Bernoulli **6**(5), 2000, 783–808

Time-invariance estimating equations

A.J. BADDELEY

Department of Mathematics & Statistics, University of Western Australia, Nedlands WA 6907, Australia. E-mail: adrian@maths.uwa.edu.au

We describe a general method for deriving estimators of the parameter of a statistical model, with particular relevance to highly structured stochastic systems such as spatial random processes and 'graphical' conditional independence models. The method is based on representing the stochastic model $\mathbf{X}$ as the equilibrium distribution of an auxiliary Markov process $\mathbf{Y} = (Y_t,\ t > 0)$ where the discrete or continuous 'time' index t is to be understood as a fictional extra dimension added to the original setting. The parameter estimate $\hat{\theta}$ is obtained by equating to zero the generator of $\mathbf{Y}$ applied to a suitable statistic and evaluated at the data $\mathbf{x}$. This produces an unbiased estimating equation for θ. Natural special cases include maximum likelihood, the method of moments, the reduced sample estimator in survival analysis, the maximum pseudolikelihood estimator for random fields and for point processes, the Takacs–Fiksel method for point processes, 'variational' estimators for random fields and multivariate distributions, and many standard estimators in stochastic geometry. The approach has some affinity with the Stein–Chen method for distributional approximation.

Keywords: censored data; conditional intensity; dead leaves model; diffusions; generator; Gibbs point processes; Gibbs random fields; Gibbs sampler; Godambe optimality; highly structured stochastic systems; infinitesimal generator; Markov random fields; maximum likelihood; maximum pseudolikelihood; method of moments; Nguyen–Zessin formula; pseudolikelihood; reduced sample estimator; spatial birth-and-death processes; Stein–Chen method; Takacs–Fiksel method; unbiased estimating equations; variational estimators

1. Introduction

This paper describes a general method of deriving parameter estimators for statistical models. It has particular relevance to highly structured stochastic models, such as spatial processes and graphical dependence models, where there is considerable interest in finding alternatives to maximum likelihood. In such contexts the likelihood is usually not known analytically, the maximum likelihood estimator may not be optimal, and sampling distributions and moments are often unknown.

For discrete Gibbs random fields, Besag (1975) proposed inference based on *pseudolikelihood*, a product of certain conditional likelihoods, which can be motivated by the dependence structure of the model. The pseudolikelihood approach was extended to spatial point processes by Besag (1977) and Jensen and Møller (1991), and a limit theorem relating these two cases was found in Besag *et al.* (1982). Takacs (1983; 1986) and Fiksel (1984; 1988) developed a completely different rationale for parameter estimation in point processes, based on equating unbiased estimators of the left- and right-hand sides of an identity (Nguyen and Zessin 1976) for the expectation of an arbitrary functional h of the

1350–7265 © 2000 ISI/BS

process. It has been shown (Diggle *et al.* 1994; Jensen and Møller 1991) that the Takacs–Fiksel method coincides with maximum pseudolikelihood for a large class of Gibbs point process models with the appropriate choice of h. For real-valued Gibbs random fields, Almeida and Gidas (1993) recently proposed a new class of estimators based on variational methods.

In all of the above-mentioned problems, it is natural and convenient to express the random process of interest, X, as the equilibrium distribution of an associated Markov process $\mathbf{Y} = (Y_t,\ t > 0)$ in discrete or continuous time. For example, the Poisson distribution is the equilibrium measure of a birth-and-death process on the non-negative integers. Under suitable conditions, a discrete random field is the equilibrium distribution of an associated Gibbs sampler (Geman and Geman 1984; Geman 1990); a finite spatial point process is the equilibrium distribution of a certain spatial birth-and-death process (Preston 1975; Møller 1989). Note that the 'time' index t is not part of the original formulation of the model, and may or may not have a simple interpretation in the context where the model is applied.

A standard result of Markov process theory asserts that if X is drawn from the equilibrium distribution of $\mathbf{Y}$, then

$$\mathbb{E}[(\mathscr{A}S)(X)] = 0 \tag{1}$$

for essentially all statistics $S(X)$, where $\mathscr{A}$ is the *generator* of $\mathbf{Y}$. For example, if $\mathbf{Y}$ is a Markov chain on a finite state space $\mathscr{X}$ with time-homogeneous transition probabilities $p(\mathbf{x}, \mathbf{y})$, its generator $\mathscr{A}$ is the operator defined on all functions $S : \mathscr{X} \to \mathbb{R}$ by

$$\mathscr{A}S(\mathbf{x}) = \sum_{\mathbf{y}\in\mathscr{X}} p(\mathbf{x}, \mathbf{y})[S(\mathbf{y}) - S(\mathbf{x})] \tag{2}$$

for all $\mathbf{x} \in \mathscr{X}$. Equation (1) is then straightforward.

In this paper we propose estimating the parameter θ of a given stochastic model as follows. For each θ, represent the distribution of X under θ as the equilibrium distribution of some $\mathbf{Y}^{(\theta)} = (Y_t^{(\theta)})$. Let $\mathscr{A}_\theta$ be the generator of $\mathbf{Y}^{(\theta)}$. Choose a statistic $S = S(X)$. Given the observed data $\mathbf{x}$, estimate θ as the solution $\hat{\theta}_T$ of

$$(\mathscr{A}_\theta S)(\mathbf{x}) = 0. \tag{3}$$

By (1), this is an unbiased estimating equation for θ.

The method can be applied in considerable generality and includes the method of moments and the maximum likelihood estimating equations as special cases arising from different choices of $\mathbf{Y}$. In this paper we apply the method to discrete random fields, spatial point processes, censored data, and the 'dead leaves model' from stochastic geometry. For discrete Markov random fields, if $\mathbf{Y}$ is the Gibbs sampler then we obtain the maximum pseudolikelihood estimator. For Markov point processes, if $\mathbf{Y}$ is the standard Gibbs sampler spatial birth-and-death process, then we obtain the Takacs–Fiksel method. For real-valued Markov random fields on a finite graph, if $\mathbf{Y}$ is a Langevin diffusion we obtain one of the Almeida–Gidas variational estimators. For random right-censored lifetime data, if $\mathbf{Y}$ is a chain which at each step randomly selects one of the observations and replaces it by a random sample from the true lifetime distribution F, we obtain the reduced-sample estimator of F. In each example, other choices of $\mathbf{Y}$ produce alternative estimators which may also be of interest.

In point process applications, this approach provides an independent explanation for the agreement between the Takacs–Fiksel and pseudolikelihood methods. It also appears to remove some of the arbitrariness encountered in the Takacs–Fiksel method, since the estimators usually favoured in applications are obtained by applying our method to the canonical sufficient statistic.

Regarding the statistical performance of these estimators, such as their consistency, asymptotic normality, and efficiency, unfortunately little can be said at this level of generality. It is also unclear how to select $\mathbf{Y}$ and S to obtain an optimal estimator $\hat{\theta}_T$. We investigate one very specific example.

The time-invariance approach can perhaps best be regarded as a useful way of generating a variety of candidate estimators for further study. In many applications the maximum likelihood estimator is intractable or requires unacceptably complex computation. An advantage of the time-invariance approach is that the computational complexity of the estimator is controlled by the choice of $\mathscr{A}$. For example, if $\mathbf{Y}$ is a pure jump process then the estimating function will be a sum or integral of terms of the form $S(\mathbf{y}) - S(\mathbf{x})$ for all possible jumps $\mathbf{x} \rightsquigarrow \mathbf{y}$. This may also be interpreted as a choice about the extent to which 'global' information should be incorporated in the estimating equation, echoing the arguments of Besag (1986). Another rationale for making particular choices of $\mathbf{Y}$ and S is to regard the equation $\mathscr{A} = 0$ as a first-order approximation to $\mathrm{e}^{t\mathscr{A}} - I = 0$, whose limit as $t \to \infty$ is the maximum likelihood normal equation.

The identity (1) is fundamental to the Stein–Chen method of distributional approximation (see for example Arratia *et al.* 1990; Barbour 1997; Barbour *et al.* 1992; Stein, 1986). Further remarks about this connection are made in the Discussion.

In the next section we give a general statement of the method, followed in Section 3 by two simple examples. Section 4 investigates the case of discrete (Markov) random fields and Section 5 finite (Markov) point processes. Section 6 discusses variational estimators. An application to survival analysis is described in Section 7. An example from stochastic geometry, the dead leaves model, is examined in Section 8. Section 9 discusses performance issues such as consistency and asymptotic normality, and Section 10 the selection of an optimal estimating equation, although little can be said about these issues at this level of generality. We conclude in Section 11 with a discussion of problems with the method and possibilities for further development.

2. General statement of method

Consider a parametric statistical model given by a family of probability distributions $\{P_\theta : \theta \in \Theta\}$ on a sample space $\mathscr{X}$ with arbitrary parameter space Θ. We assume $\mathscr{X}$ is a locally compact metric space. Typically, but not always, Θ and $\mathscr{X}$ are subsets of $\mathbb{R}^n$ and $\{P_\theta : \theta \in \Theta\}$ is an exponential family. The aim is to estimate the unknown parameter θ from a single observation $\mathbf{x}$ drawn from P_θ.

Our proposal is to find a time-homogeneous Markov process $\mathbf{Y}^{(\theta)} = (Y_t^{(\theta)}, t > 0)$, in discrete or continuous time, with states in $\mathscr{X}$, for which P_θ is an equilibrium distribution (for each $\theta \in \Theta$). Note again that $\mathbf{Y}^{(\theta)}$ is a mathematical fiction here; we do not need to simulate it, nor do we require any sample path properties.

Let $\mathscr{A}_\theta$ be the *generator* of $\mathbf{Y}^{(\theta)}$, an operator on functions $h : \mathscr{X} \to \mathbb{R}^k$ defined briefly as follows. In discrete time, set

$$(\mathscr{A}_\theta h)(\mathbf{x}) = \mathbb{E}[h(Y_{n+1}) - h(Y_n)|Y_n = \mathbf{x}]$$
$$= \mathbb{E}[h(Y_{n+1})|Y_n = \mathbf{x}] - h(\mathbf{x}) \tag{4}$$

for $\mathbf{x} \in \mathscr{X}$, where n is arbitrary and h must be integrable with respect to P_θ for all $\theta \in \Theta$. In continuous time,

$$(\mathscr{A}_\theta h)(\mathbf{x}) = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \mathbb{E}[h(Y_t)|Y_0 = \mathbf{x}] \tag{5}$$

$$= \lim_{t\downarrow 0} \frac{\mathbb{E}[h(Y_t)|Y_0 = \mathbf{x}] - h(\mathbf{x})}{t} \tag{6}$$

where the domain $D_{\mathscr{A}_\theta}$ of the generator consists of all bounded continuous $h : \mathscr{X} \to \mathbb{R}^k$ such that the limit (6) exists in the sense of uniform convergence, and the limit $\mathscr{A}_\theta h$ is bounded and continuous. See, for example, Karlin and Taylor (1981, p. 294) or Ethier and Kurtz (1986, pp. 9, 239). (It may be necessary to extend the domain in some cases, for example, to include all functions h which are L^p integrable with respect to P_θ for all θ, for some $p > 1$.)

For example, if $\mathbf{Y}^{(\theta)}$ is a continuous-time, pure jump process with transition kernel κ_θ then (Ethier and Kurtz 1986, pp. 162, 376; Kallenberg 1997, p. 314)

$$(\mathscr{A}_\theta f)(\mathbf{x}) = \int_{\mathscr{X}} [f(\mathbf{y}) - f(\mathbf{x})]\kappa_\theta(\mathbf{x}, \mathrm{d}\mathbf{y}). \tag{7}$$

Definition 1. *Choose a statistic $S : \mathscr{X} \to \mathbb{R}$ belonging to the domain of $\mathscr{A}_\theta$ for all $\theta \in \Theta$. Given observation of the data* $\mathbf{x}$, *estimate θ by the solution $\hat{\theta}_T$ of*

$$(\mathscr{A}_\theta S)(\mathbf{x}) = 0, \tag{8}$$

provided this exists and is unique. We call (8) *the time-invariance estimating equation and $\hat{\theta}_T$ the time-invariance estimator of θ. Note that these depend on the choice of S and* $\mathbf{Y}$.

Some examples are described below.

Lemma 1. *The equilibrium distribution P_θ and the generator $\mathscr{A}_\theta$ satisfy*

$$\mathbb{E}_\theta(\mathscr{A}_\theta S)(X) = 0 \tag{9}$$

where X has distribution P_θ. Thus (8) *is an unbiased estimating equation* (*MacLeish and Small 1988*).

The proof is standard, but included for clarity.

Proof. In the discrete case the result is trivial. In the continuous case, let $(T_t^{(\theta)}, t \geqslant 0)$ be the transition semigroup of **Y**,

$$(T_t^{(\theta)} S)(\mathbf{x}) = \mathbb{E}_\theta[S(Y_t) | Y_0 = \mathbf{x}] \tag{10}$$

for $t \geqslant 0$ and $\mathbf{x} \in \mathscr{X}$. Let X have distribution P_θ; then since P_θ is an equilibrium distribution of $\mathbf{Y}^{(\theta)}$ we have

$$\mathbb{E}_\theta S(X) = \mathbb{E}_\theta (T_t^{(\theta)} S)(X). \tag{11}$$

Now (Ethier and Kurtz 1986, pp. 9, 239)

$$\begin{aligned}
\mathbb{E}_\theta(\mathscr{A}_\theta S)(X) &= \int_{\mathscr{X}} (\mathscr{A}_\theta S)(\mathbf{x}) \mathrm{d}P_\theta(\mathbf{x}) \\
&= \int_{\mathscr{X}} \lim_{t \downarrow 0} \frac{(T_t^{(\theta)} S)(\mathbf{x}) - S(\mathbf{x})}{t} \, \mathrm{d}P_\theta(\mathbf{x}) \\
&= \lim_{t \downarrow 0} \frac{1}{t} \int_{\mathscr{X}} [(T_t^{(\theta)} S)(\mathbf{x}) - S(\mathbf{x})] \, \mathrm{d}P_\theta(\mathbf{x})
\end{aligned}$$

since the convergence in (6) is uniform. But by (11), the integral in the last expression is zero for all t, so we obtain the result. □

The estimator $\hat{\theta}_T$ depends on the choice of **Y** and of S. It is, however, invariant under random (data-dependent) time changes of **Y**. The estimator need not be a function of the statistic S chosen, and, in particular, need not be a function of the sufficient statistic even if S is sufficient.

3. Simple examples

Two simple examples will be given to clarify the idea.

3.1. Poisson distribution

Suppose the data consist of a single observation x of an integer random variable X with the Poisson(λ) distribution, $\lambda > 0$ unknown. Thus $\mathscr{X} = \mathbb{N}_0 = \{0, 1, 2, \ldots\}$ and $\Theta = (0, \infty)$.

The Poisson distribution can be represented in various ways as the equilibrium distribution of a birth-and-death process (Y_t) on the non-negative integers. Consider the standard immigration–death process on $\mathbb{N}_0$ in continuous time, with transition rates

$$\begin{aligned}
r(x, x+1) &= \lambda \\
r(x, x-1) &= x \qquad (x \geqslant 1)
\end{aligned}$$

and $r(x, y) = 0$ otherwise. This satisfies the detailed balance condition

$$p(x)r(x, y) = p(y)r(y, x), \qquad \forall x, y, \tag{12}$$

where $p(x) = e^{-\lambda}\lambda^x/x!$, so that (Kelly 1979, Theorem 1.3) the process $\mathbf{Y}$ has unique equilibrium distribution $(p(x), x \in \mathbb{N}_0)$. Its infinitesimal generator is, from (7),

$$(\mathscr{A}_\lambda S)(x) = \lambda[S(x+1) - S(x)] + x[S(x-1) - S(x)]$$

(interpreting the second term as 0 when $x = 0$), defined for any $S : \mathbb{N}_0 \to \mathbb{R}$. Choose $S(x) \equiv x$; then

$$(\mathscr{A}_\lambda S)(x) = \lambda - x.$$

Setting this to zero, the time-invariance estimator of λ is

$$\hat{\lambda}_T = x,$$

which is also the maximum likelihood and method-of-moments estimator.

3.2. Method of moments

Let $\mathscr{X}$, Θ and $\{P_\theta : \theta \in \mathscr{X}\}$ be arbitrary. Let $Y_n^{(\theta)}$, $n = 1, 2, \ldots,$ be independent and identically distributed samples from the distribution P_θ. For an arbitrary statistic S with finite expectation under P_θ for all $\theta \in \Theta$, the generator is well defined and equals

$$(\mathscr{A}_\theta S)(\mathbf{x}) = \mathbb{E}_\theta[S(X)] - S(\mathbf{x}).$$

Hence the time-invariance estimator $\hat{\theta}_T$ is the solution of

$$\mathbb{E}_\theta S(X) = S(\mathbf{x}),$$

that is, the time-invariance approach yields the method of moments.

4. Discrete Markov random fields

In this section we study discrete random fields $X = (X_i, i \in G)$ where the set of 'sites' G is an arbitrary finite set and the site 'labels' X_i take values in an arbitrary finite set L. The sample space $\mathscr{X}$ is the finite set L^G of all functions from G to L. Let π_θ be the distribution of X,

$$\mathbb{P}\{X = \mathbf{x}\} = \pi_\theta(\mathbf{x}), \qquad \mathbf{x} \in \mathscr{X}, \tag{13}$$

and assume $\pi_\theta(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathscr{X}$, $\theta \in \Theta$. Of particular interest are one-parameter Gibbs models (see, for example, Georgii 1988; Guyon 1996; Ripley 1989) of the form

$$\pi_\beta(\mathbf{x}) = \frac{1}{Z(\beta)} \exp\{-\beta V(\mathbf{x})\}, \qquad \mathbf{x} \in \mathscr{X}, \tag{14}$$

where $\beta \in [0, \infty)$ is the parameter, $V : \mathscr{X} \to [0, \infty)$ is the potential function, and $Z(\beta)$ the normalizing constant. Maximum likelihood for Gibbs models typically requires numerical computation of $Z(\beta)$ because this is not known analytically.

4.1. Maximum pseudolikelihood estimator

Besag (1975) defined the *pseudolikelihood* of a discrete random field by

$$\mathrm{PL}(\theta;\, \mathbf{x}) = \prod_{i\in G} P_\theta\{X_i = x_i | X_{G\setminus i} = \mathbf{x}_{G\setminus i}\} \tag{15}$$

where $X_B = (X_i,\, i \in B)$ denotes the restriction of X to $B \subset G$. The maximum pseudolikelihood estimator of θ is the value $\hat{\theta}_{\mathrm{MPL}}$ maximizing $\mathrm{PL}(\theta;\, \mathbf{x})$.

For a general random field (13) we have

$$\mathrm{PL}(\theta;\, \mathbf{x}) = \prod_{i\in G} \frac{\pi_\theta(\mathbf{x})}{\sum_{a\in L}\pi_\theta(F_i^a\mathbf{x})} \tag{16}$$

where $F_i^a : \mathscr{X} \to \mathscr{X}$ is the operator that resets the value at site $i \in G$ to be $a \in L$: if $F_i^a\mathbf{x} = \mathbf{y}$ then $y_j = x_j$ for all $j \neq i$ and $y_i = a$.

In a one-parameter Gibbs model (14),

$$\begin{aligned}\frac{\partial}{\partial\beta}\log \mathrm{PL}(\beta;\, \mathbf{x}) &= \sum_{i\in G}\left[-V(\mathbf{x}) - \frac{\sum_{a\in L}(-V(F_i^a\mathbf{x}))\mathrm{e}^{-\beta V(F_i^a\mathbf{x})}}{\sum_{a\in L}\mathrm{e}^{-\beta V(F_i^a\mathbf{x})}}\right]\\ &= \sum_{i\in G}\frac{\sum_{a\in L}\mathrm{e}^{-\beta V(F_i^a\mathbf{x})}[V(F_i^a\mathbf{x}) - V(\mathbf{x})]}{\sum_{a\in L}\mathrm{e}^{-\beta V(F_i^a\mathbf{x})}}\end{aligned}$$

so that if $\mathrm{PL}(\beta;\, \mathbf{x})$ attains its maximum at a zero of the derivative, the maximum pseudolikelihood estimator satisfies

$$\sum_{i\in G}\sum_{a\in L} V(F_i^a\mathbf{x})P_\beta\{X_i = a | X_{G\setminus i} = \mathbf{x}_{G\setminus i}\} = |G|\; V(\mathbf{x}),$$

that is, the maximum pseudolikelihood estimating equations for a one-parameter Gibbs model are

$$\frac{1}{|G|}\sum_{i\in G}\mathbb{E}_\beta[V(X)|X_{G\setminus i} = \mathbf{x}_{G\setminus i}] = V(\mathbf{x}). \tag{17}$$

See Guyon (1996) for further information.

4.2. Time-invariance estimator

Let $\mathbf{Y} = (Y_t^{(\theta)},\, t > 0)$ be the discrete-time Gibbs sampler for the random field distribution π_θ. Thus $\mathbf{Y}$ has states in $\mathscr{X}$ and each transition alters a single site only. The transition probabilities are

$$p(\mathbf{x},\, F_i^a\mathbf{x}) = \frac{\pi_\theta(F_i^a\mathbf{x})}{\sum_{b\in L}\pi_\theta(F_i^b\mathbf{x})} = P_\theta\{X_i = a | X_{G\setminus i} = \mathbf{x}_{G\setminus i}\}, \qquad a \neq x_i,$$

and $p(\mathbf{x},\, \mathbf{y}) = 0$ otherwise. The process is in detailed balance (12) with π_θ by Bayes's theorem, regardless of the form of π_θ, provided $\pi_\theta(\cdot) > 0$. Hence π_θ is an equilibrium distribution.

The generator of $\mathbf{Y}$ is, from (2),

$$\begin{aligned}(\mathscr{A}S)(\mathbf{x}) &= \sum_{i\in G}\sum_{a\in L} p(\mathbf{x},\, F_i^a\mathbf{x})[S(F_i^a\mathbf{x}) - S(\mathbf{x})] \\ &= \sum_{i\in G}\frac{1}{\sum_{b\in L}\pi_\theta(F_i^b\mathbf{x})}\sum_{a\in L}\pi_\theta(F_i^a\mathbf{x})[S(F_i^a\mathbf{x}) - S(\mathbf{x})] \\ &= \sum_{i\in G}\left[\frac{\sum_{a\in L}\pi_\theta(F_i^a\mathbf{x})S(F_i^a\mathbf{x})}{\sum_{a\in L}\pi_\theta(F_i^a\mathbf{x})} - S(\mathbf{x})\right] \\ &= \sum_{i\in G}(\mathbb{E}_\theta[S(X)|X_{G\setminus i} = \mathbf{x}_{G\setminus i}] - S(\mathbf{x})).\end{aligned}$$

Hence the time-invariance estimator $\hat{\theta}_T$ is the solution of

$$\frac{1}{|G|}\sum_{i\in G}\mathbb{E}_\theta[S(X)|X_{G\setminus i} = \mathbf{x}_{G\setminus i}] = S(\mathbf{x}). \tag{18}$$

This coincides with the normal equations for the maximum pseudolikelihood estimator (17) for a one-parameter Gibbs process (14), if we choose $S \equiv V$. More generally the following result holds.

Proposition 1. *Let X be a random field on a finite set of sites G, with strictly positive distribution π_θ, $\theta \in \Theta$. Suppose the distribution is an exponential family with canonical parameter θ and canonical sufficient statistic V. Let $\mathbf{Y}^{(\theta)}$ be the associated Gibbs sampler. Then the time-invariance estimating equation for θ derived from $\mathbf{Y}$ and V coincides with the maximum pseudolikelihood normal equations for θ.*

Other choices of $\mathbf{Y}$ lead to different estimators which may be of interest. For the one-parameter Gibbs model, consider the continuous-time pure jump Markov process on $\mathscr{X}$ with transition rates

$$r(\mathbf{x},\, F_i^a\mathbf{x}) = \mathrm{e}^{-\beta V(F_i^a\mathbf{x})}, \qquad \text{if } a \neq x_i;$$

this is clearly in detailed balance (12) with the distribution (14). The associated estimator of β is the solution of

$$\sum_{i\in G}\sum_{a\neq x_i}\mathrm{e}^{\beta V(F_i^a\mathbf{x})}[V(\mathbf{x}) - V(F_i^a\mathbf{x})] = 0,$$

or equivalently

$$\sum_{i\in G}\sum_{a\neq x_i}\mathrm{e}^{-\beta(V(\mathbf{x}) - V(F_i^a\mathbf{x}))}[V(\mathbf{x}) - V(F_i^a\mathbf{x})] = 0. \tag{19}$$

The author does not know whether (19) has been studied in the literature.

It is also possible to derive 'coding' estimators (Besag 1974, 1986) by modifying **Y**.

5. Finite point processes

Let X be a finite simple point process (Daley and Vere-Jones 1988) in a compact region $W \subset \mathbb{R}^d$. A realization of X can be regarded as an unordered set $\mathbf{x} = \{x_1, \ldots, x_n\}$ of points $x_i \in W$, where $n \geqslant 0$. Let $\mathscr{X}$ denote the space of all possible realizations – that is, the exponential space (Carter and Prenter 1972). Assume X has probability density f_θ with respect to the unit-rate Poisson process on W, with $f_\theta(\mathbf{x}) > 0$ for all $x \in \mathscr{X}$, $\theta \in \Theta$, where the parameter space Θ is arbitrary.

As a concrete example we consider the Strauss process (Kelly and Ripley 1976), which has density

$$f_\theta(\mathbf{x}) = \frac{1}{Z(\theta)} \beta^{n(\mathbf{x})} \gamma^{s(\mathbf{x})}. \tag{20}$$

Here $\theta = (\log\beta, \log\gamma)$, with $\beta > 0$ and $0 \leqslant \gamma \leqslant 1$, are the parameters; $r > 0$ is a fixed distance called the range of interaction; $n(\mathbf{x})$ denotes the number of points in $\mathbf{x}$, and $s(\mathbf{x})$ the number of pairs of r-close points, given by

$$s(\mathbf{x}) = \#\{(i, j) : x_i, x_j \in \mathbf{x}, \|x_i - x_j\| \leqslant r\}.$$

Again, maximum likelihood estimation of θ requires numerical computation of $Z(\theta)$, for example by numerical integration (Ogata and Tanemura 1981).

5.1. Pseudolikelihood approach

Besag (1977), Besag *et al.* (1982) and Jensen and Møller (1991) extended the definition of pseudolikelihood to finite point processes. Assume X has a Papangelou conditional intensity $\lambda_\theta(u, \mathbf{x})$, $u \in W$, $\mathbf{x} \in \mathscr{X}$ (Kallenberg 1983; Nguyen and Zessin 1976). This is the Radon–Nikodym derivative defined essentially uniquely by

$$\mathbb{E} \sum_{i=1}^{n(X)} g(x_i, X \backslash \{x_i\}) = \mathbb{E} \int_W \lambda_\theta(u, X) g(u, X \backslash \{u\}) \, \mathrm{d}u \tag{21}$$

for all bounded non-negative measurable functions $g : \mathbb{R}^d \times \mathscr{X} \to \mathbb{R}$. Intuitively $\lambda_\theta(u, \mathbf{x}) \, \mathrm{d}u$ is the conditional probability that X will contain a point at u given that the rest of the configuration is $X \backslash \{u\} = \mathbf{x} \backslash \{u\}$. This is the continuous analogue of the single-site conditional probabilities $P\{X_i = a | X_{G \backslash i} = \mathbf{x}_{G \backslash i}\}$ for a discrete random field. Under integrability conditions on f_θ we have

$$\lambda_\theta(u, \mathbf{x}) = \frac{f_\theta(\mathbf{x} \cup \{u\})}{f_\theta(\mathbf{x})}$$

when $u \notin \mathbf{x}$, and

$$\lambda_\theta(x_i, \mathbf{x}) = \frac{f_\theta(\mathbf{x})}{f_\theta(\mathbf{x}\backslash\{x_i\})}$$

for $x_i \in \mathbf{x}$. The pseudolikelihood is defined by

$$\mathrm{PL}(\theta; \mathbf{x}) = \left[\prod_{i=1}^{n(\mathbf{x})} \lambda_\theta(x_i, \mathbf{x})\right] \exp\left\{-\int_W \lambda_\theta(u, \mathbf{x})\, \mathrm{d}u\right\} \tag{22}$$

provided the integral exists almost surely. If $\Theta \subseteq \mathbb{R}$ then, assuming differentiability,

$$\frac{\partial}{\partial\theta} \log \mathrm{PL}(\theta; \mathbf{x}) = \sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta}\lambda_\theta(x_i, \mathbf{x})}{\lambda_\theta(x_i, \mathbf{x})} - \int_W \frac{\partial}{\partial\theta}\lambda_\theta(u, \mathbf{x})\, \mathrm{d}u. \tag{23}$$

The maximum pseudolikelihood estimator is the root of this expression, if the maximum is achieved at a zero of the partial derivative.

Example. For the Strauss process (20), the conditional intensity is

$$\lambda_\theta(u; \mathbf{x}) = \beta\gamma^{t(u,\mathbf{x}\backslash\{u\})}, \qquad u \notin \mathbf{x},$$

where

$$\begin{aligned} t(u, \mathbf{x}) &= s(\mathbf{x} \cup \{u\}) - s(\mathbf{x}) \\ &= \#\{i : \|u - x_i\| \leqslant r\} \end{aligned} \tag{24}$$

is the number of points of the realization $\mathbf{x}$ within a distance r of the point $u \in \mathbb{R}^k$. The pseudolikelihood is

$$\mathrm{PL}(\beta, \gamma, r; \mathbf{x}) = \beta^{n(\mathbf{x})}\gamma^{2s(\mathbf{x})} \exp\left(-\beta\int_W \gamma^{t(u,\mathbf{x})}\, \mathrm{d}u\right),$$

where we have used the fact that

$$\sum_{i=1}^{n(\mathbf{x})} t(x_i, \mathbf{x}\backslash\{x_i\}) = 2s(\mathbf{x}). \tag{25}$$

The stationary point of $\mathrm{PL}(\cdot; \mathbf{x})$ for the Strauss process (20) is the solution of

$$\beta\int_W \gamma^{t(u,\mathbf{x})}\, \mathrm{d}u = n(\mathbf{x}) \tag{26}$$

$$\beta\int_W t(u, \mathbf{x})\gamma^{t(u,\mathbf{x})}\, \mathrm{d}u = 2s(\mathbf{x}). \tag{27}$$

5.2. Takacs–Fiksel approach

Takacs (1983; 1986) and Fiksel (1984, 1988) proposed estimating θ, given data $\mathbf{x}$ inside a window W, by solving for θ in

$$\sum_{x_i \in B} h(x_i, \mathbf{x}\backslash\{x_i\}) = \int_B \lambda_\theta(u, \mathbf{x}) h(u, \mathbf{x})\, \mathrm{d}u \tag{28}$$

which is an unbiased estimating equation by virtue of (21) for any Borel set $B \subseteq W$ and integrable function $h : W \times \mathscr{X} \to \mathbb{R}_+$.

In fact Takacs and Fiksel consider the case where X is a partial observation $X = X_0 \cap W$ of a stationary point process X_0 in $\mathbb{R}^d$ with conditional intensity $\lambda_\theta(\cdot, \cdot)$ which is equivariant under translation,

$$\lambda_\theta(u, \mathbf{x}) = \lambda_\theta(0, T_{-u}\mathbf{x}),$$

where $T_{-u}\mathbf{x} = \{x_1 - u, \ldots, x_n - u\}$ is the configuration obtained by shifting $\mathbf{x}$ by the vector $-u$. The estimating equation (28) then reduces to

$$\sum_{x_i \in B} h(T_{-x_i}\mathbf{x}\backslash\{x_i\}) = \int_B \lambda_\theta(0, T_{-u}\mathbf{x}) h(T_{-u}\mathbf{x})\, \mathrm{d}u, \tag{29}$$

where $h : \mathscr{X} \to \mathbb{R}$ is any bounded non-negative measurable function for which the expectation of the left-hand side exists.

Here $\lambda_\theta(\cdot, \cdot)$ refers to the conditional intensity of the stationary process X_0. Hence B should be a subset of W chosen so that the conditional intensity is 'observable', $\lambda_\theta(u, \mathbf{x}) = \lambda_\theta(u, \mathbf{x} \cap W)$ for all $u \in B$. For example, for the Strauss process (20) with interaction radius r, B is typically taken to be the set of all $u \in W$ such that the ball of radius r centred on u is wholly contained in W, so that $t(u, \mathbf{x}) = t(u, \mathbf{x} \cap W)$ is observable.

It is more usual, but equivalent, to regard (29) as arising from the Nguyen–Zessin identity (Nguyen and Zessin 1976) for a stationary point process

$$\lambda \mathbb{E}^{!0}[h(X)] = \mathbb{E}[\lambda_\theta(0, X) h(X)], \tag{30}$$

where λ is the (constant) intensity of X_0 and $\mathbb{E}^{!0}$ denotes expectation with respect to the Palm distribution of X at 0.

Example. For the Strauss process (20), a finite point process in the window W, equation (28) becomes

$$\sum_{i=1}^{n(\mathbf{x})} h(x_i, \mathbf{x}\backslash\{x_i\}) = \int_W \beta\gamma^{t(u,\mathbf{x})} h(u, \mathbf{x})\, \mathrm{d}u; \tag{31}$$

this coincides with the pseudolikelihood normal equations (26) and (27) if we choose $h \equiv 1$ and $h(u, \mathbf{x}) = t(u, \mathbf{x})$, respectively. That is, the Takacs–Fiksel method produces the maximum pseudolikelihood normal equations.

A similar result is obtained when X is a partially observed stationary Strauss process.

The connection between the Takacs–Fiksel and pseudolikelihood approaches was found by Diggle *et al.* (1994).

5.3. Time-invariance estimator

Under suitable conditions a finite point process can be represented as the equilibrium distribution of a spatial birth-and-death process (see Geyer 1999; Geyer and Møller 1994; Møller 1989; 1999; Preston 1975). This is a continuous-time pure jump Markov process (Y_t) whose states are finite point patterns $\mathbf{x} \in \mathscr{X}$, with the only instantaneous transitions being 'births' $\mathbf{x} \rightsquigarrow \mathbf{x} \cup \{u\}$ in which a new point $u \in W$ is added to the existing configuration $\mathbf{x}$, and 'deaths' $\mathbf{x} \rightsquigarrow \mathbf{x} \backslash \{x_i\}$ where one of the existing points $x_i \in \mathbf{x}$ is deleted. Suppose births occur at rate $b_\theta(\mathbf{x}, u)\,\mathrm{d}u$ and deaths at rate $d_\theta(\mathbf{x}, x_i)$. Under suitable non-explosion conditions (see, for example, Preston 1975; Baddeley and Møller 1989; Møller 1989) this process exists and is in detailed balance with f_θ:

$$f_\theta(\mathbf{x})b_\theta(\mathbf{x}, u) = f_\theta(\mathbf{x} \cup \{u\})d_\theta(\mathbf{x} \cup \{u\}, u) \tag{32}$$

for all $u \in W$, $\mathbf{x} \in \mathscr{X}$, and $\mathbf{Y}$ has equilibrium density f_θ.

The infinitesimal generator of $\mathbf{Y}$ is

$$(\mathscr{A}_\theta S)(\mathbf{x}) = \int_W b_\theta(\mathbf{x}, u)[S(\mathbf{x} \cup \{u\}) - S(\mathbf{x})]\,\mathrm{d}u + \sum_{i=1}^{n(\mathbf{x})} d_\theta(\mathbf{x}, x_i)[S(\mathbf{x}\backslash\{x_i\}) - S(\mathbf{x})] \tag{33}$$

defined for all bounded Borel functions $S : \mathscr{X} \to \mathbb{R}$. The domain of the generator may be extended to functions S which are merely L^2 integrable with respect to f_θ.

Consider the standard 'constant death rate' process with

$$b_\theta(\mathbf{x}, u) = \frac{f_\theta(\mathbf{x} \cup \{u\})}{f_\theta(\mathbf{x})} = \lambda_\theta(u, \mathbf{x})),$$

$$d_\theta(\mathbf{x}, x_i) = 1, \tag{34}$$

in which births occur at a rate $b_\theta(\mathbf{x}, \cdot)$ depending on the current configuration, and points have independent exponential (mean 1) lifetimes before deletion. This satisfies (32). The time-invariance estimating equations are, from (7),

$$\sum_{i=1}^{n(x)}[S(\mathbf{x}) - S(\mathbf{x}\backslash\{x_i\})] = \int_W \lambda_\theta(u, \mathbf{x})[S(\mathbf{x} \cup \{u\}) - S(\mathbf{x})]\,\mathrm{d}u. \tag{35}$$

For example, for the finite Strauss process (20), let S be the canonical sufficient statistic $S(\mathbf{x}) = (n(\mathbf{x}), s(\mathbf{x}))^T$. Then we have

$$S(\mathbf{x} \cup \{u\}) - S(\mathbf{x}) = \begin{pmatrix} 1 \\ t(u, \mathbf{x}) \end{pmatrix},$$

so (35) reduces to (26)–(27), that is to say, the time-invariance estimator coincides with the Takacs–Fiksel and maximum pseudolikelihood estimators in this case. More generally, we have the following result.

Proposition 2. *Let X be a finite point process on a bounded domain W with strictly positive probability density f_θ. Let $\mathbf{Y}^{(\theta)}$ be the associated spatial birth-and-death process with constant death rate (34).*

(a) The time-invariance estimator derived from $\mathbf{Y}$ and any statistic h is the Takacs–Fiksel estimator derived from h as in (29).
(b) Assume the density f_θ forms an exponential family with canonical parameter θ and canonical sufficient statistic V, and that Θ contains a neighbourhood of 0. Then the time-invariance estimating equation derived from $\mathbf{Y}$ and the statistic $S = V$ is the maximum pseudolikelihood normal equation for θ.

Again we have other interesting alternative estimators. Consider the spatial birth-and-death process with constant birth rate and variable death rate:

$$b_\theta(\mathbf{x}, u) = 1,$$

$$d_\theta(\mathbf{x}, x_i) = \frac{1}{\lambda_\theta(x_i; \mathbf{x})}. \tag{36}$$

This has infinitesimal generator, from (7), given by

$$(\mathscr{A}_\theta h)(\mathbf{x}) = \int_W [h(\mathbf{x} \cup \{u\}) - h(\mathbf{x})] \, \mathrm{d}u + \sum_{i=1}^{n(\mathbf{x})} \frac{h(\mathbf{x} \backslash x_i) - h(\mathbf{x})}{\lambda_\theta(x_i, \mathbf{x})},$$

so that, for the Strauss process (20), the time-invariance estimator $\hat{\theta}_T = (\log \hat{\beta}_T, \log \hat{\gamma}_T)$ is the solution of

$$0 = (\mathscr{A}_\theta n)(\mathbf{x}) = |W| - \beta^{-1} \sum_{i=1}^{n(\mathbf{x})} \gamma^{-t(x_i, \mathbf{x} \backslash x_i)},$$

$$0 = (\mathscr{A}_\theta s)(\mathbf{x}) = \sum_{i=1}^{n(\mathbf{x})} |B(x_i, r) \cap W| - \beta^{-1} \sum_{i=1}^{n(\mathbf{x})} t(x_i, \mathbf{x} \backslash x_i) / \gamma^{-t(x_i, \mathbf{x} \backslash x_i)},$$

that is, $(\hat{\beta}_T, \hat{\gamma}_T)$ is the solution of

$$\frac{\sum_{i=1}^{n(\mathbf{x})} t(x_i, \mathbf{x} \backslash x_i) \gamma^{-t(x_i, \mathbf{x} \backslash x_i)}}{\sum_{i=1}^{n(\mathbf{x})} \gamma^{-t(x_i, \mathbf{x} \backslash x_i)}} = \frac{\sum_{i=1}^{n(\mathbf{x})} |B(x_i, r) \cap W|}{|W|}, \tag{37}$$

$$\beta = \frac{1}{|W|} \sum_{i=1}^{n(\mathbf{x})} \gamma^{-t(x_i, \mathbf{x} \backslash x_i)}. \tag{38}$$

6. Variational estimators

Let $X = (X_1, \ldots, X_n)$ be a random element of $\mathbb{R}^n$ whose distribution has an exponential family density

$$\pi_\theta(\mathbf{x}) = \frac{1}{Z(\theta)} \exp(-\theta U(\mathbf{x}))$$

for $\mathbf{x} \in \mathbb{R}^n$, where $U : \mathbb{R}^n \to \mathbb{R}$ is a C^2 function and $\theta \in \mathbb{R}$ the parameter.

An approach developed by Almeida and Gidas (1993, (1.1)–(1.4)) is to find a vector-valued statistic $W : \mathbb{R}^n \to \mathbb{R}^n$ solving the 'variational equation'

$$\int_{\mathbb{R}^n} \nabla \cdot [W(\mathbf{x})\pi_\theta(\mathbf{x})]\, \mathrm{d}x = 0. \tag{39}$$

Here $\nabla \cdot f(x) = \sum_i \partial f_i / \partial x_i$ for a function $f : \mathbb{R}^n \to \mathbb{R}^n$, writing $f(x) = (f_1(x), \ldots, f_n(x))$ and $x = (x_1, \ldots, x_n)$.

Equation (39) implies

$$\theta \mathbb{E}_\theta[W(X) \cdot \nabla U(X)] = \mathbb{E}_\theta[\nabla \cdot W(X)]. \tag{40}$$

Almeida and Gidas then propose to estimate θ by replacing the expectations on the left and right of (40) by empirical estimates, and solving for θ. In fact Almeida and Gidas consider a generalization to vector-valued θ which we will not explore here.

A special case of (39) considered in Almeida and Gidas (1993) concerns

$$W(\mathbf{x}) = \nabla U(\mathbf{x})$$

when the estimating equation becomes

$$\theta \|\nabla U(\mathbf{x})\|^2 = \Delta U(\mathbf{x}), \tag{41}$$

where $\Delta f = \sum_{i=1}^n \partial^2 f_i / \partial x_i^2$.

To compare this with the time-invariance approach, let (Y_t) be an $\mathbb{R}^n$-valued continuous-time Markov process satisfying the stochastic differential equation (Ethier and Kurtz 1986, p. 366)

$$\mathrm{d}\, Y_t = -\tfrac{1}{2}\theta \nabla U(Y_t)\, \mathrm{d}t + \mathrm{d}W_t,$$

where $\mathbf{W}_t$ is $\mathbb{R}^n$-valued Brownian motion. This is 'Langevin dynamics' which has equilibrium distribution π_θ (see Karlin and Taylor 1981, pp. 220–221). The infinitesimal generator is

$$\mathscr{A}_\theta = -\frac{1}{2}\theta \sum_{i=1}^n \frac{\partial U}{\partial x_i}\frac{\partial}{\partial x_i} + \frac{1}{2}\sum_{i=1}^n \frac{\partial^2}{\partial x_i^2}.$$

Applying the time-invariance method to the sufficient statistic $U(\mathbf{x})$, we obtain that $\hat{\theta}_T$ is the solution of

$$0 = \mathscr{A}_\theta U(\mathbf{x}) = -\frac{1}{2}\theta \sum_{i=1}^n \left(\frac{\partial U}{\partial x_i}\right)^2 + \frac{1}{2}\sum_{i=1}^n \frac{\partial^2 U}{\partial x_i^2},$$

which is equivalent to (41).

Hence the time-invariance estimator coincides with the variational estimator in this special case.

7. Survival analysis and censoring

The time-invariance approach often yields sensible estimators even when the observations are censored or there are missing data.

Consider the simplest model of independent random censoring, where the 'true' lifetimes $T_1, \ldots, T_n$ of n individuals are independent and identically distributed with unknown distribution function F which is to be estimated. The lifetimes are right-censored by censoring times $C_1, \ldots, C_n$ which are independent and identically distributed with distribution function G, and independent of the true lifetimes. We observe only the truncated lifetimes

$$\tilde{T}_i = \min(T_i, C_i)$$

and the censoring times C_i. The data consist of a vector

$$\mathbf{x} = \{(\tilde{t}_1, c_1), \ldots, (\tilde{t}_n, c_n)\}$$

in the sample space $\mathscr{X} = Q^n$, where $Q = \{(s, c) \in \mathbb{R}_+^2 : s \leqslant c\}$.

Suppose we wish to estimate $F(r)$ for a fixed $r > 0$. Choose the statistic $S : \mathscr{X} \to \mathbb{R}$ to be

$$S(\mathbf{x}) = \frac{1}{n}\#\{i : \tilde{t}_i \leqslant r\},$$

the value at r of the empirical distribution of the *observed* lifetimes. This is a severely biased estimator of $F(r)$.

For any distribution function F, let $\mathbf{Y}^F$ be the discrete time Markov process on $\mathscr{X}$ under which, when the current state is $\mathbf{x} = \{(\tilde{t}_1, c_1), \ldots, (\tilde{t}_n, c_n)\}$, the next state is determined by choosing an index $i \in \{1, 2, \ldots, n\}$ with equal probability, and replacing the entry $\tilde{t}_i$ in $\mathbf{x}$ by the value $\tilde{T}'_i = \min(T'_i, c_i)$, where T'_i is drawn according to the distribution F independently of $\mathbf{x}$. The corresponding censoring time c_i is not changed. Thus the process $\mathbf{Y}^F$ is reducible. Clearly the distribution of $X = \{(\tilde{T}_1, C_1), \ldots, (\tilde{T}_n, C_n)\}$ under F, for any G, is an equilibrium distribution of $\mathbf{Y}^F$. The generator of $\mathbf{Y}^F$ is

$$\begin{aligned}
\mathscr{A}_F S(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}[\mathbf{1}\{\tilde{t}_i \leqslant r\} - P_F(\tilde{T}'_i \leqslant r)] \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{n}\mathbf{1}\{\tilde{t}_i \leqslant r\} - \sum_{i=1}^{n}\mathbf{1}\{c_i \leqslant r\} - \sum_{i=1}^{n}\mathbf{1}\{c_i > r\}P_F(T_i \leqslant r)\right] \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{n}\mathbf{1}\{\tilde{t}_i \leqslant r, c_i > r\} - F(r)\sum_{i=1}^{n}\mathbf{1}\{c_i > r\}\right],
\end{aligned}$$

so that a time-invariance estimator of $F(r)$ is

$$\hat{F}_T(r) = \frac{\#\{i : \tilde{t}_i \leqslant r, c_i > r\}}{\#\{i : c_i > r\}}.$$

This is known as the *reduced-sample estimator* of F – see Gill (1994); Andersen *et al.* (1993). It is pointwise unbiased, although it is not the most efficient estimator in this context. Interestingly, the time-invariance approach has automatically converted a biased estimator into a sensible unbiased estimator, in the presence of censoring.

8. Dead leaves model

The dead leaves model (see, for example, Serra 1982, pp. 508–511, 560; Hall 1988, pp. 295–296) is a random partition of $\mathbb{R}^d$ which is effectively defined as the time-equilibrium distribution of a space-time process.

Consider a homogeneous Poisson process of points (x_i, t_i) in $\mathbb{R}^d \times \mathbb{R}$ with intensity λ, and an independent sequence of independent and identically distributed random compact sets ('leaves') L_i in $\mathbb{R}^d$. Intuitively the leaves 'fall' at times t_i onto $\mathbb{R}^d$ at the locations x_i so that each new arrival obscures any earlier leaves which it may overlap. Figure 1 shows a typical realization of the dead leaves model when the leaves L_i are circular discs with random radii.

Let $K_i = L_i \oplus x_i$ be the translation of L_i by the vector x_i. At time $t \in \mathbb{R}$ define Y_t to be the random partition (Matheron 1969, pp. 35–39) of $\mathbb{R}^d$ consisting of all non-empty sets of the form

$$C_i = K_i \setminus \bigcup_{t \geq t_j \geq t_i} K_j \tag{42}$$

for i such that $t_i \leq t$. That is, C_i is that part of leaf K_i which has not been covered by leaves K_j that arrived later than K_i but before the current time t.

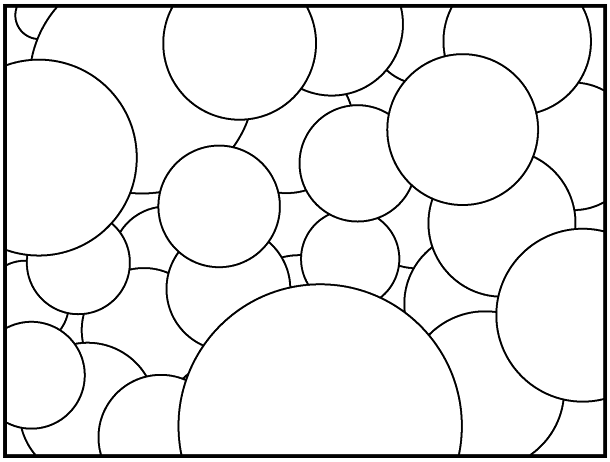

Figure 1. A typical realization of the dead leaves model with circular 'leaves'.

Clearly Y_t is stationary in time and space. The distribution of Y_t at any time t is the dead leaves model. It is of interest to estimate the distribution of the leaf size and shape L from observation of the dead leaves model inside a bounded compact window $W \subset \mathbb{R}^d$.

The infinitesimal generator of (Y_t) is

$$(\mathscr{A}_L S)(\mathbf{x}) = \lambda \mathbb{E} \int_{\mathbb{R}^d} [S(\mathbf{x} \vee (L \oplus y)) - S(\mathbf{x})] \, \mathrm{d}y \tag{43}$$

where $\mathbf{x} \vee K$ denotes the result of superimposing a new compact set K on the existing partition $\mathbf{x}$. The domain of $\mathscr{A}_L$ contains all measurable non-negative functionals S for which the right-hand side of (43) is absolutely convergent.

The integral in (43) can be evaluated in special cases using results from integral geometry (Santaló 1976). Fix $d = 2$, and assume that L has finite expected area $\mathbb{E}|L|$ and that the boundary ∂L is almost surely rectifiable with finite expected length $\mathbb{E}\ell(\partial L)$.

For a realization of the partition $\mathbf{x}$, let $\partial x = \cup_{C_i \in \mathbf{x}} \partial C_i$ denote the union of all 'visible leaf boundaries'. For the functional

$$S_1(\mathbf{x}) = \ell(W \cap \partial \mathbf{x})$$

we have, for $K = L \oplus y$,

$$S_1(\mathbf{x} \vee K) - S_1(\mathbf{x}) = \ell(W \cap \partial K) - \ell(K \cap \partial \mathbf{x}) \text{ almost surely,}$$

and a standard translational integral formula (Santaló 1976; Weil 1989; 1990) gives

$$\int_{\mathbb{R}^2} \ell(W \cap \partial(L \oplus y)) \, \mathrm{d}y = \ell(\partial L)|W|,$$

$$\int_{\mathbb{R}^2} \ell((L \oplus y) \cap \partial \mathbf{x} \cap W) \, \mathrm{d}y = \ell(\partial \mathbf{x} \cap W)|L|.$$

We obtain

$$(\mathscr{A}_L S_1)(\mathbf{x}) = \lambda |W| \mathbb{E}\ell(\partial L) - \lambda \ell(\partial \mathbf{x} \cap W) \mathbb{E}|L| \tag{44}$$

so that the time-invariance estimating equation derived from S_1 is

$$\frac{\mathbb{E}\ell(\partial L)}{\mathbb{E}|L|} = \frac{\ell(\partial \mathbf{x} \cap W)}{|W|}. \tag{45}$$

Secondly, assume the distribution of L is isotropic (rotation-invariant). Then (43) can be rewritten

$$(\mathscr{A}_L S)(\mathbf{x}) = \frac{\lambda}{2\pi} \mathbb{E} \int_{\mathrm{EM}(2)} [S(\mathbf{x} \vee gL) - S(\mathbf{x})] \, \mathrm{d}\kappa(g), \tag{46}$$

where EM(2) is the group of Euclidean motions in $\mathbb{R}^2$ (generated by rotations and translations) and κ is the standard kinematic measure on EM(2) (Santaló 1976, Chapter 6).

For a partition $\mathbf{x}$, let $v(\mathbf{x}) = \cup_{C_i \in \mathbf{x}} v_{C_i}$ be the set of all 'visible vertices', where, for a cell C_i as at (42),

$$v_{C_i} = \bigcup_{t \geqslant t_j} \left(\partial K_i \cap \partial K_j \setminus \bigcup_{t \geqslant t_k \geqslant \min\{t_i, t_j\}} \operatorname{int} K_k \right).$$

The latter is almost surely finite under the assumptions stated on L. Consider the functional

$$S_2(\mathbf{x}) = \#(W \cap v(\mathbf{x})),$$

where $\#$ denotes cardinality, that is, $S_2(\mathbf{x})$ is the number of visible vertices of $\mathbf{x}$ in W. Then

$$S_2(\mathbf{x} \vee K) - S_2(\mathbf{x}) = \#(\partial K \cap \partial \mathbf{x} \cap W) - \#(K \cap v(\mathbf{x}) \cap W) \text{ almost surely.}$$

Poincaré's formula (Santaló 1976, (7.11), p. 111) and other identities (Santaló 1976, Exercise 1, pp. 104–105) give

$$\int_{\mathrm{EM}(2)} \#(\partial(gL) \cap \partial \mathbf{x} \cap W)\, \mathrm{d}\kappa(g) = 4\ell(\partial L)\ell(\partial \mathbf{x} \cap W),$$

$$\int_{\mathrm{EM}(2)} \#(gL \cap v(\mathbf{x}) \cap W)\, \mathrm{d}\kappa(g) = 2\pi |L| \#(v(\mathbf{x}) \cap W).$$

Hence

$$(\mathscr{A}_L S_2)(\mathbf{x}) = 4\lambda \ell(\partial \mathbf{x} \cap W)\mathbb{E}\ell(\partial L) - 2\pi\lambda \#(v(\mathbf{x}) \cap W)\mathbb{E}|L|$$

and the time-invariance estimating equation derived from S_2 is

$$\frac{\mathbb{E}\ell(\partial L)}{\mathbb{E}|L|} = \frac{2}{\pi} \frac{\#(v(\mathbf{x}) \cap W)}{\ell(\partial \mathbf{x} \cap W)}. \tag{47}$$

9. Performance

So far we have avoided important questions of statistical performance of the estimators, such as consistency, asymptotic normality, and efficiency. Unfortunately, little can be said about these issues at this level of generality, for several reasons.

Firstly, our general framework does not include a limiting regime relevant to the original setting. Note especially that the 'time' index t of the Markov process (Y_t) is usually not related to the original problem. Rather, the limit behaviour of the estimator would be studied by considering a sequence of probability distributions $P_\theta^{(n)}$ on sample spaces $\mathscr{X}^{(n)}$, $n = 1, 2, \ldots,$ for a fixed parameter space Θ, with n being a measure of sample size. A corresponding sequence of time-invariance estimators $\hat{\theta}_T^{(n)}$ of θ would be derived from processes $\mathbf{Y}^{(n)}$ and statistics $S^{(n)}$ on $\mathscr{X}^{(n)}$. The limit behaviour as $n \to \infty$ depends on the structure of the specific problem, and is not related (in general) to the behaviour of trajectories of $\mathbf{Y}^{(n)}$ with respect to the 'fictional' time index t.

Secondly, especially in spatial problems, there may be several alternative limiting regimes, giving rise to different limit behaviour (Ripley 1988; Baddeley and Gill 1997; Stein 1995). Thirdly, under a particular limiting regime, consistency and asymptotic normality may hold only under regularity conditions specific to the context.

Important examples are the limit behaviour of the maximum pseudolikelihood estimator for a Markov random field (Section 4.1) and for a point process (Section 5.1). The generators of the pure jump processes **Y** (Sections 4 and 5) take the form of sums or integrals over $W \subset \mathbb{R}^d$; cf. (2) and (7). Hence in the Taylor expansion

$$-\mathscr{A}_{\theta_0} S(\mathbf{x}) = \mathscr{A}_{\hat{\theta}_T} S(\mathbf{x}) - \mathscr{A}_{\theta_0} S(\mathbf{x}) = D(\mathbf{x}, \theta_0)(\hat{\theta}_T - \theta_0) + R_{\theta_0}(\mathbf{x}, \hat{\theta}_T) \tag{48}$$

both $\mathscr{A}_{\theta_0} S(\mathbf{x})$ and its derivative $D(\mathbf{x}, \theta_0) = \partial/\partial\theta \mathscr{A}_{\theta_0} S(\mathbf{x})$ are the partial sums or integrals of random fields on $\mathbb{R}^d$. One might try to prove consistency and asymptotic normality of $\hat{\theta}_T$ by applying limit theorems for integrals of random fields. However, there are difficulties in verifying standard mixing conditions. Indeed, the random field or point process X may exhibit long-range dependence.

For the case of discrete Markov random fields, suppose we have models X_G defined on each finite subset $G \subset \mathbb{R}^d$ and consider the limit as $G \nearrow \mathbb{R}^d$. There need not exist a unique random field X on $\mathbb{R}^d$ obtained in the limit, in the sense that the conditional distributions of X on each G agree with those of X_G (see Georgii 1988, Section 6.2). There may be more than one random field satisfying these consistency relations ('phase transition') and there may be non-stationary solutions ('symmetry breakdown'). Statistical problems are discussed by Guyon (1996). Comets (1992) has proved strong consistency of $\hat{\theta}_{\text{MPL}}$ even in the case of symmetry breakdown, using a large-deviations result. Comets and Janžura (1998) derive an asymptotic normality result for $\hat{\theta}_{\text{MPL}}$ without needing asymptotic behaviour of the sample covariance. However, efficiency can only be studied properly under regularity conditions which imply uniqueness (and ergodicity) of the stationary random field X on $\mathbb{R}^d$. This is investigated by Janžura (1997).

Similarly, for the case of point processes, there may be phase transition and symmetry breakdown. The best available results on consistency and asymptotic normality of maximum pseudolikelihood estimators (Jensen and Møller 1991, Theorem 3.1; Jensen and Künsch 1994) make very restrictive assumptions on the interaction potentials.

A martingale approach can be used in at least one case. Kessler and Sørensen (1999) and Hansen and Scheinkman (1995) study diffusions X in one-dimensional time and derive estimating equations from the generator of the process X itself, rather than from the generator of an associated process **Y** indexed by another time dimension. The limiting behaviour of such estimators can be investigated using martingale limit theorems. Optimal estimating equations based on discrete-time samples of the diffusion are determined by spectral properties of the generator of the diffusion.

10. Optimality

We need a way to identify an optimal estimator amongst the wide variety of time-invariance estimators obtained under different choices of the process **Y** and statistic S. We consider two different approaches, based on the theory of estimating equations (Section 10.1) and on Markov process theory (Section 10.2), respectively.

10.1. Optimal estimating functions

10.1.1. Theory

First we recall some elements of the Godambe–Heyde theory of optimality for estimating functions; see the surveys by Godambe and Kale (1991) and MacLeish and Small (1988). Assume $\Theta \subseteq \mathbb{R}^m$ and consider an estimating equation $g(x, \theta) = 0$, where $g : \mathscr{X} \times \Theta \to \mathbb{R}^m$ is called the estimating function. Define the standardized version of g by

$$g_s(x, \theta) = \left[\mathbb{E}_\theta\left(\frac{\partial}{\partial\theta} g(X, \theta)\right)\right]^{-1} g(x, \theta).$$

Assume $g(X, \theta)$ is square-integrable for all θ and almost surely differentiable with respect to θ, and that $\mathbb{E}\partial g(X, \theta)/\partial\theta$ exists and is non-singular for all θ. Let $M(g_s, \theta)$ denote the variance–covariance matrix of $g_s(X, \theta)$. Define g^* to be *optimal* in some class of functions $\mathscr{G}$ if one of the following equivalent conditions holds: (a) $M(g_s, \theta) - M(g_s^*, \theta)$ is non-negative definite for all $g \in \mathscr{G}$ and all $\theta \in \Theta$; (b) trace$(M(g_s^*, \theta))$ is minimal over all $g \in \mathscr{G}$ for each θ; (c) det$(M(g_s^*, \theta))$ is minimal over all $g \in \mathscr{G}$ for each θ. If $\mathscr{G}$ is sufficiently large, then under mild conditions, the score function $g(x, \theta) = \partial \log L/\partial\theta$ is optimal.

10.1.2. Example

Here we develop one example of the optimal estimating functions approach, for the case of point processes. Adopt the notation of Section 5.3. Thus $\mathbf{Y}$ is a spatial birth-and-death process with birth rates $b_\theta(\mathbf{x}, u)$ and death rates $d_\theta(\mathbf{x}, x_i)$. The infinitesimal generator $\mathscr{A}_\theta S(x)$ was obtained in (33). The derivative of (33) with respect to θ is

$$\frac{\partial}{\partial\theta}(\mathscr{A}_\theta S(\mathbf{x})) = \int_W \delta S(u, \mathbf{x}) \frac{\partial}{\partial\theta} b_\theta(\mathbf{x}, u)\,\mathrm{d}u - \sum_{i=1}^{n(\mathbf{x})} \delta S(x_i, \mathbf{x}\backslash x_i) \frac{\partial}{\partial\theta} d_\theta(\mathbf{x}, x_i), \tag{49}$$

where $\delta S(u, \mathbf{x}) := S(\mathbf{x} \cup \{u\}) - S(\mathbf{x})$.

It is of interest to compare the efficiencies of time-invariance estimators derived under the constant death rate process $d_\theta \equiv 1$, $b_\theta(\mathbf{x}, u) = \lambda_\theta(u, \mathbf{x})$ and under the constant birth rate process $b_\theta \equiv 1$, $d_\theta(\mathbf{x}, x_i) = 1/\lambda_\theta(x_i, \mathbf{x})$. As we saw in Section 5.3, the former yields the maximum pseudolikelihood estimator, while the latter is an interesting alternative.

Assume the model is an exponential family so that $\lambda_\theta(u, \mathbf{x}) = \exp(\theta^{\mathrm{T}} \delta V(u, \mathbf{x}))$, with θ interpreted as a column vector and $V(\mathbf{x})$ as a row vector; and assume V is translation-invariant. Then for the constant death rate case,

$$\frac{\partial}{\partial\theta}\lambda_\theta(u, \mathbf{x}) = \lambda_\theta(u, \mathbf{x})\delta V(u, \mathbf{x})$$

so that

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}(\mathscr{A}_\theta S(\mathbf{x}))\right] = \mathbb{E}_\theta\left[\int_W \delta S(u, X)\delta V(u, \mathbf{x})\lambda_\theta(u, X)\,\mathrm{d}u\right].$$

Ignoring edge effects, the Nguyen–Zessin formula (30) yields

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}(\mathscr{A}_\theta S(\mathbf{x}))\right] = \lambda_\theta|W|\mathbb{E}_\theta^{!0}[\delta S(0, X)\delta V(0, X)],$$

where λ_θ is the intensity of the process and again $\mathbb{E}_\theta^{!0}$ denotes expectation with respect to the reduced Palm distribution of X at an arbitrary point 0. Hence the normalized version of the estimating function derived from the constant death rate process is (ignoring edge effects)

$$g_s(\mathbf{x}, \theta) = \frac{1}{\lambda_\theta|W|}\mathbb{E}_\theta^{!0}[\delta S(0, X)\delta V(0, X)]]^{-1}\left\{\int_W \delta S(u, \mathbf{x})\lambda_\theta(u, \mathbf{x})\,\mathrm{d}u - \sum_{i=1}^{n(\mathbf{x})}\delta S(x_i, \mathbf{x}\backslash x_i)\right\}. \tag{50}$$

For the constant birth rate case, on the other hand, the expectation of (49) becomes

$$\frac{\partial}{\partial\theta}(\mathscr{A}_\theta S(\mathbf{x})) = \mathbb{E}_\theta\left[\sum_{i=1}^{n(\mathbf{x})}\frac{\delta S(x_i, X\backslash x_i)\delta V(x_i, X\backslash x_i)}{\lambda_\theta(x_i, \mathbf{x}\backslash x_i)}\right];$$

applying the Nguyen–Zessin formula in the reverse direction gives that the normalized estimating function is (ignoring edge effects)

$$g_s(x, \theta) = \frac{1}{|W|}[\mathbb{E}_\theta[\delta S(0, X)\delta V(0, X)]]^{-1}\left\{\int_W \delta S(u, \mathbf{x})\,\mathrm{d}u - \sum_{i=1}^{n(\mathbf{x})}\frac{\delta S(x_i, \mathbf{x}\backslash x_i)}{\lambda_\theta(x_i, \mathbf{x}\backslash x_i)}\right\}. \tag{51}$$

We now need to compare the variance–covariance matrices of (50) and (51). There are very few instances where these can be evaluated. As an example, let X be the finite Strauss process (20) which has $\theta = (\log\beta, \log\gamma) \in \mathbb{R}\times(-\infty, 0]$ and $\delta V(u, \mathbf{x}) = t(u, \mathbf{x})$ as defined in (24). Choose $S(\mathbf{x}) = V(x)^{\mathrm{T}}$. Then (50) becomes

$$g_s(\mathbf{x}, \theta) = \frac{1}{\lambda_\theta|W|}\left[\mathbb{E}_\theta^{!0}\begin{pmatrix} 1 & t(0, X) \\ t(0, X) & t(0, X)^2\end{pmatrix}\right]^{-1}\begin{pmatrix}\int_W \beta\gamma^{t(u,\mathbf{x})}\,\mathrm{d}u - n(\mathbf{x}) \\ \int_W \beta\gamma^{t(u,\mathbf{x})}t(u, \mathbf{x})\,\mathrm{d}u - 2s(\mathbf{x})\end{pmatrix}, \tag{52}$$

while (51) becomes

$$g_s(\mathbf{x}, \theta) = \frac{1}{|W|}\left[\mathbb{E}_\theta\begin{pmatrix} 1 & t(0, X) \\ t(0, X) & t(0, X)^2\end{pmatrix}\right]^{-1}\begin{pmatrix}|W| - \beta^{-1}\sum_{i=1}^{n(\mathbf{x})}\gamma^{-t(x_i,\mathbf{x}\backslash x_i)} \\ \int_W t(u, \mathbf{x})\,\mathrm{d}u - \beta^{-1}\sum_{i=1}^{n(\mathbf{x})}t(x_i, \mathbf{x}\backslash x_i)\gamma^{-t(x_i,\mathbf{x}\backslash x_i)}\end{pmatrix}. \tag{53}$$

The variance–covariance matrices of (52) and (53) can be expressed as a sum of double integrals over W of expectations of functionals $t(u, \mathbf{x})^a t(v, \mathbf{x})^b$ with respect to the two-point reduced Palm distributions of the Strauss process. The expressions seem to be intractable; however, they would be amenable to Monte Carlo integration.

However, the special case $\gamma = 1$ is tractable. In that case X is a Poisson process of intensity β. By Slivnyak's theorem (Daley and Vere-Jones 1988) the reduced Palm distribution $P_\theta^{!0}$ is identical to the ordinary distribution P_θ. We also have $\lambda_\theta = \beta$. Thus the normalized estimating functions (52) and (53) in fact coincide when $\gamma = 1$, at the value

$$\frac{1}{\beta|W|}\begin{pmatrix} 1 & \beta\pi r^2 \\ \beta\pi r^2 & \beta\pi r^2 + \beta^2\pi^2 r^4 \end{pmatrix}^{-1}\begin{pmatrix} \beta|W| - n(\mathbf{x}) \\ \beta\pi r^2 n(\mathbf{x}) - 2s(\mathbf{x}) \end{pmatrix}. \tag{54}$$

(Here we use the approximation $\int_W t(u, \mathbf{x})\,\mathrm{d}u \approx \int_{\mathbb{R}^2} t(u, \mathbf{x})\,\mathrm{d}u = \pi r^2 n(\mathbf{x})$ which again ignores edge effects.) Hence the two estimators of the Strauss process parameter θ, obtained from the constant death rate and constant birth rate processes, are equally efficient under the Poisson process.

10.2. Rationale for choice of Y and S

Here we outline another rationale for making particular choices of the process Y and statistic S. Consider the transition semigroup T_t defined in (10). In discrete time we have $\mathscr{A}S(\mathbf{x}) = T_1 S(\mathbf{x}) - S(\mathbf{x})$ or simply $\mathscr{A} = T_1 - I$. Hence

$$T_n = (\mathscr{A} + I)^n = \sum_{k=1}^{n} \binom{n}{k} \mathscr{A}^k, \tag{55}$$

where the exponents denote n-fold composition. In continuous time, under suitable conditions on $\mathbf{Y}$ and on S, the transition operator can be expressed as

$$T_t S(\mathbf{x}) = (\mathrm{e}^{t\mathscr{A}} S)(\mathbf{x}), \tag{56}$$

where the exponential is interpreted as an operator power series (Kallenberg 1997, p. 314). In either case, if the distribution of Y_t converges weakly as $t \to \infty$ to the distribution of X from any initial state, and additionally $S(Y_t)$ converges to $\mathbb{E}S(X)$ in L^p for $p > 1$, then

$$\lim_{t\to\infty} T_t S(\mathbf{x}) = \mathbb{E}S(X).$$

Now assume that the distribution P_θ of X under θ forms an exponential family with canonical parameter θ and canonical sufficient statistic V. The maximum likelihood normal equations

$$V(\mathbf{x}) = \mathbb{E}_\theta V(X) \tag{57}$$

may then equivalently be rewritten

$$\lim_{t\to\infty} (\mathrm{e}^{t\mathscr{A}_\theta} - I)V(\mathbf{x}) = 0 \tag{58}$$

for the continuous-time case, and similarly for discrete time. In either case, taking only the first term in the expansion of the series on the left would yield

$$\lim_{t\to\infty} t\mathscr{A}_\theta V(\mathbf{x}) = 0,$$

which is equivalent to the time-invariance estimating equation (8) applied to the canonical sufficient statistic V. Thus, the latter equation can be regarded as a first-order approximation to the maximum likelihood normal equation.

If our aim is to approximate the maximum likelihood estimator as well as possible, we may argue that it is desirable to take $S = V$ and to choose a process $\mathbf{Y}$ which is rapidly mixing, so that the series in (55) or (56) converges rapidly.

11. Discussion

11.1. Arbitrary choice

The time-invariance estimator $\hat{\theta}_T$ depends on arbitrary choices, namely on the choice of stochastic evolution $\mathbf{Y}$ and on the statistic S to which the infinitesimal generator is applied. We see this as an advantage in complex models where the maximum likelihood estimator is not necessarily optimal and it is of interest to generate a variety of estimators for practical evaluation.

Consider, for example, the problem of estimating $\theta > 0$ from $n \geqslant 2$ independent and identically distributed observations from the uniform distribution on $[0, \theta]$. Let $\mathbf{Y}^{(\theta)}$ be the process in which, after an exponential waiting time with mean 1, one of the data $x_1, \ldots, x_n$ is chosen at random with equal probability and replaced by a random value uniformly distributed in $[0, \theta]$. If $S(\mathbf{x}) = \sum_i x_i$ the time-invariance estimator is $\hat{\theta}_T = 2\bar{x}$, the method-of-moments estimator. However, if $S(\mathbf{x}) = \max_i x_i$ we obtain a more interesting expression,

$$\hat{\theta}_T = x_{[n]} + n^{-1/2}(x_{[n]}^2 - x_{[n-1]}^2)^{1/2},$$

where $x_{[1]} \leqslant \ldots \leqslant x_{[n]}$ are the order statistics.

For Markov random fields, Besag (1986) argued that the choice between likelihood and pseudolikelihood depends on whether it is desired to exploit 'global' or 'local' spatial information. The same remark could be applied to the choice of $\mathbf{Y}$ here.

11.2. Other potential applications and extensions

Other potential applications of the method include classical statistical distributions for which a simple characterization exists; non-Markov random fields arising as the equilibrium distributions of interacting particle systems; and hidden Markov models.

The functional S may be assumed to depend on θ as well as $\mathbf{x}$. If $S : \mathscr{X} \times \Theta \to \mathbb{R}$ is such that for each $\theta \in \Theta$, $S_\theta := S(\cdot, \theta)$ is in the domain of $\mathscr{A}_\theta$, then $\mathbb{E}_\theta(\mathscr{A}_\theta S_\theta)(X) = \mathbb{E}_\theta \mathscr{A}_\theta S(X, \theta) = 0$ where X has distribution P_θ. For example, if $\{P_\theta\}$ has likelihood function $L(\mathbf{x}; \theta)$, choose $S(\mathbf{x}, \theta) = \log L(\mathbf{x}; \theta)$, and let $\mathbf{Y}$ be a sequence of independent and identically distributed realizations of X; then the time-invariance estimator satisfies the maximum likelihood normal equations.

11.3. Invariance

Estimators may also be required to be invariant under a group of transformations on $\mathscr{X}$. Suppose $T : \mathscr{X} \to \mathscr{X}$ is any map. If the statistic $S : \mathscr{X} \to \mathbb{R}$ is T-invariant in the sense that $S(T(\mathbf{x})) = S(\mathbf{x})$ for all $\mathbf{x} \in \mathscr{X}$, and if $\mathbf{Y}^{(\theta)}$ is T-equivariant in the sense that $P\{Y_t \in A | Y_0 = \mathbf{x}\} = P\{Y_t \in T(A) | Y_0 = t(\mathbf{x})\}$, then we have $(\mathscr{A}_\theta S)(T(\mathbf{x})) = (\mathscr{A}_\theta(S \circ T))(\mathbf{x})$ so that $(\mathscr{A}_\theta S)(T(\mathbf{x})) = (\mathscr{A}_\theta S)(\mathbf{x})$, meaning that the estimating equation derived from S and $\mathbf{Y}$ is T-invariant; in particular, the time-invariance estimator $\hat{\theta}_T$ is T-invariant.

11.4. Connection with Stein's method

The identity $\mathbb{E}[(\mathscr{A}S)(X)] = 0$ is fundamental to the Stein–Chen method of distributional approximation (see, for example, Arratia *et al.* 1990; Barbour 1997; Barbour *et al.* 1992; Stein 1986). Here X has a specified 'target' distribution P, and $\mathscr{A}$ is the infinitesimal generator of a Markov process Y_t which has equilibrium distribution P. If X' is another random variable with distribution P', the discrepancy between P' and P can be controlled by finding an upper bound on $|\mathbb{E}[(\mathscr{A}S)(X')]|$ for all S in a certain class of functionals.

While the affinity with the time-invariance estimator is clear, it is not so clear to the author whether any properties of the time-invariance estimator can be deduced using the Stein–Chen method. One may speculate that weak consistency of $\hat{\theta}_T$ could be proved if the class of functionals S for which a bound on $|\mathbb{E}[(\mathscr{A}S)(X')]|$ is available includes appropriate statistics.

Acknowledgements

The author is most grateful for valuable advice from A.D. Barbour, J.E. Besag, T.C. Brown, J.B. Chen, B. Gidas, M. Kessler, H.R. Künsch, T. Kurtz, M.N.M. van Lieshout, A.G. Pakes, C.G. Small, V.T. Stefanov, T.R. Turner and R.L. Tweedie. This research was funded partly by the Centrum voor Wiskunde en Informatica, the University of Leiden and the Australian Research Council.

References

Almeida, M.P. and Gidas, B. (1993) A variational method for estimating the parameters of MRF from complete or incomplete data. *Ann. Appl. Probab.*, **3**, 103–136.

Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (1993) *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Arratia, R., Goldstein, L. and Gordon, L. (1990) Poisson approximation and the Chen–Stein method. *Statist. Sci.*, **5**, 403–434.

Baddeley, A.J. and Gill, R.D. (1997) Kaplan–Meier estimators of interpoint distance distributions for spatial point processes. *Ann. Statist.*, **25**, 263–292.

Baddeley, A.J. and Møller, J. (1989) Nearest-neighbour Markov point processes and random sets. *Internat. Statist. Rev.*, **57**, 89–121.

Barbour, A.D. (1997) Stein's method. In S. Kotz, C. Read and D.L. Banks (eds), *Encyclopedia of Statistical Sciences, Update*, Vol. 1, pp. 513–521. New York: Wiley.

Barbour, A.D., Holst, L. and Janson, S. (1992) *Poisson Approximation*. Oxford: Oxford University Press.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B*, **36**, 192–236.

Besag, J. (1975) Statistical analysis of non-lattice data. *The Statistician*, **24**, 179–195.

Besag, J. (1977) Some methods of statistical analysis for spatial data. In *Bulletin of the International*

Statistical Institute: Proceedings of the 41st Session (New Delhi), Vol. 47, Book 2, pp. 77–92. Voorburg, Netherlands: ISI.

Besag, J. (1986) On the statistical analysis of dirty pictures (with discussion). *J. Roy. Statist. Soc. Ser. B*, **48**, 259–302.

Besag, J., Milne, R. and Zachary, S. (1982) Point process limits of lattice processes. *J. Appl. Probab.*, **19**, 210–216.

Carter, D.S. and Prenter, P.M. (1972) Exponential spaces and counting processes. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, **21**, 1–19.

Comets, F. (1992) On consistency for a class of estimators for exponential families of Markov random fields on a lattice. *Ann. Statist.*, **20**, 455–468.

Comets, F. and Janžura, M. (1998) A central limit theorem for conditionally centred random fields with an application to Markov fields. *J. Appl. Probab.*, **35**, 608–621.

Daley, D.J. and Vere-Jones, D. (1988) *An Introduction to the Theory of Point Processes*. New York: Springer-Verlag.

Diggle, P.J., Fiksel, T., Grabarnik, P., Ogata, Y., Stoyan, D. and Tanemura, M. (1994) On parameter estimation for pairwise interaction processes. *Internat. Statist. Rev.*, **62**, 99–117.

Ethier, S.N. and Kurtz, T.G. (1986) *Markov Processes: Characterization and Convergence*. New York: Wiley.

Fiksel, T. (1984) Estimation of parameterized pair potentials of marked and non-marked Gibbsian point processes. *Elektron. Informationsverarbeitung Kybernetik*, **20**, 270–278.

Fiksel, T. (1988) Estimation of interaction potentials of Gibbsian point processes. *Statistics*, **19**, 77–86.

Geman, D. (1990) Random fields and inverse problems in imaging. In P.-L. Hennequin (ed.), *Ecole d'Eté de Probabilités de Saint-Flour XVIII, 1988*, Lecture Notes in Math. 1427. Berlin: Springer-Verlag.

Geman, S. and Geman D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intelligence*, **6**, 721–741.

Georgii, H.-O. (1988) *Gibbs Measures and Phase Transitions*. Berlin: de Gruyter.

Geyer, C.J. (1999) Likelihood inference for spatial point processes. In O.E. Barndorff-Nielsen, W.S. Kendall and M.N.M. van Lieshout (eds), *Stochastic Geometry: Likelihood and Computation*, pp. 79–140. Boca Raton, FL: Chapman & Hall/CRC.

Geyer, C.J. and Møller, J. (1994) Simulation procedures and likelihood inference for spatial point processes. *Scand. J. Statist.*, **21**(4), 359–373.

Gill, R.D. (1994) Lectures on survival analysis. In P. Bernard (ed.), *Lectures on Probability Theory: École d'Eté de Probabilités de Saint-Flour XXII, 1992*, Lecture Notes in Math. 1581. Berlin: Springer-Verlag.

Godambe, V.P. and Kale, B.K. (1991) Estimating functions: an overview. In V.P. Godambe (ed.), *Estimating Functions*, pp. 3–20. Oxford: Clarendon Press.

Guyon, X. (1996) *Random Fields on a Network. Modelling, Statistics, and Applications*. New York: Springer-Verlag.

Hall, P. (1988) *An Introduction to the Theory of Coverage Processes*. New York: Wiley.

Hansen, L.P. and Scheinkman, J.A. (1995) Back to the future: generating moment implications for continuous time Markov processes. *Econometrica*, **63**, 767–804.

Janžura, M. (1997) Asymptotic results in parameter estimation for Gibbs random fields. *Kybernetika*, **33**, 133–159.

Jensen, J.L. and Künsch, H.R. (1994) On asymptotic normality of pseudo likelihood estimates for pairwise interaction processes. *Ann. Inst. Statist. Math.*, **46**, 475–486.

Jensen, J.L. and Møller, J. (1991) Pseudolikelihood for exponential family models of spatial point processes. *Ann. Appl. Probab.*, **1**, 445–461.

Kallenberg, O. (1983) *Random Measures*, 3rd edn. Berlin/New York: Akademie Verlag/Academic Press.

Kallenberg, O. (1997) *Foundations of Modern Probability.* New York: Springer-Verlag.

Karlin, S. and Taylor, H.M. (1981) *A Second Course on Stochastic Processes*. Orlando, FL: Academic Press.

Kelly, F.P. (1979) *Reversibility and Stochastic Networks*. Chichester: Wiley.

Kelly, F.P. and Ripley, B.D. (1976) On Strauss's model for clustering. *Biometrika*, **63**, 357–360.

Kessler, M. and Sørensen, M. (1999) Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli*, **5**, 299–314.

MacLeish, D.L. and Small, C.G. (1988) *The Theory and Applications of Statistical Inference Functions*, Lecture notes in Statist. 44. Berlin: Springer-Verlag.

Matheron, G. (1969) Théorie des ensembles aléatoires. Cahiers du Centre de Morphologie Mathématique de Fontainebleau 4, Ecole Nationale Supérieure des Mines de Paris.

Møller, J. (1989) On the rate of convergence of spatial birth-and-death processes. *Ann. Inst. Statist. Math.*, **41**, 565–581.

Møller, J. (1999) Markov chain Monte Carlo and spatial point processes. In O.E. Barndorff-Nielsen, W.S. Kendall and M.N.M. van Lieshout (eds), *Stochastic Geometry: Likelihood and Computation*, pp. 141–172. Boca Raton, FL: Chapman & Hall/CRC.

Nguyen, X.X. and Zessin, H. (1976) Punktprozesse mit Wechselwirkung. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, **37**, 91–126.

Ogata, Y. and Tanemura, M. (1981) Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure. *Ann. Inst. Statist. Math. B*, **33**, 315–338.

Preston, C.J. (1975) Spatial birth-and-death processes. In *Bulletin of the International Statistical Institute: Proceedings of the 40th Session (Warsaw)*, Vol. 46, Book 2, pp. 371–391. Voorburg, Netherlands: ISI.

Ripley, B.D. (1988) *Statistical Inference for Spatial Processes*. Cambridge University Press.

Ripley, B.D. (1989) Gibbsian interaction models. In D.A. Griffiths (ed.), *Spatial Statistics: Past, Present and Future*, pp. 1–19. New York: Image.

Santaló, L.A. (1976) *Integral Geometry and Geometric Probability*, Encyclopedia of Mathematics and Its Applications, Vol. 1. Reading, MA: Addison-Wesley.

Serra, J. (1982) *Image Analysis and Mathematical Morphology.* London: Academic Press.

Stein, C. (1986) *Approximate Computation of Expectations*. Hayward, CA: Institute of Mathematical Statistics.

Stein, M.L. (1995) An approach to asymptotic inference for spatial point processes. *Statist. Sinica*, **5**, 221–234.

Takacs, R. (1983) Estimator for the pair-potential of a Gibbsian point process. Institutsbericht 238, Institut für Mathematik, Johannes Kepler Universität Linz, Austria.

Takacs, R. (1986) Estimator for the pair potential of a Gibbsian point process. *Statistics*, **17**, 429–433.

Weil, W. (1989) Translative integral geometry. In A. Hübler, W. Nagel, B.D. Ripley and G. Werner (eds), *Geobild 89*, pp. 75–86. Berlin: Akademie Verlag.

Weil, W. (1990) Iterations of translative integral formulae and nonisotropic Poisson processes of particles. *Math. Z.*, **205**(4), 531–549.

Received October 1995 and revised December 1999