Bandwidth selection in kernel estimation of relative risk

Ya-Mei Chang*
Tamkang University, New Taipei City, Taiwan – yamei628@gmail.com

The motivation of this paper is arising from analyzing tree data in Pinjar study area, Western Australia. We are interested in estimating the tree death risk which is similar to estimating disease risk in epidemiology. Many literatures have used kernel estimations to estimate disease risk and discussed different criteria to select bandwidth parameters in kernel estimations. Some literatures have recommended constraining the bandwidth parameters to be equal to reduce bias. The equal-bandwidth constraint is reasonable when the underlying densities of case and control data are nearly equal. However, our tree data is on the contrary. The distribution of living trees is dense and uniform, but the distribution of dead trees is sparse and non-uniform. Whether to constrain equal bandwidths becomes a concern in estimating the tree death risk. In this work, we compare some existing cross-validation criteria of bandwidth selection and investigate the influence of the equal-bandwidth constraint through simulations. The sizes of data sets in the simulations are similar to our tree data.

Keywords: bandwidth selection; kernel estimation; relative risk; cross-validation; point pattern process; large data set.

1. Introduction

This research is motivated by the following problem in forestry. By using high-resolution airborne digital imagery of the Pinjar study area (located on the northern urban fringe of Perth, Australia), 377,171 living trees were detected from the image for 2005 (Chang et al., 2013) and 645 tree deaths occurred in a year (Wallace et al., 2008).There is an interest to estimate the tree death risk, the ratio of intensities of dead trees and all trees, which might reflect the level of ecological stress and vulnerability.

Some point process methods and marked point process methods have been used in forestry statistics (Stoyan and Penttinen, 2000; Dralle and Rudemo, 1997; Schreuder et al., 1993), where the "points" are tree locations and the "marks" are attached to the points revealing tree characteristics such as indicators of dead trees or sizes of trees.

Estimating tree death risk is similar to estimating disease risk in epidemiology. Many literatures have used kernel estimations (Silverman, 1986) to estimate relative disease risk (Kelsall and Diggle, 1995a,b, 1998). In kernel estimations, the bandwidth parameter selection is more influential than choosing the kernel function (Epanechnikov, 1969). Azzalini et al. (1989) has proposed a likelihood cross-validation method to select bandwidths. The theoretical justification of least squares cross-validation was proven by Hardle and Marron (1985). Kelsall and Diggle (1995a) provided theoretical and empirical evidence in favor of using equal bandwidths when a ratio of two intensities is close to 1. Kelsall and Diggle (1998) compared different cross-validation methods with equal-bandwidth constraint.

In epidemiology, population is usually not uniformly distributed, and a control data is often used to reflect the spatial variation in population intensity. The sizes and the distributions of case and control data are generally quite similar. However, our tree data is on the contrary. The distribution of all trees is dense and fairly uniform, but the distribution of dead trees is sparse and non-uniform. Whether to constrain equal bandwidths of two intensities in kernel estimation is arising in estimating tree death risk. In this work, we are interested in comparing some existing cross-validation criteria of bandwidth

selection and investigating the influence of the equal-bandwidth constraint in data sets similar to our tree data.

2. Methodology

Let $\Phi_1 = \{x_i : i = 1, 2, \ldots, n_1\}$ be a point process observed on a region $W \subset R^2$. Suppose there is a certain type of event occurred at some x_i's. The marks y_i are attached to x_i such that

$$y_i = \begin{cases} 1, \text{if the event occurred at } x_i, \\ 0, \text{otherwise.} \end{cases}$$

Let $n_2 = \sum_{i=1}^{n_1} y_i$ be the number of the events occurred in the region W. The process $\{(x_i, y_i) : i = 1, 2, \ldots, n_1\}$ is a marked point process (Illian et al., 2008). The sub-process of where the events are occurred is denoted by $\Phi_2 = \{x_i \in \Phi_1 : y_i = 1\}$.

Let $\lambda_1(x)$ and $\lambda_2(x)$ be the intensity functions of Φ_1 and Φ_2 respectively, and $r(x)$ be the ratio of these two intensity functions

$$r(x) = \frac{\lambda_2(x)}{\lambda_1(x)}. \quad (0.1)$$

Given location x_i, the conditional distribution of the mark y_i is

$$P(Y_i = 1 \mid X_i = x) = \frac{\lambda_2(x)}{\lambda_1(x)} = r(x).$$

We are interested in investigating the spatial variation of $r(x)$ on W. A kernel estimation method described below is used in this work.

A fixed bandwidth kernel estimator of intensity is of the form

$$\hat{\lambda}_h(x) = \frac{\sum_{i=1}^{n_1} \kappa_h(x - x_i)}{C_h(x)},$$

where $\kappa_h(x) = h^{-2}\kappa(h^{-1}x)$, κ is a symmetric kernel function and

$$C_h(x) = \int_W \kappa_h(x - u)du$$

is an edge-correction function (Diggle, 1985). For more discussions of kernel estimation, see Silverman (1986).

A kernel estimator for $r(x)$ in (1.1) is thus

$$\hat{r}_{h_1,h_2}(x) = \frac{\hat{\lambda}_{2,h_2}(x)}{\hat{\lambda}_{1,h_1}(x)} = \frac{\sum_{i=1}^{n_1} \kappa_{h_2}(x - x_i) y_i / C_{h_2}(x)}{\sum_{i=1}^{n_1} \kappa_{h_1}(x - x_i) / C_{h_1}(x)} \quad (0.2)$$

where $\hat{\lambda}_{1,h_1}(x)$ and $\hat{\lambda}_{2,h_2}(x)$ are the kernel estimators of intensity functions $\lambda_1(x)$ and $\lambda_2(x)$ respectively, with different bandwidth parameters h_1 and h_2. In this work, we use the bivariate Gaussian kernel as the kernel function, $\kappa(x) = (2\pi)^{-1} \exp(-(1/2) \| x \|^2)$.

Kelsall and Diggle (1995a) has derived the approximation for the mean integrated square error of a log density ratio. Apart from an additive constant, the density ratio in Kelsall and Diggle (1995a) and our intensity ratio $r(x)$ are equal. Kelsall and Diggle (1995a) recommended choosing the two bandwidths h_1 and h_2 jointly and constraining the bandwidths to be equal to reduce bias.

Kelsall and Diggle (1998) introduced some leave-one-out cross-validation methods for selecting the bandwidth parameters. Those leave-one-out cross-validation methods are intuitively attractive and easily implemented. Kelsall and Diggle (1998) have taken the advice of Kelsall and Diggle (1995a) to constrain $h_1 = h_2$ for reducing the bias, which is very reasonable when two intensities are nearly equal. However, our tree data is not of the same situation. The numbers and the spatial variations of all trees and dead trees are hugely different. In this work, we adopt some cross-validation methods used by Kelsall and Diggle (1998) but release the constraint $h_1 = h_2$. The adopted cross-validation methods are described in the following.

Let $\hat{r}^{-k}_{h_1,h_2}(x)$ be the kernel estimate (1.2) of $r(x)$ without the k-th observation. The cross-validation methods adopted from Kelsall and Diggle (1998) are to minimize the criteria as follows:
Likelihood cross-validation

$$CV_1 = \left\{ \prod_{k=1}^{n_1} \left[\hat{r}^{-k}_{h_1,h_2}(x_k) \right]^{y_k} \left[1 - \hat{r}^{-k}_{h_1,h_2}(x_k) \right]^{1-y_k} \right\}^{1/n_1} \qquad (0.3)$$

Least squares cross-validation

$$CV_2 = n_1^{-1} \sum_{k=1}^{n_1} \left[y_k - \hat{r}^{-k}_{h_1,h_2}(x_k) \right]^2 \qquad (0.4)$$

Weighted least squares cross-validation}

$$CV_3 = n_1^{-1} \sum_{k=1}^{n_1} \frac{\left[y_k - \hat{r}^{-k}_{h_1,h_2}(x_k) \right]^2}{\hat{r}^{-k}_{h_{01},h0_2}(x_k) \left[1 - \hat{r}^{-k}_{h_{01},h0_2}(x_k) \right]} \qquad (0.5)$$

where h_{01} and h_{02} are obtained by the least squares cross-validation (1.4).

When the size of the point process is large, computing the kernel regression estimation is very computationally intensive. To solve this problem, Baddeley et al. (2000) has developed an efficient algorithm by using a fast Fourier transform (FFT). The kernel regression estimation $\hat{r}^{-k}_{h_1,h_2}(x)$ can be computed through functions in *spatstat* (Baddeley and Turner, 2005), an R package for point process. Thus the cross-validation criteria $CV_1 - CV_3$ can be easily obtained. The performance of the three cross-validation methods and the effect of the constraint $h_1 = h_2$ are investigated through simulation in this research.

References

Azzalini, A., Bowman, A. W., and Hardle, W. (1889).On the use of nonparametric regression for model checking. *Biometrika*,76(1), 1-11.

Baddeley, A., Moller, J., and Waagepetersen, R. (2000).Non-and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*,54(3), 329-350.

Baddeley, A., and Turner, T. R. (2005).Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6), 1-42.

Diggle, P. (1985). A kernel method for smoothing point process data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*,34(2), 138-147.

Chang, Y.-M., Baddeley, A., Wallace, J. and Canci, M.. Spatial statistical analysis of tree deaths using airborne digital imagery (2013). *International Journal of Applied Earth Observation and Geoinformation*, 21, 418-426.

Dralle, K., and Rudemo, M. (1997). Automatic estimation of individual tree positions from aerial photos. *Canadian Journal of Forest Research*, 27(11),1728-1736.

Epanechnikov, V. A. (1969). Nonparametric estimates of a multivariate probability density. *Theory Probability Application*,14, 153-158.

Hardle, W., and Marron, J. S. (1985). Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression. *Biometrika*, 72(2), 481-484.

Illian, J., Penttinen, A., and Stoyan, H. (2008). *Statistical analysis and modelling of spatial point patterns*, Wiley-Interscience.

Kelsall, J. E., and Diggle, P. J. (1995a). Kernel estimation of relative risk. *Bernoulli*,1, 3-16.

Kelsall, J. E., and Diggle, P. J. (1995b). Non-parametric estimation of spatial variation in relative risk. *Statistics in Medicine*,14,2335-2342.

Kelsall, J. E., and Diggle, P. J. (1998). Spatial variation in risk of disease: a nonparametric binary regression approach. *Applied Statistics*,47(4), 559-573.

Kingman, J. F. C. (1993). *Poisson processes*, Oxford University Press, USA.

Schreuder, H. T., Wood, G. B., and Gregoire, T. G. (1993). *Sampling methods for multiresource forest inventory*, John Wiley & Sons Inc.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, London: Chapman & Hall.

Stoyan, D., and Penttinen, A. (2000). Recent applications of point process methods in forestry statistics. *Statistical Science*,15(1), 61-78.

Wallace, J. F., Cance, M., Wu, X., and Baddeley, A. J. (2008). Monitoring native vegetation on an urban groundwater supply mound using airborne digital imagery. *Spatial Science*, 53(1), 63-74.