The Annals of Statistics
2000, Vol. 28, No. 6, 1503–1532

ESTIMATING THE *K* FUNCTION OF A POINT PROCESS WITH AN APPLICATION TO COSMOLOGY[1]

BY MICHAEL L. STEIN[1], JEAN M. QUASHNOCK[2] AND JI MENG LOH[1]

University of Chicago, University of Chicago and Carthage College and University of Chicago

Motivated by the study of an important data set for understanding the large-scale structure of the universe, this work considers the estimation of the reduced second-moment function, or K function, of a stationary point process on $\mathbb{R}$ observed over a large number of segments of possibly varying lengths. Theory and simulation are used to compare the behavior of isotropic and rigid motion correction estimators and some modifications of these estimators. These results generally support the use of modified versions of the rigid motion correction. When applied to a catalog of astronomical objects known as absorbers, the proposed methods confirm results from earlier analyses of the absorber catalog showing clear evidence of clustering up to 50 h^{-1} Mpc and marginal evidence for clustering of matter on spatial scales beyond 100 h^{-1} Mpc, which is beyond the distance at which clustering of matter is now generally accepted to exist.

1. Introduction. One way to describe a stationary spatial point process is through some measure of clumpiness of the events of the process. A commonly used measure of clumpiness is the reduced second-moment function $K(t)$, defined as the expected number of events within distance t of a typical event of the process divided by the intensity of the process. For a homogeneous Poisson process on $\mathbb{R}^d$, $K(t) = \mu_d t^d$, where μ_d is the volume of a unit ball in d dimensions. Thus, values of $K(t)$ greater than $\mu_d t^d$ are indicative of a process that is clumpier than Poisson and values less than $\mu_d t^d$ are indicative of a process that is more regular than Poisson. When estimating $K(t)$ based on observing a process within a bounded window W, a central problem is that for any event in W that is within t of the boundary of W, we do not know for sure how many other events are within t of it. Baddeley (1998) describes a number of ways of accounting for these edge effects. Although there is quite a bit of asymptotic theory for how these estimators behave when the underlying process is Poisson [Ripley (1988) and Stein (1993)], much less is known for non-Poisson processes.

An interesting aspect of asymptotic theory for point processes is how one should take limits. Ripley (1988) and Stein (1993) consider a single growing window, which might appear to be the obvious way to take limits. However, Baddeley, Moyeed, Howard and Boyde (1993) describe applications in which

Received March 1999; revised May 2000.
[1]Supported in part by NSF Grants DMS 95-04470 and DMS 99-71127.
[2]Supported in part by NASA Grant NAG 5-4406 and NSF Grant DMS 97-09696.
AMS 2000 *subject classifications*. Primary 62M30; secondary 62P35, 60G55.
Key words and phrases. Reduced second-moment function, bootstrapping, large-scale structure of the universe, heavy-element absorption-line systems.

point processes are observed in many well-separated windows. For this setting, Baddeley and Gill (1997) argue that it is natural to consider taking limits by keeping the size of these windows fixed and letting their number increase. As they point out, one advantage of this approach is that the edge effects do not become negligible in the limit, since for any fixed t, the fraction of events that are within t of a window boundary does not tend to 0. Thus, for comparing different approaches for handling edge effects, increasing the number of windows may be more informative than allowing a single region to grow in all dimensions, for which the fraction of events that are within t of a window boundary does tend to 0. Another advantage of taking limits by letting the number of windows increase is that if the process is independent in different regions, then limit theorems are easier to prove. This is particularly the case when the windows are all well-separated translations of the same set so that the observations of the process on the multiple windows can be reasonably modeled as iid realizations. Baddeley and Gill (1997) use this approach to obtain weak convergence results for estimators of K and other functions describing point process behavior. The resulting limiting variances are difficult to evaluate and Baddeley and Gill (1997) only give explicit results for what they call the sparse Poisson limit, in which the intensity of a homogeneous Poisson process tends to 0.

This work studies the estimation of K for a process on $\mathbb{R}$ when the windows are segments of varying lengths. The fact that the windows are one-dimensional greatly simplifies the calculation of estimators and permits the explicit derivation of some of their properties. The fact that the segment lengths vary provides for an interesting wrinkle on the approach of Baddeley and Gill (1997). Notably, simulation results in Section 6 show that the differences between certain estimators are much greater when the segment lengths the unequal.

Section 2 describes a cosmological problem that motivated the present study. Vanden Berk, Quashnock, York and Yanny (1996) put together a catalog of what are known as absorption-line systems, or absorbers, detected along the lines of sight of QSO's (quasi-stellar objects, or quasars). This catalog, a preliminary version of which can be obtained from Daniel Vanden Berk (danvb@fnal.gov), provides important evidence about the large-scale structure of the universe. To a first approximation, in appropriate units, the locations of these absorbers along the lines of sight can be viewed as multiple realizations of a stationary point process along segments of varying length.

Section 3 describes the estimators of K used in this paper and gives explicit expressions for the commonly used rigid motion correction and isotropic correction estimators when the observation region is a collection of line segments of varying lengths. In addition, Section 3 provides an explicit expression for a modification to the rigid motion correction advocated in Stein (1993). The fact that this estimator can be calculated explicitly is in contrast to the situation in more than one dimension, in which case calculating this modified rigid motion correction requires numerous numerical integrations even for simple regions such as circles and rectangles. Finally, following on an idea of Picka

(1996), Section 3 introduces another approach to modifying the rigid motion correction and isotropic correction. When the underlying process is homogeneous Poisson, Picka's modification of the rigid motion correction has similar properties to the estimator proposed in Stein (1993), but theoretical results in Section 5 and simulation results in Section 6 suggest that his approach may have some advantage and we recommend the adoption of the resulting estimator for routine use.

When the underlying process is homogeneous Poisson, Section 4 derives some asymptotic theory for the various estimators as the number of segments on which the process is observed increases. As in the case of a single growing observation window studied in Stein (1993), the modified rigid motion correction asymptotically minimizes the variance of the estimator of $K(t)$ among a large class of estimators possessing a type of unbiasedness property. Furthermore, if the segments are of equal length, then it is possible to give explicit comparisons between various estimators. In particular, the ratio of the asymptotic mean squared error of the ordinary rigid motion correction to that of the modified rigid motion correction equals 1 plus a positive term proportional to the expected number of events per line segment. Thus, the benefit of the modification is modest when this expectation is small, around 1, say, but can be quite substantial when this expectation is large.

Section 5 considers asymptotic results when the underlying process is not necessarily homogeneous Poisson, the segments are all of equal length and the processes on different segments are independent. In this case, it is essentially trivial to obtain a central limit theorem for the estimators of K used here. From the general result, it is difficult to make comparisons between the various estimators. However, if the processes on the different segments are each homogeneous Poisson but with intensities that vary from segment to segment according to some sequence of iid positive random variables, it is possible to give simple expressions for the asymptotic variances of the rigid motion correction and the two modifications of this estimator. These results show that the modification in Stein (1993) has strictly smaller asymptotic variance than the ordinary rigid motion correction. Furthermore, the modification of Picka (1996) has strictly smaller asymptotic variance than the modification in Stein (1993) unless the random intensities have zero variance, in which case the two modified estimators have equal asymptotic variance.

Section 6 reports on the results of a simulation study comparing the ordinary rigid motion correction and the two modifications for both Poisson and non-Poisson processes and equal and unequal segment lengths. While there is no theory showing the general superiority of the modified estimators for non-Poisson processes, the modified estimators do, for the most part, outperform the unmodified estimator. The advantage of the modified estimators tends to be larger when the process is more regular than Poisson, when the segment lengths are unequal and when t is near the length of the longest available segment.

Section 7 applies the rigid motion correction and the two modifications of it described in Section 3 to the estimation of K for the absorber catalog. In addition, approximate confidence intervals are obtained using bootstrapping based on viewing the segments as the sampling units. All three estimates are similar and confirm the finding in Quashnock and Stein (1999) of clear evidence of clustering up to at least 50 h^{-1} Mpc. In addition, the confidence intervals based on the modified procedures produce a slightly stronger case for clustering of absorbers beyond 100 h^{-1} Mpc. Whether there is clustering of matter at such large scales and for the high redshifts in the absorber catalog is a critical issue in modern cosmology, since presently used models for the evolution of the universe have difficulty explaining such clustering [Steidel, Adelberger, Dickinson, Giavalisco, Pettini and Kellogg (1998) and Jing and Suto (1998)].

2. The absorber catalog. The cosmological principle, which states that on large enough spatial scales the distribution of matter in the universe is homogeneous and isotropic, is a central tenet of modern cosmology [Peebles (1993)]. In cosmology, it is convenient to measure distances in units of h^{-1} Mpc, where Mpc, or megaparsec, is 3.26×10^6 light years and h is an inexactly known dimensionless number that is believed to be between 0.5 and 0.75. As is common in the cosmological literature, in reporting distances determined from redshifts, we will assume that Hubble's constant, H_0, equals 100 h km s^{-1} Mpc^{-1}. To help calibrate one's thinking about such distances, 1 h^{-1} Mpc is a typical distance between neighboring galaxies. It is now generally agreed that galaxies cluster up to scales of 10–20 h^{-1} Mpc [Davis and Peebles (1983) and Loveday, Maddox, Efstathiou and Peterson (1995)]. Furthermore, clustering on such scales can be reproduced by computer simulations of the evolution of the universe based on our present understanding of this evolution [see Zhang, Meiksin, Anninos and Norman (1998) and the references therein]. However, there is some evidence of clustering of matter on scales of up to 100 h^{-1} Mpc [see Quashnock, Vanden Berk and York (1996) and the references therein] and a few cosmologists have speculated that clustering may exist at all spatial scales [Coleman and Pietronero (1992) and Sylos Labini, Montuori and Pietronero (1998)], despite the fact that clustering at all scales contradicts both the cosmological principle and the considerable evidence that supports it [Peebles (1993), pages 20, 45 and 221]. Thus, determining the extent to which clustering of matter is present is of fundamental importance to modern cosmology.

One way to measure the clustering of matter is through the direct observation of large numbers of galaxies. Several galaxy surveys in various regions of the sky have been done in recent years [Martínez (1997)]; Pons-Bordería, Martínez, Stoyan, Stoyan and Saar (1999) describe recent work on estimating second-moment structures of galaxy locations from such surveys. The presently ongoing Sloan Digital Sky Survey will be by far the largest such survey and will contain roughly 10^8 galaxies, approximately 10^6 of which will have spectroscopically measured redshifts [Margon (1999)]. An object's

redshift gives its velocity relative to the Earth, which, using Hubble's law, yields its approximate distance from the Earth. Galaxy surveys are limited by the fact that galaxies are difficult to observe directly beyond several hundred h^{-1} Mpc. QSO's, on the other hand, are extremely bright and focused objects that can be readily detected at distances of several thousand h^{-1} Mpc, going back to nearly the beginning of the universe. Matter that falls on the line of sight between the QSO and the Earth can absorb light from the QSO and thus be detected from the Earth, even though this matter cannot be directly observed. Certain types of matter absorb light in a characteristic pattern of frequencies that can be used to identify the matter and, through the redshift of this absorption pattern, the relative velocity of this matter to the Earth. Astronomical objects detected in this way are called absorption-line systems, or absorbers. As noted by Crotts, Melott and York (1985), catalogs of absorbers provide a means for estimating the clustering of matter over very large spatial scales. Vanden Berk, Quashnock, York and Yanny (1996), Quashnock, Vanden Berk and York (1996) and Quashnock and Vanden Berk (1998) make use of an extensive catalog of heavy-element absorption-line systems drawn from the literature to investigate the clustering of matter at various scales. York, Yanny, Crotts, Carilli, Garrison and Matheson (1991) describe an earlier version of this catalog and a preliminary version of an updated catalog is available from Daniel Vanden Berk (danvb@fnal.gov). Here we will use the same absorber catalog as in Quashnock and Stein (1999), who examined clustering in 352 C IV absorbers (absorbers detected from the absorption-line patterns of C IV, or triply ionized carbon) along 274 QSO lines of sight. Although the relationship between C IV absorbers and galaxies is unclear, they do appear to track the general spatial patterns of galaxies [Lanzetta, Bowen, Tytler and Webb (1995) and Quashnock and Vanden Berk (1998)], and hence provide a plausible means for assessing the clustering of visible matter on large scales.

Because the universe expands over time and, due to the finite velocity of light, the more distant an object the further in the past we observe it, the method used for converting redshifts into distances from Earth is critical to the analysis of this catalog. Redshifts are generally denoted by z and, according to Hubble's law, an object observed at redshift z is seen at a time when distances between objects were approximately $(1 + z)^{-1}$ times their present values. To correct for the expansion, here, as in Quashnock and Stein (1999), we use what are called comoving coordinates, which scale up all distances to what they would be today if all the matter in the universe moved exactly with the Hubble flow [Peebles (1993)]. Thus, in examining the clustering of absorbers in comoving coordinates, we have removed the most important effects of the universe's expansion. If one did not make this correction, the volume density of absorbers would drop approximately like $(1 + z)^3$ as z decreases and we move toward the present.

For various reasons, it is only possible to detect C IV absorbers along a segment of each line of sight. The mean length of these segments in comoving units is 303.3 h^{-1} Mpc, with a range of 7.5 h^{-1} Mpc to 439.8 h^{-1} Mpc. For this catalog, the median redshift of the absorbers is about 2.2, with the

bulk of absorbers having redshifts from about 1.5 to 3. Our analysis acts as if clustering is both stationary in time and homogeneous in space. We are more accurately examining an average clustering over the range of redshifts in the sample at a cosmic epoch corresponding to a characteristic redshift of 2.2 (when the universe was about 1/3 its present scale and about 1/6 its present age). Section 7 provides further discussion of this issue and its possible influence on our results.

As in Quashnock and Stein (1999), we will act as if the absorber catalog can be viewed as multiple partial realizations of some stationary point process on $\mathbb{R}$ along a series of segments. In particular, we will not attempt to use any information about the physical location of these segments in three-dimensional space. Using this simplification, we will then be able to apply the methods described in the next section to the absorber catalog.

3. Methodology. Suppose $M_1, \ldots, M_p$ are simple, stationary point processes on $\mathbb{R}$ with a common probability law having intensity λ and reduced second-moment function K. We do not necessarily assume that $M_1, \ldots, M_p$ are independent. For a Borel subset A of $\mathbb{R}$, let $M_j(A)$ be the number of events of M_j contained in A. If $[0, Q_j]$ is the interval on which we observe M_j, then we can write the observation domain as $D = \bigcup_{j=1}^{p} \{[0, Q_j], j\}$, so that $(x, l) \in D$ implies $l \in \{1, \ldots, p\}$ and $x \in [0, Q_l]$. Define $N_j = M_j([0, Q_j])$, $N_+ = \sum_{j=1}^{p} N_j$ and denote the realized value of N_+ by n. For $j = 1, \ldots, N_+$, let (X_j, L_j) be the random locations of these observed events with realized values (x_j, l_j) for $j = 1, \ldots, n$.

The basic principle behind all edge-corrected estimators of K described by Ripley (1988) is to first find an exactly unbiased estimator of $\lambda^2 \times$ volume of observation domain $\times K(t)$ and then to divide by an estimator of $(\lambda^2 \times \text{volume})$. Here, the volume of the observation domain is $Q_+ = \sum_{j=1}^{p} Q_j$. For a symmetric function ϕ on $D \times D$, define $T(\phi) = \sum_{j \neq k} \phi((X_j, L_j), (X_k, L_k))$. Then the unbiasedness constraint requires that

$$ET(\phi) = \lambda^2 Q_+ K(t) \tag{1}$$

for any reduced moment function K. Estimating λ^2 by $N_+(N_+ - 1)/Q_+^2$ yields

$$\widetilde{K}(t) = \begin{cases} \dfrac{Q_+ T(\phi)}{N_+(N_+ - 1)}, & \text{if } N_+ > 1, \\ 0, & \text{otherwise,} \end{cases}$$

as a natural estimator of $K(t)$.

There is an infinite array of functions ϕ satisfying (1). Two popular choices are the rigid motion correction [Ohser and Stoyan (1981)] and the isotropic correction [Ripley (1976)]. Asymptotic results in Sections 4 and 5 suggest that modified versions of the rigid motion correction have good large-sample properties when the underlying process is Poisson, so we focus on this correction here, although we also give some results for the isotropic correction for comparison.

It is fairly elementary to prove that the rigid motion correction satisfies (1) when the observation domain D is a subset of $\mathbb{R}$. First, for a stationary point process M on $\mathbb{R}$ with intensity λ, define the reduced second-moment measure $\mathscr{K}$ by $\lambda^2\mathscr{K}(ds)dx = 2E\{M(dx)M(x+ds)\}$, in which case the reduced second-moment function K is given by $K(t) = \int_{(0,\,t]}\mathscr{K}(ds)$. Denote the indicator function by $1\{\cdot\}$ and use $|A|$ to indicate the Lebesgue measure of the set $A \subset \mathbb{R}$ and A_s to indicate the set A translated by the amount s. The rigid motion correction is given by

$$\phi(x,\,y) = \frac{1\{|x-y| \le t\}|D|}{|D \cap D_{x-y}|}.$$

we can then write

$$\begin{aligned} T(\phi) &= \int_{s\in[-t,\,0)\cup(0,\,t]}\int_{x\in\mathbb{R}} M(dx)M(x+ds)\frac{1\{x\in D,\ x+s\in D\}}{|D\cap D_s|} \\ &= 2\int_{s\in(0,\,t]}\int_{x\in\mathbb{R}} M(dx)M(x+ds)\frac{1\{x\in D,\ x+s\in D\}}{|D\cap D_s|}, \end{aligned}$$

so that

$$\begin{aligned} E\{T(\phi)\} &= 2\int_{s\in(0,\,t]}\int_{x\in\mathbb{R}} \frac{1}{2}\lambda^2\mathscr{K}(ds)\frac{1\{x\in D,\ x+s\in D\}}{|D\cap D_s|}\,dx \\ &= 2\int_{s\in(0,\,t]}\frac{1}{2}\lambda^2\frac{|D\cap D_s|}{|D\cap D_s|}\mathscr{K}(ds) \\ &= \lambda^2 K(t). \end{aligned}$$

One way to view the setting where D is a collection of line segments is to think of these segments as being widely spaced intervals on $\mathbb{R}$, in which case we just have a special case of the treatment in the preceding paragraph. However, it will be helpful in the subsequent development to think of D as $\bigcup_{j=1}^{p}\{[0,\,Q_j],\,j\}$. The rigid motion correction can then be defined by taking ϕ to be

$$\phi^R((x,\,k),\,(y,\,l)) = \frac{Q_+1\{|x-y|\le t,\ k=l\}}{\sum_{j=1}^{p}(Q_j-|x-y|)^+}.$$

To write the isotropic correction in terms of a symmetric function, let

$$\phi^I((x,\,k),\,(y,\,l)) = \frac{Q_+1\{x-y|\le t,\ k=l\}\{\alpha_l(x,\,y)+\alpha_l(y,\,x)\}}{Q_+ - \sum_{j=1}^{p}\min\{(2|x-y|-Q_j)^+,\,Q_j\}}, \tag{2}$$

where $\alpha_l(x,\,y)^{-1} = 1\{x+|y-x|<Q_l\}+1\{x-|y-x|>0\}$. Define $\widetilde{K}_R(t) = Q_+T(\phi^R)/\{N_+(N_+-1)\}$ and $\widetilde{K}_I(t) = Q_+T(\phi^I)/\{N_+(N_+-1)\}$, where it is understood that $\widetilde{K}_R(t) = \widetilde{K}_I(t) = 0$ for $N_+ \le 1$. We have used Ohser's extension of the isotropic correction to cover the case $t > \frac{1}{2}\min(Q_1,\ldots,Q_p)$ [Ohser (1983)]. As Ripley [(1988), page 32] notes, this extension is generally not of much practical value when there is a single contiguous observation

window. However, when there are multiple windows of various sizes, the extension is critical. For the absorber catalog, for example, one is certainly interested in estimating K at distances greater than 3.75 h^{-1} Mpc, the value of $\frac{1}{2}\min(Q_1, \ldots, Q_p)$ in the catalog.

Note that $\phi^I((x, k), (y, l)) = \phi^R((x, k), (y, l)) = 0$ if $k \neq l$, which just says that pairs of observations on different segments do not contribute to the estimate of $K(t)$. Since we have made no assumption about the joint distribution of $M_1, \ldots, M_p$, for (1) to be valid, it is necessary to assume $\phi((x, k), (y, l)) = 0$ whenever $k \neq l$. Thus, throughout this work, we will only consider ϕ satisfying

(A) $\phi((x, k), (y, l)) = 0$ for $k \neq l$.

We next show how to apply to the present setting the method developed in Stein (1993) for improving upon any estimator of K of the form $Q_+T(\phi)/\{N_+ \times (N_+ - 1)\}$ with ϕ satisfying (1). Suppose (X, L) is uniformly distributed on D in the sense that $P(L = l) = Q_l/Q_+$ and the density of X given $L = l$ is uniform on $[0, Q_l]$. Then $M_1, \ldots, M_p$ stationary with common distribution imply that, for any real-valued function g for which $E|g(X, L)| < \infty$, $E\sum_{j=1}^{N_+} g(X_j, L_j) = \lambda Q_+ Eg(X, L)$, so that $\sum_{j=1}^{N_+}\{g(X_j, L_j) - Eg(X, L)\}$ is an unbiased estimator of 0. The idea in Stein (1993) is to choose g to minimize

$$\mathrm{var}_n\left[T(\phi) - \sum_{j=1}^{n}\{g(X_j, L_j) - Eg(X, L)\}\right],$$

where var_n means to compute the variance under binomial sampling: $N_+ = n$ is fixed and, for $j = 1, \ldots, n$, (X_j, L_j) are independent and all have the same distribution as (X, L). Proposition 1 in Stein (1993) shows that, for $n \geq 1$ and $(y, m) \in D$, a minimizing g is $2(n-1)h(y, m; \phi)/Q_+$, where $h(y, m; \phi) = \sum_{l=1}^{p}\int_0^{Q_l} \phi((x, l), (y, m))\,dx$. Under (A), $h(y, m; \phi) = \int_0^{Q_m} \phi((x, m), (y, m))\,dx$.

Now define

$$T^*(\phi) = T(\phi) - \frac{2(N_+ - 1)}{Q_+}\sum_{j=1}^{N_+}\{h(X_j, L_j; \phi) - Eh(X, L; \phi)\}.$$

Note that if ϕ satisfies (1), $Eh(X, L; \phi) = 2t$. Under binomial sampling, we always have $\mathrm{var}_n\{T^*(\phi)\} \leq \mathrm{var}_n\{T(\phi)\}$. This suggests that the estimator $\widehat{K}(t) = Q_+T^*(\phi)/\{N_+(N_+ - 1)\}$ for $N_+ > 1$ and 0 otherwise may be preferred over $\widetilde{K}(t)$. As with the unmodified estimators, $\widehat{K}_R(t)$ indicates that $\phi = \phi^R$ and $\widehat{K}_I(t)$ indicates that $\phi = \phi^I$.

Picka (1996) suggests another approach to modifying estimates of second-moment measures. He considered random sets for which the probability of any fixed point being in the random set is positive, but his approach can also be applied to point processes, for which this probability is 0. For point processes, his idea corresponds to using an estimator of λQ_+ other than N_+ in $\widetilde{K}$. For any real-valued function c on D satisfying $\sum_{l=1}^{p}\int_0^{Q_l} c(x, l)\,dx =$

Q_+, $\hat{\lambda}_c = Q_+^{-1}\sum_{j=1}^{N_+} c(X_j, L_j)$ is an unbiased estimator of λ. Let us consider estimators of $K(t)$ of the form $Q_+T(\phi)/\{\hat{\lambda}_c Q_+(\hat{\lambda}_c Q_+ - 1)\}$. It is not generally possible to calculate the exact variance of such estimators under binomial sampling. However, for Q_+ sufficiently large, $\hat{\lambda}_c - \lambda$ and $Q_+^{-1}T(\phi) - \lambda^2 K(t)$ should be small in probability, which suggests using a first-order Taylor series approximation to obtain

$$\frac{Q_+T(\phi)}{\hat{\lambda}_c Q_+(\hat{\lambda}_c Q_+ - 1)} \approx \frac{1}{\lambda^2 Q_+}T(\phi) - \frac{2K(t)}{\lambda}(\hat{\lambda}_c - \lambda). \tag{3}$$

For a given ϕ and subject to c satisfying the unbiasedness constraint, now consider minimizing the variance of the right-hand side of (3) when $M_1, \ldots, M_p$ are iid Poisson processes with intensity λ. It is a straightforward variational problem to show that a minimizing c is given by $c(x, l; \phi) = h(x, l; \phi)/(2t)$. Define

$$\check{K}(t) = \frac{Q_+T(\phi)}{\sum_{j=1}^{N_+} c(X_j, L_j; \phi)\Big\{\sum_{j=1}^{N_+} c(X_j, L_j; \phi) - 1\Big\}}$$

for $N_+ > 1$ and $\check{K}(t) = 0$ otherwise. As with $\widetilde{K}$ and $\widehat{K}$, subscripts R or I on $\check{K}$ indicate that $\phi = \phi^R$ or $\phi = \phi^I$.

When $M_1, \ldots, M_p$ are iid Poisson processes, $\widehat{K}(t)$ and $\check{K}(t)$ should behave similarly. To see this, first use Taylor series to obtain

$$\widehat{K}(t) \approx \frac{1}{\lambda^2 Q_+}T(\phi) - \frac{2}{\lambda Q_+}\sum_{j=1}^{N_+} h(X_j, L_j; \phi) + 2\big\{2t - K(t)\big\}\frac{N_+}{\lambda Q_+} + 2K(t).$$

From this approximation and (3), when $K(t) = 2t$, both $\widehat{K}$ and $\check{K}$ are approximately

$$\frac{1}{\lambda^2 Q_+}T(\phi) - \frac{2}{\lambda Q_+}\sum_{j=1}^{N_+} h(X_j,\ L_j;\ \phi) + 4t.$$

Thus, for Q_+ large, the two estimators will be similar when $M_1, \ldots, M_p$ are iid Poisson processes, but they are not necessarily similar otherwise.

Even for simple regions in two or more dimensions, calculating $h(\cdot; \phi)$ requires numerical integrations. However, when the observation region is $D = \bigcup_{j=1}^{p}\{[0, Q_j], j\}$, it is possible to give an explicit expression for $h(x, l; \phi^R)$ for $(x, l) \in D$. For convenience, we will assume that the Q_j's have been arranged in increasing order. For $r < Q_p$, define $j(r) = \min_{1\le j\le p}\{j : Q_j \ge r\}$ and let $U(r) = \sum_{j=1}^{p}(Q_j - r)^+$. For $j = 1, \ldots, p$, let $U_j = U(Q_j)$ and set $Q_0 = 0$ so that $U_0 = Q_+$. Furthermore, define

$$\kappa(x, t) = \sum_{j=1}^{j(x\wedge t)-1} \frac{1}{p - j + 1}\log\left(\frac{U_{j-1}}{U_j}\right) + \frac{1}{p - j(x\wedge t) + 1}\log\left\{\frac{U_{j(x\wedge t)-1}}{U(x\wedge t)}\right\},$$

where a sum whose upper limit is less than its lower limit is defined to be 0 and $x \wedge t$ is the minimum of x and t. Then

$$Q_+^{-1} h(x, l; \phi^R) = \kappa(x, t) + \kappa(Q_l - x, t) \tag{4}$$

(see the Appendix). If the segment lengths are all equal, then $\kappa(s, t) = p^{-1} \log[Q/\{Q - (x \wedge t)\}]$.

It is also possible to evaluate $h(x, l; \phi^I)$ explicitly, but the resulting expression is rather cumbersome. If $t < \frac{1}{2} \min(Q_1, \ldots, Q_p)$, then the denominator in the definition of ϕ^I in (2) equals Q_+ whenever $|x - y| \le t$, which greatly simplifies matters. In this case, it is possible to show that

$$h(x, l; \phi^I) = t + (x \wedge t) + \{(Q_l - x) \wedge t\} - \frac{1}{2}\left(\frac{x}{2} \wedge t\right) - \frac{1}{2}\left(\frac{Q_l - x}{2} \wedge t\right).$$

A second special case yielding a simple result is when $Q_1 = \cdots = Q_p = Q$. When $t < \frac{1}{2}Q$, the preceding expression for h applies and, for $t \ge \frac{1}{2}Q$,

$$h(x, l; \phi^I) = \frac{3Q}{4} + \{x \wedge (Q - x)\} + Q \log\left[\frac{\frac{1}{2}Q}{\{x \wedge (Q - x)\} \vee (Q - t)}\right],$$

where $x \vee y$ is the maximum of x and y.

There is a considerable literature in astrophysical journals on estimating second-order characteristics of galaxy locations based on galaxy surveys in large, contiguous regions of the sky. Martínez (1997) and Stoyan and Stoyan (2000) provide two recent reviews of this work. Astrophysicists have generally focused on estimating the pair correlation function, which is, after a normalization, just the derivative of the K function. For example, for a stationary point process M on $\mathbb{R}$, assuming K is differentiable, the pair correlation function is $\frac{1}{2}K'$. Similar to $\widehat{K}$ here, Landy and Szalay (1993) make use of unbiased estimators of 0 to modify estimators of second-order characteristics. Moreover, similar to $\check{K}$, Hamilton (1993) describes estimators of the pair correlation function of the form $T(\phi)/\hat{\lambda}^2$ in which λ^2 is estimated by something other than the obvious estimator. We prefer to estimate K rather than the pair correlation function because it separates the problem of handling edge effects from that of density estimation and the consequent smoothing problem. If one wants to estimate the pair correlation function, we recommend first computing an appropriately edge-corrected estimate of K and then differentiating a smoothed version of this estimate.

4. Asymptotic theory when the truth is Poisson. There are a number of ways one might take limits to study the properties of the estimators proposed in the previous section. One possibility would be to fix p and let the Q_j's tend to ∞. In this approach, the fraction of the observation region within a fixed distance of an endpoint of a segment tends to 0, and, as in Ripley (1988) and Stein (1993), the variance of all reasonable estimators of $K(t)$ for fixed t have the same first-order asymptotic behavior under binomial sampling. However, for the absorber catalog, in which $p = 274$ and the number

of absorbers per line is 1.28, a more relevant choice is to uniformly bound the Q_j's and let $p \to \infty$. This limiting approach keeps the fraction of the observation region within a fixed distance of an endpoint of a segment bounded away from 0 with the result that the differences between various estimators under binomial sampling show up in the leading terms for the asymptotic variance. Hansen, Gill and Baddeley (1996) and Baddeley and Gill (1997) take a similar asymptotic approach for studying estimators of properties of spatial point processes based on observing the process in an increasing number of identical and distantly spaced windows.

We now consider adapting the asymptotic results in Ripley (1988) and Stein (1993) to the present setting. First, we give exact expressions for the variance under binomial sampling of both $\widetilde{K}(t)$ and $\widehat{K}(t)$. Following Ripley (1988), for a symmetric function ϕ on $D \times D$ satisfying (A), define

$$S(\phi) = \sum_{j=1}^{p} \int_0^{Q_j} \int_0^{Q_j} \phi((x, j), (y, j))\, dx\, dy,$$

$$S_1(\phi) = \sum_{j=1}^{p} \int_0^{Q_j} \left\{ \int_0^{Q_j} \phi((x, j), (y, j))\, dx \right\}^2 dy$$

and

$$S_2(\phi) = \sum_{j=1}^{p} \int_0^{Q_j} \int_0^{Q_j} \phi((x, j), (y, j))^2\, dx\, dy.$$

Under (A) [Ripley (1988)],

$$\begin{aligned} &\operatorname{var}_n\{T(\phi)\} \\ &\quad = \frac{2n(n-1)}{Q_+^2}\left\{ S_2(\phi) + \frac{2n-4}{Q_+} S_1(\phi) - \frac{2n-3}{Q_+^2} S(\phi)^2 \right\} \end{aligned} \tag{5}$$

and [Stein (1993)]

$$\operatorname{var}_n\{T^*(\phi)\} = \frac{2n(n-1)}{Q_+^2}\left\{ S_2(\phi) - \frac{2}{Q_+} S_1(\phi) + \frac{1}{Q_+^2} S(\phi)^2 \right\}. \tag{6}$$

We now want to study what happens as $p \to \infty$. Suppose $Q_1, Q_2, \ldots$ is a sequence of positive numbers and the subscript p is used to indicate the dependence of a term on the number of segments observed, so that $D_p = \bigcup_{j=1}^{p} \{[0, Q_j], j\}$, $Q_{+p} = \sum_{j=1}^{p} Q_j$ and N_{+p} is the total number of events on D_p. Suppose $\{\phi_p\}$ is a sequence of functions for which the domain of ϕ_p is $D_p \times D_p$ and ϕ_p is symmetric for all p. In addition to ϕ_p satisfying (A) for all p, we will assume the following regularity conditions:

(B) The ϕ_p's are uniformly bounded.
(C) For each p, ϕ_p satisfies the unbiasedness constraint in (1).
(D) The Q_j's are bounded away from 0 and ∞.

Under (A)–(D), we have $S(\phi_p) = 2tQ_{+p} = O(p)$, $S_1(\phi_p) = O(p)$ and $S_2(\phi_p) = O(p)$ but not $o(p)$. It follows that, as $p \to \infty$,

$$S_2(\phi_p) - \frac{2}{Q_{+p}} S_1(\phi_p) + \frac{1}{Q_{+p}^2} S(\phi_p)^2 = S_2(\phi_p)\{1 + O(p^{-1})\}. \tag{7}$$

Comparing (6) and (7) suggests that minimizing $S_2(\phi_p)$ subject to (A)–(D) is nearly the same as minimizing $\mathrm{var}_n\{T^*(\phi_p)\}$. Stein (1993) shows that, subject to (C), the rigid motion correction gives a minimizer of $S_2(\phi_p)$. The Appendix gives an explicit expression for $S_2(\phi^R)$ in terms of elementary functions.

We next obtain an analog to Proposition 2 in Stein (1993), which demonstrates the asymptotic optimality under the Poisson model for $\widehat{K}_R$ among a certain class of estimators as the dimensions of a single observation window increase. For a sequence of functions $\{\phi_p\}$ on $D_p \times D_p$ and a sequence of functions $\{g_p\}$ on $D_p \times \{0, 1, \ldots\}$, define the statistic $\Theta(\phi_p, g_p)$ by

$$\Theta(\phi_p, g_p) = \frac{Q_{+p}}{N_{+p}(N_{+p}-1)}\Bigg[T(\phi_p) - \sum_{j=1}^{N_{+p}}\Bigg\{g_p((X_j, L_j), N_{+p}) - \frac{1}{Q_{+p}}\sum_{l=1}^{p}\int_0^{Q_l} g_p((x, l), N_{+p})\,dx\Bigg\}\Bigg]$$

if $N_{+p} > 1$ and 0 otherwise. Write E_λ to indicate expectations assuming $M_1, M_2, \ldots$ are independent Poisson processes with constant intensity λ independent of p. All ensuing asymptotic results in the rest of this section involve expectations over the Poisson model and can be proven by first conditioning on N_{+p}, by using the fact that under this model the conditional distribution of the observed events on D_p follows binomial sampling and, finally, by averaging over the distribution of N_{+p}, which follows a Poisson distribution with mean λQ_{+p}.

PROPOSITION 1. *Suppose* $\{\phi_p\}$ *satisfies* (A)–(C), $E_\lambda\{\sum_{j=1}^{N_{+p}} |g_p((X_j, L_j), N_{+p})|\} < \infty$ *for all* p, *the* Q_j*'s satisfy* (D) *and* $p^{-1}\sum(Q_j - t)^+$ *is bounded away from* 0 *as* $p \to \infty$. *Then*

$$p^2\Big[E_\lambda\{\widehat{K}_R(t) - 2t\}^2 - E_\lambda\big\{\Theta(\phi_p, g_p) - 2t\big\}^2\Big]$$

is bounded from above as $p \to \infty$.

The assumption that $p^{-1}\sum(Q_j - t)^+$ is bounded away from 0 as $p \to \infty$ guarantees that $\{\phi_p^R\}$ satisfies (B). Since, under the conditions of Proposition 1, $E_\lambda\{\widehat{K}_R(t) - 2t\}^2 = O(p^{-1})$ as $p \to \infty$, this result says that, when the underlying processes are independent Poisson with equal intensity, $\widehat{K}_R$ asymptotically minimizes the mean squared error among all sequences of estimators of the form considered in the proposition.

Let us now make some comparisons of the asymptotic mean squared errors of some estimators of $K(t)$ under the Poisson model when all Q_j's equal Q and $s = t/Q$. From (6), we get

$$E_\lambda\{\widehat{K}(t) - 2t\}^2 \sim \frac{2}{\lambda^2 p^2 Q^2} S_2(\phi_p).$$

Thus, (17) in the Appendix implies

$$E_\lambda\{\widehat{K}_R(t) - 2t\}^2 \sim -\frac{4}{\lambda^2 p}\log(1-s) \tag{8}$$

and (20) in the Appendix implies

$$E_\lambda\{\widehat{K}_I(t) - 2t\}^2 \sim \frac{4}{\lambda^2 p} \times \begin{cases} s + \frac{3}{4}s^2, & \text{if } 0 < s \le \frac{1}{3}, \\ \frac{1}{12} + \frac{1}{2}s + \frac{3}{2}s^2, & \text{if } \frac{1}{3} \le s \le \frac{1}{2}, \\ \frac{17}{24} - \log 2 - \log(1-s), & \text{if } \frac{1}{2} \le s < 1. \end{cases} \tag{9}$$

From Proposition 1, the right-hand side of (9) must be at least as large as the right-hand side of (8) for all $s \in (0, 1)$. In fact, it is a straightforward exercise to show analytically that the right-hand side of (9) is strictly greater than the right-hand side of (8) for all $s \in (0, 1)$. Thus, as $p \to \infty$, the modified rigid motion estimator $\widehat{K}_R$ performs nonnegligibly better than either the ordinary or the modified isotropic estimator for any $t \in (0, Q)$ under the Poisson model, although the improvement over the modified isotropic estimator is minor. Figure 1 shows the ratio of the asymptotic variances for $\widehat{K}_I(t)$ and $\widehat{K}_R(t)$ under the Poisson model, which reaches a maximum of approximately 1.032 near $t = 0.247Q$. The asymptotic results in (8) and (9) are unchanged if $\check{K}_R$ and $\check{K}_I$ replace $\widehat{K}_R$ and $\widehat{K}_I$.

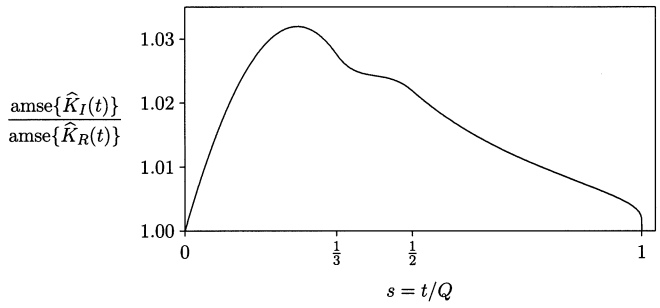

FIG. 1. *Ratio of asymptotic mean squared error (amse) of $\widehat{K}_I(t)$ to that of $\widehat{K}_R(t)$ for p segments of length Q as $p \to \infty$ under Poisson model.*

We next compare the modified and unmodified rigid motion estimators as $p \to \infty$ when all Q_j's equal Q. From (5),

$$E_\lambda\{\widetilde{K}(t) - 2t\}^2 \sim \frac{2}{\lambda^2 p^2 Q^2} S_2(\phi^p) + \frac{4}{\lambda p^2 Q^2} S_1(\phi^p) - \frac{16t^2}{\lambda p Q}.$$

Using (17) and (18) in the Appendix then yields

$$E_\lambda\{\widetilde{K}_R(t) - 2t\}^2 \sim \frac{4}{\lambda^2 p}\big[-\log(1-s) + 4\lambda Q\{\gamma(s) - s^2\}\big], \tag{10}$$

where

$$\gamma(s) = \frac{1}{4}\int_0^1 \left[\int_0^1 \frac{1\{|x-y| \le s\}}{1-|x-y|}\, dy\right]^2 dx. \tag{11}$$

Equation (19) in the Appendix gives a more explicit expression for γ. Note that

$$\gamma(s) - s^2 = \frac{1}{4}\int_0^1 \left[\int_0^1 \frac{1\{|x-y| \le s\}}{1-|x-y|}\, dy - 2s\right]^2 dx,$$

which is strictly positive for all $s \in (0, 1]$.

Comparing (8) and (10) shows that, in terms of mean squared error, the asymptotic relative advantage of either modified rigid motion estimator over the unmodified rigid motion estimator is proportional to λQ, the expected number of events per segment. Figure 2 plots $4\{\gamma(s) - s^2\}/\{-\log(1-s)\}$, which is less than 0.124 for all $s \in (0, 1)$ and is less than 0.061 for all $s < 0.9$. Thus, at least for equal Q_j's, we should not expect a large improvement under the Poisson model due to the modifications when there are only 1.28 events per segment as in the absorber catalog. Simulation results in Section 6 show that larger improvements can occur with unequal Q_j's.

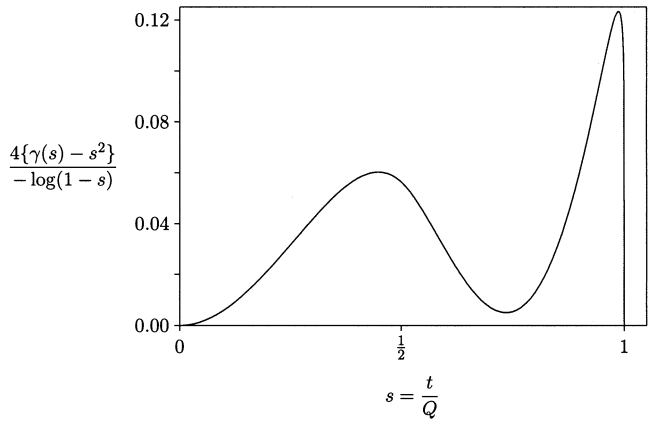

FIG. 2. *Plot of* $4\{\gamma(s)-s^2\}/\{-\log(1-s)\}$. *Multiplying this ratio by* λQ *gives the relative increase in asymptotic mean squared error as* $p \to \infty$ *due to using* $\widetilde{K}_R(t)$ *rather than* $\widehat{K}_R(t)$ *for* p *segments of length* Q *under the Poisson model with* $s = t/Q$.

5. Some asymptotic theory for non-Poisson processes. There is a decided lack of asymptotic theory that permits useful comparisons of estimators of K when the underlying process is not Poisson. Stein (1995) derives results showing the advantage of estimators like $\widehat{K}$ over those like $\widetilde{K}$, but the asymptotic approach taken there requires that the distance t at which one is estimating K be large compared to the distances at which the underlying process shows nontrivial dependence. When the observation window is made up of many segments, especially if the Q_j's are equal and the M_j's are independent, it appears feasible to develop some useful asymptotic results for non-Poisson processes. This section describes some general asymptotic results for the estimators $\widetilde{K}$, $\widehat{K}$ and $\check{K}$ described in Section 3. These results are used to demonstrate that if $M_1, M_2, \ldots$ are, conditional on $\Lambda_1, \Lambda_2, \ldots$, independent Poisson processes with M_j having intensity Λ_j, where the Λ_j's are iid positive random variables, then, as $p \to \infty$, $\check{K}_R(t)$ is superior to $\widehat{K}_R(t)$, which is in turn superior to $\widetilde{K}_R(t)$.

Suppose $M_1, M_2, \ldots$ are iid simple, stationary point processes on $\mathbb{R}$ with intensity λ and reduced second-moment function K. Assume $Q = Q_1 = Q_2 = \ldots$ and let $X_{1j}, \ldots, X_{N_j j}$ be the locations of the N_j events from M_j on $(0, Q)$. For a bounded, symmetric function ϕ on $(0, Q) \times (0, Q)$, define $\Phi_j = \sum_{k \neq l} \phi(X_{kj}, X_{lj})$. Analogous to (1), suppose $E\Phi_j = \lambda^2 Q K(t)$ for any reduced second-moment function K for the M_j's. Define

$$G_j = (2t)^{-1} \sum_{k=1}^{N_j} \int_0^Q \phi(X_{kj}, y)\, dy,$$

so that $EG_j = \lambda Q$. Using these definitions, the estimators described in Section 3 are given by

$$\widetilde{K}(t) = \frac{pQ \sum_{j=1}^p \Phi_j}{\sum_{j=1}^p N_j \left(\sum_{j=1}^p N_j - 1\right)},$$

$$\widehat{K}(t) = \widetilde{K}(t) - \frac{4t \sum_{j=1}^p G_j}{\sum_{j=1}^p N_j} + 4t$$

and

$$\check{K}(t) = \frac{pQ \sum_{j=1}^p \Phi_j}{\sum_{j=1}^p G_j \left(\sum_{j=1}^p G_j - 1\right)}.$$

Furthermore, since $\{N_j, \Phi_j, G_j\}_{j=1}^{\infty}$ is an iid trivariate sequence, we can readily derive the limiting distribution of these estimators. Specifically, if $E(N_1^4) <$

∞, then Φ_1 and G_1 have finite second moments, so, as $p \to \infty$,

$$p^{1/2} \begin{pmatrix} \frac{1}{p}\sum_{j=1}^{p} N_j - \lambda Q \\ \frac{1}{p}\sum_{j=1}^{p} \Phi_j - \lambda^2 Q K(t) \\ \frac{1}{p}\sum_{j=1}^{p} G_j - \lambda Q \end{pmatrix} \xrightarrow{\mathcal{L}} N(0, \Sigma),$$

where $\xrightarrow{\mathcal{L}}$ indicates convergence in distribution and $\sum$ is the 3×3 covariance matrix of (N_1, Φ_1, G_1). Using first-order Taylor series, we get $\lambda Q p^{1/2}\{\widetilde{K}(t) - K(t)\} \xrightarrow{\mathcal{L}} N(0, \widetilde{V})$, $\lambda Q p^{1/2}\{\widehat{K}(t) - K(t)\} \xrightarrow{\mathcal{L}} N(0, \widehat{V})$ and $\lambda Q p^{1/2}\{\check{K}(t) - K(t)\} \xrightarrow{\mathcal{L}} N(0, \check{V})$, where

$$\widetilde{V} = 4K(t)^2 \operatorname{var}(N_1) + \frac{1}{\lambda^2}\operatorname{var}(\Phi_1) - \frac{4K(t)}{\lambda}\operatorname{cov}(N_1, \Phi_1), \tag{12}$$

$$\begin{aligned} \widehat{V} = {} & 4\big\{K(t) - 2t\big\}^2 \operatorname{var}(N_1) + \frac{1}{\lambda^2}\operatorname{var}(\Phi_1) \\ & + 16t^2 \operatorname{var}(G_1) - \frac{4\{K(t) - 2t\}}{\lambda}\operatorname{cov}(N_1, \Phi_1) \\ & - \frac{8t}{\lambda}\operatorname{cov}(\Phi_1, G_1) + 16\big\{K(t) - 2t\big\}\operatorname{cov}(N_1, G_1) \end{aligned} \tag{13}$$

and

$$\check{V} = 4K(t)^2 \operatorname{var}(G_1) + \frac{1}{\lambda^2}\operatorname{var}(\Phi_1) - \frac{4K(t)}{\lambda}\operatorname{cov}(\Phi_1, G_1). \tag{14}$$

As expected, $\check{V} = \widehat{V}$ when $K(t) = 2t$.

To calculate the limiting behavior of these estimators for any given ϕ, Q and law of M_1, we only have to compute the covariance matrix $\sum$ and plug the results into (12)–(14). In some limited cases this computation can be done analytically or more often by numerical integration; otherwise, $\sum$ is easily approximated by simulation whenever M_1 can be readily simulated.

We now consider a simple setting in which $\sum$ can be explicitly derived. Suppose $M_1, M_2, \ldots$ are, conditional on $\Lambda_1, \Lambda_2, \ldots$, independent Poisson processes with M_j having intensity Λ_j, where the Λ_j's are iid positive random variables. Such a model could serve as an approximation for a Cox process [Daley and Vere-Jones (1988), Section 8.5] observed over widely spaced segments where the random intensity function $\Lambda(\cdot)$ of the process has little variation over distances of length Q but the segments are sufficiently spaced so that the behavior of $\Lambda(\cdot)$ in different segments is essentially independent.

Next, suppose $\phi(x, y) = Q1\{|x - y| \le t\}/(Q - |x - y|)$, so that we are using the rigid motion estimator. In this case, the elements of $\sum$ can be readily

calculated in terms of the moments of Λ_1. Writing m_j for $E(\Lambda_1^j)$, we have $\lambda = m_1$, $K(t) = 2tm_2/m_1^2$,

$$\begin{aligned}
\operatorname{var}(N_1) &= Qm_1 + Q^2(m_2 - m_1^2),\\
\operatorname{var}(\Phi_1) &= 16Q^3\gamma\left(\frac{t}{Q}\right)m_3 - 4Q^2\log\left(1 - \frac{t}{Q}\right)m_2\\
&\quad + 4t^2Q^2(m_4 - m_2^2),\\
\operatorname{var}(G_1) &= \frac{Q^3}{t^2}\gamma\left(\frac{t}{Q}\right)m_1 + Q^2(m_2 - m_1^2),\\
\operatorname{cov}(N_1, \Phi_1) &= 4tQm_2 + 2tQ^2(m_3 - m_1m_2),\\
\operatorname{cov}(N_1, G_1) &= Qm_1 + Q^2(m_2 - m_1^2)
\end{aligned}$$

and

$$\operatorname{cov}(\Phi_1, G_1) = \frac{4Q^3}{t}\gamma\left(\frac{t}{Q}\right)m_2 + 2tQ^2(m_3 - m_1m_2).$$

Each of these results can be obtained by conditioning on Λ_1. For example,

$$\begin{aligned}
\operatorname{var}(\Phi_1) &= E\{\operatorname{var}(\Phi_1|\Lambda_1)\} + \operatorname{var}\{E(\Phi_1|\Lambda_1)\}\\
&= E\left[4\Lambda_1^3\int_0^Q\left\{\int_0^Q \phi(x, y)\,dy\right\}^2 dx + 2\Lambda_1^2\int_0^Q\int_0^Q \phi(x, y)^2 dx\,dy\right]\\
&\quad + \operatorname{var}(2t\Lambda_1^2Q)\\
&= 16Q^3\gamma\left(\frac{t}{Q}\right)m_3 - 4Q^2\log\left(1 - \frac{t}{Q}\right)m_2 + 4t^2Q^2(m_4 - m_2^2),
\end{aligned}$$

where the second step follows from (10) in Ripley [(1988), page 30] and the last step uses (17) and (18) in the Appendix.

Plugging these results into (12)–(14) yields

$$\begin{aligned}
\widetilde{V}_R &= \frac{1}{m_1^2}\operatorname{var}(\Phi_1) - 16t^2Q\frac{m_2^2}{m_1^3} + 16t^2Q^2\frac{m_2(m_2^2 - m_1m_3)}{m_1^4},\\
\widehat{V}_R &= \frac{1}{m_1^2}\operatorname{var}(\Phi_1) - 16t^2Q\frac{(m_2 - m_1^2)^2}{m_1^3}\\
&\quad - 16Q^3\gamma\left(\frac{t}{Q}\right)\left(\frac{2m_2}{m_1} - m_1\right) + 16t^2Q^2\frac{m_2(m_2^2 - m_1m_3)}{m_1^4}
\end{aligned}$$

and

$$\check{V}_R = \frac{1}{m_1^2}\operatorname{var}(\Phi_1) - 16Q^3\gamma\left(\frac{t}{Q}\right)\frac{m_2^2}{m_1^3} + 16t^2Q^2\frac{m_2(m_2^2 - m_1m_3)}{m_1^4},$$

where the subscript R indicates that the asymptotic variance is for the appropriate version of the rigid motion estimator. Thus,

$$\widetilde{V}_R - \widehat{V}_R = 16Q^3\left(\frac{2m_2}{m_1} - m_1\right)\left\{\gamma\left(\frac{t}{Q}\right) - \frac{t^2}{Q^2}\right\}, \tag{15}$$

which is positive on $(0, 1)$ since $\gamma(s) - s^2 > 0$ for $s \in (0, 1)$ and $m_2 \geq m_1^2$. Furthermore,

$$\widehat{V}_R - \check{V}_R = 16Q^3\frac{(m_2 - m_1^2)^2}{m_1^3}\left\{\gamma\left(\frac{t}{Q}\right) - \frac{t^2}{Q^2}\right\}, \tag{16}$$

which is positive on $(0, 1)$ whenever $m_2 > m_1^2$. Thus, $\widehat{V}_R > \check{V}_R$ unless $\operatorname{var}\Lambda_1 = 0$, in which case $m_2 = m_1^2$ and $\widehat{V}_R = \check{V}_R$.

The arguments in this section largely carry over to estimators for the reduced second-moment function of iid point processes on $\mathbb{R}^d$ observed over $\bigcup_{j=1}^p \{A, j\}$ for some $A \subset \mathbb{R}^d$. In particular, (12)–(14) still hold if, at the appropriate places, $2t$ is replaced by $\mu_d t^d$, the volume of a ball of radius t in $\mathbb{R}^d$. Furthermore, the comparisons between $\widetilde{V}_R$, $\widehat{V}_R$ and $\check{V}_R$ in (15) and (16) still hold after replacing $\gamma(t/Q) - t^2/Q^2$ by $\int_A\{\int_A \phi(x, y)\,dy - \mu_d t^d\}^2\,dx$.

6. Simulation study. The asymptotic results in the preceding two sections provide only limited information about the relative advantages of the various estimators, especially for non-Poisson processes or unequal Q_j's. Because the estimators $\widetilde{K}_R$, $\widehat{K}_R$ and $\check{K}_R$ can all be explicitly calculated, it is fairly straightforward to study the behavior of these estimators via simulation. This section reports some results from a simulation study that considers equal and unequal Q_j's and three models for the law of the point processes. For the unequal segment length case, $p = 50$ and $Q_j = 0.1j$ for $j = 1, \ldots, p$ and for the equal segment length case, $p = 50$ and each $Q_j = 2.55$, so that $Q_+ = 127.5$ in both cases. The three processes reported on here are all stationary renewal processes; that is, the waiting times between consecutive events are iid random variables. In each case, the intensity of the process is 1, so that $EN_+ = 127.5$ in all simulations. Stationary renewal processes are straightforward to simulate on an interval $[0, Q]$. If F is the cdf (cumulative distribution function) for the waiting times and $\mu < \infty$ is the mean waiting time, then to obtain a stationary process on $[0, \infty)$, use $\mu^{-1}\int_0^x\{1 - F(y)\}\,dy$ for the cdf of the time of the first event after 0 [Daley and Vere-Jones (1988), page 107]. Simulate a random variable from this distribution; if it is greater than Q, then one is done and there are no events in $[0, Q]$ for this realization of the process. If not, simulate random waiting times with cdf F until one gets the first event after Q and use the preceding events as the realization of the process on $[0, Q]$. Here, we consider waiting time densities f that are exponential with mean 1 (in which case the M_j's are Poisson processes), $f(x) = 4xe^{-2x}$ for $x > 0$ (a gamma density with parameters 2 and $\frac{1}{2}$) and $f(x) = 24/(2 + x)^4$ for $x > 0$. Figure 3 plots $K(t) - 2t$ for renewal processes

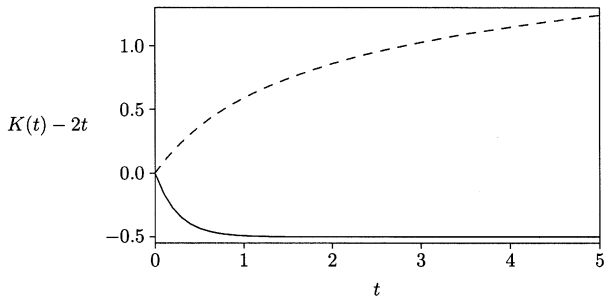

FIG. 3. *Plots of* $K(t) - 2t$ *for the renewal processes with waiting time densities* $4xe^{-2x}$ *for* $x > 0$ (*solid line*) *and* $24/(2+x)^4$ *for* $x > 0$ (*dashed line*).

with the last two waiting time densities, which shows that the first of these corresponds to a process more regular than the Poisson and the second is more clumped than the Poisson. For the gamma waiting times, it is possible to show that, for $x \neq 0$, $P\{M_1(dx) = 1 \mid M_1(\{0\}) = 1\} = 1 - e^{-4x}$ and hence that $K(t) = 2t - \frac{1}{2}(1 - e^{-4t})$. For the third waiting time density, we cannot give an analytic expression for $K(t)$, although Theorem 1 in Feller (1971), page 366, implies that $K(t) - 2t \to 2$ as $t \to \infty$. The values for $K(t)$ in Figure 3 for this process were obtained by simulation. Since the mean waiting times are all equal, the variances of the waiting times provide another measure of clumpiness with larger variances corresponding to a clumpier process. For the exponential waiting times, the variance is 1, for the gamma case, the variance is $\frac{1}{2}$ and for the last case, the variance is 3.

Figures 4–6 show the results of simulations for both sets of segment lengths and all three processes. For each scenario, the three estimators were calculated at a range of distances for 10,000 simulations. Generally speaking, $\widehat{K}_R$ and

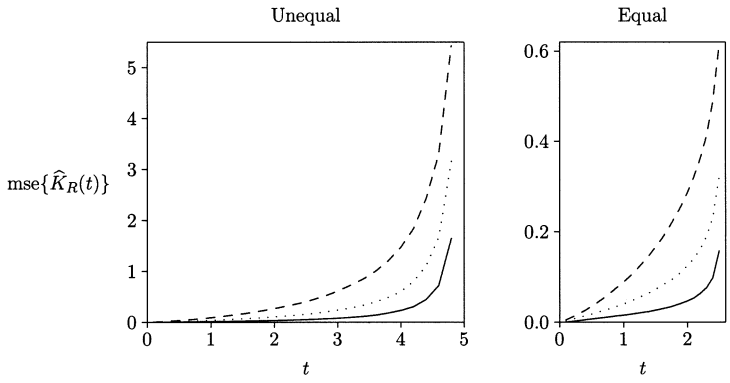

FIG. 4. *Mean squared errors for* $\widehat{K}_R$ *for three renewal processes and unequal or equal segment lengths. The waiting time densities are*: e^{-x} *for* $x > 0$ (*dotted line*), $4xe^{-2x}$ *for* $x > 0$ (*solid line*) *and* $24/(2+x)^4$ *for* $x > 0$ (*dashed line*).

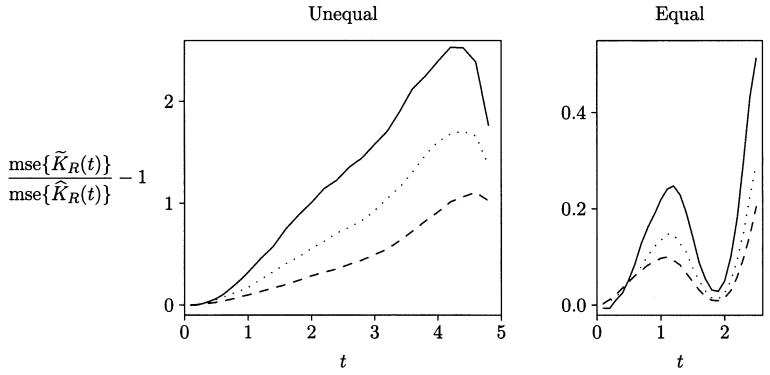

FIG. 5. *Relative differences in mean squared errors of* $\widetilde{K}_R$ *and* $\widehat{K}_R$: $mse(\widetilde{K}_R)/mse(\widehat{K}_R) - 1$. *Line types have same meaning as in Figure* 4.

$\breve{K}_R$ behave similarly and are superior to $\widetilde{K}_R$, especially at longer distances when the Q_j's are unequal. Figure 4 shows the mean squared errors for $\widehat{K}_R$. In all cases, the contributions of the squared biases to the mean squared errors are practically negligible and are always less than 0.5%. As expected, the mean squared errors grow with t, especially for the unequal segment length case as t gets near 5, the longest segment length available. Another expected result is that the mean squared errors increase with increasing clumpiness of the underlying process. Figure 5 compares $\widetilde{K}_R$ and $\widehat{K}_R$. We see that $\widehat{K}_R$ is generally superior, although $\widetilde{K}_R$ is sometimes slightly better for smaller t.

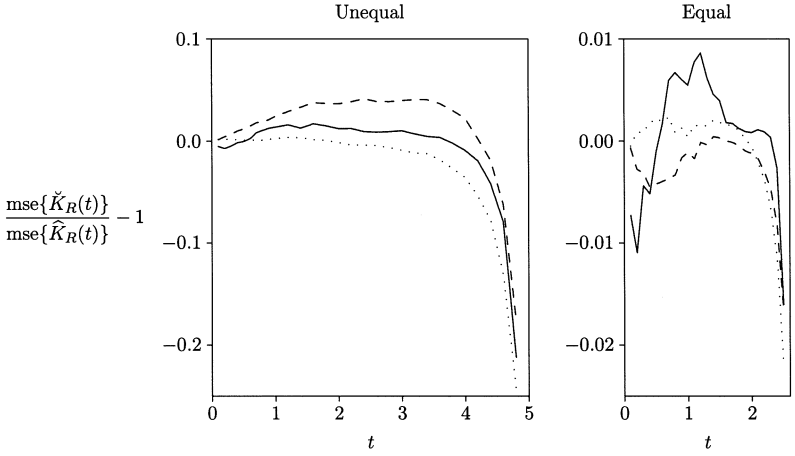

FIG. 6. *Relative differences in mean squared errors of* $\breve{K}_R$ *and* $\widehat{K}_R$: $mse(\breve{K}_R)/mse(\widehat{K}_R) - 1$. *Line types have same meaning as in Figure* 4.

The relative advantage of $\widehat{K}_R$ (and $\check{K}_R$) over $\widetilde{K}_R$ tends to be greater for more regular processes, which qualitatively agrees with the asymptotic results in Stein (1995). The advantage also tends to be greater for unequal segment lengths, demonstrating that theoretical results obtained for equal segment lengths may not accurately reflect the differences between estimators when segment lengths are unequal. Figure 6 compares $\widehat{K}_R$ and $\check{K}_R$. From the theoretical results in the previous section, we should expect these estimators to behave similarly when the waiting time density is exponential so that the underlying model is Poisson. The simulations show that the estimators also tend to behave very similarly for some non-Poisson models, especially when the segment lengths are equal. Neither estimator dominates the other, although $\check{K}$ tends to be silghtly superior for t nearly as large as the longest segment length.

For highly regular processes, $\widehat{K}_R$ can be substantially inferior to either $\widetilde{K}_R$ or $\check{K}_R$ for t sufficiently small. The problem is caused by the fact that, in such circumstances, having a pair of events within t of each other is rare, so that $\operatorname{var}\{T(\phi)\}$ is much smaller than under a Poisson model with the same intensity, whereas the variance of

$$T(\phi) - T^*(\phi) = \frac{2(N_+ - 1)}{Q_+} \sum_{j=1}^{N_+} \{h(X_j, L_j; \phi) - Eh(X, L; \phi)\}$$

is not much different for a highly regular process than for a Poisson process. As a consequence, subtracting off $T(\phi) - T^*(\phi)$ from $T(\phi)$ tends to inflate the variance of the estimator. As an example of a highly regular process, consider the stationary renewal process with waiting time density $(6^6/5!)x^5e^{-x/6}$ for $x > 0$, a gamma density with parameters 6 and $\frac{1}{6}$. This waiting time distribution has mean 1 and variance $\frac{1}{6}$ and corresponds to a highly regular point process. It is possible to show that

$$K(t) = 2t - \frac{5}{6} + \frac{1}{6}e^{-12t} + \frac{1}{3}\cos(3^{3/2}t)(e^{-9t} + e^{-3t}) + \frac{1}{3^{1/2}}\sin(3^{3/2}t)\left(\frac{1}{3}e^{-9t} + e^{-3t}\right)$$

for this process. Figure 7 shows that $\widehat{K}_R$ is notably inferior to either $\widetilde{K}_R$ or $\check{K}_R$ for t sufficiently small; for larger t, it is competitive with $\check{K}_R$ and clearly superior to $\widetilde{K}_R$. The overall winner is $\check{K}_R$, which performs well for all t.

We are unaware of any circumstances in which $\check{K}_R$ performs substantially worse than either $\widehat{K}_R$ or $\widetilde{K}_R$. Thus, we recommend routinely using $\check{K}_R$ to estimate K, although routine adoption for processes in more than one dimension will require the development of the necessary software.

7. Application to the absorber catalog. Figure 8 displays the estimators $\widetilde{K}_R$, $\widehat{K}_R$ and $\check{K}_R$ as applied to the absorber catalog described in Section 2. The three estimators are very similar and, as expected, show clear evidence of

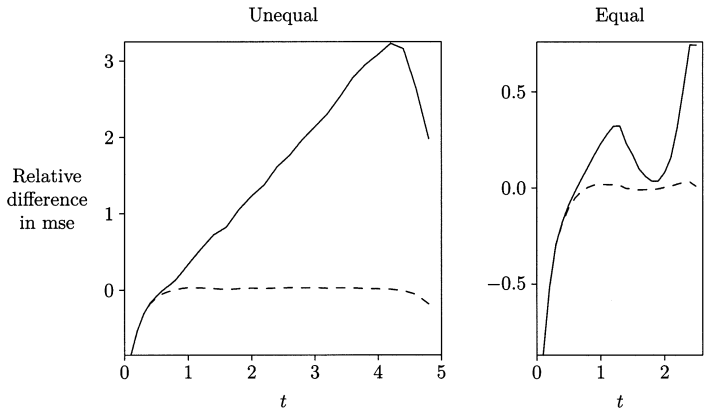

FIG. 7. *Relative differences in mean squared errors of* $\widetilde{K}_R$ *and* $\widehat{K}_R$ *(solid line) and of* $\check{K}_R$ *and* $\widehat{K}_R$ *(dashed line) for renewal process with waiting times having density* $(6^6/5!)x^5e^{-x/6}$ *for* $x > 0$ *and for equal and unequal segment lengths. Where the dashed line is not visible, it coincides with the solid line.*

clustering of absorbers. To obtain some idea about the uncertainty of these estimates, as in Quashnock and Stein (1999), approximate 95% pointwise confidence intervals were obtained by bootstrapping using the 274 segments as the sampling units. Specifically, using the notation in Section 5, simulated absorber catalogs were produced by sampling with replacement from $(Q_j; X_{1j}, \ldots, X_{N_j j})$ for $j = 1, \ldots, 274$, so that when one selects a segment,

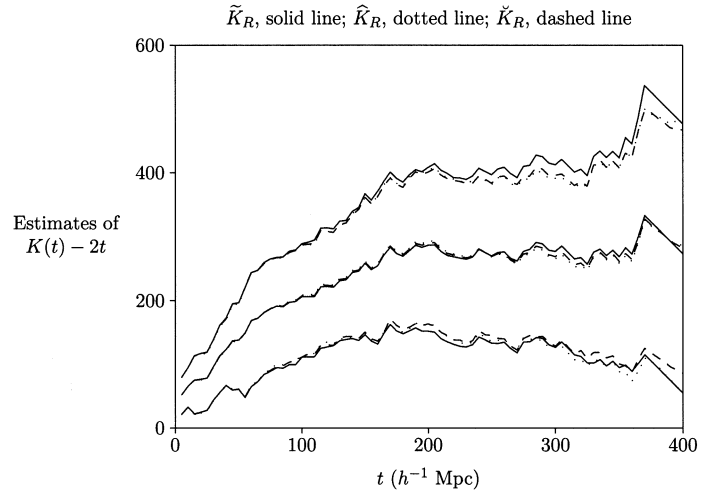

FIG. 8. *Three estimates of* $K(t) - 2t$ *for the absorber catalog and 95% confidence intervals based on bootstrapping lines of sight.*

one automatically selects the absorber locations that go with this segment. The confidence bands displayed in Figure 8 are then what Davison and Hinkley [(1997), page 29] call the basic bootstrap confidence limits and are based on 999 simulated catalogs. All three estimators yield similar confidence intervals, which is disappointing but perhaps not unexpected given the strong clustering that exists in the absorber catalog and the finding in the simulation study that the advantage of the modifications decreases as clustering increases. For these bootstrapping intervals to be appropriate, $(Q_j; X_{1j}, \ldots, X_{N_j j})$ for $j = 1, \ldots, 274$ should be iid random objects. Since the segments are of widely varying lengths, if the Q_j's are viewed as fixed, the identically distributed assumption is false. However, if we view the Q_j's as being a sequence of iid positive random variables that are independent of the locations of absorbers, then the identically distributed assumption may be reasonable. Whether or not the independence assumption is reasonable depends on the spatial extent of clustering among absorbers. If there is no spatial dependence in absorber locations beyond, say, 100 h^{-1} Mpc, then the independence assumption is not seriously in error, since few pairs of segments are within this distance of each other. If, however, nonnegligible clustering exists well beyond 100 h^{-1} Mpc, then the independence assumption is more problematic.

Analyses of galaxy surveys [Davis and Peebles (1983) and Loveday, Maddox, Efstathiou and Peterson (1995)] show that visible matter clusters on scales of up to 20 h^{-1} Mpc. Thus, it is more interesting to investigate how $K(t) - 2t$ changes at distances beyond 20 h^{-1} Mpc than to look at K itself. Figure 8 shows that $\widehat{K}_R(t) - 2t$ generally increases until about 200 h^{-1} Mpc and it is important to assess the uncertainty in this pattern. Applying the bootstrapping procedure to $\widetilde{K}_R(t) - \widetilde{K}_R(t_0)$ for $t_0 = 20, 50, 100$ and 150 h^{-1} Mpc, Quashnock and Stein (1999) concluded that there was strong evidence for clustering from 20 to 50 h^{-1} Mpc and from 50 to 100 h^{-1} Mpc, but at best marginal evidence for clustering beyond 100 h^{-1} Mpc. The results with the modified estimates (not shown) confirm the clear evidence for clustering from 20 to 50 h^{-1} Mpc and from 50 to 100 h^{-1} Mpc. Figure 9 shows the lower bounds for pointwise 95% confidence intervals for $K(t) - K(100) - 2(t - 100)$. The modified estimators yield slightly stronger evidence of clustering beyond 100 h^{-1} Mpc, which is mostly due to the fact that the modified estimates of $K(t) - K(100) - 2(t - 100)$ are slightly larger than the unmodified estimates for t around 200 and not because the modified intervals are narrower. If one used 99% pointwise confidence intervals in Figure 9, then, for all $t > 100$ and all three estimators, the lower confidence bounds are negative. Thus, the conclusion in Quashnock and Stein (1999) that there is perhaps marginal evidence for clustering beyond 100 h^{-1} Mpc is not altered by using the modified estimators.

As discussed in Section 2, the broad range of redshifts in the absorber catalog implies that we are looking at the universe at a broad range of times. The use of comoving units largely equalizes the intensity of absorbers across redshifts, but it does not equalize the clustering. Indeed, by dividing the absorber

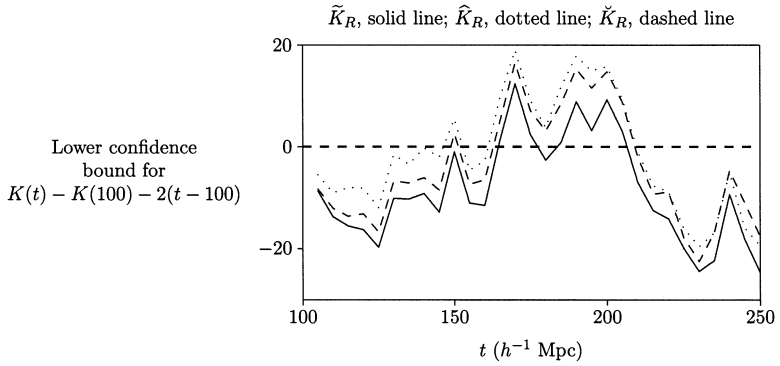

FIG. 9. *Lower limits of* 95% *confidence intervals for* $K(t) - K(100) - 2(t - 100)$ *for the absorber catalog*.

catalog into groups based on their redshift, Quashnock and Vanden Berk (1998) found evidence that as the redshift decreases, clustering on the scales of 1 to 16 h^{-1} Mpc strongly increases across the range of redshifts in the absorber catalog. Quashnock and Vanden Berk (1998) further note that this increase in clustering with decreasing redshift is consistent with what is known through theory and simulations about how gravity should affect the evolution of the clustering of absorbers over time. Using the various forms of the rigid motion estimator of K described here on groups of the absorber catalog with similar redshifts, we also find that on the scale of a few tens of h^{-1} Mpc, clustering increases substantially with decreasing redshift over the range of redshifts in the absorber catalog (results not shown). Thus, on these shorter scales, our estimates of K measure an average clustering over the range of redshifts in the absorber catalog.

In contrast, Quashnock, Vanden Berk and York (1996) found no evidence that clustering at scales of 100 h^{-1} Mpc changes over the redshift range in the absorber catalog. Similarly, when looking at, say, $\check{K}_R(t) - \check{K}_R(100)$ for $t > 100$ based on higher and lower redshift parts of the catalog, we find no systematic difference in the estimates as a function of redshift. For example, dividing the 274 segments in the catalog into two groups of size 137 based on redshift, $\check{K}_R(150) - \check{K}_R(100)$ equals 150.8 for the lower redshift group and 151.4 for the higher redshift group. Thus, we do not believe that the modest evidence we find for clustering at these larger scales is due to inhomogeneities across time in the distribution of absorbers.

8. Summary. For studying the behavior of edge-corrected estimators of the K function of a point process, taking the observation domain to be a sequence of segments has a number of desirable consequences. First, explicit expressions are available for a number of the more popular estimators, which is often not the case for regions in more than one dimension. The availability of such explicit expressions eases the study of the properties of these estima-

tors via both theory and simulation. In addition, studying settings in which the number of segments is large yields results that highlight the differences between the various methods of edge correction. In particular, simulation results show that allowing the segment lengths to vary generally increases the differences between estimators. The overall conclusion about the merits of the various estimators is that $\check{K}_R$, a modification of the rigid motion estimator based on an approach suggested by Picka (1996), is the estimator of choice.

The absorber catalog studied here shows that multiple windows of varying size can arise in practice. Although it is somewhat disappointing that the bootstrap confidence intervals for the ordinary rigid motion corrected estimator and its modifications are very similar, this result is not too surprising in light of the simulation results showing that the benefit of the modifications is smaller for clustered processes. The simulation results indicate that the modified estimators can have substantially smaller mean squared errors for Poisson or more regular processes, especially if the segment lengths vary substantially.

APPENDIX: PROOFS

We first derive (4) assuming, for convenience, that the Q_j's have been arranged in increasing order. We have

$$\begin{aligned}\frac{1}{Q_+}h(x,l;\phi^R) &= \frac{1}{Q_+}\sum_{j=1}^{p}\int_0^{Q_j}\phi^R((x,l),(y,j))\,dy\\ &= \int_0^{Q_l}\frac{1\{|x-y|\le t\}}{U(|x-y|)}\,dy\\ &= \int_0^{x}\frac{1\{|x-y|\le t\}}{U(|x-y|)}\,dy+\int_x^{Q_l}\frac{1\{|x-y|\le t\}}{U(|x-y|)}\,dy\\ &= \int_0^{x}\frac{1\{|x-y|\le t\}}{U(|x-y|)}\,dy+\int_0^{Q_l-x}\frac{1\{|Q_l-x-y|\le t\}}{U(|Q_l-x-y|)}\,dy.\end{aligned}$$

Thus, to verify (4), we need to show that

$$\kappa(x,t)=\int_0^x\frac{1\{|x-y|\le t\}}{U(|x-y|)}\,dy.$$

Now

$$\begin{aligned}\int_0^x\frac{1\{|x-y|\le t\}}{U(|x-y|)}\,dy &= \int_{(x-t)^+}^{x}\frac{dy}{U(x-y)}\\ &= \sum_{k=1}^{j(x\wedge t)-1}\int_{x-Q_k}^{x-Q_{k-1}}\frac{dy}{\sum_{j=k}^{p}(Q_j-x+y)}\\ &\quad+\int_{(x-t)^+}^{x-Q_{j(x\wedge t)-1}}\frac{dy}{\sum_{j=j(x\wedge t)}^{p}(Q_j-x+y)},\end{aligned}$$

which equals $\kappa(x,t)$ by calculus.

We next derive $S_2(\phi^R)$, again assuming that the Q_j's have been arranged in increasing order. By the symmetry of ϕ^R,

$$S_2(\phi^R) = 2Q_+^2 \sum_{j=1}^{p} \int_0^{Q_j} \int_0^x \frac{1\{x-y \le t\}}{U(x-y)^2}\, dy\, dx,$$

so taking $v = x - y$ and then switching the order of integration yields

$$\begin{aligned}
\frac{S_2(\phi^R)}{2Q_+^2} &= \sum_{j=1}^{p} \int_0^{Q_j} \int_0^{x\wedge t} \frac{1}{U(v)^2}\, dv\, dx \\
&= \sum_{j=1}^{p} \int_0^{Q_j \wedge t} \frac{Q_j - v}{U(v)^2}\, dv \\
&= \sum_{j=1}^{p} \sum_{l=1}^{j\wedge\{j(t)-1\}} \int_{Q_{l-1}}^{Q_l} \frac{Q_j - v}{\{\sum_{k=l}^{p}(Q_k - v)\}^2} dv \\
&\quad + \sum_{j=j(t)}^{p} \int_{Q_{j(t)-1}}^{t} \frac{Q_j - v}{\{\sum_{k=j(t)}^{p}(Q_k - v)\}^2}\, dv \\
&= \sum_{j=1}^{p} \sum_{l=1}^{j\wedge\{j(t)-1\}} \left\{ \frac{Q_j - Q_l}{(p-l+1)U_l} - \frac{Q_j - Q_{l-1}}{(p-l+1)U_{l-1}} \right. \\
&\qquad \left. - \frac{1}{(p-l+1)^2} \log\left(\frac{U_l}{U_{l-1}}\right) \right\} \\
&\quad + \sum_{j=j(t)}^{p} \left[\frac{Q_j - t}{\{p-j(t)+1\}U(t)} - \frac{Q_j - Q_{j(t)-1}}{\{p-j(t)+1\}U_{j(t)-1}} \right. \\
&\qquad \left. - \frac{1}{\{p-j(t)+1\}^2} \log\left\{\frac{U(t)}{U_{j(t)-1}}\right\} \right].
\end{aligned}$$

Using the definition of $U(t)$, the second sum simplifies to $\{p-j(t)+1\}^{-1} \times \log\{U_{j(t)-1}/U(t)\}$ and by switching the order of summation and using the definition of U_l, the first sum equals

$$\begin{aligned}
&\sum_{l=1}^{j(t)-1} \sum_{j=l}^{P} \left\{ \frac{Q_j - Q_l}{(p-l+1)U_l} - \frac{Q_j - Q_{l-1}}{(p-l+1)U_{l-1}} + \frac{1}{(p-l+1)^2} \log\left(\frac{U_{l-1}}{U_l}\right) \right\} \\
&\quad = \sum_{l=1}^{j(t)-1} \left\{ \frac{U_l}{(p-l+1)U_l} - \frac{U_{l-1}}{(p-l+1)U_{l-1}} + \frac{1}{p-l+1} \log\left(\frac{U_{l-1}}{U_l}\right) \right\} \\
&\quad = \sum_{l=1}^{j(t)-1} \frac{1}{p-l+1} \log\left(\frac{U_{l-1}}{U_l}\right).
\end{aligned}$$

Thus,

$$S_2(\phi^R) = 2Q_+^2 \sum_{l=1}^{j(t)-1} \frac{1}{p-l+1} \log\left(\frac{U_{l-1}}{U_l}\right) + \frac{2Q_+^2}{p-j(t)+1} \log\left\{\frac{U_{j(t)-1}}{U(t)}\right\}.$$

If $Q_1 = \cdots = Q_p = Q$, then, for $t < Q$, $j(t) = 1$, so

$$S_2(\phi^R) = -2pQ^2 \log\left(1 - \frac{t}{Q}\right). \tag{17}$$

Calculating $S_1(\phi^R)$ is more difficult and we only give the special case $Q_1 = \cdots = Q_p = Q$. Setting $s = t/Q$, we then have

$$S_1(\phi^R) = p\int_0^Q \left\{\int_0^Q \frac{Q1\{|x-y| \le t\}}{Q-|x-y|}\, dy\right\}^2 dx = 4pQ^3\gamma(s), \tag{18}$$

where γ is defined in (11). To evaluate γ, write

$$\begin{aligned}\gamma(s) &= \frac{1}{2}\int_0^1 \left[\int_0^x \frac{1\{x-y \le s\}}{1-x+y}\, dy\right]^2 dx \\ &\quad + \frac{1}{2}\int_0^1 \left[\int_0^x \frac{1\{x-y \le s\}}{1-x+y}\, dy\right]\left[\int_x^1 \frac{1\{z-x \le s\}}{1-z+x}\, dz\right] dx \\ &= \frac{1}{2}\int_0^1 \log^2\{1-(x\wedge s)\}\, dx \\ &\quad + \frac{1}{2}\int_0^1 \log\{1-(x\wedge s)\}\log\{(1-s)\vee x\}\, dx.\end{aligned}$$

Now

$$\int_0^1 \log^2\{1-(x\wedge s)\}\, dx = 2s + 2(1-s)\log(1-s)$$

and, for $s \le \frac{1}{2}$,

$$\int_0^1 \log\{1-(x\wedge s)\}\log\{(1-s)\vee x\}\, dx = -\log^2(1-s) - 2s\log(1-s)$$

whereas, for $s > \frac{1}{2}$,

$$\begin{aligned}&\int_0^1 \log\{1-(x\wedge s)\}\log\{(1-s)\vee x\} dx \\ &\quad = -2(1-s)\log(1-s) - 2s\log s\log(1-s) \\ &\qquad + \int_{1-s}^s \log(1-y)\log y\, dy.\end{aligned}$$

Hence,

$$\begin{aligned}\gamma(s) &= s + (1-2s)^+\log(1-s) - 1\{s \le \tfrac{1}{2}\}\tfrac{1}{2}\log^2(1-s) \\ &\quad - 1\{s > \tfrac{1}{2}\}s\log s\log(1-s) \\ &\quad + \int_0^{(s-1/2)^+} \log(\tfrac{1}{2}-y)\log(\tfrac{1}{2}+y)dy.\end{aligned} \tag{19}$$

Let us next consider computing $S_2(\phi^I)$. Defining $R(v) = Q_+ - \sum_{j=1}^p (2v - Q_j)^+$, then, for $y < x < Q_l$, we have

$$\phi^I((x,l),(y,m)) = \frac{1\{x-y \le t, l=m\}Q_+}{R(x-y)} \times \left[\frac{1}{1+1\{2x-y<Q_l\}} + \frac{1}{1+1\{2y-x>0\}}\right].$$

Thus, taking $v = x - y$,

$$\begin{aligned}\frac{S_2(\phi^I)}{2Q_+^2} &= \sum_{l=1}^p \int_0^{Q_l}\int_0^x \frac{1\{x-y\le t\}}{R(x-y)^2} \times \left[\frac{1}{1+1\{2x-y<Q_l\}} + \frac{1}{1+1\{2y-x>0\}}\right]^2 dy\,dx \\ &= \sum_{l=1}^p \int_0^{Q_l}\int_0^{x\wedge t} \frac{1}{R(v)^2}\left[\frac{1}{1+1\{x+v<Q_l\}} + \frac{1}{1+1\{x>2v\}}\right]^2 dv\,dx \\ &= \sum_{l=1}^p \int_0^{t\wedge Q_l} \frac{1}{R(v)^2}\int_v^{Q_l}\left[\frac{1}{1+1\{x+v<Q_l\}} + \frac{1}{1+1\{x>2v\}}\right]^2 dx\,dv.\end{aligned}$$

Now $[1+1\{x+v<Q_l\}]^{-1} + [1+1\{x>2v\}]^{-1}$ takes on values 2, $\frac{3}{2}$ and 1 depending on, respectively, whether none, one or both of $x+v<Q_l$ and $x>2v$ are true. Thus,

$$\begin{aligned}\frac{S_2(\phi^I)}{2Q_+^2} &= \sum_{l=1}^p \left\{\int_0^{t\wedge\frac{1}{3}Q_l} \frac{\frac{9}{4}2v + 1(Q_l-3v)}{R(v)^2}dv + \int_{t\wedge\frac{1}{3}Q_l}^{t\wedge\frac{1}{2}Q_l} \frac{\frac{9}{4}(2Q_l-4v)+4(3v-Q_l)}{R(v)^2}dv + \int_{t\wedge\frac{1}{2}Q_l}^{t\wedge Q_l}\frac{4(Q_l-v)}{R(v)^2}dv\right\} \\ &= \sum_{l=1}^p \left\{\int_0^{t\wedge\frac{1}{3}Q_l}\frac{Q_l+\frac{3}{2}v}{R(v)^2}dv + \int_{t\wedge\frac{1}{3}Q_l}^{t\wedge\frac{1}{2}Q_l}\frac{\frac{1}{2}(Q_l+3v)}{R(v)^2}dv + \int_{t\wedge\frac{1}{2}Q_l}^{t\wedge Q_l}\frac{4(Q_l-v)}{R(v)^2}dv\right\}.\end{aligned}$$

While it is possible to evaluate these integrals explicitly, the resulting expressions do not appear to simplify as in the case for the rigid motion estimator. When $Q_1 = \cdots = Q_p = Q$, we do obtain a fairly simple explicit result. By taking $u = v/Q$, we get

$$\begin{aligned}S_2(\phi^I) &= 2pQ^2\left[\int_0^{s\wedge\frac{1}{3}}\frac{1+\frac{3}{2}u}{\{1-(2u-1)^+\}^2}du + \int_{s\wedge\frac{1}{3}}^{s\wedge\frac{1}{2}}\frac{\frac{1}{2}+3u}{\{1-(2u-1)^+\}^2}du + \int_{s\wedge\frac{1}{2}}^{s}\frac{4-4u}{\{1-(2u-1)^+\}^2}du\right] \\ &= 2pQ^2\left\{\int_0^{s\wedge\frac{1}{3}}\left(1+\frac{3}{2}u\right)du + \int_{s\wedge\frac{1}{3}}^{s\wedge\frac{1}{2}}\left(\frac{1}{2}+3u\right)du + \int_{s\wedge\frac{1}{2}}^{s}\frac{1}{1-u}du\right\},\end{aligned}$$

so that, for $s = t/Q < 1$,

$$S_2(\phi^I) = 2pQ^2 \times \begin{cases} s + \frac{3}{4}s^2, & \text{if } 0 < s \le \frac{1}{3}, \\ \frac{1}{12} + \frac{1}{2}s + \frac{3}{2}s^2, & \text{if } \frac{1}{3} \le s \le \frac{1}{2}, \\ \frac{17}{24} - \log 2 - \log(1-s), & \text{if } \frac{1}{2} \le s < 1. \end{cases} \tag{20}$$

REFERENCES

BADDELEY, A. (1998). Spatial sampling and censoring. In *Stochastic Geometry: Likelihood and Computation* (O. E. Barndorff-Nielsen, W. S. Kendall and M. N. M. van Lieshout, eds.) Chap. 2. Chapman and Hall, London.

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

BADDELEY, A. J., MOYEED, R. A., HOWARD, C. V. and BOYDE, A. (1993). Analysis of a three-dimensional point pattern with replication. *Appl. Statist.* **42** 641–668.

COLEMAN, P. H. and PIETRONERO, L. (1992). The fractal structure of the universe. *Phys. Rep.* **213** 311–389.

CROTTS, A. P. S., MELOTT, A. L. and YORK, D. G. (1985). QSO metal-line absorbers: the key to large-scale structure? *Phys. Lett. B* **155** 251–254.

DALEY, D. J. and VERE-JONES, D. (1988). *An Introduction to the Theory of Point Processes.* Springer, New York.

DAVIS, M. and PEEBLES, P. J. E. (1983). A survey of galaxy redshifts. V. The two-point position and velocity correlations. *Astrophys. J.* **267** 465–482.

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application.* Cambridge Univ. Press.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* **2**. Wiley, New York.

HAMILTON, A. J. S. (1993). Toward better ways to measure the galaxy correlation function. *Astrophys. J.* **417** 19–35.

HANSEN, M. B., GILL, R. D. and BADDELEY, A. J. (1996). Kaplan–Meier type estimators for linear contact distributions. *Scand. J. Statist.* **23** 129–155.

JING, Y. P. and SUTO, Y. (1998). Confronting cold dark matter cosmologies with strong clustering of Lyman break galaxies at $z \sim 3$. *Astrophys. J.* **494** L5–L8.

LANDY, S. L. and SZALAY, A. S. (1993). Bias and variance of angular correlation functions. *Astrophys. J.* **412** 64–71.

LANZETTA, K. M., BOWEN, D. B., TYTLER, D. and WEBB, J. K. (1995). The gaseous extent of galaxies and the origin of Lyman-alpha absorption systems: a survey of galaxies in the fields of Hubble Space Telescope spectroscopic target QSOs. *Astrophys. J.* **442** 538–568.

LOVEDAY, J., MADDOX, S. J., EFSTATHIOU, G. and PETERSON, B. A. (1995). The Stromlo–APM redshift survey. II. Variation of galaxy clustering with morphology and luminosity. *Astrophys. J.* **442** 457–468.

MARGON, B. (1999). The Sloan Digital Sky Survey. *Philos. Trans. Roy. Soc. London Ser. A* **357** 93–103.

MARTÍNEZ, V. J. (1997). Recent advances in large-scale structure statistics. In *Statistical Challenges in Modern Astronomy II* (G. J. Babu and E. D. Feigelson, eds.) 153–166. Springer, New York.

OHSER, J. (1983). On estimators for the reduced second moment measure of point process. *Math. Oper. Statist. Ser. Statist.* **14** 63–71.

OHSER, J. and STOYAN, D. (1981). On the second-order and orientation analysis of planar stationary point processes. *Biometrical J.* **23** 523–533.

PEEBLES, P. J. E. (1993). *Principles of Physical Cosmology.* Princeton Univ. Press.

PICKA, J. (1996). Variance-reducing modifications for estimators of dependence in random sets. Ph.D. dissertation, Dept. Statistics, Univ. Chicago.

PONS-BORDERÍA, M.-J., MARTÍNEZ, V. J., STOYAN, D., STOYAN, H. and SAAR, E. (1999). Comparing estimators of the galaxy correlation function. *Astrophys. J.* **523** 480–491.

QUASHNOCK, J. M. and STEIN, M. L. (1999). A measure of clustering of QSO heavy-element absorption-line systems. *Astrophys. J.* **515** 506–511.

QUASHNOCK, J. M. and VANDEN BERK, D. E. (1998). The form and evolution of the clustering of QSO heavy-element absorption-line systems. *Astrophys. J.* **500** 28–36.

QUASHNOCK, J. M., VANDEN BERK, D. E. and YORK, D. G. (1996). High-redshift superclustering of quasi-stellar object absorption-line systems on 100 h^{-1} Mpc scales. *Astrophys. J.* **472** L69–L72.

RIPLEY, B. D. (1976). The second-order analysis of stationary point processes. *J. Appl. Probab.* **13** 255–266.

RIPLEY, B. D. (1988). *Statistical Inference for Spatial Processes.* Cambridge Univ. Press.

STEIDEL, C. C., ADELBERGER, K. L., DICKINSON, M., GIAVALISCO, M., PETTINI, M. and KELLOGG, M. (1988). A large structure of galaxies at redshift $z \sim 3$ and its cosmological implications. *Astrophys. J.* **492** 428–438.

STEIN, M. L. (1993). Asymptotically optimal estimation for the reduced second moment measure of point processes. *Biometrika* **80** 443–449.

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

STOYAN, D. and STOYAN, H. (2000). Improving ratio estimators of second order point process characteristics. *Scand. J. Statist.* **27** 641–656.

SYLOS LABINI, F., MONTUORI, M. and PIETRONERO, L. (1998). Scale-invariance of galaxy clustering. *Phys. Rep.* **293** 61–226.

VANDEN BERK, D. E., QUASHNOCK, J. M., YORK, D. G., YANNY, B. (1996). An excess of C IV absorbers in luminous quasars: evidence for gravitational lensing? *Astrophys. J.* **469** 78–83.

YORK, D. G., YANNY, B., CROTTS, A., CARILLI, C., GARRISON, E. and MATHESON, L. (1991). An inhomogeneous reference catalogue of identified intervening heavy element systems in spectra of QSOs. *Mon. Not. Roy. Astron. Soc* **250** 24–49.

ZHANG, Y., MEIKSIN, A., ANNINOS, P. and NORMAN, M. L. (1998). Physical properties of the Lyα forest in cold dark matter cosmology. *Astrophys. J.* **495** 63–79.

M. L. STEIN
DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637
E-MAIL: stein@galton.uchicago.edu

J. M. QUASHNOCK
DEPARTMENT OF ASTRONOMY AND ASTROPHYSICS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637
AND
DEPARTMENT OF PHYSICS
CARTHAGE COLLEGE
KENOSHA, WISCONSIN 53140

J. M. LOH
DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637