SMOOTHING TECHNIQUES IN SPATIAL STATISTICS

Wenceslao González Manteiga and Manuel Febrero Bande
Universidade de Santiago de Compostela, Spain
wences@zmat.usc.es

The variogram is one of the most important tools in the assessment of spatial variability of a spatial statistical model. Estimation and testing on this function is a crucial problem in random processes inference, with several applications in a broad spectrum of areas such as geostatistics, hydrology, atmospheric sciences, etc. We show in this work how a generalized family of variogram estimators can be built based on the classical ideas of smoothing techniques in nonparametric regression. Some examples will be given in order to compare the performance of Nadaraya-Watson and Local Linear estimators with the empirical variogram. The proper choice of the bandwidth for these methods will be discussed. Some applications to atmospheric and/or environmental data will also be provided. Finally, some extensions to the space-time setting will be considered. Special emphasis will be placed on teaching aspects in this talk.

INTRODUCTION

Spatial statistics constitutes nowadays one of the fundamental topics in teaching Statistical Inference. It is part of the academic program in specialized courses (especially PhD programs) in Mathematics Schools. It also appears as a study topic in engineering courses (Mining Engineering, Geostatistics…) or in MSc Programs, connected with other fields as Epidemiology, Geography, etc.

In this talk, we focus on the revision of a part of the Course on Environmental Statistics, held during the ISI&EH Conference (July 2003, University of Santiago de Compostela, Spain). This course was taught by different experts and a part of it was devoted to smoothing techniques in Spatial Statistics. This part was developed by Prof. Wenceslao González-Manteiga. Professors, students and researchers from all over the world attended the course. The outline of the course was the following:

- Geostatistical Modeling
- Spatio-temporal Modeling
- Non-stationary spatial processes

Here, we only introduce the analysis of the spatial dependence through smoothing techniques, completely new in this type of courses.

SPATIAL DEPENDENCE: EMPIRICAL VARIOGRAM ESTIMATION.

Let $\{Z(s)/\ s \in D \subset \mathbb{R}^d\}$ be a spatial random process where D is a bounded region with positive d-dimensional volume. Suppose that n data, $Z(s_1)$, $Z(s_2)$,…, $Z(s_n)$, are collected, at known spatial locations $s_1, s_2,…, s_n$, respectively. A random process is defined as intrinsic or intrinsically stationary if the following conditions are satisfied:

a) $\mathrm{E}(Z(s_i)–Z(s_j)) = 0$ for all $s_i, s_j \in \mathrm{D}$

b) $\mathrm{Var}(Z(s_i)–Z(s_j)) = 2\gamma(s_i - s_j)$, for all $s_i, s_j \in \mathrm{D}$

The latter assumptions convey to the fact that the first two moments of the difference $Z(s_i)–Z(s_j)$ depend only on the relative location, $s_i - s_j$, of the two variables. The function γ is called the semivariogram (and 2γ is the variogram).

As it has already been mentioned in the abstract, estimation of the semivariogram is a fundamental problem in intrinsic random processes inference, with applications in a broad spectrum of areas such as geostatistics, hydrology, atmospheric science, etc; see, for instance, Cressie (1991) and references therein. In particular, the semivariogram estimation plays a crucial role for spatial prediction, since the kriging equations depend on the semivariogram function which is, in general, unknown.

The semivariogram γ must satisfy the conditionally negative definiteness property:

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j \gamma(s_i - s_j) \leq 0 \quad (m \leq n) \tag{1}$$

for any $\{s_i \in \mathbb{R}^d / \ 1 \leq i \leq m\}$ and for any $\{a_i \in \mathbb{R} / \ 1 \leq i \leq m\}$, such that $\sum_{i=1}^{m} a_i = 0$. Otherwise, negative mean squared prediction errors may be obtained; therefore, property (1) will be also required from the semivariogram estimator.

Condition (b) may be replaced by the more restrictive condition:

c) Var(Z(s$_i$)–Z(s$_j$)) = 2γ(||s$_i$ – s$_j$||), for all s$_i$, s$_j$$\in$D

Then, the intrinsic random process is said to be isotropic. In this case, the first two moments of $Z(s_i)$–$Z(s_j)$ will be dependent only on the distance of the spatial locations, ||s$_i$ – s$_j$||.

For the sake of simplicity, we have considered isotropic models in this presentation; however, this assumption is not so restrictive in practice. In fact, a different semivariogram may fit in each direction, in case that the random process proves to be anisotropic. A natural and unbiased estimator based on the method of moments, due to Matheron (1963), is the empirical semivariogram given by:

$$\hat{\gamma}(r) = \frac{1}{2|N(r)|}\sum_{(s_i,s_j)\in N(r)} \left(Z(s_i) - Z(s_j)\right)^2, \, r \geq 0 \tag{2}$$

and |N(r)| is:

$$N(r) = \left\{(s_i, s_j) / \left\| s_i - s_j \right\| = r, 1 \leq i, j \leq n\right\} \tag{3}$$

When data are irregularly spaced, the latter estimator is usually smoothed by considering a tolerance region T(r) around r, rather than N(r).

An alternative estimator has been proposed in Cressie and Hawkins (1980), by using instead the root square of the differences |$Z(s_i)$–$Z(s_j)$|; however, this estimator may be destroyed by a single outlier in the data. In this sense, a more robust variogram estimator is suggested in Genton (1988a), based on a highly robust estimator of scale.

The semivariogram estimators mentioned above cannot be used directly for spatial prediction, since condition (1) typically fails. In that case, the estimation procedure must be modified in order to obtain a semivariogram estimator with this requirement. A first approach is based on choosing a parametric family first, as proposed in Cressie (1985) or in Genton (1988b), and then selecting a semivariogram in the family considered which best fits the data. However, one should take care when judging the quality of a parametric estimator obtained from the empirical semivariogram, due in part to the fact that the latter estimator is a poor tool for distinguishing the degree of smoothness of a differentiable process; see Stein (1999), and also because one may misspecify the underlying model.

A GENERALIZED FAMILY OF VARIOGRAM ESTIMATORS BASED ON SMOOTHING TECHNIQUES

The idea of averaging the square differences (Z(s$_i$)–Z(s$_j$))^2, leads to a very general construction of semivariogram nonparametric estimators as follows:

$$\tilde{\gamma}(r) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}(r)\left(Z(s_i) - Z(s_j)\right)^2}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}(r)}, \, r \geq 0 \tag{4}$$

where $w_{i,j}(r) \geq 0$, for all i, j and $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}(r) > 0.$

For instance, taking $w_{i,j}(r) = I_{\{\|s_i - s_j\| = r\}}$ yields to the empirical semivariogram (Matheron 1963).

An alternative estimator is suggested in García-Soidán *et al* (2004), as a result of adapting the Nadaraya-Watson regression estimator to the context of spatial data. The estimator obtained may be written as given in (4) by selecting:

$$w_{i,j}(r) = K\left(\frac{\|s_i - s_j\| - r}{h}\right) \tag{5}$$

where *K* denotes a symmetric density function and *h*=*h*$_n$ is the called bandwidth parameter.

Another general estimator studied in García-Soidán *et al* (2003) is constructed by using the local polynomial fitting, since it provides a kernel method with attractive properties; see Fan and Gijbels (1996) for a description of this procedure in a regression setting. For the sake of simplicity, we will apply the local linear estimation: we suppose that the semivariogram function can be locally approximated by $\gamma(r) \approx \sum_{k=0}^{1} \gamma^{(k)}(r_0)(r - r_0)^k$ for *r* in a neighborhood of *r*$_0$, by using Taylor's expansion. The latter polynomial may be fitted locally by a weighted least-squares problem, say, by obtaining β_0 and β_1 that minimizes

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{1}{2}\left(Z(s_i) - Z(s_j)\right)^2 - \sum_{k=0}^{1}\beta_k\left(\|s_i - s_j\| - r\right)^k\right]K\left(\frac{\|s_i - s_j\| - r}{h}\right). \tag{6}$$

Define

$$w_{0,i,j}(r) = K\left(\frac{\|s_i - s_j\| - r}{h}\right)\sum_{k=1}^{n}\sum_{l=1}^{n}K\left(\frac{\|s_k - s_l\| - r}{h}\right)\left(\|s_k - s_l\| - r\right)\left(\|s_k - s_l\| - \|s_i - s_j\|\right)$$

$$w_{1,i,j}(r) = K\left(\frac{\|s_i - s_j\| - r}{h}\right)\sum_{k=1}^{n}\sum_{l=1}^{n}K\left(\frac{\|s_k - s_l\| - r}{h}\right)\left(\|s_i - s_j\| - \|s_k - s_l\|\right).$$

The minimizers of (6) will be given by

$$\hat{\beta}_k = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{k,i,j}(r)\left(Z(s_i) - Z(s_j)\right)^2}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{k,i,j}(r)}.$$

Then, $\hat{\beta}_k$ may be considered as an estimator of $\gamma^{(k)}(r)$, for *k*=0,1.

The expression above means that the local linear estimator will be constructed by taking $w_{i,j}(r)$ in (4) as $w_{0,i,j}(r)$. From the two kernels, several features make advisable the use of the local linear estimation in practice. The bias at the boundary is of the same order as that in the interior, unlike the Nadaraya-Watson estimator; this is an important advantage. Moreover, boundary modifications may be a difficult task, specially in higher dimensions; for instance, the Nadaraya-Watson estimation requires the use of an specific combination of boundary kernels, to retain rates of convergence. In addition, the perfomance of the local linear semivariogram outside the boundary may be better than that kernel method for an appropiate selection of the kernel function. The properties of both estimators are detailed in García-Soidán *et al* (2003, 2004). Real environmental data applications will be shown in the talk.

SPATIO-TEMPORAL VARIOGRAM MODELS

For a stationary spatio-temporal process *Z*(*s*,*t*), denote the observed values of the process as *Z*(s$_1$, t$_1$),…, *Z*(s$_n$, t$_n$). Then, the empirical estimator of the semivariogram will be given by

$$\hat{\gamma}(r,u) = \frac{1}{2|N(r,u)|}\sum_{N(r,u)}\left(Z(s_i,t_i) - Z(s_j,t_j)\right)^2,$$

where $N(r,u) = \left\{(i,j) : \|s_i - s_j\| \in T(r), |t_i - t_j| = u\right\}$ and *T*(*r*) is a tolerance region about *r*.

In practice, if we use this estimator of the semivariogram, then the estimates may be highly variable. In order to avoid this "overfit" problem, we propose again the use of multivariate nonparametric estimators.

As in the spatial case, we can obtain an estimate of γ(*r*,*u*) by multivariate local linear least squares regression, minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{1}{2}\left(Z(s_i,t_i)-Z(s_j,t_j)\right)^2-(\beta_0,\beta_{10},\beta_{01})\begin{pmatrix}1\\ \left\|s_i-s_j\right\|-r\\ \left|t_i-t_j\right|-u\end{pmatrix}\right]^2 K_H\left[\begin{pmatrix}\left\|s_i-s_j\right\|-r\\ \left|t_i-t_j\right|-u\end{pmatrix}\right] \tag{7}$$

where $K(\cdot)$ is a bidimensional kernel and H is a bandwidth matrix. If $(\beta_0,\beta_{10},\beta_{01})$ is the solution of the straightforward weighted linear least squares problem of equation (7), then the pilot estimate of $\gamma(r,u)$ will be $\hat{\gamma}(r,u)=\hat{\beta}_0$.

The estimates obtained with this method depend highly on the bandwidth matrix *H*; thus the main concern when using this estimator is choosing an appropiate bandwidth matrix.

As well as for the purely spatial setting, several illustrations based on different criteria for the bandwidth selection will be shown in the talk, with applications to real data.

REFERENCES

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Mathematical Geology,* 17, 563-586.

Cressie, N. and Hawkins, D. (1980). Robust estimation of the variogram. *Mathematical Geology,* 12, 115-125.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. London: Chapman and Hall.

García Soidán, P., Febrero Bande, M. and Gonzalez Manteiga, W. (2003). Nonparametric kernel estimation of an isotropic semivariogram. *Statistics and Probability Letters,* 64, 169-179.

García Soidán, P., Febrero Bande, M. and Gonzalez Manteiga, W. (2004). Local linear regression estimation of the semivariogram. *Journal of Statistical Planning and Inference,* 121, 65-92.

Genton, M. (1998a). Highly robust variogram estimation. *Mathematical Geology,* 30, 213-221

Genton, M. (1998b). Variogram fitting by generalized least squares using an explicit formula for the covariance structure. *Mathematical Geology,* 30, 323-345.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58, 1246-1266.

Stein, M. L. (1999). *Interpolation of Spatial Data, Some Theory for Kriging*. Springer: New York.