Statistica Sinica **27** (2017), 1401-1418
doi:https://doi.org/10.5705/ss.202015.0426

BANDWIDTH SELECTION FOR ESTIMATING THE TWO-POINT CORRELATION FUNCTION OF A SPATIAL POINT PATTERN USING AMSE

Woncheol Jang and Ji Meng Loh

Seoul National University and New Jersey Institute of Technology

Abstract: The two-point correlation function (2PCF) is a measure of the second-order properties of a spatial point pattern and is commonly used in astronomy. We introduce an asymptotic mean squared error (AMSE) approach to obtain closed-form expressions of adaptive optimal bandwidths for estimating the two-point correlation function of a homogeneous spatial point pattern. This approach provides a simple, quick method for optimal bandwidth selection for 2PCF estimation. Using optimal bandwidths allows more information to be extracted from the data. Numerical studies suggest that the mean squared error of estimates obtained using AMSE optimal bandwidths are close to those obtained with the empirical optimal bandwidths. We illustrate this with an application to a galaxy cluster catalog from the Sloan Digital Sky Survey.

Key words and phrases: Bandwidth selection, second-order properties, spatial point patterns, two-point correlation function.

1. Introduction

In the analysis of data consisting of the random locations of objects, it is often of interest to quantify the clustering of the points, which can be thought of as the degree of clumpiness or regularity observed in the point pattern. The clustering in the points often reflect the underlying process driving the placement of points. For example, in astronomy, the clustering of galaxies is believed to be due to the effects of gravitation on matter over an extremely long period of time. Hence, the degree and range of galaxy clustering estimated from a catalog or data set of galaxy locations can help astronomers refine their physical models for, say, the evolution of the universe (Dodelson (2003)). In ecological studies, the locations of a species of trees in a forest form a spatial point pattern. Clustering or regularity in the tree locations may be reflective of the seed dispersal mechanism employed by the specific tree species (Waagepetersen and Guan (2009)).

The clustering of a spatial point pattern is often studied via the properties of point pairs, in particular, the distribution of the inter-point distances in the

point pattern. Several related measures can be used to quantify this second-order property: the second-order product density $\rho^{(2)}$, Ripley's K function, the pair correlation function g, and the two-point correlation function (2PCF) ξ; see Section 1.1.

Each of these quantities is a function of the inter-point distance r, and estimates are obtained by counting the numbers of point pairs in the data that are separated by distance r. With the exception of the K function, which is an integral measure, this involves selecting a bin size or bandwidth. In this paper, we consider the problem of finding an optimal bandwidth for estimating the 2PCF of an observed spatial point pattern.

Stoyan and Stoyan (1994) introduced a simple rule-of-thumb for selecting the bandwidth, while Loh and Jang (2010) introduced a computationally intensive bootstrap procedure for bandwidth selection. Our goal is to introduce a method that performs better than the simple rule-of-thumb (as measured by mean square errors of estimates) but does not require a lot of computation.

The rest of the paper is organized as follows. Section 1.1 introduces second-order properties of a spatial point process, estimation of the 2PCF, and the bandwidth selection rule of Stoyan and Stoyan (1994). Section 1.2 highlights key ideas in bandwidth selection for density estimation, including the asymptotic mean integrated square error (AMISE) method. In Section 2 we describe the application of the AMISE approach to the 2PCF, removing the integration from the original AMISE approach to obtain an AMSE method that allows us to select adaptive bandwidths depending on r. Section 3 describes the results of a simulation study comparing the bandwidths obtained using AMSE with empirically optimal bandwidths. In Section 4 we apply the AMSE method to a Sloan Digital Sky Survey dataset of galaxy clusters. Section 5 concludes.

1.1. Second-order property in spatial point patterns

Let N represent a spatial point process, say in $\mathbb{R}^2$. For an observation window W with area $|W|$, a realization of N consists of a random number of points within W. The intensity function, $\lambda(s)$, for $s \in W$ is defined as

$$\lambda(s) = \lim_{|\mathrm{d}S| \to 0} \frac{N(\mathrm{d}S)}{|\mathrm{d}S|},$$

where $\mathrm{d}S$ represents a small region centered at s and $N(\mathrm{d}S)$ the number of points in $\mathrm{d}S$ (see e.g. Diggle (2003)). It can also be characterized in integral form as

$$\mathsf{E}\left[N(B)\right] = \int_B \lambda(s)\, ds,$$

for Borel sets B. In this work we assume that N is stationary, so that $\lambda(s) \equiv \lambda$, with an estimate of λ given by $N(W)/|W|$.

For a stationary spatial point process, the second-order product density, $\rho^{(2)}$, is defined as

$$\rho^{(2)}(\mathbf{r}) = \lim_{|\mathrm{d}S|\to 0} \frac{N(\mathrm{d}S)N(\mathrm{d}S+\mathbf{r})}{|\mathrm{d}S|^2},$$

where $\mathrm{d}S$ denotes a small region in W and $\mathrm{d}S+\mathbf{r}$ represents $\mathrm{d}S$ shifted by $\mathbf{r}$, or, equivalently,

$$\mathsf{E}\left[N(B)(N(B)-1)\right] = \int_B \int_B \rho^{(2)}(s_1-s_2)\, ds_1\, ds_2.$$

We assume isotropy, so $\rho^{(2)}(\mathbf{r}) = \rho^{(2)}(r)$, where $r = |\mathbf{r}|$. Intuitively, $\rho^{(2)}(r)|\mathrm{d}S_1||\mathrm{d}S_2|$ can be thought of as the probability of finding a pair of points, one in each of the two small regions $\mathrm{d}S_1$ and $\mathrm{d}S_2$ separated by distance r.

There are other second-order quantities of spatial point processes. The pair correlation function, g, is $\rho^{(2)}$ normalized by the intensity,

$$g(r) = \frac{\rho^{(2)}(r)}{\lambda^2},$$

the K function is an integrated version of g,

$$K(r) = \int_0^r 2\pi u\, g(u)\, \mathrm{d}u, \quad \text{for a process in } \mathbb{R}^2,$$

and the two-point correlation function (2PCF) is the "overdensity" relative to the unclustered Poisson process,

$$\xi(r) = g(r) - 1,$$

so that $\xi \equiv 0$ for the Poisson process, and $\xi(r) > 0$ or $\xi(r) < 0$ indicates respectively clustering or regularity at scale r, compared to the Poisson.

We focus on bandwidth selection for estimating the 2PCF. The 2PCF is routinely used by astronomers to quantify the clustering of astronomical objects such as galaxies and quasi-stellar objects (Martínez and Saar (2001)), and new catalogs of astronomical objects routinely contain estimates of the 2PCF. Cosmological models describing the evolution of the universe link the two-point correlation of astronomical objects, such as galaxies, to cosmological parameters that affect the early universe's evolution. The prediction of a bump in the 2PCF due to an event called "recombination", which occurred shortly after the universe began, was subsequently observed in the empirical 2PCF estimated from data (Ryden (2003)).

Estimating the two-point correlation function from data requires the count-

ing of pairs of points that are distance r apart. This requires specifying a bandwidth, and effects of the observation region boundary also has to be taken into account. There are several analytical methods for dealing with edge effects (Baddeley et al. (1993); Illian et al. (2008)). Astronomers deal with edge effects by using pair counts with a second, randomly generated dataset R, in what Kerscher, Szapudi and Szalay (2000) refers to as pairwise estimators.

Specifically, let D and R represent the data and a randomly generated set of points, with N_D and N_R numbers of points, respectively. Define, for $r_0 > h$,

$$DD = DD(r_0) = \sum_{x\in D}\sum_{\substack{y\in D\\ y\neq x}} 1\{r_0 - h \le |x-y| \le r_0 + h\},$$

$$DR = DR(r_0) = \sum_{x\in D}\sum_{y\in R} 1\{r_0 - h \le |x-y| \le r_0 + h\},$$

$$RR = RR(r_0) = \sum_{x\in R}\sum_{\substack{y\in R\\ y\neq x}} 1\{r_0 - h \le |x-y| \le r_0 + h\}.$$

Hence DD counts the number of point pairs that are within $r_0 - h$ and $r_0 + h$ apart, while RR and DR have the same interpretation, with both points from R and one each from D and R, respectively. In the definitions of DD, DR, and RR, a boxcar kernel is used. We consider the choice of optimal bandwidths assuming the boxcar kernel. It is known (Silverman (1986)) that the choice of bandwidth is more important than the choice of kernel.

If we set $DD^* = 2DD/(N_D(N_D-1)), RR^* = 2RR/(N_R(N_R-1))$ and $DR^* = DR/(N_D N_R)$, then a simple estimator of the 2PCF at r_0 is

$$\widehat{\xi}(r_0) = \frac{DD^*}{RR^*} - 1. \tag{1.1}$$

The quantity DD is related to the second-order product density (see Section 2) and DD^* to the pair correlation function. The role of RR^* in (1.1) is to empirically account for the edge effect. The analytical corrections (see e.g. Baddeley et al. (1993)) account for edge effects of each point pair individually, while RR^* provides an average correction for the edge effect. This might be less effective than using individual analytical corrections, though Kerscher, Szapudi and Szalay (2000) did not find it so. It has the advantage of being straightforward to apply in the context of astronomy where the observation windows can be complex.

The estimator (1.1) can be improved to provide lower mean square errors. Kerscher, Szapudi and Szalay (2000) compares the pairwise estimators used by astronomers, as well as kernel-type estimators with analytical edge corrections.

Labatie, Starck and Lachièze-Rey (2012) studied the bias and uncertainty of these estimators. The Landy-Szalay estimator (Landy and Szalay (1993)) and the Hamilton estimator (Hamilton, 1993) are the best pairwise estimators of the 2PCF, with the Landy-Szalay estimator being slightly more popular:

$$\widehat{\xi}_{LS}(r_0) = \frac{(DD^* - 2DR^* + RR^*)}{RR^*}. \tag{1.2}$$

Its standard error is reduced compared with the other estimators because the $-2DR^* + RR^*$ term is negatively correlated with DD^*. Stein (1993) used a similar idea to reduce the variance of estimators of the K function. In principle, analytically edge-corrected estimators of the 2PCF can be similarly improved, though we are not aware of any work on this.

We propose a method to obtain optimal bandwidths for estimating the 2PCF using (1.2), by adapting an asymptotic mean integrated squared error (AMISE) bandwidth selection method used in density estimation.

Third and higher-order measures of clustering are now more commonly studied in astronomy and in other areas (e.g. Szapudi et al. (2001); Kim et al. (2011)), but second-order measures, and especially the 2PCF in astronomy, are still the norm. Our method for finding the optimal bandwidth can be extended to apply to third and higher-order measures if necessary.

1.2. Bandwidth selection for density estimation

We give an overview of bandwidth selection in density estimation here. Given n independent identically distributed random variables, $X_1, \ldots, X_n$ with an unknown density function f, the kernel density estimator of $f(x)$ is

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where h is the bandwidth and K is a kernel function. We assume that K satisfies

$$\int K(x)\,\mathrm{d}x = 1, \int xK(x) = 0, \text{ and } \int x^2 K(x) < \infty.$$

For choosing the bandwidth in density estimation, a common criterion is the mean integrated square error (MISE) of an estimator $\hat{f}$,

$$\mathrm{MISE}(\hat{f}) = \mathsf{E} \int \left(\hat{f}(x) - f(x)\right)^2 \mathrm{d}x.$$

Then the asymptotic mean integrated square error (AMISE) is

$$\mathrm{AMISE}(\hat{f}) = \frac{1}{nh} \int K(t)^2\,\mathrm{d}t + \frac{1}{4} h^4 \mu_2(K)^2 \int f''(x)^2\,\mathrm{d}x,$$

where $\mu_2(K) = \int t^2 K^2(t)\,\mathrm{d}t$. See Silverman (1986) for details.

With simple algebra, the bandwidth that minimizes the AMISE is

$$h_{opt} = \left(\frac{R(K)}{\mu_2(K)^2 R(f'')n} \right)^{1/5},$$

where $R(g) = \int g^2(x)\,\mathrm{d}x$. As $R(f'')$ is unknown, this cannot be calculated in practice. Popular bandwidth selectors include least squares cross-validation and likelihood cross-validation (Silverman (1986)). One can choose a *plug-in* bandwidth selector based on assuming f is Normal that provides a closed form of the optimal bandwidth and plug-in methods have an advantage in computation over cross-validation-based bandwidth selectors.

2. AMSE for the 2PCF

To estimate the 2PCF, we need a bandwidth h for computing DD, DR, and RR. Stoyan and Stoyan (1994) proposed using the bandwidth $h = c\lambda^{-1/d}$ for the Epanechnikov kernel, where d is the dimension of the observational area and c is a constant. For planar point patterns with about 50-300 points, they suggested $c \in (0.1, 0.2)$ with $c = .15$ being a common choice (Guan (2007)). Stoyan (2006) recommended the boxcar kernel, but did not provide guidelines on how to select the optimal bandwidth. For three dimensions, Pons-Borderia et al. (1999) recommended $c = 0.05$ and 0.1 for clustered and Poisson point processes respectively. We develop a simple bandwidth selection method that is less sensitive to assumptions on the underlying process and which allows for adaptive bandwidth selection depending on r.

Let $K_h(t) = 1/2h1\{|t| \le h\}$, so that $RR = RR(r_0) = 2h \sum_{x\in R} \sum_{y\in R: y\neq x} K_h$ $(|x - y| - r_0)$, similarly for DR and DD. Here, $\int_{-h}^{h} K_h(t)\,\mathrm{d}t = 1$. Let λ_R and λ_D be the intensities of the point processes generating R and D. For $r_0 > h$, we have

$$\begin{aligned}
\mathsf{E}\,(RR) &= 2h \int_W \int_W K_h(|x - y| - r_0)\lambda_R^2\,\mathrm{d}y\,\mathrm{d}x \\
&= 2h\lambda_R^2 \int_W \int_0^\infty \int_\Theta K_h(r - r_0) 1_W\{x + (r, \theta)\}\, r^2 \sin\theta\,\mathrm{d}\theta\,\mathrm{d}r\,\mathrm{d}x \\
&= 2h\lambda_R^2 \int_W \int_0^\infty K_h(r - r_0) \left[\int_\Theta 1_W\{x + (r, \theta)\}\, r^2 \sin\theta\,\mathrm{d}\theta \right] \mathrm{d}r\,\mathrm{d}x \\
&= 2h\lambda_R^2 \int_W \int_0^\infty K_h(r - r_0) C_x(r)\,\mathrm{d}r\,\mathrm{d}x,
\end{aligned}$$

wherein we have expressed y in spherical coordinates: $y = x + (r, \theta)$, with $r \in (0, \infty), \theta \in \Theta$, and included the indicator function $1_W\{x + (r, \theta)\}$ to ensure that

$y \in W$. The quantity $C_x(r)$ represents the edge effect due to the boundary of W. It is equal to $4\pi r^2$ in three dimensions if x is more than r away from the boundary of W. For given r_0, take

$$|W_0| = \int_W C_x(r_0)\,\mathrm{d}x, \quad |W_0'| = \int_W C_x'(r_0)\,\mathrm{d}x, \quad |W_0''| = \int_W C_x''(r_0)\,\mathrm{d}x.$$

We need some regularity conditions similar to the conditions in kernel density estimation (Silverman (1986)):

(C1) C_x'' and g'' are absolutely continuous and $C_x''', g''' \in L_2$.

(C2) $|W_0|$, $|W_0'|$ and $|W_0''|$ are of the same order.

Note that $C_x''' = 0$ in three dimensions. We also assume $h = h_n \to 0$, $|W_0| = |W_{n,0}| \to \infty$, and $|W_{n,0}|h_n^2 \to \infty$ as the number of realizations goes to ∞.

For small h, $r \approx r_0$ and, using a Taylor's expansion of $C_x(r)$ about $C_x(r_0)$, we obtain

$$\begin{aligned}
\mathsf{E}\,(RR) &= 2h\lambda_R^2 \int_W \int_{-1}^{1} K(t) C_x(r_0 + th)\,\mathrm{d}t\,\mathrm{d}x \\
&= 2h\lambda_R^2 \int_W \int_{-1}^{1} K(t) \left\{ C_x(r_0) + C_x'(r_0)th + C_x''(r_0)\frac{t^2h^2}{2} \right\} \mathrm{d}t\,\mathrm{d}x \\
&= 2h\lambda_R^2 \left(|W_0| + |W_0''|\frac{h^2}{3} \right),
\end{aligned}$$

where we used $\int_0^\infty K_h(r - r_0)\,\mathrm{d}r = \int_{r_0-h}^{r_0+h} 1\{|r - r_0| \le h\}/2h\,\mathrm{d}r = 1$ if $r_0 > h$, along with $\int t^2 K(t)\,\mathrm{d}t = 2/3$. Similarly, since the sets D and R are independent, we have

$$\begin{aligned}
\mathsf{E}\,(DR) &= 2h \int_W \int_W K_h(|x - y| - r_0)\lambda_R\lambda_D\,\mathrm{d}y\,\mathrm{d}x \\
&= 2h\lambda_D\lambda_R \left(|W_0| + |W_0''|\frac{h^2}{3} \right),
\end{aligned}$$

where the steps from the first and second equality here follow those for $\mathsf{E}\,(RR)$. Furthermore, we have

$$\mathsf{E}\,(DD) = 2h \int_W \int_W K_h(|x - y| - r_0)\lambda_D^2\, g(|x - y|)\,\mathrm{d}y\,\mathrm{d}x,$$

for the data set D, since point pairs are correlated, with the correlation represented by the pair correlation function g. Using a Taylor expansion for both $C_x(r)$ and g, we obtain

$$\mathsf{E}\,(DD) = 2h\lambda_D^2 \int_W \int_{-1}^{1} K(t) C_x(r_0 + th) g(r_0 + th)\,\mathrm{d}t\,\mathrm{d}x$$

$$
\begin{aligned}
&= 2h\lambda_D^2 \int_W \int_{-1}^{1} K(t) \left\{ C_x(r_0)g(r_0) + th(C'_x(r_0)g(r_0) + C_x(r_0)g'(r_0)) \right\} \mathrm{d}t \, \mathrm{d}x \\
&\quad + 2h\lambda_D^2 \int_W \int_{-1}^{1} K(t) \Bigg\{ \frac{t^2h^2}{2} (C''_x(r_0)g(r_0) + 2C'_x(r_0)g'(r_0) \\
&\quad + C_x(r_0)g''(r_0)) + O(h^3) \Bigg\} \mathrm{d}t \, \mathrm{d}x \\
&= 2h\lambda_D^2 g(r_0)|W_0| + \frac{2h^3\lambda_D^2}{3} \left\{ |W''_0| g(r_0) + 2|W'_0| g'(r_0) + |W_0| g''(r_0) \right\} \\
&\quad + O(|W_0| h^5).
\end{aligned}
$$

It is common in astronomy to take $N_R = N_D$. If we set $\lambda_R = \lambda_D$, then we have $E(DR) = E(RR)$, and the difference between $E(DD)$ and $E(RR)$ is due to the presence of $g(r)$. From

$$
\begin{aligned}
\frac{1}{\mathsf{E}\,(RR)} &= \left(2h\lambda^2|W_0|\right)^{-1} \left(1 + \frac{|W''_0|}{|W_0|} \frac{h^2}{3}\right)^{-1} \\
&= \left(2h\lambda^2|W_0|\right)^{-1} \left(1 - \frac{|W''_0|}{|W_0|} \frac{h^2}{3} + O(h^4)\right), \\
\frac{\mathsf{E}\,(DD - 2DR + RR)}{\mathsf{E}\,(RR)} &= \frac{E(DD) - E(RR) + O(h^4)}{E(RR)} = \frac{E(DD) + O(h^4)}{E(RR)} - 1.
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
&\frac{E(DD)}{E(RR)} \\
&= \left(g(r_0) + \frac{1}{3}h^2 \left(\frac{|W''_0|}{|W_0|} g(r_0) + 2\frac{|W'_0|}{|W_0|} g'(r_0) + g''(r_0) \right) \right) \left(1 - \frac{|W''_0|}{|W_0|} \frac{h^2}{3} + O(h^4) \right) \\
&= g(r_0) + \frac{1}{3}h^2 \left(2\frac{|W'_0|}{|W_0|} g'(r_0) + g''(r_0) \right) + O(h^4),
\end{aligned}
$$

and so

$$
\frac{\mathsf{E}\,(DD - 2DR + RR)}{\mathsf{E}\,(RR)} = \xi(r_0) + \frac{1}{3}h^2 \left(2\frac{|W'_0|}{|W_0|} g'(r_0) + g''(r_0) \right) + O(h^4).
$$

We can then express the bias of $\widehat{\xi}_{LS}(r_0)$ as

$$
\begin{aligned}
\mathrm{Bias}(\hat{\xi}_{LS}(r_0)) &= \mathsf{E}\,(\hat{\xi}_{LS}(r_0)) - \xi(r_0) \\
&= \mathsf{E}\,(\hat{\xi}_{LS}(r_0)) - \frac{\mathsf{E}\,(DD - 2DR + RR)}{\mathsf{E}\,(RR)} + \frac{\mathsf{E}\,(DD - 2DR + RR)}{\mathsf{E}\,(RR)} - \xi(r_0) \\
&= \frac{1}{3}h^2 \left(2\frac{|W'_0|}{|W_0|} g'(r_0) + g''(r_0) \right) + O(h^4) + O\left(\frac{1}{h^2|W_0|} \right),
\end{aligned}
$$

where we also replaced the expected value of the ratio of $DD - 2DR + RR$ to

RR with the ratio of their expected values,

$$\mathsf{E}\left(\hat{\xi}_{LS}(r_0)\right) - \frac{\mathsf{E}\left(DD - 2DR + RR\right)}{\mathsf{E}\left(RR\right)} = O\left(\frac{1}{h^2|W_0|}\right). \tag{2.1}$$

See the online Supplementary Materials for the derivation of (2.1).

Under the assumption that $\xi = o(1)$, Landy and Szalay (1993) derived an approximation for the variance of $\hat{\xi}_{LS}$,

$$\begin{aligned}
\mathsf{Var}\left[\hat{\xi}_{LS}(r_0)\right] &= \frac{\mathsf{E}\left[1 + \hat{\xi}_{LS}(r_0)\right]^2}{\mathsf{E}\left(RR\right)} + o\left(h^{-1}|W_0|^{-1}\right) \\
&= \frac{\mathsf{E}\left[1 + \hat{\xi}_{LS}(r_0)\right]^2}{2h\lambda^2|W_0|}\left(1 + O(h^2)\right) + o\left(h^{-1}|W_0|^{-1}\right) \\
&= \frac{g(r_0)^2}{(2h\lambda^2|W_0|)} + o\left(h^{-1}|W_0|^{-1}\right).
\end{aligned}$$

Theorem 1. *If C_x'' and ξ'' are absolutely continuously and $C_x''', \xi''' \in L_2$, and if $\xi = o(1)$, then*

$$MSE(r_0) = \frac{h^4}{9}A_0^2 + \left(\frac{1}{h}\frac{g(r_0)^2}{2\lambda^2|W_0|}\right) + o\left(h^4 + \frac{1}{|W_0|h}\right), \tag{2.2}$$

where $A_0 = 2|W_0'|g'(r_0)/|W_0| + g''(r_0)$. The optimal bandwidth is

$$h_{opt}(r_0) = \left[\frac{9g(r_0)^2}{8\lambda^2|W_0|A_0^2}\right]^{1/5}. \tag{2.3}$$

The sample size appears in (2.3), through $\lambda|W_0|$, which is related to the size of the observation window W and hence the number of points of the point process observed in W.

The expression in (2.3) depends on the unknown function $g = 1 + \xi$ as well as its second deriviative. A simple procedure to get an optimal bandwidth is then to refer to a standard model with a specific form for g (or ξ). For example, we can choose as g that of the modified Thomas process (Thomas (1949)), a Neyman-Scott process. With homogeneous parent intensity κ, Poisson mean number of offspring μ, and standard deviation σ for the Gaussian density of offspring around parents, the pair correlation function of the 2D modified Thomas process is

$$g(r) = 1 + \frac{\mu}{4\pi\lambda\sigma^2}e^{-r^2/4\sigma^2},$$

with $\lambda = \kappa\mu$. Using this expression for g, we get

$$h_{opt}(r_0) = \left(\frac{18\sigma^8}{\lambda^2|W_0|}\left[\frac{4\pi\lambda\sigma^2 + \mu e^{-r^2/4\sigma^2}}{\mu(r^2 - 6\sigma^2)e^{-r^2/4\sigma^2}}\right]^2\right)^{1/5}. \tag{2.4}$$

We use this optimal bandwidth in our simulation study in Section 3.

Table 1. Parameters used for the Thomas modified process in the simulation study.

Thomas model	1	2	3	4	5	6	7	8	9	10	11	12
κ	50	50	50	50	50	50	100	100	100	100	100	100
σ	0.05	0.05	0.1	0.1	0.2	0.2	0.05	0.05	0.1	0.1	0.2	0.2
μ	2	8	2	8	2	8	1	4	1	4	1	4

In astronomy, a commonly used functional form for the 2PCF is the power-law, $\xi(r) = (r/s_0)^{-\gamma}$ that is known to fit a wide range of empirical data well. If we use this in (2.3), we get

$$h_{opt}(r_0) = \left(\frac{9}{8\lambda^2 |W_0| \gamma^2 (\gamma - 1)^2} \left[1 + \left(\frac{r_0}{s_0} \right)^{\gamma} \right]^2 r_0^4 \right)^{1/5}. \tag{2.5}$$

The modified Thomas process is well-known in the spatial statistics community and we recommend using the pair correlation function based on this model, for clustered point patterns. It may not be ideal for regular point processes, and there we recommend using (2.3) based on an appropriate regular point model.

In applications, values of the parameters in g have to be estimated from the point pattern data. We can do this using a minimum contrast method (Diggle (2003)) with an initial estimate of either the K function (which does not require a bandwidth), or of ξ using Stoyan's bandwidth. Instead of $\lambda^2|W_0|$, we can use the average of $RR/2h$ evaluated over a range of values of h, where RR is as defined before, using $\lambda_R = \lambda_D$.

3. Simulation Study

We performed a simulation study to show the performance of the AMSE method for bandwidth selection to estimate the 2PCF. We compared it with (a) the simple rule of thumb $c\lambda^{-1/d}$ (Stoyan and Stoyan, 1994) for the bandwidth, using $c = .15$ and $d = 2$, and (b) the optimal bandwidths obtained empirically (described below).

We considered the Thomas modified process with 12 different sets of parameters (see Table 1). For each parameter set, we simulated 500 realizations and estimated the 2PCF, $\hat{\xi}_b^i(r), i = 1, \ldots, 500$, over a set of values r each using a range of bandwidths. For each value of r, we found the bandwidth b_r that minimized the mean squared error $\text{MSE}(r) = \sum_i [\hat{\xi}_b^i(r) - \xi(r)]^2$. The bandwidth b_r was then the empirically obtained optimal bandwidth for estimating $\xi(r)$. For each parameter set, we simulated a new set of 500 realizations and for each realiza-

Table 2. Parameters used for the Matérn cluster process in the simulation study.

Matérn	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
κ	50	50	50	50	50	50	50	50	100	100	100	100	100	100	100	100
R	0.05	0.05	0.1	0.1	0.2	0.2	0.4	0.4	0.05	0.05	0.1	0.1	0.2	0.2	0.4	0.4
μ	2	8	2	8	2	8	2	8	1	4	1	4	1	4	1	4

Table 3. Parameters used for the Log-Gaussian Cox Process model in the simulation study.

LGCP	1	2	3	4	5	6	7	8	9	10	11	12
μ	4.5	5.9	4.5	5.9	4.5	5.9	4.4	5.75	4.4	5.75	4.4	5.75
σ^2	0.25	0.25	0.25	0.25	0.25	0.25	0.5	0.5	0.5	0.5	0.5	0.5
ρ	0.05	0.05	0.1	0.1	0.15	0.15	0.05	0.05	0.1	0.1	0.15	0.15

tion, estimated $\xi(r)$ using the optimal bandwidths b_r obtained above, the AMSE bandwidth using (2.4), and Stoyan's bandwidth. We considered two methods for obtaining the AMSE and Stoyan bandwidths, one based on the true parameters of the process, and the other using estimated parameters. For the AMSE bandwidth, the *thomas.estK* function in the *spatstat* R package was used to obtain estimates of the parameters. For the Stoyan bandwidth, the estimated intensity is used. Thus, five bandwidths were used. We computed MSE(r) for each of them.

We also considered the performance of the method when the underlying point process was different from the model for the pair correlation function used to obtain the AMSE optimal bandwidths, so an incorrect g was used in (2.3). We generated realizations from the Matérn cluster point process model using the parameters listed in Table 2.

Upon a reviewer's suggestion, we also considered realizations from a Log-Gaussian Cox Process (LGCP) model where the random process Λ is such that $\log \Lambda$ is a Gaussian random field, and, given Λ, the point process N is inhomogeneous Poisson with intensity function Λ. The point process is stationary if Λ is stationary. We used the stationary Gaussian process with mean μ and exponential covariance function $C(r) = \sigma^2 \exp(-r/\rho)$, with parameter values given in Table 3.

We obtained AMSE optimal bandwidths using the optimal bandwidths specified by (2.4) and by (2.5). These bandwidths were then used to estimate the pair correlation function and the MSE's were computed. The parameters in (2.4) and (2.5) had to be estimated. In each case, we used minimum contrast between the model and the estimated K functions.

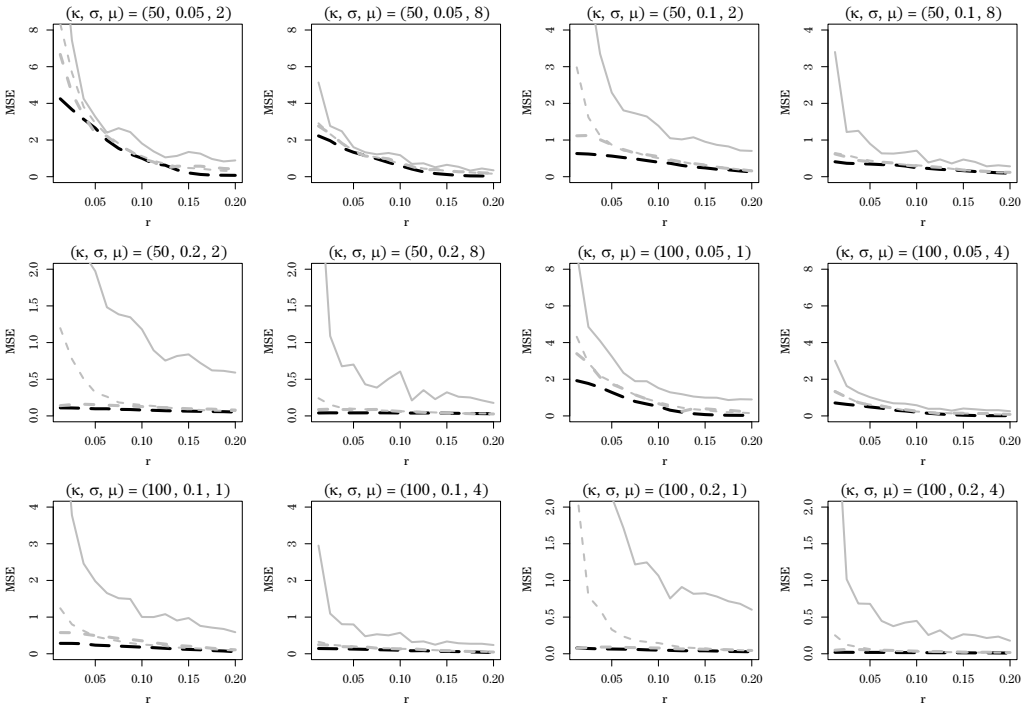

Figure 1. Plots showing the mean squared error (MSE) of estimates of the two-point correlation function of the Thomas process using empirically obtained optimal bandwidths (black dashed: – – –), bandwidths obtained using our AMSE method (gray thick – – – and thin dashed – – –) and using the Stoyan's rule of thumb (gray solid: ———). The thick and thin dashed lines represent MSE's from using, respectively, the true and estimated model parameters in the AMSE method.

3.1. Results

Figures 1 to 3 show the results of our simulation study. Each figure shows plots of the mean squared error for estimates of the two-point correlation function at distance$r$. The MSE of estimates obtained with the empirically obtained optimal bandwidths are the smallest in each case. When the true model parameters were used the MSE's of estimators were close to the MSE's of estimators based on the empirically obtained optimal bandwidths. Using estimated parameters results in a higher MSE, but not by much. MSE's obtained for estimates using Stoyan's simple formula for the bandwidth were the highest.

Results suggest that the performance of our AMSE approach can vary a bit with the particular formula used for the pair correlation function g, but is not overly sensitive to the choice. This is encouraging, since the optimal bandwidth

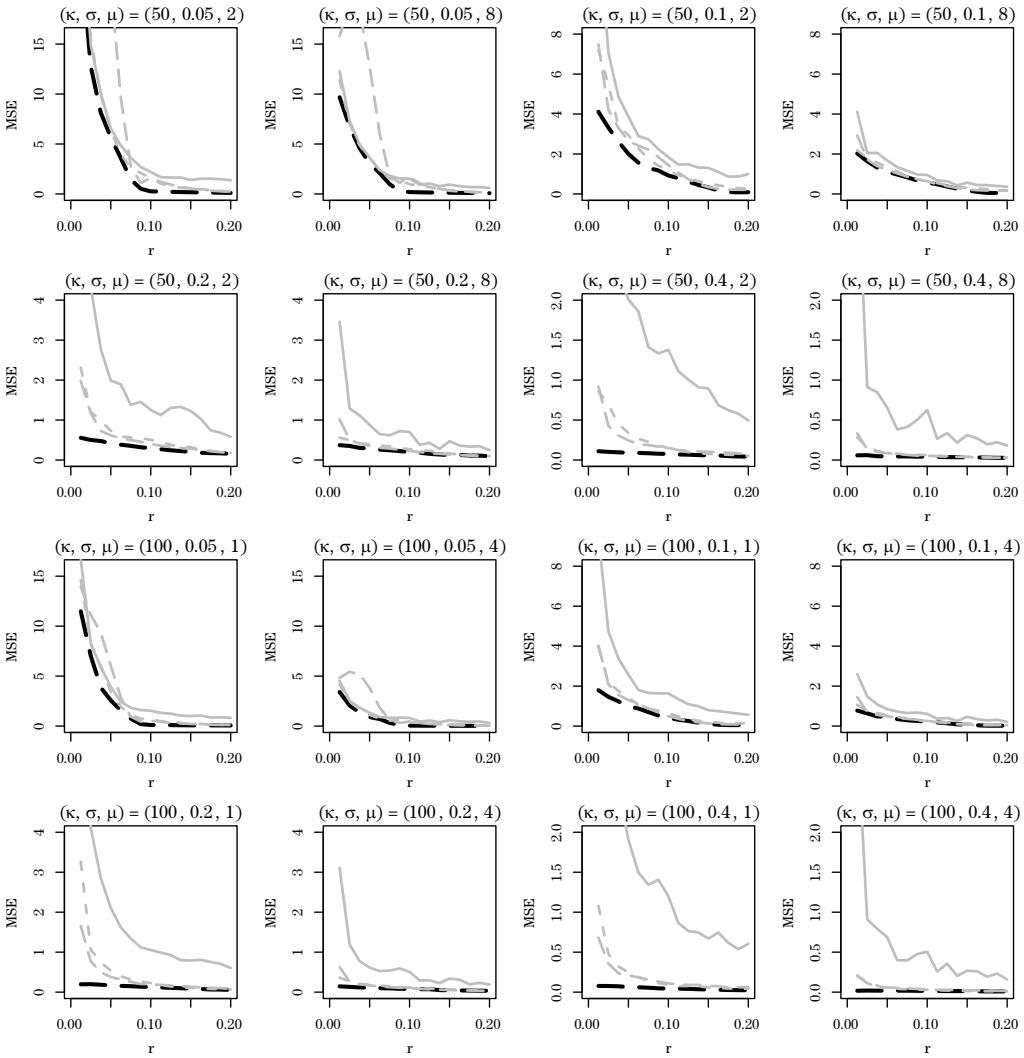

Figure 2. Plots showing the mean squared error (MSE) of estimates of the two-point correlation function of the Matérn cluster process using empirically obtained optimal bandwidths (black dashed: - - -), bandwidths obtained using our AMSE method based on the incorrect Thomas model (gray dashed: - - -), based on the incorrect power law model (gray long-dashed - - -) and using the Stoyan's rule of thumb (gray solid: ———).

h_{opt} for the 2PCF using Further, a functional form for g, like the power-law model, can be used and still achieve reasonable results.

We also did a small study with realizations from a determinantal point process and using AMSE optimal bandwidths based on the Thomas model, the

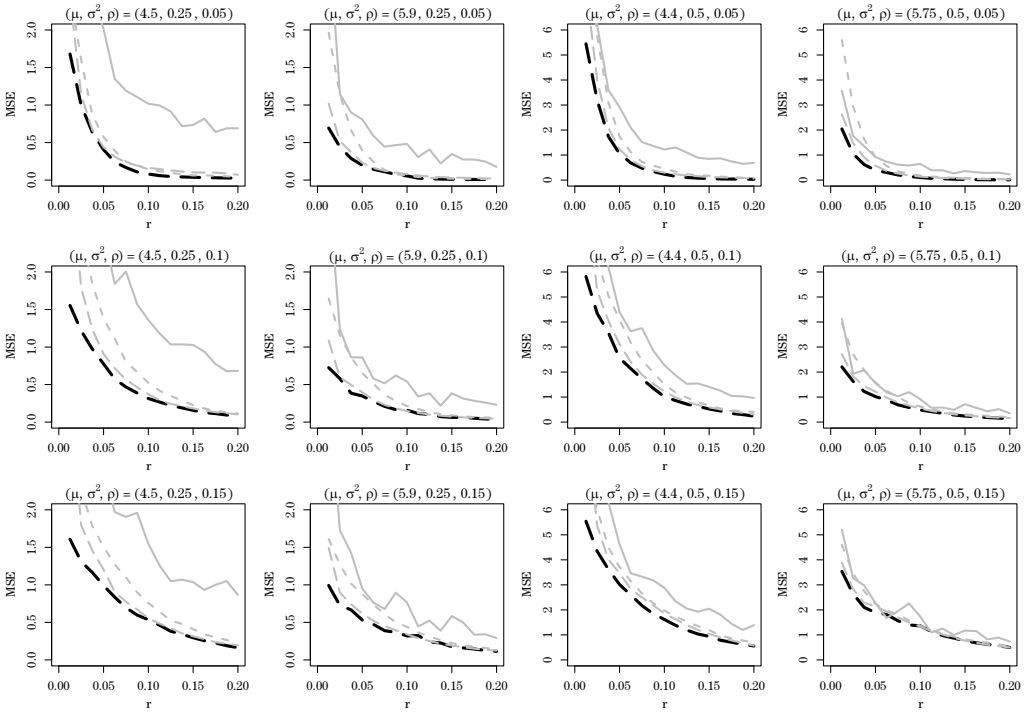

Figure 3. Plots showing the mean squared error (MSE) of estimates of the two-point correlation function of the Log-Gaussian Cox Process using empirically obtained optimal bandwidths (black dashed: – – –), bandwidths obtained using our AMSE method based on the incorrect Thomas model (gray dashed: - - - -), based on the incorrect power law model (gray long-dashed – – –) and using the Stoyan's rule of thumb (gray solid: ———).

power-law model, and the Stoyan bandwidth. It turned out that the empirically optimal bandwidths were large, and since we set a maximum bandwidth of 0.3 for the AMSE method, it did fairly well. The Stoyan rule-of-thumb suggested much smaller bandwidths and did not perform well.

4. Application to SDSS Data

The Sloan Digital Sky Survey (York et al. (2000)) was a major survey that began in 2000, covering about 35% of the sky, and collected observations on more than a million objects consisting of different types of objects such as galaxies and quasars. Goto et al. (2002) introduced a "cut-and-enhance" method for selecting clusters of galaxies from raw SDSS data and Basilakos and Plionis (2004) analyzed a subset of 200 of these galaxies. Loh and Jang (2010) used this data set in conjunction with a bootstrap bandwidth selection procedure. More

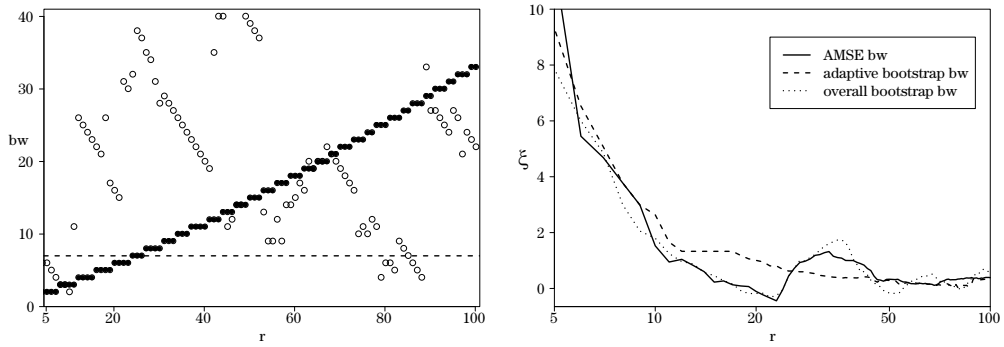

Figure 4. Plots showing, on the left, the overall (dashed line) and adaptive bandwidths (clear dots) obtained by a bootstrap bandwidth procedure (Loh and Jang, 2010) and adaptive bandwidths using AMSE (solid dots), and, on the right, estimates of the two-point correlation function using these bandwidths, on the log scale.

detailed description of the galaxy catalog, such as the regions in the sky in which they are located, can be found in Goto et al. (2002), Basilakos and Plionis (2004), and Loh and Jang (2010).

We obtained results from applying our AMSE bandwidth selection procedure to this galaxy catalog. We used (2.5) to obtain adaptive bandwidths for obtaining the Landy-Szalay estimator of ξ. We considered distances from 5 to 100 h^{-1} Mpc and used values of 20.7 and 1.6 for s_0 and γ, respectively. These values were obtained as estimates of s_0 and γ by Basilakos and Plionis (2004) through fitting a power-law model for ξ.

Figure 4 shows the bandwidths found from (2.5) and from the bootstrap bandwidth procedure of Loh and Jang (2010), and estimates of ξ using these bandwidths. We find that the adaptive bootstrap bandwidths using the method in Loh and Jang (2010) (black dots in Figure 4) are more variable, but also tend to be larger, producing much smoother estimates of ξ. The overall bootstrap bandwidth, represented in the left plot of Figure 4 by a horizontal dashed line, is a single value applying to all values of r. Its value of 7 is on the lower end of the range of the adaptive bootstrap bandwidths and produces a more jagged curve for ξ.

The AMSE bandwidths fall between these two. The bandwidths increase as r increase, with smaller bandwidths than both the overall and adaptive bootstrap bandwidths for $r < 40h^{-1}$ Mpc. For $r > 40h^{-1}$ Mpc, the AMSE bandwidths were larger than the overall bootstrap bandwidth, having values comparable to those of the adaptive bootstrap bandwidths. This behavior in the values of the

AMSE bandwidths is reflected in the resulting estimate of ξ, shown on the log scale in the right-hand plot of Figure 4. The estimate is slightly more jagged for $r < 40h^{-1}$ Mpc, but still comparable with the estimate obtained with the overall bootstrap bandwidth. Its smoothness for $r > 40h^{-1}$ Mpc lies between the two bootstrap bandwidth versions, closer to the estimate obtained using the adaptive bootstrap bandwidths.

In all, we find that the procedure performs reasonably well. The AMSE bandwidths are easy to obtain, unlike the bootstrap optimal bandwidths that require a more computationally expensive procedure.

5. Conclusion

We introduced the use of the AMSE approach as a method for obtaining optimal bandwidths for estimating the 2PCF that is widely used in astronomy. The AMSE optimal bandwidth method introduced here has a closed form solution that is easily computed. Our simulation studies suggest that it can lead to estimates of the 2PCF with substantially smaller mean square errors than other methods.

Supplementary Materials

The online supplementary material contains the derivation of (2.1).

Acknowledgment

We are grateful to the referees for their ideas and suggestions toward improving this work. Woncheol Jang's work was supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea grant by the Ministry of Education (No. 2013R1A1A2010065) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2014R1A4A1007895).

References

Baddeley, A., Moyeed, R., Howard, C. and Boyde, A. (1993). Analysis of a three-dimensional point pattern with replication. *Journal of the Royal Statistical Society C* **42**, 641–668.

Basilakos, S. and Plionis, M. (2004). Modeling the two-point correlation function of galaxy clusters in the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* **349**, 882–888.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*. 2nd edn, Arnold, London.

Dodelson, S. (2003). *Modern Cosmology*. Academic Press, New York.

Goto, T., Sekiguchi, M., Nichol, R. C., Bahcall, N. A., Kim, R. S. J., Annis, J., Ivezic, Z., Brinkmann, J., Hennessy, G. S., Szokoly, G. P. and Tucker, D. L. (2002). The cut-an-enhance method: selecting clusters of galaxies from the Sloan Digital Sky Survey commissioning data. *Astronomical Journal* **123**, 1807–1825.

Guan, Y. (2007). A least squares cross-validation bandwidth selection approach in pair correlation measures. *Statistics and Probability Letters* **77**, 1722–1729.

Hamilton, A. J. S. (1993). Towards better ways to measure the galaxy distribution. *Astrophysical Journal* **417**, 19–35.

Illian, J., Penttinen, A., Stoyan, H. and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*, John Wiley & Sons, Chichester.

Kerscher, M., Szapudi, I. and, Szalay, A. S. (2000). A comparison of estimators for the two point correlation function. *Astrophysical Journal Letters* **535**, 13–16.

Kim, S., Nowozin, S., Kohli, P. and Yoo, C. D. (2011). Higher-order correlation clustering for image segmentation. *Advances in Neural Information Processing Systems* **24**, 1530–1538.

Labatie, A., Starck, J.-L. and Lachièze-Rey, M. (2012). Uncertainty in 2-point correlation function estimators and baryon acoustic oscillation detection in galaxy surveys. *Statistical Methodology* **9**, 85–100.

Landy, S. D. and Szalay, A. S. (1993). Bias and variance of angular correlation functions. *Astrophysical Journal* **412**, 64–71.

Loh, J. M. and Jang, W. (2010). Estimating a cosmological mass bias parameter with semi-parametric bootstrap bandwidth selection. *Journal of the Royal Statistical Society, Series C* **59**, 761–779.

Martínez, V. J. and Saar, E. (2001). *Statistics of the Galaxy Distribution*. Chapman & Hall/CRC Press, New York.

Pons-Borderia, M.-J., Marinez, V. J., Stoyan, D., Stoyan, H. and Saar, E. (1999). Comparing estimators of the galaxy correlation function. *Astrophysical Journal* **523**, 480–491.

Ryden, B. (2003). *Introduction to Cosmology*. Addison-Wesley, Boston.

Sliverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC Press, New York.

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

Stoyan, D. (2006). Fundamentals of point process statistics. In *Case Studies in Spatial Point Process Modeling* (A. Baddeley, P. Gregori, J. Mateu, R. Stoica and D. Stoyan, eds.), 3–22, Springer, New York.

Stoyan, D. and Stoyan, H. (1994). *Fractals, Random Shapes and Point Fields*. Wiley, New York.

Szapudi, I., Postman, M., Lauer, T. R. and Oegerle, W. (2001). Observational constraints on higher order clustering up to $z \simeq 1$. *Astrophysical Journal* **548**, 114–126.

Thomas, M. (1949). A generalisation of Poisson's binomial limit for use in ecology. *Biometrika* **36**, 18–25.

Waagepetersen, R. and Guan, Y. (2009). Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society B* **71**, 685–702.

York, D., Adelman, J., Anderson, J. and others. (2000). The Sloan Digital Sky Survey: Technical summary. *Astronomical Journal* **120**, 1579–1587.

Room 323, Building 25 Department of Statistics College of Natural Sciences Seoul National University 1 Gwanak-ro, Gwanak-gu Seoul, 08826, Korea

E-mail: wcjang@snu.ac.kr

Dept of Mathematical Sciences 323 Martin Luther King Jr Blvd New Jersey Institute of Technology Newark, New Jersey 07102, USA

E-mail: loh@njit.edu

(Received November 2015; accepted July 2016)