Computational Statistics and Data Analysis 66 (2013) 19–31

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Inference for variograms[☆]

Adrian W. Bowman [a,*], Rosa M. Crujeiras [b]

[a] School of Mathematics & Statistics, The University of Glasgow, UK

[b] Department of Statistics and OR, University of Santiago de Compostela, Spain

ARTICLE INFO

Article history:
Received 9 December 2011
Received in revised form 19 February 2013
Accepted 21 February 2013
Available online 26 March 2013

Keywords:
Isotropy
Nonparametric smoothing
Standard error
Stationarity
Variogram

ABSTRACT

The empirical variogram is a standard tool in the investigation and modelling of spatial covariance. However, its properties can be difficult to identify and exploit in the context of exploring the characteristics of individual datasets. This is particularly true when seeking to move beyond description towards inferential statements about the structure of the spatial covariance which may be present. A robust form of empirical variogram based on a fourth-root transformation is used. This takes advantage of the normal approximation which gives an excellent description of the variation exhibited on this scale. Calculations of mean, variance and covariance of the binned empirical variogram then allow useful computations such as confidence intervals to be added to the underlying estimator. The comparison of variograms for different datasets provides an illustration of this. The suitability of simplifying assumptions such as isotropy and stationarity can then also be investigated through the construction of appropriate test statistics and the distributional calculations required in the associated *p*-values can be performed through quadratic form methods. Examples of the use of these methods in assessing the form of spatial covariance present in datasets are shown, both through hypothesis tests and in graphical form. A simulation study explores the properties of the tests while pollution data on mosses in Galicia (north-west Spain) are used to provide a real data illustration.

1. Introduction

The variogram is a very well established tool in the analysis of spatial data as it describes the nature of spatial covariance in a very convenient manner. Cressie (1993) gives a broad overview of spatial methods in general, including a detailed treatment of the variogram, while Diggle and Ribeiro (2007) approach the topic in the context of regression and other forms of spatial modelling. From this latter approach, spatial covariance can be viewed as a nuisance parameter which is not of primary interest but which needs to be accommodated appropriately in a model in order to avoid compromising the estimation and interpretation of regression and other parameters. This is a very appealing perspective. However, it remains clear that there is considerable motivation for assessing the characteristics of spatial covariance in a dataset or model and the variogram offers a natural route to doing so. The abundance of papers in both the methodological and the applied literature, with Marchant and Lark (2007) and Emery and Ortiz (2007) as indicative examples, testifies to the popularity and value of the variogram.

Suppose that the random variable $Z(\mathbf{s})$ denotes the value of a process at spatial location $\mathbf{s}$ within a region $\mathcal{D}$. A simple model of spatial variation assumes that the process is Gaussian and intrinsically stationary with mean zero.

☆ Supplementary material is available online.

* Corresponding author. Tel.: +44 1413304046.
E-mail addresses: adrian.bowman@glasgow.ac.uk (A.W. Bowman), rosa.crujeiras@usc.es (R.M. Crujeiras).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.02.027

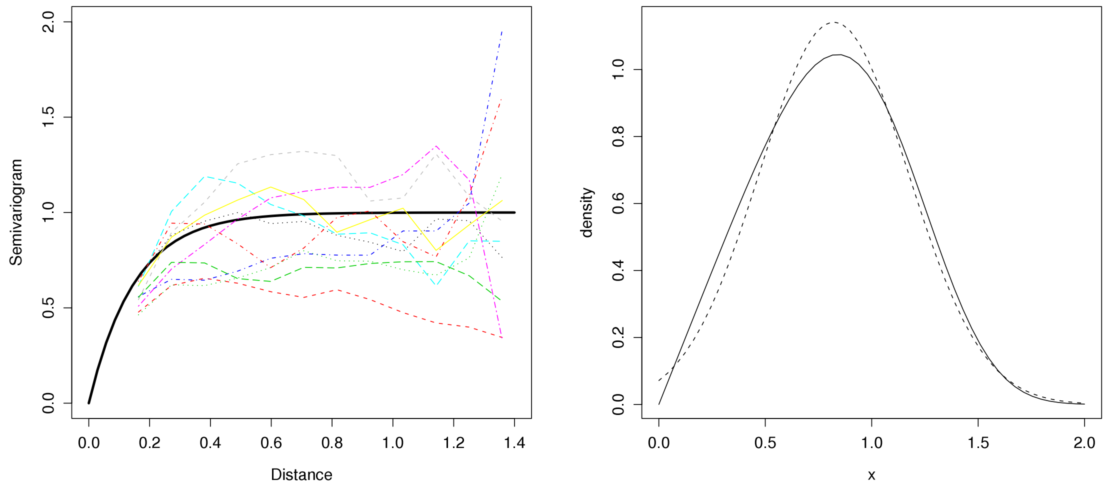

Fig. 1. The left hand panel shows ten empirical semivariograms (dashed lines), each constructed from data simulated from a Gaussian spatial process with exponential semivariogram (full line). One hundred data points were arranged in a regular grid and the sill and range parameters were set to 1 and 0.15 respectively. The right hand panel shows the density function of the fourth-root transformation of a χ^2_1 random variable (full line) and the density function of a normal random variable with the same mean and standard deviation (dashed line).

Spatial covariance can then be characterised through the variogram $2\gamma(\mathbf{h}, \mathbf{s})$ as

$$2\,\gamma(\mathbf{h}, \mathbf{s}) = \mathrm{var}\,\{Z(\mathbf{s}) - Z(\mathbf{s} + \mathbf{h})\}\,.$$

The function γ is referred to as the semivariogram. When the process is weakly stationary, with mean and covariance independent of location $\mathbf{s}$, this reduces to the simpler form

$$2\,\gamma(\mathbf{h}) = \mathrm{var}\,\{Z(\mathbf{s}) - Z(\mathbf{s} + \mathbf{h})\}\,.$$

An assumption of isotropy, where covariance depends on the spatial separation vector $\mathbf{h}$ only through its size $h = \|\mathbf{h}\|$, allows the further simplification

$$2\,\gamma(h) = \mathrm{var}\,\{Z(\mathbf{s}) - Z(\mathbf{s} + h\mathbf{e})\}\,,$$

for any unit vector $\mathbf{e}$. When data are available in the form of observations $Z(\mathbf{s}_i) = z_i$ at spatial locations $\{\mathbf{s}_i,\ i = 1, \ldots, n\}$ with inter-point distances denoted by $h_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|$, the variogram is usually estimated by its empirical version based on binning the data, as

$$2\,\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{(i,j)\in N(h)} (z_i - z_j)^2, \tag{1}$$

where $N(h) = \{(i, j) : h_{ij} \in b(h)\}$, $b(h)$ denotes an interval or 'bin' containing h and $|N(h)|$ denotes the cardinality of $N(h)$.

The large variability which the empirical variogram can exhibit is not always fully appreciated. The left hand panel of Fig. 1 shows ten simulations of 100 observations from a stationary and isotropic Gaussian spatial process in $[0, 1]^2$ with exponential semivariogram $\gamma(h) = \sigma^2\{1 - \exp(-h/\phi)\}$, with the sill ($\sigma^2$) and range ($\phi$) parameters set to 1 and 0.15 respectively. We expect large variability at large distances, because of the sparsity of data; however, considerable variability is also displayed at smaller distances. There is also a strong degree of correlation across the values in the estimate, giving in some cases a false impression of a well estimated function from a single realisation.

The properties of the sample variogram are not easy to evaluate. The finite sampling distribution of $2\hat{\gamma}(h)$ has been tabulated by Davis and Borgman (1979) for univariate data using Fourier inversion methods, although the procedure could in principle be applied in higher dimensions. Asymptotic analysis has been carried out by Davis and Borgman (1982) and Cressie (1985), who established the sampling distribution for the empirical variogram and a robust version based on the fourth-root transform, introduced by Cressie and Hawkins (1980). Pardo-Igúzquiza and Dowd (2001) and Marchant and Lark (2004) discuss the computation of the covariance matrix of the empirical variogram, in the original scale, as a potential route to confidence intervals based on the asymptotically normal distribution derived by Davis and Borgman (1979). However, even in the Gaussian scenario, skewness is an issue as demonstrated by Baczkowski and Mardia (1987) who propose a log-normal approximation for the distribution of the sample variogram. Cressie (1993, Section 2.4.2) shows that the distribution can be written as a weighted sum of independent χ^2_1 random variables.

These issues have hampered the use of the sample variogram as an inferential tool. In the case of independent spatial data, the distributional properties of the sample variogram can be evaluated and exploited to form the basis of a test for the presence of spatial correlation, as described by Diblasi and Bowman (2001). Similarly, a test of goodness-of-fit assuming