Unbiased risk estimation method for covariance estimation

Hélène Lescornel, Jean-Michel Loubes, Claudie Chabriac

▶ **To cite this version:**

HAL Id: hal-00653892

https://hal.archives-ouvertes.fr/hal-00653892

Submitted on 20 Dec 2011

Unbiased risk estimation method for covariance estimation

Hélène Lescornel *, Jean-Michel Loubes,† Claudie Chabriac‡

Abstract

We consider a model selection estimator of the covariance of a random process. Using the Unbiased Risk Estimation (URE) method, we build an estimator of the risk which allows to select an estimator in a collection of model. Then, we present an oracle inequality which ensures that the risk of the selected estimator is close to the risk of the oracle. Simulations show the efficiency of this methodology.

Keywords: covariance estimation, model selection, URE method.

1 Introduction

Estimating the covariance function of stochastic processes is a fundamental issue in statistics with many applications, ranging from geostatistics, financial series or epidemiology for instance (we refer to [Ste99], [Jou77] or [Cre93] for general references). While parametric methods have been extensively studied in the statistical literature (see [Cre93] for a review), nonparametric procedures have only recently received attention, see for instance [EPTA08, BBLMA10, BBLLMA10, BBLA11] and references therein. One of the main difficulty in this framework is to impose that the estimator is also a covariance function, preventing the direct use of usual nonparametric statistical methods.

In this paper, we propose to construct a non parametric estimator of the covariance function of a stochastic process by using a model selection procedure based on the Unbiased Risk Estimation (U.R.E.) method. We work under general assumptions on the process, that is, we do not assume Gaussianity nor stationarity of the observations.

Consider a stochastic process $(X(t))_{t \in T}$ taking its values in $\mathbb{R}$ and indexed by $T \subset \mathbb{R}^d$, $d \in \mathbb{N}$. We assume that $\mathbb{E}[X(t)] = 0 \; \forall t \in T$ and we aim at estimating its covariance function $\sigma(s, t) = \mathbb{E}[X(s)X(t)] < \infty$ for all $t, s \in T$. We assume we observe $X_i(t_j)$ where $i \in \{1 \dots n\}$ and $t \in \{1 \dots n\}$. Note that the observation points t_j are fixed and that the X_i's are independent copies of the process X. Set set $x_i = (X_i(t_1), \dots, X_i(t_p)) \; \forall i \in \{1 \dots n\}$ and denote by Σ the covariance matrix of these vectors.

Following the methodology presented in [BBLMA10], we approximate the process X by its projection onto some finite dimensional model. For this, consider a countable set of functions $(g_\lambda)_{\lambda \in \Lambda}$ which may be for instance a basis of $L^2(T)$ and choose a collection of models $\mathcal{M} \subset \mathcal{P}(\Lambda)$. For $m \subset \mathcal{M}$, a finite number of indices, the process can be approximated by

$$X(t) \approx \sum_{\lambda \in m} a_\lambda g_\lambda(t).$$

*Institut de Mathématiques de Toulouse UMR 5219 31062 Toulouse, Cedex 9, France. Email: lescornel@math.univ-toulouse.fr

†loubes@math.univ-toulouse.fr

‡chabriac@univ-tlse2.fr

Such an approximation leads to an estimator of Σ depending on the collection of functions m, denoted by $\hat{\Sigma}_m$. Our objective is to select in a data driven way, the best model, i.e the one close to an oracle m_0 defined as a minimizer of the quadratic risk, namely

$$m_0 \in \underset{m \in \mathcal{M}}{\arg \min} R(m) = \underset{m \in \mathcal{M}}{\arg \min} \mathbb{E}\left[\left\|\Sigma - \hat{\Sigma}_m\right\|^2\right].$$

A model selection procedure will be performed using the U.R.E. method, which has been introduced in [Ste81] and fully described in [Tsy04]. The idea is to find an estimator $\hat{R}(m)$ of the risk which is unbiased, and to select $\hat{m}$ by minimizing this estimator. Hence, if $\hat{R}$ is close to its expectation, $\hat{\Sigma}_{\hat{m}}$ will be an estimator with a small risk, nearly as the best quantity $\hat{\Sigma}_{m_0}$.

In this work, following the U.R.E. method, we build an estimator of the risk which allows to select an estimator of the covariance function. Then, we present an oracle inequality for the covariance estimator which ensures that the risk of the selected estimator is not too large with respect to the risk of the oracle.

The paper is organized as follows. In Section 2 we present the statistical framework and recall some useful algebraic tools for matrices. The following section, Section 3 is devoted to the approximation of the process and the construction of the covariance estimator. Section 4 is devoted to the U.R.E. method, and provides an oracle inequality. Some numerical experiments are exposed in Section 5, while the proofs are postponed to the Appendix.

2 The statistical framework

Recall that we consider an $\mathbb{R}$-valued stochastic process, $X = (X(t))_{t \in T}$, where T is some subset of $\mathbb{R}^d$, $d \in \mathbb{N}$. We assume that X has finite moments up to order 4 and zero mean. Our aim is to study the covariance function of X denoted by $\sigma(s,t) = \mathbb{E}[X(s)X(t)]$.

Let $X_1, ... X_n$ be independent copies of the process X, and assume that we observe these copies at some determinist points $t_1, ..., t_p$ in T. We set $x_i = (X_i(t_1), \ldots, X_i(t_p))\top$, and denote the empirical covariance of the data by

$$S = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top$$

with expectation $\Sigma = (\sigma(t_j, t_k))_{1 \leqslant j, k \leqslant p}$.

Hence, the observation model can be written, in a matrix regression framework, as

$$x_i x_i^\top = \Sigma + U_i \quad \in \mathbb{R}^{p \times p} \quad, 1 \leqslant i \leqslant n \tag{1}$$

Where U_i are i.i.d. error matrices with $\mathbb{E}[U_i] = 0$.

We now recall some notations related to the study of matrices, which will be used in the following.
Denote by $\mathcal{S}_t$ the subset composed of symmetric matrix in $\mathbb{R}^{t \times t}$.

For any matrix $A = (a_{ij}) \in \mathbb{R}^{s \times t}$, $\|A\|^2 = tr(AA^\top)$ is the Frobenius norm of the matrix which is associated to the inner scalar product $\langle A, B \rangle = tr(AB^\top)$.

$A^- \in \mathbb{R}^{t \times s}$ is a reflexive generalized inverse of A, that is, some matrix such as $A^- A A^- = A$ and $AA^- A = A^-$.

In the following, we will consider matrix data as a natural extension of the vectorial data, with different correlation structure. For this, we introduce a natural linear transformation, which converts any matrix into a column vector. The vectorization of a $k \times n$ matrix $A = (a_{ij})_{1 \leq i \leq k, 1 \leq j \leq n}$ is the $kn \times 1$ column vector denoted by $vec(A)$, obtained by stacking the columns of the matrix A on top of one another. That is $vec(A) = [a_{11}, ..., a_{k1}, a_{12}, ..., a_{k2}, ..., a_{1n}, ..., a_{kn}]^\top$.

If $A = (a_{ij})_{1 \leq i \leq k, 1 \leq j \leq n}$ is a $k \times n$ matrix and $\mathbf{B} = (b_{ij})_{1 \leq i \leq p, 1 \leq j \leq q}$ is a $p \times q$ matrix, then the Kronecker product of the two matrices, denoted by $A \otimes B$, is the $kp \times nq$ block matrix

$$A \otimes B = \begin{bmatrix} a_{11}B & . & . & . & a_{1n}B \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ a_{k1}B & . & . & . & a_{kn}B \end{bmatrix}.$$

For A, B and C some real matrices, we recall the following properties that will be useful in our settings.

Proposition 2.1.

$$vec(ABC) = \left(C^\top \otimes A \right) vec(B) \tag{2}$$

$$\|A\| = \|vec(A)\| = \|vec(A)\|_{\ell_2} \tag{3}$$

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD) \tag{4}$$

$$(A \otimes B)^\top = A^\top \otimes B^\top \tag{5}$$

These identities can be found in [Seb08].

Let $m \in \mathcal{M}$, and recall that to the finite set $\{g_\lambda\}_{\lambda \in m}$ of functions $g_\lambda : T \to \mathbb{R}$ we associate the $n \times |m|$ matrix $\mathbf{G}$ with entries $g_{j\lambda} = g_\lambda(t_j)$, $j = 1, ..., n$, $\lambda \in m$. Furthermore, for each $t \in T$, we write $\mathbf{G}_t = (g_\lambda(t), \lambda \in m)^\top$. For $k \in \mathbb{N}$, $\mathcal{S}_k$ denotes the linear subspace of $\mathbb{R}^{k \times k}$ composed of symmetric matrices. For $\mathbf{G} \in \mathbb{R}^{n \times |m|}$, $\mathcal{S}(\mathbf{G})$ is the linear subspace of $\mathbb{R}^{n \times n}$ defined by

$$\mathcal{S}(\mathbf{G}) = \left\{ \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top : \mathbf{\Psi} \in \mathcal{S}_m \right\}.$$

This set will be the natural projection space for the corresponding covariance estimator.

3   Model selection approach

The estimation procedure is a two step procedure. First we consider a functional expansion of the process and approximate it by its projection onto some finite collection of functions. Then, we construct a rule to pick out the best of these estimators among the collection of estimated, based on the U.R.E. method.

In this section, we explain the construction of a projection based estimator for the covariance of a process and point out its properties. More details can be found in [BBLMA10].

Consider a process X with an expansion on a set of functions $(g_\lambda)_{\lambda \in \Lambda}$ of the following form

$$X(t) = \sum_{\lambda \in \Lambda} a_\lambda g_\lambda(t)$$

where Λ is a countable set, and $(a_\lambda)_{\lambda \in \Lambda}$ are random coefficients in $\mathbb{R}$ of the process X.

This situation occurs in large number of cases. If we assume that the process takes its values in $L^2(T)$ or an Hilbert space, a natural choice of the functions is given by the corresponding Hilbert basis $(g_\lambda)_{\lambda \in \Lambda}$ of $L^2(T)$. Alternatively, the Karhunen-Loeve expansion of the covariance provides a natural basis. However, since it relies on the nature of the process X, this expansion is usually unknown or require additional information on the process. We refer to [Adl90] for more references on this expansion. Under other kind of regularity assumptions on the process, for instance assuming that the paths of the process belong to some RKHS, other expansions can be considered as in [CY10] for instance.

Now consider the projection of the process onto a finite number of functions. For this, let m be a finite subset of Λ and consider the corresponding approximation of the process in the following form

$$\tilde{X}(t) = \sum_{\lambda \in m} a_\lambda g_\lambda(t) \tag{6}$$

We note $G_m \in \mathbb{R}^{p \times |m|}$ where $(G_m)_{j\lambda} = g_\lambda(t_j)$ and a_m the random vector of $\mathbb{R}^{|m|}$ with coefficients $(a_\lambda)_{\lambda \in m}$.

Hence, we obtain that

$$\tilde{x} = \left(\tilde{X}(t_1), .., \tilde{X}(t_p) \right)^\top = G_m a_m$$

and

$$\tilde{x}\tilde{x}^\top = G_m a_m a_m^\top G_m^\top.$$

Thus, approximating the process X by $\tilde{X}$ its projection onto the model m implies approximating the covariance matrix Σ by $G_m \Psi G_m^\top \quad \Psi \in \mathbb{R}^{|m| \times |m|}$ where $\Psi = \mathbb{E}\left[a_m a_m^\top \right]$ is some symmetric matrix. With previous definitions, that amounts to saying that we want to choose an estimator in the subset $\mathcal{S}(G_m)$ for some subset m of Λ.

Assume that the subset m is fixed. The best approximation of Σ in $\mathcal{S}(G_m)$ for the Frobenius norm is its projection denoted by Σ_m. But Σ is unknown, hence we can not determinate this quantity. A natural idea is to study the projection of S on $\mathcal{S}(G_m)$. We denote this quantity by $\hat{\Sigma}_m$.

Proposition 3.1 in [BBLMA10] gives an explicit form for these projections. We recall it for sake of completeness.

Proposition 3.1. *Let A in $\mathbb{R}^{p \times p}$ and $G \in \mathbb{R}^{p \times |m|}$. The infimum*

$$\inf \left\{ \|A - \Gamma\| ; \Gamma \in \mathcal{S}(G) \right\}$$

is achieved at

$$\hat{\Gamma} = G \left(G^\top G \right)^- G^\top \left(\frac{A + A^\top}{2} \right) G \left(G^\top G \right)^- G^\top$$

In particular, if $A \in \mathcal{S}_p$, the projection of A on $\mathcal{S}(G)$ is $\Pi A \Pi$ with the projection matrix $\Pi = G \left(G^\top G \right)^- G^\top \in \mathbb{R}^{p \times p}$.

It amounts to saying that $\inf \left\{ \left\| A - G \Psi G^\top \right\| ; \Psi \in \mathcal{S}_{|m|} \right\}$ is reached at

$$\hat{\Psi} = \left(G^\top G \right)^- G^\top \left(\frac{A + A^\top}{2} \right) G \left(G^\top G \right)^-.$$

Remark 3.2. *Thanks to the properties of the reflexive generalized inverse given in [Rao73], the projection of a non-negative definite matrix $A \in \mathcal{S}_p$ on $\mathcal{S}(G)$ will be also a non-negative definite matrix. Moreover, the matrix Π does not depend on the choice of the generalized inverse.*

Thanks to this result, the projection of Σ on $\mathcal{S}(G_m)$ can be characterized as

$$\Sigma_m = \Pi_m \Sigma \Pi_m \tag{7}$$

and the same for S (that is, our candidate for estimating Σ)

$$\hat{\Sigma}_m = \Pi_m S \Pi_m \tag{8}$$

where $\Pi_m = G_m \left(G_m^\top G_m \right)^- G_m^\top$.

Hence, the estimator $\hat{\Sigma}_m$ is a covariance matrix. Now, our aim is to choose the best subset m among a collection of candidates.

4 Model selection with the U.R.E. method

Let $\mathcal{M}$ be a finite collection of models m. In this section, we focus on picking the best model among this collection by following the U.R.E. method. Since the law of $\left\|\Sigma - \hat{\Sigma}_m\right\|$ is unknown, we thus aim at finding an estimator of its expectation.

We consider that the best subset m is m_0 defined by

$$m_0 \in \underset{m \in \mathcal{M}}{argmin} \mathbb{E}\left[\left\|\Sigma - \hat{\Sigma}_m\right\|^2\right]$$

Then the oracle is defined as the best estimate knowing all the information, namely $\hat{\Sigma}_{m_0}$.

Set $R(m) = \mathbb{E}\left[\left\|\Sigma - \hat{\Sigma}_m\right\|^2\right]$. First, we compute this quantity.

Proposition 4.1.

$$\mathbb{E}\left[\left\|\Sigma - \hat{\Sigma}_m\right\|^2\right] = \|\Sigma - \Pi_m \Sigma \Pi_m\|^2 + \frac{tr\left((\Pi_m \otimes \Pi_m)\Phi\right)}{n} \tag{9}$$

where $\Phi = Var\left(vec\left(xx^\top\right)\right)$.

Here we can note the similarity with the usual risk for standard estimation models. For instance, assume that we observe a Gaussian model with observations a vector $Y \in \mathbb{R}^n$ such as

$$Y = \theta + \epsilon\xi \quad \xi \sim \mathcal{N}(0, I_n)$$

where $\epsilon \in \mathbb{R}$ and $\theta \in \mathbb{R}^n$ is the unknown quantity to estimate, using the projection $\hat{\theta}_m$ of the vector Y onto some subspace S_m. If the subspace dimension is denoted by D_m, the risk of a such estimator is given by

$$\mathbb{E}\left[\left\|\theta - \hat{\theta}_m\right\|^2\right] = \|\theta_m - \theta\|^2 + \epsilon^2 D_m.$$

We thus recognize the same kind of decomposition with a bias term and with $\frac{tr((\Pi_m \otimes \Pi_m)\Phi)}{n}$ playing the role of the variance term D_m/n with $\epsilon = 1/\sqrt{n}$. Hence it is natural to extend the Unbiased Risk Estimation procedure of previous Gaussian model to the matrix model obtained by the vectorization of Model (1).

Now, we present an estimator of the risk. We assume $n \geqslant 3$, and we set :

$$\hat{\gamma}_m^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left\|\Pi_m x_i x_i^\top \Pi_m - \hat{\Sigma}_m\right\|^2$$

Proposition 4.2. $\left\|S - \hat{\Sigma}_m\right\|^2 + 2\frac{\hat{\gamma}_m^2}{n} + C$ is an unbiased estimator of the risk, where C does not depend on m. More precisely :

$$\mathbb{E}\left[\left\|S - \hat{\Sigma}_m\right\|^2 + 2\frac{\hat{\gamma}_m^2}{n}\right] = \mathbb{E}\left[\left\|\Sigma - \hat{\Sigma}_m\right\|^2\right] + \frac{tr(\Phi)}{n}$$

Note that the constant $\frac{tr(\Phi)}{n}$ is unknown but does not depend on m. So in the URE procedure, minimizing $\left\|S - \hat{\Sigma}_m\right\|^2 + 2\frac{\hat{\gamma}_m^2}{n}$ with respect to m is equivalent to minimizing $\left\|S - \hat{\Sigma}_m\right\|^2 + 2\frac{\hat{\gamma}_m^2}{n} + C$ which is unbiased.

Then we can define the estimator $\hat{\Sigma}$ of Σ by

$$\hat{\Sigma} = \Pi_{\hat{m}} S \Pi_{\hat{m}} = \hat{\Sigma}_{\hat{m}}$$

$$\text{with } \hat{m} \in \underset{m \in \mathcal{M}}{argmin}\left(\left\|S - \hat{\Sigma}_m\right\|^2 + 2\frac{\hat{\gamma}_m^2}{n}\right)$$

The next theorem establishes an oracle inequality for this estimator.

Theorem 4.3. *For all $A > 0$, we have :*

$$\mathbb{E}\left[\left\|\tilde{\Sigma} - \Sigma\right\|^2\right] \leqslant \left(1 + A^{-1}\right) \inf_{m \in \mathcal{M}} \mathbb{E}\left[\left\|\Sigma - \hat{\Sigma}_m\right\|^2\right] + \frac{tr\left(\Phi\right)}{n}\left(4 + A\right)$$

Hence we have obtained a model selection procedure which enables to recover the best covariance model among a given collection. This method works without strong assumptions on the process, in particular stationarity is not assumed, but at the expend of necessary i.i.d observations of the process at the same points.

We point out that this study requires a large number of replications n with respect to the number of observation points p. Actually our method is not designed to tackle the problem of covariance estimation in the high dimensional case $p >> n$. This topic has received a growing attention over the past years and we refer to [BL08] and references therein for a survey.

The proof of these results are using the vectorization of the matrices involved here. That is why we must deal with the matrix $\Phi = var\left(vec\left(xx^\top\right)\right)$. It is postponed to the appendix.

5   Numerical examples

In this section we illustrate the behaviour of the covariance estimator $\hat{\Sigma}$ with programs implemented using SCILAB. We aim at knowing if our procedure leads to choose the best model, that is the model minimizing the risk.

Recall that n is the number of copies of the process and p is the number of points where we observe these copies. Here, we consider the case where $T = [0; 1]$ and Λ is a subset of $\mathbb{N}$. For sake of simplicity, we identify m and the set $1, \ldots, m$. Moreover, the points $(t_j)_{1 \leqslant j \leqslant p}$ are equi-spaced in $[0; 1]$.

For a given process X, we must start by the choice of the functions of its expansion. Their knowledge is needed for the matrix G_m. Indeed, $(G_m)_{j\lambda} = g_\lambda(t_j)$.

The method is the following: First, we simulate a sample for p and n given. Second, for m between 1 to some integer M, we compute the unbiased risk estimator related to the model m. Finally, we pick out a $\hat{m}$ minimizing this estimator and we compute $\hat{\Sigma}$.

For each example, we plot the curve of the risk function and give its minimum m_0. We plot also the curve of the function of the risk estimator and give its minimum $\hat{m}$. Finally we compare the true covariance and the estimator.

Example 1

Here we work with the numerical examples of [BBLMA10]. We choose the Fourier basis functions:

$$g_\lambda(t) = \begin{cases} \frac{1}{\sqrt{p}} \text{ si } \lambda = 1 \\ \sqrt{2}\frac{1}{\sqrt{p}}\cos(2\pi\frac{\lambda}{2}t) \text{ si } \lambda \text{ est pair} \\ \sqrt{2}\frac{1}{\sqrt{p}}\sin(2\pi\frac{\lambda-1}{2}t) \text{ si } \lambda \text{ est impair} \end{cases}$$

And we study the following process :

$$X(t) = \sum_{\lambda=1}^{m^\star} a_\lambda g_\lambda(t)$$

where a_λ are independent Gaussian variables with mean zero and variance $V(a_\lambda)$. Let $D(V)$ the diagonal matrix in $m^\star \times m^\star$ such as $D(V)_{\lambda\lambda} = V(a_\lambda)$. Then we have

$$\Sigma = G_{m^\star}D(V)G_{m^\star}^\top$$

Here are the results for $V(a_\lambda) = 1 \quad \forall \quad \lambda$. We choose $m^\star = 35 = p$, $n = 50$, $M = 31$. Here it can be shown that the minimum of the risk is achieved at $\frac{n}{2} - 1$, so in this setting we have $m_0 = 24$. Here the minimum of the estimator is the same : $\hat{m} = 24$.
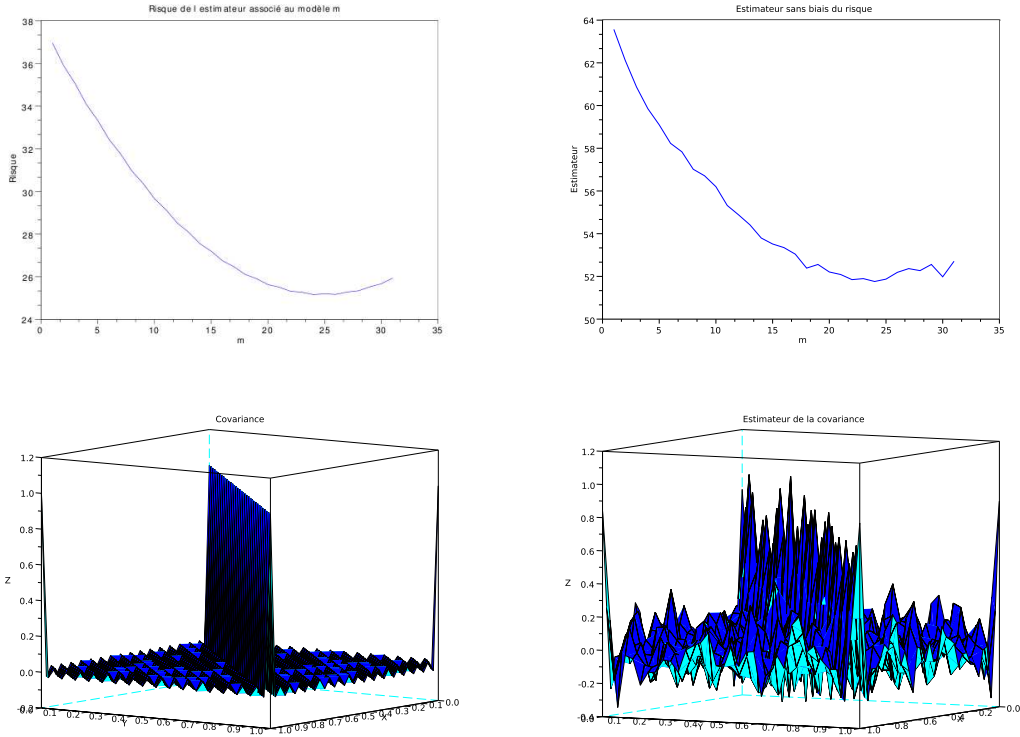
Here are the results for $V(a_\lambda) = 0.0475 + 0.95^\lambda \quad \forall \quad \lambda$, and $m^\star = 35 = p$, $n = 60$, $M = 34$. Here the figures show that $m_0 = \hat{m} = 18$.

Example 2

Now we test our estimator with the process studied in [CY10].

We consider the functions
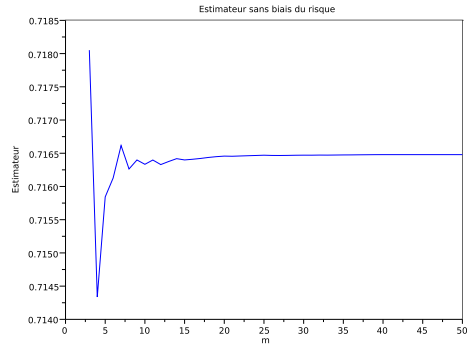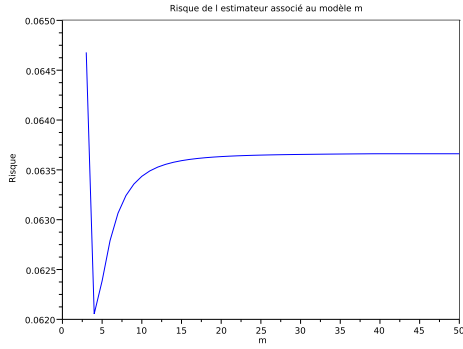
$$g_\lambda(t) = \cos(\lambda \pi t)$$

And the process X studied is :

$$X(t) = \sum_{\lambda=1}^{m^\star} a_\lambda \zeta_\lambda g_\lambda(t)$$

where a_λ are i.i.d. random variables following the uniform law on $\left[-\sqrt{3}; \sqrt{3}\right]$ and $\zeta_\lambda = \frac{(-1)^{\lambda+1}}{\lambda^2}$. If D is the diagonal matrix with entries $D_{\lambda\lambda} = \frac{1}{\lambda^4}$, as before we have that

$$\Sigma = G_{m^\star} D G_{m^\star}^\top$$

Here we choose $m^\star = 50$, $n = 1000$, $p = 40$ and $M = 20$. We found $m_0 = 4 = \hat{m}$.
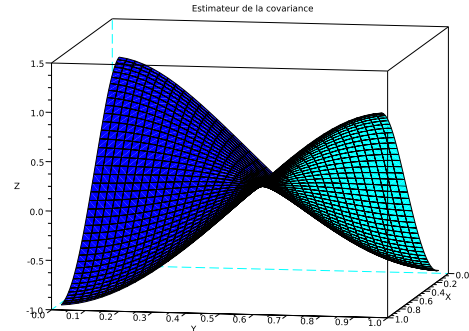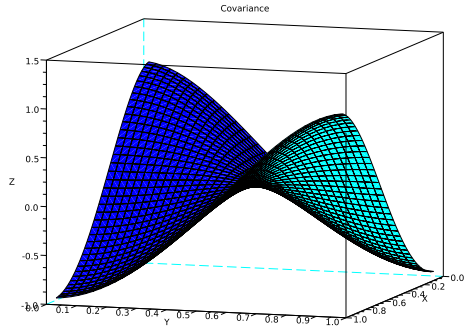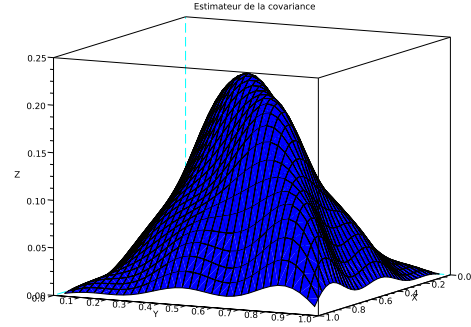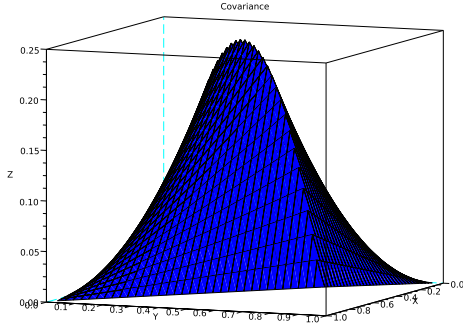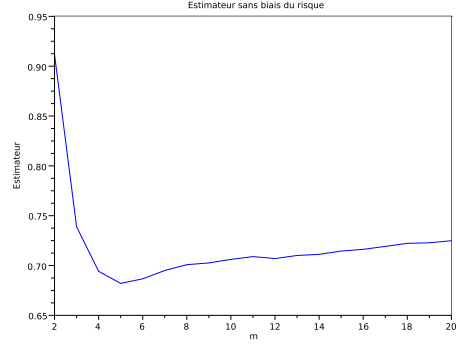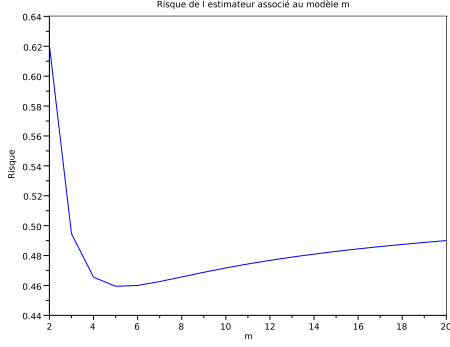
Example 3

Here we consider the case of the brownian bridge with its Karhunen Loeve expansion. Indeed, this expansion

$$X(t) = \sum_{\lambda \geqslant 1} Z_\lambda \sqrt{\nu_\lambda} g_\lambda(t)$$

is computed in [SW86], p.213-215 : $\nu_\lambda = \left(\frac{1}{\lambda\pi}\right)^2$, and $g_\lambda(t) = \sqrt{2}\sin(\lambda\pi t)$.

The covariance function of the brownian bridge is $K(s,t) = s(1-t)$ for $s \leqslant t$. Simulate the sample is the same as simulate n gaussian vectors of covariance matrix $\Sigma = (K(t_i, t_j))_{1 \leqslant i,j \leqslant p}$. Here $n = 100$, $p = 35$ and $M = 20$. We found $m_0 = 5 = \hat{m}$.

8

Note that for the first and the last example, the size of the sample is not so large. However, for each of the simulated examples, the covariance estimator shows good performances. Indeed, the procedure introduced in this paper leads each time to the selection of the best model, in the sense that the chosen model minimizes the risk.

6   Appendix

Recall that $\Sigma_m = \Pi_m \Sigma \Pi_m$, $\hat{\Sigma}_m = \Pi_m S \Pi_m$ and

$$\hat{\gamma}_m^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left\| \Pi_m x_i x_i^\top \Pi_m - \hat{\Sigma}_m \right\|^2$$

We start by proving the proposition 4.1.

Proof. Using the orthogonality, we have

$$\left\| \Sigma - \hat{\Sigma}_m \right\|^2 = \| \Sigma - \Sigma_m \|^2 + \left\| \Sigma_m - \hat{\Sigma}_m \right\|^2$$

With the proposition 2.1 we deduce

$$\left\| \Sigma_m - \hat{\Sigma}_m \right\|^2 = \left\| vec(\Sigma_m - \hat{\Sigma}_m) \right\|^2 = \left\| \left(\Pi_m^\top \otimes \Pi_m \right) vec\left(\Sigma - S \right) \right\|^2$$

Since Π_m is a projection matrix,

$$\left\| \Sigma_m - \hat{\Sigma}_m \right\|^2 = tr\left(\left(\Pi_m \otimes \Pi_m \right) vec\left(\Sigma - S \right) vec\left(\Sigma - S \right)^\top \right)$$

Hence

$$\mathbb{E}\left[\left\| \Sigma - \hat{\Sigma}_m \right\|^2 \right] = \| \Sigma - \Sigma_m \|^2 + \mathbb{E}\left[tr\left(\left(\Pi_m \otimes \Pi_m \right) vec\left(\Sigma - S \right) vec\left(\Sigma - S \right)^\top \right) \right]$$

9

$$= \|\Sigma - \Sigma_m\|^2 + tr\left((\Pi_m \otimes \Pi_m) \mathbb{E}\left[vec\left(\Sigma - S\right) vec\left(\Sigma - S\right)^\top\right]\right)$$

$$= \|\Sigma - \Sigma_m\|^2 + \frac{tr\left((\Pi_m \otimes \Pi_m) \mathbb{E}\left[vec\left(\Sigma - xx^\top\right) vec\left(\Sigma - xx^\top\right)^\top\right]\right)}{n}$$

$\square$

Proof of Proposition 4.2.

Proof. We start by the proof of the following lemma

Lemma 6.1. $\hat{\gamma}_m^2$ *is an unbiased estimator of* $tr\left((\Pi_m \otimes \Pi_m)\Phi\right)$.

Proof. We deduce from the proposition 2.1 and the fact that Π_m is a projection matrix that:

$$(n-1)\mathbb{E}\left[\hat{\gamma}_m^2\right] = \sum_{i=1}^n \mathbb{E}\left[\left\|vec\left(\Pi_m x_i x_i^\top \Pi_m\right) - vec\left(\hat{\Sigma}_m\right)\right\|^2\right]$$

$$= \sum_{i=1}^n \mathbb{E}\left[\left\|(\Pi_m \otimes \Pi_m)\left(vec\left(x_i x_i^\top\right) - vec\left(S\right)\right)\right\|^2\right]$$

$$= \sum_{i=1}^n \mathbb{E}\left[tr\left((\Pi_m \otimes \Pi_m)\left(vec\left(x_i x_i^\top\right) - vec\left(S\right)\right)\left(vec\left(x_i x_i^\top\right) - vec\left(S\right)\right)^\top (\Pi_m \otimes \Pi_m)^\top\right)\right]$$

$$= \sum_{i=1}^n tr\left((\Pi_m \otimes \Pi_m)\mathbb{E}\left[\left(vec\left(x_i x_i^\top\right) - vec\left(S\right)\right)\left(vec\left(x_i x_i^\top\right) - vec\left(S\right)\right)^\top\right]\right)$$

But if $(v_i)_{1 \leqslant i \leqslant n}$, are some i.i.d. vectors with covariance matrix V and mean $\bar{v} = \frac{1}{n}\sum_{i=1}^n v_i$, we have

$$\mathbb{E}\left[(v_i - \bar{v})(v_i - \bar{v})^\top\right] = \frac{1}{n^2}\sum_{j,k=1}^n \mathbb{E}\left[(v_i - v_k)(v_i - v_j)^\top\right]$$

$$= \frac{1}{n^2}\sum_{\substack{j,k=1 \\ j,k \neq i}}^n \mathbb{E}\left[(v_i - v_k)(v_i - v_j)^\top\right]$$

$$= \frac{1}{n^2}\left\{(n-1)\mathbb{E}\left[(v_1 - v_2)(v_1 - v_2)^\top\right] + (n-2)(n-1)\mathbb{E}\left[(v_1 - v_2)(v_1 - v_3)^\top\right]\right\}$$

$$= \frac{1}{n^2}\left\{(n-1)2V + (n-2)(n-1)V\right\}$$

$$= \frac{1}{n^2}\left((n-1)nV\right)$$

Hence

$$\mathbb{E}\left[(v_i - \bar{v})(v_i - \bar{v})^\top\right] = \frac{1}{n}\left((n-1)V\right)$$

this identity gives

$$(n-1)\mathbb{E}\left[\hat{\gamma}_m^2\right] = \sum_{i=1}^n tr\left((\Pi_m \otimes \Pi_m)\frac{1}{n}\left((n-1)\Phi\right)\right)$$

Finally

$$\mathbb{E}\left[\hat{\gamma}_m^2\right] = tr\left((\Pi_m \otimes \Pi_m)\Phi\right)$$

$\square$

Now, it remains to show that

$$\mathbb{E}\left[\left\|S - \hat{\Sigma}_m\right\|^2\right] = \|\Sigma - \Pi_m\Sigma\Pi_m\|^2 - \frac{tr\left((\Pi_m \otimes \Pi_m)\,\Phi\right)}{n} + \frac{tr\,(\Phi)}{n}$$

We have that

$$\left\|S - \hat{\Sigma}_m\right\|^2 = \|S - \Sigma\|^2 + 2\left\langle S - \Sigma, \Sigma - \hat{\Sigma}_m\right\rangle + \left\|\Sigma - \hat{\Sigma}_m\right\|^2$$

And using the orthogonality we deduce that

$$\left\|S - \hat{\Sigma}_m\right\|^2 = \|S - \Sigma\|^2 + 2\left\langle S - \Sigma, \Sigma - \hat{\Sigma}_m\right\rangle + \|\Sigma - \Sigma_m\|^2 + \left\|\Sigma_m - \hat{\Sigma}_m\right\|^2$$

For the same reason :

$$\left\langle S - \Sigma, \Sigma - \hat{\Sigma}_m\right\rangle = \langle S - \Sigma, \Sigma - \Sigma_m\rangle + \left\langle S - \Sigma, \Sigma_m - \hat{\Sigma}_m\right\rangle$$

$$= \langle S - \Sigma, \Sigma - \Sigma_m\rangle - \left\|\Sigma_m - \hat{\Sigma}_m\right\|^2$$

And because the expectation of S is equal to Σ we obtain that

$$\mathbb{E}\left[\left\|S - \hat{\Sigma}_m\right\|^2\right] = \|\Sigma - \Sigma_m\|^2 + \mathbb{E}\left[\|S - \Sigma\|^2\right] - \mathbb{E}\left[\left\|\Sigma_m - \hat{\Sigma}_m\right\|^2\right]$$

First

$$\mathbb{E}\left[\|S - \Sigma\|^2\right] = \frac{1}{n^2}\mathbb{E}\left[\sum_{i,j=1}^{n}\left\langle x_ix_i^\top - \Sigma, x_jx_j^\top - \Sigma\right\rangle\right] = \frac{1}{n}\mathbb{E}\left[\left\|xx^\top - \Sigma\right\|^2\right]$$

And with the properties of the Frobenius norm

$$\mathbb{E}\left[\left\|xx^\top - \Sigma\right\|^2\right] = \mathbb{E}\left[\left\|vec\left(xx^\top - \Sigma\right)\right\|^2\right]$$

$$= tr\left(\mathbb{E}\left[\left(vec\left(xx^\top\right) - vec\,(\Sigma)\right)\left(vec\left(xx^\top\right) - vec\,(\Sigma)\right)^\top\right]\right)$$

then we derive that

$$\mathbb{E}\left[\left\|xx^\top - \Sigma\right\|^2\right] = tr\,(\Phi)$$

Thus

$$\mathbb{E}\left[\|S - \Sigma\|^2\right] = \frac{tr\,(\Phi)}{n} \tag{10}$$

Second

$$\mathbb{E}\left[\left\|\Sigma_m - \hat{\Sigma}_m\right\|^2\right] = \frac{1}{n^2}\mathbb{E}\left[\sum_{i,j=1}^{n}\left\langle \Pi_m\left(x_ix_i^\top - \Sigma\right)\Pi_m, \Pi_m\left(x_jx_j^\top - \Sigma\right)\Pi_m\right\rangle\right]$$

$$= \frac{1}{n}\mathbb{E}\left[\left\langle \Pi_m\left(xx^\top - \Sigma\right)\Pi_m, \Pi_m\left(xx^\top - \Sigma\right)\Pi_m\right\rangle\right] = \frac{1}{n}\mathbb{E}\left[\left\|\Pi_m\left(xx^\top - \Sigma\right)\Pi_m\right\|^2\right]$$

And using the proposition 2.1 and the specificity of Π_m, we obtain that

$$\mathbb{E}\left[\left\|\Pi_m\left(xx^\top - \Sigma\right)\Pi_m\right\|^2\right] = \mathbb{E}\left[\left\|vec\left(\Pi_m\left(xx^\top - \Sigma\right)\Pi_m\right)\right\|^2\right]$$

$$= \mathbb{E}\left[\left\|\Pi_m \otimes \Pi_m \left(vec\left(xx^\top - \Sigma\right)\right)\right\|^2\right]$$

$$= \mathbb{E}\left[tr\left(\Pi_m \otimes \Pi_m \left(vec\left(xx^\top - \Sigma\right)\left(vec\left(xx^\top - \Sigma\right)\right)^\top \left(\Pi_m \otimes \Pi_m\right)^\top\right)\right)\right]$$

$$= \mathbb{E}\left[tr\left(\left(\Pi_m \otimes \Pi_m\right)\left(vec\left(xx^\top - \Sigma\right)\right)\left(vec\left(xx^\top - \Sigma\right)\right)^\top\right)\right]$$

Hence

$$\mathbb{E}\left[\left\|\Pi_m\left(xx^\top - \Sigma\right)\Pi_m\right\|^2\right] = tr\left(\left(\Pi_m \otimes \Pi_m\right)\Phi\right)$$

And we obtain

$$\mathbb{E}\left[\left\|\Sigma_m - \hat{\Sigma}_m\right\|^2\right] = \frac{tr\left(\left(\Pi_m \otimes \Pi_m\right)\Phi\right)}{n} \tag{11}$$

Finally, we have

$$\mathbb{E}\left[\left\|S - \hat{\Sigma}_m\right\|^2\right] = \|\Sigma - \Sigma_m\|^2 - \frac{tr\left(\left(\Pi_m \otimes \Pi_m\right)\Phi\right)}{n} + \frac{tr\left(\Phi\right)}{n}$$

$\square$

Proof of Proposition 4.3.

Proof. As $\frac{\hat{\gamma}_m^2}{n} \geqslant 0$, we have:

$$\mathbb{E}\left[\left\|\hat{\Sigma} - \Sigma\right\|^2\right] \leqslant \mathbb{E}\left[\left\|\hat{\Sigma} - S\right\|^2 + 2\frac{\hat{\gamma}_{\hat{m}}^2}{n}\right] + 2\mathbb{E}\left[\left\langle\hat{\Sigma} - S, S - \Sigma\right\rangle\right] + \mathbb{E}\left[\|S - \Sigma\|^2\right]$$

Let $m_0 \in \underset{m\in\mathcal{M}}{argmin}\,\mathbb{E}\left[\left\|\Sigma - \hat{\Sigma}_m\right\|^2\right]$ an oracle. By definition of $\hat{m}$,

$$\left\|S - \hat{\Sigma}_{\hat{m}}\right\|^2 + 2\frac{\hat{\gamma}_{\hat{m}}^2}{n} \leqslant \left\|S - \hat{\Sigma}_{m_0}\right\|^2 + 2\frac{\hat{\gamma}_{m_0}^2}{n}$$

Then

$$\mathbb{E}\left[\left\|\hat{\Sigma} - \Sigma\right\|^2\right] \leqslant \mathbb{E}\left[\left\|S - \hat{\Sigma}_{m_0}\right\|^2 + 2\frac{\hat{\gamma}_{m_0}^2}{n}\right] + \mathbb{E}\left[\|S - \Sigma\|^2\right] + 2\mathbb{E}\left[\left\langle\hat{\Sigma} - S, S - \Sigma\right\rangle\right]$$

we derive from the previous proposition and (10)

$$\mathbb{E}\left[\left\|\hat{\Sigma} - \Sigma\right\|^2\right] \leqslant \mathbb{E}\left[\left\|\Sigma - \hat{\Sigma}_{m_0}\right\|^2\right] + 2\frac{tr\left(\Phi\right)}{n} + 2\mathbb{E}\left[\left\langle\hat{\Sigma} - S, S - \Sigma\right\rangle\right]$$

Moreover by the Cauchy-Schwarz inequality we have that

$$\left\langle\hat{\Sigma} - S, S - \Sigma\right\rangle \leqslant \left\|\hat{\Sigma} - S\right\|\|S - \Sigma\|$$

And using again this inequality

$$\mathbb{E}\left[\left\langle\hat{\Sigma} - S, S - \Sigma\right\rangle\right] \leqslant \sqrt{\mathbb{E}\left[\left\|\hat{\Sigma} - S\right\|^2\right]}\sqrt{\mathbb{E}\left[\|S - \Sigma\|^2\right]}$$

$$\leqslant \sqrt{\mathbb{E}\left[\left\|\hat{\Sigma} - S\right\|^2 + 2\frac{\hat{\gamma}_{\hat{m}}^2}{n}\right]}\sqrt{\frac{tr\left(\Phi\right)}{n}}$$

12

For the same reasons as before we obtain

$$\mathbb{E}\left[\left\langle \hat{\Sigma} - S, S - \Sigma \right\rangle\right] \leqslant \sqrt{\mathbb{E}\left[\left\|\Sigma - \hat{\Sigma}_{m_0}\right\|^2\right]} + \frac{tr\left(\Phi\right)}{n}\sqrt{\frac{tr\left(\Phi\right)}{n}}$$

$$\leqslant \frac{tr\left(\Phi\right)}{n} + \sqrt{\mathbb{E}\left[\left\|\Sigma - \hat{\Sigma}_{m_0}\right\|^2\right]}\sqrt{\frac{tr\left(\Phi\right)}{n}}$$

Thus

$$\mathbb{E}\left[\left\|\hat{\Sigma} - \Sigma\right\|^2\right] \leqslant \mathbb{E}\left[\left\|\Sigma - \hat{\Sigma}_{m_0}\right\|^2\right] + 4\frac{tr\left(\Phi\right)}{n} + 2\sqrt{\mathbb{E}\left[\left\|S - \hat{\Sigma}_{m_0}\right\|^2\right]}\sqrt{\frac{tr\left(\Phi\right)}{n}}$$

With the following inequality which holds $\forall a, b \in \mathbb{R}$ et $\forall A > 0$

$$2ab \leqslant \frac{a^2}{A} + Ab^2$$

We obtain for all $A > 0$:

$$\mathbb{E}\left[\left\|\hat{\Sigma} - \Sigma\right\|^2\right] \leqslant \mathbb{E}\left[\left\|\Sigma - \hat{\Sigma}_{m_0}\right\|^2\right]\left(1 + A^{-1}\right) + \frac{tr\left(\Phi\right)}{n}\left(4 + A\right)$$

The definition of m_0 gives the result. $\qquad\square$

References

[Adl90]   Robert J. Adler. *An introduction to continuity, extrema, and related topics for general gaussian processes.* Institute of mathematical statistics, 1990.

[BBLA11]   Jrmie Bigot, Rolando Biscay, Jean-Michel Loubes, and Lilian Muniz Alvarez. Group lasso estimation of high-dimensional covariance matrices. *Journal of Machine Learning Resarch*, 2011.

[BBLLMA10] Jérémie Bigot, Rolando Biscay Lirio, Jean-Michel Loubes, and Lilian Muniz Alvarez. Adaptive estimation of spectral densities via wavelet thresholding and information projection. May 2010.

[BBLMA10]   Jérémie Bigot, Rolando Biscay, Jean-Michel Loubes, and Lilian Muñiz-Alvarez. Nonparametric estimation of covariance functions by model selection. *Electron. J. Stat.*, 4:822–855, 2010.

[BL08]   Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *Ann. Statist.*, 36(6):2577–2604, 2008.

[Cre93]   Noel A. C. Cressie. *Statistics for spatial data.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1993. Revised reprint of the 1991 edition, A Wiley-Interscience Publication.

[CY10]   Tony Cai and Ming Yuan. Nonparametric covariance function estimation for functional and longitudinal data. Technical report, 2010.

[EPTA08]   S. N. Elogne, O. Perrin, and C. Thomas-Agnan. Non parametric estimation of smooth stationary covariance functions by interpolation methods. *Stat. Inference Stoch. Process.*, 11(2):177–205, 2008.

[Jou77]   A. G. Journel. Kriging in terms of projections. *J. Internat. Assoc. Mathematical Geol.*, 9(6):563–586, 1977.

[Rao73]   C. Radhakrishna Rao. *Linear statistical inference and its applications*. Wiley, second edition, 1973.

[Seb08]   Georges A. F. Seber. *A matrix handbook for statisticians*. Wiley, 2008.

[Ste81]   Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

[Ste99]   Michael L. Stein. *Interpolation of spatial data*. Springer Series in Statistics. Springer-Verlag, New York, 1999. Some theory for Kriging.

[SW86]   Galen R. Shorack and Jon A. Wellner. *Empirical processes with applications to statistics*. Wiley, 1986.

[Tsy04]   Alexandre B. Tsybakov. *Introduction à l'estimation non-paramétrique*. Springer, 2004.