Principal Component Analysis for Spatial Point Process Data

Janine Illian¹, Erica Benson², Harry Staines¹ and John Crawford¹

¹ SIMBIOS, University of Abertay Dundee
² School of Contemporary Sciences, University of Abertay Dundee

1 Introduction

Spatial point processes are complex stochastic models. Their complexity increases with the number of types in a multi-type point pattern due to the increasing number of potential inter-and intra-type interactions and thus the increasing number of parameters complicating simulation and estimation. This calls for a statistical method which reduces the dimensionality of the dataset, i.e. groups the point processes. But even though multivariate statistical methods such as principal component analysis (PCA) and factor analysis have been around for a long time and are a popular tool for dimension reduction in many contexts, no similar method exists for spatial point pattern data.

This talk will illustrate how functional data analysis tools can be applied to the second order statistics of multi-type point processes, in particular to (inhomogeneous) L-functions, to derive a PCA method for spatial point pattern data. We will briefly introduce functional data analysis, in particular functional PCA and discuss how these methods can be extended to work in the context of spatial point patterns. A detailed simulation study was used to validate the approach and showed promising results, giving rise to some general guidelines and results. We conclude with an application, investigating a multi-type spatial point pattern formed by a natural plant community in the heathlands of Western Australia, comprising 6378 plants from 68 species (see Illian (200x)).

Key words: spatial point processes, functional data analysis, inhomogeneous L-function, multi-type spatial point patterns, spline smoothing, principal component analysis

2 Methods

2.1 Functional data analysis

For a detailed introduction to functional data analysis see Ramsay and Silverman (1997). Functional data analysis deals with functional data, i.e. observations are functions and these are interpreted as single entities rather than as consecutive measurements. Generally speaking, the record of a functional observation \( x \) consists of \( n \) pairs \((t_j, y_j)\), where \( y_j \) is an observation of \( x(t_j) \) at time \( t_j \). Since the functions are usually observed at a countable number of values of \( t \) only, interpolation or smoothing techniques have to be applied to yield a functional representation of the data. For a functional PCA context we consider function values \( x_i(t) \) and define
\[ f_i = \int \beta(t)x_i(t) \, dt, \] where \( \beta(t) \) is a weight function \( \beta(t) \). We want to maximise \( N^{-1} \sum f_i^2 \) under the constraint \( \|w_1\|^2 = \int w_1(t)^2 \, dt = 1 \) and get an eigen-equation
\[
\int v(t, s)w(s) \, ds = \lambda w(t)
\] (1)
with variance-covariance function \( v(t, s) = N^{-1} \sum x_i(t)x_i(s) \). Subsequent steps in the analysis will then mainly examine the scores \( f_{ik} \) for each of the curves on the first \( p \) components.

### 2.2 Spatial point processes

Spatial point processes describe the locations of objects in space. A spatial point pattern \( x \) is a realisation of a spatial point process \( X \) which can be defined as follows. Let \( A \) be some metric space and assume this space to be Polish, i.e. complete and separable. For each bounded Borel set \( B \subset A \), let \( \phi(B) \) be the number of events in \( B \). Thus we can identify a point configuration with a counting measure \( \phi \) on Borel sets on \( A \). Let \( N \) be the set of all such measures and \((\Omega, \mathcal{A}, P)\) some probability space. On \( N \) define \( \mathcal{N} \) as the smallest \( \sigma \)-algebra generated by sets of the form \( \{ \phi \in N : \phi(B) = n \}, \forall n \in \{0, 1, 2, \ldots\} \). A spatial point process \( X \subset A \) can then be regarded as a mapping from \((\Omega, \mathcal{A})\) into \((N, \mathcal{N})\).

Let \( Z \) be a simple point process in \( \mathbb{R} \) as defined above. Attaching a random mark \( m_\zeta \in \mathcal{M} \) with \( \zeta \) some mark space to each point \( \zeta \) yields a marked point processes \( X = \{ (\zeta, m_\zeta) : \zeta \in Z \} \). If \( \mathcal{M} = \{1, \ldots, k\}, X \) is a multi-type point process with \( k \) different types of points.

For a point process \( X \) the intensity measure is given by
\[
\Lambda(B) = E[\phi(B)], \quad \text{for any Borel set } B.
\]

Often a density function, the so-called intensity function, \( \lambda : A \rightarrow \mathbb{R}^+ \) exists, such that
\[
\Lambda(B) = \int_B \lambda(x) \, dx.
\]

For a stationary process with \( \lambda(x) = \lambda_0 \) and with finite intensity define a second order summary statistic, commonly called Ripley’s \( K \)-function (Ripley, 1976) as
\[
K(r) = E \sum_{\zeta \in X, \xi \in X} 1[||\xi - \zeta|| \leq r]/(\lambda^2(W)).
\]

Often, the variance stabilising \( L \)-function is used instead, where \( L(r) = \sqrt{K(r)} \). The \( K \)-function can be used to distinguish between clustered, random and regular patterns. Note that for a homogeneous Poisson process with complete spatial randomness \( K(r) = \pi r^2 \) and \( L(r) = r \). If \( L(r) > r \) we have a clustered process and if \( L(r) < r \) a regular process.

Baddeley, Møller andWaagepetersen (2000) introduce an inhomogeneous version of the \( K \)-function taking the local intensities into account. Here \( \lambda^2(W) \) is replaced by \( \lambda(\xi)\lambda(\zeta) \) to yield
\[
K_{\text{inhom}}(r) = E \sum_{\zeta \in X, \xi \in X} 1[||\xi - \zeta|| \leq r]/(\lambda(\xi)\lambda(\zeta)).
\]
2.3 Functional principal component analysis of L-functions

Consider a multitype point process $X = \{(\zeta, m_\zeta) : \zeta \in Z\}$ with $m_\zeta \in \mathcal{M}$ and $\mathcal{M} = \{1, \ldots, k\}$ as defined above, where no other additional data marks are available. We use the second order summary statistics to characterise the spatial behaviour of the individual subprocesses $X_i \subset X$ with $X_i = \{(\zeta, m_\zeta) : \zeta \in Z$ and $m_\zeta = i\}$ and $i = 1, \ldots, k$ and suggest a functional principal component analysis on the smoothed L-functions to group the point processes by their spatial behaviour.

In the homogeneous case we estimate the L-functions using the following estimator (Ripley, 1976)

$$\hat{K} = n^{-2}|A| \sum_{\zeta \neq \xi} w_{\zeta, \xi}^{-1} I_d(d_{\zeta, \xi})$$

with $n$ the number of points in region $A \in \mathbb{R}$ with area $|A|$, $d_{\zeta, \xi}$ the distance between point $\zeta$ and $\xi$ and $w_{\zeta, \xi}$ an edge correction factor- the proportion of the circle with centre $\zeta$ passing through $\xi$ which lies in $A$. In the inhomogeneous case, we use the estimator introduced in Baddeley et al. (2000):

$$\hat{K}_{\text{inom}}(r) = 2 \sum_{\xi, \zeta \in X_W} 1[||\xi - \zeta|| < r]/(|W_\xi \cap W_\zeta|/|\bar{\lambda}(\xi)\bar{\lambda}(\zeta)|),$$

where $\bar{\lambda}$ is the intensity estimator

$$\bar{\lambda}(s) = \sum_{\xi \in X_W \setminus \zeta} \kappa(\xi - \zeta)/c(\zeta)ds, \ \zeta \in W,$$

$c(\zeta) = \int_W \kappa(\xi - \zeta)d\xi$ is an edge correction factor (Diggle, 1985) and $\kappa$ a kernel function.

We smooth the estimated L-functions using B-splines (see Green and Silverman, 1994), i.e. splines with compact support, as they are capable of picking up local features, and subsequently perform a functional PCA on the smoothed functions. We then group the point processes on the basis of their scores on the principal components and plot the scores on the first and second PC. Cluster analysis methods are applied to these to detect clusters of similar L-functions and hence groups of point processes with similar spatial behaviour in the data set.

L-functions for regular patterns can take values in $[r, 0]$ whereas L-functions for clustered pattern can take values in $[r, \infty)$. Hence, if we want to distinguish between clustered, random and regular patterns, the difference between the L-function for a clustered pattern and a random pattern tends to be larger than the difference between the L-function for a regular pattern and a random pattern. This problem is similar to measuring different variables on different scales in a standard PCA context, where the usual solution is to perform a PCA on the correlation matrix rather than on the covariance matrix. Thus a FPCA based on the correlation matrix is applied here instead, i.e. equation (1) now becomes

$$\int v^*(t, s)w(s)\,ds = \lambda w(t)$$

where $v^*$ is the correlation function $v(t, s) = N^{-1} \sum_{i=1}^{N} x_i^*(t)x_i^*(s)$, i.e. the covariance function of a standardised data matrix $x^*$.

3 Conclusions

An extended simulation study examined the capability of the approach described in this paper to separate groups of simulated homogeneous (and also inhomogeneous) point patterns with
different a spatial behaviour. These groups could clearly be separated and even if the differences between the patterns were very small, there was only a small number of misclassifications.

Nevertheless, as predicted the method had some problems when regular, clustered and random patterns were analysed. The suggested modification which uses the correlation matrix rather than the covariance matrix did indeed yield better results. But, still some improvement is needed here and currently various methods to rectify this problem are being considered.

When the inhomogeneous L-function (see equation (2)) was applied the method was strongly influenced by the bias in the estimation for larger distances. Restriction to smaller distances only improved the results. Since the method apparently strongly depends on the summary statistics chosen as well as on the quality of the estimator, other second order summary statistics or even summary statistics of higher order might have to be taken into account.

References


Illian, J.B. Multivariate methods for spatial point processes (in preparation)
