# A Least-Squares Cross-Validation Bandwidth Selection Approach in Pair Correlation Function Estimations

Yongtao Guan*

April 19, 2007

## ABSTRACT

The pair correlation function is a useful tool to analyze spatial point patterns. It is often estimated nonparametrically by a procedure such as kernel smoothing. This article develops a data-driven method for the selection of the bandwidth involved in the estimation. The proposed method uses the idea of least-squares cross-validation which has been often applied for bandwidth selection in density estimation and many other nonparametric estimations. The asymptotic property of the proposed approach will be investigated under an increasing-domain setting in this article.

KEY WORDS: Bandwidth Selection, Least-Squares Cross-Validation, Pair Correlation Function.

---

*Yongtao Guan is Assistant Professor, Division of Biostatistics, Yale School of Public Health, Yale University, 60 College Street, New Haven, CT 06520-8034, e-mail yongtao.guan@yale.edu. This research was supported by NSF grant DMS-0603673. The author thanks the Editor and a referee for their comments which have greatly improved the paper.

# 1. INTRODUCTION

Consider a two-dimensional spatial point process $N$ that is observed over a domain of interest $D$. For an arbitrary Borel set $B \subset \mathbb{R}^2$, let $|B|$ denote the area of $B$, and $N(B)$ denote the number of events of $N$ that fall in $B$. Let $d\mathbf{x}$ be an infinitesimal region containing $\mathbf{x} \in \mathbb{R}^2$. Following Diggle (2003), we define the first- and second-order intensity functions of $N$ as

$$\lambda(\mathbf{x}) \equiv \lim_{|d\mathbf{x}| \to 0} \left\{ \frac{\mathrm{E}[N(d\mathbf{x})]}{|d\mathbf{x}|} \right\}, \text{ and}$$

$$\lambda_2(\mathbf{x}, \mathbf{y}) \equiv \lim_{|d\mathbf{x}|, |d\mathbf{y}| \to 0} \left\{ \frac{\mathrm{E}[N(d\mathbf{x})N(d\mathbf{y})]}{|d\mathbf{x}||d\mathbf{y}|} \right\},$$

respectively. An important summary function based on the foregoing two intensity functions is the pair correlation function (PCF), which is defined as follows:

$$g(\mathbf{x}, \mathbf{y}) = \frac{\lambda_2(\mathbf{x}, \mathbf{y})}{\lambda(\mathbf{x})\lambda(\mathbf{y})}.$$

A process is said to be second-order intensity reweighted stationary (SOIRS) if $g(\mathbf{x}, \mathbf{y}) = g(\mathbf{x} - \mathbf{y})$ following Møller and Waagepetersen (2004). It is said to be second-order reweighted isotropic (SOIRI) if $g(\mathbf{x}, \mathbf{y}) = g(||\mathbf{x} - \mathbf{y}||)$, where $|| \cdot ||$ stands for Euclidean norm.

The empirical PCF is often estimated by some nonparametric approach such as kernel smoothing. A typical kernel estimator of $g(\cdot)$ for an SOIRS process admits the following general form:

$$\hat{g}(\mathbf{t}; h) = \sum_{\mathbf{x} \neq \mathbf{y}} \sum \frac{k[(\mathbf{t} - \mathbf{x} + \mathbf{y})/h]}{D(\mathbf{x}, \mathbf{y})\lambda(\mathbf{x})\lambda(\mathbf{y})h^2}, \tag{1}$$

where $k(\cdot)$ is a two-dimensional kernel function and $D(\mathbf{x}, \mathbf{y})$ is an edge correction term, e.g. $D(\mathbf{x}, \mathbf{y}) = |(D - \mathbf{x}) \cap (D - \mathbf{y})|$ as in Stoyan and Stoyan (2000). For a SOIRI process, Møller and Waagepetersen (2004) suggested the following improved estimator for $g(\cdot)$:

$$\hat{g}(t; h) = \frac{1}{2\pi} \sum_{\mathbf{x} \neq \mathbf{y}} \sum \frac{k[(t - ||\mathbf{x} - \mathbf{y}||)/h]}{D(\mathbf{x}, \mathbf{y})\lambda(\mathbf{x})\lambda(\mathbf{y})||\mathbf{x} - \mathbf{y}||h}, \tag{2}$$

where $k$ now becomes a one-dimensional kernel function. In the stationary case, i.e. $\lambda(\mathbf{x}) = \lambda$ for some positive constant $\lambda$, Stoyan and Stoyan (1994) gave the following expression for the variance of $\hat{g}(t;h)$:

$$Var[\hat{g}(t;h)] = \frac{g(t)}{\pi t \bar{\gamma}(t)\lambda^2} \int_{-h}^{h} k_h^2(x)dx, \tag{3}$$

where $k_h(x) = k(x/h)/h$ and $\bar{\gamma}(t) = 1 - 4r/\pi + r^2/\pi$.

As in any nonparametric smoothing applications, the statistical properties of the resulting estimators, i.e., $\hat{g}(t)$ given in (1) and (2), are highly dependent on the choice of the bandwidth $h$. An inappropriate value of $h$ may lead to an estimator with a large bias or variance or both. It is thus necessary and important to develop data-driven methods which can be used to automatically and objectively select the bandwidth $h$.

In the case that $N$ is stationary, some general guidelines on the selection of $h$, although not completely satisfactory, are available. For example, Stoyan and Stoyan (1994, p.285) recommended using $h = c\lambda^{-1/2}$ for Epanechnikov kernels with $c = .1 - .2$ for planar point patterns of 50-300 points. Stoyan and Stoyan (2000) suggested an alternative method which uses an approximation of the variance of the empirical PCF for given bandwidth and lag; so the optimal $h$ may depend on lag. Guan *et al.* (2005) recently proposed a subsampling approach in selecting the bandwidth. However, all these procedures were developed under the assumption that the underlying process is stationary and thus are not appropriate for a more general SOIRS process. The purpose of this article is to develop a data-driven procedure to select the bandwidth used to estimate the PCF for SOIRS processes. The proposed procedure is a familiar least-squares cross-validation (LSCV) type of procedure which can be easily applied in practice.

## 2. LEAST-SQUARES CROSS-VALIDATION

Let $r$ be the largest lag for which the PCF is to be estimated and $\hat{g}^{-(\mathbf{x},\mathbf{y})}(\cdot)$ be the empirical PCF estimated by deleting events $\mathbf{x}$ and $\mathbf{y}$. We impose an upper limit $r$ on the lags here due

to the fact that events separated by a large lag are often independent and thus the estimation of the PCF for such lags may not be interesting. We propose to select $h$ as the minimizer of the following LSCV criterion:

$$M(h) = \int_{||\mathbf{u}||\le r} [\hat{g}(\mathbf{u}; h)]^2 d\mathbf{u} - 2 \sum_{0<||\mathbf{x}-\mathbf{y}||\le r} \sum \frac{\hat{g}^{-(\mathbf{x},\mathbf{y})}(\mathbf{x}-\mathbf{y}; h)}{|D \cap D - \mathbf{x} + \mathbf{y}|\lambda(\mathbf{x})\lambda(\mathbf{y})}, \tag{4}$$

where $|D \cap D - \mathbf{x} + \mathbf{y}|$ is a translation edge correction introduced by Ohser and Stoyan (1981). Note that (4) has a similar form as the LSCV criterion used in density estimations (e.g., Silverman 1998, p.49) and other nonparametric estimation in general (e.g., Hart 1997). However, it's worth noting that a pair of events are left out each time while calculating the double sums in (4), which is in contrast to traditional cross-validation criteria where only one observation is omitted each time.

We study the asymptotic properties of the LSCV criterion and the resulting bandwidth under an increasing-domain setting. Specifically, consider a sequence of domains of interest $D_n$. Let $\partial D_n$ denote the boundary of $D_n$ and $|\partial D_n|$ denote the length of $\partial D_n$. We assume

$$|D_n| = O(n^2) \text{ and } |\partial D_n| = O(n). \tag{5}$$

Condition (5) simply says the domains of interest are truly spatial and need to increase in all directions as $n$ increases.

To quantify the dependence strength in $N$, we first define the $k$th-order cumulant function as follows:

$$Q_k(\mathbf{x}_1, \cdots, \mathbf{x}_k) = \lim_{|d\mathbf{x}_i|\to 0} \left\{ \frac{\text{Cum}[N(d\mathbf{x}_1), \cdots, N(d\mathbf{x}_k)]}{|d\mathbf{x}_1| \cdots |d\mathbf{x}_k|} \right\}, \quad i = 1, \cdots, k,$$

where $\text{Cum}(Y_1, \cdots, Y_k)$ is the coefficient of $i^k t_1 \cdots t_k$ in the Taylor series expansion of $\log[\text{E}[\exp(i \sum_{j=1}^k Y_j t_j)]]$ about the origin (e.g., Brillinger 1975). Further define

$$C_k(\mathbf{x}_1, \cdots, \mathbf{x}_k) = \frac{Q_k(\mathbf{x}_1, \cdots, \mathbf{x}_k)}{\lambda(\mathbf{x}_1) \cdots \lambda(\mathbf{x}_k)}.$$

We assume the following mild conditions on the spatial point process $N$:

$$C_k(\mathbf{x}_1, \cdots, \mathbf{x}_k) = C_k(\mathbf{x}_2 - \mathbf{x}_1, \cdots, \mathbf{x}_k - \mathbf{x}_1) \text{ for } k = 1, \cdots, 4, \tag{6}$$

$$C_2(\mathbf{u}) \text{ is bounded, and } \int_{\mathbb{R}^2} |C_2(\mathbf{u})| d\mathbf{u} < \infty, \tag{7}$$

$$\int_{\mathbb{R}^2} |C_3(\mathbf{u}_1, \mathbf{u}_2)| d\mathbf{u}_1 < \infty, \ \int_{\mathbb{R}^2} |C_3(\mathbf{u}_1, \mathbf{u}_1 + \mathbf{u}_2)| d\mathbf{u}_1 < \infty \text{ for all } \mathbf{u}_2, \tag{8}$$

$$\int_{\mathbb{R}^2} |C_4(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_2 + \mathbf{u}_3)| d\mathbf{u}_2 < \infty \text{ for any } \mathbf{u}_1, \mathbf{u}_3. \tag{9}$$

Let $M_n(h)$ denote $M(h)$ defined in (4) but calculated on $D_n$. The following theorem establishes the connection between $M_n(h)$ and the integrated squared error defined as:

$$R_n(h) = \int_{||\mathbf{u}|| \leq r} [\hat{g}_n(\mathbf{u}; h) - g(\mathbf{u})]^2 d\mathbf{u}. \tag{10}$$

*Theorem 1.* Assume that that $N$ is an SOIRS spatial point process and that conditions (5)-(10) hold. Then

$$E[M_n(h)] \to E[R_n(h)] - \int_{||\mathbf{u}|| \leq r} [g(\mathbf{u})]^2 d\mathbf{u} \text{ as } n \to \infty.$$

*Proof.* First we define the following two functions of $h$:

$$A_n(h) = \int_{||\mathbf{u}|| \leq r} \hat{g}_n(\mathbf{u}; h) g(\mathbf{u}) d\mathbf{u}, \text{ and}$$

$$B_n(h) = \sum_{0 < ||\mathbf{x}-\mathbf{y}|| \leq r} \sum \frac{\hat{g}_n^{-(\mathbf{x},\mathbf{y})}(\mathbf{x} - \mathbf{y}; h)}{|D_n \cap D_n - \mathbf{x} + \mathbf{y}| \lambda(\mathbf{x}) \lambda(\mathbf{y})}.$$

Note that in order to prove Theorem 1, we only need to show $E[B_n(h)] \to E[A_n(h)]$. Also note that

$$\begin{aligned} A_n(h) &= \int_{||\mathbf{u}|| \leq r} \left\{ \sum_{\mathbf{x} \neq \mathbf{y}} \sum \frac{k[(\mathbf{u} - \mathbf{x} + \mathbf{y})/h]}{|D_n \cap D_n - \mathbf{x} + \mathbf{y}| \lambda(\mathbf{x}) \lambda(\mathbf{y}) h^2} \right\} g(\mathbf{u}) d\mathbf{u} \\ &= \sum_{\mathbf{x} \neq \mathbf{y}} \sum \frac{\int_{||\mathbf{u}|| \leq r} k[(\mathbf{u} - \mathbf{x} + \mathbf{y})/h] g(\mathbf{u}) d\mathbf{u}}{|D_n \cap D_n - \mathbf{x} + \mathbf{y}| \lambda(\mathbf{x}) \lambda(\mathbf{y}) h^2}. \end{aligned}$$

Thus

$$E[A_n(h)] = \int_{\mathbf{x}, \mathbf{y} \in D_n} \int \left\{ \int_{||\mathbf{u}|| \leq r} k[(\mathbf{u} - \mathbf{x} + \mathbf{y})/h] g(\mathbf{u}) d\mathbf{u} \right\} \frac{g(\mathbf{y} - \mathbf{x})}{|D_n \cap D_n - \mathbf{x} + \mathbf{y}| h^2} d\mathbf{x} d\mathbf{y}$$

$$
\begin{aligned}
&= \int_{||\mathbf{u}||\le r}\int_{\mathbf{v}\in D_n - D_n} \frac{k[(\mathbf{u}-\mathbf{v})/h]g(\mathbf{u})g(\mathbf{v})}{h^2} d\mathbf{u}d\mathbf{v} \\
&= \int\int\int\int_{\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{w}\in D_n, ||\mathbf{z}-\mathbf{w}||\le r} \frac{k[(\mathbf{z}-\mathbf{w}-\mathbf{x}+\mathbf{y})/h]g(\mathbf{y}-\mathbf{x})g(\mathbf{z}-\mathbf{w})}{|D_n\cap D_n - \mathbf{x}+\mathbf{y}||D_n\cap D_n-\mathbf{z}+\mathbf{w}|h^2} d\mathbf{z}d\mathbf{w}d\mathbf{x}d\mathbf{y}.
\end{aligned}
$$

Clearly $E[A_n(h)] = O(1)$. To prove Theorem 1, we only need to show that

$$E[A_n(h)] - E[B_n(h)] \to 0 \text{ as } n\to\infty.$$

To show this, first note that

$$
B_n(h) = \sum\sum\sum\sum_{0<||\mathbf{x}-\mathbf{y}||\le r, \mathbf{w}\neq\mathbf{z}} \frac{k[(\mathbf{x}-\mathbf{y}-\mathbf{w}+\mathbf{z})/h]}{|D_n\cap D_n-\mathbf{x}+\mathbf{y}||D_n\cap D_n-\mathbf{w}+\mathbf{z}|\lambda(\mathbf{x})\lambda(\mathbf{y})\lambda(\mathbf{w})\lambda(\mathbf{z})h^2}.
$$

Now define

$$
g_4(\mathbf{x},\mathbf{y},\mathbf{z}) \equiv \lim_{|d\mathbf{0}|,|d\mathbf{x}|,|d\mathbf{y}|,|d\mathbf{z}|\to 0} \left\{ \frac{\mathrm{E}[N(d\mathbf{0})N(d\mathbf{x})N(d\mathbf{y})N(d\mathbf{z})]}{|d\mathbf{0}||d\mathbf{x}||d\mathbf{y}||d\mathbf{z}|\lambda(\mathbf{0})\lambda(\mathbf{x})\lambda(\mathbf{y})\lambda(\mathbf{z})} \right\},
$$

Then

$$
E[B_n(h)] = \int\int\int\int_{\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{w}\in D_n, ||\mathbf{z}-\mathbf{w}||\le r} \frac{k[(\mathbf{z}-\mathbf{w}-\mathbf{x}+\mathbf{y})/h]g_4(\mathbf{y}-\mathbf{x},\mathbf{z}-\mathbf{x},\mathbf{w}-\mathbf{x})}{|D_n\cap D_n-\mathbf{x}+\mathbf{y}||D_n\cap D_n-\mathbf{z}+\mathbf{w}|h^2} d\mathbf{z}d\mathbf{w}d\mathbf{x}d\mathbf{y}.
$$

Define $f(\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3) = g_4(\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3) - g(\mathbf{y}_1)g(\mathbf{u}_2-\mathbf{u}_3)$. In terms of the defined cumulant functions, lengthy but elementary algebra yields

$$
\begin{aligned}
f(\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3) &= C_4(\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3) \\
&+ C_3(\mathbf{u}_1,\mathbf{u}_2) + C_3(\mathbf{u}_1,\mathbf{u}_3) + C_3(\mathbf{u}_2,\mathbf{u}_3) + C_3(\mathbf{u}_2-\mathbf{u}_1,\mathbf{u}_3-\mathbf{u}_1) \\
&+ C_2(\mathbf{u}_2)C_2(\mathbf{u}_3-\mathbf{u}_1) + C_2(\mathbf{u}_3)C_2(\mathbf{u}_2-\mathbf{u}_1) \\
&+ C_2(\mathbf{u}_2) + C_2(\mathbf{u}_3) + C_2(\mathbf{u}_2-\mathbf{u}_1) + C_2(\mathbf{u}_3-\mathbf{u}_1).
\end{aligned}
$$

Thus

$$
\begin{aligned}
&E[A_n(h)] - E[B_n(h)] \\
&= \int\int\int\int_{\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{w}\in D_n, ||\mathbf{z}-\mathbf{w}||\le r} \frac{k[(\mathbf{z}-\mathbf{w}-\mathbf{x}+\mathbf{y})/h]f(\mathbf{y}-\mathbf{x},\mathbf{z}-\mathbf{x},\mathbf{w}-\mathbf{x})}{|D_n\cap D_n-\mathbf{x}+\mathbf{y}||D_n\cap D_n-\mathbf{z}+\mathbf{w}|h^2} d\mathbf{z}d\mathbf{w}d\mathbf{x}d\mathbf{y} \\
&= \int\int\int_{||\mathbf{u}_3-\mathbf{u}_2||\le r} \frac{k[(\mathbf{u}_3-\mathbf{u}_2+\mathbf{u}_1)/h]f(\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3)|D_n\cap D_n-\mathbf{u}_1\cap D_n-\mathbf{u}_2\cap D_n-\mathbf{u}_3|}{|D_n\cap D_n+\mathbf{u}_1||D_n\cap D_n-\mathbf{u}_3+\mathbf{u}_2|h^2} d\mathbf{u}_1 d\mathbf{u}_2 d\mathbf{u}_3.
\end{aligned}
$$

Further

$$
\begin{aligned}
|E[A_n(h)] - E[B_n(h)]| \;\leq\; & \int\int\int_{||\mathbf{u}_3-\mathbf{u}_2||\leq r} \frac{k[(\mathbf{u}_3-\mathbf{u}_2+\mathbf{u}_1)/h]|f(\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3)|}{|D_n\cap D_n-\mathbf{u}_3+\mathbf{u}_2|h^2}d\mathbf{u}_1d\mathbf{u}_2d\mathbf{u}_3 \\
\leq\; & \int\int\int_{||\mathbf{u}_3-\mathbf{u}_2||\leq r} \frac{k[(\mathbf{u}_3-\mathbf{u}_2+\mathbf{u}_1)/h]|C_4(\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3)|}{|D_n\cap D_n-\mathbf{u}_3+\mathbf{u}_2|h^2}d\mathbf{u}_1d\mathbf{u}_2d\mathbf{u}_3 \\
+\; & \int\int\int_{||\mathbf{u}_3-\mathbf{u}_2||\leq r} \frac{k[(\mathbf{u}_3-\mathbf{u}_2+\mathbf{u}_1)/h]|C_3(\mathbf{u}_1,\mathbf{u}_2)|}{|D_n\cap D_n-\mathbf{u}_3+\mathbf{u}_2|h^2}d\mathbf{u}_1d\mathbf{u}_2d\mathbf{u}_3 \\
+\; & \int\int\int_{||\mathbf{u}_3-\mathbf{u}_2||\leq r} \frac{k[(\mathbf{u}_3-\mathbf{u}_2+\mathbf{u}_1)/h]|C_3(\mathbf{u}_1,\mathbf{u}_3)|}{|D_n\cap D_n-\mathbf{u}_3+\mathbf{u}_2|h^2}d\mathbf{u}_1d\mathbf{u}_2d\mathbf{u}_3 \\
+\; & \int\int\int_{||\mathbf{u}_3-\mathbf{u}_2||\leq r} \frac{k[(\mathbf{u}_3-\mathbf{u}_2+\mathbf{u}_1)/h]|C_3(\mathbf{u}_2,\mathbf{u}_3)|}{|D_n\cap D_n-\mathbf{u}_3+\mathbf{u}_2|h^2}d\mathbf{u}_1d\mathbf{u}_2d\mathbf{u}_3 \\
+\; & \int\int\int_{||\mathbf{u}_3-\mathbf{u}_2||\leq r} \frac{k[(\mathbf{u}_3-\mathbf{u}_2+\mathbf{u}_1)/h]|C_3(\mathbf{u}_2-\mathbf{u}_1,\mathbf{u}_3-\mathbf{u}_1)|}{|D_n\cap D_n-\mathbf{u}_3+\mathbf{u}_2|h^2}d\mathbf{u}_1d\mathbf{u}_2d\mathbf{u}_3 \\
+\; & \int\int\int_{||\mathbf{u}_3-\mathbf{u}_2||\leq r} \frac{k[(\mathbf{u}_3-\mathbf{u}_2+\mathbf{u}_1)/h]|C_2(\mathbf{u}_2)C_2(\mathbf{u}_3-\mathbf{u}_1)|}{|D_n\cap D_n-\mathbf{u}_3+\mathbf{u}_2|h^2}d\mathbf{u}_1d\mathbf{u}_2d\mathbf{u}_3 \\
+\; & \int\int\int_{||\mathbf{u}_3-\mathbf{u}_2||\leq r} \frac{k[(\mathbf{u}_3-\mathbf{u}_2+\mathbf{u}_1)/h]|C_2(\mathbf{u}_3)C_2(\mathbf{u}_2-\mathbf{u}_1)|}{|D_n\cap D_n-\mathbf{u}_3+\mathbf{u}_2|h^2}d\mathbf{u}_1d\mathbf{u}_2d\mathbf{u}_3 \\
+\; & \int\int\int_{||\mathbf{u}_3-\mathbf{u}_2||\leq r} \frac{k[(\mathbf{u}_3-\mathbf{u}_2+\mathbf{u}_1)/h]|C_2(\mathbf{u}_2)|}{|D_n\cap D_n-\mathbf{u}_3+\mathbf{u}_2|h^2}d\mathbf{u}_1d\mathbf{u}_2d\mathbf{u}_3 \\
+\; & \int\int\int_{||\mathbf{u}_3-\mathbf{u}_2||\leq r} \frac{k[(\mathbf{u}_3-\mathbf{u}_2+\mathbf{u}_1)/h]|C_2(\mathbf{u}_3)|}{|D_n\cap D_n-\mathbf{u}_3+\mathbf{u}_2|h^2}d\mathbf{u}_1d\mathbf{u}_2d\mathbf{u}_3 \\
+\; & \int\int\int_{||\mathbf{u}_3-\mathbf{u}_2||\leq r} \frac{k[(\mathbf{u}_3-\mathbf{u}_2+\mathbf{u}_1)/h]|C_2(\mathbf{u}_2-\mathbf{u}_1)|}{|D_n\cap D_n-\mathbf{u}_3+\mathbf{u}_2|h^2}d\mathbf{u}_1d\mathbf{u}_2d\mathbf{u}_3 \\
+\; & \int\int\int_{||\mathbf{u}_3-\mathbf{u}_2||\leq r} \frac{k[(\mathbf{u}_3-\mathbf{u}_2+\mathbf{u}_1)/h]|C_2(\mathbf{u}_3-\mathbf{u}_1)|}{|D_n\cap D_n-\mathbf{u}_3+\mathbf{u}_2|h^2}d\mathbf{u}_1d\mathbf{u}_2d\mathbf{u}_3.
\end{aligned}
$$

Denote the eleven terms on the right hand side of equality in the above as $T_1$ until $T_{11}$ in turn and let $\mathbf{u}=\mathbf{u}_1$, $\mathbf{w}=\mathbf{u}_2$ and $\mathbf{v}=\mathbf{u}_3-\mathbf{u}_2$. A simple change of variable yields

$$
T_1 \leq \int_{||\mathbf{v}||\leq r}\int_{\mathbb{R}^2} \frac{k[(\mathbf{v}+\mathbf{u})/h]}{|D_n\cap D_n-\mathbf{v}|h^2}\left[\int_{\mathbb{R}^2}|C_4(\mathbf{u},\mathbf{v}+\mathbf{w},\mathbf{w})|d\mathbf{w}\right]d\mathbf{u}d\mathbf{v}\to 0,
$$

$$T_2 \leq \int_{||\mathbf{v}||\leq r} \int_{\mathbb{R}^2} \frac{k[(\mathbf{v}+\mathbf{u})/h]}{|D_n \cap D_n - \mathbf{v}|h^2} \left[ \int_{\mathbb{R}^2} |C_3(\mathbf{u},\mathbf{w})| d\mathbf{w} \right] d\mathbf{u} d\mathbf{v} \to 0,$$

$$T_6 \leq \int_{||\mathbf{v}||\leq r} \int_{\mathbb{R}^2} \frac{k[(\mathbf{v}+\mathbf{u})/h]}{|D_n \cap D_n - \mathbf{v}|h^2} \left[ \int_{\mathbb{R}^2} |C_2(\mathbf{w}) C_2(\mathbf{v}+\mathbf{w}-\mathbf{u})| d\mathbf{w} \right] d\mathbf{u} d\mathbf{v} \to 0,$$

$$T_8 \leq \int_{||\mathbf{v}||\leq r} \int_{\mathbb{R}^2} \frac{k[(\mathbf{v}+\mathbf{u})/h]}{|D_n \cap D_n - \mathbf{v}|h^2} \left[ \int_{\mathbb{R}^2} |C_2(\mathbf{w})| d\mathbf{w} \right] d\mathbf{u} d\mathbf{v} \to 0.$$

Note that $T_3$, $T_4$ and $T_5$ all go to zero following the result $T_2 \to 0$; $T_7$ goes to zero following $T_6 \to 0$; and $T_9$, $T_{10}$ and $T_{11}$ all go to zero following $T_8 \to 0$. Thus $|E[A_n(h)] - E[B_n(h)]| \to 0$. □

## 3. SIMULATION

### 3.1 A modified estimator for the PCF

The PCF estimator given by (2) tends to have a large bias at small lags. To reduce the bias, we consider the following modified version of (2):

$$\hat{g}^*(t;h) = \frac{\hat{g}(t;h)}{\int_{-1}^{\min(1,t/h)} k(x)dx}, \tag{11}$$

where $\hat{g}(t;h)$ is as defined in (2). The denominator of (11) serves as a bias correction term. It is equal to one if $t \geq h$ and thus $\hat{g}^*(t;h) = \hat{g}(t;h)$, but is smaller than one if $t < h$, which accounts for the fact that the distance between any two distinct events is larger than zero. In the homogeneous Poisson process case, the corresponding estimator of (11) for $\lambda_2(t)$ $(= \lambda^2)$ is

$$\hat{\lambda}_2^*(t;h) = \frac{1}{2\pi \int_{-1}^{\min(1,t/h)} k(x)dx} \sum_{\mathbf{x}} \sum_{\neq \mathbf{y}} \frac{k[(t - ||\mathbf{x}-\mathbf{y}||)/h]}{D(\mathbf{x},\mathbf{y})||\mathbf{x}-\mathbf{y}||h}.$$

Note that $\hat{\lambda}_2^*(t;h)$ is unbiased for $\lambda_2(t)$ since

$$\begin{aligned} E[\hat{\lambda}_2^*(t;h)] &= \frac{\lambda^2}{2\pi \int_{-1}^{\min(1,t/h)} k(x)dx} \int_D\int_D \frac{k[(t-||\mathbf{x}-\mathbf{y}||)/h]}{D(\mathbf{x},\mathbf{y})||\mathbf{x}-\mathbf{y}||h} d\mathbf{x}d\mathbf{y} \\ &= \frac{\lambda^2}{2\pi \int_{-1}^{\min(1,t/h)} k(x)dx} \int_{D-D} \frac{k[(t-||\mathbf{u}||)/h]}{||\mathbf{u}||h} d\mathbf{u} \\ &= \frac{\lambda^2}{\int_{-1}^{\min(1,t/h)} k(x)dx} \int_0^{t+h} \frac{k[(t-r)/h]}{h} d\mathbf{r} \\ &= \lambda^2. \end{aligned}$$

Thus $\hat{g}^*(t;h)$ is approximately unbiased for $g(t)$. By applying the same argument that was used to derive (3) in Stoyan *et al.* (1993), the variance of $\hat{g}^*(t;h)$ can be analogously obtained as follows:

$$Var[\hat{g}^*(t;h)] = \frac{g(t) \int_{-1}^{\min(1,t/h)} k^2(x)dx}{\pi h t \bar{\gamma}(t) \lambda^2 [\int_{-1}^{\min(1,t/h)} k(x)dx]^2}. \tag{12}$$

### 3.2 Results

We applied the proposed bandwidth selection method to data generated by a Poisson process, a Poisson cluster process (PCP), and a simple inhibition process (SIP) of the Matérn's first type (e.g., Diggle 2003). In each case, we simulated 500 realizations on a unit square. The expected number of events per realization was 100 for the Poisson and PCP case but 400 for the SIP case. In the PCP case, we set the number of expected parents, $\rho = 25$, and used a radially symmetric normal distribution (see, e.g., Diggle 2003) for the dispersion of offspring relative to the parent. The spread parameter, $\sigma$, was set at .02 and .04, which in turn correspond to relatively strong and weak clustering. In the SIP case, we set the inhibition parameter $\delta = .015$, i.e. no two events had an inter-event distance less than $.015$. Note that the "optimal" bandwidth recommended by Stoyan and Stoyan (1994) is between $.01$ and $.02$ in the Poisson and PCP case, but between $.005$ and $.01$ in the SIP case.

In the Poisson case, recall that $\hat{g}^*(t;h)$ is approximately unbiased for $g(t)$. Following the expression of the variance of $\hat{g}^*(t;h)$ in (12), we expected the "optimal" $h$ to be as

large as possible. In the simulation, we imposed an arbitrary upper limit ($= .2$) for $h$ so as to mimic the reality that we do not know if a process is Poisson or not in practice. For the upper bound for the lags, we set $r = .2$ in the Poisson case, $r = 4\sigma$ in the PCP case, and $r = 3\delta$ in the SIP case.

Figure 1 presents the histograms for the bandwidths that were selected by minimizing (4). In the PCF case, there was a clear pattern that as the strength of clustering weakened, the selected "optimal" bandwidth became larger. In the Poisson process case, the selected bandwidths were at or very close to $.2$ for an overwhelmingly large percentage of the time. In this case, $h = .2$ indeed was the true "optimal" bandwidth among the bandwidthds being considered in the simulation. In the SIP case, the selected bandwidths had a tendency to be smaller than what was implied by Stoyan and Stoyan's recommendation. We thus expected that the PCF estimator using a bandwidth selected by the proposed procedure could better capture the jump of the PCF at the hard-core distance (i.e. $\delta$).

Figure 2 plots the variance and mean squared error (MSE) for $\hat{g}^*(t; h)$ for three different bandwidth values. The first was obtained by minimizing (4); the second was from Stoyan and Stoyan's recommendation with $c = .15$; and the third was the true "optimal" bandwidth. For the ease of presentation, we denote the three bandwidths by $h_{cv}$, $h_{ss}$ and $h_{op}$, respectively. From Figure 2, we see that $\hat{g}^*(t; h_{cv})$ performed similarly to $\hat{g}^*(t; h_{ss})$ in the PCP case with $\sigma = .02$, but performed much better than $\hat{g}^*(t; h_{ss})$ in the PCP case with $\sigma = .04$ and in the Poisson case. The latter was because $h_{ss}$ was generally much smaller than $h_{cv}$, which in turn led to a much larger variance. In the SIP case, the MSE for $\hat{g}^*(t; h_{cv})$ was smaller than that for $\hat{g}^*(t; h_{ss})$ around the hard core jumps. This confirmed confirmed our conjecture that $\hat{g}^*(t; h_{cv})$ could better capture this jump than $\hat{g}^*(t; h_{ss})$.

## REFERENCES


Brillinger, D. R. (1975), *Time Series: Data Analysis and Theory*, Holt, Rinehart & Winston.

Diggle, P. J. (2003), *Statistical Analysis of Spatial Point Patterns*, New York: Oxford University Press Inc.

Guan, Y., Sherman, M. and Calvin, J. A. (2006), "Assessing Isotropy for Spatial Point Processes", *Biometrics*, 119-125.

Hart, J. (1997), *Nonparametric Smoothing and Lack-of-Fit Tests*, New York, Springer.

Møller, J. and Waagepetersen, R. P. (2004), *Statistical Inference and Simulation for Spatial Point Processes*, New York: Chapman & Hall.

Ohser, J. and Stoya, D. (1981), "On the Second-Order and Orientation Analysis of Planar Stationary Point Processes", *Biometrical Journal*, 23, 523-533.

Silverman, B. W. (1998), *Density Estimation for Statistics and Data Analysis*, New York: Chapman & Hall.

Stoyan, D., Bertram, U. and Wendrock, H. (1993) "Estimation Variances for Estimators of Product Densities and Pair Correlation Functions of Planar Point Processes", *Annals of the Institute of Statistical Mathematics*, 45, 211–221

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

Stoyan, D. and Stoyan, H. (2000), "Improving Ratio Estimators of Second Order Point Process Characteristics", *Scandinavian Journal of Statistics*, 28, 641–656.

Waagepetersen, R. P. (2007), "An Estimating Function Approach to Inference for Inhomogeneous Neyman-Scott Processes", 252-258.

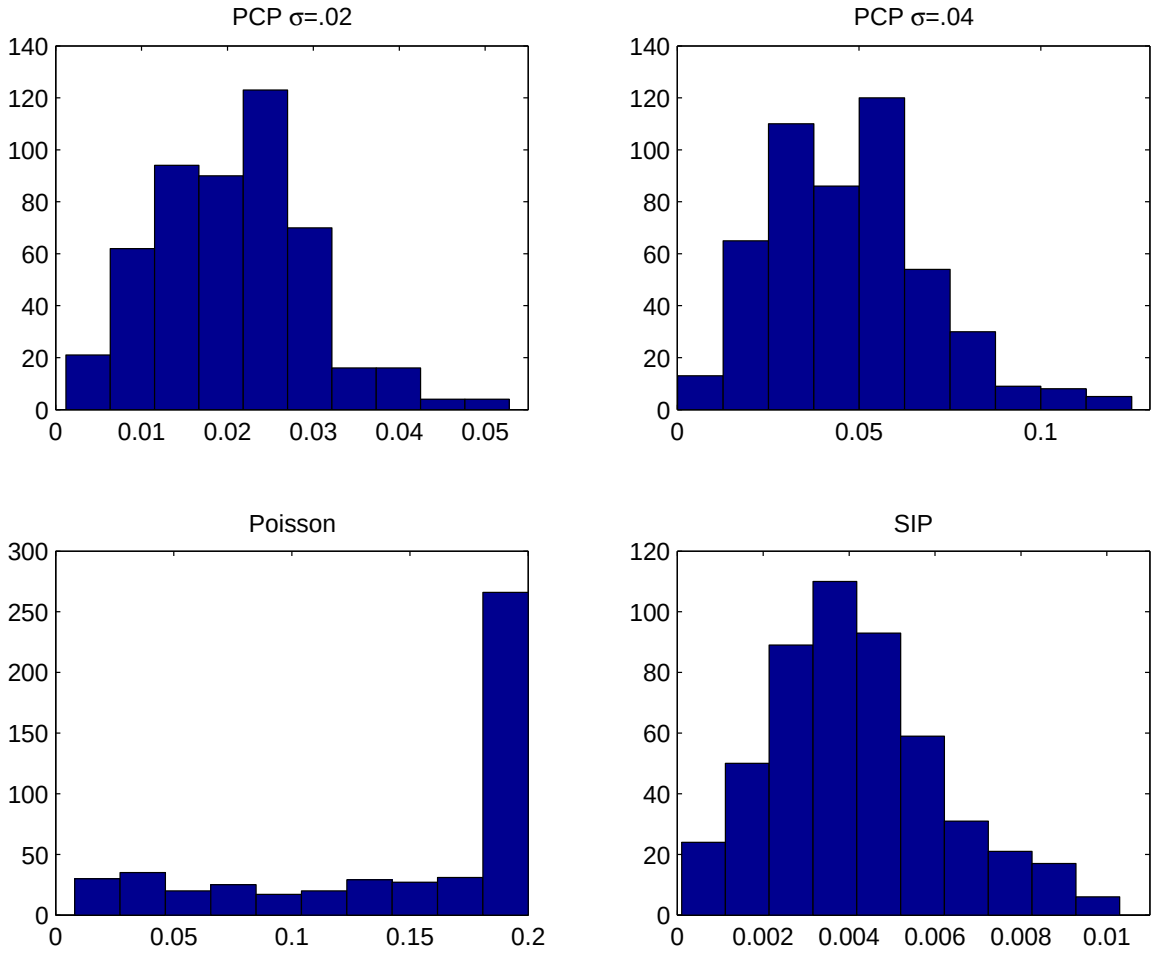


*Figure 1. Histograms for bandwidths that were selected by minimizing (4) for the Poisson cluster process (PCP) with $\sigma = .02$, $\sigma = .04$, the Poisson process, and the simple inhibition process (SIP).*

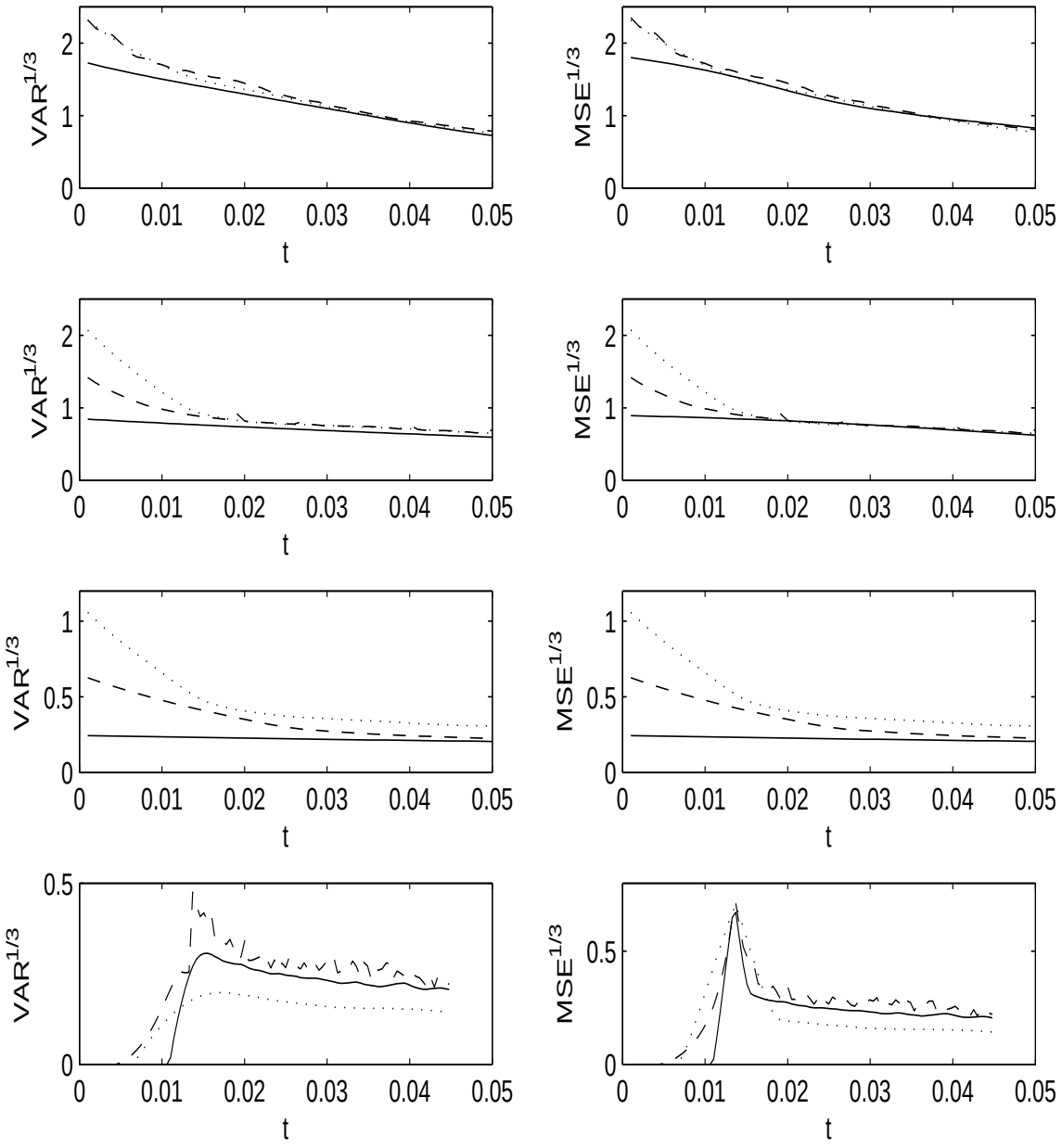


*Figure 2. Variances (left panel) and mean squared errors (right panel) of the empirical PCF. From top to bottom, the plots are for the Poisson cluster process with $\sigma = .02$, $\sigma = .04$, the Poisson process, and the simple inhibition process. The bandwidth being used are Stoyan & Stoyan's $h$ with $c = .15$ $(\cdots)$, the estimated optimal $h$ by minimizing (4) $(---)$, and the true optimal $h$ (——).*