



**HAL**  
open science

## Mis-parametrization subsets for a penalized least squares model selection

Xavier Guyon, Cécile Hardouin

► **To cite this version:**

Xavier Guyon, Cécile Hardouin. Mis-parametrization subsets for a penalized least squares model selection. 2011. hal-00648151

**HAL Id: hal-00648151**

**<https://hal.science/hal-00648151>**

Preprint submitted on 5 Dec 2011

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Mis-parametrization subsets for a penalized least squares model selection

Xavier Guyon, and Cécile Hardouin

*Xavier Guyon*

*SAMM*

*University Paris 1*

*90 rue de Tolbiac*

*75013 Paris, France*

*e-mail: [guyon@univ-paris1.fr](mailto:guyon@univ-paris1.fr)*

*Cécile Hardouin*

*MODAL'X*

*University Paris Ouest Nanterre La Defense*

*200 avenue de la République*

*92000 Nanterre, France*

*e-mail: [cecile.hardouin@u-paris10.fr](mailto:cecile.hardouin@u-paris10.fr)*

**Abstract:** When identifying a model by a penalized minimum contrast procedure, we give a description of the over and under fitting parametrization subsets for a least squares contrast. This allows to determine an accurate sequence of penalization rates ensuring good identification. We present applications for the identification of the covariance for a general time series, and for the variogram identification of a geostatistical model.

**AMS 2000 subject classifications:** Primary 62F12; secondary 62M10 62M40.

**Keywords and phrases:** least squares contrast, penalized contrast, model selection, misfitting, AIC, BIC, mixture models, geostatistics.

## 1. Introduction

Let  $X = (X_1, X_2, \dots)$  be a sequence of observations associated to a semi-parametric model  $\mathbb{P}_{\theta_0}$  where  $\theta_0$  denotes the (finite dimensional) parameter of the true model. We assume that  $\theta_0$  lies in the interior of  $\Theta$ , a compact subset of  $\mathbb{R}^m$ .

We associate a subset  $P$  of  $M = \{1, 2, \dots, m\}$  to the sub model  $\Theta_P = \{\theta \in \Theta \text{ s.t. for } i \notin P, \theta_i = 0\}$  :  $p$ , the cardinal of  $P$ , is the dimension of the model  $\Theta_P$ , while  $M$  is the “support” of the dominating model  $\Theta_M$  of dimension  $m$ . We study in this paper the problem of identifying  $P_0$ , the support of the true model  $\Theta_{P_0}$ :

$$P_0 = \{i \in M : \theta_{i,0} \neq 0\}.$$

The framework is that of model selection by penalized contrast ([2], [16], [14], [12]). The selection criterion minimizes in  $P \subseteq M$  an estimating contrast  $U_n(\theta)$  in  $\Theta_P$ , this contrast being penalized at a rate  $c_n$  by the dimension  $p$  of model  $\Theta_P$ . Precisely, the procedure is the following:

(i) estimate  $\theta_0$  in  $\Theta_P$  by minimizing contrast  $U_n$ :

$$\widehat{\theta}_P \equiv \widehat{\theta}_{P,n} = \arg \min_{\theta \in \Theta_P} U_n(\theta) \quad (1.1)$$

(ii) estimate  $P_0$  by:

$$\widehat{P}_n = \arg \min \left\{ U_n(\widehat{\theta}_P) + \frac{c_n}{n} p : P \subseteq M \right\}. \quad (1.2)$$

For instance, if  $U_n$  is  $-\frac{2}{n}$  times the log-likelihood, we meet the AIC criterion with  $c_n = 2$  ([2]) and the BIC with  $c_n = \log n$  ([15]).

Therefore, we prefer  $P$  than the true model  $P_0$  if and only if:

$$\Delta_n(P, P_0) \doteq U_n(\widehat{\theta}_P) - U_n(\widehat{\theta}_{P_0}) < \frac{c_n}{n} (p_0 - p). \quad (1.3)$$

Let us define the underfitting subset  $M_n^-$  and the overfitting one  $M_n^+$  in the following way:

$$M_n^- = \{\widehat{P}_n \not\supseteq P_0\}, \quad M_n^+ = \{\widehat{P}_n \supsetneq P_0\}. \quad (1.4)$$

Such a penalization criterion for model selection was first introduced by Akaike ([2]). The literature on the subject is huge, but essentially deals with consistency or asymptotic properties of the estimator  $\widehat{\theta}_P$ , or with the choice of the penalization rate. There are less studies on the mis-fitting subsets (see for instance [3], [4], [12]). Following Guyon and Yao's approach ([12]), the aim of this work is to describe those two subsets when  $U_n$  is a least squares contrast.

Let us precise our framework: let  $\theta \mapsto \gamma(\theta)$  be a continuous and injective function from  $\Theta = \Theta_M \subseteq \mathbb{R}^m$  in  $\mathbb{R}^K$  and let  $T_n = T_n(X(n)) \in \mathbb{R}^K$  be an estimator of  $\gamma(\theta)$ . In all this work, we consider the least squares contrast :

$$U_n(\theta) = \|T_n - \gamma(\theta)\|^2.$$

where  $\|\cdot\|$  is the euclidian norm on  $\mathbb{R}^K$ .

Without any particular assumption, this simple specification on the type of contrast allows us the description of the poor parametrization subsets. The least squares contrast is commonly used and our result applies in many applications. As an illustration, we give hereafter some examples.

**Moments method:**  $\gamma(\theta)$  is a vector of  $K$  moments of a real random variable with distribution  $P_\theta$ ,  $\gamma(\cdot)$  allowing  $\theta$ 's identifiability, and  $T_n$  is the empirical moments estimator.

**Linear regression:**  $Y = X\theta + \varepsilon$ , and denoting  $T_n = ({}^tXX)^{-1} {}^tXY$ ,

$$U_n(\theta) = n^{-1} \|Y - X\theta\|^2 = u_n + \|\theta - T_n\|^2$$

**Mixture models:** The dominating model  $\Theta_m$  is associated to a family of component models in the following way:

$$\Theta_m = \cup_{\mathcal{L}} \Theta_l(\alpha_l) \text{ with } \mathcal{L} = \{1, 2, \dots, L\}.$$

A sub model is then identified in two steps : (i) first choose a part  $V$  of  $\mathcal{L}$ , (ii) then choose a sub model in each  $\Theta_l$ ,  $l \in V$ .

For instance, in the regression model  $E[Y] = a + b(x + cx^2) + d(z + exz)$ , sub-models correspond to the choices  $\Theta_1 = \{a\}$ ,  $\Theta_2 = \{(b, c) : b \neq 0 \text{ if } c \neq 0\}$ , and  $\Theta_3 = \{(d, e) : d \neq 0 \text{ if } e \neq 0\}$ . Here, for instance the choice of  $\Theta_2$  means that  $x^2$  can occur only together with  $x$  (and same for  $\Theta_3$ ).

### Covariance of a time series, variogram mixture for an intrinsic random field

Let  $(X_n)_{n \in \mathbb{Z}}$  be a zero mean stationary real valued time series, with covariance  $C$  a mixture of possible accurate covariances. For instance, for  $C(h) = C_1(h, \theta_1) + C_2(h, \theta_2) = a_1 c_1(h, \alpha_1) + a_2 c_2(h, \alpha_2)$ , sub-models are based on  $\Theta_1 = \{a_1, \alpha_1 : a_1 \neq 0, \alpha_1 \in \mathbb{R}^l\}$ ,  $\Theta_2 = \{a_2, \alpha_2 : a_2 \neq 0, \alpha_2 \in \mathbb{R}^q\}$  and  $\Theta_1 \cup \Theta_2$ , with the possibility of choosing sub-models in each  $\Theta_i$ ,  $i = 1, 2$ .

$T_n$  is the empirical estimator of the vector  $\gamma(\theta) = (C(h, \theta) : h \in \mathcal{H})^T$  where  $\mathcal{H} = \{h_k, k = 1, K\}$  is a set of lags allowing identifiability of  $\theta$ .

In the spatial case, we consider  $X = \{X_s, s \in S\}$  an intrinsic zero mean random field defined on  $\mathbb{R}^d$  with variogram  $2\gamma(h) = E[(X_{s+h} - X_s)^2]$  ([6], [5], [10];  $\gamma(h) = C(0) - C(h)$  when  $X$  is stationary with covariance  $C$ ). As previously, we determine the right variogram, a mixture of different predefined variogram, using a penalized least squares contrast procedure (see § 3). Identification of the variogram is the key function in geostatistics as it will be used to fit a model of the temporal/spatial correlation of the observed phenomenon.

The paper is organized as follows. We present the evaluation of the misfitting subsets  $M_n^-$  and  $M_n^+$  in section 2. Then, coming back to the examples of a mixture of covariances or of variograms in the spatial framework, we give in section 3 general assumptions on the underlying processes and on the sequence of penalization rates that lead to true model identification.

## 2. Description of the over and under parametrization subsets

The identification procedure is given by (1.2) and over and under parametrization subsets are given by (1.4).

### 2.1. The nearest neighbour distance for a model $P$

Let us note  $\gamma_0 = \gamma(\theta_0)$  and, for  $P \subseteq M$ ,  $\Gamma_P = \{\gamma(\theta) : \theta \in \Theta_P\}$  the  $\gamma$ -image of  $\Theta_P$ . Since the map  $\gamma(\cdot)$  is continuous,  $\Gamma_P$  is a compact submanifold of  $\mathbb{R}^K$ ,  $\Gamma_M$  being the dominating manifold: if  $P \subseteq Q \subseteq M$ ,  $\Gamma_P \subseteq \Gamma_Q \subseteq \Gamma_M \subseteq \mathbb{R}^K$ .

For any  $g \in \mathbb{R}^K$  and  $P \subseteq M$ , because of the continuity of  $\delta \mapsto \|\delta - g\|^2$  together with  $\Gamma_P$ 's compactness, there exists  $\gamma_P(g) \in \Gamma_P$ , a *nearest neighbour* of  $g$  in  $\Gamma_P$ , and  $\theta_P(g) \in \Theta_P$  such that:

$$\sigma_P(g) = \|\gamma_P(g) - g\|^2 = \min_{\delta \in \Gamma_P} \|\delta - g\|^2 = \min_{\theta \in \Theta_P} \|\gamma(\theta) - g\|^2 = U(\theta_P(g), g).$$

$\theta_P(g)$  can be seen as a least squares estimator of  $\theta$  in  $\Theta_P$  if  $g$  estimates  $\gamma(\theta)$ , and  $\gamma_P(g) = \gamma(\theta_P(g))$  as a nearest neighbour of  $g$  for the model  $P$ .

Let us underline the different behaviours of  $g \mapsto \sigma_P(g)$  depending on whether  $P$  corresponds to an under or over model:

- if  $P \not\supseteq P_0$ ,  $\sigma_P(\gamma_0) > 0$ ; indeed, there exists  $i \in P_0$  such that  $i \notin P$ , in which case  $\theta_{i,0} \neq 0$  and  $\theta_i = 0$  for all  $\theta \in \Theta_P$ ; thus  $\delta \neq \gamma_0$  for all  $\delta \in \Gamma_P$ .
- On the contrary, if  $P \supseteq P_0$ ,  $\gamma_0 \in \Gamma_P$  and  $\sigma_P(\gamma_0) = 0$ .

The continuity control of map  $g \mapsto \sigma_P(g)$  is the main point, leading to the description of  $M_n^-$  and  $M_n^+$ . Coming back to their writings (1.3), we have to compare  $\Delta_n(P, P_0) \doteq U_n(\hat{\theta}_P) - U_n(\hat{\theta}_{P_0})$  with  $\frac{c_n}{n}(p_0 - p)$ . Thus, we use the following decomposition:

$$\begin{aligned} \Delta_n(P, P_0) &\doteq U_n(\hat{\theta}_P) - U_n(\hat{\theta}_{P_0}) = \sigma_P(T_n) - \sigma_{P_0}(T_n) \\ &= \xi_1 + \xi_2(P, P_0) + \xi_3, \text{ where} \end{aligned} \quad (2.1)$$

$$\xi_2(P, P_0) = \sigma_P(\gamma_0) - \sigma_{P_0}(\gamma_0), \text{ and} \quad (2.2)$$

$$\xi_1 = \sigma_P(T_n) - \sigma_P(\gamma_0), \xi_3 = \sigma_{P_0}(\gamma_0) - \sigma_{P_0}(T_n). \quad (2.3)$$

## 2.2. The underfitting subset $M_n^-$

### 2.2.1. Continuity of the nearest neighbour distance

For  $g, g' \in \mathbb{R}^K$  and  $P \subseteq M$ , denote the points  $g, \gamma_P(g), g'$  and  $\gamma_P(g')$  of  $\mathbb{R}^K$  by  $A, B, C$  and  $D$  respectively, and let  $UV$  be the length of  $[U, V]$ .  $g, g'$  are any points in  $\mathbb{R}^K$  and  $\gamma_P(g), \gamma_P(g')$  are in  $\Gamma_P$ . We have:

$$\sigma_P(g') - \sigma_P(g) = CD^2 - AB^2 = (CD - AB)(CD + AB). \quad (2.4)$$

Since  $D$  is a nearest neighbour of  $C$  in  $\Gamma_P$ ,  $CD \leq CB$ . Moreover, the triangle inequality gives  $CB \leq CA + AB$ . Then we can write:

$$CD - AB \leq CB - AB \leq CA.$$

Analogously, we have  $AB - CD \leq CA$ , that is:

$$|CD - AB| \leq CA.$$

Let us define  $r = \sup\{\|\delta\| : \delta \in \Gamma_M\}$ . Since  $\Gamma_M$  is compact,  $r < \infty$  and  $CD + AB$  is upper bounded by  $\|g\| + \|g'\| + 2 \times r$ . This ensures:

$$\forall P \subseteq M, \forall g, g' \in \mathbb{R}^k : |\sigma_P(g') - \sigma_P(g)| \leq \{\|g\| + \|g'\| + 2 \times r\} \times \|g' - g\|. \quad (2.5)$$

### 2.2.2. Control for $\Delta_n(P, P_0)$

From (2.5) and since  $\gamma_0 \in \Gamma_M$ ,

$$|\xi_1| \text{ and } |\xi_3| \leq \delta(T_n, \gamma_0) = \{\|T_n\| + 3 \times r\} \|T_n - \gamma_0\|.$$

Besides, if  $P \not\supseteq P_0$ ,  $\sigma_P(\gamma_0) > 0$  and

$$\Delta = \inf\{\sigma_P(\gamma_0) - \sigma_{P_0}(\gamma_0) : M \supseteq P \not\supseteq P_0\} = \inf\{\sigma_P(\gamma_0) : M \supseteq P \not\supseteq P_0\} > 0. \quad (2.6)$$

Then we have  $\xi_2(P, P_0) \geq \Delta > 0$  and  $\Delta_n(P, P_0) \geq \Delta - 2\delta(T_n, \gamma_0)$ .

We then set:

$$\eta_0 = \frac{\Delta}{m} \text{ and } \delta_0 = \frac{1}{2}(\Delta - m \frac{c_n}{n}). \quad (2.7)$$

Therefore,  $\delta_0$  is strictly positive if  $c_n$  satisfies:

$$\frac{c_n}{n} < \eta_0, \quad (2.8)$$

in which case, if  $\delta(T_n, \gamma_0) < \delta_0$ ,

$$\Delta_n(P, P_0) \geq \Delta - 2\delta(T_n, \gamma_0) > m \times \frac{c_n}{n} \geq (p_0 - p) \times \frac{c_n}{n},$$

meaning, from (1.3), that  $P_0$  is preferred than  $P$ .

This gives a first description of  $M_n^-$  :

$$\text{if } \frac{c_n}{n} < \eta_0, \text{ then } M_n^- \subseteq \{(\|T_n\| + 3 \times r) \|T_n - \gamma_0\| \geq \delta_0\}.$$

Now, let us note  $A_n = \{(\|T_n\| + 3 \times r) \|T_n - \gamma_0\| \geq \delta_0\}$ ,  $B_n = \{\|T_n - \gamma_0\| \geq \delta_0\}$ ,  $\overline{B_n} = \{\|T_n - \gamma_0\| < \delta_0\}$ . We have:

$$A_n = (A_n \cap B_n) \cup (A_n \cap \overline{B_n}) \subseteq B_n \cup (A_n \cap \overline{B_n}).$$

But  $\overline{B_n} \subseteq C_n = \{\|T_n\| < \|\gamma_0\| + \delta_0\}$ . Therefore we get the final description of  $M_n^-$ :

$$\begin{aligned} M_n^- &\subseteq A_n \subseteq B_n \cup \{A_n \cap \overline{B_n}\} \subseteq B_n \cup \{A_n \cap C_n\} \\ &\subseteq \{\|T_n - \gamma_0\| \geq \delta_0\} \cup \{(\|\gamma_0\| + \delta_0 + 3 \times r) \|T_n - \gamma_0\| \geq \delta_0\} \\ &\subseteq \{\|T_n - \gamma_0\| \geq \delta_0^*\} \text{ where } \delta_0^* = \inf\{\delta_0, \frac{\delta_0}{\|\gamma_0\| + \delta_0 + 3 \times r}\}. \end{aligned}$$

### 2.3. The overfitting subset $M_n^+$

We assume that  $P \supseteq P_0$ . We denote the points  $\gamma_0$ ,  $\gamma_P(\gamma_0)$ ,  $g'$  and  $\gamma(g')$  of  $\mathbb{R}^K$  by  $A$ ,  $B$ ,  $C$  and  $D$  respectively. Since  $\gamma_P(\gamma_0) = \gamma_0$ ,  $A$  and  $B$  are identical and  $\sigma_P(\gamma_0) = 0$ . Let us then evaluate  $|\sigma_P(g') - \sigma_P(\gamma_0)| = \sigma_P(g')$ .

Since  $D$  is a point of  $\Gamma_P$  nearest of  $C$ ,  $CD^2 \leq CA^2$ . So, uniformly in  $P$ ,  $P \supsetneq P_0$  :

$$|\sigma_P(g') - \sigma_P(\gamma_0)| = \sigma_P(g') \leq \|g' - \gamma_0\|^2.$$

Moreover,  $\xi_2(P, P_0) = 0$  in formula (2.1). Therefore,  $\|T_n - \gamma_0\| < \sqrt{\frac{c_n}{2n}}$  leads to:

$$\Delta_n(P, P_0) \geq -2\|T_n - \gamma_0\|^2 > -\frac{c_n}{n} \geq (p_0 - p)\frac{c_n}{n},$$

that is  $P_0$  is preferred to  $P$ . We deduce that, without any condition on  $c_n$ :

$$M_n^+ \subseteq \{\|T_n - \gamma_0\| \geq \sqrt{\frac{c_n}{2n}}\}.$$

#### 2.4. Descriptions of $M_n^-$ and $M_n^+$

Finally we have the following result.

**Theorem 2.1.** *Let  $U_n$  be the least squares contrast*

$$U_n(\theta) = \|T_n - \gamma(\theta)\|^2$$

where  $\theta \mapsto \gamma(\theta)$  is injective and continuous from the compact set  $\Theta_m$  in  $\mathbb{R}^K$  and  $T_n \in \mathbb{R}^K$  is an estimator of  $\gamma(\theta)$ .

Let  $\Theta_m$  be the dominating model with support  $M = \{1, \dots, m\}$ ,  $\theta_0$  the true value of the parameter,  $\gamma_0 = \gamma(\theta_0)$ , and  $P_0$  the true model's support. We have the following estimations of mis parametrization subsets (1.4) for the model selection criterion (1.2):

(i) Underfitting subset  $M_n^-$  :

We set  $\Delta = \inf_P \{\min_{\theta \in \Theta_P} \|\gamma(\theta) - \gamma_0\|^2 : M \supseteq P \not\supseteq P_0\} > 0$ . If  $\frac{c_n}{n} < \frac{\Delta}{m}$  there exists a strictly positive threshold  $\delta_0^*$  such that

$$M_n^- \subseteq \{\|T_n - \gamma_0\| \geq \delta_0^*\} \tag{2.9}$$

(ii) Overfitting subset  $M_n^+$  : Without any condition on  $c_n$ , we have:

$$M_n^+ \subseteq \{\|T_n - \gamma_0\| \geq \sqrt{\frac{c_n}{2n}}\}. \tag{2.10}$$

#### Comments.

1. As in Shibata ([16], see also ([12])), the result points the important dissymmetry of  $M_n^-$  and  $M_n^+$ .
2. Similar results hold for the contrast  $V_n(\theta) = \sqrt{U_n(\theta)} = \|T_n - \gamma(\theta)\|$ . In fact, the previous step can be applied replacing  $g \mapsto \sigma_P(g)$  by  $\tau_P(g) = \|\gamma_P(g) - g\|$ . For the same constants  $\eta_0$  and  $\delta_0$  (2.7), we get the estimations:

$$\text{if } \frac{c_n}{n} < \eta_0 : M_n^- \subseteq \{\|T_n - \gamma_0\| \geq \delta_0\},$$

and without any assumption on  $c_n$ :

$$M_n^+ \subseteq \{\|T_n - \gamma_0\| \geq \frac{c_n}{2n}\}.$$

### 3. Applications

#### 3.1. A covariance mixture

Let  $(X_n)_{n \in \mathbb{Z}}$  be a zero mean real valued stationary time series with covariance function  $C$ . The goal is to determine  $C$  under very mild assumptions on  $X$ . In fact, we will assume that  $X$  is an  $\eta$ -weakly dependent process ([8]); this class of dependent processes generalizes and avoids some difficulties linked with strong mixing properties. It includes a lot of models like linear or bilinear strong mixing processes, non causal time series or gaussian weakly dependant processes particularly.

We assess  $C$  is a mixture of three covariances, those of a white noise, an exponential covariance and a Gaussian one. We set  $C(h) = C_1(h, \theta_1) + C_2(h, \theta_2) + C_3(h, \theta_3)$ , with

$$\begin{aligned} C_1(h, \theta_1) &= \sigma_1^2 \mathbf{1}_{\{0\}}(h), \theta_1 = \sigma_1^2 > 0, \\ C_2(h, \theta_2) &= \sigma_2^2 \exp\{-\rho_2 |h|\}, \theta_2 = (\sigma_2^2, \rho_2), \sigma_2^2 > 0, \rho_2 > 0 \text{ and} \\ C_3(h, \theta_3) &= \sigma_3^2 \exp\{-\rho_3 |h|^2\}, \theta_3 = (\sigma_3^2, \rho_3), \sigma_3^2 > 0, \rho_3 > 0. \end{aligned}$$

We consider the sub-models  $P_1$  to  $P_7 = M$  linked to the following settings:  $\theta_1, \theta_2, \theta_3, (\theta_1, \theta_2), (\theta_1, \theta_3), (\theta_2, \theta_3), (\theta_1, \theta_2, \theta_3)$ .

For a  $n$ - sample  $(X_1, \dots, X_n)$ , we denote  $\hat{C}(h) = \frac{1}{n-h} \sum_{i=1}^{n-h} X_i X_{i+h}$  the empirical covariance at lag  $0 \leq h < n$ . Let  $\mathcal{H} = \{h_1, \dots, h_K\}$  be a subset of  $K$  lags ( $K \geq m$ ),  $\gamma(\theta) = (C(h), h \in \mathcal{H})^T \in \mathbb{R}^K$  and  $T_n = (\hat{C}(h), h \in \mathcal{H})^T$ . The set  $\mathcal{H}$  is chosen such that it ensures to identify  $\theta$ . It's not always easy to check analytically this condition, but we can increase  $K$  to make it sure.

For each sub-model  $P$ , the least squares contrast and associated estimator are defined by

$$U_n(\theta) = \|T_n - \gamma(\theta)\|^2 = \sum_{i=1, K} \left( \hat{C}(h_i) - C(h_i; \theta) \right)^2 \text{ and } \hat{\theta}_P = \arg \min_{\theta \in \Theta_P} U_n(\theta).$$

We set the following assumptions on the  $X$  process : there exists  $q > 4$  such that  $E[|X|^q] < \infty$ , and  $X$  is an  $\eta$ -weakly dependent process ([8], [9]) such that the sequence  $\eta = (\eta_r)_{r \in \mathbb{N}}$  verifies:  $0 < \eta_r = O(r^{-\alpha})$  with  $\alpha > \max(3, \frac{2m-1}{m-4})$ . This condition is not very restrictive and is satisfied, for example, for a Gaussian process whose covariance decreases exponentially.

Denoting  $\theta_0$  the true value of the parameter, then the vector  $\sqrt{n}(T_n - \gamma(\theta_0))$  converges in distribution to a zero mean Gaussian variable ([9] theorem 1). This implies that the penalized least squares criterion identifies the true sub-model with probability 1 as soon as the sequence of penalization rates is such that  $c_n \rightarrow \infty$  and  $c_n = o(n)$ .

We can consider many variants on the previous covariance model ([17], [6], [5]) as, for instance,  $C_2$  extend to the 3-dimensional model  $C_2(h, \theta_2) = \sigma_2^2 \cos(\tau h) \exp\{-\rho_2 |h|\}$ ,  $\theta_2 = (\sigma_2^2, \rho_2, \tau)$ .



**Example 2 - A variogram mixture in geostatistics**

Let  $X = \{X_s, s \in S\}$  be a real valued intrinsic random field on  $\mathbb{R}^d$ , with variogram  $v(h) = v_1(h, \theta_1) + v_2(h, \theta_2) + v_3(h, \theta_3)$ , where  $v_1$  is the pure nugget variogram,  $v_2$  is Matérn’s variogram, and  $v_3$  is the power variogram ([6], [5], [10]):

$$\begin{aligned} v_1(h, \theta_1) &= \sigma_1^2 \text{ pour } \|h\| > 0, \gamma_1(0, \theta_1) = 0, \theta_1 = \sigma_1^2 \\ v_2(h, \theta_2) &= \sigma_2^2 \left\{ 1 - \frac{2^{1-\nu}}{\Gamma(\nu)} (b \|h\|)^\nu \mathcal{K}_\nu(b \|h\|) \right\}, b > 0, \nu > -1, \theta_2 = (\sigma_2^2, \nu, b), \\ v_3(h, \theta_3) &= \sigma_3^2 \|h\|^c, 0 < c \leq 2, \theta_3 = (\sigma_3^2, c). \end{aligned}$$

Here  $\|\cdot\|$  is the euclidian norm on  $\mathbb{R}^d$  and  $\mathcal{K}_\nu$  is the Bessel function of the second kind of parameter  $\nu$  ([1]);  $\nu = \frac{1}{2}$  (resp.  $\nu = \infty$ ) correspond to the exponential (resp. gaussian) variograms.

If the first two components of  $v$  can be associated to a stationary covariance, it is not the case for the third one since it is not bounded. This mixture model includes 6 parameters, and involves standardly seven sub-models. We can also extend the model to anisotropy, for instance, geometric anisotropy where  $\|h\|^2$  is replaced by  $\|Ah\|^2$  for  $A$  definite positive.

Suppose that  $X$  is observed on a sequence of increasing domains  $D_n$  of cardinal  $d_n$ ; for instance,  $D_n = (nB(0, r)) \cap \mathbb{Z}^d$  where  $B(0, r)$  is the ball of radius  $r$  of  $\mathbb{R}^d$ , and  $d_n = O(n^d)$ . We choose a family of lags  $\mathcal{H} = \{h_1, \dots, h_K\}$  such that the map  $\gamma : \theta \rightarrow \gamma(\theta) = (v(h), h \in \mathcal{H})^T$  is injective. For each  $h \in \mathcal{H}$  we define the empirical estimator linked to the variogram cloud  $\{(s_i - s_j), s_i, s_j \in D_n\}$ ,

$$\hat{\gamma}_n(h) = \frac{1}{2|N_{n,h}|} \sum_{s_i, s_j \in N_{n,h}} (X_{s_i} - X_{s_j})^2$$

where  $N_{n,h} = \{(s_i, s_j) \in D_n : \|s_i - s_j\| \simeq h\}$  approximates the class of couples of sites  $(s_i, s_j)$  at distance  $h$  ([6]). Therefore, we note  $T_n = (\hat{\gamma}_n(h), h \in \mathcal{H})^T$ .

For each sub-model  $P$ , we consider the least squares contrast and the associated estimator

$$\begin{aligned} U_n(\theta) &= \|T_n - \gamma(\theta)\|^2 = \sum_{i=1}^K (\hat{\gamma}_n(h_i) - \gamma(h_i; \theta))^2 \\ \hat{\theta}_{n,P} &= \arg \min_{\theta \in \Theta_P} U_n(\theta). \end{aligned}$$

Let us note for each lag  $h \in \mathcal{H}$ ,  $Z(h) = \{Z_s(h) = (X_{s+h} - X_s), s \in \mathbb{R}^d\}$  the field of the  $h$ -increments. We make the following assumptions. There exists  $\eta > 0$  such that  $E[Z_s(h)^{4+\eta}] < \infty$ , and  $Z(h)$  is  $\alpha$ -mixing ([7]) satisfying:

$$\exists C < \infty \text{ and } \tau > \frac{(4 + \eta)d}{\eta} \text{ such that for all } k, l, m : \alpha_{k,l}(m) \leq Cm^{-\tau}. \quad (3.1)$$

Then the vector  $d_n^{-\frac{1}{2}}(T_n - \gamma(\theta))^T$  is asymptotically zero mean Gaussian distributed ([10], [13]). Therefore, the penalized least squares criterion well-identifies

the true sub-model with probability tending to 1 as soon as the sequence of penalization rates tends to infinity and verifies  $c_n = o(d_n^{-1})$ . Condition (3.1) above is not very restrictive and is verified in many cases; for instance it applies for gaussian stationary fields  $Z(s)$  with exponentially decreasing covariance.

## References

- [1] ABRAMOVWITZ, M. and STEGUN, I.A. (1970). *Handbook of Mathematical Functions*, Wiley, New York.
- [2] AKAIKE, H. (1969). Fitting autoregressive models for prediction. *Theory Probab. Appl.* **21** 243–247.
- [3] BAI, Z.D., SUBRAMANYAN, K. and ZHAO, L.C. (1988). On determination of the order of an autoregressive model. *J. Multivariate Analysis* **27** 40–52.
- [4] BAI, Z.D., KRISHNAIAH, P.R., SAMBAMOORTHY, N. and ZHAO, L.C. (1992). Model selection for non linear models. *Sankya: Indian J. Statist., Ser. B* **54** 200–219.
- [5] CHILÈS, J.P. and DELFINER, P. (1999). *Geostatistics*, Wiley.
- [6] CRESSIE, N. (1993). *Statistics for spatial data*, Wiley.
- [7] DOUKHAN, P. (1994). *Mixing: properties and examples*, Lecture Notes in Statistics **85**, Springer-Verlag.
- [8] DOUKHAN, P., and LOUHICHI, S. (1999). A new weak dependence condition and applications to moment inequalities. *Stoch. Proc. Appl.* **84** 313–342.
- [9] BARDET, J.-M., DOUKHAN, P., and LEON, J. (2008). A functional limit theorem for weakly dependent processes and its applications. *Statistical inference for Stochastic processes* **11,3** 265–280.
- [10] GAETAN, C. and GUYON, X. (2010). *Spatial statistics and modeling*, Springer.
- [11] GUYON, X. (1995). *Random fields on a network : Modeling, Statistics and Applications*, Springer.
- [12] GUYON, X., and YAO, J.F. (1999). On the underfitting and overfitting sets of models chosen by order selection criteria. *J. Mult. Anal.* **70** 221–249.
- [13] LAHIRI, S.N., LEE, Y., and CRESSIE, N. (2002). On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters. *J. Stat. Planning and Inf.* **103** 65–85.
- [14] SENOUSI, R., (1990). Statistique asymptotique presque sûre des modèles statistiques convexes. *Ann. Inst. H. P.* **26** 19–44.
- [15] SCHWARZ, G., (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.
- [16] SHIBATA, R., (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63** 117–126.
- [17] YAGLOM, A.M. (1987). *Correlation Theory of Stationary and Related Random Functions, Vol. I, Basic Results*, Springer.