Empirical variogram of the underlying Gaussian fields in the plurigaussian models

Nicolas Desassis, D Renard, Hélène Beucher

Empirical variogram of the underlying Gaussian fields in the plurigaussian models

N. Desassis[1], D. Renard[1], and H. Beucher[1]

[1]Centre de Géosciences - MINES-ParisTech

Abstract

The plurigaussian model is particularly suited to describe categorical regionalized variables. Starting from a simple principle, the thresholding of one or several Gaussian random fields (GRFs) to obtain categories, the plurigaussian model is well adapted for a wide range of situations. By acting on the form of the thresholding rule and/or the threshold values (which can vary along space) and the variograms of the underlying GRFs, one can generate many spatial configurations for the categorical variables. One difficulty arising with the use of this model is to choose variogram model for the underlying GRFs. Indeed, these latter are hidden by the truncation and we only observe the simple and cross-variograms of the category indicators. In this paper, we propose a method based on the pairwise likelihood to estimate the empirical variogram of the GRFs. It provides an exploratory tool in order to choose a suitable model for each GRF and later to estimate its parameters. We illustrate the efficiency of the method with a Monte-Carlo simulation study. The method presented in this paper is implemented in the R package RGeostats.

Keywords: Plurigaussian models ; empirical variography ; pairwise likelihood (PL) ; underlying Gaussian Random Fields (GRFs)

1 Introduction

Regionalized categorical variables often appear in several scientific domains. For instance, in the earth sciences, some continuous soil properties (e.g the permeability, the grade of an element, ...) can be better described by first categorizing the rock types into lithofacies (or facies) which have a certain homogeneity regarding to the studied variable. Then, the continuous variables are studied separatly in each category. In the scope of conditional simulations, the lithofacies are first simulated conditionally to the observed lithofacies, then the continuous variables are simulated inside each simulated category according to their associated spatial distribution (see

for instance Dubrule, 1993). To model and simulate a categorical random field, the plurigaussian model is particularly appealing. Starting from a simple principle, the truncation of one (Matheron et al., 1987) or several Gaussian Random Fields (Le Loc'h and Galli, 1996; Le Loc'h et al., 1994), it allows to reproduce a wide range of patterns. Applications of the plurigaussian model can be found for mineral resources evaluation (Talebi et al., 2015), in hydrology (Mariethoz et al., 2009). In petroleum, some authors use the plurigaussian models in link with history matching (Hu, 2000; Liu and Oliver, 2004; Romary, 2010). In this paper, we suppose that we directly observe the categorical variable.

When the underlying Gaussian Random Fields (GRFs) are supposed to be stationary, two ingredients are necessary to fully specify the plurigaussian model: the coding function (or truncation rule) which defines the sets associated to each category and which can vary along the space and the multivariate covariance function.

Concerning the coding function, some authors use a simple parametric form, for instance a cartesian product of intervals and they allow the threshold values to vary along the space. It is often the case vertically through the vertical proportion curves (Felletti, 2004) but also laterally, for instance when one want to use auxiliary information as seismic data in the model. Other authors concentrate on the estimation of more complex coding functions constant in space (Astrakova and Oliver, 2014). Finally, Allard et al. (2012) estimate complex coding function varying in space by using auxiliary information.

As mentionned by Mariethoz et al. (2009), one of the main difficulty arising from the use of the plurigaussian model is the inference of the variogram models of the underlying GRFs. Indeed, the available empirical variograms are the variograms of the indicator functions of the categories (one simple variogram per category and the cross-variograms for all the bivariate combinations) while the variograms required by the model are the variograms of the underlying GRFs whose realizations are hidden by the truncation.

Until now, most of the methods to estimate the variogram of the underlying GRFs rely on the indicator variograms. For instance Mariethoz et al. (2009) determine the variogram model of the underlying GRFs by using simulations. More precisely, they choose a parametric model for the underlying GRFs and they compute the parameters value such as the indicator simple variograms of the simulations are the closest to the data indicator variograms. The optimization is performed with simulated annealing. Armstrong et al. (2011) exploits the mathematical relationships between the underlying GRFs variogram models, the coding function and the indicator simple and cross-variograms. Some industrial softwares (as Isatis®, 2014) also use these relations and the users have to choose the parameters of the variogram models of the underlying GRFs by visual inspection of the resulting indicator variograms. It is made by-trial-and-error (Galli et al., 1994). Emery (2007) performs the numerical integration of the Gaussian density by using its expansion into the normalized Hermite polynomials. All these methods are rather tedious as they have a high computational cost or require a lot of trials. Dowd et al. (2003) and Xu et al. (2006) propose to find the range parameters automatically by minimizing a squared differences with a grid-search but the choice of the covariance models of the underlying GRFs remains arbitrary and limited.

In this paper, we will supppose that the coding function is known and we will concentrate on the estimation of the variograms of the underlying GRFs. We propose an original methodology based on the pairwise likelihood (PL) maximization principle to directly compute the empirical variograms of the underlying and hidden GRFs. More precisely, we consider the variogram at a given distance, or a given vector for a directional variogram, as a parameter of the model and we maximize the PL by selecting only the pairs of points approximately separated by this distance or vector. We iterate this calculation on all distances (respectively vectors). Thereby, we obtain an empirical variogram which helps the user to choose a suitable valid model that can then be fitted by least squares or estimated with a likelihood based approach. Then, the simple and cross-variograms in the indicator scale can be deduced and compared to the empirical variograms of the indicators to check the quality of the resulting models.

In the first part, we will give the main notations of the paper and we will recall the definition of the plurigaussian model. In section 3.1, the relationships between variograms of GRFs and variograms of indicators are recalled for comparison purposes. Then we present our method in section 3.2. First, we describe the general principle which should make possible the estimation of a complex multivariate spatial model. Then we describe with more details the implementation in the case where the underlying GRFs are supposed to be independent. To assess the efficiency of the method and to evaluate the uncertainty associated to the variogram estimation, a Monte-Carlo study is performed and its results are summarized in section 4. Our results are discussed in a conclusion where some perspectives are given.

2 The data model

2.1 General formulation of the plurigaussian model

Let $\mathcal{F} = \{f_1, \ldots, f_K\}$ a finite set with K categories. For a set of n sites $\{x_i\}_{1 \leq i \leq n}$ of a domain $\mathcal{D} \subset \mathbb{R}^d$, we observe $\mathbf{f} = (f(x_1), \ldots, f(x_n))$ a $\mathcal{F}$-valued vector. We suppose that for a given location $x \in \mathcal{D}$, the value $f(x)$ is the realization of a $\mathcal{F}$-valued random variable $F(x)$.

To characterize the spatial distribution of $F(.)$, we use the plurigaussian model. To describe this model, we adopt the same formulation as Armstrong et al. (2011). Let

$$\mathbf{Y}(.) = \{\mathbf{Y}(x), x \in \mathcal{D}\}$$

a q-variate centered and standardized GRF on $\mathcal{D}$. In other words, for all $x \in \mathcal{D}$, $\mathbf{Y}(x) = (Y_1(x), \ldots, Y_q(x))$ is a random vector with q components and for all $N \in \mathbb{N}^\star$ and for all $(x_1, \ldots, x_N) \in \mathcal{D}^N$, the $N \times q$-vector

$$(\mathbf{Y}(x_1), \ldots, \mathbf{Y}(x_N)) = (Y_1(x_1), \ldots, Y_1(x_N), \ldots, Y_q(x_1), \ldots, Y_q(x_N))$$

is a standard Gaussian vector with $E[Y_r(x_i)] = 0$ and $\mathrm{Var}[Y_r(x_i)] = 1$ for all $r \in [\![1, q]\!]$ and $i \in [\![1, N]\!]$. In this paper, we will suppose that $\mathbf{Y}(.)$ is a second-order

stationary multivariate function. In other words, there exists a matricial cross-covariance function $\mathbf{C}$ such as $\text{Cov}(Y_r(x), Y_s(x')) = \mathbf{C}_{rs}(x'-x)$ for $(r,s) \in [\![1,q]\!]^2$ (see Wackernagel, 2003, for an introduction on multivariate spatial random functions).

Let $\mathcal{C}$ a coding function on $\mathcal{D}$ such as, for all $x \in \mathcal{D}$, $\mathcal{C}(x) = (\mathcal{C}_1(x), \ldots, \mathcal{C}_K(x))$ where, for $k \in [\![1,K]\!]$, the subsets $\mathcal{C}_k(x)$ form a (measurable) partition of $\mathbb{R}^q$. The model is defined by the following equivalence

$$F(x) = f_k \text{ if and only if } \mathbf{Y}(x) \in \mathcal{C}_k(x). \tag{1}$$

Note that the formulation given by (1) provides a quite general class of models. Indeed, it also contains the models defined by

$$F(x) = f_k \text{ if and only if } \varphi(\mathbf{Y}(x)) \in \tilde{\mathcal{C}}_k(x)$$

for any surjective function φ from $\mathbb{R}^q$ to any set E where the sets $\tilde{\mathcal{C}}_k(x)$ for $[\![1,K]\!]$ form a partition of E. The subsets $\varphi^{-1}(\tilde{\mathcal{C}}_k(x))$ have to be some measurable sets of $\mathbb{R}^q$. This remark aims to highlight the fact that the marginal gaussianity of the random variables $Y_r(x)$ is arbitrary. Nevertheless, the multi-gaussian assumption is a convenient way to describe the spatial multivariate relationships of the underlying random function. It also provides a multivariate random function easy to simulate (see e.g Lantuejoul, 2002).

We will note $c(x)$ the set defined as:

$$c(x) = \mathcal{C}_k(x)$$

where $k \in [\![1,K]\!]$ is the index of the category at location x. In other words, $f(x) = f_k$. In all the sequel, we will suppose that the classes $\mathcal{C}_k(x)$ are known.

3 Estimating the spatial structure

3.1 Indicators cross-variograms based methods

As already mentionned in the introduction, most of the methods to choose the simple and cross-covariance models of the underlying GRFs rely on the indicator simple and cross-variograms or covariances. Armstrong et al. (2011) or Isatis® (2014) use the mathematical relationships between the simple and cross-covariances (or variograms) of the GRFs and the simple and cross-covariances (or variograms) of the indicators of each category. In the current paper, we only use these relationships to check the quality of the results given by our proposed method. We recall these relationships below. For that purpose, we will note the random indicator function of the category $f_k \in \mathcal{F}$ as follows:

$$\mathbf{1}_{f_k}(x) = \begin{cases} 1 & \text{if } F(x) = f_k \\ 0 & \text{otherwise} \end{cases}$$

and $1_{f_k}(x)$ the associated true value.

3.1.1 Variogram between two points

For $1 \leq k, l \leq K$, one can define the cross-variogram between indicators of facies k and l, between two locations x and x' of $\mathcal{D}$:

$$\gamma_{kl}(x, x') = \frac{1}{2} E[(\mathbf{1}_{f_k}(x') - \mathbf{1}_{f_k}(x))(\mathbf{1}_{f_l}(x') - \mathbf{1}_{f_l}(x))]$$

When $k = l$, we have

$$\gamma_{kk}(x, x') = \frac{E[\mathbf{1}_{f_k}(x)] + E[\mathbf{1}_{f_k}(x')]}{2} - E[\mathbf{1}_{f_k}(x')\mathbf{1}_{f_k}(x)] \tag{2}$$

When $k \neq l$, we have

$$\gamma_{kl}(x, x') = -\frac{E[\mathbf{1}_{f_k}(x')\mathbf{1}_{f_l}(x)] + E[\mathbf{1}_{f_l}(x')\mathbf{1}_{f_k}(x)]}{2} \tag{3}$$

We note Σ_x and $\Sigma_{x,x'}$ the respective correlation matrices of the vectors $\mathbf{Y}(x)$ and $(\mathbf{Y}(x), \mathbf{Y}(x'))$. Furthermore, $g_{\Sigma}^{(q)}(\mathbf{u})$ stands for the centered and standardized Gaussian density of dimension q and correlation matrix Σ computed for the q-vector $\mathbf{u}$.

With these notations, we can establish the link between $\gamma_{kl}(x, x')$ and the correlations between the underlying GRFs. Indeed, the expectation of the indicator of facies k (which corresponds to its proportion at location x) is equal to:

$$E[\mathbf{1}_{f_k}(x)] = \int_{\mathcal{C}_k(x)} g_{\Sigma_x}^{(q)}(\mathbf{u}) d\mathbf{u} \tag{4}$$

and

$$E[\mathbf{1}_{f_k}(x)\mathbf{1}_{f_l}(x')] = \int_{\mathcal{C}_k(x)} \int_{\mathcal{C}_l(x')} g_{\Sigma_{x,x'}}^{(2q)}((\mathbf{u}, \mathbf{v})) d\mathbf{u} d\mathbf{v} \tag{5}$$

where each integration symbol represents an integration over a q-dimensional space. These integrations and all the others mentionned in the current paper are integral of the Gaussian probability density function. They can be computed numerically with the efficient algorithm proposed by Genz (1992).

Note that it is sometimes useful to work with the non-centered covariances $E[\mathbf{1}_{f_k}(x)\mathbf{1}_{f_l}(x')]$ which can be computed in the same way. Indeed, it has the advantage to capture asymetry in the model.

When the GRFs are stationary and the coding function $\mathcal{C}$ is constant over $\mathcal{D}$, the variograms of all the involved random functions only depend on the lag between the points. Therefore, we can deduce the simple and cross-variograms of the indicator for a given lag from the variograms value of the underlying GRFs by using formulas (2), (3), (4) and (5). However, when the coding function varies over $\mathcal{D}$, the theoretical simple and cross-variograms of the indicators for a given lag don't exist anymore. Nevertheless, it is still possible to compute the associated empirical variograms and compare them with an averaged version of the variograms between two points computed in the indicators domain as described below.

3.1.2 Variogram for a specific lag

For a given vector $\mathbf{h} \in \mathbb{R}^d$, we will note $(i,j) \in \mathcal{V}(\mathbf{h})$ when $x_j - x_i \simeq \mathbf{h}$, i.e when the pair (x_i, x_j) should be used to compute the empirical variogram for lag $\mathbf{h}$ (see e.g Chilès and Delfiner, 2012, for details). $N(\mathbf{h})$ stands for the number of pairs in $\mathcal{V}(\mathbf{h})$

$$\hat{\gamma}_{kl}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{(i,j)\in\mathcal{V}(\mathbf{h})} (1_{f_k}(x_j) - 1_{f_k}(x_i))(1_{f_k}(x_j) - 1_{f_k}(x_i)).$$

and try to fit them with the variogram models of indicators associated to the observation locations $\{x_1, \ldots, x_n\}$ defined for $(k,l) \in [\![1, K]\!]^2$ by:

$$\gamma_{kl}(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{(i,j)\in\mathcal{V}(\mathbf{h})} \gamma_{kl}(x_i, x_j). \tag{6}$$

Its behaviour results from the spatial characteristics of $\mathbf{Y}(.)$ (defined through its multivariate cross-covariance function in the stationary case) and from the spatial variability of the set functions $\mathcal{C}_k$.

3.2 Pairwise likelihood maximization

In this part, we describe a new methodology to perform the multivariate empirical variography of the underlying gaussian random functions from the category observations. This methodology is based on the pairwise likelihood (PL) maximization. We first recall the principle of the more general composite likelihood based approach. Then we show how to apply it for the plurigaussian model. Finally, we describe more precisely the algorithm in two particular cases: the monogaussian case ($q = 1$) and the plurigaussian case in which the q-Gaussian random functions are independent and the sets $c(x)$ are cartesian products of real subsets.

3.2.1 General presentation of the methodology

The PL approach belongs to the family of the composite likelihood methods (see e.g. Varin et al., 2011, for a comprehensive review). It is generally used to estimate a parameters vector θ of a statistical model, for instance when the usual maximization of the full likelihood is computationally cumbersome. In these cases, the full likelihood is replaced by a weighted product of marginal or conditional likelihoods. Lindsay (1988) defines the composite likelihood as follows: if W is a random vector of size m with multivariate density $f(w;\theta)$ and $\{\mathcal{A}_1, \ldots, \mathcal{A}_k\}$ is a set of marginal or conditional events with associated likelihoods $\mathcal{L}_k(\theta; w) \propto f(w \in \mathcal{A}_k; \theta)$, the composite likelihood is the weighted product

$$\mathcal{L}_C(\theta; w) = \prod_{d=1}^{D} \mathcal{L}_d(\theta; w)^{\lambda_d}$$

where λ_d are nonnegative weights to be chosen.

Here we focus on the PL which corresponds to the particular case in which

$$\mathcal{L}_d(\theta; w) = f(w_i, w_j; \theta)$$

are the bivariate densities for all $(i,j) \in [\![1,m]\!]$.

One of the advantages of the composite likelihood based approach is that they enable to estimate only some components of θ. For instance, in the plurigaussian model, we would like to estimate $\Sigma(\mathbf{h}_\alpha)$, the cross-covariance matrices of the vectors

$$(Y_1(x), Y_1(x+\mathbf{h}_\alpha), \ldots, Y_q(x), Y_q(x+\mathbf{h}_\alpha))$$

for a set of n_l separation vectors $\mathbf{h}_\alpha, \alpha \in [\![1,n_l]\!]$. We consider that $\theta = (\Sigma(\mathbf{h}_\alpha))_{1\leq\alpha\leq n_l}$ is the set of parameters. Then, we group pairs of sites according to their separation vector in the same way as the empirical variogram computation and we write the log PL as follows:

$$\mathcal{L}_d(\theta; \mathbf{f}) = \sum_{\alpha=1}^{n_l} \sum_{(i,j)\in\mathcal{V}(\mathbf{h}_\alpha)} \log\, p_{ij}^{(q)}(\Sigma(\mathbf{h}_\alpha)) \tag{7}$$

where

$$p_{ij}^{(q)}(\Sigma) = \int_{c(x_i)} \int_{c(x_j)} g_\Sigma^{(2q)}((\mathbf{u},\mathbf{v}))d\mathbf{u}d\mathbf{v}$$

is the probability that $F(x_i) = f(x_i)$ and $F(x_j) = f(x_j)$ when the cross-covariance matrix of the vector $(Y_1(x), Y_1(x+\mathbf{h}_\alpha), \ldots, Y_q(x), Y_q(x+\mathbf{h}_\alpha)$ is Σ. Note that the weights λ_d attached to a pair (i,j) have been set to 1 if there exists $\alpha \leq n_l$ such as the pairs belongs to $\mathcal{V}(\mathbf{h}_\alpha)$ and to 0 otherwise.

Then, the maximum PL estimator is obtained by maximizing $\mathcal{L}_d$ with respect to all the matrices $\Sigma(\mathbf{h}_\alpha)$. Note that to satisfy the stationarity of the resulting model, the condition

$$\mathrm{Cov}(Y_r(x_i), Y_s(x_i)) = \mathrm{Cov}(Y_r(x_j), Y_s(x_j))$$

is required for all locations x_i and x_j and all variable indices $(r,s) \in [\![1,q]\!]^2$. It implies that the $2q \times 2q$-matrices $\Sigma(\mathbf{h}_\alpha)$ belong to the set noted $\mathcal{S}_{2q}$ and defined by

$$B \in \mathcal{S}_{2q} \Leftrightarrow b_{2r,2s} = b_{2r-1,2s-1}$$

for all $(r,s) \in [\![1,q]\!]^1$, where $b_{r,s}$ stands for the $(r,s)^{\text{th}}$ element of the matrix B.

Furthermore, it is important to remark that all the matrices $\Sigma(\mathbf{h}_\alpha)$ share some common terms to estimate, the ones corresponding to $\mathbf{C}_{rs}(0)$. These two constraints on the global solution make the problem numerically difficult to solve. For this reason, we will focus on the simplified cases where the q GRFs are independent. This assumption is generally made in most of the applications of the plurigaussian model.

With this assumption, we have $\mathbf{C}_{rs}(0) = 0$ for all $(r,s) \in [\![1,q]\!]^2$ with $r \neq s$, so the matrices $\Sigma(\mathbf{h}_\alpha)$ does not share some common terms to estimate simultaneously.

It results that the maximization of the log PL can be achieved by solving $q \times n_l$ simpler maximization problems:

$$\hat{\Sigma}(\mathbf{h}_\alpha) = \arg\max_{\Sigma\in\mathcal{E}_{2q}} \sum_{(i,j)\in\mathcal{V}(\mathbf{h}_\alpha)} \log\, p_{ij}^{(q)}(\Sigma) \tag{8}$$

where $\mathcal{E}_{2q}$ is the subset of S_{2q} such as all the terms $\mathrm{Cov}(Y_r(x_i), Y_s(x_j))$ are equal to 0 as soon as $r \neq s$.

We show more precisely how to apply this method in sections 3.2.2 and 3.2.3.

3.2.2 Application in the monogaussian case

In this part, we assume that $q = 1$. The only quantity to estimate for a given lag $\mathbf{h}_\alpha$ is the spatial correlation of the underlying univariate Gaussian random function $\rho(\mathbf{h}_\alpha) = \mathrm{Cor}(Y(x), Y(x + \mathbf{h}_\alpha))$ (or equivalently $\gamma(\mathbf{h}_\alpha) = 1 - \rho(\mathbf{h}_\alpha)$).

The estimator of $\rho(\mathbf{h}_\alpha)$ by PL maximization is obtained by

$$\rho^\star(\mathbf{h}_\alpha) = \underset{\rho \in]-1,1[}{\arg\max} \sum_{(i,j) \in \mathcal{V}(\mathbf{h}_\alpha)} \log p_{ij}^{(1)}(\rho) \tag{9}$$

where

$$p_{ij}^{(1)}(\rho) = \int_{c(x_i)} \int_{c(x_j)} g_\rho^{(2)}(u, v) du dv.$$

In this formula, each integral symbol represents a single integral and $g_\rho^{(2)}(u, v)$ is a simplified notation standing for the centered standardized bi-Gaussian density with correlation coefficient ρ computed for $(u, v) \in \mathbb{R}^2$.

Hence, the PL maximization problem (8) is reduced to a one dimensional optimization problem over a bounded interval. Therefore, it can easily be solved, for instance with the golden section search algorithm (Press et al., 2007).

3.2.3 Generalization to $q \geq 2$

In this part, we assume that the random functions $Y_1(.), \ldots, Y_q(.)$ are independent. In other words, for any $(r, s) \in [\![1, q]\!]^2$ and any x and x' of $\mathbb{R}^d$, we have

$$\mathbf{C}_{rs}(x - x') = 0$$

as soon as $r \neq s$.

So the covariance matrices $\Sigma(\mathbf{h}_\alpha)$ are block diagonal and contain q two-dimensional blocks corresponding to the correlation matrices of the sub-vectors $(Y_r(x), Y_r(x + \mathbf{h}_\alpha))$, $r \leq q$. If we note for all $r \in [\![1, q]\!]$

$$\rho_r(\mathbf{h}_\alpha) = \mathrm{Cor}(Y_r(x), Y_r(x + \mathbf{h}_\alpha)),$$

it results that

$$g_{\Sigma(\mathbf{h}_\alpha)}^{(2q)}(\mathbf{u}, \mathbf{v}) = \prod_{r=1}^{q} g_{\rho_r(\mathbf{h}_\alpha)}^{(2)}(u_r, v_r)$$

where $\mathbf{u} = (u_1, \ldots, u_q)$ and $\mathbf{v} = (v_1, \ldots, v_q)$.

Furthermore, we assume that all the sets $\mathcal{C}_k(x)$ are cartesian products of subsets of $\mathbb{R}$:

$$\mathcal{C}_k(x) = \bigtimes_{r=1}^{q} T_k^r(x)$$

with $T_k^r(x) \subset \mathbb{R}$.

We will note $t_r(x) = T_k^r(x)$ where k is such that $f(x) = f_k$ is the actual category at site x.

Hence,

$$p_{ij}^{(q)}(\Sigma(\mathbf{h}_\alpha)) = \prod_{r=1}^{q} \int_{t_r(x_i)} \int_{t_r(x_j)} g^{(2)}_{\rho_r(\mathbf{h}_\alpha)}(u,v)dudv.$$

In other word, each $\rho_r(\mathbf{h}_\alpha)$ is estimated by:

$$\rho_r^\star(\mathbf{h}_\alpha) = \underset{\rho\in]-1,1[}{\operatorname{argmax}} \sum_{(i,j)\in\mathcal{V}(\mathbf{h}_\alpha)} \log \int_{t_r(x_i)} \int_{t_r(x_j)} g_\rho^{(2)}(u,v)dudv.$$

which is equivalent to solve q problems similar to the problem (9) presented in section 3.2.2.

4 Simulation results

In this section, we present two simulation studies to assess the efficiency of the proposed method.

4.1 $q = 1$ and $\mathcal{C}$ is constant

On a 1-dimensional regular grid with mesh size 1 and 2000 nodes, 1000 realizations of a GRF $Y(.)$ with covariance function

$$C(h) = e^{-h^2/40^2}$$

have been drawn. For each realization, $y(.)$, one category among the set $\mathcal{F} = \{\text{black}, \text{red}, \text{green}\}$ is assigned to each node x of the grid according to the following rule:

$$f(x) = \begin{cases} \text{black} & \text{if} \quad y(x) \in \mathcal{C}_1(x) = (-\infty, s_1) \\ \text{red} & \text{if} \quad y(x) \in \mathcal{C}_2(x) = (s_1, s_2) \\ \text{green} & \text{if} \quad y(x) \in \mathcal{C}_3(x) = (s_2, +\infty) \end{cases}$$

where $s_1 = -s_2$ are chosen such as the probability that $P(Y(x) \in \mathcal{C}_i(x)) = \frac{1}{3}$ for all $i = 1, 2, 3$. On figure 1, one realization of the resulting categories is displayed. We also represent the underlying realization of the GRF, only known here as a by-product of the simulation workflow.

The empirical variogram of the underlying GRF is computed by pairwise likelihood from the categories as described in section 3.2.2, for 150 distances ranging regularly from 1 to 150. For comparison purpose, the traditional empirical variogram has been computed directly on the realizations $y(.)$ for the same set of distances. The results are summarized on figure 2 (a) for the PL computed from the categories and 2(b) for the traditional empirical variogram computed from the realizations of the GRF. The average over all the simulations display a negligible

bias. As expected, the variability of the estimator increases with the distance. The variogram seems to be better estimated when computed from categories by PL than with the original Gaussian values despite the loss of information due to the truncation. The reason is that we provide additional information by fixing the sill to 1 in the computation by PL.

4.2 $q = 1$ and $\mathcal{C}$ is not constant

We use the same simulation scheme as in section 4.1 except that the covariance model of the GRF is now given by:

$$C(h) = e^{-h/20}$$

and the categories are assigned to each node x of the grid according to the following rule:

$$f(x) = \begin{cases} \text{black} & \text{if} \quad y(x) \in \mathcal{C}_1(x) = (-\infty, s_1(x)) \\ \text{red} & \text{if} \quad y(x) \in \mathcal{C}_2(x) = (s_1(x), s_2(x)) \\ \text{green} & \text{if} \quad y(x) \in \mathcal{C}_3(x) = (s_2(x), +\infty) \end{cases}$$

where $s_1(x)$ and $s_2(x)$ have been simulated once for all simulations. The figure 3 displays one realizations of this process with the two functions $s_1(.)$ and $s_2(.)$. The figure 4 shows results which are similar to the constant coding function case of section 4.1.

4.3 $q = 2$, $\mathcal{C}$ is a constant cartesian product of intervals

In this example, we consider the same categories as previously. They are generated by using two independent GRFs $Y_1(.)$ and $Y_2(.)$ with respective covariance functions

$$C_1(h) = e^{-100h^2}$$

and

$$C_2(h) = e^{-20h}$$

The categories are assigned to a point x according to the following rule:

$$f(x) = \begin{cases} \text{black} & \text{if} \quad y(x) \in \mathcal{C}_1(x) = (s_1, +\infty) \times \mathbb{R} \\ \text{red} & \text{if} \quad y(x) \in \mathcal{C}_2(x) = (-\infty, s_1) \times (-\infty, t_1) \\ \text{green} & \text{if} \quad y(x) \in \mathcal{C}_3(x) = (-\infty, s_1) \times (t_1, +\infty) \end{cases}$$

where $s_1 = t_1 = 0$ such that
$P((Y_1(x), Y_2(x) \in \mathcal{C}_1(x)) = \frac{1}{2}$ and $P((Y_1(x), Y_2(x) \in \mathcal{C}_i(x)) = \frac{1}{4}$ for $i = 2, 3$.
A scheme of this coding function is displayed fig. 5.

Then 1000 simulations are performed on 800 locations chosen uniformely on the square $[0, 1] \times [0, 1]$ one time for all the simulations. A realization is displayed fig. 5.

The results are summarized fig.6 and again, they are rather good compared to the empirical variograms computed directly from the Gaussian data. Note that for each simulation, the computation of the empirical variogram of the second Gaussian from $y(.)$ has been computed by using only the subset of locations for which the first Gaussian is greater than 0.

5 Discussion

In this paper, we propose to use the pairwise likelihood principle to estimate empirical variograms of the underlying GRFs in the plurigaussian model. The explicit use of a composite likelihood based approach as an exploratory data analysis tool seems original. Note that some existing tools as the classical empirical variogram (Matheron, 1962) can be viewed as a maximum of a composite likelihood. Indeed, let consider the marginal likelihood based on pairwise differences as a particular case of composite likelihood (see for instance Curriero and Lele, 1999, for parametric estimation of the variogram). We can easily show that the usual estimator of the semivariogram at a given distance maximizes this quantity under a bigaussian assumption. Although, it is not explicity stated and the fundations of these methods are rather based on moments considerations.

Once the empirical variograms of the underlying GRFs has been computed, we can use it to choose a valid variogram model which can be fitted by least squares, for instance by using the algorithm proposed in Desassis and Renard (2013) or estimated by a likelihood based method. The likelihood will probably remain intractable since it involves an integral on $\mathbb{R}^n$ where n is the number of samples. A composite likelihood based approach should be used instead. Again, the PL seems well suited.

To conclude, note that the method presented in this paper is implemented in the R-package RGeostats (Renard et al., 2015) in the function named *vario.pgs*. Some demonstration scripts are provided through a tutorial on the dedicated website.

Further researches will concentrate on the generalization of the approach presented in this paper to the case where no independence assumption is made between the underlying GRF. Indeed, one can model more complex transitions between categories with more general multivariate spatial models (see Galli et al., 2006). In that case, one have to estimate all the elements (except the diagonal) of the correlation matrices $\Sigma(\mathbf{h}_\alpha)$ with the constraints mentionned in section 3.2.1. This is computationally much more challenging.

Finally, the PL likelihood approach to compute empirical variograms seems to be a promising idea which could be applied to other similar context of hidden variable, or variable known after a transformation.

To cite some of them:

- compute the empirical variogram of the underlying GRF in the hierarchical geostatistical models (see e.g Diggle et al., 1998). Some authors have already proposed a way to compute empirical variogram of underlying random fields in hierarchical models: Oliver et al. (1993) treats the binomial case, Monestiez et al. (2006) the poisson case. However, these estimators are based on the method of moments and the distribution of the underlying random function is

not specified. Thus, the underlying intensity can only be predicted by kriging but they can not be simulated. An approach based on the PL in a distribution based framework could be a good alternative;

- perform the multivariate empirical variography of the underlying GRFs when one have to deal with a continuous variable vs. discrete variable (Emery and Silva, 2009), or even two discrete variables (Renard et al., 2008);
- compute the empirical variogram of a variable at punctual level when the observations are some regularizations with different supports.

Acknowledgments: We are very grateful to Christian Lantuejoul for the idea to directly compute the empirical variogram of the underlying GRF. We would also like to thank Geovariances and the School of Earth Science of the University of Queensland to have partly funded this research during the visit of the first author in Brisbane in 2012/2013.

References

Allard, D., D. D'Or, P. Biver, and R. Froidevaux (2012). Non-parametric diagrams for pluri-gaussian simulations of lithologies. In *9th international geostatistical congress*, Oslo, Norway, pp. 11–15.

Armstrong, M., A. Galli, H. Beucher, G. Le Loc'h, D. Renard, B. Doligez, R. Eschard, and F. Geffroy (2011). *Plurigaussian simulation in geosciences*. Berlin: Springer-Verlag.

Astrakova, A. and D. S. Oliver (2014). Truncation map estimation for the truncated bigaussian model based on univariate unit-lag probabilities. In *Proceedings of ECMOR 14th European Conference on the Math. of Oil Recovery*.

Chilès, J. and P. Delfiner (2012). *Geostatistics : Modeling spatial uncertainty, 2nd edn.* New York: Wiley.

Curriero, F. C. and S. Lele (1999). A composite likelihood approach to semivariogram estimation. *Journal of Agricultural, Biological, and Environmental Statistics 4*(1), 9–28.

Desassis, N. and D. Renard (2013). Automatic variogram modeling by iterative least squares: Univariate and multivariate cases. *Mathematical Geosciences 45*(4), 453–470.

Diggle, P. J., J. A. Tawn, and R. A. Moyeed (1998). Model based geostatistics. *Journal of the Royal Statistics Society, Serie B. Applied Statistics 47*(3), 299–350.

Dowd, P. A., E. Pardo-Igúzquiza, and C. Xu (2003). Plurigau: a computer program for simulating spatial facies using the truncated plurigaussian method. *Computers & Geosciences 29*(2), 123–141.

Dubrule, O. (1993). Introducing more geology in stochastic reservoir modelling. In A. Soares (Ed.), *Geostatistics Troia'92*, Dordrecht, pp. 351–369. Kluwer Academic Publishers.

Emery, X. (2007). Simulation of geological domains using the plurigaussian model: New developments and computer programs. *Computers & Geosciences 33*, 1189–1201.

Emery, X. and D. A. Silva (2009). Conditional co-simulations of continuous and categorical variables for geostatistical applications. *Computers and Geosciences 35*(6), 1234–1246.

Felletti, F. (2004). Statistical modelling and validation of correlation in turbidites: an example from the tertiary piedmont basin (castagnola fm., north italy). *Marine Petroleum Geology 21*, 23–39.

Galli, A., H. Beucher, G. Le Loc'h, B. Doligez, and . Heresim Group (1994). The pros and cons of the truncated gaussian method. In *Geostat. Simul.* Kluwer, Dordrecht.

Galli, A., G. Le Loc'h, F. Geffroy, and R. Eschard (2006). An application of the truncated plurigaussian method for modeling geology. In T. Coburn, J. M. Yarus, and R. L. Chambers (Eds.), *Stochastic modeling and geostatistics: Principles, methods, and case studies II*, Number 5 in AAPG Computer Applications in Geology, pp. 109–122.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics 1*, 141–149.

Hu, L. Y. (2000). Gradual deformation and iterative calibration of gaussian-related stochastic models. *Mathematical Geology 32*(1), 87–108.

Isatis® (2014). *Geostatistical Software by Geovariances^{TM}*. Avon - France: Geovariances.

Lantuejoul, C. (2002). *Geostatistical Simulation: models and algorithms*. Berlin: Springer.

Le Loc'h, G. and A. Galli (1996). Truncated plurigaussian method: theoretical and practical point of view. In E. Y. Baafi and N. A. Schofield (Eds.), *Proceedings of the Fitfth International Geostatics Congress, Wollongong'96*. Kluwer, Dordrecht.

Le Loc'h, G., A. Galli, B. Doligez, and . Heresim Group (1994). Improvement in the truncated gaussian method: combining several gaussian functions. In *EC-MOR IV, 4th European Conference on the Mathematics of Oil Recovery, Røros, Norway*.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics 80*, 221–240.

Liu, N. and D. S. Oliver (2004). Automatic history matching of geologic facies. *SPE journal*, 188–195.

Mariethoz, G., P. Renard, F. Cornaton, and O. Jacquet (2009). Truncated plurigaussian simulations to characterize aquifer heterogeneity. *Groundwater 47*(1), 13–24.

Matheron, G. (1962). *Traité de Géostatistique appliquée. Tome 1* (Technip ed.). Number 14 in Mémoires du BRGM. Paris.

Matheron, G., H. Beucher, C. de Fouquet, A. Galli, D. Guérillot, and C. Ravenne (1987). Conditional simulation of the geometry of fluvio deltaic reservoirs. In *SPE 16753*, pp. 123–131.

Monestiez, P., L. Dubroca, E. Bonnin, J.-P. Durbec, and C. Guinet (2006). Geostatistical modelling of spatial distribution of *balaenoptera physalus* in the northwestern merditerranean sea from sparse count data and heterogeneous observation efforts. *Ecological Modelling 193*, 615–628.

Oliver, M. A., R. Webster, C. Lajaunie, K. R. Muir, S. E. Parkes, A. H. Cameron, M. C. G. Stevens, and J. R. Mann (1993). Estimating the risk of childhood cancer. In A. Soares (Ed.), *Geostatistics Troia'92*, Dordrecht, pp. 899–910. Kluwer Academic Publishers.

Press, W., S. Teukolsky, W. Vetterling, and B. Flannery (2007). *Numerical Recipes: The Art of Scientific Computing.* (3rd ed.)., Chapter 10.2 Golden Section Search in One Dimension. New-York: Cambridge University Press.

Renard, D., H. Beucher, and B. Doligez (2008). Heterotopic bi-categorical variables in plurigaussian simulation. In *VIII International Geostatistics Congress 1*, Santiago Chile, pp. 289–298.

Renard, D., N. Bez, N. Desassis, H. Beucher, and F. Ors (2015). RGeostats: The geostatistical package [11.0.0]. *Ecole des Mines de Paris*. Free download from http://www.geosciences.mines-paristech.fr.

Romary, T. (2010). History matching of approximated lithofacies models under uncertainty. *Computational Geosciences 14*(2).

Talebi, H., O. Asghari, and X. Emery (2015). Stochastic rock type modeling in a porphyry copper deposit and its application to copper grade evaluation. *Journal of Geochimical Exploration 157*, 162–168.

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica 21*, 5–42.

Wackernagel, H. (2003). *Multivariate Geostatistics - An Introduction with Application, 3rd Edition*. New York: Springer-Verlag.

Xu, C., P. A. Dowd, K. V. Mardia, and R. J. Fowell (2006). A flexible true plurigaussian code for spatial facies simulations. *Computers & Geosciences 32*.

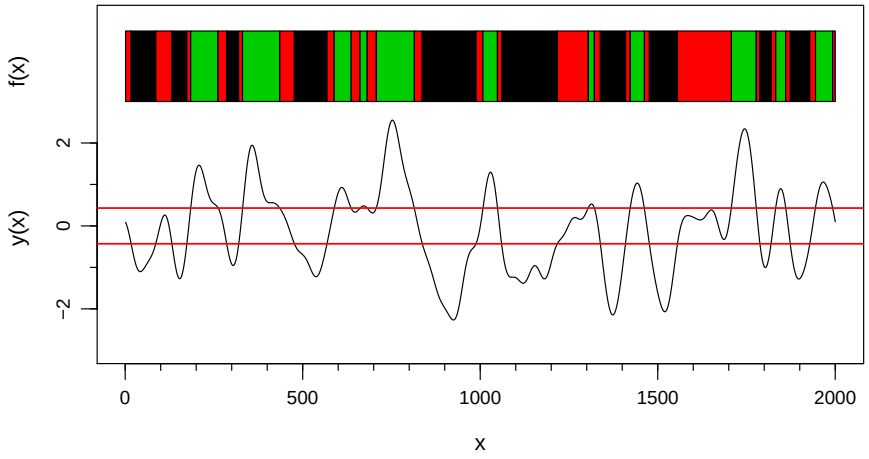


Figure 1: One realization of F (up) and the associated y (down ———), s_1 and s_2 (———)

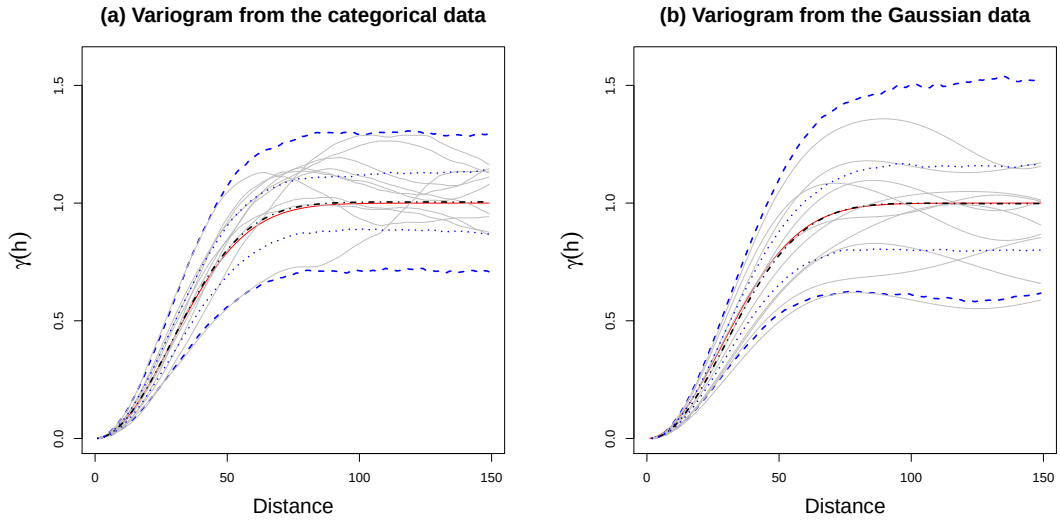


Figure 2: Actual model (———), empirical variogram of ten arbitrary simulations (———), average of the empirical variograms over all the simulations (— · — ·), 25th and 75th percentiles (·····), 5th and 95th percentiles (- - - -).

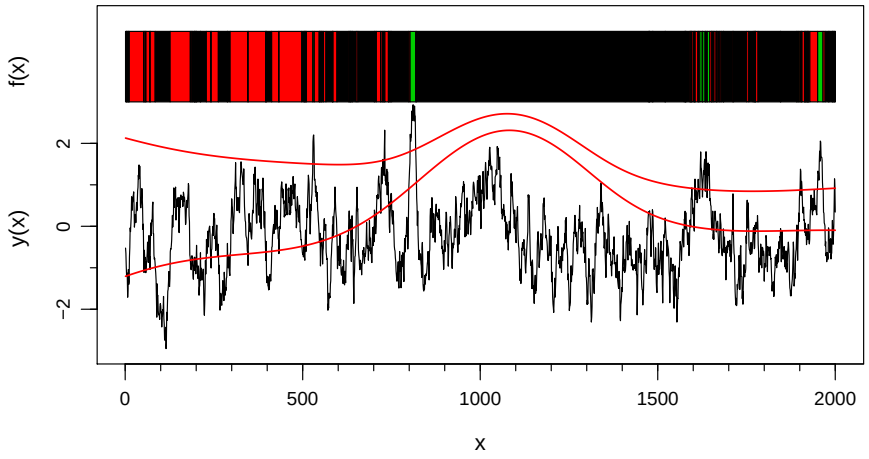


Figure 3: One realization of F (up) and the associated y (down ——), $s_1(x)$ and $s_2(x)$ (——)

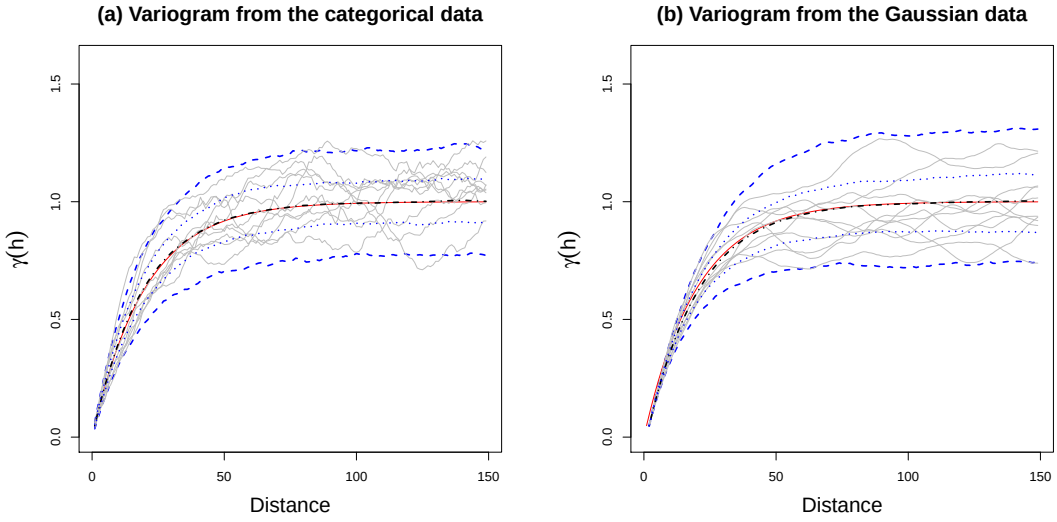


Figure 4: Actual model (——), empirical variogram of ten arbitrary simulations (——), average of the empirical variograms over all the simulations (— · — ·), 25th and 75th percentiles (·····), 5th and 95th percentiles (- - - -).

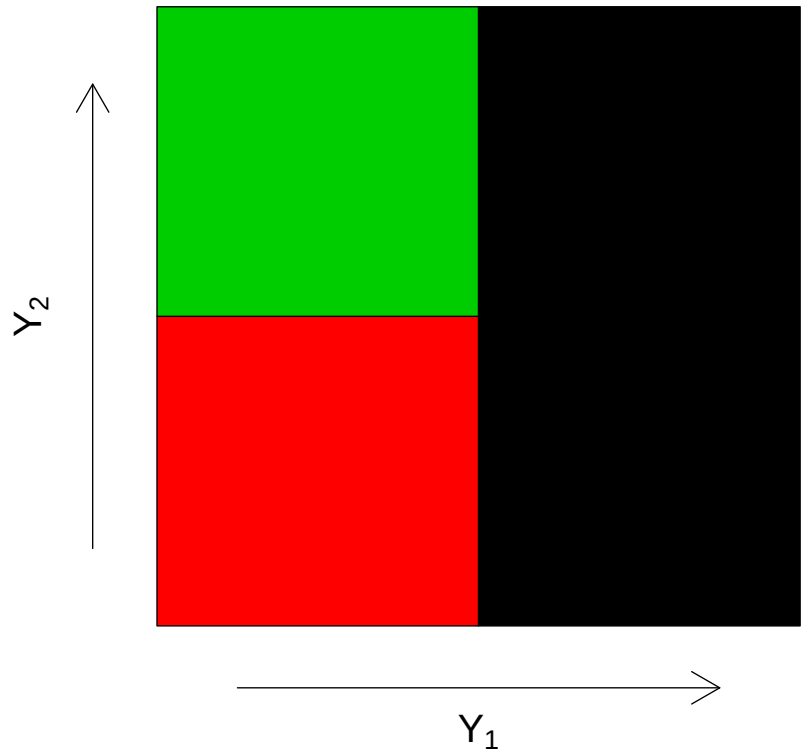
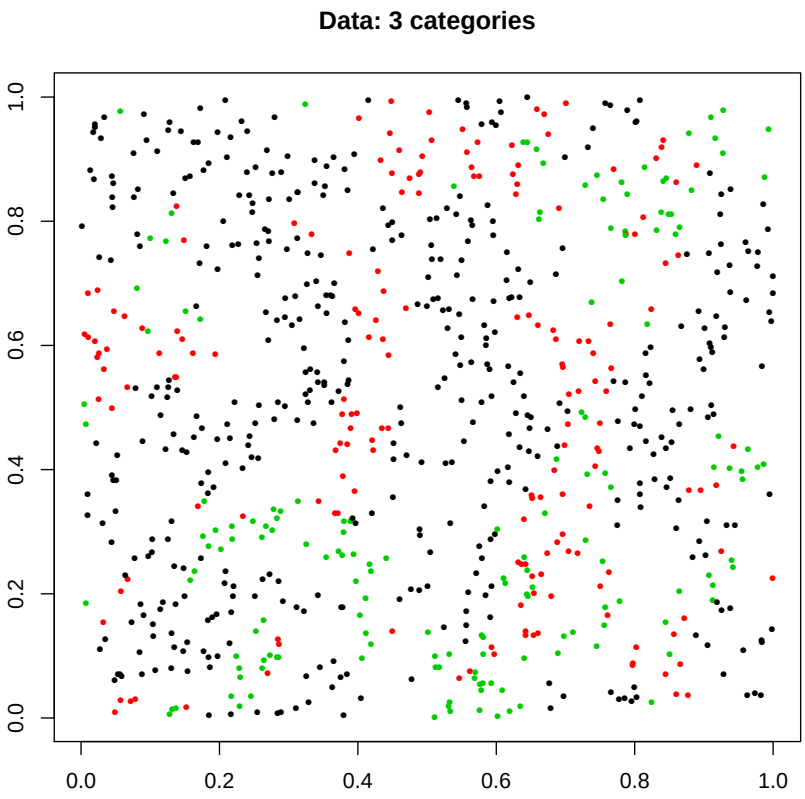


Figure 5: Representation of the coding function $\mathcal{C}$ used for the simulation study, case $q = 2$


Data: 3 categories
0.0
0.2
0.4
0.6
0.8
1.0

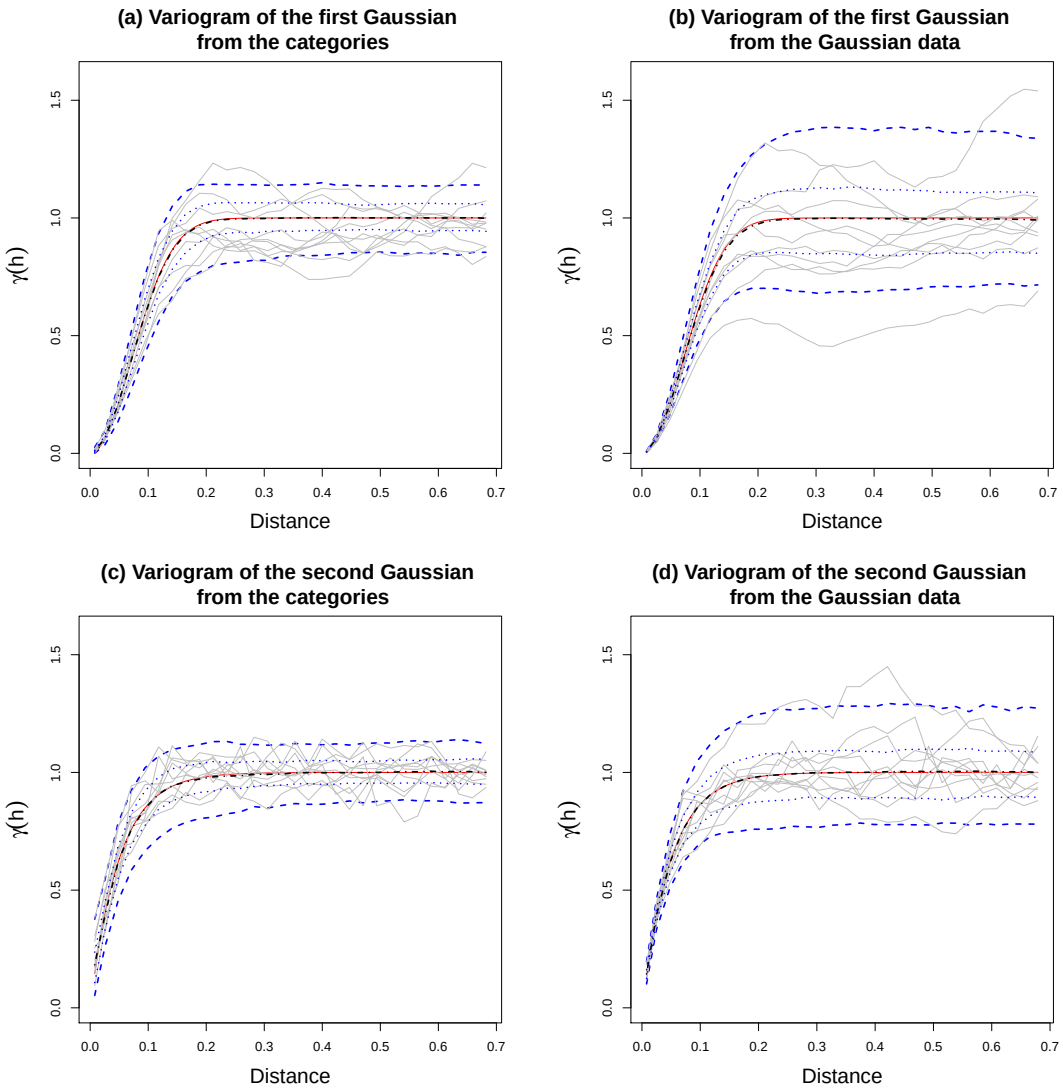


Figure 6: Actual model (——), empirical variogram of ten arbitrary simulations (——), average of the empirical variograms over all the simulations (— · — ·), 25th and 75th percentiles (·····), 5th and 95th percentiles (- - - -).