

NON PARAMETRIC ANALYSIS OF SPATIAL DATA

Analisi non Parametrica per Dati Territoriali

Rosa Capobianco

Dipartimento di Studi Giuridici, Politici e Sociali “Persona, Mercato e Istituzioni” -
Sezione di Statistica

Facoltà di Economia – Università degli Studi del Sannio

Via Calandra – 82100 Benevento

capobianco@unisannio.it

Riassunto: Data una realizzazione finita di un processo puntuale con marchi qualitativi, in questo articolo si propone una generalizzazione dello stimatore di Kendall per valutare l’auto-associazione esistente tra i marchi. Tale approccio, oltre a consentire un’analisi esplorativa di un pattern spaziale con marchi sia qualitativi che quantitativi, presenta interessanti possibilità applicative: a) è possibile fare inferenza sul processo stocastico generatore dei marchi dal momento che, sotto opportune ipotesi, è nota la distribuzione asintotica dello stimatore; b) dati k marchi per ogni punto, è possibile considerare l’estensione dello stimatore al caso multivariato. Le performance dello stimatore sono state valutate attraverso l’applicazione a due set di dati simulati.

Keywords: Auto-association, qualitative data, marked point processes, Kendall estimator.

1. Introduction

Spatial data observed on a plane can be thought as a finite realization of an isotropic and stationary marked point process (m.p.p.), defined through the set of values $\Psi = \{p_i, m_i\}$, where p_i is the location of the point on $\mathbb{R}^2$ while m_i is the corresponding mark, i.e. one of the point’s characteristics (Stoyan and Stoyan, 1994). Let us consider, for example, a forest: trees are the points localised on the plane while possible marks can be the trunk diameter or the colour of the leaves. In both cases it is interesting to know whether there is a relationship among the marks and to investigate the nature of the association, i.e. if marks are independent, aggregated or inhibited.

Many estimators have been defined to analyse the correlation among quantitative marks (Stoyan and Stoyan, 1994; Capobianco and Renshaw, 1998), while not much has been done for the qualitative case. In this paper we propose a non parametric approach to estimate the association among pairs of qualitative marks. Our proposal is based on a generalization to spatial data of Kendall estimator (Landenna and Marasini, 1990), defined to evaluate the association between two qualitative variables, observed for the same statistical unit, whose modalities can be ranked. Kendall estimator varies between $[-1,1]$: if it assumes values greater than zero (less than zero) the association is positive (negative), if it is equal to zero there is no association between the variables. Here we have only one variable (the mark) observed on different locations of the plane, pairs of

marks are determined on the basis of the Euclidean distance r among the corresponding points. We are interested to verify whether, as r varies, there is “auto-association” among pairs of marks.

The outline of the paper is the following. In section 2 we illustrate the generalization of Kendall estimator to spatial data, while in section 3 the estimator is applied to simulated point patterns with qualitative marks.

2. Kendall estimator

Let $\{p_i, m_i\}$, for $i=1, \dots, n$, be a finite realization of an isotropic and stationary m.p.p., observed on a torus. For a given Euclidean distance r , we are interested to estimate the auto-association among the pairs of marks (m_i, m_j) and (m_k, m_l) , for $i, j, k, l=1, \dots, n$, whose correspondent points are r distance apart, i.e. $r - e < \|p_i - p_j\| < r + e$ and $r - e < \|p_k - p_l\| < r + e$, where e is an arbitrary positive constant.

Kendall estimator is defined as follows

$$t(r) = \frac{1}{\{[n_r(n_r - 1) - B_1][n_r(n_r - 1) - B_2]\}^{1/2}} \times \sum_{i=1}^n \sum_{k=1}^n \sum_{j=i+1}^n \sum_{l=k+1}^n I_r f[(M_i - M_k)(M_j - M_l)] \quad (1)$$

for $r > 0$, where n_r is the number of pairs of points r distance apart; B_1 and B_2 are two constants equal to

$$B_1 = \sum_{s=1}^q {}_1b_s ({}_1b_s - 1) \quad \text{and} \quad B_2 = \sum_{h=1}^p {}_2b_h ({}_2b_h - 1)$$

where ${}_1b_s$, for $s=1, \dots, q$, and ${}_2b_h$, for $h=1, \dots, p$, count the number of times that (M_s, M_h) are repeated among the pairs of marks, respectively; I_r is the indicator function such that

$$I_r = \begin{cases} 1 & \text{if } r - e < \|p_i - p_j\| < r + e \quad \text{and} \quad r - e < \|p_k - p_l\| < r + e \\ 0 & \text{otherwise} \end{cases}$$

while

$$f[(M_i - M_k)(M_j - M_l)] = \begin{cases} 1 & \text{if } (M_i - M_k)(M_j - M_l) > 0 \\ -1 & \text{if } (M_i - M_k)(M_j - M_l) < 0 \\ 0 & \text{if } (M_i - M_k)(M_j - M_l) = 0 \end{cases}$$

$t(r)$ can vary in the range $[-1,1]$. If marks are independent then $t(r)=0$, if marks are aggregated [inhibited] $t(r)$ assumes positive [negative] values in the range $(0,1)$ $[(-1,0)]$; the auto-association is perfect when $t(r) = \pm 1$.

When we have defined the estimator we have assumed that the m.p.p. that generated the data set is isotropic. This assumption can be easily removed and the estimator can be expressed in terms of the distances between x - and y -coordinates, as in Capobianco and Renshaw (1998). We have also assumed that the realization has been observed on a torus, so that the edge correction factor has not to be introduced in the estimator. The extension of the estimator to the case when the observation window is rectangular should be straightforward.

The non parametric approach, proposed to carry out explorative analysis of qualitative marks, is also appealing because it gives the opportunity to do inference on the unknown generating process in order to check whether the association among the marks is significative. In fact it has been proved (Landenna and Marasini, 1990) that as the sample size increases, Kendall estimator is asymptotically distributed as a normal random variable. To this proposal other estimators have been defined in order to study the association among qualitative characters, like Spearman correlation coefficient or Gini cograduation index. Their asymptotic distribution too is a normal random variable, however the rate of convergence is slower compared to that of Kendall estimator.

Another important property of (1) is the possibility to consider for each point a vector of marks and verify whether there is auto-association or not. The extension to the multivariate case is interesting because the only attempt to consider at the same time several variables has been carried out by Goulard *et al.* (1995). They proposed an estimator to study the correlation between two different marks (m_i, l_k) , observed for the pair of points (p_i, p_k) r distance apart.

The last consideration concerns the possibility to apply the estimator even when quantitative marks are observed. Obviously there is a lack of information given that the ranks, and not the values of the marks, are used in the estimation. However the advantage of this approach is that a non linear association is estimated, which is often the case for the relationship among the marks.

3. A simulation study

In order to check the performance of the estimator $t(r)$, two point patterns with independent and aggregated qualitative marks have been generated; the sample size is $n=300$. In both cases the point process is Poisson type, whose x - and y -coordinates have been generated from a Uniform random variable defined on the range $[0,1]$.

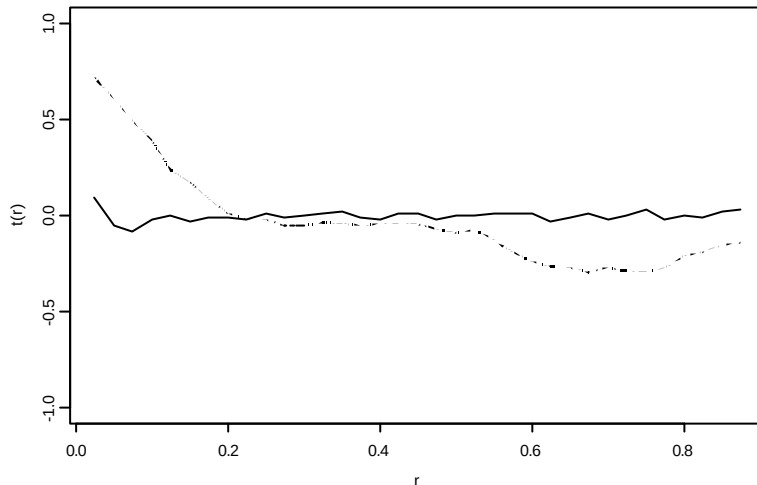
Independent marks have been generated from a uniform random variable. In particular the values A, B, C, D and E have been attributed to marks in the ranges $[0,0.20]$, $(0.20,0.40]$, $(0.40,0.60]$, $(0.60,0.80]$ and $(0.80,1.00]$, respectively. In this way we have five qualitative marks to rank in increasing order.

In order to generate aggregated marks, each point of the data set has been considered as the centre of a circle with ray equal to 0.02; the associated mark is the number of points in the circle (the centre is not included), then we proceeded as for the independent case.

Figure 1 shows the results of the application of the estimators to the two data sets. The continuous line refers to the case of independent marks. As expected, the estimator

assumes values close to zero as the Euclidean distance varies in the range (0,1). The dot line corresponds to the case of aggregated marks. For small values of r , the positive association among the marks is quite strong, the peak in fact is about 0.7, then it tends to zero as r tends to 0.5. For $r > 0.5$ $t(r)$ assumes negative values; a possible explanation of the negative association could be the presence of a mild inhibition among big marks.

Figure 1: Kendall estimator for simulated m.p.p.



References

- Capobianco, R., Renshaw, E. (1998) The autocovariance function for marked point processes: a comparison between two different approaches, *Biometrical Journal*, 40, 431-446.
- Goulard, M., Pagès, L., Cabanettes, A. (1995) Marked point processes: using correlation functions to explore data set, *Biometrical Journal*, 37, 837-853.
- Landenna, G., Marasini, D. (1990) *Metodi Statistici Non Parametrici*, Il Mulino, Bologna.
- Stoyan, D., Stoyan, H. (1994) *Fractals, Random Shapes and Point Fields*, Wiley, New York.