HOW TO ESTIMATE THE PRECISION OF THE VARIOGRAM

K.G. v.d. Boogaart
TU Bergakademie Freiberg, Germany; *boogaart@grad.tu-freiberg.de*

Introduction

The accuracy of the modeled variogram is essential for any geostatistical analysis. While geostatistics usually involves no assumptions on the underlying distributions, typical considerations on the precision of variogram estimations are base on a multivariate Gaussian distribution of the underlying random field. A new method purely based on moments and independent of any assumption of normality is introduced and used to estimate the precision of the empirical variogram and estimated variogram parameters based on the information provided by the dataset itself. The method is based on a new method to estimate prediction variances of means of spatially correlated observations (Boogaart 2002), which is itself an interesting task.

Estimating the empirical variogram

The semivariogram $\gamma(h)$ for a given lags h is usually estimated by half the mean of squared increments for that given lag:

$$\hat{\gamma}(h) := \frac{1}{2|N(h)|}\sum_{(i,j)\in N(h)}(Z(x_i)-Z(x_j))^2$$

The estimator is obviously unbiased, but how precise is this estimate. Due to the spatial correlation of the different $(Z(x_i)-Z(x_j))^2$ the variance of this mean is not just 1/n times the variance of $(Z(x_i)-Z(x_j))^2$ which itself is unknown. The variability of this estimate can be inferred by simulation or by calculations based on the forth order moments of a distribution model (Gento 1998, Bogaert and Russo 1999, Pardo-Iguzquiza 1998, Pardo-Iduzquiza and Dowd 2001). Both approaches are based on a known forth order structure of the regionalized variable. The same is true for the approximate formula

$$\mathrm{var}(\hat{\gamma}(h)) \approx \frac{2\gamma(h)^2}{|N(h)|}$$

given by Cressie (1985), which uses the assumption of gaussian distributions and independent increments. When the fourth order structure is specified, the variance of the mean of a set of random variables is given by

$$\mathrm{var}(\bar{y}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{cov}(y_i, y_j)$$

which reduces to the well known

$$\mathrm{var}(\bar{y}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\mathrm{var}(y)$$

for i.i.d. variables y_i due to $\mathrm{cov}(y_i,y_j)=\mathrm{var}(y)$ and $\mathrm{cov}(y_i,y_j)=0$ for all $i \neq j$.

However I want to give an approach to estimate the variability of the variogram without doubtful assumptions about the distribution or higher order moments of the process. Anyway these parametric approaches stay useful and superior to the approach presented here, when the distribution of the random field is known.

Estimating the variance of the mean of spatially correlates measurments

In (Boogaart 2002) it is shown, that the variance of a mean $\bar{y}$ of correlated random variables y_i with the same expectation can be estimated unbiasedly by

$$\hat{\mathrm{var}}(\bar{y}) := \bar{y}^2 - \frac{1}{|N^C|}\sum_{(i,i')\in N^C} y_i y_{i'} \qquad (1)$$

when we know a set N^C of uncorrelated pairs of observations.

$$(i,i') \in N^C \Rightarrow \mathrm{cov}(y_i, y_{i'}) = 0$$

The unbiasedness of the estimator is easily proved by:

$$\begin{aligned} E[\hat{\mathrm{var}}(\bar{y})] &= E[\bar{y}^2] - \frac{1}{N^C} \sum_{(i,i' \in N^C)} E[y_i, y_{i'}] \\ &= E[\bar{y}^2] - \frac{1}{N^C} \sum_{(i,i' \in N^C)} E[y]^2 \\ &= E[\bar{y}^2] - E[\bar{y}]^2 = \mathrm{var}(\bar{y}) \end{aligned}$$

When the pairs in N^C are not precisely uncorrelated but only weekly correlated:

$$(i,i') \in N^C \Rightarrow \left\| \mathrm{cov}(y_i, y_{i'}) \right\| \le \varepsilon$$

then $\hat{\mathrm{var}}(\bar{y})$ is slightly biased. The bias is limited by ε:

$$E[\hat{\mathrm{var}}(\bar{y})] = E[\bar{y}^2] - \frac{1}{N^C} \sum_{(i,i' \in N^C)} (E[y]^2 + \mathrm{cov}(y_i, y_{i'})) \in [\mathrm{var}(\bar{y}) - \varepsilon, \mathrm{var}(\bar{y} + \varepsilon)] \quad (2)$$

Thus small (typically positive) correlations in the pairs in N^C result in a small (typically negative) bias in the estimated variance. For further details the reader is referred to (Boogaart 2002).

Covariance of spatial means

A similar method can be used to estimate the covariance of means of correlated observations unbiasedly.

$$\hat{\mathrm{cov}}(\bar{y}, \bar{z}) := \bar{y}\bar{z} - \frac{1}{N^C} \sum_{(i,i') \in N^C} y_i z_{i'} \quad (3)$$

with

$$(i,i') \in N^C \Rightarrow \mathrm{cov}(y_i, z_{i'}) = 0$$

The unbiasedness is established by:

$$\begin{aligned} E[\hat{\mathrm{cov}}(\overline{yz})] &= E[\overline{yz}] - \frac{1}{N^C} \sum_{(i,i') \in N^C} E[y_i z_{i'}] \\ &= E[\bar{y}^2] - \frac{1}{N^C} \sum_{(i,i') \in N^C} E[y]E[z] \\ &= E[\bar{y}^2] - E[\bar{y}]E[\bar{z}] \\ &= \mathrm{cov}(\bar{y}, \bar{z}) \end{aligned}$$

and again the bias is bounded by ε

$$E[\hat{\mathrm{cov}}(\bar{y}, \bar{z})] = E[\bar{y}^2] - \frac{1}{N^C} \sum_{(i,i') \in N^C} (E[y]E[z] + \mathrm{cov}(y_i, z_{i'})) \in [\mathrm{cov}(\bar{y}, \bar{z}) - \varepsilon, \mathrm{cov}(\bar{y}, \bar{z}) + \varepsilon]$$

in case of a remanent covariance of ε

$$(i,i') \in N^C \Rightarrow \left\| \mathrm{cov}(y_i, z_{i'}) \right\| \le \varepsilon$$

The empirical variogram as mean correlated quantities

Defining $y_{(i,i')} := (Z(x_i) - Z(x_{i'}))^2$ the empirical variogram can be seen as a mean of correlated quantities:

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j) \in N(h)} y_{(i,i')}$$

Since the $y_{(i,i')}$ and $y_{(i,i')}$ functionally depend on $Z(x_i)$, $Z(x_{i'})$, $Z(x_j)$, $Z(x_{j'})$ they are stochastically independent from each other, whenever $Z(x_i)$ and $Z(x_{i'})$ are jointly independent of $Z(x_j)$ and $Z(x_{j'})$. We assume the $y_{(i,i')}$ to be stochastically independent or to be weakly correlated when the minimum distance

$$d((i,i'),(j,j')) := \min\left(\left\| x_i - x_j \right\|, \left\| x_{i'} - x_j \right\|, \left\| x_i - x_{j'} \right\|, \left\| x_{i'} - x_{j'} \right\| \right)$$

between the pairs exceeds some known distance R. The stochastic independence can be proved for any $R > R_0$ larger than the stochastic range R_0 to be defined as:

$$R_0 = \inf\{R \geq 0 : \forall n : \forall m : \forall (x_i)_{i=1,\ldots,n} : \forall (y_i)_{j=1,\ldots,m} : \inf \|x_i - y_i\| \geq R \Rightarrow$$
$$(Z(x_i))_{i=1,\ldots,n} \text{ is jointly stochastic independent to } (Z(y_j))_{j=1,\ldots,m}\}$$

For Gaussian random fields this definition of a stochastic range is equivalent to the true range of the covariance function, since too jointly Gaussian vectors $(Z(x_i))_{i=1,\ldots,n}$ and $(Z(x_j))_{j=1,\ldots,m}$ are stochastically independent if and only if all pair correlations are 0. Thus they are independent if and only if the minimum distance $\inf_{i,j} |x_i-y_j|$ equals the range of the covariance function. For non gaussian random fields the stochastic range R_0 is larger or equal to the range of the covariance function. The range of the covariance function in this context is the true range, which is infinit for exponential or Gaussian covariogram and not the pseudo range used to define their scale parameter.

However even for fields with infinite range such as Gaussian fields with exponential or Gaussian covariance function any general nongaussian fields with weak mixing properties, the correlation of $y_{(i,i')}$ and $y_{(j,j')}$ asymptotically vanishes for increasing distance of the pairs. Thus we need to guess an upper limit R_0 for the range and an upper limit ε of the remanent correlation in that distance.

With this setting a set of pairs increments with no or little covariance $(\leq \varepsilon)$ can be given by:

$$N^C(h_1,h_2,R) := \{((i,i'),(j,j')) : (i,i') \in N(h), (j,j') \in N(h),$$
$$\|x_i - x_j\| > R, \|x_{i'} - x_j\| > R, \|x_i - x_{j'}\| > R, \|x_{i'} - x_{j'}\| > R\}$$

Estimation of the variance of the empirical variogram

Setting $y_{(i,j)} := (Z(x\_i)-Z(x\_j))^2$ formula (1) can be applied to estimate the variance of empirical variogram estimator, where (i,j) the double index of increments replaces the single index i:

$$\hat{\text{var}}(\hat{\gamma}(h)) := \hat{\gamma}(h)^2 - \frac{1}{4\left|N^C(h,h,R)\right|} \sum_{((i,j),(i',j'))\in N^C(h,h,R)} (Z(x_i)-Z(x_j))^2 (Z(x_{i'})-Z(x_{j'}))^2$$

It is only necessary to find some set containing only uncorrelated pairs or nearly uncorrelated pairs, when the bias bound given in formula (2) is used.

In a similar way we can estimate the covariance of $\hat{\gamma}(h_1)$ and $\hat{\gamma}(h_2)$ unbiasedly by

$$\hat{\text{cov}}(\hat{\gamma}(h_1),\hat{\gamma}(h_2)) := \hat{\gamma}(h_1)\hat{\gamma}(h_2) - \frac{1}{4\left|N^C(h_1,h_2,R)\right|} \sum_{((i,j),(i',j'))\in N^C(h_1,h_1,R)} y_{(i,i')} y_{(j,j')}$$

Estimated precision of fitted variogram models

When we fit a variogram models to the empirical semivariogram the variability error is propagated through the fitting procedure. Although the fitting of variogram model parameters is a nonlinear problem a linear we can linearize it to assess the propageted error. Let $\gamma(h;p_1,\ldots,p_d)$ be a semivariogram model fitted by generalized least squares with weight matrix $W = (w_{ij})_{i=1,\ldots,e,j=1,\ldots,e}$ to a empirical semi variogram $g = (g_i)_i = \left(\hat{\gamma}(h_i)\right)_i$, i = 1,…,e, such that

$$\sum_{i,j} (\gamma(h_i;p_1,\ldots,p_d) - g_i) w_{ij} (\gamma(h_j;p_1,\ldots,p_d) - g_j) \to \min$$

Let $\hat{p}_1,\ldots,\hat{p}_d$ denote the estimated parameters. The minimum implies, that the derivative is 0 and thus it holds:

$$\sum_{i,j} \frac{\partial}{\partial p_k} \gamma(h_i;\hat{p}_1(g),\ldots,\hat{p}_d(g)) w_{ij} (\gamma(h_j;\hat{p}_1(g),\ldots,\hat{p}_d(g)) - g_j) = 0$$

deriving this for g_l yields:

$$\sum_{i,j}\frac{\partial}{\partial p_k}\gamma(h_i;\hat{p}_1(g),...,\hat{p}_d(g))w_{ij}\left(\sum_{m=1}^{d}\frac{\partial}{\partial p_m}(\gamma(h_j;\hat{p}_1(g),...,\hat{p}_d(g))\frac{\partial}{\partial g_l}\hat{p}_m(g)-\delta_{jl}\right)$$

$$+\sum_{i,j}\sum_{m=1}^{d}\frac{\partial^2}{\partial p_k\partial p_m}\gamma(h_i;\hat{p}_1(g),...,\hat{p}_d(g))\frac{\partial}{\partial g_l}\hat{p}_m(g)w_{ij}\left(\gamma(h_j;\hat{p}_1(g),...,\hat{p}_d(g))-g_j\right)=0$$

or in matrix notation:

$$\mathrm{A^tW(AD-I)+UD=0}$$

with:

$$A:=\left(\frac{\partial}{\partial p_k}\gamma(h_i;\hat{p}_1,...;\hat{p}_d)\right)_{ik},\ I=(\delta_{ij})_{ij}$$

$$a_k:=\left(\frac{\partial}{\partial p_k}\gamma(h_i;\hat{p}_1,...;\hat{p}_d)\right)_i,\ \gamma:=\left(\gamma(h_j,p_1,...,p_d)\right)_j,\ g:=(g_j)$$

$$H_{kmi}:=\left(\frac{\partial^2}{\partial p_k\partial p_m}\gamma(h_i;\hat{p}_1(g),...,\hat{p}_d(g))\right)_{mi},\ D_{lm}:=\left(\frac{\partial}{\partial g_l}\hat{p}_m(g)\right)_l$$

$$U:=\sum_{ij}H_{kmi}w_{ij}(\gamma_i-g_i)_{km}$$

and thus

$$(A^tWA+U)D-A^tW=0$$

which results in an explicit formula for the derivative of the estimators $\hat{p}(g)$ for the semivariogram values g_i.

$$D=(A^tWA+U)^{-1}A^tW \qquad (4)$$

Note that this formula is slightly different from the corresponding formula in linear models, which lacks the curvature term U:

$$D=(A^tWA)^{-1}A^tW \qquad (5)$$

This second linearization also corresponds to the linearization in the point g=γ with the reestimated values as data. It is therefore preferable, when we really believe the model we are fitting, since the observed values have larger error. On the other hand when the model is fitted to an empirical variogram, which might have some systematic deviations from the model, it could be better to believe the class means more than the refitted values. It is thus a philosophical question whether to choose formula (4) or (5). Personally I would always prefer the one yielding a larger variance, to be on the save side.

Now using a first order the Taylor series approximation

$$\hat{p}(g)\approx\hat{p}(g)+D(g-\hat{g})$$

the variance of $\hat{p}(g)$ can be estimated by error propagation from the estimated variance-covariance structure of the empirical variogram by

$$\hat{\mathrm{var}}(\hat{p}(g)):=D(\hat{\mathrm{cov}}(\hat{\gamma}(h_i),\hat{\gamma}(h_j)))D^t$$

and the correspondingly the variance of $\gamma(h,\hat{p}_1,...,\hat{p}_d)$ by

$$\hat{\mathrm{var}}(\gamma(h,\hat{p}_1,...,\hat{p}_d)):=v(h)^tD(\hat{\mathrm{cov}}(\hat{\gamma}(h_i),\hat{\gamma}(h_j)))D^tv(h)$$

with

$$v(h)^t:=\left(\frac{\partial}{\partial p_k}\gamma(h,p_1,...,p_d)\right)_k$$

Comparison to distribution based approaches

Distribution based methods in general perform better or much better, when their assumptions are met. Especially for the variogram distributions based methods calculate the fourth order moment structure from an estimated second order moment structure, while this methods directly uses the fourth order

moment structure to estimate fourth order moments of the process. Normally an estimate based on a lower order moment is much more stable and efficient that the estimate based on the higher order moment. Anyway the conclusion from the second order moment structure to the fourth order moment structure is purely based on the distributional assumption which might or might not be true. Thus parametric methods will be clearly superior as long as their assumptions are met or at least not to badly violated. Otherwise, when their assumptions are clearly violated, their outcome is arbitrary and we need to use the nonparametric method.

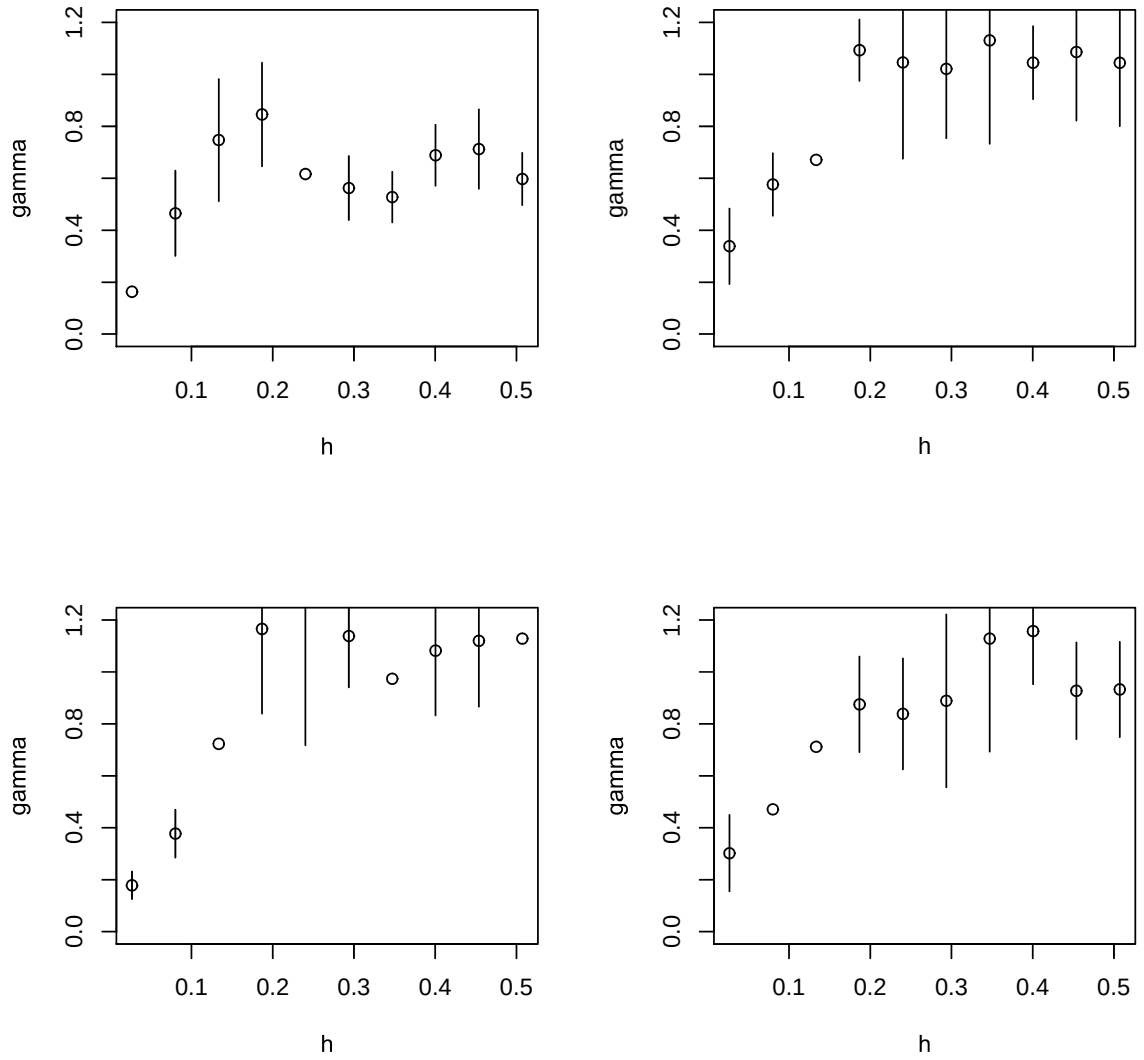

Figure 1: Example calculations: The dots represent the estimated values of the semivariogram and the lines the estimated standard deviation. Points without lines correspond to negative variance estimates. The estimator is not guaranteed to be positive.

Example Calculation

Example calculations for the empirical variogram are shown in figure 1. They all correspond to the same set of 100 measurement locations uniformly distributed in the unit square and a simulated gaussian random field with spherical covariance function with sill=1, range=0.2 and nugget=0.

Although the variance the estimated variances seem to express the variability in the between the four pictures well, the estimate is for only 100 points still not stable and eventually gets negative.

Conclusions

The method presented here does not rely on second order stationarity, known distributions, or estimated covariances. The basic assumption is that we know or estimate a distance in which correlation vanishes or nearly vanishes. Since no variogram modelling step is required this methods suits for the estimation of the precision of the variogram itself.

On the other hand as a non-parametric method the method needs a large amount of data for good results. It should therefore always be used together with one of the distribution based methods to detect and handle situations of nongaussianity. However reliable results are to be expected only for relatively small stochastic ranges and a larger amount of data.

References

Bogaert, P., Russo, D., 1999, Optimal spatial sampling design for the estimation of the variogram based on least squares approach: Water Resources Research, v. 35, no. 4, p. 1275-1289

Boogaart, K.G. v.d., 2002, Estimating the variance of the spatial mean, submitted to Mathematical Geology.

Gento, M.G., 1998, Variogram fitting by generalized least squares using explicit formula for the covariance structure, Mathematical Geology, v. 30, no. 4, p. 323-345

Pardo-Iguzquiza, E., 1998, Maximum likelihood estimation of spatial covariance parameters, Mathematical Geology, v. 30, no. 1, p. 95-108

Pardo-Iguzquiza, E. and Dowd, P.A., 2001, The variance-covariance matrix of the experimental variogram: assessing variogram uncertainty, Mathematical Geology, v 33, no 4, p. 397-419.