Bootstrapped Spatial Point Processes

Brian S. Yandell
University of Wisconsin–Madison

1. Introduction

The bootstrap idea of Efron (1979, 1982) can be carried over directly to spatial point processes. Quite often we consider statistics which have complicated distributions under complete spatial randomness, and nearly impossible distributions under suggested alternatives of clustering or a regular pattern. Further, the parametric alternatives may be too restrictive to adequately characterize the process at hand.

Consider points $Z_1, \cdots, Z_n \sim PP(\lambda)$, in which $PP(\lambda)$ is a spatial point process with intensity $\{\lambda(\mathbf{x}) : \mathbf{x} \varepsilon A\}$ for some index set A (usually a compact subset of R^2). Denote by $\Lambda(A) = \int_A \lambda(\mathbf{x}) d\mathbf{x}$. The distribution of Z_1 can be uniquely characterized by $\{\lambda\}$, which we will do since this is most natural. Complete spatial randomness (CSR) is the isotropic, stationary "null model" $\lambda(\mathbf{x}) \equiv \lambda_o$.

We have some statistic in mind,

$$R(\mathbf{Z}, \lambda)$$

which we wish to study. We fit $\hat{\lambda}$, or $\hat{\Lambda}$ in some cases. Then we draw a "bootstrap sample" $Z_1^*, \cdots, Z_n^* \sim PP(\hat{\lambda})$, and compute

$$R^* = R(\mathbf{Z}^*, \hat{\lambda}) \ .$$

Repeatedly draw N such samples and summarize the findings, say in terms of the bootstrap mean, bias, and standard deviation (*cf.* Efron (1982)).

Example 1. We might count the number of points in quadrats $\{S_1, \cdots, S_q\}$, and use Fisher's variance to mean ratio,

$$R(\mathbf{Z}, \lambda) = \sum_{i=1}^{q} (C_i - \overline{C})^2 / \overline{C}$$

in which $C_i = \#\{j : Z_j \varepsilon S_i\}$ and $\overline{C} = \sum C_i / q$. This is known to have, asymptotically, a χ^2_{q-1} distribution under complete spatial randomness (or more generally, under stationarity). However, the distribution is unknown under alternatives.

Example 2. We may focus on the k-th smallest nearest neighbor distance d_k; Ripley and Silverman (1978) showed that

$$n(n-1)\pi d_k^2 \int_A f^2(\mathbf{x}) d\mathbf{x} \rightarrow \chi^2_{2k} \quad as \quad n \rightarrow \infty \ ,$$

with $f(\mathbf{x}) = \lambda(\mathbf{x}) / \Lambda(A)$, the density of the point process. Under CSR, this simplifies nicely, but one would need to know f under alternatives.

Example 3. Another type of statistic is the empirical distribution function of nearest neighbor distances, as considered by Diggle (1983) and others. Rather than perform a number of simulations to develop an envelope "test" of the null hypothesis *edf*, one could draw a number of bootstrap samples and determine a confidence envelope for the true distribution function. Due the the undercoverage problem with such procedures (Loh, 1985) one may want to consider summary statistics such as means and von Mises functionals, again as considered by Diggle (1983).

2. Bootstrap distribution

Several possible estimates of λ present themselves. The discrete estimate is simply a point mass at each point $Z_1, \cdots, Z_n$, or

$$\hat{\Lambda}(B) = \#\{j : Z_j \varepsilon B\}, \ B \subset A \ . \qquad (2.1)$$

This implies that $\hat{\Lambda}(A) = n$. A second possibility is a parametric class, such as Neyman type A or Strauss processes, with the parameters estimated from the data. One could convolute such a process with the simple bootstrap, as described by Efron (1982).

A third class of estimates arises from nonparametric estimation via kernels or splines. In other words, we find the penalized maximum likelihood estimate of

$$\log\lambda(\mathbf{x}) = \log f(\mathbf{x}) + \log\Lambda(A)$$

subject to a penalty, say

$$J(f) = \int_A \left[f''(\mathbf{x})\right]^2 d\mathbf{x}$$

being bounded. This is simply a density estimation problem (*cf.* Silverman (1982)) of minimizing

$$\sum_{i=1}^{n} \log f(x_i) + \alpha J(f) \ .$$

Note that $\hat{\Lambda}(A) = n$, as before. Any other nonparametric density estimator (*e.g.* kernel or nearest neighbor) could also be used here.

In addition there is the question of perspective: does one view the sample size as fixed given the data, or as random? The former case is sometimes referred to as a Binomial process, and one would take bootstrap samples of fixed size m ($m = n$ is not required). For the latter,

$$N \sim Poisson(\Lambda(A)) \text{ and } N^* \sim Poisson(N) \ ,$$

noting that $\hat{\Lambda}(A) = N$ for most practical situations. The fixed case corresponds very closely to the ordinary bootstrap, while the latter introduces a new element which does not appear to have been examined. Note also that the sense in which sample size grows is different. One way to allow n to increase is to have a fixed area A, but let λ be a function of "time",

$$\lambda(\mathbf{x}, t) = t\,\lambda(\mathbf{x}) \ .$$

and to take the limit as $t \to \infty$. Thus $EN = t\,\Lambda(A)$. Another option is to increase the size of the area. These have different practical consequences, and rather different applications.

One might consider other bootstrap approaches to Example 1, where the data are summarized by quadrat. One could consider bootstrapping the counts $C_1, \cdots, C_q$. However, they are *not* i.i.d. if the process is a compound Poisson process, since some quadrats would have higher expected counts than others. Another possible bootstrap is a multinomial of the points, with probability (C_i/n) for S_i. But this is equivalent to the simple bootstrap first proposed, conditional on the number of points being n.

While we may be able to discern new properties of the classical spatial statistics tools, we are still left with the identifiability dilemna of the London busdrivers: we cannot distinguish between a doubly stochastic and a compound point process on the basis of one realization.

3. Worked Examples

We explore a few examples to see what we have gained. Consider Example 1, and a fixed number of points. Then

$$(C_1, \cdots, C_q) \sim Multinomial\,(n, p_1, \cdots, p_q) \ ,$$

with $\pi = \Lambda(S_i)/\Lambda(A)$. A bootstrap sample of size n will have distribution

$$(C_1^*, \cdots, C_q^*) \sim Multinomial\,(n, \hat{p}_1, \cdots, \hat{p}_q) \ ,$$

with $\hat{p}_i = C_i/n$. Thus trivially the bootstrap sample will provide decent estimates of the underlying distribution, as they have the same form (see Remark G of Efron (1979)). In particular we could investigate the variance to mean ratio. Similarly we could examine the distribution with a random sample size N. Here, $C_i \sim Poisson\,(\Lambda(S_i))$ and $C_i^* \sim Poisson\,(\hat{\Lambda}(S_i) = C_i)$. Again, the distributions are of the same form.

Of more interest, perhaps, is the second example. While the limiting distribution is proportional to a χ^2, we must estimate the proportionality (i.e., $\int f^2$) in all but the simplest models. This statistic depends on pairs of points, specifically the k-th smallest nearest neighbor distance. Therefore, with the simple bootstrap we end up sampling from

$$d_{ij}, \ 1 \le i \le j \le n \ ,$$

with d_{ij} this distance between Z_i and Z_j, and $d_{ii} = 0$. Our minimal distance is likely to be 0! Further, the results cited by Ripley and Silverman (1978) require a continuous density. Thus there is some argument for using a parametric or smoothed bootstrap, particularly if one is interested in small k.

The third example could be attacked using any kind of bootstrap, although some continuous form is probably preferable.

References

Diggle, P. J. (1983) *Statistical Analysis of Spatial Point Patterns*. New York: Academic Press.

Efron, B. (1982) *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: CBMS-NSF Regional Conf. Ser. in Appl. Math..

Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Statist.*, **7**, 1-26.

Loh, W.-Y. (1985) A new method for testing separate familites of hypotheses. *J. Amer. Statist. Assoc.*, **80**, 362-368.

Ripley, B. D. and Silverman, B. W. (1978) Quick tests for spatial regularity. *Biometrika*, **65**, 641-642.

Silverman, B. W. (1982) On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.*, **10**, 795-810.

---

inference:
· can we apply iid dist. results directly? (1-D, 2-D)
· what happens with random N?

simulation
Ex 2+3 with 3 data sets in Diggle
interpret as power?