A note on the consistent estimation of spatial-temporal point process parameters

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 spatial-temporal marked point processes with covariates, the high number of parameters typically involved can make model evaluation, construction, and estimation using maximum likelihood quite difficult. Conditions are explored here under which parameters governing a space-time marked point process may be estimated simply and consistently by maximizing a partial likelihood. The resulting estimates are, under the given conditions, similar to maximum likelihood estimates for the full model.

Key words: conditional intensity, consistency, maximum likelihood estimation, weighted least squares estimation, Poisson process, spatial-temporal point process.

1 Introduction.

Recent increases in spatial-temporal marked point process data with multiple covariates have led to the development of point process models with relatively large numbers of parameters. The high dimensionality of the models can pose problems when it comes to the improvement, evaluation, and estimation of these models. In such situations, one may initially choose to focus on a portion of the process individually, e.g. modeling first the purely temporal aspects of the process, and then adding spatial, mark, and covariate components. In seismology, for instance, a decade elapsed between the introduction of temporal marked Epidemic-Type Aftershock Sequence (ETAS) models (Ogata 1988) and the development of spatial-temporal marked versions (Ogata 1998) that are now commonly used. When examining just a portion of the process, however, one may inquire whether the modeled process may be accurately
estimated in the absence of the components being ignored. In modeling the times and marks of a spatial-temporal marked point process, for instance, it is important to determine under what conditions the spatial components of the process may be ancillary to the parameters governing the temporal and mark components, or when ignoring the spatial coordinates of the observations would substantially bias estimates of temporal and mark parameters.

In addition, even when a space-time marked point process model with covariates is estimated, it is often the case in practice that numerous covariates that might affect the process are not recorded, and the absence of these confounding factors from the model may bias conventional estimates of model parameters. Hence it is important to explore conditions under which the errors introduced by such missing covariates are substantial and when they are negligible.

The current paper explores conditions under which such partial models may be consistently estimated despite missing information. Certain special cases are well known. In point process models for earthquake occurrences, for instance, the distribution of earthquake magnitudes is typically modeled as constant over time. That is, while earthquake magnitudes can affect the times and locations of future earthquakes, the magnitude distribution of an event at a particular location and time, given that an event occurs, is typically thought to be constant. 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 process, when estimated separably
using maximum likelihood methods, yield consistent parameter estimates.

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 certain parameters can be consistently estimated separably, i.e. without optimizing over all values of the other parameters.

Tests for separability of point process models have been proposed by Schoenberg (2004) and Chang and Schoenberg (2010). 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 consistently. 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 consistently 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, $D = D_0 \times D_1 \times \ldots \times D_k$, equipped with Lebesgue measure $\mu$. 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 $\mu_i$, and that in particular $D_0 = [0, T]$ is a portion of the real (time) line.

For any point $x = (t, m_1, m_2, ..., m_k)$ in $D$, let $\lambda(t, m_1, m_2, ..., m_k)$ denote the conditional intensity of the point process. That is, following Brown et al. (1986) or Merzbach and Nualart (1986), beginning with a filtration $F_x$ on $D$, we define $F_x^1$ as the filtration generated by the $F_x$-adapted, left-continuous processes, and say a process is predictable if it is $F_x^1$-adapted. Then the conditional intensity (or 1-intensity) $\lambda$ is any non-negative, $F^\infty$-predictable process such that for any measurable subset $S$ of $D_1 \times D_2 \times \ldots \times D_k$,

$$N([0, t] \times S) \leq \int_0^t \int_S \lambda(u, m_1, m_2, ..., m_k) d\mu_0 d\mu_1 \ldots d\mu_k$$

if an $F$-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 $\lambda$ is governed by a parameter vector $\theta$ from some compact parameter space.
Θ, and that Θ is a product of compact parameter spaces Θ₀, Θ₁, . . . , Θₖ, Θₖ₊₁. We assume in what follows that Θₖ₊₁ is a compact subset of \( \mathbb{R}^+ \), but each of the other spaces Θᵢ may be multi-dimensional.

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

We say λ is completely separable if

\[
\lambda(t, m_1, \ldots, m_k; \theta) = \theta_{k+1}\lambda_0(t; \theta_0)\lambda_1(t, m_1; \theta_2) \ldots \lambda_k(t, m_k; \theta_k),
\]

where \( \theta_i \in \Theta_i \), and each \( \lambda_i \) is \( \mathcal{F}^t \)-predictable. \( \theta_{k+1} \) simply represents a multiplicative constant; if this is not desired, \( \Theta_{k+1} \) 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

\[
\lambda(t, m_1, \ldots, m_k; \theta) = \theta_{k+1}\lambda_i(t, m_i; \theta_i)\lambda_{-i}(t, m_{-i}; \theta_{-i}).
\]

Note that mark \( m_i \) may be multiplicative and yet may influence the conditional rates \( \lambda_i \) and \( \lambda_{-i} \) 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 (2) is that the parameter \( \theta_i \) only influences the process \( \lambda_i \).
For point processes in general, the loglikelihood for the full parameter vector $\theta$ 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 $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k, \hat{\theta}_{k+1})$ is called the maximum likelihood estimate (MLE) of $\theta$.

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

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

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 $\lambda$ is separable in mark $m_i$ as in (2). Suppose that both $L(\theta)$ and $\tilde{L}_i(\theta_i)$ are differentiable with respect to $\theta_i$, and that $\tilde{\theta}_i$ is the unique value of $\theta_i$ satisfying $\frac{\partial \tilde{L}_i}{\partial \theta_i} = 0$. Suppose also that (at least) one of the following
three conditions hold, for some scalar $\gamma$:

$$\int_{D_{-i}} \lambda_{-i}(t, m_{-i}; \theta_{-i})d\mu_{-i} = \gamma, \forall \theta_{-i}$$  \hspace{1cm} (5)$$

$$\int_{D_i} \lambda_i(t, m_i; \theta_i)d\mu_i = \gamma, \forall \theta_i$$ \hspace{1cm} (6)$$

$$\int_{D} \lambda(t, m_1, \ldots, m_k; \theta)d\mu = \theta_{k+1} \int_{D_0} \tilde{\lambda}(t, m_i; \theta_i)d\mu_i dt = \gamma, \forall \theta. \hspace{1cm} (7)$$

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

**Proof.**

When $\lambda$ is separable in mark $m_i$, (3) becomes

$$L_i(\theta_1, \theta_{k+1}) = N(D) \log(\theta_{k+1}) + \int_D \log \lambda_i(t, m_i; \theta_i) + \log \lambda_{-i}(t, m_{-i}; \theta_{-i}) dN$$

$$- \theta_{k+1} \int_{D_i} \lambda_i(t, m_i; \theta_i)d\mu_i \int_{D_{-i}} \lambda(t, m_{-i}; \theta_{-i})d\mu_{-i}.$$ 

Hence

$$0 = \frac{\partial L(\theta)}{\partial \theta_i}$$

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

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

$$0 = \frac{\partial L(\theta)}{\partial \theta_i}$$

$$= \frac{\partial}{\partial \theta_i} \int_D \log \lambda_i(t, m_i; \theta_i)dN - \theta_{k+1} \int_{D_i} \frac{\partial \lambda_i(t, m_i; \theta_i)}{\partial \theta_i} d\mu_i dt. \hspace{1cm} (8)$$

Therefore under condition (5), $\left(\hat{\theta}_{k+1}, \hat{\theta}_i\right)$ is the unique solution to (8). If (6) or (7) 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 $\theta_i$, so both $\hat{\theta}_i$ and the MLE $\hat{\theta}_i$ must uniquely satisfy $\frac{\partial}{\partial \theta_i} \int_D \log \lambda_i(t, m_i; \theta_i)dN = 0.$
Equations (5)-(7) 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 $\lambda$ is separable with respect to magnitude, and can be written $\lambda(t, m, x) = \lambda_1(t, x) f(m)$, where $f(m)$ is the magnitude density, which is posited not to change over time. Thus the LHS of (6) 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 (2) 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 integrates 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 (2). Hence (6) is satisfied for such models with $m_i$ the burn area of a fire (or
equivalently (5) 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 (3) 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 (7) 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 of 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, (7) holds with $\gamma = n$.

Recall that in the parameterization of each component $\lambda_i(t, m_i; \theta_i)$, the parameter $\theta_i$ need not be a scalar, but may instead be a vector in $\mathbb{R}^d$. (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 $\lambda$. 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

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

(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

$$
H(x, y) = H_1(x)H_2(y).
$$

(10)

Let $\hat{\beta}_1$ denote the maximum likelihood estimate when the parameter (vector) $\beta_1$ is estimated separately, i.e. the value of $\beta_1$ maximizing

$$
\hat{L}(\beta_1) := \int_{D_0} \int_{D_i} \log[\theta_0f_1(X(t, m_i); \beta_1)]dN(t, m_i) - \theta_0 \int_{D_0} \int_{D_i} f_1(X(t, m_i); \beta_1)d\mu_idt.
$$

(11)

**Theorem 3.2.** Suppose that the conditions of Lemma 3.1 hold and that $\lambda_i$ is multiplicative as in (9). Suppose that $\hat{L}$ is differentiable with respect to $\beta_1$, and that there exists
a unique solution \((\hat{\theta}_0, \hat{\beta}_1)\) satisfying \(\frac{d\tilde{L}}{d\beta_1} = 0\). If \(H\) has the multiplicative form (10), then \(\hat{\beta}_1\) is the MLE of \(\beta_1\).

**Proof.** Reparameterizing the second term in \(\tilde{L}_i\), one may write

\[
\theta_0 \int_{D_0} \int_{D_i} \lambda_i(t, m_i; \theta_1) d\mu_i dt = \theta_0 \int_x \int_y f_1(x; \beta_1) f_2(y; \beta_2) dH(x, y)
= \theta_0 \int_x f_1(x; \beta_1) dH_1(x) \int_y f_2(y; \beta_2) dH_2(y).
\]

Hence \(\hat{\beta}_1\) satisfies

\[
0 = \frac{d}{d\beta_1} \tilde{L}(\hat{\theta}_i)
= \frac{d}{d\beta_1} \int_{D_0} \int_{D_i} \log f_1(X(t, m_i; \beta_1)) dN(t, m_i) - \theta_0 \int_y f_2(y; \theta_2) dH_2(y) \frac{d}{d\beta_1} \int_x f_1(x; \theta_1) dH_1(x).
\]

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

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

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

**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 (2) and (9). According to Theorem 3.2, if two of the covariates \(X\) and \(Y\) satisfy (10), then the parameters governing their components in the conditional rate \(\lambda\) may equivalently be estimated separately.
4 Additive Models

The result in Theorem 3.2 may seem intuitively obvious given condition (10), 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 (10) holds, but the estimate \( \hat{\beta} \) obtained by separately estimating the coordinate \( f_2(Y(t)) = \beta Y(t) \) is simply the MLE of \( \beta \) for the model \( \lambda(t) = \beta t \), which is obviously different from the MLE of \( \beta \) 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 \( \lambda \) that are additive rather than multiplicative. As an alternative to the product form in (9), suppose instead that \( \lambda_i \) is parameterized as a sum of functions of the covariates \( X \) and \( Y \). That is,

\[
\lambda_i(t, m_i; \theta_i) = f_1(X(t, m_i); \beta_1) + f_2(Y(t, m_i); \beta_2),
\]

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) \( \beta_1 \) is estimated individually, using observations on \([0, T] \times D_1 \times \ldots \times D_k\). That is, \( \hat{\beta}_1(T) \) is the value of \( \beta_1 \) maximizing

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

General conditions for the convergence in probability of the MLE \( \hat{\theta} \) to the true parameter vector \( \theta^{*} \) 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 \( \lambda_i \) has the additive form (12), where \( f_1 \) and \( f_2 \) are continuous in \( \beta_1 \) and \( \beta_2 \), respectively. Suppose also that

\[
1_T \int \int \int \log \left[ f_1(X(t, m_i); \beta_1) + f_2(Y(t, m_i); \beta_2) \right] dN(t, m_i) \to 0 \quad a.s.
\]

Thus we can write

\[
\frac{\tilde{L}_i(T)(\theta_i)}{T} = \frac{1}{T} \int \int \log \left[ f_1(X(t, m_i); \beta_1) + f_2(Y(t, m_i); \beta_2) \right] dN(m_i, t)
\]
\[- \frac{1}{T} \int_0^T \left[ f_1(X(t, m_i); \beta_1) + f_2(Y(t, m_i); \beta_2) \right] d\mu_i dt \]
\sim \frac{1}{T} \int_0^T \left[ \log \left[ f_1(X(t, m_i); \beta_1) + f_2(Y(t, m_i); \beta_2) \right] \lambda(t, m_i; \theta_i^*) d\mu_i dt \right. \\
\left. - \frac{1}{T} \int_0^T \left[ f_1(X(t, m_i); \beta_1) + f_2(Y(t, m_i); \beta_2) \right] d\mu_i dt, \]

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

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

Hence, for \( \theta \in U \), \( \frac{\hat{L}^{(T)}(\theta_i)}{T} - \frac{\hat{L}^{(T)}(\beta_1)}{T} = \)
\[
\frac{1}{T} \int_0^T \int_{D_i} \lambda(t, m_i; \theta_i^*) [\log (f_1(X(t, m_i); \beta_1) + f_2(Y(t, m_i); \beta_2)) - \log (f_1(X(t, m_i); \beta_1))] d\mu_i dt \\
- \frac{1}{T} \int_0^T \int_{D_i} f_2(Y(t, m_i); \beta_2) d\mu_i dt + o(T).
\]

But by assumption, \( \frac{1}{T} \int_0^T \int_{D_i} f_2(Y(t, m_i); \beta_2) d\mu_i dt \to 0 \) 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 \( \hat{L}_i^{(T)}(\theta_i)/T - \hat{L}_i^{(T)}(\beta_1)/T \) and this convergence is uniform in \( \theta_i \) due to the continuity of \( f_1 \) and \( f_2 \) and the compactness of \( \Theta_i \). Thus for any \( \epsilon > 0 \), \( \sup_{\theta \in U} \hat{L}_i^{(T)}(\theta_i)/T - \sup_{\theta \in U} \hat{L}_i^{(T)}(\beta_1)/T \)
and \( \sup_{\theta \in U} \tilde{L}_i(T)(\theta_i)/T - \sup_{\tilde{\theta} \in U} \hat{L}_i(T)(\tilde{\theta}_1)/T \) are each less than \( \epsilon/2 \) with probability going to one, as \( T \to \infty \).

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 \( T > T_1 \),

\[
\sup_{\theta \in U} \tilde{L}_i(T)(\theta_i) \geq \sup_{\tilde{\theta} \in U} \hat{L}_i(T)(\tilde{\theta}_1) + \epsilon T.
\]

Hence with probability going to one as \( t \to \infty \),

\[
\sup_{\beta_1 \in U_1} \frac{\dot{L}_i(T)(\beta_1)}{T} - \sup_{\beta_1 \not\in U_1} \frac{\dot{L}_i(T)(\beta_1)}{T} > \epsilon/2 - \sup_{\theta \in U} \frac{\tilde{L}_i(T)(\theta_i)}{T} - \epsilon/2 \geq 0.
\]

\[\square\]

**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 \( \theta \) in a neighborhood \( U \) of \( \theta^* \), \( |\lambda| \) is bounded in absolute value by some value \( b \) with probability going to one, and that \( \int f_2 d\mu_i \) and \( \int f_2/f_1 d\mu_i \) converge to zero in probability. Then so do \( \frac{1}{T} \int_0^T \int_{D_i} f_2 d\mu_i dt \) and \( \frac{1}{T} \int_0^T \frac{\lambda(t, m_i; \theta^*_i)f_2}{f_1} d\mu_i dt \), 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 \( \int f_2 d\mu_i \to_p 0 \), 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 \( \hat{\beta} \) 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


