Nonparametric estimation of the indicator variogram with applications

P. García-Soidán[1*] and R. Menezes[2]

[1] *Dept. of Statistics and Operations Research, University of Vigo, Spain; pgarcia@uvigo.es*
[2] *Dept. of Mathematics and Applications, University of Minho, Portugal; rmenezes@mct.uminho.pt*
[*] *Corresponding author*

Abstract. *Estimation of the distribution function of a spatial random process can be addressed in a parametric way, by imposing a shape or analytical expression for the distribution function. However, the data provided do not always support the distribution model assumption. An additional option is to proceed via the indicator kriging approach, which demands estimation of the indicator variogram (or the indicator covariance function). In this paper, we suggest a kernel-type estimator for the latter aim, as a nonparametric alternative to the empirical indicator variogram, typically used in this setting. Consistency of the kernel indicator variogram will be proved, under several assumptions. In addition, we will check that approximation of the sill of the kernel indicator variogram provides another mechanism for estimation of the distribution function.*

Keywords. *Distribution function; Indicator kriging; Kernel method; Variogram.*

1 Introduction

There are practical situations where approximation of the distribution function of a spatial random process $\{Z(\mathrm{s}) : \mathrm{s} \in D \subset \mathbb{R}^d\}$ is the issue of interest. For instance, in the estimation of metal deposits or recoverable reserves, in assessing soil contamination or in the classification schemes for map analysis, among others. Typically, a finite number of spatial locations s_i is selected, $1 \leq i \leq n$, where measurements of the variable involved are taken and used to derive information for the whole observation region, including the non-sampled locations. In this setting, estimation of the distribution function can be addressed in a parametric way, by imposing a shape or analytical expression for it. However, the data provided do not always support the distribution model assumption, so that a nonparametric approach must be adopted instead. Then, an alternative is provided by the indicator kriging, as described in [4], which proves to be an efficient method.

The indicator approach is based on the interpretation of the distribution function as the expectation of an indicator random variable, namely:

$$\mathrm{P}\left(Z(\mathrm{s}) \leq x\right) = F_{\mathrm{s}}(x) = \mathrm{E}\left[I(\mathrm{s},x)\right]$$

with $I(\mathrm{s},x) = 1$ if $Z(\mathrm{s}) \leq x$ and zero otherwise. In practice, the distribution is approximated at Q thresholds x_q, previously fixed, and the remainder values are obtained by interpolation.

The least-squares (kriging) estimator of the indicator function is also the least-squares estimator of its expectation, according to the projection theorem, as noticed in [2]. Consequently, an approximation of the distribution function at location s and threshold x is given by the indicator kriging predictor of $I(\mathrm{s},x)$, expressed as:

$$\hat{I}(\mathrm{s},x) = \sum_{i=1}^{n} \lambda_i I(\mathrm{s}_i,x) \tag{1}$$

where $\{\lambda_i : 1 \leq i \leq n\}$ are obtained by solving the corresponding kriging equations. The latter implies that a variogram (or covariance function) needs to be inferred for each threshold, referred to as indicator variogram (or indicator covariance function). This paper deals with this issue, which is called indicator structural analysis.

To develop this theory, we will assume that the random process is strictly stationary, so that $F_{\mathrm{s}}(x) = F_{\mathrm{s}'}(x) = F(x)$, for all $x \in \mathbb{R}$ and all $\mathrm{s},\mathrm{s}' \in D$. Then, the indicator variogram is defined as:

$$2\gamma_I(\mathrm{t},x) = \mathrm{Var}\left[I(\mathrm{s},x) - I(\mathrm{s}+\mathrm{t},x)\right] = \mathrm{E}\left[\left(I(\mathrm{s},x) - I(\mathrm{s}+\mathrm{t},x)\right)^2\right]$$

for each $t \in \mathbb{R}^d$ and $x \in \mathbb{R}$.

An estimator of the indicator variogram is given by the experimental or empirical variogram, derived from the method of moments:

$$2\hat{\gamma}_I(\mathrm{t},x) = \frac{1}{N(\mathrm{t})} \sum_{(i,j)\in N(\mathrm{t})} \left(I(\mathrm{s}_i,x) - I(\mathrm{s}_j,x)\right)^2$$

where $N(\mathrm{t})$ denotes the set of distinct pairs (i,j) satisfying that $\mathrm{s}_i - \mathrm{s}_j = \mathrm{t}$.

The indicator kriging also demands using a valid variogram estimator, satisfying the conditionally negative-definiteness property. The maximum likelihood method is applied in [5] with this purpose, although it is noticed that the appearance of the empirical indicator variogram can be noisy as the threshold moves away from the median, resulting in significant uncertainty in the fitted model.

In this paper, we suggest a nonparametric alternative to the empirical variogram, similar to that analyzed in [1] and adapted to the indicator setting, defined as follows:

$$2\hat{\gamma}_{I,h}(\mathrm{t},x) = \frac{\sum_{i\neq j} K\left(\frac{\mathrm{t}-(\mathrm{s}_i-\mathrm{s}_j)}{h}\right)\left(I(\mathrm{s}_i,x) - I(\mathrm{s}_j,x)\right)^2}{\sum_{i\neq j} K\left(\frac{\mathrm{t}-(\mathrm{s}_i-\mathrm{s}_j)}{h}\right)} \tag{2}$$

where K represents a d-dimensional kernel function and h is the bandwidth parameter.

The kernel indicator variogram provides a smoother estimator, whose consistency will be derived under several assumptions. In addition, we will check that a direct estimation of the distribution function can be obtained through that of the sill of the indicator variogram, as proposed in [4].

2 Main results

Let $\{Z(\mathrm{s}) : \mathrm{s} \in D \subset \mathbb{R}^d\}$ be a spatial random process and denote by $Z(\mathrm{s}_1)$, ..., $Z(\mathrm{s}_n)$, the n data collected at the spatial locations s_1, ..., s_n. An increasing observation region D will be considered and a random design will be assumed for the spatial locations, as suggested in [3] to achieve consistent estimation.

In addition, a dependence condition will be required from the random process, similar to that imposed in [7]. For this purpose, given $\mathrm{S}, \mathrm{S}' \subset \mathbb{R}^d$, take $Z[\mathrm{S}]$ as the σ-field generated by $\{Z(\mathrm{s})/\mathrm{s} \in \mathrm{S}\}$ and $d(\mathrm{S}, \mathrm{S}') = \inf\{\|\mathrm{s} - \mathrm{s}'\| : \mathrm{s} \in \mathrm{S}, \mathrm{s}' \in \mathrm{S}'\}$, where $\|\cdot\|$ denotes the l_1-norm on $\mathbb{R}^d$. Write $\alpha_1(\mathrm{S}, \mathrm{S}') = \sup\{|\mathrm{P}(A \cap B) - \mathrm{P}(A)\mathrm{P}(B)| : A \in Z[\mathrm{S}], B \in Z[\mathrm{S}']\}$. The α-mixing coefficient is defined as:

$$\alpha(k, b) = \sup\left\{\alpha_1\left(\mathrm{S}, \mathrm{S}'\right) : \mathrm{S}, \mathrm{S}' \in \mathcal{R}_p(b), d\left(\mathrm{S}, \mathrm{S}'\right) \geq k\right\}$$

with $\mathcal{R}_p(b) = \left\{\cup_{i=1}^{p} D_i : D_i's \text{ are disjoint and } \sum_{i=1}^{p} \|D_i\| \leq b\right\}$.

Next, we will describe the main hypotheses to be assumed.

(H1) $F_{\mathrm{s}_1,\ldots,\mathrm{s}_j}(x_1, \ldots, x_j) = F_{\mathrm{s}_1+d,\ldots,\mathrm{s}_j+d}(x_1, \ldots, x_j)$, for all $d \in \mathbb{R}^d$ and $j \geq 1$, with $F_{\mathrm{s}_1,\ldots,\mathrm{s}_j}(x_1, \ldots, x_j) = \mathrm{P}(Z(\mathrm{s}_1) \leq x_1, \ldots, Z(\mathrm{s}_j) \leq x_j)$

(H2) For all $j \leq 4$ and $(\mathrm{s}_1, \ldots, \mathrm{s}_j) \in D^j$, $F_{\mathrm{s}_1,\ldots,\mathrm{s}_j}(x_1, \ldots, x_j)$ admits two continuous derivatives in a neighborhood of $(\mathrm{s}_1, \ldots, \mathrm{s}_j)$, as a function of $(\mathrm{s}_1, \ldots, \mathrm{s}_j)$.

(H3) For all $(\mathrm{s}_1, \mathrm{s}_2) \in D^2$, $F_{\mathrm{s}_1,\mathrm{s}_2}(x_1, \ldots, x_j)$ admits three continuous derivatives in a neighborhood of $(\mathrm{s}_1, \mathrm{s}_2)$, as a function of $(\mathrm{s}_1, \mathrm{s}_2)$.

(H4) $D = D_n = \beta D_0$, for some $\beta = \beta_n$ diverging to $+\infty$ and some bounded region $D_0 \subset \mathbb{R}^d$ containing a sphere with positive d-dimensional volume.

(H5) The spatial locations will be taken as $\mathrm{s}_i = \beta \mathrm{u}_i$, for $1 \leq i \leq n$, where u_1, ..., u_n represents a realization of a random sample of size n drawn from g_0, where g_0 is the density function considered on D_0.

(H6) For a given $\mathrm{t} \in \mathbb{R}^d$, g_0 admits three continuous derivatives in a neighborhood of t.

(H7) $\left\{h + (nh)^{-1} + \beta^{-1} + n^{-2}\beta^d h^{-d}\right\} \xrightarrow{n\to\infty} 0$.

(H8) $\alpha(k, b) \leq c_1 k^{-c_2} b^{c_3}$, for some positive real numbers c_1, c_2, c_3.

(H9) K is a d-variate, compactly supported, symmetric and bounded density function, with $K(0) > 0$.

Under assumptions (H1)-(H9), estimator $\hat{\gamma}_{I,h}(\mathrm{t}, x)$ satisfies several properties, such as asymptotically unbiasedness and consistency, for all $\mathrm{t} \in \mathbb{R}^d$ and $x \in \mathbb{R}$. More specifically, we can check that:

$$\mathrm{Bias}\left[2\hat{\gamma}_{I,h}(\mathrm{t}, x)\right] = 2h^2 \sum_{i,j} \frac{\partial^2 \gamma_{I,h}}{\partial \mathrm{t}^{(i)} \partial \mathrm{t}^{(j)}}\bigg|_{(\mathrm{t},x)} \int \mathrm{z}^{(i)} \mathrm{z}^{(j)} K(\mathrm{z}) d\mathrm{z} + o\left(h^2\right) \text{ and } \mathrm{Var}\left[2\check{\gamma}_h(\mathrm{s})\right] = o(1) \qquad (3)$$

In consequence, the MSE and the MISE of the kernel indicator variogram tend to zero as the sample size increases, so that minimization of the above quantities can provide asymptotically optimal bandwidth parameters. An alternative for selection of h may be that of considering a balloon estimator, namely, a kernel estimator where the bandwidth is allowed to vary with the lag t, as developed in [6] for density

estimation. For instance, we could take $h = h_k(\mathrm{t})$ as the euclidean distance from t to the k-nearest distances between locations in the sample.

Denote by F the univariate distribution function, namely, $F_{\mathrm{s}} = F$ for all $\mathrm{s} \in \mathbb{R}^d$. The kernel indicator variogram can be used for approximation of F either in an indirect way, by applying the kriging techniques, or directly, as an application of the proposal given in [4]. For both approaches, estimation of the distribution function F will be discretized at Q thresholds x_q, previously fixed, and the remainder values will be approximated by interpolation. To proceed in the first way, the kernel indicator variogram $2\hat{\gamma}_{I,h}(\cdot, x_q)$ must be obtained for each q and used to solve the kriging equations, which provides Q values of the distribution function F.

The second alternative for estimation of F will be derived from that of the sill of the indicator variogram. With this aim, bear in mind that the sill $S(x)$ of the indicator variogram is linked to the distribution function as follows:

$$S(x) = \lim_{\|\mathrm{t}\| \to \infty} \gamma_{I,h}(\mathrm{t}, x) = F(x) - F(x)^2$$

Furthermore, $S(x)$ is increasing in $(-\infty, x_M]$ and decreasing in $[x_M, \infty)$ and takes values in $[0, 0.25]$, where x_M stands for the median of the distribution F. Then, we propose to proceed as follows:

- Approximate the sill at each threshold, $S^*(x_q)$.
- Determine the value of the median, x_M^*, by selecting the value x_q for which $S^*(x_q)$ is maximum and close to 0.25, so that $F(x_M^*) \approx 0.5$.
- For each x_q, take $F(x_q) \approx 0.5\left(1 + \varepsilon(x_q)\sqrt{1 - 4S^*(x_q)}\right)$, with $\varepsilon(x) = sign(x - x_M^*)$.

Numerical studies will be included to illustrate the performance of both approaches for approximation of the distribution function.

Acknowledgments. This work has been supported in part by grant INCITE-08-PXIB-322219-PR from Consellería de Innovación e Industria (Xunta de Galicia, Spain).

References

[1] García-Soidán, P. (2007). Asymptotic normality of the Nadaraya-Watson semivariogram estimator. *TEST* **16**, 479–503.

[2] Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Oxford University Press. Oxford.

[3] Hall, P. and Patil, P. (1994). Properties of nonparametric estimators of autocovariance for stationary random fields. *Probability Theory and Related Fields* **99**, 399–424.

[4] Journel, A. G. (1983). Nonparametric estimation of spatial distribution. *Mathematical Geology* **15**, 445–468.

[5] Pardo-Igúzquiza, E. (1998). Inference of spatial indicator covariance parameters by maximum likelihood using MLREML. *Computers & Geosciences* **24**, 453–464.

[6] Terrell, G. . and Scott, D. W. (1992). Variable Kernel Density Estimation. *Annals of Statistics* **20**, 1236–1265.

[7] Zhu, J. and Lahiri, S. N. (2007). Bootstrapping the Empirical Distribution Function of a Spatial Process. *Statistical Inference for Stochastic Processes* **10**, 107–145.