Moment estimation for statistics from marked point processes

Dimitris N. Politis
Department of Mathematics
University of California, San Diego
La Jolla, CA 92093-0112, U.S.A.

Michael Sherman
Department of Statistics
Texas A&M University
College Station, TX 77843-3143, U.S.A.

Abstract

In spatial statistics the data typically consist of measurements of some quantity at irregularly scattered locations; in other words, the data form a realization of a marked point process. In this paper, we formulate subsampling estimators of the moments of general statistics computed from marked point process data, and establish their L_2 consistency. The variance estimator in particular can be used for the construction of confidence intervals for estimated parameters. A practical data-based method of choosing a subsampling parameter is given and illustrated on a data set. Finite-sample simulation examples are also presented.

AMS 1980 subject classifications: primary 62M20; secondary 62G05.

Keywords and phrases: Confidence intervals, large sample inference, nonparametric estimation, point processes, Poisson process, random fields, subsampling.

1 Introduction

Let $\{Z(s), s \in \mathbf{R}^d\}$ be a homogeneous (strictly stationary) random field in d dimensions, with $d \in \mathbf{Z}^+$, that is, a collection of random variables $Z(s)$ taking values in a general state space $\mathcal{Z}$ that are indexed by the continuous parameter $s \in \mathbf{R}^d$, such that the joint distribution of any finite number of variables $Z(s)$ is translation invariant. In the important special case where $d = 1$, the random field $\{Z(s)\}$ is just a continuous time, stationary stochastic process. The unknown probability law of the random field $\{Z(s), s \in \mathbf{R}^d\}$ will be denoted by P_Z.

Let $S(K_n)$ be a statistic which estimates a parameter of interest β associated with P_Z; the statistic $S(K_n)$ is computed from data of the type $\{(s_1, Z(s_1)), (s_2, Z(s_2)), \ldots\}$, where the 'locations' $s_1, s_2, \ldots$ are some points in the compact set $K_n \subset \mathbf{R}^d$. If our objective is to draw inferences from $S(K_n)$ regarding β, an estimate of the variability of $S(K_n)$ will typically be required. The problem is that in many practical situations $Var(S(K_n))$ is not available to the user; this is often the case for spatial data where the presence of correlation and the complicated nature of the statistic makes a theoretical derivation of $Var(S(K_n))$ intractable.

In the case where the data are of the form $\{Z(s), s \in \mathbf{E}\}$, with $\mathbf{E}$ being a finite subset of the rectangular lattice $\mathbf{Z}^d$, there is a large literature on resampling/subsampling variance estimators; see, for example, Hall (1985), Carlstein (1986), Künsch (1989), Liu and Singh (1992), Politis and Romano (1992, 1993), Sherman and Carlstein (1994), and Sherman (1996). However, in many important cases, the data correspond to observations of $Z(s)$ at non-lattice, irregularly spaced points. For instance, if $d = 2$, $Z(s)$ might represent the measurement of a health characteristic of an individual at location s, the quality or quantity of the ore found in location s, or the precipitation at location s during a fixed time interval, etc. When $d > 1$ irregularly spaced data seem to be the rule rather than the exception; see, for example, Cressie (1993), Karr (1991), Ripley (1981).

A useful way to model the irregularly scattered s-points is to assume they are generated by a homogeneous point process observable on the compact subset $K_n \subset \mathbf{R}^d$.

The simplest example is the Poisson process; see, e.g., Karr (1986, 1991). However, different point processes —for instance, cluster point process— are also of interest; see, e.g., Jensen (1993a,b).

The joint (product) probability law of the random field $\{Z(s)\}$ and the point process N will be denoted by P. The observations then are described via the 'marked point process' defined as the collection of pairs $\{(s_j, Z(s_j)),\ j = 1, ..., N(K_n)\}$, where $\{s_j\}$ are the points at which the $\{Z(s_j)\}$ 'marks' happen to be observed; see Daley and Vere-Jones (1988), Karr (1991), Krickeberg (1982) or Stoyan et al. (1995) for more details on marked point processes.

Politis et al. (1999) studied the problem of variance estimation in the case where the statistic $S(K_n)$ is the sample mean of a continuous time random field observed at random points generated by a Poisson point process; to this effect, they proposed a block-resampling methodology analogous to the block-bootstrap schemes in Künsch (1989), Liu and Singh (1992), and Politis and Romano (1992, 1993). In the present paper, we address the issue of variance estimation for marked point processes in the case where the statistic $S(K_n)$ is of an arbitrary, general form. For example, we may desire to assess the variability of a robust estimate of location or scale. Further, we allow for more general stationary point processes, including the Poisson as a special case. Finally, our subsampling approach is adaptable to more general settings (arbitrary rates of convergence, see Sec. 3.2) and potentially to long range dependence (see Hall, Jing, and Lahiri (1998) for the equally spaced lattice data case). The block bootstrap has difficulty with long range dependence (see, e.g., Lahiri (1993)).

In the following section, we discuss our mixing assumptions and the domains in which variables are observed. In Section 3 we define the subsampling estimators of moments of the general statistic $S(K_n)$, and establish their L_2 consistency. Section 4 contains some simulation results, as well as the construction of practical approximations to the integrals defining our subsampling estimators, while Section 5 applies our methods to a data set on Longleaf pines; all technical proofs are deferred to Section 6.

2 Quantifying dependence and spatial domains

2.1 Mixing assumptions

The continuous parameter random field $\{Z(s), s \in \mathbf{R}^d\}$ will be assumed to satisfy a certain weak dependence condition that will be quantified in terms of strong mixing coefficients. Let $d(\cdot,\cdot)$ denote sup-distance (i.e., the distance arising from the l_∞ norm) on $\mathbf{R}^d$. Following Politis et al. (1998, 1999) we will make use of a particular type of strong mixing coefficients defined by

$$\alpha_Z(k;l) \equiv \sup\{|P(A_1 \cap A_2) - P(A_1)P(A_2)| : A_i \in \mathcal{F}_Z(E_i), i = 1,2, E_2 = E_1 + s,$$

$$|E_1| = |E_2| \leq l, d(E_1, E_2) \geq k\}, \tag{1}$$

where the supremum is taken over all *compact and convex* sets $E_1 \subset \mathbf{R}^d$, and over all $s \in \mathbf{R}^d$ such that $d(E_1, E_2) \geq k$; in the above, $\mathcal{F}_Z(E)$ denotes the σ–algebra generated by the random variables $\{Z(s) : s \in E\}$, and $|E|$ denotes Lebesgue measure (volume) of set E.

We assume throughout that

$$\alpha_Z(l; l^d) \to 0 \text{ as } l \to \infty. \tag{2}$$

In the $d = 1$ case, $\alpha_Z(l;\infty) \to 0$ holds for AR(1) processes with normal, double-exponential, or Cauchy errors; cf. Gastwirth and Rubin (1975). This property also holds for linear time series, e.g., MA (∞) models, with MA coefficients rapidly decaying to zero and independent and identically distributed innovations possessing an absolutely continuous distribution; cf. Doukhan (1994).

In the $d = 2$ case, Bradley (1993a) gives simple examples of random fields for which $\alpha_Z(l;\infty)$ does not tend to 0, as $l \to \infty$ but with $\alpha_Z(l;l^2) \to 0$, thus illustrating the need to account for the sizes of the sets involved in defining the strong mixing coefficients. From Künsch's (1982) Proposition 3.1 and Remark 3.5iv, it can be deduced that equation (2) holds for a natural class of Gibbs (i.e., Markov) random fields which are useful in statistical mechanics and in image processing.

In this paper, we explicitly address the case when the homogeneous point process N is Poisson distributed. Karr (1986) makes a strong case for the plausibility of the Poisson assumption for many practical situations, e.g., marked point processes arising from meteorological data. Nevertheless, the point process literature abounds with different (non-Poisson) point process models and classes of models, one of the most prominent of which is the general notion of cluster processes.

For this reason, we also generalize to the case where N is possibly non-Poisson: our results actually hold as long as the point process N satisfies a certain weak dependence condition based on the notion of maximal correlation mixing; cf. Doukhan (1994). Let

$$\rho_N(k;l) \equiv \sup\{Corr(\xi_1,\xi_2) : \xi_i \text{ is } \mathcal{F}_N(E_i)\text{−measurable and } E|\xi_i|^2 < \infty \text{ for } i = 1,2,$$

$$E_2 = E_1 + s, |E_1| = |E_2| \leq l, d(E_1,E_2) \geq k\}, \tag{3}$$

where again the supremum is taken over all *compact and convex* sets $E_1 \subset \mathbf{R}^d$, and over all $s \in \mathbf{R}^d$ such that $d(E_1,E_2) \geq k$. Note that in the above, $\mathcal{F}_N(E)$ denotes the σ–algebra generated by the random points of the point process N that happen to fall in set E; in other words, ξ_i is a (measurable) function of the random points $s_1, ..., s_{N(E_i)}$ that are generated by the point process N in set E_i.

The literature on mixing properties for point processes is incomplete at the moment, but rapidly evolving. Daley and Vere-Jones (1988, Proposition 10.3.IX) give the general result that a cluster process is mixing if (but not only if) the cluster–center process is itself mixing. A very interesting concrete application is given by the work of Jensen (1993a,b) who showed that the Strauss point process satisfies

$$\alpha_N(k;l) \leq Ale^{-Bk}$$

for some constants A, B; here the strong mixing coefficients $\alpha_N(k;l)$ are defined by(1) with the understanding that $\mathcal{F}_N(E)$ denotes the σ–algebra generated by the random points of the point process N that happen to fall in set E. Thus, a Strauss process satisfies $\alpha_N(l;l^d) \to 0$ as $l \to \infty$ which is exactly the analog of (2) for point processes.

However, for our theoretical results, we will require

$$\rho_N(l;l^d) \to 0 \text{ as } l \to \infty. \tag{4}$$

Equation (4) is seen to be stronger than (2) because of the inequality $\alpha_N(k;l) \leq \rho_N(k;l)$. However, it is not much stronger; in fact, (4) is sometimes equivalent to (2) as maximal correlation mixing and strong mixing are sometimes equivalent to each other. For example, results of Bradley (1993b) imply that, at least for dimension $d \geq 2$, a reverse inequality holds: $\rho_N(k;\infty) \leq 2\pi\alpha_N(k;\infty)$, thus rendering maximal correlation mixing equivalent to strong mixing. A similar equivalence holds in the case of Gaussian random fields; cf. Rosenblatt (1985, p. 74).

To give a concrete example of a point process satisfying (4), consider the following construction of a cluster point process: Let $\{M_s, s \in \mathbf{Z}^d\}$, be an independent and identically distributed random field taking nonnegative integer values. Consider the points $p \in \mathbf{Z}^d$, and replace each point p by a cluster of M_p points generated by some arbitrary distribution on the l_∞ ball in $\mathbf{R}^d$ that has center p and radius $\epsilon \in (0, 1/2)$. Finally, add U to each of those points—i.e., displace each point by the same amount given by U— where U is a random variable with uniform distribution on $[0,1]^d$. The cluster process constucted by the above recipe is homogeneous and trivially satisfies (4); in particular, "offspring" points at distance more than 2ϵ (corresponding to different "parents") are exactly independent, leading to

$$\rho_N(l;l^d) = 0 \text{ for all } l > 1.$$

One may conjecture that a similar independence between "offspring" from different "parents" might also hold for Poisson cluster processes but this is not true. Interestingly though, Poisson cluster processes satisfy a nearest-neighbour Markov property; hence it is expected that they will satisfy condition (4), since Markov processes are typically mixing with fast (geometric) mixing rates. For more details see Baddeley et al. (1996) and the references therein.

2.2 Domain of Locations

Let $K \subset [0,1]^d$ be a compact, convex set, and let $K_n \subset [0,n]^d$ be the region where we observe the marked point process, where K_n is simply the shape K 'inflated' by the factor n; in other words, $K_n = \{y : y = ns,\ s \in K\}$. Thus, our marked point process data are of the form $\{(s_1, Z(s_1)), ..., (s_{N(K_n)}, Z(s_{N(K_n)}))\}$, where $N(K_n)$ represents our (random) sample size. Note that in the important special case of irregularly spaced data in 2-dimensional space, the convexity assumption can be slightly relaxed; in that $(d = 2)$ case, the region K can be the interior of a rectifiable curve, thus allowing for a broader class of regions on which data can be observed, e.g., starshapes.

3 Variance estimation based on subsampling

3.1 The regular case

Let $S(K_n) := S_{N(K_n)}\{(s_1, Z(s_1)), ..., (s_{N(K_n)}, Z(s_{N(K_n)}))\}$ be our statistic which estimates a parameter of interest β associated with probability law P_Z. Also let

$$t(K_n) := |K_n|^{1/2}(S(K_n) - E\{S(K_n)\}) \tag{5}$$

denote the standardized statistic and assume that:

$$Var(t(K_n)) \to \theta \geq 0 \ \text{ as } \ n \to \infty. \tag{6}$$

Since our (expected) sample size is proportional to $|K_n|$, the standardization given in equations (5) and (6) corresponds to the *'regular'* case where the convergence of our statistic $S(K_n)$ occurs at the rate of the square-root of the sample size; see e.g. Karr (1991) for examples.

As explained in Section 1, our goal is to get an estimate of the sampling variability of $S(K_n)$ in order to draw inferences from $S(K_n)$ regarding β. Assuming a large sample, our goal may therefore be recast as estimating the unknown asymptotic variance θ without specifying the dependence structure in the process $Z(\cdot)$.

For $c \in (0,1)$, let $B_n = K_{\lfloor cn \rfloor}$ be of the same shape as K_n but rescaled, where $\lfloor r \rfloor$ is the largest integer less than or equal to the real number r, and let $B_n + y = \{t + y : t \in B_n\}$ be its shifted (translated by y) copy; here $y \in K_n^{1-c}$, where $K_n^{1-c} := \{y \in K_n : B_n + y \subset K_n\}$ is the set of 'allowed' displacements. By retaining the same shape as the original region, K_n, the subregions $B_n + y$ have approximately the same underlying dependence structure as that generating the original data due to stationarity of the random field and the point process. This is an important aspect of valid resampling schemes, as has been pointed out in Cressie (1993, p.478) who argues that "asymptotic theory should allow the domain to increase, but still respect the geometric configuration of lattice sites..."; see also Lahiri et al. (1999). For asymptotic considerations we need to allow K_n to become large but such that there are also many subregions. Thus, we assume that

$$c \to 0 \quad \text{and} \quad cn \to \infty \quad \text{as} \quad n \to \infty, \tag{7}$$

i.e., c is a function of n. A more precise notation would be $c = c_n$, but in what follows we will write c under the understanding that c is a function of n. An example is $c = n^{-1/2}$. *All* convergences in the following will be taken under condition (7).

From the (standardized) variability of the associated replicate statistics, $S(B_n + y)$, we assess the variability of (the standardized) $S(K_n)$ as follows:

$$\hat{\theta} := \int_{K_n^{1-c}} |B_n| [S(B_n + y) - \overline{S}(B_n)]^2 dy / |K_n^{1-c}|, \tag{8}$$

where $\overline{S}(B_n) := \int_{K_n^{1-c}} S(B_n + y) dy / |K_n^{1-c}|$.

This is our theoretical subsampling estimator, whose asymptotic consistency properties will be studied in our Theorem 2. In practice, a gridded or stochastic approximation to $\hat{\theta}$ will be useful as discussed in Section 4.1. However, before considering the issue of variance estimation, we establish the consistency of estimating the moments of a general statistic $h(K_n)$ using subsampling.

Let $h(K_n) := h_{N(K_n)}\{(s_1, Z(s_1)), ..., (s_{N(K_n)}, Z(s_{N(K_n)}))\}$ be a statistic calculated from our marked point process, and let $h(B_n + y)$, for $y \in K_n^{1-c}$, be the associated subregion replicates.

Our basic result, Theorem 1, gives conditions under which the subsampling estimator of $E[h(K_n)]$ is consistent; in other words, Theorem 1 allows the user to estimate any population moment of an arbitarily complicated statistic computed from irregularly spaced, correlated data under mild regularity conditions.

Theorem 1: *Assume that $E[h(K_n)] \to \gamma \in \mathbf{R}$, as well as condition (7). Also assume that for all n we have*

$$E|h(K_n)|^{2+\delta} \leq C_\delta < \infty \text{ for some } \delta > 0,$$

for some constant C_δ. If mixing conditions (2) and (4) hold

then

$$\overline{h}(B_n) \xrightarrow{L_2} \gamma,$$

where

$$\overline{h}(B_n) := \int_{K_n^{1-c}} h(B_n + y)dy/|K_n^{1-c}|. \tag{9}$$

To see how Theorem 1 can be used to obtain consistency of the variance estimator $\hat{\theta}$, defined in equation (8), note that:

$$\hat{\theta} = \int_{K_n^{1-c}} t^2(B_n + y)dy/|K_n^{1-c}| - [\int_{K_n^{1-c}} t(B_n + y)dy/|K_n^{1-c}|]^2.$$

To verify the above observe that

$$\int_{K_n^{1-c}} [S(B_n + y) - \overline{S}(B_n)]^2 dy$$

$$= \int_{K_n^{1-c}} [S(B_n + y) - E[S(B_n)] + E[S(B_n)] - \overline{S}(B_n)]^2 dy$$

$$= \int_{K_n^{1-c}} t^2(B_n + y)dy/|B_n| - (|K_n^{1-c}|^{-1})[\int_{K_n^{1-c}} t(B_n + y)dy/|B_n|^{1/2}]^2.$$

Thus, $\hat{\theta}$ is a function of moments so we can hope to use Theorem 1 to obtain a result for the variance estimator $\hat{\theta}$. This is indeed true, and we now state the L_2 consistency of $\hat{\theta}$.

Theorem 2: *Assume equations (2), (4), (5), (6), (7), and that for all n we have*

$$E|t(K_n)|^{4+\delta} \leq C'_\delta < \infty \text{ for some } \delta > 0, \tag{10}$$

for some constant C'_δ. Then

$$\hat{\theta} \xrightarrow{L_2} \theta.$$

As an important special case suppose our homogeneous point process N is Poisson. Sometimes the given context may justify the Poisson assumption (Sec. 2.1); see also, e.g., Karr (1986, 1991) for a thorough discussion. Note that a Poisson point process is characterized by independence of points dropped at disjoint subsets. An examination of the proof of our Theorems 1 and 2 indicates that, if N is Poisson, the condition $\rho_N(l; l^d) \to 0$ is redundant; thus, we have the following corollary.

Corollary 1 *: Theorems 1 and 2 remain valid if condition (4) is replaced by the assumption that N is a Poisson process.*

As a simple example, consider the case where our statistic $S(K_n)$ is the sample mean $\bar{Z}_{K_n} = N(K_n)^{-1} \sum_{i=1}^{N(K_n)} Z(s_i)$. Note here that $\bar{Z}_{K_n}$ can also be expressed compactly as an integral with respect to the point mass measure N that puts mass 1 on each of the observed s-points; in other words, $\bar{Z}_{K_n} = N(K_n)^{-1} \int_{K_n} Z(s)N(ds)$. Karr (1986) has shown that, under some regularity assumptions, $\bar{Z}_{K_n}$ is consistent for the common mean $EZ(0)$, and asymptotically normal at the 'regular' rate $\sqrt{|K_n|}$, with asymptotic variance equal to $\int R(s)ds + \lambda^{-1}R(0)$, where $R(s) := Cov(Z(0), Z(s))$. Nevertheless, to actually use this asymptotic normality to construct confidence intervals for the mean, the asymptotic variance must be estimated. Though $R(s)$ is typically unknown, nonparametric estimators of $R(s)$ (say, $\hat{R}(s)$) are available, and are consistent under some conditions (Karr, 1986, 1991). However, even though $\hat{R}(s) \to R(s)$ for each s (in probability as $n \to \infty$), the plug-in estimate $\int \hat{R}(s)ds$ will generally be *inconsistent* for $\int R(s)ds$, and thus direct estimation of the asymptotic variance is not trivial. Our subsampling variance estimator $\hat{\theta}$ will perform the required consistent estimation of $\theta = \int R(s)ds + \lambda^{-1}R(0)$, as our simulation in the next section will also confirm. Note, however, that our results are applicable to estimate the moments (and variance) of other, arbitrarily complicated, statistics as well.

3.2 Variance estimation: the general case

As before, let $S(K_n) := S_{N(K_n)}\{(s_1, Z(s_1)), ..., (s_{N(K_n)}, Z(s_{N(K_n)}))\}$ be our statistic which estimates a parameter of interest β associated with probability law P_Z. However, we now let

$$t(K_n) := \tau_{|K_n|}(S(K_n) - E\{S(K_n)\}) \tag{11}$$

denote the general standardized statistic; in the above, $\tau_x = x^{\gamma}L(x)$ for some $\gamma > 0$, and some function $L(\cdot)$ that is normalized, and slowly varying at infinity, that is, $L(1) = 0$, and for any $\lambda > 0$, $\lim_{x\to\infty} \frac{L(\lambda x)}{L(x)} = 1$ (see Bingham, Goldie, and Teugels (1987)).

Again we assume equation (6), i.e., that $\tau_{|K_n|}$ is the proper standardization for $S(K_n)$, and we construct the $B_n + y$ regions and the replicate statistics, $S(B_n + y)$. Our Theorem 1 holds *verbatim* in this general, nonregular case. Nevertheless, Theorem 2 has to be modified to account for the general rate of convergence.

Our subsampling estimator of the asymptotic variance θ is now defined as

$$\hat{\theta}_{gen} := \int_{K_n^{1-c}} \tau^2_{|B_n|}[S(B_n + y) - \overline{S}(B_n)]^2 dy/|K_n^{1-c}|, \tag{12}$$

where $\overline{S}(B_n) := \int_{K_n^{1-c}} S(B_n + y)dy/|K_n^{1-c}|$ as before. Minor modifications to the proof of Theorems 1 and 2 show that $\hat{\theta}_{gen}$ is consistent, and the following corollary ensues.

Corollary 2 *: Assume equations (2), (4), (6), (7), (10), and (11). Then:*

$$\hat{\theta}_{gen} \xrightarrow{L_2} \theta.$$

As an example, consider the case where we are interested in estimating the autocovariance $R(s) = Cov(Z(0), Z(s))$ at some fixed point s. Assuming for simplicity that

$EZ(s) = 0$, the usual estimator of $R(s)$ is given as

$$S(K_n) := \frac{|K_n|}{N(K_n)^2} \int_{K_n} \int_{K_n} W_n(s - s_1 + s_2) Z(s_1) Z(s_2) N(ds_1)(N - \epsilon_{s_1})(ds_2),$$

where ϵ_s is a point mass measure that puts mass 1 on point s; in the above $W_n(s) = a_n^{-d} W(s/a_n)$, and the kernel W is assumed to be a positive, bounded, isotropic probability function on $\mathbf{R}^d$. Under regularity conditions, and if the bandwidth a_n satisfies $a_n \to 0$ but with $a_n^d |K_n| \to \infty$ as $n \to \infty$, Karr (1986) showed asymptotic normality of $S(K_n)$ at rate $\tau_{|K_n|} = \sqrt{a_n^d |K_n|}$. Karr (1986) also calculated the asymptotic variance of $S(K_n)$ which is found to depend on the 4th order cumulant function of the random field $Z(s)$. Thus, to obtain confidence intervals for $\beta := R(s)$ using the asymptotic normal distribution it is necessary to estimate the 4th order cumulant function—unless this cumulant function is known to vanish, e.g., if the random field $Z(s)$ is known to be Gaussian. Nevertheless, our subsampling estimate of the variance of $S(K_n)$ applies immediately, and the difficult task of estimating the 4th order cumulant function is side-stepped.

4 Some finite-sample simulations and practical concerns

4.1 Practical approximations

In the previous section it was shown that subsampling marked point processes can be successfully used for moment estimation. However, the subsampling moment estimators such as $\overline{h}(B_n)$ are defined by integrals which have to be approximated by finite sums in practice. There are two general ways of performing this approximation, namely:

(a) Deterministic approximation: For example, the region K_n^{1-c} can be 'tiled' by a grid consisting of k 'small' cells, and then the integral $\overline{h}(B_n)$ can be approximated by the corresponding Riemann sum denoted by $\overline{h}_{Riemann,k}(B_n)$.

(b) Monte-Carlo or stochastic approximation: Points $y_1, \ldots, y_k$ can be dropped at random on K_n^{1-c}, (i.e., $y_1, \ldots, y_k$ are independent and identically distributed with the Uniform distribution on K_n^{1-c}), and the average $\overline{h}_{MC,k}(B_n) := k^{-1} \sum_{i=1}^{k} h(B_n + y_i)$ will then be a consistent approximation to $\overline{h}(B_n)$ as long as $k \to \infty$.

Both (a) and (b) are valid provided the number of points over which the approximations are computed is big enough; this is the issue of the following theorem, whose proof is standard. To state it, let us define P_k^* to be the probability law of the i.i.d. sequence $y_1, \ldots, y_k$, and let $\tilde{P}_k$ be the product probability measure of P with P_k^*. Finally, let $\tilde{E}_k$ denote expectation under $\tilde{P}_k$.

Theorem 3: *Under condition (7), and the assumptions of Theorem 1 it follows that*

$$\lim_{k,n \to \infty} E(\overline{h}_{Riemann,k}(B_n) - \gamma)^2 \longrightarrow 0,$$

and

$$\lim_{k,n \to \infty} \tilde{E}_k(\overline{h}_{MC,k}(B_n) - \gamma)^2 \longrightarrow 0.$$

4.2 On the choice of the subsample size

Some comments on proper choice of c are also in order. The assumption that $c \to 0$ but $cn \to \infty$ as $n \to \infty$ in our asymptotic results suffices for consistency of our subsampling moment estimators. Nevertheless, it is expected that a particular choice of the sequence $c = c_n$ may be 'optimal' with respect to some criterion; this choice however may depend on unknown features of the marked point process such as the strength of dependence in the $\{Z(s)\}$ random field.

Choosing the Mean Squared Error (MSE) as our optimality criterion, we note that the MSE-optimal choice of c strikes a balance in the well-known trade-off between the bias (squared) and the variance of $\hat{\theta}$. See also Künsch (1989), Lahiri (1996), and Hall *et al.* (1996, 1998) for 'block-size' choice in the one-dimensional lattice case. By analogy to the case of a random field observed on a lattice (see, e.g., Politis and Romano, 1993) we conjecture that to minimize the asymptotic order of the MSE of $\hat{\theta}$ we would need to take c asymptotically proportional to $n^{-2/(d+2)}$ in the case our statistic $S(K_n)$ is the sample mean or a closely related statistic (e.g., a smooth function of the sample mean or a smooth function of the first empirical marginal distribution). Unfortunately, the constant of proportionality required for direct practical implementation is generally intractable as it typically depends on unknown higher-order cumulant spectra. We now give a method for determining c which bypasses estimation of the constant of proportionality.

The algorithm to determine c is adapted from that in Hall and Jing (1996) who use a similar algorithm for equally spaced time series data; see also Hall et al. (1996). Assume that for a given shape K_n, we have that the optimal value of c, $c_n \simeq Cn^{-2/(d+2)}$, for some constant C . This implies that for a smaller (but still relatively large shape) K_m we have that $c_m \simeq Cm^{-2/(d+2)}$, and thus that $c_n \simeq (m/n)^{2/(d+2)}c_m$. But there is a natural empirical method to estimate c_m. For some "pilot" value of c, c' say, estimate θ by $\hat{\theta}$ using c'. Now consider the subshape K_m; by a procedure similar to the deterministic approximation of Sec. 4.1, consider N_m points on a grid, and the corresponding N_m subshapes of size K_m obtained by displacements. For each of those N_m subshapes of size K_m, estimate θ using a subshape of size K_{m^*} obtaining $\hat{\theta}_j^{m^*}$, for $j = 1, ..., N_m$. Then

$$R(m, m^*) := N_m^{-1} \sum_{j=1}^{N_m} (\hat{\theta}_j^{m^*} - \hat{\theta})^2$$

is an estimate of the risk of $\hat{\theta}^{m^*}$ as an estimate of θ. $\hat{c}_m$ is the value corresponding to the value of m^* that minimizes $R(m, m^*)$, i.e., $\hat{c}_m = m^{-1}\ \mathrm{argmin}_{m^*} R(m, m^*)$. The estimated value of c_n for the entire shape K_n is then simply,

$$\hat{c}_n = (m/n)^{2/(d+2)} \hat{c}_m.$$

This algorithm is a promising method of choosing block length as reported by Hall and Jing (1996); see also Sherman (1998). Unfortunately, in the current setting of irregularly spaced indices it is quite computationally intensive. Thus, we use this method to determine the choice of c in the data example in Section 5 but not in the simulations in the next section. We will use the results of the next section to study the efficacy of our estimator as c varies.

4.3 Some finite-sample simulations

Our asymptotic results in Section 2 show that our moment and variance estimators of general statistics are L_2 consistent. In this section we study the finite sample performance of the variance estimator $\hat{\theta}$ in estimating the variance of two location statistics: the sample mean ($\overline{X}_n$) computed from a marked point process on $\mathbf{R^2}$, i.e., in the $d = 2$ case and the sample median ($\tilde{X}_n$) computed from a marked point process on $\mathbf{R^1}$, i.e., in the $d = 1$ case.

Specifically, consider the marked point process with locations generated by a homogenous Poisson process with intensity λ observed on $K_n := [0, n]^d$. Thus, the expected number of points in K_n is λn^d. We chose $\lambda = 1.0$ without loss of generality in all cases. The underlying continuous time process generating the Z–'marks' is taken to be Gaussian, with covariance function $Cov[Z(s), Z(t)] = \exp(-\gamma|t - s|)$, for some $\gamma > 0$. The Gaussian process $Z(t)$ is therefore necessarily strong mixing, since the exponential decay of the autocovariances is sufficient to ensure that the $Z(t)$ possesses a (continuous) spectral density; see Theorem 7 of Rosenblatt (1985, p. 73)

Large γ corresponds to approximate independence, while γ near 0 corresponds to strong correlation if $|t - s|$ is small; regardless of the value of γ, if $|t - s|$ is large, $Cov[Z(s), Z(t)] \approx 0$. To assess the importance of the strength of correlation on the performance of the variance estimator $\hat{\theta}$ we consider two different values of γ in each of the following two examples.

4.3.1 The Sample Mean

The statistic $S(K_n) := N(K_n)^{-1} \sum_{i=1}^{N(K_n)} Z(s_i)$ is computed from two regions $K_n = [0, 100] \times [0, 100]$ and $K_n = [0, 500] \times [0, 500]$ to empirically verify the consistency stated in Theorem 2. The target variances for the two different strengths of dependence are estimated empirically from 10,000 simulated marked point processes to be 2.34 and 6.01, for $\gamma = 2.0$ and $\gamma = 1.0$, respectively. The former corresponds to a weakly dependent process, while the latter corresponds to a more strongly correlated process.

Two different subregion sizes are considered corresponding to $c = n^{-1/2}$ and $c = 2n^{-1/2}$. This gives insight into how performance of the estimator depends on subregion size. A deterministic 'tiled' approximation (see Section 4.1) with cell length equal to 1.0 was used in all cases. The results including the mean (and associated standard error), variance, and MSE's are given in Table 1. Each row is based on 1000 simulations.

From Table 1, we see that the theoretical consistency stated in Theorem 2 is empirically demonstrated. For fixed strength of correlation, γ, and subregion size, characterized by c, both the bias and the variance of the variance estimator decrease as the region size, n, increases thus giving finite-sample evidence of the theoretical L_2 convergence. For fixed strength of correlation and region size, however, there is a trade-off between the bias and variance of the estimator as the subregion size varies. The bias decreases and the variance increases as the subregion size increases. This is analogous to the results observed for equally spaced lattice data; see, e.g., Künsch (1989) in the $d = 1$ case.

4.3.2 The Sample Median

In this example the statistic is the sample median $S(K_n) := Med(Z(s_i), i = 1, ..., N(K_n))$ in the case d=1. The mechanism generating the marked point process uses the same covariance function as that in Section 4.3.1, with $\gamma = 1.0$ and $\gamma = 0.2$. Two values of c are $c = n^{-2/3}$ and $c = n^{-1/2}$. The results are given in Table 2. The conclusions are qualitatively similar to those obtained in Table 1. L_2 consistency is exhibited as n increases. Also, for fixed strength of correlation, the same bias/variance trade-off for varying subregion sizes occurs. We expect that other similar robust estimators of location, e.g., trimmed means, or other quantile estimators will exhibit qualitatively similar behavior.

5 A Real Data Example: Assessing the Width of Longleaf Pines

Cressie (1993) gives the exact locations and diameters at breast height (cm) of $N = 584$ Longleaf pine trees from the Wade Tract, an old growth forest in Thomas County, Georgia, U.S.A. The set K_n is a square region $200m \times 200m$, and thus we take $n = 200$. A variety of procedures show that there is significant positive clustering of the locations of the trees (see e.g. Cressie (1993) or Sherman (1996)). While allowing for this correlation we assume that the locations are generated by a stationary point process that satisfies (4). There is some debate about the stationarity assumption due to some apparent spatial trends, and thus we consider the following more as an illustration than as a substantive analysis. One statistic that crudely assesses the health of the forest is the sample mean of the breast heights, which is $\overline{x} = 26.9cm$ (Stoyan et al. (1995) arrive at $26.8cm$ based on $N = 583$ trees). In order to assess the stability of this estimate we desire an estimate of the standard deviation $SD(\overline{x})$. The sample variance of the heights is $s^2 = 336cm^2$ and thus a naive estimate of $SD(\overline{x})$ assuming independence is $\hat{SD}_{nai} = (336/584)^{1/2} = 0.76$. We now use the algorithm given in Section 4.2 to determine c and then use the resulting c to estimate θ. We will see shortly that $\hat{SD}_{nai}$ seriously underestimates the true variability in average height.

We apply the algorithm for each of two different pilot values c', taking $n^{-1/2} = .071$ and $2n^{-1/2} = .141$, as used in the simulations in Section 4.3.1, where we also have $d = 2$, and for $m = n/10 = 20$ and $m = n/5 = 40$. The resulting values are given in Table 3. All four tuning parameters give qualitatively similar results for $\hat{c}_n$, between $\hat{c}_n = .20$ and $\hat{c}_n = .25$. The former corresponds to $40m \times 40m$ subshapes while the latter corresponds to $50m \times 50m$. The first value gives $\hat{\theta}_1 = 315000$ while the second gives $\hat{\theta}_2 = 377000$. The former gives an estimated standard deviation of $\hat{SD}_1 = (\hat{\theta}_1/|K_n|)^{1/2} = (315000)^{1/2}/200 = 2.81$ while the latter gives an estimate of $\hat{SD}_2 = (377000)^{1/2}/200 = 3.07$. Both estimates are approximately 4 times larger than $\hat{SD}_{nai}$, which erroneously assumed an underlying structure of independent observations.

Acknowledgement. Many thanks are due to Prof. R. Bradley (Indiana University) for a discussion related to the reference Bradley (1993b), and to the associate editor for suggesting the connection between cluster point processes and Markov processes alluded to at the end of Sec. 2.1.

6 Technical proofs

We will first need a preliminary lemma; its proof is straightforward and thus it is omitted.

Lemma 1: *Let $h(K_n)$ be a statistic computed from the marked point process $(N, Z(\cdot))$. If N and $Z(\cdot)$ are each stationary, then $h(K_n)$ is also 'stationary', i.e., for any $x \in \mathbf{R}^d$ and all $y \in \mathbf{R}$:*

$$P[h(K_n) \leq y] = P[h(K_n + x) \leq y].$$

Proof of Theorem 1:

We consider the bias and variance of $\overline{h}(B_n)$, separately. Firstly, $E(\overline{h}(B_n)) = E(h(B_n)) \to \gamma$ by Lemma 1 and the definition of γ.

We next need to show that $Var(\overline{h}(B_n)) \to 0$. Note that we have for all $x, y \in \mathbf{R}^d$, and $l \in \mathbf{R}^1$:

$$Var(\overline{h}(B_n)) = \int_{K_n^{1-c}} \int_{K_n^{1-c}} Cov[h(B_n + x), h(B_n + y)] dxdy / |K_n^{1-c}|^2 = U_n + V_n,$$

where

$$U_n := \int\int_{d(B_n+x,B_n+y)\le l} Cov[h(B_n+x),h(B_n+y)]dxdy/|K_n^{1-c}|^2$$

and

$$V_n := \int\int_{d(B_n+x,B_n+y)>l} Cov[h(B_n+x),h(B_n+y)]dxdy/|K_n^{1-c}|^2.$$

Now, let $l = |B_n|^{1/d}$. Note that for each fixed y, $B_n + y$ is contained in a d-dimensional cube with sides of length cn, and thus for each x we have that

$$\int_{d(B_n+x,B_n+y)\le l} dy \le [2l+3cn)]^d \le 5^d c^d n^d.$$

Hence,

$$\begin{aligned} U_n &\le [Var(h(B_n))/|K_n^{1-c}|^2][\int\int_{d(B_n+x,B_n+y)\le l} dxdy] \\ &\le \text{const.}c^d\ Var(h(B_n))|K_n^{1-c}||K_n|/|K_n^{1-c}|^2 \to 0 \end{aligned}$$

since $c \to 0$, $|K_n|/|K_n^{1-c}| \to 1$, and since $Var(h(B_n)) \le C_\delta^{1/(2+\delta)}$ by Jensen's inequality and our assumption on $E[|h(K_n)|^{2+\delta}$.

Next we consider V_n. For any random elements X, Y, Z we have

$$Cov(X,Y) = E[Cov^Z(X,Y)] + Cov[E^Z X, E^Z Y] \tag{13}$$

where E^Z, Var^Z and Cov^Z denotes expectation, variance and covariance conditional on Z. Thus, letting $X = h(B_n + x)$, $Y = h(B_n + y)$, $Z = N$, we have

$$Cov[h(B_n{+}x),h(B_n{+}y)] = E[Cov^N(h(B_n{+}x),h(B_n{+}y))]{+}Cov[E^N h(B_n{+}x),E^N h(B_n{+}y)].$$

For any x, y in the integral defining V_n we have

$$Cov^N(h(B_n+x),h(B_n+y)) \le C_\delta \alpha_Z^{\delta/(2+\delta)}(l;|B_n|),$$

by a moment inequality – see Roussas and Ioannides (1987) or Doukhan (1994). Now taking expectations we have that

$$E[Cov^N(h(B_n+x),h(B_n+y))] \le C_\delta \alpha_Z^{\delta/(2+\delta)}(l;|B_n|)$$

as well.

Similarly, it is easy to see that

$$Var(E^N h(B_n + x)) = E[E^N h(B_n + x)]^2 - [Eh(B_n + x)]^2$$

$$\leq E[E^N h^2(B_n + x)] - [Eh(B_n + x)]^2 = Var(h(B_n + x)) \leq C_\delta^{1/(2+\delta)}.$$

Thus,

$$Cov[E^N h(B_n + x), E^N h(B_n + y)] \leq C_\delta^{1/(2+\delta)} \rho_N(l; |B_n|),$$

since $E^N h(B_n + x)$ and $E^N h(B_n + y)$ are measurable functions of the subcollections of random points generated by the point process N in the sets $B_n + x$ and $B_n + y$ respectively.

Putting this all together we see that $Var(\overline{h}(B_n)) \to 0$, due to the assumed mixing conditions and the fact that by definition $|B_n| = l^d$. □

Proof of Theorem 2:

We will need the following:

Lemma 2 [Chung, 1974]:

Let Y_n be a sequence of random variables such that $Y_n \xrightarrow{p} Y$, and for $0 < r < \infty$, $E|Y|^r < \infty$. Then, $|Y_n|^r$ is uniformly integrable if and only if $Y_n \xrightarrow{L_r} Y$.

The proof is now entirely analogous to that of Theorem 2 in Sherman (1996). Specifically, recall that

$$\hat{\theta} = \int_{K_n^{1-c}} t^2(B_n + y)dy/|K_n^{1-c}| - [\int_{K_n^{1-c}} t(B_n + y)dy/|K_n^{1-c}|]^2 = U_n - V_n$$

(say) where $t(B_n)$ was defined in Sec. 3.1.

For U_n, note that $E[t^2(K_n)] \to \theta$ and thus $U_n \xrightarrow{L_2} \theta$ by applying Theorem 1 with $h(K) = t^2(K)$.

For V_n, we let $Y_n := \int_{K_n^{1-c}} t(B_n + y)dy/|K_n^{1-c}|$, and note that $V_n = Y_n^2$. Note that $E(Y_n) = 0$ implies that $Y_n \xrightarrow{L_2} 0$ by Theorem 1, and thus that $Y_n \xrightarrow{p} 0$. Also, $E(Y_n^4) \leq E(U_n^2)^2$, and $U_n \xrightarrow{L_2} \theta$ implies $U_n \xrightarrow{p} \theta$. U_n is uniformly integrable (by

Lemma 2), and hence Y_n^4 is uniformly integrable. Finally, using Lemma 2 once again, we have that Y_n^4 is uniformly integrable and $Y_n \xrightarrow{p} 0$ together imply that $Y_n \xrightarrow{L_4} 0$, i.e., $V_n \xrightarrow{L_2} 0$ as desired. $\square$

References

[1] Baddeley, A. J., van Lieshout, M. N. M., and Moeller, J. (1996), Markov properties of cluster processes. *Adv. Appl. Probab.*, 28, 346–355.

[2] Bingham, N.H., Goldie, C.M. and Teugels, J.L. (1987), *Regular Variation*, Cambridge University Press.

[3] Bradley, R.C. (1993a), Some examples of mixing random fields, *Rocky Mountain Journal of Mathematics*, 23, 495-519.

[4] Bradley, R.C. (1993b), Equivalent mixing conditions for random fields, *Ann. Prob.*, 21, 1921-1926.

[5] Carlstein, E. (1986), The use of subseries values for estimating the variance of a general statistic from a stationary sequence, *Ann. Statist.*, 14, 1171-1179.

[6] Chung, K.L. (1974), A course in probability theory, 2nd ed., Harcourt Brace Jovanovich, New York.

[7] Cressie, N. (1993), *Statistics for Spatial Data*, John Wiley, New York.

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

[9] Doukhan, P. (1994), *Mixing: Properties and Examples*, Lecture Notes in Statistics # 85, Springer, New York.

[10] Gastwirth, J. L. and Rubin, H. (1975), The behavior of robust estimators on dependent data, *Annals of Statistics*, 3, 1070-1100.

[11] Hall, P. (1985), Resampling a coverage pattern, *Stoch. Process. Appl.*, 20, 231-246.

[12] Hall, P., Jing, B. (1996), On sample re-use methods for dependent data, *J. of the Royal Stat. Soc., Ser.B*, 58, 727-738.

[13] Hall, P., Horowitz J.L., and Jing, B. (1996). On blocking rules for the bootstrap with dependent data, *Biometrika*, 50, 561–574.

[14] Hall, P., Jing, B., Lahiri, S.N (1998), On the sampling window method under long range dependence, *Statistica Sinica*, 8, 1189-1204.

[15] Jensen, J.L. (1993a), A note on asymptotic expansions for sums over a weakly dependent random field with application to the Poisson and Strauss processes, *Ann. Inst.Statist. Math.*, 45, 353-360.

[16] Jensen, J.L. (1993b), Asymptotic normality of estimates in spatial point processes, *Scand. J. Statist.*, 20, 97-109.

[17] Karr, A. F. (1986), Inference for stationary random fields given Poisson samples, *Adv. Appl. Prob.*, 18, 406-422.

[18] Karr, A. F. (1991), *Point Processes and their Statistical Inference*, 2nd ed., Marcel Dekker, New York.

[19] Krickeberg, K. (1982), Processus ponctuels en statistique, *Ècole d' été de probabilités de Saint-Flour X – 1980*, P.L. Hennequin (Ed.), *Lecture Notes in Mathematics No. 929*, Springer, Berlin, pp. 205-313.

[20] Künsch, H.R. (1982), Decay of correlations under Dobrushin's uniqueness condition and its applications, *Communications in Mathematical Physics*, 84, 207-222.

[21] Künsch, H.R. (1989), The jackknife and the bootstrap for general stationary observations, *Ann. Statist.*, 17, 1217-1241.

[22] Lahiri, S.N. (1993), On the moving block bootstrap under long range dependence, *Statistic and Prob. Letters*, 18, 405-413.

[23] Lahiri, S.N. (1996), On empirical choice of the optimal block size for block bootstrap methods, Technical Report 96-14, Dept. of Statistics, Iowa State University.

[24] Lahiri, S.N., Kaiser, M.S., Cressie, N.A., and Hsu, N.J., (1999), Prediction of spatial cumulative distribution functions using subsamples, *J. Amer. Statist. Assoc.*, 94, 86-110.

[25] Liu, R.Y. and Singh, K. (1992), Moving blocks jackknife and bootstrap capture weak dependence, in *Exploring the Limits of Bootstrap*, (edited by Raoul LePage and Lynne Billard), John Wiley, pp. 225-248.

[26] Politis, D.N., Paparoditis, E. and Romano, J.P. (1998), Large sample inference for irregularly spaced dependent observations based on subsampling, *Sankhya, Series A.*, 60, 274-292.

[27] Politis, D.N., Paparoditis, E. and Romano, J.P. (1999), Resampling marked point processes, in *Multivariate Analysis, Design of Experiments, and Survey Sampling: a Tribute to J. N. Srivastava*, Subir Ghosh (Ed.), Marcel Dekker, Inc., New York, pp. 163-185.

[28] Politis, D.N. and Romano, J.P. (1992), A circular block-resampling procedure for stationary data, in *Exploring the Limits of Bootstrap*, (edited by Raoul LePage and Lynne Billard), John Wiley, pp. 263-270.

[29] Politis, D.N., and Romano, J.P. (1993). Nonparametric resampling for homogeneous strong mixing random fields, *J. Multivar. Anal.*, 47, 301-328.

[30] Politis, D.N. and Romano, J.P. (1994), 'Large sample confidence regions based on subsamples under minimal assumptions', *Ann. Statist.*, 22, 2031-2050.

[31] Ripley, B.D. (1981), *Spatial Statistics*, Wiley, New York.

[32] Roussas, G.G. and Ioannides, D. (1987), Moment inequalities for mixing sequences of random variables, *Stoch. Analysis and Applications*, 5, p.61-120, Marcel Dekker.

[33] Sherman, M. and Carlstein, E. (1994), Nonparametric estimation of the moments of a general statistic computed from spatial data, *J. Amer. Statist. Assoc.*, 89, 496-500.

[34] Sherman, M. (1996), Variance estimation for statistics computed from spatial lattice data, *J. of the Royal Stat. Soc., Ser. B*, 58, 509-523.

[35] Sherman, M. (1998), Data-based choice of batch size for simulation output analysis, *Simulation*, 71, 38-47.

[36] Stoyan, D., Kendall, W.S., and Mecke, J. (1995), *Stochastic Geometry and its Applications*, Wiley, Chichester.