate of β_1 .

Proof. First, note that by martingale convergence (see e.g. Theorem A3.4 of Daley and Vere-Jones 1988 or 3.3 of Lipster and Shiryaev 1977), for any $\theta_i = (\beta_1, \beta_2)$,

$$\frac{1}{T} \left[\int_0^T \int_{D_i} \log \lambda(t, m_i; \theta_i) dN(t, m_i) - \int_0^T \int_{D_i} \log \lambda(t, m_i; \theta_i) \lambda(t, m_i; \theta_i^*) d\mu_i dt \right] \longrightarrow 0 \text{ a.s.}$$

and

$$\frac{1}{T} \left[\int_0^T \int_{D_i} \log [f_1(X(t, m_i); \beta_1)] dN(t, m_i) - \int_0^T \int_{D_i} \log [f_1(X(t, m_i); \beta_1)] \lambda(t, m_i; \theta_i^*) d\mu_i dt \right] \longrightarrow 0 \text{ a.s.}$$

Thus we can write

$$\begin{aligned} \frac{\tilde{L}_i^{(T)}(\theta_i)}{T} &= \frac{1}{T} \int_0^T \int_{D_i} \log [f_1(X(t, m_i); \beta_1) + f_2(Y(t, m_i); \beta_2)] dN(m_i, t) \\ &\quad - \frac{1}{T} \int_0^T \int_{D_i} [f_1(X(t, m_i); \beta_1) + f_2(Y(t, m_i); \beta_2)] d\mu_i dt \\ &\sim \frac{1}{T} \int_0^T \int_{D_i} \log [f_1(X(t, m_i); \beta_1) + f_2(Y(t, m_i); \beta_2)] \lambda(t, m_i; \theta_i^*) d\mu_i dt \\ &\quad - \frac{1}{T} \int_0^T \int_{D_i} [f_1(X(t, m_i); \beta_1) + f_2(Y(t, m_i); \beta_2)] d\mu_i dt, \end{aligned}$$

where by $a \sim b$ we mean that $a - b$ converges to zero a.s. as $T \rightarrow \infty$.

Similarly,

$$\begin{aligned} \frac{\dot{L}^{(T)}(\beta_1)}{T} &= \frac{1}{T} \int_0^T \int_{D_i} \log [f_1(X(t, m_i); \beta_1)] dN(t, m_i) - \frac{1}{T} \int_0^T \int_{D_i} f_1(X(t, m_i); \beta_1) d\mu_i dt \\ &\sim \frac{1}{T} \int_0^T \int_{D_i} \log [f_1(X(t, m_i); \beta_1)] \lambda(t, m_i; \theta_i^*) d\mu_i dt - \frac{1}{T} \int_0^T \int_{D_i} f_1(X(t, m_i); \beta_1) d\mu_i dt. \end{aligned}$$

Hence, for $\theta \in U$, $\frac{\tilde{L}_i^{(T)}(\theta_i)}{T} - \frac{\dot{L}^{(T)}(\beta_1)}{T} =$

$$\begin{aligned} &\frac{1}{T} \int_0^T \int_{D_i} \lambda(t, m_i; \theta_i^*) [\log (f_1(X(t, m_i); \beta_1) + f_2(Y(t, m_i); \beta_2)) - \log (f_1(X(t, m_i); \beta_1))] d\mu_i dt \\ &\quad - \frac{1}{T} \int_0^T \int_{D_i} f_2(Y(t, m_i); \beta_2) d\mu_i dt + o(T). \end{aligned}$$

But by assumption, $\frac{1}{T} \int_0^T \int_{D_i} f_2(Y(t, m_i); \beta_2) d\mu_i dt \longrightarrow 0$ in probability. Furthermore, abbreviating $f_1(X(t, m_i); \beta_1)$ and $f_2(Y(t, m_i); \beta_2)$ to f_1 and f_2 , respectively, for the moment,

$$\log(f_1 + f_2) - \log(f_1) = \log\left(\frac{f_1 + f_2}{f_1}\right) \leq \frac{f_1 + f_2}{f_1} - 1 = \frac{f_2}{f_1},$$

using the well-known relation $\log(x) \leq x-1$, for positive x (see e.g. Abramowitz, 1964). Thus, since by assumption $\frac{1}{T} \int_0^T \int_{D_i} \lambda(t, m_i; \theta_i^*) f_2/f_1 d\mu_i dt$ converges to zero in probability, the same is true of $\tilde{L}_i^{(T)}(\theta_i)/T - \dot{L}^{(T)}(\beta_1)/T$ and this convergence is uniform in θ_i due to the continuity of f_1 and f_2 and the compactness of Θ_i . Thus for any $\epsilon > 0$, $|\sup_{\theta \in U} \tilde{L}_i^{(T)}(\theta_i)/T - \sup_{\theta \in U} \dot{L}^{(T)}(\beta_1)/T|$ and $|\sup_{\theta \notin U} \tilde{L}_i^{(T)}(\theta_i)/T - \sup_{\theta \notin U} \dot{L}^{(T)}(\beta_1)/T|$ are each less than $\epsilon/2$ with probability going to one, as $T \rightarrow \infty$.

By Lemma 3.1, $\tilde{\beta}_1 = \hat{\beta}_1$. By relation 3.6 of Ogata (1978), for any $\epsilon > 0$, there exists T_1 such that for $T > T_1$,

$$\sup_{\theta \in U} \tilde{L}_i^{(T)}(\theta_i) \geq \sup_{\theta \notin U} \tilde{L}_i^T(\theta_i) + \epsilon T.$$

Hence with probability going to one as $t \rightarrow \infty$,

$$\sup_{\beta_1 \in U_1} \frac{\dot{L}^{(T)}(\beta_1)}{T} - \sup_{\beta_1 \notin U_1} \frac{\dot{L}^{(T)}(\beta_1)}{T} > \sup_{\theta \in U} \frac{\tilde{L}_i^{(T)}(\theta_i)}{T} - \epsilon/2 - \sup_{\theta \notin U} \frac{\tilde{L}_i^{(T)}(\theta_i)}{T} - \epsilon/2 \geq 0. \quad \square$$

Example 4.1. The conditions on f_1 and f_2 in Theorem 4.1 may be satisfied when f_2 is small, both in absolute terms and relative to f_1 . Let f_1 and f_2 be shorthand for $f_1(X(t, m_i); \beta_1)$ and $f_2(Y(t, m_i); \beta_2)$, respectively. Suppose that, for θ in a neighborhood U of θ^* , $|\lambda|$ is bounded in absolute value by some value b with probability going to one, and that $\int_{D_i} f_2 d\mu_i$ and $\int_{D_i} f_2/f_1 d\mu_i$ converge to zero in probability. Then so do $\frac{1}{T} \int_0^T \int_{D_i} f_2 d\mu_i dt$ and $\frac{1}{T} \int_0^T \int_{D_i} \frac{\lambda(t, m_i; \theta_i^*) f_2}{f_1} d\mu_i dt$, satisfying the last conditions in Theorem 4.1.

Example 4.2. If f_1 is bounded away from zero and $|\lambda|$ is bounded above, then the conditions on f_1, f_2 in Theorem 4.1 simply amount to the convergence to zero in probability of $\frac{1}{T} \int_0^T \int_{D_i} f_2 d\mu_i dt$. In particular, if $\int_{D_i} f_2 d\mu_i \rightarrow_p 0$, then these conditions in Theorem 4.1 are trivially satisfied.

5 Discussion

While the result in Lemma 3.1 is hardly surprising, Theorems 3.2 and 4.1 imply that parameters governing individual covariates in multi-dimensional point process models may often be estimated separately. Indeed, the parameters governing a given covariate's effect on the conditional intensity will hardly be influenced by the omission of other covariates, even if these other covariates may influence the conditional intensity overall and may even interact with the given covariate in an additive or multiplicative way. The conditions in both results essentially mandate that the interactions between covariates are not too large.

As mentioned in the Introduction, these results may have implications for point process estimation. It is typically far easier (and faster) to obtain an SMLE $\check{\beta}$ or $\hat{\beta}$ than to search over values of all parameters in order to find the value $\hat{\beta}$ maximizing the full likelihood.

In addition, the results in Sections 3 and 4 may have implications for model *building* as well. It is typically extremely difficult to construct realistic models for multi-dimensional point processes with many covariates. Ideally such models should be based on well-understood physical principles and subject-matter expertise, of course. However, in some situations empirically-based models may be sought, and one method for constructing such a model would be to investigate individually the distribution of each coordinate, and the individual contribution to the conditional intensity of each (or perhaps small collections of) covariates. These marginal distributions of the process could then be estimated separately, and the parametric forms for each could readily be inspected for goodness-of-fit. The results above suggest circumstances under which a model may be thus constructed and estimated.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 9978318. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

- Abramowitz, M. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. M. Abramowitz and I. Stegun (Eds.), U.S. Government Printing Office, Washington D.C.
- Baddeley, A., and Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns (with discussion). *Australian and New Zealand Journal of Statistics*, **42**(3):283-322.
- Baddeley, A., and Turner, R. (2005). Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, **12**(6):1-42.
- Brown, T., Ivanoff, B.G., and Weber, N.C. (1986). Poisson convergence in two dimensions with application to row and column exchangeable arrays. *Stochastic Processes and their Applications* **23**, 307–318.
- Daley, D., and Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes*. Springer, New York.
- Lipster, R.S. and Shiryaev, A.N. (1977). *Statistics of Random Processes, I: General Theory*. Springer-Verlag, New York.
- Merzbach, E., and Nualart, D. (1986). A characterization of the spatial Poisson process and changing time. *Annals of Probability*, **14**, 1380–1390.
- Ogata, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Ann. Inst. Statist. Math.*, **30**, 243–261.

- Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.*, **83**, 9–27.
- Ogata, Y. (1998). Space-time point process models for earthquake occurrences. *Ann. Inst. Statist. Math.*, **50**, 379–402.
- Peng, R. D., Schoenberg, F. P., and Woods, J. (2005). A space-time conditional intensity model for evaluating a wildfire hazard index. *J. Amer. Statist. Assoc.*, **100** (469), 26–35.
- Rathbun, S.L. (1996). Asymptotic properties of the maximum likelihood estimator for spatio-temporal point processes. *J. Statist. Plan. Inf.*, **51**, 55–74.
- Schoenberg, F.P. (2003). Multi-dimensional residual analysis of point process models for earthquake occurrences. *J. Amer. Statist. Assoc.*, **98**(464), 789–795.
- Schoenberg, F.P. (2004). Testing separability in multi-dimensional point processes. *Biometrics*, **60**, 471–481.