Monografías del Seminario Matemático García de Galdeano **33**, 253–258 (2006)

GOODNESS OF FIT TESTS FOR ISOTROPIC VARIOGRAM MODELS

Pilar García-Soidán and Carmen Iglesias-Pérez

Abstract. The aim of this work is to provide procedures to check if the theoretical semivariogram of an intrinsic and isotropic random process follows a parametric model. For this purpose, several tests based on measuring the L_2 distance between the parametric fit and a nonparametric kernel semivariogram are proposed, which are proved to have normal limit distributions.

Keywords: Goodness of fit test, intrinsic random process, isotropy, variogram.

AMS classification: 62G05, 62G10.

§1. Introduction

An adequate estimation of the semivariogram is fundamental to perform inference on an intrinsic random process; see, for instance, N. Cressie [1] and references therein. For the sake of simplicity, we will restrict our attention to the isotropic semivariograms.

Definition 1. A random process $\{Z(\mathrm{t}) \mid \mathrm{t} \in D \subset \mathbb{R}^d\}$ is defined as intrinsic with semivariogram γ if the following conditions are satisfied:

(i) $\mathrm{E}[Z(\mathrm{t}_1) - Z(\mathrm{t}_2)] = 0$, for all $\mathrm{t}_1, \mathrm{t}_2 \in D$.

(ii) $\mathrm{Var}[Z(\mathrm{t}_1) - Z(\mathrm{t}_2)] = 2\gamma(\mathrm{t}_1 - \mathrm{t}_2)$, for all $\mathrm{t}_1, \mathrm{t}_2 \in D$.

Definition 2. The intrinsic random process is said to be isotropic if hypothesis (ii) above is replaced by the more restrictive condition:

(ii') $\mathrm{Var}[Z(\mathrm{t}_1) - Z(\mathrm{t}_2)] = 2\gamma(\|\mathrm{t}_1 - \mathrm{t}_2\|)$, for all $\mathrm{t}_1, \mathrm{t}_2 \in D$.

Suppose that n data $Z(\mathrm{s}_1), \ldots, Z(\mathrm{s}_n)$, are collected at $\mathrm{s}_1, \ldots, \mathrm{s}_n$. A natural nonparametric estimator of γ is the empirical semivariogram. An alternative may be that of considering the Nadaraya-Watson (NW) estimator in this setting, defined as follows:

$$\hat{\gamma}_h(s) = \frac{\sum_{i=1}^n \sum_{j=1}^n K\left(\frac{s - \|\mathrm{s}_i - \mathrm{s}_j\|}{h}\right) \left(Z(\mathrm{s}_i) - Z(\mathrm{s}_j)\right)^2}{2\sum_{i=1}^n \sum_{j=1}^n K\left(\frac{s - \|\mathrm{s}_i - \mathrm{s}_j\|}{h}\right)}, \; s \geq 0.$$

In P. García-Soidán et al. [2], some properties of $\hat{\gamma}_h(s)$ are established; in particular, that it is asymptotically unbiased as well as consistent, under several conditions.

The aim of this work is to provide procedures to check whether or not the theoretical semivariogram of an intrinsic and isotropic random process follows a parametric model, by carrying out the following contrast:

$$\begin{array}{l} H_0 : \gamma \in \Gamma_\theta = \{\gamma_\theta(\cdot) \mid \theta \in \Theta \subset \mathbb{R}^p\} \text{ versus} \\ H_1 : \ \gamma \notin \Gamma_\theta . \end{array} \tag{1}$$

For this purpose, several tests based on measuring the L_2 distance between the parametric fit and the Nadaraya-Watson semivariogram are proposed, which are proved to be asymptotically normal distributed.

§2. Hypotheses

(S1) $D = \lambda D_0$, for some $\lambda = \lambda_n \xrightarrow{n\to\infty} +\infty$ and some bounded region $D_0 \subset \mathbb{R}^d$ containing a sphere with positive d-dimensional volume.

(S2) Let f_0 be a density on D_0. Then, f_0 is bounded and strictly positive on D_0.

(S3) $\mathrm{s}_i = \lambda\, \mathrm{u}_i$, for $1 \le i \le n$, where $\mathrm{u}_1,\ldots,\mathrm{u}_n$ represents a realization of a random sample of size n drawn from f_0, which will be denoted by $\mathrm{U}_1,\ldots,\mathrm{U}_n$.

(S4) Denote by f_i the density of $(\mathrm{U}_1-\mathrm{U}_2,\ldots,\mathrm{U}_1-\mathrm{U}_{i+1})$. Then, $f_1(0) > 0$ and f_i is continuously differentiable in a neighborhood of 0^+, for all $i \le 7$.

(S5) K is a compactly supported, symmetric and bounded density function.

(S6) $\{h + (nh)^{-1} + \lambda^d n^{-1} + n^2\lambda^{-2d}h\} \xrightarrow{n\to\infty} 0$. Moreover, $\lim_{n\to\infty} h^5 n^2 \lambda^{-d} = c \ge 0$.

(S7) $\{Z(\mathrm{t}) \mid \mathrm{t}\in D\subset\mathbb{R}^d\}$ is an intrinsic and isotropic random process with semivariogram γ, satisfying that $\mathrm{E}[Z(\mathrm{t})^8] < \infty$, for all $\mathrm{t}\in D$.

(S8) γ admits three continuous derivatives in a neighborhood of s, for all $s \in (0,y)$.

(S9) $\mathrm{Var}[(Z(\mathrm{t}_1) - Z(\mathrm{t}_2))^2] = g(\|\mathrm{t}_1 - \mathrm{t}_2\|)$, for all $\mathrm{t}_1,\mathrm{t}_2 \in D$ and some $g : \mathbb{R} \to \mathbb{R}$.

(S10) g admits two continuous derivatives in a neighborhood of s, for all $s \in (0,y)$.

(S11) Assuming a parametric model $\Gamma_\theta = \{\gamma_\theta(\cdot) \mid \theta \in \Theta \subset \mathbb{R}^p\}$ for γ and given a set $\{s_i\}_{i=1}^k$ with $s_i > 0$, we will ask that, for any $\varepsilon > 0$, there exists a $\delta > 0$ such that:

$$\inf\left\{ \sum_{i=1}^k (\gamma_{\theta_1}(s_i) - \gamma_{\theta_2}(s_i))^2 \,\middle|\, \|\theta_1 - \theta_2\| \ge \varepsilon \right\} > \delta.$$

(S12) γ_θ is bounded and r-times continuously differentiable with respect to θ.

(S13) $V(\theta)$ is a positive definite $k \times k$ matrix, which is s-times continuously differentiable on Θ, with $\sup\left\{\|V(\theta)\| + \left\|V(\theta)^{-1}\right\| \;\middle|\; \theta \in \Theta\right\} < \infty$.

§3. Main results

Theorem 1. *Assume that conditions (S1)–(S9) are satisfied. It follows that*

$$(hv(y))^{-1/2}\left(n^2\lambda^{-d}h\int_0^y (\hat{\gamma}_h(s)-\gamma(s))^2\,ds - m(y)\right) \xrightarrow{d} N(0,1),$$

with

$$m(y) = \frac{c\left(\int_{\mathbb{R}} z^2K(z)\,dz\right)^2}{4}\int_0^y \gamma''(s)^2\,ds + \frac{K*K(0)}{2f_1(0)A_{0,d}}\int_0^y \frac{g(s)}{s^{d-1}}\,ds,$$

$$v(y) = \frac{c\int_{\mathbb{R}} z^2K(z)dz}{f_1(0)A_{0,d}}\int_0^y \frac{\gamma''(s)^2g(s)}{s^{d-1}}\,ds + \frac{K*K*K*K(0)}{2\left(f_1(0)A_{0,d}\right)^2}\int_0^y \frac{g(s)^2}{s^{2(d-1)}}\,ds,$$

$$A_{0,d} = \int_0^{\pi}\cdots\int_0^{\pi}\int_0^{2\pi}(\sin\theta_1)^{d-2}(\sin\theta_2)^{d-3}\cdots\sin\theta_{d-2}\,d\theta_1\cdots d\theta_{d-2}\,d\theta_{d-1},$$

where constant c is given in condition (S6) and $*$ denotes convolution.

Proof. From Theorems 3.1 and 3.2 in P. García-Soidán et al. [2], it is straightforward to check that

$$\begin{aligned} &n^2\lambda^{-d}h\int_0^y \left(\hat{\gamma}_h(s)-\gamma(s)\right)^2 ds \\ &\quad = \frac{c\left(\int_{\mathbb{R}} z^2K(z)\,dz\right)^2}{4}\int_0^y \gamma''(s)^2\,ds + \frac{cn\lambda^{-d/2}h^{1/2}\int_{\mathbb{R}} z^2K(z)\,dz}{f_1(0)A_{0,d}}\int_0^y \frac{\gamma''(s)}{s^{d-1}}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X_{i,j}(s)\,ds \\ &\quad + \frac{n^2\lambda^{-d}h}{\left(f_1(0)A_{0,d}\right)^2}\int_0^y\left(\frac{1}{s^{d-1}}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X_{i,j}(s)\right)^2 ds + o_P\left(h^{1/2}\right), \end{aligned}$$

where

$$X_{i,j}(s) = K\left(\frac{s-\lambda\left\|\mathrm{U}_i-\mathrm{U}_j\right\|}{h}\right)\left[\left(Z(\lambda\mathrm{U}_i)-Z(\lambda\mathrm{U}_j)\right)^2 - 2\gamma\left(\lambda\left\|\mathrm{U}_i-\mathrm{U}_j\right\|\right)\right].$$

Then, Theorem 1 would be proved if we checked that

$$\left\{\frac{n\lambda^{-d/2}h^{1/2}}{f_1(0)A_{0,d}\,s^{d-1}}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X_{i,j}(s)\;\middle|\; s\in(0,y)\right\}$$

converges to a gaussian process $\{X(s)\mid s\in(0,y)\}$, with zero mean and covariance function given by

$$\operatorname{Cov}[X(s),X(t)] = \frac{(s+t)^{d-1}K*K\left(\frac{t-s}{h}\right)g\left(\frac{s+t}{2}\right)}{2^d f_1(0)A_{0,d}\,s^{d-1}t^{d-1}}.$$

For the latter purpose, it would be enough to establish the asymptotic normality of

$$S = n\lambda^{-d/2}h^{1/2}\sum_{l=1}^{m}\beta_l\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X_{i,j}(s_l) = n\lambda^{-d/2}h^{1/2}\sum_{i=2}^{n} Z_i$$

for any set of positive distances s_l and real parameters β_l, $\beta_l \neq 0$, with $1 \leq l \leq m$ and $m \in \mathbb{N}$, where

$$Z_i = n\lambda^{-d/2}h^{1/2}\sum_{j=1}^{i-1}\sum_{l=1}^{m}\beta_l X_{i,j}(s_l).$$

Bear in mind that $\mathrm{E}[Z_i/\mathrm{U}_1,\ldots,\mathrm{U}_{i-1}] = 0$ and that the random variables $Z_2,\ldots, Z_n$ may be considered as differences of the martingales $S_2,\ldots, S_n$, given by

$$S_2 = Z_2,\ S_3 = Z_2 + Z_3,\ldots,\ S_n = Z_2 + \cdots + Z_n.$$

After some algebra, we might see that $\mathrm{E}\left[\sum_{i=2}^{n} Z_i^2\right]$ and $\mathrm{E}\left[\sum_{i=2}^{n} Z_i^4\right]$ are of the respective exact orders 1 and $n^{-2}\lambda^d h^{-1}$.

From the relations above, it follows the Lyapunov condition and, therefore, the Lindeberg condition. Then, Corollary (2.13) in D. L. McLeish [4] allows to state the normal limit distribution of S. □

Remark 1. Note that $\hat{\gamma}_h(s)$ is not well-defined for large s; therefore, the integral considered in Theorem 1, $\int_0^y (\hat{\gamma}_h(s) - \gamma(s))^2\, ds$, cannot be extended to the case $y = \infty$. In the covariance estimation setting, the latter extension may be easily obtained by introducing a weight function, since the covariance function is usually assumed to tend to zero as the distance increases. However, this is not the point when the variogram estimation is considered; on the contrary, the variogram is typically required to have a positive sill which, in addition, should be estimated in practice.

Remark 2. The second part of condition (S6) is introduced in order to guarantee that the bandwidth h considered in Theorem 1 is of the optimal order; see P. García-Soidán et al. [2] for details.

Remark 3. For a gaussian process, one has that $g(s) = 8\gamma(s)^2$. Thus, as an application of Theorem 1, we can test:

$$\begin{aligned} &H_0 : \gamma = \gamma_\theta \text{ versus} \\ &H_1 : \gamma \neq \gamma_\theta \end{aligned}$$

for a fixed $\theta \in \mathbb{R}^p$ and a gaussian process, at an approximate level α. For the latter purpose, write m_θ and v_θ for those functions obtained by substituting γ_θ and $8\gamma_\theta^2$ for γ and g, respectively, in m and v. We will reject when

$$\int_0^y (\hat{\gamma}_h(s) - \gamma_\theta(s))^2\, ds \geq n^{-2}\lambda^d h^{-1}\Big(m_\theta(y) + z_{1-\alpha}\,(hv_\theta(y))^{1/2}\Big),$$

where z_β denotes the β-quantile of the $N(0,1)$ distribution.

Bear in mind that our interest is to test the general contrast given in (1). The latter requires the choice of an estimator $\hat{\theta}_n$ of the true parameter θ_0, under the null hypothesis H_0. This issue will be addressed by applying the least squares criteria, which will guarantee an appropriate rate of convergence of $\hat{\theta}_n$.

Definition 3. Given a parametric family Γ_θ, a set $\{s_i\}_{i=1}^k$ with $s_i > 0$ and a positive definite $k \times k$ matrix $V(\theta)$, the least squares estimator $\hat{\theta}_n$ will be defined as

$$\hat{\theta}_n = \arg\min\Big\{ \big(\vec{\hat{\gamma}} - \vec{\gamma}_\theta\big)^T (\mathrm{V}(\theta))^{-1} \big(\vec{\hat{\gamma}} - \vec{\gamma}_\theta\big) \;\Big|\; \theta \in \Theta \subset \mathbb{R}^p \Big\},$$

where $\vec{\hat{\gamma}} = (\hat{\gamma}_h(s_1),\dots,\hat{\gamma}_h(s_k))^T$, $\vec{\gamma}_\theta = (\gamma_\theta(s_1),\dots,\gamma_\theta(s_k))^T$ and $\gamma_\theta \in \Gamma_\theta$.

Theorem 2. *Assume the conditions required in Theorem 1 and that (S11)–(S13) hold for $r = s = 1$. Then, it follows that $\|\hat{\theta}_n - \theta_0\| = o_P(n^{-1}\lambda^{d/2})$, where θ_0 denotes the true parameter under the null hypothesis in (1).*

Proof. This result follows from Theorem 3.2 in S. Lahiri et al. [3]. □

Theorem 3. *Under the assumptions of Theorem 2, if (S12) holds for $r = 3$, then*

$$\left(hv_{\hat{\theta}_n}(y)\right)^{-1/2}\left(n^2\lambda^{-d}h\int_0^y \left(\hat{\gamma}_h(s) - \gamma_{\hat{\theta}_n}(s)\right)^2 ds - m_{\hat{\theta}_n}(y)\right) \xrightarrow{d} N(0,1),$$

with $m_{\hat{\theta}_n}$ and $v_{\hat{\theta}_n}$ obtained by replacing γ by $\gamma_{\hat{\theta}_n}$ in m and v given in Theorem 1.

Proof. We may apply Theorem 2 and Taylor expand about θ_0 to yield that

$$\begin{aligned}
\gamma_{\hat{\theta}_n}(s) - \gamma_{\theta_0}(s) &= o_P(n^{-1}\lambda^{d/2}),\\
m_{\hat{\theta}_n}(y) - m_{\theta_0}(y) &= o_P(n^{-1}\lambda^{d/2}),\\
v_{\hat{\theta}_n}(y) - v_{\theta_0}(y) &= o_P(n^{-1}\lambda^{d/2}),
\end{aligned}$$

for all $s \in (0,y)$. Consequently, Theorem 3 follows by combining the latter relations and Theorem 1. □

Remark 4. An immediate consequence of Theorem 3 is that we can test the contrast given in (1) for a gaussian process, at an approximate level α. From Theorem 3, the rejection region would be given by

$$\int_0^y \left(\hat{\gamma}_h(s) - \gamma_{\hat{\theta}_n}(s)\right)^2 ds \ge n^{-2}\lambda^d h^{-1}\left(m_{\hat{\theta}_n}(y) + z_{1-\alpha}\left(hv_{\hat{\theta}_n}(y)\right)^{1/2}\right).$$

Recall that our aim is to provide a procedure to check the contrast (1) for a general intrinsic random process (not necessarily gaussian). Then, function g must be estimated in practice, since the form of this function is in general unknown. With this idea, we may consider the NW estimator of $g(s)$, given by

$$\hat{g}_h(s) = \frac{\sum_{i=1}^n\sum_{j=1}^n K\left(\frac{s-\|\mathrm{s}_i-\mathrm{s}_j\|}{h}\right)\left((Z(\mathrm{s}_i)-Z(\mathrm{s}_j))^2 - 2\hat{\gamma}_h\left(\|\mathrm{s}_i-\mathrm{s}_j\|\right)\right)^2}{\sum_{i=1}^n\sum_{j=1}^n K\left(\frac{s-\|\mathrm{s}_i-\mathrm{s}_j\|}{h}\right)},\ s \ge 0.$$

Theorem 4. *Suppose that conditions (S1)–(S10) hold. Then, for all $s \in (0,y)$, one has:*

$$\begin{aligned}
\mathrm{E}[\hat{g}_h(s)] &= g(s) + O(h^2),\\
\mathrm{Var}[\hat{g}_h(s)] &= O(n^{-2}\lambda^d h^{-1} + h^4).
\end{aligned}$$

Proof. To derive this proof, we might proceed similarly as in the proofs of Theorems 3.1 and 3.2 in P. García-Soidán et al. [2]. □

Theorem 5. *Assume the conditions required in Theorem 3 together with hypothesis (S10). Then, one has*

$$\left(h\hat{v}_{\hat{\theta}_n}(y)\right)^{-1/2}\left(n^2\lambda^{-d}h\int_0^y \left(\hat{\gamma}_h(s)-\gamma_{\hat{\theta}_n}(s)\right)^2 ds - \hat{v}_{\hat{\theta}_n}(y)\right) \xrightarrow{d} N(0,1),$$

with $\hat{m}_{\hat{\theta}_n}$ and $\hat{v}_{\hat{\theta}_n}$ obtained by substituting $\hat{g}_h$ for g in $m_{\hat{\theta}_n}$ and $v_{\hat{\theta}_n}$ defined in Theorem 3.

Proof. This result follows straightforwardly from Theorems 3 and 4. □

Remark 5. For a general random process, we can test the contrast given in (1), at an approximate level α. From Theorem 5, we will reject when:

$$\int_0^y \left(\hat{\gamma}_h(s)-\gamma_{\hat{\theta}_n}(s)\right)^2 ds \geq n^{-2}\lambda^d h^{-1}\left(\hat{m}_{\hat{\theta}_n}(y)+z_{1-\alpha}\left(h\hat{v}_{\hat{\theta}_n}(y)\right)^{1/2}\right).$$

Remark 6. For small sample sizes, the unsatisfactory behavior near endpoints may affect the performance of the NW estimator or, even, the asymptotic distributions achieved may be inappropriate to approximate the critical values. To avoid the first problem, a boundary kernel might be used instead of a symmetric kernel in the NW semivariogram; a solution for the second one could be based on the use of the Bootstrap techniques.

References

[1] CRESSIE, N. *Statistics for spatial data*. Wiley, New York, 1993.

[2] GARCÍA-SOIDÁN, P., FEBRERO-BANDE, M., AND GONZÁLEZ-MANTEIGA, W. Nonparametric kernel estimation of an isotropic semivariogram. *Journal of Statistical Planning and Inference 121*, 1 (2004), 65–92.

[3] LAHIRI, S., LEE, Y., AND CRESSIE, N. On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters. *Journal of Statistical Planning and Inference 103* (2002), 65–85.

[4] MCLEISH, D. L. Dependent central limit theorems and invariance principles. *Annals of Probability 2* (1974), 620–628.

Pilar García-Soidán
Fac. CC. Sociales y de la Comunicación
Universidad de Vigo
Campus A Xunqueira
36005 Pontevedra, Spain
pgarcia@uvigo.es

Carmen Iglesias-Pérez
E.U. Ingeniería Técnica Forestal
Universidad de Vigo
Campus A Xunqueira
36005 Pontevedra, Spain
mcigles@uvigo.es