ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Flexible and efficient estimating equations for variogram estimation

Ying Sun [a], Xiaohui Chang [b,*], Yongtao Guan [c]

[a] CEMSE Division, King Abdullah University of Science and Technology, Saudi Arabia
[b] College of Business, Oregon State University, USA
[c] Department of Management Science, University of Miami, USA

ARTICLE INFO

Article history:
Received 23 January 2017
Received in revised form 20 December 2017
Accepted 30 December 2017
Available online xxxx

Keywords:
Estimating equations
Lag effect
Low rank approximation
Statistical efficiency

ABSTRACT

Variogram estimation plays a vastly important role in spatial modeling. Different methods for variogram estimation can be largely classified into least squares methods and likelihood based methods. A general framework to estimate the variogram through a set of estimating equations is proposed. This approach serves as an alternative approach to likelihood based methods and includes commonly used least squares approaches as its special cases. The proposed method is highly efficient as a low dimensional representation of the weight matrix is employed. The statistical efficiency of various estimators is explored and the lag effect is examined. An application to a hydrology data set is also presented.

1. Introduction

The variogram is a fundamental tool in the modeling of spatial processes. For a spatial stochastic process $Z(\mathbf{s})$ defined on $\mathcal{D} \subset \mathbb{R}^2$, if it has a constant mean: $\mathrm{E}(Z(\mathbf{s})) = \mu$ for all $\mathbf{s} \in \mathcal{D}$, and its variogram $2\gamma(\mathbf{s}_1 - \mathbf{s}_2) \equiv \mathrm{Var}(Z(\mathbf{s}_1) - Z(\mathbf{s}_2))$ for $\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}$ only depends on the displacement between the locations, $\mathbf{s}_1 - \mathbf{s}_2$, then Z is *intrinsically stationary*. A process is *isotropic* if its variogram is dependent on the absolute distance, not the relative direction, between locations, i.e., $2\gamma(\mathbf{s}_1 - \mathbf{s}_2) = 2\gamma(\|\mathbf{s}_1 - \mathbf{s}_2\|)$ where $\|\cdot\|$ denotes the Euclidean distance, and *geometric anisotropic* if $Z(\mathbf{Hs})$ is isotropic for some invertible matrix $\mathbf{H} \neq \mathbf{I}$. The variogram is widely used to quantify the spatial variability of many physical processes and is essential for obtaining accurate spatial predictions. Some examples are from mining (Matheron, 1963; Journel and Huijbregts, 1978), forestry (Moeur, 1993), hydrogeology (Kitanidis, 1997; Yu et al., 2003), soil science (Heuvelink and Webster, 2001; McGrath et al., 2004), epidemiology (Kleinschmidt et al., 2000; Wong et al., 2004), meteorology (Haylock et al., 2008) and many others (Isaaks and Srivastava, 1989; Cressie, 1993; Webster and Oliver, 2007). In practice, the variogram is usually unknown and needs to be estimated from the data. To model the spatial relationships, especially for spatial predictions, it is crucial to obtain variogram estimates with high statistical efficiency and also low variance.

A graphical assessment of the variogram estimate can be achieved using a variogram cloud, that is a plot of squared differences of observations versus pair-wise distances. The variogram cloud has the advantage of displaying the individual contributions to the overall variogram from all pair-wise distances (Müller, 1999). Compared to a variogram cloud, the empirical variogram is widely used in variogram estimation because it is more robust to modeling assumptions and provides more information in identifying an appropriate model form for the variogram. The classical empirical variogram estimator

* Correspondence to: College of Business, Oregon State University, Corvallis, OR 97331, USA.
E-mail addresses: ying.sun@kaust.edu.sa (Y. Sun), xiaohui.chang@oregonstate.edu (X. Chang), yguan@bus.miami.edu (Y. Guan).

https://doi.org/10.1016/j.csda.2017.12.006
0167-9473/

is based on the method of moments (Matheron, 1962),

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|}\sum_{N(\mathbf{h})}(Z(\mathbf{s}_1) - Z(\mathbf{s}_2))^2,$$

where $N(\mathbf{h}) = \{(\mathbf{s}_1, \mathbf{s}_2) \in \mathcal{D} \subset \mathbb{R}^2 : \mathbf{s}_1 - \mathbf{s}_2 = \mathbf{h}\}$ with cardinality $|N(\mathbf{h})|$. To eliminate the constraint of considering pairs of sites with exactly $\mathbf{h}$ lag apart, some modifications of the classical empirical variogram are proposed to use a tolerance region or group distances into bins (Cressie, 1993). There are also other variations of the empirical variogram to take care of the susceptibility to outliers through robust variogram estimators based on $|Z(\mathbf{s}_1) - Z(\mathbf{s}_2)|^{1/2}$ (Cressie and Hawkins, 1980) or other highly robust scale estimators (Genton, 1998a).

Fitting the empirical variogram directly using nonparametric methods may not necessarily produce a valid variogram. In practice, variogram estimation is usually done by first assuming a parametric form that is known to be conditionally negative definite and then estimating the parameters to ensure that the estimated variogram is close to the empirical variogram of the data. Refer to Cressie (1993) for a list of popular choices of variogram models.

Least squares and likelihood based methods are the two approaches that are commonly adopted for fitting variogram models to spatial data. The least squares method fits a parametric model by minimizing a quadratic distance measure between the empirical variogram $2\hat{\gamma}(\mathbf{h})$ (or some other nonparametric variogram estimators) and the parametric model $2\gamma(\mathbf{h}, \boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \mathbb{R}^p$ denotes p unknown parameters. More specifically, for a variogram estimated at K discrete lags, the parameter vector $\boldsymbol{\theta}$ is estimated by minimizing

$$\{2\hat{\boldsymbol{\gamma}} - 2\boldsymbol{\gamma}(\boldsymbol{\theta})\}' V(\boldsymbol{\theta})^{-1} \{2\hat{\boldsymbol{\gamma}} - 2\boldsymbol{\gamma}(\boldsymbol{\theta})\},$$

where $2\hat{\boldsymbol{\gamma}}$ is a $K \times 1$ vector of the empirical variogram at K lags, $V(\boldsymbol{\theta})$ is a $K \times K$ non-negative definite matrix characterizing the dependence structure and different variabilities of $2\hat{\boldsymbol{\gamma}}$ among the K lags.

The weights assigned to different lags are embedded in the inverse matrix $V(\boldsymbol{\theta})^{-1}$, and how these weights are allocated is in fact directly linked with the type of least squares estimation. For example, when the empirical variogram values are assumed to be uncorrelated with same variance, $V(\boldsymbol{\theta})$ reduces to the identity matrix, corresponding to having equal weight to all lags and leading to the ordinary least squares (OLS) estimate. The weighted least squares (WLS) estimate is when $V(\boldsymbol{\theta})$ is chosen to be some diagonal matrix with specified variances along the diagonal; see Cressie (1985) and Genton (1998c), among others, for different choices of weights. The generalized least squares (GLS) estimate is when $V(\boldsymbol{\theta})$ is set to be the asymptotic covariance matrix of the empirical variogram, i.e., $V(\boldsymbol{\theta}) = \mathrm{Var}(2\hat{\boldsymbol{\gamma}})$, and only assigning equal weight to all pairs with equal lag. Zhu and Stein (2002) proposed to define generalized variograms using linear filters where the horizontal, vertical and diagonal increments can be treated separately under a variogram setting. Under an increasing domain regime, all least squares estimates, including the OLS, WLS and GLS estimates, are consistent and asymptotically normal, with the GLS estimates being asymptotically statistically efficient among all (Lahiri et al., 2002). Despite the statistical efficiency of the GLS estimates, the facts that GLS requires the full covariance matrix that is difficult to obtain and consists of a nonlinear optimization make GLS less appealing to implement in practice. WLS is often employed as an alternative as it is a compromise between OLS and GLS.

Likelihood based methods are convenient tools for variogram estimation when a distributional assumption, usually normality, is valid. Maximum likelihood (ML) methods have been developed for stationary isotropic Gaussian process models for regularly and irregularly spaced locations, and large data sets problem using different approximation methods, many of which have been reviewed by Sun et al. (2012). When the mean parameter of the process is unknown and needs to be estimated, ML usually underestimates the variability of the process as it assumes the mean parameter is known (Stein, 1999, Section 6.6). Restricted maximum likelihood (REML) is useful for this problem as it maximizes the likelihood for linear combinations of the observations whose means are independent of the unknown mean parameter (Patterson and Thompson, 1971; Kitanidis, 1983). REML produces estimates with less bias compared to that from the ML estimation, especially when the number of parameters is large relative to the sample size (McGilchrist, 1989; Tunnicliffe-Wilson, 1989; Kang et al., 2003), and is also less computationally involved in practice. Compared to likelihood based methods, an alternative is via estimating equations. For example, for stationary Gaussian processes, Kaufman et al. (2008), Stein (2013) and Sun and Stein (2016) proposed different types of biased or unbiased estimating equations for covariance function estimation. In this paper, we develop new and flexible unbiased estimating equations to fit variogram models. The proposed method provides an alternative approach to the likelihood based methods, and includes the commonly used OLS, WLS and GLS as its special cases. Our method is highly efficient as a low dimensional representation of the weight matrix is adopted. The asymptotic properties of the estimators and the effect of lag set are explored. We illustrate our methodology for both lattice data and irregularly spaced data.

The remainder of the paper is organized as follows. In Section 2, we first describe the estimating equations approach, present our low rank approximation procedure, and then demonstrate how this framework can be used to generalize many widely used estimators. We end this section by presenting their theoretical properties. We investigate the statistical efficiency of different estimators constructed using our approach and examine the lag effect using a series of numerical studies and simulations in Section 3. The application of the proposed method is illustrated using a hydrology data set in Section 4. We conclude with a discussion of the proposed method.

Please cite this article in press as: Sun Y., et al., Flexible and efficient estimating equations for variogram estimation. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2017.12.006.