ESAIM: PS 18 (2014) 251–264
DOI: 10.1051/ps/2013034

ESAIM: Probability and Statistics
www.esaim-ps.org

UNBIASED RISK ESTIMATION METHOD FOR COVARIANCE ESTIMATION

Hélène Lescornel[1], Jean-Michel Loubes[1] and Claudie Chabriac[1]

Abstract. We consider a model selection estimator of the covariance of a random process. Using the Unbiased Risk Estimation (U.R.E.) method, we build an estimator of the risk which allows to select an estimator in a collection of models. 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.

Mathematics Subject Classification. 62G05.

Received June 7, 2012. Revised December 7, 2012.

1. Introduction

The estimation of covariance function of stochastic processes lies at the core of many statistical applications, ranging from geostatistics, financial series to epidemiology for instance (we refer to [15], [10] or [8] for general references). A large literature exists for parametric methods, see for instance in [8] for a review. Non-parametric procedures have only recently received attention, see for instance [3–5, 9] and references therein. Besides the estimation issue, we will focus on estimators which are true covariance functions, preventing the direct use of usual non-parametric 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 $T \subset \mathbb{R}^d$, where $d \in \mathbb{N}$. 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 $j \in \{1 \dots p\}$. Note that the observation points t_j are fixed and that the X_i's are independent copies of the process X. Set $\mathbf{x}_i = (X_i(t_1), \dots, X_i(t_p))^\top \ \forall i \in \{1 \dots n\}$ and denote by $\mathbf{\Sigma}$ the covariance matrix of these vectors. In this work, p is fixed while the asymptotic depends on n, the number of replications of the process.

Following the methodology presented in [4], 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 $\mathcal{M}_n \subset \mathcal{P}(\Lambda)$ allowed to grow with the number n of replications of the process. For $m \subset \mathcal{M}_n$, a finite number of indices, the process can be approximated by

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

Keywords and phrases. Covariance estimation, model selection, U.R.E. method.

[1] Institut de Mathématiques de Toulouse UMR 5219, 31062 Toulouse, Cedex 9, France.
helene.lescornel@math.univ-toulouse.fr; loubes@math.univ-toulouse.fr; chabriac@univ-tlse2.fr

Article published by EDP Sciences

Such an approximation leads to an estimator of $\mathbf{\Sigma}$ depending on the collection of functions m, denoted by $\hat{\mathbf{\Sigma}}_m$. 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

$$m_0 \in \arg \min_{m \in \mathcal{M}_n} R(m) = \arg \min_{m \in \mathcal{M}_n} \mathbb{E}\left[\left\| \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_m \right\|^2 \right],$$

where $\|\mathbf{A}\|$ denotes the Frobenius norm of the matrix $\mathbf{A}$. The Frobenius matrix norm provides a meaningful metric for comparing covariance matrices, widely used in multivariate analysis, in particular in the theory of principal components analysis. See also [6] and references therein for other applications of this loss function.

A model selection procedure will be performed using the U.R.E. method, which has been introduced in [14] and fully described in [16]. 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, $\hat{\mathbf{\Sigma}}_{\hat{m}}$ will be an estimator with a small risk, nearly as the best quantity $\hat{\mathbf{\Sigma}}_{m_0}$.

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 estimators. In Section 4 we apply the U.R.E. method to select one of them, and provide 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 fixed points $t_1, ..., t_p$ in T. We set $\mathbf{x}_i = (X_i(t_1), \ldots, X_i(t_p))^\top$, and denote the empirical covariance of the data by

$$\mathbf{S} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top$$

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

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

$$\mathbf{x}_i \mathbf{x}_i^\top = \mathbf{\Sigma} + \mathbf{U_i} \quad \in \mathbb{R}^{p \times p}, \quad 1 \leqslant i \leqslant n \tag{2.1}$$

where $\mathbf{U_i}$ are independent and identically distributed (i.i.d.) error matrices with $\mathbb{E}[\mathbf{U_i}] = 0$.

We now recall some notations related to the study of matrices, which will be used in the following. More details can be found in [12] and in [11].

For any matrix $\mathbf{A} = (a_{ij})_{1 \leqslant i \leqslant s, 1 \leqslant j \leqslant t} \in \mathbb{R}^{s \times t}$, $\|\mathbf{A}\|^2 = \mathrm{Tr}\left(\mathbf{A}\mathbf{A}^\top\right)$ is the Frobenius norm of the matrix which is associated to the inner scalar product $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}\left(\mathbf{A}\mathbf{B}^\top\right)$.

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

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 $s \times t$ matrix $\mathbf{A} = (a_{ij})_{1 \leq i \leq s, 1 \leq j \leq t}$ is the $st \times 1$ column vector denoted by $vec(\mathbf{A})$, obtained by stacking the columns of the matrix on top of one another, that is $vec(\mathbf{A}) = (a_{11}, ..., a_{s1}, a_{12}, ..., a_{s2}, ..., a_{1t}, ..., a_{st})^\top$.

If $\mathbf{A} = (a_{ij})_{1\le i\le s, 1\le j\le t}$ is a $s\times t$ matrix and $\mathbf{B} = (b_{ij})_{1\le i\le p,1\le j\le q}$ is a $p\times q$ matrix, then the Kronecker product of the two matrices, denoted by $\mathbf{A}\otimes\mathbf{B}$, is the $sp\times tq$ block matrix

$$\mathbf{A}\otimes\mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1t}\mathbf{B} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ a_{s1}\mathbf{B} & \dots & a_{st}\mathbf{B} \end{bmatrix}.$$

For $q\in\mathbb{N}$, $\mathcal{S}_q$ denotes the linear subspace of $\mathbb{R}^{q\times q}$ composed of symmetric matrices. For $\mathbf{G}\in\mathbb{R}^{p\times q}$, $\mathcal{S}(\mathbf{G})$ is the linear subspace of $\mathbb{R}^{p\times p}$ defined by

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

This set will be a natural candidate to select covariance estimators.

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. This leads to an estimator of the covariance of the process depending on the functions of this finite basis. By choosing different basis, we obtain several covariance estimators. Then, we construct a rule to pick out the best of these estimators among this collection of estimates, 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 [4].

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

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

where $\varLambda$ is a countable set, and $(a_\lambda)_{\lambda\in\varLambda}$ are the 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 in a Hilbert space, a natural choice of the functions is given by the corresponding Hilbert basis $(g_\lambda)_{\lambda\in\varLambda}$. Alternatively, the Karhunen-Loeve expansion of the covariance also provides a natural basis. However, since it relies on the nature of the process X, this expansion is usually unknown or requires additional information on the process. We refer to [1] for more references on this expansion. Under other regularity assumptions on the process, for instance assuming that the paths of the process belong to some R.K.H.S. (Reproducing Kernel Hilbert Space), other expansions can be considered as in [7] for instance.

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

$$\tilde{X}(t) = \sum_{\lambda\in m} a_\lambda g_\lambda(t). \tag{3.1}$$

We note $\mathbf{G}_m\in\mathbb{R}^{p\times|m|}$ where $(\mathbf{G}_m)_{j\lambda} = g_\lambda(t_j)$ and $\mathbf{a}_m$ the random vector of $\mathbb{R}^{|m|}$ with coefficients $(a_\lambda)_{\lambda\in m}$. Hence, we obtain that

$$\tilde{\mathbf{x}} = \left(\tilde{X}(t_1), ..., \tilde{X}(t_p)\right)^\top = \mathbf{G}_m\mathbf{a}_m$$

and

$$\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top = \mathbf{G}_m\mathbf{a}_m\mathbf{a}_m^\top\mathbf{G}_m^\top.$$

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

Assume that the subset m is fixed. The best approximation of $\mathbf{\Sigma}$ in $\mathcal{S}\left(\mathbf{G}_m\right)$ for the Frobenius norm is its projection denoted by $\mathbf{\Sigma}_m$. Since $\mathbf{\Sigma}$ is unknown, we consider the estimator built using the projection of $\mathbf{S}$ onto $\mathcal{S}\left(\mathbf{G}_m\right)$. We denote this quantity by $\hat{\mathbf{\Sigma}}_m$.

Proposition 3.1 in [4] gives an explicit form for these projections. We recall it for sake of completeness.

Proposition 3.1 (Description of the projected covariance)**.**
Let $\mathbf{A}$ *in* $\mathbb{R}^{p\times p}$ *and* $\mathbf{G}\in\mathbb{R}^{p\times|m|}$. *The infimum*

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

is achieved at

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

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

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

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

Remark 3.2. Thanks to the properties of the reflexive generalized inverse given in [11], the projection of a non–negative definite matrix $\mathbf{A}\in\mathcal{S}_p$ on $\mathcal{S}\left(\mathbf{G}\right)$ will be also a non–negative definite matrix. Moreover, the matrix $\mathbf{\Pi}$ does not depend on the choice of the generalized inverse.

With this result, the projection of $\mathbf{\Sigma}$ on $\mathcal{S}\left(\mathbf{G}_m\right)$ can be characterized as

$$\mathbf{\Sigma}_m=\mathbf{\Pi}_m\mathbf{\Sigma}\mathbf{\Pi}_m \tag{3.2}$$

and the same for $\mathbf{S}$ (that is, our candidate for estimating $\mathbf{\Sigma}$)

$$\hat{\mathbf{\Sigma}}_m=\mathbf{\Pi}_m\mathbf{S}\mathbf{\Pi}_m \tag{3.3}$$

where $\mathbf{\Pi}_m=\mathbf{G}_m\left(\mathbf{G}_m^\top\mathbf{G}_m\right)^-\mathbf{G}_m^\top$.

Note that the previous remark implies that the estimator $\hat{\mathbf{\Sigma}}_m$ 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 $\mathcal{M}_n$ be a finite collection of models m whose size may grow with the number n of replications of the process. In this section, we focus on picking the best model among this collection by following the U.R.E. method. Since the law of $\left\|\mathbf{\Sigma}-\hat{\mathbf{\Sigma}}_m\right\|$ is unknown, we thus aim at finding an estimator of its expectation.

We consider that the best subset m is m_0 defined by

$$m_0\in\arg\min_{m\in\mathcal{M}_n}\mathbb{E}\left[\left\|\mathbf{\Sigma}-\hat{\mathbf{\Sigma}}_m\right\|^2\right].$$

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

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

Proposition 4.1.

$$\mathbb{E}\left[\left\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}_m\right\|^2\right]=\left\|\boldsymbol{\Sigma}-\boldsymbol{\Pi}_m\boldsymbol{\Sigma}\boldsymbol{\Pi}_m\right\|^2+\frac{\operatorname{Tr}\left(\left(\boldsymbol{\Pi}_m\otimes\boldsymbol{\Pi}_m\right)\boldsymbol{\Phi}\right)}{n} \tag{4.1}$$

where $\boldsymbol{\Phi}=Var\left(vec\left(\mathbf{x}\mathbf{x}^\top\right)\right)$.

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

$$\mathbf{Y}=\theta+\epsilon\xi\quad\xi\sim\mathcal{N}\left(0,\mathbf{I}_n\right)$$

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 $\mathbf{Y}$ onto some subspace S_m. If the subspace dimension is denoted by D_m, the risk of such an estimator is given by

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

We thus recognize the same kind of decomposition with a bias term and with $\frac{\operatorname{Tr}((\boldsymbol{\Pi}_m\otimes\boldsymbol{\Pi}_m)\boldsymbol{\Phi})}{n}$ 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 (2.1).

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

$$\hat{\gamma}_m^2=\frac{1}{n-1}\sum_{i=1}^n\left\|\boldsymbol{\Pi}_m\mathbf{x}_i\mathbf{x}_i^\top\boldsymbol{\Pi}_m-\hat{\boldsymbol{\Sigma}}_m\right\|^2$$

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

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

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

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

$$\hat{\boldsymbol{\Sigma}}=\boldsymbol{\Pi}_{\hat{m}}\mathbf{S}\boldsymbol{\Pi}_{\hat{m}}=\hat{\boldsymbol{\Sigma}}_{\hat{m}}$$
$$\text{with }\hat{m}\in\arg\min_{m\in\mathcal{M}_n}\left(\left\|\mathbf{S}-\hat{\boldsymbol{\Sigma}}_m\right\|^2+2\frac{\hat{\gamma}_m^2}{n}\right)\cdot$$

The next theorem establishes an oracle inequality for this estimator.

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

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\right\|^2\right]\leqslant\left(1+\varepsilon^{-1}\right)\inf_{m\in\mathcal{M}_n}\mathbb{E}\left[\left\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}_m\right\|^2\right]+\frac{\operatorname{Tr}\left(\boldsymbol{\Phi}\right)}{n}\left(4+\varepsilon\right).$$

Recall that $\boldsymbol{\Phi}=\operatorname{Var}\left(vec\left(\mathbf{x}\mathbf{x}^\top\right)\right)\in\mathbb{R}^{p^2\times p^2}$ which does not depend on $\mathcal{M}_n$ nor n. More precisely, some standard computations leads to $\operatorname{Tr}\left(\boldsymbol{\Phi}\right)=\sum_{j=1}^p\sum_{i=1}^p\operatorname{Var}\left(X\left(t_i\right)X\left(t_j\right)\right)$. Hence, increasing the collection of models does not affect the remainder term in the oracle inequality. So increasing the size of $\mathcal{M}_n$ leads to a better estimation. However, this may lead to computational problems since the estimator $\hat{\Sigma}$ is built by minimizing some functional on $\mathcal{M}_n$.

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 expense of replicate i.i.d observations of the process. Hence, this study requires a large number of replications n with respect to the number of observation points p and it is illustrated by the previous remarks on the penalty term $\frac{\mathrm{Tr}(\mathbf{\Phi})}{n}\cdot$ Actually our method is not designed to tackle the problem of covariance estimation in the high dimensional case $p \gg n$. This topic has received a growing attention over the past years and we refer to [2] and references therein for a survey.

We also stress that, in this work, an appropriate data-based subset of indices $m \in \mathcal{M}_n$ is chosen in order to obtain a good approximation for the covariance. This *dimension* corresponds to Model (2.1) and is very distinct in Model (3.1). Indeed, model selection for (3.1) depends on the variability of the vectors $\mathbf{x}_i$'s while for (2.1), it depends on the variability of the matrices $\mathbf{x}_i\mathbf{x}_i^\top$' s.

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

5. Numerical examples

In this section we illustrate the behaviour of the covariance estimator $\hat{\mathbf{\Sigma}}$ with programs implemented using SCILAB. We want to assess whether our procedure selects 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 at which 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 $\mathbf{G}_m$: recall that $(\mathbf{G}_m)_{j\lambda} = g_\lambda(t_j)$.

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 the model selection estimator $\hat{\mathbf{\Sigma}}$.

For each example, we plot the curve of the risk function (respectively the curve of the estimator of the risk) and give the value where it reaches its minimum m_0 (respectively $\hat{m}$). Next we compare the true covariance and the estimator. Finally, we repeat the computation procedure for estimating $\hat{m}$ by simulating 100 different samples. The histogram of the empirical distribution of our estimator illustrates its asymptotic behaviour.

Example 1. Here we work with the numerical examples of [4]. We choose the Fourier basis functions

$$g_\lambda(t) = \begin{cases} \frac{1}{\sqrt{p}} \text{ if } \lambda = 1 \\ \sqrt{2}\frac{1}{\sqrt{p}}\cos(2\pi\frac{\lambda}{2}t) \text{ if } \lambda \text{ is even} \\ \sqrt{2}\frac{1}{\sqrt{p}}\sin(2\pi\frac{\lambda-1}{2}t) \text{ if } \lambda \text{ is odd} \end{cases}$$

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 $\mathbf{D}(\mathbf{V})$ the diagonal matrix in $m^\star \times m^\star$ such as $D(V)_{\lambda\lambda} = V(a_\lambda)$. Then we have

$$\mathbf{\Sigma} = \mathbf{G}_{m^\star}\mathbf{D}(\mathbf{V})\mathbf{G}_{m^\star}^\top.$$

Here are the results for $V(a_\lambda) = 0.0475 + 0.95^\lambda \quad \forall \quad \lambda$, and $m^\star = 35 = p$, $n = 60$, $M = 34$. The figures show that $m_0 = \hat{m} = 18$ for the sample considered for the estimation of Σ.

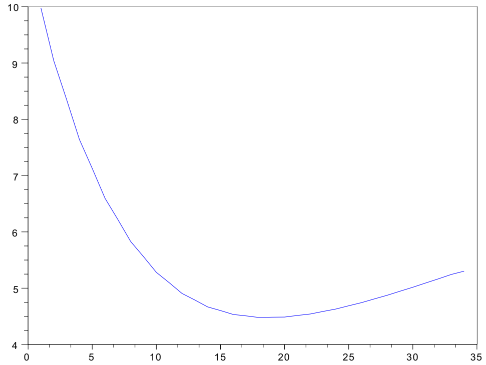

FIGURE 1. Risk function $R\,(m) = \mathbb{E}\left[\left\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_m\right\|^2\right]$.

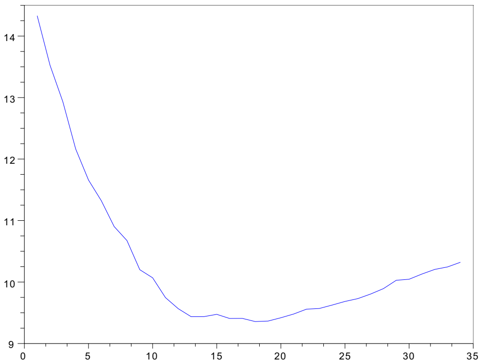

FIGURE 2. Estimator of the risk function $\hat{R}\,(m) = \left\|\mathbf{S} - \hat{\mathbf{\Sigma}}_m\right\|^2 + 2\frac{\hat{\gamma}_m^2}{n}$.

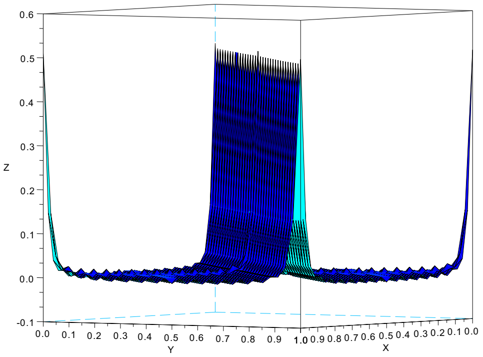

FIGURE 3. Covariance Σ.

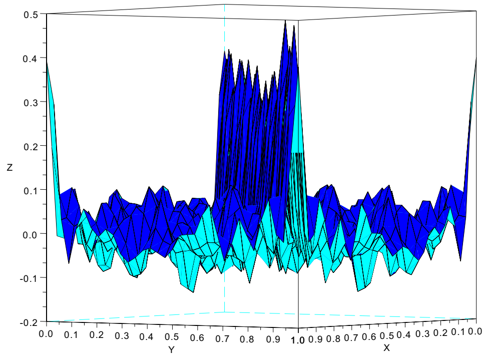

FIGURE 4. Estimator $\hat{\Sigma}$.

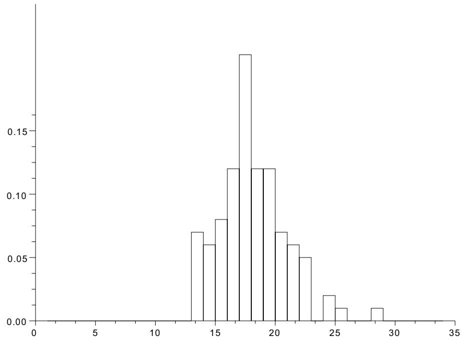

FIGURE 5. Distribution of $\hat{m}$.

Example 2. Now we test our estimator with the process studied in [7].

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 $\zeta_\lambda = \frac{(-1)^{\lambda+1}}{\lambda^2}\cdot$ If $\mathbf{D}$ is the diagonal matrix with entries $D_{\lambda\lambda} = \frac{1}{\lambda^4}$, as before we have that

$$\mathbf{\Sigma} = \mathbf{G}_{m^\star}\mathbf{D}\mathbf{G}_{m^\star}^\top.$$

Here we choose $m^\star = 50$, $n = 1000$, $p = 40$ and $M = 20$. We found $m_0 = 4 = \hat{m}$ for the sample used for the estimation.

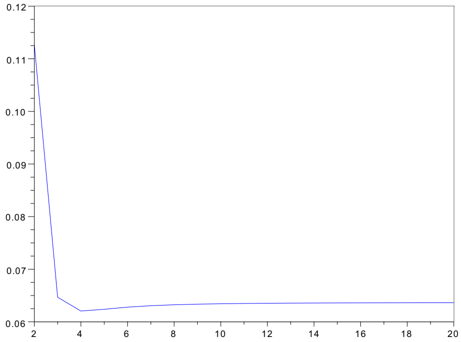

Figure 6. Risk function $R(m) = \mathbb{E}\left[\left\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_m\right\|^2\right]$.

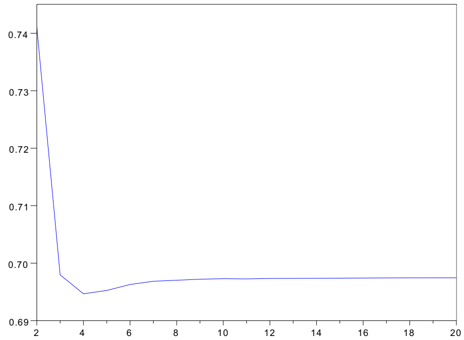

Figure 7. Estimator of the risk function $\hat{R}(m) = \left\|\mathbf{S} - \hat{\mathbf{\Sigma}}_m\right\|^2 + 2\frac{\hat{\gamma}_m^2}{n}$.

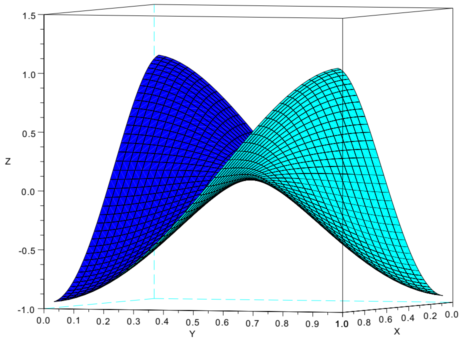

Figure 8. Covariance $\varSigma$.

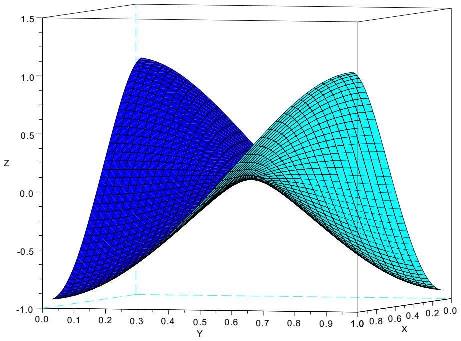

Figure 9. Estimator $\hat{\varSigma}$.

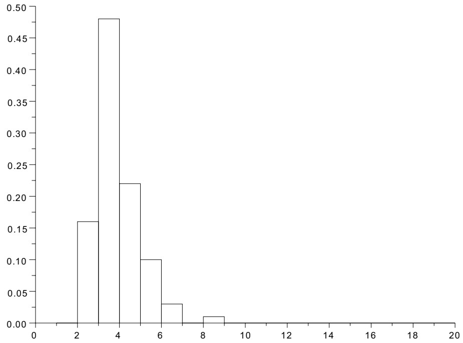

Figure 10. Distribution of $\hat{m}$.

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

Recall that the covariance function of the Brownian bridge is $K(s,t) = s(1-t)$ for $s \leqslant t$.

Here $n = 100$, $p = 35$ and $M = 20$. We found $m_0 = 5 = \hat{m}$.

In each case, the curve of the risk R and its estimator $\hat{R}$ have the same form but differ by a translation along the vertical axis. In fact, the parameter of this translation corresponds to the quantity $\frac{\mathrm{Tr}(\mathbf{\Phi})}{n}$ in the equation $\mathbb{E}\left[\hat{R}(m)\right] = R(m) + \frac{\mathrm{Tr}(\mathbf{\Phi})}{n}$ of Proposition 4.2. Moreover, even in the case where the size of the sample is not so large (the first and the last example), yet the covariance estimator shows good performances.

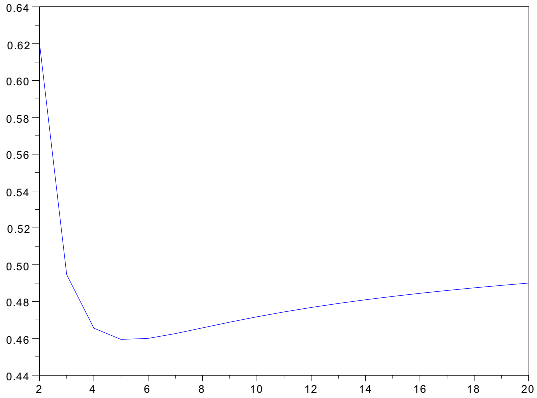

FIGURE 11. Risk function $R(m) = \mathbb{E}\left[\left\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_m\right\|^2\right]$.

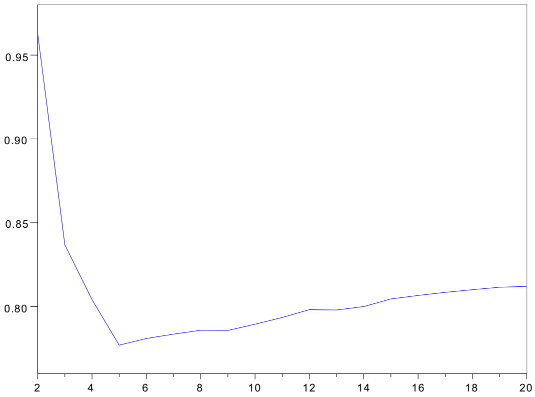

FIGURE 12. Estimator of the risk function $\hat{R}(m) = \left\|\mathbf{S} - \hat{\mathbf{\Sigma}}_m\right\|^2 + 2\frac{\hat{\gamma}_m^2}{n}$.

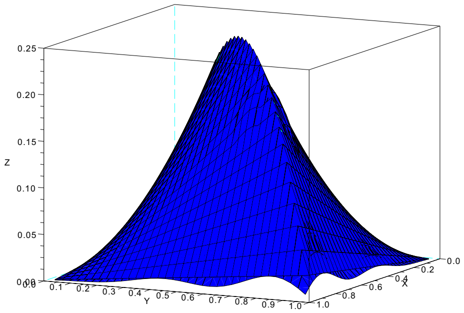

FIGURE 13. Covariance Σ.

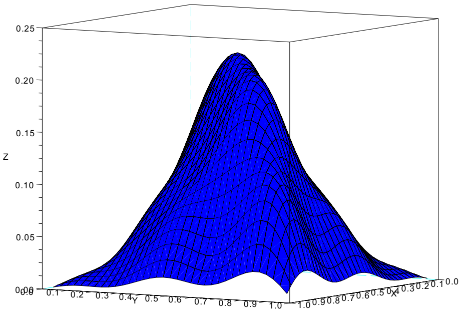

FIGURE 14. Estimator $\hat{\Sigma}$.

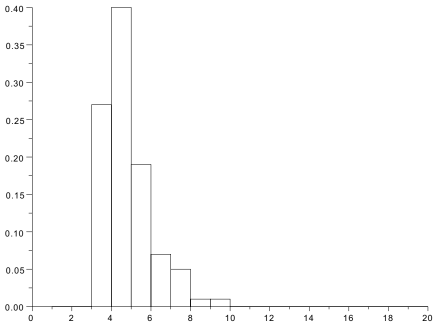

FIGURE 15. Distribution of $\hat{m}$.

6. APPENDIX

In the proofs, we will use the following results of linear algebra where $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ some real matrices

$$\begin{aligned} &vec\left(\mathbf{ABC}\right) = \left(\mathbf{C}^\top \otimes \mathbf{A}\right) vec\left(\mathbf{B}\right)\\ &\|\mathbf{A}\| = \|vec\left(\mathbf{A}\right)\| = \|vec\left(\mathbf{A}\right)\|_{\ell_2}\\ &\left(\mathbf{A}\otimes\mathbf{B}\right)\left(\mathbf{C}\otimes\mathbf{D}\right) = \left(\mathbf{AC}\right)\otimes\left(\mathbf{BD}\right)\\ &\left(\mathbf{A}\otimes\mathbf{B}\right)^\top = \mathbf{A}^\top\otimes\mathbf{B}^\top. \end{aligned} \tag{6.1}$$

We refer to [12] for their proofs.

Recall that $\mathbf{\Sigma}_m = \mathbf{\Pi}_m\mathbf{\Sigma}\mathbf{\Pi}_m$, $\hat{\mathbf{\Sigma}}_m = \mathbf{\Pi}_m\mathbf{S}\mathbf{\Pi}_m$ and

$$\hat{\gamma}_m^2 = \frac{1}{n-1}\sum_{i=1}^n \left\|\mathbf{\Pi}_m\mathbf{x}_i\mathbf{x}_i^\top\mathbf{\Pi}_m - \hat{\mathbf{\Sigma}}_m\right\|^2.$$

We start by proving Proposition 4.1.

Proof. Using the orthogonality, we have

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

With the equalities (6.1) we deduce

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

and since $\mathbf{\Pi}_m$ is a projection matrix,

$$\left\|\mathbf{\Sigma}_m - \hat{\mathbf{\Sigma}}_m\right\|^2 = \mathrm{Tr}\left(\left(\mathbf{\Pi}_m\otimes\mathbf{\Pi}_m\right) vec\left(\mathbf{\Sigma}-\mathbf{S}\right) vec\left(\mathbf{\Sigma}-\mathbf{S}\right)^\top\right).$$

Hence

$$\begin{aligned} \mathbb{E}\left[\left\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_m\right\|^2\right] &= \|\mathbf{\Sigma}-\mathbf{\Sigma}_m\|^2 + \mathbb{E}\left[\mathrm{Tr}\left(\left(\mathbf{\Pi}_m\otimes\mathbf{\Pi}_m\right) vec\left(\mathbf{\Sigma}-\mathbf{S}\right) vec\left(\mathbf{\Sigma}-\mathbf{S}\right)^\top\right)\right]\\ &= \|\mathbf{\Sigma}-\mathbf{\Sigma}_m\|^2 + \mathrm{Tr}\left(\left(\mathbf{\Pi}_m\otimes\mathbf{\Pi}_m\right)\mathbb{E}\left[vec\left(\mathbf{\Sigma}-\mathbf{S}\right) vec\left(\mathbf{\Sigma}-\mathbf{S}\right)^\top\right]\right)\\ &= \|\mathbf{\Sigma}-\mathbf{\Sigma}_m\|^2 + \frac{\mathrm{Tr}\left(\left(\mathbf{\Pi}_m\otimes\mathbf{\Pi}_m\right)\mathbb{E}\left[vec\left(\mathbf{\Sigma}-\mathbf{x}\mathbf{x}^\top\right) vec\left(\mathbf{\Sigma}-\mathbf{x}\mathbf{x}^\top\right)^\top\right]\right)}{n}. \end{aligned}$$

□

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* $\mathrm{Tr}\left(\left(\mathbf{\Pi}_m \otimes \mathbf{\Pi}_m\right)\mathbf{\Phi}\right)$.

Proof. We deduce from the equations (6.1) and the fact that $\mathbf{\Pi}_m$ is a projection matrix that

$$(n-1)\,\mathbb{E}\left[\hat{\gamma}_m^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[\left\| vec\left(\mathbf{\Pi}_m \mathbf{x}_i \mathbf{x}_i^\top \mathbf{\Pi}_m\right) - vec\left(\hat{\mathbf{\Sigma}}_m\right)\right\|^2\right]$$

$$= \sum_{i=1}^{n} \mathbb{E}\left[\left\|\left(\mathbf{\Pi}_m \otimes \mathbf{\Pi}_m\right)\left(vec\left(\mathbf{x}_i \mathbf{x}_i^\top\right) - vec\left(\mathbf{S}\right)\right)\right\|^2\right]$$

$$= \sum_{i=1}^{n} \mathbb{E}\left[\mathrm{Tr}\left(\left(\mathbf{\Pi}_m \otimes \mathbf{\Pi}_m\right)\left(vec\left(\mathbf{x}_i \mathbf{x}_i^\top\right) - vec\left(\mathbf{S}\right)\right)\left(vec\left(\mathbf{x}_i \mathbf{x}_i^\top\right) - vec\left(\mathbf{S}\right)\right)^\top \left(\mathbf{\Pi}_m \otimes \mathbf{\Pi}_m\right)^\top\right)\right]$$

$$= \sum_{i=1}^{n} \mathrm{Tr}\left(\left(\mathbf{\Pi}_m \otimes \mathbf{\Pi}_m\right)\mathbb{E}\left[\left(vec\left(\mathbf{x}_i \mathbf{x}_i^\top\right) - vec\left(\mathbf{S}\right)\right)\left(vec\left(\mathbf{x}_i \mathbf{x}_i^\top\right) - vec\left(\mathbf{S}\right)\right)^\top\right]\right).$$

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

$$\mathbb{E}\left[\left(\mathbf{v}_i - \bar{\mathbf{v}}\right)\left(\mathbf{v}_i - \bar{\mathbf{v}}\right)^\top\right] = \frac{1}{n^2}\sum_{j,k=1}^{n} \mathbb{E}\left[\left(\mathbf{v}_i - \mathbf{v}_k\right)\left(\mathbf{v}_i - \mathbf{v}_j\right)^\top\right]$$

$$= \frac{1}{n^2}\sum_{\substack{j,k=1\\ j,k\neq i}}^{n} \mathbb{E}\left[\left(\mathbf{v}_i - \mathbf{v}_k\right)\left(\mathbf{v}_i - \mathbf{v}_j\right)^\top\right]$$

$$= \frac{1}{n^2}\left\{(n-1)\,\mathbb{E}\left[\left(\mathbf{v}_1 - \mathbf{v}_2\right)\left(\mathbf{v}_1 - \mathbf{v}_2\right)^\top\right] + (n-2)\,(n-1)\,\mathbb{E}\left[\left(\mathbf{v}_1 - \mathbf{v}_2\right)\left(\mathbf{v}_1 - \mathbf{v}_3\right)^\top\right]\right\}$$

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

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

Hence

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

and this identity gives

$$(n-1)\,\mathbb{E}\left[\hat{\gamma}_m^2\right] = \sum_{i=1}^{n} \mathrm{Tr}\left(\left(\mathbf{\Pi}_m \otimes \mathbf{\Pi}_m\right)\frac{1}{n}\left((n-1)\,\mathbf{\Phi}\right)\right).$$

Finally

$$\mathbb{E}\left[\hat{\gamma}_m^2\right] = \mathrm{Tr}\left(\left(\mathbf{\Pi}_m \otimes \mathbf{\Pi}_m\right)\mathbf{\Phi}\right).$$

□

Now, it remains to show that

$$\mathbb{E}\left[\left\|\mathbf{S}-\hat{\boldsymbol{\Sigma}}_m\right\|^2\right]=\left\|\boldsymbol{\Sigma}-\boldsymbol{\Pi}_m\boldsymbol{\Sigma}\boldsymbol{\Pi}_m\right\|^2-\frac{\operatorname{Tr}\left(\left(\boldsymbol{\Pi}_m\otimes\boldsymbol{\Pi}_m\right)\boldsymbol{\Phi}\right)}{n}+\frac{\operatorname{Tr}\left(\boldsymbol{\Phi}\right)}{n}\cdot$$

We have that

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

and using the orthogonality we deduce

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

For the same reason

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

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

and because the expectation of $\mathbf{S}$ is equal to $\boldsymbol{\Sigma}$ we obtain that

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

First

$$\mathbb{E}\left[\left\|\mathbf{S}-\boldsymbol{\Sigma}\right\|^2\right]=\frac{1}{n^2}\mathbb{E}\left[\sum_{i,j=1}^{n}\left\langle\mathbf{x}_i\mathbf{x}_i^\top-\boldsymbol{\Sigma},\mathbf{x}_j\mathbf{x}_j^\top-\boldsymbol{\Sigma}\right\rangle\right]=\frac{1}{n}\mathbb{E}\left[\left\|\mathbf{x}\mathbf{x}^\top-\boldsymbol{\Sigma}\right\|^2\right].$$

With the properties of the Frobenius norm

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

$$=\operatorname{Tr}\left(\mathbb{E}\left[\left(vec\left(\mathbf{x}\mathbf{x}^\top\right)-vec\left(\boldsymbol{\Sigma}\right)\right)\left(vec\left(\mathbf{x}\mathbf{x}^\top\right)-vec\left(\boldsymbol{\Sigma}\right)\right)^\top\right]\right),$$

then we derive that

$$\mathbb{E}\left[\left\|\mathbf{x}\mathbf{x}^\top-\boldsymbol{\Sigma}\right\|^2\right]=\operatorname{Tr}\left(\boldsymbol{\Phi}\right),$$

thus

$$\mathbb{E}\left[\left\|\mathbf{S}-\boldsymbol{\Sigma}\right\|^2\right]=\frac{\operatorname{Tr}\left(\boldsymbol{\Phi}\right)}{n}. \tag{6.2}$$

Second

$$\mathbb{E}\left[\left\|\boldsymbol{\Sigma}_m-\hat{\boldsymbol{\Sigma}}_m\right\|^2\right]=\frac{1}{n^2}\mathbb{E}\left[\sum_{i,j=1}^{n}\left\langle\boldsymbol{\Pi}_m\left(\mathbf{x}_i\mathbf{x}_i^\top-\boldsymbol{\Sigma}\right)\boldsymbol{\Pi}_m,\boldsymbol{\Pi}_m\left(\mathbf{x}_j\mathbf{x}_j^\top-\boldsymbol{\Sigma}\right)\boldsymbol{\Pi}_m\right\rangle\right]$$

$$=\frac{1}{n}\mathbb{E}\left[\left\langle\boldsymbol{\Pi}_m\left(\mathbf{x}\mathbf{x}^\top-\boldsymbol{\Sigma}\right)\boldsymbol{\Pi}_m,\boldsymbol{\Pi}_m\left(\mathbf{x}\mathbf{x}^\top-\boldsymbol{\Sigma}\right)\boldsymbol{\Pi}_m\right\rangle\right]=\frac{1}{n}\mathbb{E}\left[\left\|\boldsymbol{\Pi}_m\left(\mathbf{x}\mathbf{x}^\top-\boldsymbol{\Sigma}\right)\boldsymbol{\Pi}_m\right\|^2\right],$$

and using the equalities (6.1) and the specificity of $\mathbf{\Pi}_m$, we obtain that

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{\Pi}_m\left(\mathbf{x}\mathbf{x}^\top-\mathbf{\Sigma}\right)\mathbf{\Pi}_m\right\|^2\right] &= \mathbb{E}\left[\left\|vec\left(\mathbf{\Pi}_m\left(\mathbf{x}\mathbf{x}^\top-\mathbf{\Sigma}\right)\mathbf{\Pi}_m\right)\right\|^2\right]\\ &= \mathbb{E}\left[\left\|\mathbf{\Pi}_m\otimes\mathbf{\Pi}_m\left(vec\left(\mathbf{x}\mathbf{x}^\top-\mathbf{\Sigma}\right)\right)\right\|^2\right]\\ &= \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{\Pi}_m\otimes\mathbf{\Pi}_m\left(vec\left(\mathbf{x}\mathbf{x}^\top-\mathbf{\Sigma}\right)\left(vec\left(\mathbf{x}\mathbf{x}^\top-\mathbf{\Sigma}\right)\right)^\top\left(\mathbf{\Pi}_m\otimes\mathbf{\Pi}_m\right)^\top\right)\right)\right]\\ &= \mathbb{E}\left[\mathrm{Tr}\left(\left(\mathbf{\Pi}_m\otimes\mathbf{\Pi}_m\right)\left(vec\left(\mathbf{x}\mathbf{x}^\top-\mathbf{\Sigma}\right)\right)\left(vec\left(\mathbf{x}\mathbf{x}^\top-\mathbf{\Sigma}\right)\right)^\top\right)\right].\end{aligned}$$

Hence

$$\mathbb{E}\left[\left\|\mathbf{\Pi}_m\left(\mathbf{x}\mathbf{x}^\top-\mathbf{\Sigma}\right)\mathbf{\Pi}_m\right\|^2\right] = \mathrm{Tr}\left(\left(\mathbf{\Pi}_m\otimes\mathbf{\Pi}_m\right)\mathbf{\Phi}\right),$$

and we obtain

$$\mathbb{E}\left[\left\|\mathbf{\Sigma}_m-\hat{\mathbf{\Sigma}}_m\right\|^2\right] = \frac{\mathrm{Tr}\left(\left(\mathbf{\Pi}_m\otimes\mathbf{\Pi}_m\right)\mathbf{\Phi}\right)}{n}\cdot \tag{6.3}$$

Finally, we have

$$\mathbb{E}\left[\left\|\mathbf{S}-\hat{\mathbf{\Sigma}}_m\right\|^2\right] = \left\|\mathbf{\Sigma}-\mathbf{\Sigma}_m\right\|^2 - \frac{\mathrm{Tr}\left(\left(\mathbf{\Pi}_m\otimes\mathbf{\Pi}_m\right)\mathbf{\Phi}\right)}{n} + \frac{\mathrm{Tr}\left(\mathbf{\Phi}\right)}{n}\cdot$$

□

Proof of Theorem 4.3.

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

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

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

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

then

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

We derive from the previous proposition and (6.2)

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

Moreover by the Cauchy–Schwarz inequality we have that

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

and using again this inequality

$$\begin{aligned}\mathbb{E}\left[\left\langle\hat{\mathbf{\Sigma}}-\mathbf{S},\mathbf{S}-\mathbf{\Sigma}\right\rangle\right] &\leqslant \sqrt{\mathbb{E}\left[\left\|\hat{\mathbf{\Sigma}}-\mathbf{S}\right\|^2\right]}\sqrt{\mathbb{E}\left[\left\|\mathbf{S}-\mathbf{\Sigma}\right\|^2\right]}\\ &\leqslant \sqrt{\mathbb{E}\left[\left\|\hat{\mathbf{\Sigma}}-\mathbf{S}\right\|^2 + 2\frac{\hat{\gamma}_{\hat{m}}^2}{n}\right]}\sqrt{\frac{\mathrm{Tr}\left(\mathbf{\Phi}\right)}{n}}.\end{aligned}$$

For the same reasons as before we obtain

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

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

Thus

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

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

$$2ab \leqslant \frac{a^2}{\varepsilon} + \varepsilon b^2$$

we obtain for all $A > 0$:

$$\mathbb{E}\left[\left\| \hat{\mathbf{\Sigma}} - \mathbf{\Sigma} \right\|^2\right] \leqslant \mathbb{E}\left[\left\| \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_{m_0} \right\|^2\right] \left(1 + \varepsilon^{-1}\right) + \frac{\operatorname{Tr}(\mathbf{\Phi})}{n} (4 + \varepsilon).$$

The definition of m_0 gives the result. □

References

[1] R.J. Adler, An introduction to continuity, extrema, and related topics for general gaussian processes. *Lect. Note Ser.* Institute of Mathematical Statistics (1990).

[2] P.J. Bickel and E. Levina, Covariance regularization by thresholding. *Ann. Statist.* **36** (2008) 2577–2604.

[3] J. Bigot, R. Biscay, J.-M. Loubes and L.M. Alvarez, Group lasso estimation of high-dimensional covariance matrices. *J. Machine Learn. Res.* (2011).

[4] J. Bigot, R. Biscay, J.-M. Loubes and L. Muñiz-Alvarez, Nonparametric estimation of covariance functions by model selection. *Electron. J. Statis.* **4** (2010) 822–855.

[5] J. Bigot, R. Biscay Lirio, J.-M. Loubes and L. Muniz Alvarez, Adaptive estimation of spectral densities via wavelet thresholding and information projection (2010).

[6] R. Biscay, L.M. Rodrguez and E. Daz-Frances, Cross-validation of covariance structures using the frobenius matrix distance as a discrepancy function. *J. Stat. Comput. Simul.* **58** (1997) 195–215.

[7] T. Cai and M. Yuan, *Nonparametric covariance function estimation for functional and longitudinal data.* Technical report (2010).

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

[9] P.J. Diggle and A.P. Verbyla, Nonparametric estimation of covariance structure in longitudinal data. *Biometrics* **54** (1998) 401–415.

[10] A.G. Journel, Kriging in terms of projections. *J. Int. Assoc. Math. Geol.* **9** (1977) 563–586.

[11] C.R. Rao, Linear statistical inference and its applications. *Wiley ser. Probab. Stastis.* Wiley, 2nd edn. (1973).

[12] G.A.F. Seber, A matrix handbook for statisticians. *Wiley ser. Probab. Stastis.* Wiley (2008).

[13] G.R. Shorack and J.A. Wellner, *Empirical processes with applications to statistics.* Wiley (1986).

[14] C.M. Stein, Estimation of the mean of a multivariate normal distribution. *Ann. Statis.* **9** (1981) 1135–1151.

[15] M.L. Stein. Interpolation of spatial data. Some theory for Kriging. *Springer Ser. Statis.* Springer-Verlag, New York (1999).

[16] A.B. Tsybakov, Introduction à l'estimation non-paramétrique. Vol. 41 of *Math. Appl.* Springer (2004).