Electronic Journal of Statistics
Vol. 4 (2010) 822–855
ISSN: 1935-7524
DOI: 10.1214/09-EJS493

Nonparametric estimation of covariance functions by model selection

Jérémie Bigot

Institut de Mathématiques de Toulouse
Université Toulouse 3, France
e-mail: bigot@math.univ-toulouse.fr

Rolando Biscay

CIMFAV-DEUV, Facultad de Ciencias,
Universidad de Valparaiso, Chile
e-mail: rolando.biscay@uv.cl

Jean-Michel Loubes*

Institut de Mathématiques de Toulouse
Université Toulouse 3, France
e-mail: loubes@math.univ-toulouse.fr

and

Lilian Muñiz-Alvarez

Facultad de Matemàtica y Computaciòn,
Universidad de la Habana Cuba
e-mail: llilian@matcom.uh.cu

Abstract: We propose a model selection approach for covariance estimation of a stochastic process. Under very general assumptions, observing i.i.d replications of the process at fixed observation points, we construct an estimator of the covariance function by expanding the process onto a collection of basis functions. We study the non asymptotic property of this estimate and give a tractable way of selecting the best estimator among a possible set of candidates. The optimality of the procedure is proved via an oracle inequality which warrants that the best model is selected.

AMS 2000 subject classifications: Primary 62G05, 62G20.
Keywords and phrases: Covariance estimation, model selection, oracle inequality.

Received September 2009.

Contents

*Corresponding author

1. Introduction

Estimating the covariance function of stochastic process is a fundamental issue with many applications, ranging from geostatistics, financial series or epidemiology for instance (we refer to [23], [13] or [8] for general references for applications). While parametric methods have been extensively studied in the statistical literature (see [8] for a review), nonparametric procedures have only recently received a growing attention. 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 use a model selection procedure to construct a nonparametric estimator of the covariance function of a stochastic process under general assumptions for the process. In particular we will not assume Gaussianity nor stationarity.

Consider a stochastic process $X(t)$ with values in $\mathbb{R}$, indexed by $t \in T$, a subset of $\mathbb{R}^d$, $d \in \mathbb{N}$. Throughout the paper, we assume that its covariance function is finite, i.e $\sigma(s,t) = cov(X(s), X(t)) < +\infty$ for all $s,t \in T$ and, for sake of simplicity, zero mean $\mathbb{E}(X(t)) = 0$ for all $t \in T$. The observations are $X_i(t_j)$ for $i = 1,\dots,N$, $j = 1,\dots,n$, where the observation points $t_1,...,t_n \in T$ are fixed, and $X_1,\dots,X_N$ are independent copies of the process X.

Functional approximations of the processes $X_1,\dots,X_N$ from data $(X_i(t_j))$ are involved in covariance function estimation. When dealing with functional data analysis (see, e.g., [20]), smoothing the processes $X_1,\dots,X_N$ is sometimes carried out as a first step before computing the empirical covariance such as spline interpolation for example (see for instance in [9]) or projection onto a general finite basis. Let $\mathbf{x}_i = (X_i(t_1),\dots,X_i(t_n))^\top$ be the vector of observations at the points $t_1,\dots,t_n$ with $i \in \{1,\dots,N\}$. Let $\{g_\lambda\}_{\lambda\in\mathcal{M}}$ be a collection of possibly independent functions $g_\lambda : T \to \mathbb{R}$ where $\mathcal{M}$ denote a generic countable set of indices. Then, let $m \subset \mathcal{M}$ be a subset of indices of size $|m| \in \mathbb{N}$ and define the $n \times |m|$ matrix $\mathbf{G}$ with entries $g_{j\lambda} = g_\lambda(t_j)$, $j = 1,\dots,n$, $\lambda \in m$. $\mathbf{G}$ will be called the design matrix corresponding to the set of basis functions indexed by m.

In such setting, usual covariance estimation is a two-step procedure: first, for each $i = 1, \dots, N$, fit the regression model

$$\mathbf{x}_i = \mathbf{G}\mathbf{a}_i + \epsilon_i \tag{1.1}$$

(by least squares or regularized least squares), where ϵ_i are random vectors in $\mathbb{R}^n$, to obtain estimates $\widehat{\mathbf{a}}_i = (\hat{a}_{i,\lambda})_{\lambda \in m} \in \mathbb{R}^{|m|}$ of $\mathbf{a}_i$ where in the case of standard least squares estimation (assuming for simplicity that $\mathbf{G}^\top \mathbf{G}$ is invertible)

$$\widehat{\mathbf{a}}_i = (\mathbf{G}^\top \mathbf{G})^{-1} \mathbf{G}^\top \mathbf{x}_i, i = 1, \dots, N.$$

Then, estimation of the covariance is obtained by computing the following estimate

$$\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top, \tag{1.2}$$

where

$$\widehat{\mathbf{\Psi}} = \frac{1}{N}\sum_{i=1}^{N} \widehat{\mathbf{a}}_i \widehat{\mathbf{a}}_i^\top = (\mathbf{G}^\top \mathbf{G})^{-1} \mathbf{G}^\top \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\top \right) \mathbf{G} (\mathbf{G}^\top \mathbf{G})^{-1}. \tag{1.3}$$

This corresponds to approximate the process X by a truncated process $\tilde{X}_i$ defined as

$$\widetilde{X}_i(t) = \sum_{\lambda \in m} \hat{a}_{i,\lambda} g_\lambda(t), i = 1, \dots, N,$$

and to choose the empirical covariance of $\tilde{X}$ as an estimator of the covariance of X, defined by

$$\widehat{\sigma}(s,t) = \frac{1}{N}\sum_{i=1}^{N} \widetilde{X_i}(s) \widetilde{X_i}(t).$$

In this paper, we consider the estimator (1.2) as the least squares estimator of the following matrix regression model

$$\mathbf{x}_i \mathbf{x}_i^\top = \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top + \mathbf{U}_i, \quad i = 1, \dots, N, \tag{1.4}$$

where $\mathbf{\Psi}$ is a symmetric matrix and $\mathbf{U}_i$ are i.i.d matrix errors. Fitting the models (1.1) and (1.4) by least squares naturally leads to the definition of different contrast and risk functions as the estimation is not performed in the same space ($\mathbb{R}^{|m|}$ for model (1.1) and $\mathbb{R}^{|m|\times|m|}$ for model (1.4)). By choosing an appropriate loss function, least squares estimation in model (1.4) also leads to the natural estimate (1.2) derived from least square estimation in model (1.1). A similar estimate can be found in [11]. However, in this paper, we tackle the problem of model selection, i.e. choosing an appropriate data-based subset of indices $m \in \mathcal{M}$, which is very distinct in model (1.1) and model (1.4). Indeed, model selection for (1.1) depends on the variability of the vectors $\mathbf{x}_i$'s while for (1.4) it depends on the variability of the matrices $\mathbf{x}_i \mathbf{x}_i^\top$'s. One of the main contributions of this paper is to show that considering model (1.4) enables to handle a large variety of cases and to build an optimal model selection estimator

of the covariance without too strong assumptions on the model. Moreover it will be shown that considering model (1.4) leads to the estimator $\widehat{\mathbf{\Psi}}$ defined in (1.3) which lies in the class of non-negative definite matrices and thus provides a proper covariance matrix $\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top$.

A similar method has been developed for smooth interpolation of covariance functions in [6], but restricted to basis functions that are determined by reproducing kernels in suitable Hilbert spaces and a different fitting criterion. Similar ideas are also tackled in [19]. These authors deal with the estimation of $\mathbf{\Sigma}$ within the covariance class $\mathbf{\Gamma} = \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top$ induced by an orthogonal wavelet expansion. However, their fitting criterion is not general since they choose the Gaussian likelihood as a contrast function, and thus their method requires specific distributional assumptions. We also point out that computation of the Gaussian likelihood requires inversion of $\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top$, which is not directly feasible if $rank\,(\mathbf{G}) < n$ or some diagonal entities of the non-negative definite (n.n.d) matrix $\mathbf{\Psi}$ are zero.

Hence, to our knowledge, no previous work has proposed to use the matrix regression model (1.4) under general moments assumptions of the process X using a general basis expansion for nonparametric covariance function estimation. We point out that the asymptotic behaviour will be taken with respect to the number of replications N while the observation points $t_i,\ i = 1, \ldots, n$ remain fixed.

The paper falls into the following parts. The description of the statistical framework of the matrix regression is given in Section 2. Section 3 is devoted to the main statistical results. Namely we study the behavior of the estimator for a fixed model in Section 3.1 while Section 3.2 deals with the model selection procedure and provide the oracle inequality. Section 4 states a concentration inequality that is used in all the paper, while some numerical experiments are described in Section 5. The proofs are postponed to a technical Appendix.

2. Nonparametric model selection for covariance estimation

Recall that $X = (X(t))_{t\in T}$ is an $\mathbb{R}$-valued stochastic process, where T denotes some subset of $\mathbb{R}^d$, $d \in \mathbb{N}$. Assume that X has finite moments up to order 4, and zero mean, i.e $\mathbb{E}(X(t)) = 0$ for all $t \in T$. The covariance function of X is denoted by $\sigma(s,t) = cov(X(s), X(t))$ for $s,t \in T$ and recall that $X_1, \ldots, X_N$ are independent copies of the process X.

In this work, we observe at different points $t_1, \ldots, t_n \in T$ independent copies of the process, denoted by $X_i(t_j)$, with $i = 1, \ldots, N$, $j = 1, \ldots, n$. Set $\mathbf{x}_i = (X_i(t_1), \ldots, X_i(t_n))^\top$ the vector of observations at the points $t_1, \ldots, t_n$ for each $i = 1, \ldots, N$. The matrix $\mathbf{\Sigma} = \mathbb{E}\left(\mathbf{x}_i\mathbf{x}_i^\top\right) = (\sigma(t_j,t_k))_{1\le j\le n, 1\le k\le n}$ is the covariance matrix of X at the observations points. Let $\overline{\mathbf{x}}$ and $\mathbf{S}$ denote the sample mean and the sample covariance (non corrected by the mean) of the data $\mathbf{x}_1, \ldots, \mathbf{x}_N$, i.e.

$$\overline{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i, \qquad \mathbf{S} = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^\top.$$

Our aim is to build a model selection estimator of the covariance of the process observed with N replications but without additional assumptions such as stationarity nor Gaussianity. The asymptotics will be taken with respect to N, the number of copies of the process.

2.1. Notations and preliminary definitions

First, define specific matrix notations. We refer to [18] or [14] for definitions and properties of matrix operations and special matrices. As usual, vectors in $\mathbb{R}^k$ are regarded as column vectors for all $k \in \mathbb{N}$. For any matrix $\mathbf{A}$, $\mathbf{A}^\top$ is the transpose of $\mathbf{A}$, $tr(\mathbf{A})$ is the trace of $\mathbf{A}$, $\|\mathbf{A}\|$ is the Frobenius matrix norm defined as $\|\mathbf{A}\|^2 = tr(\mathbf{A}\mathbf{A}^\top)$, $\lambda_{\max}(\mathbf{A})$ is the maximum eigenvalue of $\mathbf{A}$, and $\rho(\mathbf{A})$ is the spectral norm of $\mathbf{A}$, that is $\rho(\mathbf{A}) = \lambda_{\max}(\mathbf{A})$ for $\mathbf{A}$ a n.n.d matrix.

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 $\mathbf{A} = (a_{ij})_{1\le i\le k, 1\le j\le n}$ is the $kn \times 1$ column vector denoted by $vec(\mathbf{A})$, obtained by stacking the columns of the matrix $\mathbf{A}$ on top of one another. That is $vec(A) = [a_{11}, \ldots, a_{k1}, a_{12}, \ldots, a_{k2}, \ldots, a_{1n}, \ldots, a_{kn}]^\top$.

For a symmetric $k \times k$ matrix $\mathbf{A}$, the vector $vec(\mathbf{A})$ contains more information than necessary, since the matrix is completely determined by the lower triangular portion, that is, the $k(k+1)/2$ entries on and below the main diagonal. Hence, we introduce the symmetrized vectorization, which corresponds to a half-vectorization, denoted by $vech(\mathbf{A})$. More precisely, for any matrix $\mathbf{A} = (a_{ij})_{1\le i\le k, 1\le j\le k}$, define $vech(\mathbf{A})$ as the $k(k+1)/2 \times 1$ column vector obtained by vectorizing only the lower triangular part of $\mathbf{A}$. That is $vech(\mathbf{A}) = [a_{11}, \ldots, a_{k1}, a_{22}, \ldots, a_{n2}, \ldots, a_{(k-1)(k-1)}, a_{(k-1)k}, a_{kk}]^\top$. There exist a unique linear transformation which transforms the half-vectorization of a matrix to its vectorization and vice-versa called, respectively, the duplication matrix and the elimination matrix. For any $k \in \mathbb{N}$, the $k^2 \times k(k+1)/2$ duplication matrix is denoted by $\mathbf{D}_k$, $\mathbf{1}_k = (1, \ldots, 1)^\top \in \mathbb{R}^k$ and $\mathbf{I}_k$ is the identity matrix in $\mathbb{R}^{k\times k}$.

If $\mathbf{A} = (a_{ij})_{1\le i\le k, 1\le j\le n}$ is a $k \times n$ 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 $kp \times nq$ block matrix

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

For any random matrix $\mathbf{Z} = (Z_{ij})_{1\le i\le k, 1\le j\le n}$, its expectation is denoted by $\mathbb{E}(\mathbf{Z}) = (\mathbb{E}(Z_{ij}))_{1\le i\le k, 1\le j\le n}$. For any random vector $\mathbf{z} = (Z_i)_{1\le i\le k}$, let $V(\mathbf{z}) = (cov(Z_i, Z_j))_{1\le i,j\le k}$ be its covariance matrix. With this notation, $V(\mathbf{x}_1) = V(\mathbf{x}_i) = (\sigma(t_j, t_k))_{1\le j\le n, 1\le k\le n}$ is the covariance matrix of X.

Let $m \in \mathcal{M}$, and recall that to the finite set $\mathcal{G}_m = \{g_\lambda\}_{\lambda \in m}$ of functions $g_\lambda : T \to \mathbb{R}$ we associate the $n \times |m|$ matrix $\mathbf{G}$ with entries $g_{j\lambda} = g_\lambda(t_j), j = 1, \ldots, n$, $\lambda \in m$. 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

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

Let $\mathcal{S}_N(\mathbf{G})$ be the linear subspace of $\mathbb{R}^{nN \times n}$ defined by

$$\mathcal{S}_N(\mathbf{G}) = \left\{ \mathbf{1}_N \otimes \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top : \mathbf{\Psi} \in \mathcal{S}_m \right\} = \{ \mathbf{1}_N \otimes \mathbf{\Gamma} : \mathbf{\Gamma} \in \mathcal{S}(\mathbf{G}) \}$$

and let $\mathcal{V}_N(\mathbf{G})$ be the linear subspace of $\mathbb{R}^{n^2 N}$ defined by

$$\mathcal{V}_N(\mathbf{G}) = \left\{ \mathbf{1}_N \otimes vec\left(\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right) : \mathbf{\Psi} \in \mathcal{S}_m \right\} = \{ \mathbf{1}_N \otimes vec(\mathbf{\Gamma}) : \mathbf{\Gamma} \in \mathcal{S}(\mathbf{G}) \}.$$

All these spaces are regarded as Euclidean spaces with the scalar product associated to the Frobenius matrix norm.

2.2. Model selection approach for covariance estimation

The approach that we will develop to estimate the covariance function σ is based on the following two main ingredients: first, we consider a functional expansion $\tilde{X}$ to approximate the underlying process X and take the covariance of $\tilde{X}$ as an approximation of the true covariance σ.

For this, let $m \in \mathcal{M}$ and consider an approximation to the process X of the following form:

$$\widetilde{X}(t) = \sum_{\lambda \in m} a_\lambda g_\lambda(t), \tag{2.1}$$

where a_λ are suitable random coefficients. For instance if X takes its values in $L^2(T)$ (the space of square integrable real-valued functions on T) and if $(g_\lambda)_{\lambda \in \mathcal{M}}$ are orthonormal functions in $L^2(T)$, then one can take

$$a_\lambda = \int_T X(t) g_\lambda(t) dt.$$

Several basis can thus be considered, such as a polynomial basis on $\mathbb{R}^d$, Fourier expansion on a rectangle $T \subset \mathbb{R}^d$ (i.e. $g_\lambda(t) = e^{i2\pi\langle \omega_\lambda, t \rangle}$, using a regular grid of discrete set of frequencies $\{\omega_\lambda \in \mathbb{R}^d, \lambda \in m\}$ that do not depend on $t_1, \ldots, t_n$). One can also use, as in [9], tensorial product of B-splines on a rectangle $T \subset \mathbb{R}^d$, with a regular grid of nodes in $\mathbb{R}^d$ not depending on $t_1, \ldots, t_n$ or a standard wavelet basis on $\mathbb{R}^d$, depending on a regular grid of locations in $\mathbb{R}^d$ and discrete scales in $\mathbb{R}_+$. Another class of natural expansion is provided by Karhunen-Loeve expansion of the process X (see [1] for more references).

Therefore, it is natural to consider the covariance function ρ of $\widetilde{X}$ as an approximation of σ. Since the covariance ρ can be written as

$$\rho(s,t) = \mathbf{G}_s^\top \overline{\mathbf{\Psi}} \mathbf{G}_t, \tag{2.2}$$

where, after reindexing the functions if necessary, $\mathbf{G}_t = (g_\lambda(t), \lambda \in m)^\top$ and

$$\overline{\mathbf{\Psi}} = (\mathbb{E}(a_\lambda a_\mu)), \text{ with } (\lambda,\mu) \in m \times m.$$

Hence we are led to look for an estimate $\widehat{\sigma}$ of σ in the class of functions of the form (2.2), with $\mathbf{\Psi} \in \mathbb{R}^{|m|\times|m|}$ some symmetric matrix. Note that the choice of the function expansion in (2.1), in particular the choice of the subset of indices m, will be crucial in the approximation properties of the covariance function ρ. This estimation procedure has several advantages: it will be shown that an appropriate choice of loss function leads to the construction of symmetric n.n.d matrix $\widehat{\mathbf{\Psi}}$ (see Proposition 3.1) and thus the resulting estimate

$$\widehat{\sigma}(s,t) = \mathbf{G}_s^\top \widehat{\mathbf{\Psi}} \mathbf{G}_t,$$

is a covariance function, so the resulting estimator can be plugged in other procedures which requires working with a covariance function. We also point out that the large amount of existing approaches for function approximation of the type (2.1) (such as those based on Fourier, wavelets, kernel, splines or radial functions) provides great flexibility to the model (2.2).

Secondly, we use the Frobenius matrix norm to quantify the risk of the covariance matrix estimators. Recall that $\mathbf{\Sigma} = (\sigma(t_j,t_k))_{1\le j,k\le n}$ is the true covariance matrix while $\mathbf{\Gamma} = (\rho(t_j,t_k))_{1\le j,k\le n}$ will denote the covariance matrix of the approximated process $\tilde{X}$ at the observation points. Hence

$$\mathbf{\Gamma} = \mathbf{G}\overline{\mathbf{\Psi}}\mathbf{G}^\top. \tag{2.3}$$

Comparing the covariance function ρ with the true one σ over the design points t_j, implies quantifying the deviation of $\mathbf{\Gamma}$ from $\mathbf{\Sigma}$. For this consider the following loss function

$$L(\mathbf{\Psi}) = \mathbb{E}\left\|\mathbf{x}\mathbf{x}^\top - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|^2,$$

where $\mathbf{x} = (X(t_1),\ldots,X(t_n))^\top$ and $\|.\|$ is the Frobenius matrix norm. Note that

$$L(\mathbf{\Psi}) = \left\|\mathbf{\Sigma} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|^2 + C,$$

where the constant C does not depend on $\mathbf{\Psi}$. The Frobenius matrix norm provides a meaningful metric for comparing covariance matrices, widely used in multivariate analysis, in particular in the theory on principal components analysis. See also [5], [22] and references therein for other applications of this loss function.

To the loss L corresponds the following empirical contrast function L_N, which will be the fitting criterion we will try to minimize

$$L_N(\mathbf{\Psi}) = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|^2.$$

We point out that this loss is exactly the sum of the squares of the residuals corresponding to the matrix linear regression model

$$\mathbf{x}_i\mathbf{x}_i^\top = \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top + \mathbf{U}_i, \quad i = 1, \ldots, N, \tag{2.4}$$

with i.i.d. matrix errors $\mathbf{U}_i$ such that $\mathbb{E}(\mathbf{U}_i) = \mathbf{0}$. This remark provides a natural framework to study the covariance estimation problem as a matrix regression model. Note also that the set of matrices $\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top$ is a linear subspace of $\mathbb{R}^{n\times n}$ when $\mathbf{\Psi}$ ranges over the space of symmetric matrices $\mathcal{S}_m$.

To summarize our approach, we finally propose following two-step estimation procedure: in a first step, for a given design matrix $\mathbf{G}$, define

$$\widehat{\mathbf{\Psi}} = \arg\min_{\mathbf{\Psi}\in\mathcal{S}_m} L_N(\mathbf{\Psi}),$$

and take $\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top$ as an estimator of $\mathbf{\Sigma}$. Note that $\widehat{\mathbf{\Psi}}$ will be shown to be a n.n.d matrix (see Proposition 3.1) and thus $\widehat{\mathbf{\Sigma}}$ is also a n.n.d matrix. Since the minimization of $L_N(\mathbf{\Psi})$ with respect to $\mathbf{\Psi}$ is done over the linear space of symmetric matrices $\mathcal{S}_m$, it can be transformed to a classical least squares linear problem, and the computation of $\widehat{\mathbf{\Psi}}$ is therefore quite simple. For a given design matrix $\mathbf{G}$, we will construct an estimator for $\mathbf{\Gamma} = \mathbf{G}\overline{\mathbf{\Psi}}\mathbf{G}^\top$ which will be close to $\mathbf{\Sigma} = V(\mathbf{x}_1)$ as soon as $\tilde{X}$ is a sharp approximation of X. So, the role of $\mathbf{G}$ and thus the choice of the subset of indices m is crucial since it determines the behavior of the estimator.

Hence, in second step, we aim at selecting the best design matrix $\mathbf{G} = \mathbf{G}_m$ among a collection of candidates $\{\mathbf{G}_m, m \in \mathcal{M}\}$. For this, methods and results from the theory of model selection in linear regression can be applied to the present context. In particular the results in [2], [7] or [16, 17] will be useful in dealing with model selection for the framework (2.4). Note that only assumptions about moments, not specific distributions of the data, are involved in the estimation procedure.

Remark 2.1. *We consider here a least-squares estimates of the covariance. Note that suitable regularization terms or constraints could also be incorporated into the minimization of $L_N(\mathbf{\Psi})$ in order to impose desired properties for the resulting estimator, such as smoothness or sparsity conditions as in [15].*

3. Oracle inequality for covariance estimation

The first part of this section describes the properties of the least squares estimator $\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top$ while the second part builds a selection procedure to pick automatically the best estimate among a collection of candidates.

3.1. Least squares covariance estimation

Given some $n \times |m|$ fixed design matrix $\mathbf{G}$ associated to a finite family of $|m|$ basis functions, the least squares covariance estimator of $\mathbf{\Sigma}$ is defined by

$$\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top = \arg\min\left\{\frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{\Gamma}\right\|^2 : \mathbf{\Gamma} = \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top, \mathbf{\Psi} \in \mathcal{S}_m\right\}. \quad (3.1)$$

The corresponding estimator of the covariance function σ is

$$\widehat{\sigma}(s,t) = \mathbf{G}_s^\top \widehat{\mathbf{\Psi}}\mathbf{G}_t. \quad (3.2)$$

Proposition 3.1. *Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_N \in \mathbb{R}^{n\times n}$ and $\mathbf{G} \in \mathbb{R}^{n\times|m|}$ be arbitrary matrices Then, (a) The infimum*

$$\inf\left\{\frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{Y}_i - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|^2 : \mathbf{\Psi} \in \mathcal{S}_m\right\}$$

is achieved at

$$\widehat{\mathbf{\Psi}} = \left(\mathbf{G}^\top\mathbf{G}\right)^- \mathbf{G}^\top\left(\frac{\overline{\mathbf{Y}} + \overline{\mathbf{Y}}^\top}{2}\right)\mathbf{G}\left(\mathbf{G}^\top\mathbf{G}\right)^-, \quad (3.3)$$

where $\left(\mathbf{G}^\top\mathbf{G}\right)^-$ is any generalized inverse of $\mathbf{G}^\top\mathbf{G}$ (see [10] for a general definition), and

$$\overline{\mathbf{Y}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{Y}_i.$$

(b) Furthermore, $\mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top$ is the same for all the generalized inverses $\left(\mathbf{G}^\top\mathbf{G}\right)^-$ of $\mathbf{G}^\top\mathbf{G}$. In particular, if $\mathbf{Y}_1, \ldots, \mathbf{Y}_N \in \mathcal{S}_n$ (i.e., if they are symmetric matrices) then any minimizer has the form

$$\widehat{\mathbf{\Psi}} = \left(\mathbf{G}^\top\mathbf{G}\right)^- \mathbf{G}^\top\overline{\mathbf{Y}}\mathbf{G}\left(\mathbf{G}^\top\mathbf{G}\right)^-.$$

If $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ are n.n.d then these matrices $\widehat{\mathbf{\Psi}}$ are n.n.d.

If we assume that $(\mathbf{G}^\top\mathbf{G})^{-1}$ exists, then Proposition 3.1 shows that we retrieve the expression (1.3) for $\widehat{\mathbf{\Psi}}$ that has been derived from least square estimation in model (1.1).

Theorem 3.2. *Let $\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^\top$. Then, the least squares covariance estimate defined by (3.1) is given by the n.n.d matrix*

$$\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top = \Pi\mathbf{S}\Pi,$$

where

$$\widehat{\mathbf{\Psi}} = \left(\mathbf{G}^\top\mathbf{G}\right)^- \mathbf{G}^\top\mathbf{S}\mathbf{G}\left(\mathbf{G}^\top\mathbf{G}\right)^-, \quad (3.4)$$
$$\Pi = \mathbf{G}\left(\mathbf{G}^\top\mathbf{G}\right)^-\mathbf{G}^\top.$$

Moreover $\widehat{\mathbf{\Sigma}}$ has the following interpretations in terms of orthogonal projections:

i) $\widehat{\mathbf{\Sigma}}$ *is the projection of* $\mathbf{S} \in \mathbb{R}^{n\times n}$ *on* $\mathcal{S}(\mathbf{G})$.

ii) $\mathbf{1}_N \otimes \widehat{\mathbf{\Sigma}}$ *is the projection of* $\mathbf{Y} = \left(\mathbf{x}_1\mathbf{x}_1^\top, \ldots, \mathbf{x}_N\mathbf{x}_N^\top\right)^\top \in \mathbb{R}^{nN\times n}$ *on* $\mathcal{S}_N(\mathbf{G})$.

iii) $\mathbf{1}_N \otimes vec\left(\widehat{\mathbf{\Sigma}}\right)$ *is the projection of* $\mathbf{y} = \left(vec^\top\left(\mathbf{x}_1\mathbf{x}_1^\top\right), \ldots, vec^\top\left(\mathbf{x}_N\mathbf{x}_N^\top\right)\right)^\top \in \mathbb{R}^{n^2N}$ *on* $\mathcal{V}_N(\mathbf{G})$.

The proof of this theorem is a direct application of Proposition 3.1. Hence for a given design matrix $\mathbf{G}$, the least squares estimator $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}(\mathbf{G})$ is well defined and has the structure of a covariance matrix. It remains to study how to pick automatically the estimate when dealing with a collection of design matrices coming from several approximation choices for the random process X.

3.2. Main result

Consider a collection of indices $m \in \mathcal{M}$ with size $|m|$. Let also $\{\mathbf{G}_m : m \in \mathcal{M}\}$ be a finite family of design matrices $\mathbf{G}_m \in \mathbb{R}^{n\times|m|}$, and let $\widehat{\mathbf{\Sigma}}_m = \hat{\mathbf{\Sigma}}(\mathbf{G}_m)$, $m \in \mathcal{M}$, be the corresponding least squares covariance estimators. The problem of interest is to select the best of these estimators in the sense of the minimal quadratic risk $\mathbb{E}\left\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_m\right\|^2$.

The main theorem of this section provides a non-asymptotic bound for the risk of a penalized strategy for this problem. For all $m \in \mathcal{M}$, write

$$\begin{aligned} \mathbf{\Pi}_m &= \mathbf{G}_m\left(\mathbf{G}_m^\top\mathbf{G}_m\right)^- \mathbf{G}_m^\top, \\ D_m &= Tr\left(\mathbf{\Pi}_m\right), \end{aligned} \tag{3.5}$$

We assume that $D_m \geq 1$ for all $m \in \mathcal{M}$. The estimation error for a given model $m \in \mathcal{M}$ is given by

$$\mathbb{E}\left(\left\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_m\right\|^2\right) = \left\|\mathbf{\Sigma} - \mathbf{\Pi}_m\mathbf{\Sigma}\mathbf{\Pi}_m\right\|^2 + \frac{\delta_m^2 D_m}{N}, \tag{3.6}$$

where

$$\begin{aligned} \delta_m^2 &= \frac{\mathrm{Tr}\left(\left(\mathbf{\Pi}_m \otimes \mathbf{\Pi}_m\right)\mathbf{\Phi}\right)}{D_m}, \\ \mathbf{\Phi} &= V\left(vec\left(\mathbf{x}_1\mathbf{x}_1^\top\right)\right). \end{aligned}$$

Given $\theta > 0$, define the penalized covariance estimator $\widetilde{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}_{\widehat{m}}$ by

$$\widehat{m} = \arg\min_{m\in\mathcal{M}}\left\{\frac{1}{N}\sum_{i=1}^N\left\|\mathbf{x}_i\mathbf{x}_i^\top - \widehat{\mathbf{\Sigma}}_m\right\|^2 + pen(m)\right\},$$

where

$$pen(m) = (1+\theta)\frac{\delta_m^2 D_m}{N}. \tag{3.7}$$

Theorem 3.3. *Let $q > 0$ be given such that there exists $p > 2(1+q)$ satisfying $\mathbb{E}\left\|\mathbf{x}_1\mathbf{x}_1^\top\right\|^p < \infty$. Then, for some constants $K(\theta) > 1$ and $C'(\theta, p, q) > 0$ we have that*

$$\left(\mathbb{E}\left\|\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}}\right\|^{2q}\right)^{1/q} \leq 2^{\left(q^{-1}-1\right)_+}\left[K(\theta)\inf_{m\in\mathcal{M}}\left(\left\|\boldsymbol{\Sigma} - \boldsymbol{\Pi}_m\boldsymbol{\Sigma}\boldsymbol{\Pi}_m\right\|^2 + \frac{\delta_m^2 D_m}{N}\right) + \frac{\Delta_p}{N}\delta_{\sup}^2\right],$$

where

$$\Delta_p^q = C'(\theta, p, q)\,\mathbb{E}\left\|\mathbf{x}_1\mathbf{x}_1^\top\right\|^p\left(\sum_{m\in\mathcal{M}}\delta_m^{-p}D_m^{-(p/2-1-q)}\right)$$

and

$$\delta_{\sup}^2 = \max\left\{\delta_m^2 : m\in\mathcal{M}\right\}.$$

In particular, for $q = 1$ we have

$$\mathbb{E}\left(\left\|\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}}\right\|^2\right) \leq K(\theta)\inf_{m\in\mathcal{M}}\mathbb{E}\left(\left\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_m\right\|^2\right) + \frac{\Delta_p}{N}\delta_{\sup}^2. \tag{3.8}$$

For the proof of this result, we first restate this theorem in a a vectorized form which turns to be a k-variate extensions of results in [2] (which are covered when $k = 1$) and are stated in Section 4.1. Their proof rely on model selection techniques and a concentration tool stated in Section 4.2.

Remark 3.4. *Note that the penalty depends on the quantity δ_m which is unknown in practice. Indeed, the penalty relies on $\boldsymbol{\Phi}$=$V\left(vec\left(\mathbf{x}_1\mathbf{x}_1^\top\right)\right)$, which reflects the correlation structure of the data. In the original paper by Baraud [3], an estimator of the variance is proposed to overcome this issue. However, the consistency proof relies on a concentration inequality which turns to be a χ^2 like inequality. Extending this inequality to our case would mean to be able to construct concentration bounds for matrices $\mathbf{x}\mathbf{x}^\top$, implying Wishart distributions. Some results exist in this framework [21], but adapting this kind of construction to our case is a hard task which falls beyond the scope of this paper.*

However, we point out that for practical purpose, when N is large enough, this quantity can be consistently estimated using the empirical version of $\boldsymbol{\Phi}$ since the $\mathbf{x}_i, i = 1, \ldots, N$ are i.i.d observed random variables, which is given by

$$\widehat{\boldsymbol{\Phi}} = \frac{1}{N}\sum_{i=1}^N vec\left(\mathbf{x}_i\mathbf{x}_i^\top\right)\left(vec\left(\mathbf{x}_i\mathbf{x}_i^\top\right)\right)^\top - vec(\mathbf{S})\left(vec(\mathbf{S})\right)^\top. \tag{3.9}$$

Hence, there is a practical way of computing the penalty. The influence of the use of such an estimated penalty is studied in Section 5. Note also that if $\tau > 0$ denotes any bound of δ_m^2 such that $\delta_m^2 \leq \tau$ for all m, then Theorem 3.3 remains true with δ_m^2 replaced by τ in all the statements.

We have obtained in Theorem 3.3 an oracle inequality since, using (3.6) and (3.8), one immediately sees that $\widetilde{\mathbf{\Sigma}}$ has the same quadratic risk as the "oracle" estimator except for an additive term of order $O\left(\frac{1}{N}\right)$ and a constant factor. Hence, the selection procedure is optimal in the sense that it behaves as if the true model were at hand. To describe the result in terms of rate of convergence, we have to pay a special attention to the bias terms $\|\mathbf{\Sigma} - \mathbf{\Pi}_m \mathbf{\Sigma} \mathbf{\Pi}_m\|^2$. In a very general framework, it is difficult to evaluate such approximation terms. If the process has bounded second moments, i.e for all $j = 1, \ldots, n$, we have $\mathbb{E}\left(X^2(t_j)\right) \leq C$, then we can write

$$\|\mathbf{\Sigma} - \mathbf{\Pi}_m \mathbf{\Sigma} \mathbf{\Pi}_m\|^2 \leq C \sum_{j=1}^{n} \sum_{j'=1}^{n} \left[\mathbb{E}\left(X(t_j) - \widetilde{X}(t_j) \right)^2 + \mathbb{E}\left(X(t_{j'}) - \widetilde{X}(t_{j'}) \right)^2 \right]$$
$$\leq 2Cn^2 \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}\left(X(t_j) - \widetilde{X}(t_j) \right)^2.$$

Since n is fixed and the asymptotics are given with respect to N, the number of replications of the process, the rate of convergence relies on the quadratic error of the expansion of the process.

To compute the rate of convergence, this approximation error must be controlled. From a theoretical point of view, take $d = 1$, $T = [a, b]$, and consider a process $X(t)$ with $t \in [a, b]$, for which the basis of its Karhunen-Loève expansion is known. Set $\mathcal{M} = \mathcal{M}_N = \{m = \{1, \ldots, |m|\}, |m| = 1, \ldots, N\}$. Then we can write $X(t) = \sum_{\lambda=1}^{\infty} Z_\lambda g_\lambda(t)$, where Z_λ are centered random variables with $\mathbb{E}\left(Z_\lambda^2\right) = \gamma_\lambda^2$, where γ_λ^2 is the eigenvalue corresponding to the eigenfunction g_λ of the operator $(Kf)(t) = \int_a^b \sigma(s,t) f(s)\, ds$. If $X(t)$ is a Gaussian process then the random variables Z_λ are Gaussian and stochastically independent. Hence, a natural approximation of $X(t)$ is given by $\widetilde{X}(t) = \sum_{\lambda=1}^{|m|} Z_\lambda g_\lambda(t)$. So we have that

$$\mathbb{E}\left(X(t) - \widetilde{X}(t) \right)^2 = \mathbb{E}\left(\sum_{\lambda=|m|+1}^{\infty} Z_\lambda g_\lambda(t) \right)^2 = \sum_{\lambda=|m|+1}^{\infty} \gamma_\lambda^2 g_\lambda^2(t).$$

therefore, if $\|g_\lambda\|^2_{L_2([a,b])} = 1$ then $\mathbb{E}\big\|X(t) - \widetilde{X}(t)\big\|^2_{L_2([a,b])} = \sum_{\lambda=|m|+1}^{\infty} \gamma_\lambda^2$. Assume that the γ_λ's have a polynomial decay of rate $\alpha > 0$, namely $\gamma_\lambda \sim \lambda^{-\alpha}$, then we get an approximation error of order $O\big((|m|+1)^{-2\alpha}\big)$. Hence, we get that (under appropriate conditions on the design points $t_1, \ldots, t_n$)

$$\|\mathbf{\Sigma} - \mathbf{\Pi}_m \mathbf{\Sigma} \mathbf{\Pi}_m\|^2 = O\left((|m|+1)^{-2\alpha} \right).$$

Finally, since in this example $\mathbb{E}\|\mathbf{\Sigma} - \widetilde{\mathbf{\Sigma}}\|^2 \leq K(\theta) \inf_{m \in \mathcal{M}_N} \big(\|\mathbf{\Sigma} - \mathbf{\Pi}_m \mathbf{\Sigma} \mathbf{\Pi}_m\|^2 + \frac{\delta_m^2 m}{N} \big) + O\big(\frac{1}{N}\big)$ then the quadratic risk is of order $N^{-\frac{2\alpha}{2\alpha+1}}$ as soon as $|m| \sim N^{1/(2\alpha+1)}$ belongs to the collection of models $\mathcal{M}_N$. In another framework, if we consider a spline expansion, the rate of convergence for the approximation given in [9] are of the same order.

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 n. Moreover, since for a practical use of this methodology, an estimator of the penalty must be computed, relying on the estimation of the 4-th order moment, the need for a large amount of data is crucial even if the simulations are still, quite satisfactory, for not so large sample. This settings is quite common in epidemiology where a phenomenon is studied at a large number of locations but only during a short time. Hence our method is not designed to tackle the problem of covariance estimation in the high dimensional case $n >> N$. This topic has received a growing attention over the past years and we refer to [4] and references therein for a survey.

4. Model selection for multidimensional regression

4.1. Oracle inequality for multidimensional regression model

Recall that we consider the following model

$$\mathbf{x}_i\mathbf{x}_i^\top = \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top + \mathbf{U}_i, \quad i = 1, \dots, N,$$

with i.i.d. matrix errors $\mathbf{U}_i \in \mathbb{R}^{n\times n}$ such that $\mathbb{E}(\mathbf{U}_i) = \mathbf{0}$.

The key point is that previous model can be rewritten in vectorized form in the following way

$$\mathbf{y}_i = \mathbf{A}\beta + \mathbf{u}_i, \quad i = 1, \dots, N, \tag{4.1}$$

where $\mathbf{y}_i = vec\left(\mathbf{x}_i\mathbf{x}_i^\top\right) \in \mathbb{R}^{n^2}$, $\mathbf{A} = (\mathbf{G}\otimes\mathbf{G})\,\mathbf{D}_m \in \mathbb{R}^{n^2\times\frac{|m|(|m|+1)}{2}}$, where $\mathbf{D}_m \in \mathbb{R}^{|m|^2\times\frac{|m|(|m|+1)}{2}}$ is the duplication matrix, $\beta = vech(\mathbf{\Psi}) \in \mathbb{R}^{\frac{|m|(|m|+1)}{2}}$, and $\mathbf{u}_i = vec(\mathbf{U}_i) \in \mathbb{R}^{n^2}$.

Note that this model is equivalent to the following regression model

$$\mathbf{y} = (\mathbf{1}_N \otimes \mathbf{A})\,\beta + \mathbf{u}, \tag{4.2}$$

where $\mathbf{y} = \left((\mathbf{y}_1)^\top, \dots, (\mathbf{y}_N)^\top\right)^\top \in \mathbb{R}^{Nn^2}$ is the data vector, $(\mathbf{1}_N \otimes \mathbf{A}) \in \mathbb{R}^{Nn^2\times\frac{|m|(|m|+1)}{2}}$ is a known fixed matrix, $\beta = vech(\mathbf{\Psi})$ is an unknown vector parameter as before, and $\mathbf{u} = \left((\mathbf{u}_1)^\top, \dots, (\mathbf{u}_N)^\top\right)^\top \in \mathbb{R}^{Nn^2}$ is such that $\mathbb{E}(\mathbf{u}) = \mathbf{0}$. It is worth of noting that this regression model has several peculiarities in comparison with standard ones.

i) The error $\mathbf{u}$ has a specific correlation structure, namely $\mathbf{I}_N \otimes \mathbf{\Phi}$, where $\mathbf{\Phi} = V\left(vec\left(\mathbf{x}_1\mathbf{x}_1^\top\right)\right)$.

ii) In contrast with standard multivariate models, each coordinate of $\mathbf{y}$ depends on all the coordinates of β.

iii) For any estimator $\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top$ that be a linear function of the sample covariance $\mathbf{S}$ of the data $\mathbf{x}_1,\ldots,\mathbf{x}_N$ (and so, in particular, for the estimator minimizing L_N) it is possible to construct an unbiased estimator of its quadratic risk $\mathbb{E}\left\|\mathbf{\Sigma}-\widehat{\mathbf{\Sigma}}\right\|^2$.

More generally, assume we observe $\mathbf{y}_i$, $i = 1,\ldots,N$ random vectors of $\mathbb{R}^k$, with $k \geq 1$ ($k = n^2$ in the particular case of model (4.1)), such that

$$\mathbf{y}_i = \mathbf{f}^i + \varepsilon_i, \quad i = 1,\ldots,N, \tag{4.3}$$

where $\mathbf{f}^i \in \mathbb{R}^k$ are nonrandom and $\varepsilon_1,\ldots,\varepsilon_N$ are i.i.d. random vectors in $\mathbb{R}^k$ with $\mathbb{E}(\varepsilon_1) = \mathbf{0}$ and $V(\varepsilon_1) = \mathbf{\Phi}$. For sake of simplicity, we identify the function $g : \mathcal{X} \to \mathbb{R}^k$ with vectors $(g(x_1),\ldots,g(x_N))^\top \in \mathbb{R}^{Nk}$ and we denote by $\langle a,b\rangle_N = \frac{1}{N}\sum_{i=1}^N a_i^\top b_i$ the inner product of $\mathbb{R}^{Nk}$ associated to the norm $\|.\|_N$, where $a = (a_1 \ldots a_N)^\top$ and $b = (b_1 \ldots b_N)^\top$ with $a_i, b_i \in \mathbb{R}^k$ for all $i = 1,\ldots,N$.

Given $N, k \in \mathbb{N}$, let $(\mathcal{L}_m)_{m\in\mathcal{M}}$ be a finite family of linear subspaces of $\mathbb{R}^{Nk}$. For each $m \in \mathcal{M}$, assume $\mathcal{L}_m$ has dimension $D_m \geq 1$. Let $\widehat{\mathbf{f}}_m$ be the least squares estimator of $\mathbf{f} = \left((\mathbf{f}^1)^\top,\ldots,(\mathbf{f}^N)^\top\right)^\top$ based on the data $\mathbf{y} = \left(\mathbf{y}_1^\top,\ldots,\mathbf{y}_N^\top\right)^\top$ under the model $\mathcal{L}_m$, i.e.

$$\widehat{\mathbf{f}}_m = \arg\min_{\mathbf{v}\in\mathcal{L}_m}\left\{\|\mathbf{y}-\mathbf{v}\|_N^2\right\} = \mathbf{P}_m\mathbf{y},$$

where $\mathbf{P}_m$ is the orthogonal projection matrix from $\mathbb{R}^{Nk}$ on $\mathcal{L}_m$. Write

$$\delta_m^2 = \frac{\mathrm{Tr}\left(\mathbf{P}_m\left(\mathbf{I}_N \otimes \mathbf{\Phi}\right)\right)}{D_m},$$
$$\delta_{\sup}^2 = \max\left\{\delta_m^2 : m \in \mathcal{M}\right\}.$$

Given $\theta > 0$, define the penalized estimator $\widetilde{\mathbf{f}} = \widehat{\mathbf{f}}_{\widehat{m}}$, where

$$\widehat{m} = \arg\min_{m\in\mathcal{M}}\left\{\left\|\mathbf{y}-\widehat{\mathbf{f}}_m\right\|_N^2 + pen(m)\right\},$$

with

$$pen(m) = (1+\theta)\frac{\delta_m^2 D_m}{N}.$$

Proposition 4.1. *Let $q > 0$ be given such that there exists $p > 2(1+q)$ satisfying $\mathbb{E}\|\varepsilon_1\|^p < \infty$. Then, for some constants $K(\theta) > 1$ and $c(\theta,p,q) > 0$ we have that*

$$\mathbb{E}\left(\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 - K(\theta)M_N^*\right)_+^q \leq \Delta_p^q\frac{\delta_{\sup}^{2q}}{N^q}, \tag{4.4}$$

where

$$\Delta_p^q = C(\theta,p,q)\,\mathbb{E}\|\varepsilon_1\|^p\left(\sum_{m\in\mathcal{M}}\delta_m^{-p}D_m^{-(p/2-1-q)}\right),$$
$$M_N^* = \inf_{m\in\mathcal{M}}\left\{\|\mathbf{f}-\mathbf{P}_m\mathbf{f}\|_N^2 + \frac{\delta_m^2 D_m}{N}\right\}.$$

This theorem is equivalent to Theorem 3.3 using the vectorized version of the model (4.3) and turns to be an extension of Theorem 3.1 in [2] to the multivariate case. In a similar way, the following result constitutes also a natural extension of Corollary 3.1 in [2]. It is also closely related to the recent work in [12].

Corollary 4.2. *Under the assumptions of Proposition 4.1 it holds that*

$$\left(\mathbb{E}\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^{2q}\right)^{1/q} \leq 2^{\left(q^{-1}-1\right)_+}\left[K(\theta)\inf_{m\in\mathcal{M}}\left(\|\mathbf{f}-\mathbf{P}_m\mathbf{f}\|_N^2+\frac{\delta_m^2 D_m}{N}\right)+\frac{\Delta_p}{N}\delta_{\sup}^2\right],$$

where Δ_p was defined in Proposition 4.1.

Under regularity assumptions for the function $\mathbf{f}$, depending on a smoothness parameter s, the bias term is of order

$$\|\mathbf{f}-\mathbf{P}_m\mathbf{f}\|_N^2 = O(D_m^{-2s}).$$

Hence, for $q=1$ we obtain the usual rate of convergence $N^{-\frac{2s}{2s+1}}$ for the quadratic risk as soon as the optimal choice $D_m = N^{\frac{1}{2s+1}}$ belongs to the collection of models, yielding the optimal rate of convergence for the penalized estimator.

4.2. Concentration bound for random processes

Recall that $k \geq 1$. The following result is a k-variate extension of results in [2] (which are covered when $k=1$). Its proof is deferred to the Appendix.

Proposition 4.3. *(Extension of Corollary 5.1 in [2]). Given $N, k \in \mathbb{N}$, let $\widetilde{\mathbf{A}} \in \mathbb{R}^{Nk\times Nk}\setminus\{\mathbf{0}\}$ be a non-negative definite and symmetric matrix and $\varepsilon_1,\ldots,\varepsilon_N$ i.i.d random vectors in $\mathbb{R}^k$ with $\mathbb{E}(\varepsilon_1)=0$ and $V(\varepsilon_1)=\mathbf{\Phi}$. Write $\varepsilon=\left(\varepsilon_1^\top,\ldots,\varepsilon_N^\top\right)^\top$, $\zeta(\varepsilon)=\sqrt{\varepsilon^\top\widetilde{A}\varepsilon}$, and $\delta^2=\frac{\mathrm{Tr}\left(\widetilde{\mathbf{A}}(\mathbf{I}_N\otimes\mathbf{\Phi})\right)}{\mathrm{Tr}\left(\widetilde{\mathbf{A}}\right)}$. For all $p\geq 2$ such that $\mathbb{E}\|\varepsilon_1\|^p<\infty$ it holds that, for all $x>0$,*

$$\mathbb{P}\left(\zeta^2(\varepsilon)\geq\delta^2\mathrm{Tr}\left(\widetilde{\mathbf{A}}\right)+2\delta^2\sqrt{\mathrm{Tr}\left(\widetilde{\mathbf{A}}\right)\rho\left(\widetilde{\mathbf{A}}\right)x}+\delta^2\rho\left(\widetilde{\mathbf{A}}\right)x\right)$$
$$\leq C(p)\frac{\mathbb{E}\|\varepsilon_1\|^p\,\mathrm{Tr}\left(\widetilde{\mathbf{A}}\right)}{\delta^p\rho\left(\widetilde{\mathbf{A}}\right)x^{p/2}}, \tag{4.5}$$

where the constant $C(p)$ depends only on p.

Proposition 4.3 reduces to Corollary 5.1 in [2] when we only consider $k=1$, in which case $\delta^2=(\mathbf{\Phi})_{11}=\sigma^2$ is the variance of the univariate i.i.d. errors ε_i.

5. Numerical examples

In this section we illustrate the practical behaviour of the covariance estimator by model selection proposed in this paper. In particular, we study its performance when computing the criterion using the estimated penalty described in

Section 3.2. The programs for our simulations were implemented using MATLAB and the code is available on request.

We will consider i.i.d copies $X_1, \ldots, X_n$ of different Gaussian processes X on $T = [0,1]$ with values in $\mathbb{R}$, observed at fixed equi-spaced points $t_1, \ldots, t_n$ in $[0,1]$ for a fixed n, generated according to

$$X(t_j) = \sum_{\lambda=1}^{m^*} a_\lambda g_\lambda^*(t_j), \ j = 1, \ldots, n, \tag{5.1}$$

where m^* denotes the true model dimension, $(g_\lambda^*)_\lambda$ $(\lambda = 1, \ldots, m^*)$ are orthonormal functions on $[0,1]$, and the coefficients $a_1, \ldots, a_{m^*}$ are independent and identically distributed Gaussian variables with zero mean. Note that $\mathbb{E}(X(t_j)) = 0$ for all $j = 1, \ldots, n$, and that the covariance function of the process X at the points $t_1, \ldots, t_n$ is given by

$$\sigma(t_j, t_k) = cov(X(t_j), X(t_k)) = \sum_{\lambda=1}^{m^*} V(a_\lambda) g_\lambda^*(t_j) g_\lambda^*(t_k)$$

for all $1 \leq j, k \leq n$. The corresponding covariance matrix is $\Sigma = (\sigma(t_j, t_k))_{1 \leq j,k \leq n}$. We will write $\mathbf{X} = (X_i(t_j))$ an $n \times N$ matrix. The columns of $\mathbf{X}$ are denoted by $\mathbf{x}_i = (X_i(t_1), \ldots, X_i(t_n))^\top$, $i = 1, \ldots, N$.

The covariance estimation by model selection is computed as follows. Let $(g_\lambda)_\lambda$ be an orthonormal basis on $[0,1]$ (wich may differ from the original basis functions g_λ^*). For a given $M > 0$, candidate models are chosen among the collection $\mathcal{M} = \{\{1, \ldots, m\} : m = 1, \ldots, M\}$. To each set indexed by m we associate the matrix (model) $\mathbf{G}_m \in \mathbb{R}^{n \times m}$, with entries $(g_\lambda(t_j))_{1 \leq j \leq n, 1 \leq \lambda \leq m}$, which corresponds to a number m of basis functions $g_1, \ldots, g_m$ in the expansion to approximate the process X. We aim at choosing a good model among the family of models $\{\mathbf{G}_m : m \in \mathcal{M}\}$ in the sense of achieving the minimum of the quadratic risk

$$R(m) = \mathbb{E}\left\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_m\right\|^2 = \|\mathbf{\Sigma} - \mathbf{\Pi}_m \mathbf{\Sigma} \mathbf{\Pi}_m\|^2 + \frac{\delta_m^2 D_m}{N}. \tag{5.2}$$

The ideal model m_0 is the minimizer of the risk function $m \mapsto R(m)$.

Note that for all $m = 1, \ldots, M$,

$$\begin{aligned} L_N(m) &= \frac{1}{N} \sum_{i=1}^{N} \left\|\mathbf{x}_i \mathbf{x}_i^\top - \widehat{\mathbf{\Sigma}}_m\right\| = \mathcal{L}_N(m) + C \\ L_N(m) + pen(m) &= PC(m) + C, \end{aligned}$$

where

$$\begin{aligned} \mathcal{L}_N(m) &= \left\|\mathbf{S} - \widehat{\mathbf{\Sigma}}_m\right\|^2 = \|\mathbf{S} - \mathbf{\Pi}_m \mathbf{S} \mathbf{\Pi}_m\|^2 \\ PC(m) &= \left\|\mathbf{S} - \widehat{\mathbf{\Sigma}}_m\right\|^2 + pen(m) = \|\mathbf{S} - \mathbf{\Pi}_m \mathbf{S} \mathbf{\Pi}_m\|^2 + (1+\theta) \frac{\delta_m^2 D_m}{N}, \end{aligned} \tag{5.3}$$

$\widehat{\mathbf{\Sigma}}_m$ is the least squares covariance estimator of $\mathbf{\Sigma}$ corresponding to the model m as in Theorem 3.2 and the constant C does not depend on m. Thus, $\mathcal{L}_N$ and PC can be regarded as the empirical contrast function and the penalized criterion respectively that will be used for visual presentations of the results. For each model $m = 1, \ldots, M$ we evaluate the penalized criterion (5.3) with $\theta = 1$ and expect that the minimum of PC is attained at a value $\widehat{m}\left(\delta^2\right)$ close to m_0.

The quantity $\delta_m^2 = Tr\left(\left(\Pi_m \otimes \Pi_m\right)\mathbf{\Phi}\right)/D_m$ depends on the matrix $\mathbf{\Phi} = \mathbf{V}\left(vec(\mathbf{x}_1\mathbf{x}_1^\top)^\top\right)$, as pointed out in Section 3.2, which is unknown in practice but can be consistently estimated by (3.9), yielding the plug-in estimate $\widehat{\delta}_m^2 = Tr\left(\left(\Pi_m \otimes \Pi_m\right)\widehat{\mathbf{\Phi}}\right)/D_m$ for δ_m^2. We study the influence of using $\widehat{\delta}^2 = \left(\widehat{\delta}_m^2\right)_{1\leq m\leq M}$ rather than $\delta^2 = \left(\delta_m^2\right)_{1\leq m\leq M}$ on the model selection procedure. Actually, we first compute the following approximation of the risk R,

$$\widehat{R}\left(m\right) = \left\|\mathbf{\Sigma} - \mathbf{\Pi}_m\mathbf{\Sigma}\mathbf{\Pi}_m\right\|^2 + \frac{\widehat{\delta}_m^2 D_m}{N},$$

and then, compute the estimator of the penalized criterion PC

$$\widehat{PC}\left(m\right) = \left\|\mathbf{S} - \mathbf{\Pi}_m\mathbf{S}\mathbf{\Pi}_m\right\|^2 + \left(1+\theta\right)\frac{\widehat{\delta}_m^2 D_m}{N}.$$

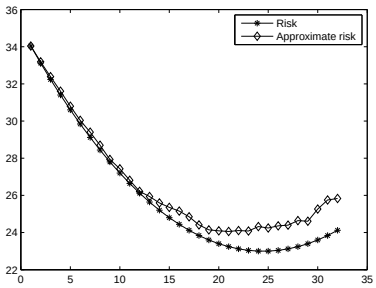
We denote by $\widehat{m}\left(\widehat{\delta}^2\right)$ the point at which the penalized criterion estimate $\widehat{PC}$ attains its minimum value, i.e., the model selected by minimizing $\widehat{PC}$.

In the following examples we plot the empirical contrast function $\mathcal{L}_N$ $(m = 1, \ldots, M)$, the risk function R, the approximate risk function $\widehat{R}$, the penalized criterion PC and the penalized criterion estimate $\widehat{PC}$. We also show figures of the true covariance function $\sigma\left(t,s\right)$ for $s,t \in [0,1]$ and the penalized covariance estimate based on $\widehat{PC}$, i.e., $\widehat{\sigma}\left(t,s\right) = \mathbf{G}_{\widehat{m},t}^\top \widehat{\mathbf{\Psi}}_{\widehat{m}}\mathbf{G}_{\widehat{m},s}$, where $\widehat{m} = \widehat{m}\left(\widehat{\delta}^2\right)$, $\widehat{\mathbf{\Psi}}_{\widehat{m}}$ is obtained as in Theorem 3.2 and $\mathbf{G}_{\widehat{m},t} = \left(g_1\left(t\right), \ldots, g_{\widehat{m}}\left(t\right)\right)^\top \in \mathbb{R}^{\widehat{m}}$ for all $t \in [0,1]$. Furthermore, we will focus attention on finite sample settings, i.e., those in which the number of repetitions N is not notably large (in comparison with the number n of design points t_j).

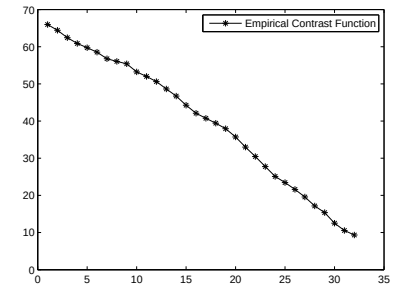
Example 1. *Let $g_1^*, \ldots, g_{m^*}^*$ be the Fourier basis functions*

$$g_\lambda^*\left(t\right) = \begin{cases} \frac{1}{\sqrt{n}} \text{ if } \lambda = 1 \\ \sqrt{2}\frac{1}{\sqrt{n}}\cos\left(2\pi\frac{\lambda}{2}t\right) \text{ if } \frac{\lambda}{2} \in \mathbb{Z} \\ \sqrt{2}\frac{1}{\sqrt{n}}\sin\left(2\pi\frac{\lambda-1}{2}t\right) \text{ if } \frac{\lambda-1}{2} \in \mathbb{Z}_* \end{cases} \tag{5.4}$$

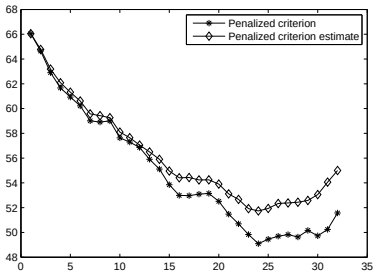
We simulate a sample of size $N = 50$ according to (5.1) with $n = m^* = 35$ and $V\left(a_\lambda\right) = 1$ for all $\lambda = 1, \ldots, m^*$. Set $M = 31$ and consider the models obtained by choosing m Fourier basis functions. In this setting, it can be shown that the minimum of the quadratic risk R is attained at $m_0 = \frac{N}{2} - 1$, which for $N = 50$ gives $m_0 = 24$. Figures 1a, 1b, 1c and 1d present the results obtained for a simulated sample. Figure 1a shows that the approximate risk function $\widehat{R}$ reproduces the shape of the risk function R, so replacing δ^2 by $\widehat{\delta}^2$ into the



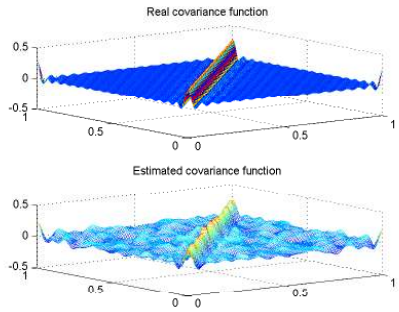
1a. Risk function R and approximate risk $\widehat{R}$.



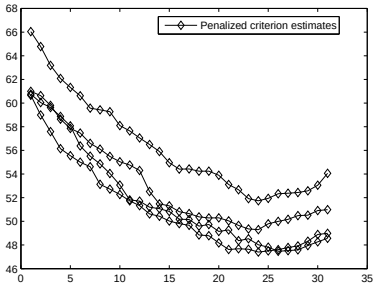
1b. Empirical contrast function $\mathcal{L}_N$.



1c. Penalized criterion PC and its estimate $\widehat{PC}$.



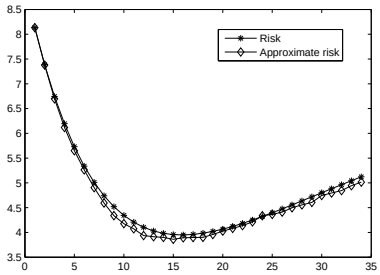
1d. True covariance function σ and estimated covariance function $\widehat{\sigma}$.



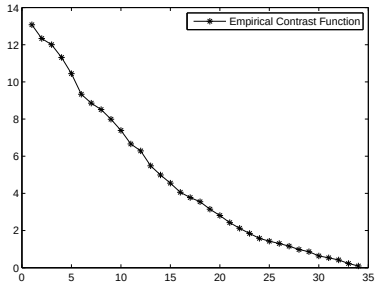
1e. Variability of the simulation process.

FIG 1. *Results of Example 1*

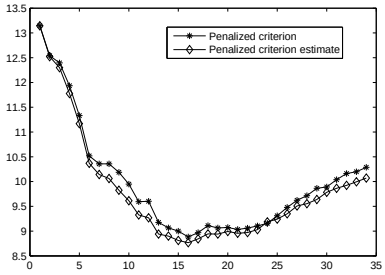
risk does not have a too drastic effect. It can be observed in Figure 1b that, as expected, the empirical contrast function $\mathcal{L}_N$ is strictly decreasing over the whole range of possible models, hence its minimization would lead to choose the largest model $M = 31$. Note that, unlike to what is quite common in univariate linear regression with i.i.d. errors, the empirical contrast curve does not have an "elbow" (i.e., a change of curvature) around the optimal model $m_0 = 24$, which could provide by visual inspection some hint for selecting a suitable model. On the



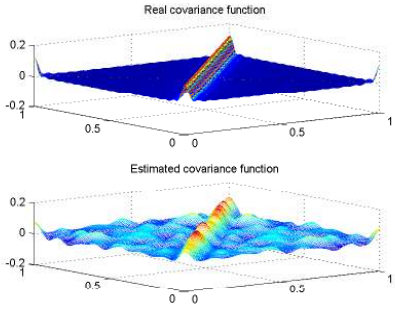
2a. Risk function R and approximate risk $\widehat{R}$.



2b. Empirical contrast function $\mathcal{L}_N$.



2c. Penalized criterion PC and its estimate $\widehat{PC}$.

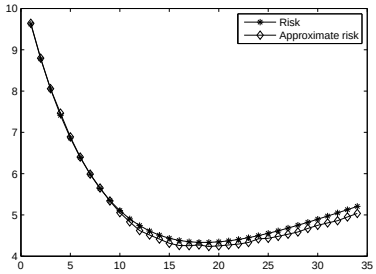


2d. True covariance function σ and estimated covariance function $\widehat{\sigma}$.

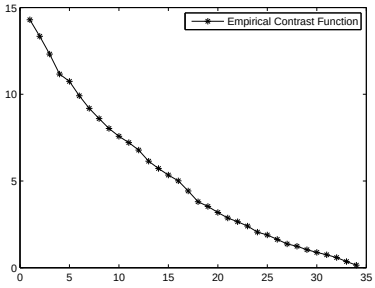
FIG 2. *Results of Example 2*

contrary, both minimization of the penalized criterion PC and its estimate $\widehat{PC}$ lead to select the best model, i.e., $\widehat{m}\left(\delta^2\right) = 24$ and $\widehat{m}\left(\widehat{\delta}^2\right) = 24$ (see Figure 1c). This also demonstrates that replacing δ^2 by $\widehat{\delta}^2$ into the penalized criterion does not notably deteriorate the performance of the model selection procedure in this example. Figure 1d shows that, in spite of the small sample size $N = 50$, a quite nice approximation to the true covariance function σ is achieved by its penalized covariance estimate $\widehat{\sigma}$ based on $\widehat{PC}$. It is clear that the selected model $\widehat{m}\left(\widehat{\delta}^2\right)$ is a random variable that depends on the observed sample $\mathbf{X}$ through the penalized criterion estimate $\widehat{PC}$. Figure 1e illustrates such a variability by plotting the curves $\widehat{PC}$ corresponding to several simulated samples. It can be observed that the selected model $\widehat{m}\left(\widehat{\delta}^2\right)$ is close to the ideal model m_0, and the risk R evaluated at the selected model is much less than that of the largest model $M = 31$ that would be chosen by using the empirical contrast function.

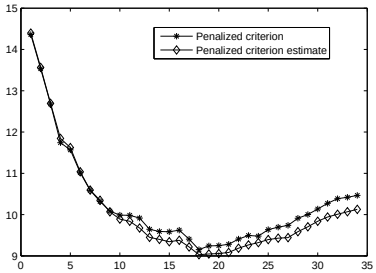
Example 2. *Using the Fourier basis (5.4) we simulate a sample of size $N = 50$ according to (5.1) with $n = m^* = 35$ as in the previous example, but now we set a geometric decay of the variances (or eigenvalues of the covariance operator of the process X) $V(a_\lambda)$, $\lambda = 1, \ldots, m^*$; namely, $V(a_1) = r$ and*



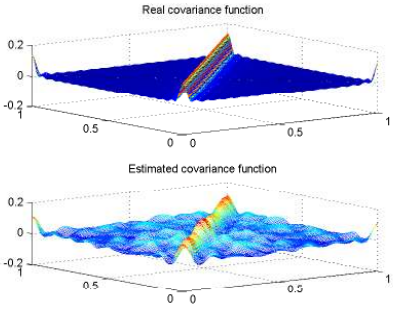
3a. Risk function R and approximate risk $\widehat{R}$.



3b. Empirical contrast function $\mathcal{L}_N$.



3c. Penalized criterion PC and its estimate $\widehat{PC}$.

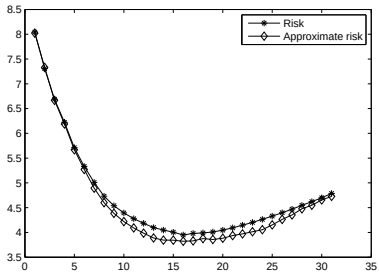


3d. True covariance function σ and estimated covariance function $\widehat{\sigma}$.

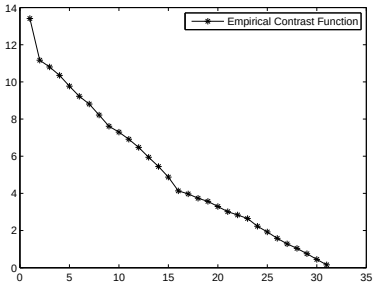
FIG 3. *Results of Example 3.*

$V(a_{\lambda+1}) = V(a_\lambda)\, r$ *for* $\lambda = 2, 3, ..,$ *where* $r = 0.95$. *We consider a collection of models up to* $M = 34$, *with* m *Fourier basis functions. In this setting it can be proved that the minimum of the risk* R *is attained at* $m_0 = (\log(2/[(1-r)(N-2)+2]))/\log(r)$, *which yields* $m_0 = 16$ *for the actual values* $N = 50$ *and* $r = 0.95$. *The results obtained from a simulated sample are shown in Figures 2a, 2b, 2c and 2d. It can be noted that the empirical contrast function is strictly decreasing without any "elbow" effect, while the selected model by both the penalized criterion and the penalized criterion estimate is* $\widehat{m}(\delta^2) = \widehat{m}(\widehat{\delta^2}) = 16$, *which is the best model* m_0 *according to the risk* R.

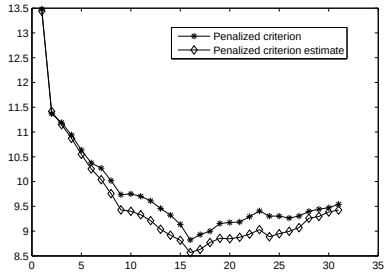
Example 3. *Using the Fourier basis (5.4) we simulate a sample of size* $N = 60$ *according to (5.1) with* $n = m^* = 35$, *but now we set the variances (eigenvalues) as follows:* $V(a_\lambda) = \sigma^2 + r^\lambda$ *for all* $\lambda = 1, \ldots, m^*$, *where* $r = 0.95$ *and* $\sigma^2 = 0.0475$. *This decay of the eigenvalues is common in the factor models. Actually all the eigenvalues have almost the same small value* σ^2 *(corresponding to "noise") except for a few first eigenvalues that have larger values (corresponding to some "factors"). The collection of models considered corresponds to a number* m $(1 \le m \le M)$ *of Fourier basis functions up to* $M = 34$. *The results from a simulated sample are shown in Figures 3a, 3b, 3c and 3d. Fig-*



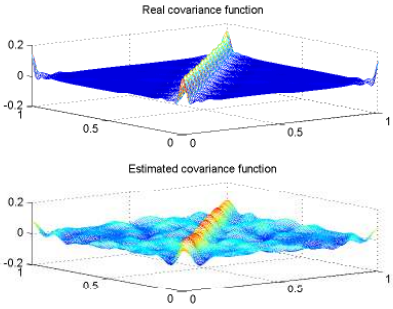
4a. Risk function R and approximate risk $\widehat{R}$.



4b. Empirical contrast function $\mathcal{L}_N$.



4c. Penalized criterion PC and its estimate $\widehat{PC}$.



4d. True covariance function σ and estimated covariance function $\widehat{\sigma}$.

FIG 4. *Results of Example 4*

ure 3a shows that the minimum of the risk function R is attained at $m_0 = 18$. Likewise the previous examples, the empirical contrast function is strictly decreasing without any "elbow", while the model selection procedure chooses the model $\widehat{m}(\delta^2) = \widehat{m}(\widehat{\delta^2}) = 18$, which is the value of m_0.

Example 4. *In this example we use different basis functions for generating the data and for estimating the covariance. Specifically, the process is generated using a wavelet basis and the collection of models considered in the model selection procedure corresponds to different numbers of Fourier basis functions up to $M = 31$. We simulate a sample of size $N = 50$ according to (5.1) using the Symmlet 8 wavelet basis, with $n = m^* = 32$. We set the variances of the random coefficients a_λ with a geometric decay likewise in Example 2, i.e., $V(a_1) = r$ and $V(a_{\lambda+1}) = V(a_\lambda)\, r$, where $r = 0.95$. The results of one simulation are displayed in Figures 4a, 4b, 4c and 4d. Here it can be also observed that the penalized estimation procedure shows good performance even when using an estimate of the penalized criterion, leading to choosing the model $\widehat{m}(\widehat{\delta^2}) = \widehat{m}(\delta^2) = 16 = m_0$.*

Summarizing the results of these simulated examples, we may conclude that for not so large sample sizes N:

a) The empirical contrast function $\mathcal{L}_N$ is useless to select a model that attains a low risk. It is a strictly decreasing function whose minimization leads to simply choose the largest model M within the set of candidate models $m = 1, \ldots, M$. Furthermore, frequently the curve $\mathcal{L}_N$ does not have an "elbow" that could guide researchers to choose a suitable model by exploratory analysis.

b) The covariance function estimator by model selection introduced in this paper shows good performance in a variety of examples when based on the penalized criterion PC but also when using the estimated penalty $\widehat{PC}$.

6. Appendix

6.1. Proofs of preliminary results

Proof of Proposition 3.1

Proof. *a*) The minimization problem is equivalent to minimize

$$h(\mathbf{\Psi}) = \left\| \overline{\mathbf{Y}} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top \right\|^2.$$

The Frobenius norm $\|.\|$ is invariant by the vec operation. Furthermore, $\mathbf{\Psi} \in \mathcal{S}_m$ can be represented by means of $vec(\mathbf{\Psi}) = \mathbf{D}_m \beta$ where $\beta \in \mathbb{R}^{|m|(|m|+1)/2}$. These facts and the identity

$$vec(\mathbf{ABC}) = \left(\mathbf{C}^\top \otimes \mathbf{A}\right) vec(\mathbf{B}) \tag{6.1}$$

allow one to rewrite

$$h(\mathbf{\Psi}) = \|\overline{\mathbf{y}} - (\mathbf{G} \otimes \mathbf{G})\mathbf{D}_m \beta\|^2,$$

where $\overline{\mathbf{y}} = vec\left(\overline{\mathbf{Y}}\right)$. Minimization of this quadratic function with respect to β in $\mathbb{R}^{|m|(|m|+1)/2}$ is equivalent to solve the normal equation

$$\mathbf{D}_m^\top (\mathbf{G} \otimes \mathbf{G})^\top (\mathbf{G} \otimes \mathbf{G})\mathbf{D}_m \beta = \mathbf{D}_m^\top (\mathbf{G} \otimes \mathbf{G})^\top \overline{\mathbf{y}}.$$

By using the identities

$$\mathbf{D}_m^\top vec(\mathbf{A}) = vech\left(\mathbf{A} + \mathbf{A}^\top - diag(\mathbf{A})\right)$$

and 6.1, said normal equation can be rewritten

$$vech\left(\mathbf{G}^\top\mathbf{G}\left(\mathbf{\Psi} + \mathbf{\Psi}^\top\right)\mathbf{G}^\top\mathbf{G} - diag\left(\mathbf{G}^\top\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\mathbf{G}\right)\right) =$$
$$vech\left(\mathbf{G}^\top\left(\overline{\mathbf{Y}} + \overline{\mathbf{Y}}^\top\right)\mathbf{G} - diag\left(\mathbf{G}^\top\overline{\mathbf{Y}}\mathbf{G}\right)\right).$$

Finally, it can be verified that $\widehat{\mathbf{\Psi}}$ given by (3.3) satisfies this equation as a consequence of the fact that such $\widehat{\mathbf{\Psi}}$ it holds that

$$vech\left(\mathbf{G}^\top\mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\mathbf{G}\right) = vech\left(\mathbf{G}^\top\left(\frac{\overline{\mathbf{Y}} + \overline{\mathbf{Y}}^\top}{2}\right)\mathbf{G}\right).$$

b) It straightforwardly follows from part *a*). □

6.2. Proofs of main results

Proof of Proposition (4.1)

Proof. The proof follows the guidelines of the proof in [2]. More generally we will prove that for any $\eta > 0$ and any sequence of positive numbers L_m, if the penalty function $pen : \mathcal{M} \longrightarrow \mathbb{R}_+$ is chosen to satisfy:

$$pen(m) = (1 + \eta + L_m) \frac{\delta_m^2}{N} D_m \text{ for all } m \in \mathcal{M}, \tag{6.2}$$

then for each $x > 0$ and $p \geq 2$

$$\mathbb{P}\left(\mathcal{H}(\mathbf{f}) \geq \left(1 + \frac{2}{\eta}\right) \frac{x}{N} \delta_m^2\right) \leq c(p, \eta) \, \mathbb{E}\|\varepsilon_1\|^p \sum_{m \in \mathcal{M}} \frac{1}{\delta_m^p} \frac{D_m \vee 1}{(L_m D_m + x)^{p/2}}, \tag{6.3}$$

where we have set

$$\mathcal{H}(\mathbf{f}) = \left[\left\|\mathbf{f} - \widetilde{\mathbf{f}}\right\|_N^2 - \left(2 - \frac{4}{\eta}\right) \inf_{m \in \mathcal{M}} \left\{d_N^2(\mathbf{f}, \mathcal{L}_m) + pen(m)\right\}\right]_+.$$

To obtain (4.4), take $\eta = \frac{\theta}{2} = L_m$. As for each $m \in \mathcal{M}$,

$$\begin{aligned} d_N^2(\mathbf{f}, \mathcal{L}_m) + pen(m) &\leq d_N^2(\mathbf{f}, \mathcal{L}_m) + (1 + \theta) \frac{\delta_m^2}{N} D_m \\ &\leq (1 + \theta) \left(d_N^2(\mathbf{f}, \mathcal{L}_m) + \frac{\delta_m^2}{N} D_m\right) \end{aligned}$$

we get that for all $q > 0$,

$$\mathcal{H}^q(\mathbf{f}) \geq \left[\left\|\mathbf{f} - \widetilde{\mathbf{f}}\right\|_N^2 - \left(2 + \frac{8}{\theta}\right)(1 + \theta) M_N^*\right]_+^q = \left[\left\|\mathbf{f} - \widetilde{\mathbf{f}}\right\|_N^2 - K(\theta) M_N^*\right]_+^q, \tag{6.4}$$

where $K(\theta) = \left(2 + \frac{8}{\theta}\right)(1 + \theta)$.

Since

$$\mathbb{E}\left(\mathcal{H}^q(\mathbf{f})\right) = \int_0^\infty q u^{q-1} \mathbb{P}\left(\mathcal{H}(\mathbf{f}) > u\right) du,$$

we derive from (6.4) and (6.3) that for all $p > 2(1 + q)$

$$\begin{aligned} &\mathbb{E}\left[\left(\left\|\mathbf{f} - \widetilde{\mathbf{f}}\right\|_N^2 - K(\theta) M_N^*\right)_+^q\right] \\ &\leq \mathbb{E}\left(\mathcal{H}^q(\mathbf{f})\right) \\ &\leq c(p, \theta) \left(1 + \frac{4}{\theta}\right)^q \frac{\mathbb{E}\|\varepsilon_1\|^p}{N^q} \sum_{m \in \mathcal{M}} \frac{\delta_m^{2q}}{\delta_m^p} \int_0^\infty q x^{q-1} \left[\frac{D_m \vee 1}{\left(\frac{\theta}{2} D_m + x\right)^{p/2}} \wedge 1\right] dx \\ &\leq c'(p, q, \theta) \frac{\mathbb{E}\|\varepsilon_1\|^p}{N^q} \delta_{\sup}^{2q} \left[\sum_{m \in \mathcal{M}} \delta_m^{-p} D_m^{-(p/2 - 1 - q)}\right] \end{aligned}$$

using that $\mathbb{P}\left(\mathcal{H}\left(\mathbf{f}\right) > u\right) \leq 1$. Indeed, for $m \in \mathcal{M}$ such that $D_m \geq 1$, using that $q - 1 - p/2 < 0$, we get

$$
\begin{aligned}
&\frac{\delta_m^{2q}}{\delta_m^p} \int_0^\infty q x^{q-1} \left[\frac{D_m \vee 1}{\left(\frac{\theta}{2} D_m + x\right)^{p/2}} \wedge 1 \right] dx \\
&\quad \leq \delta_{\sup}^{2q} \delta_m^{-p} \int_0^\infty q x^{q-1} \left[\frac{D_m}{\left(\frac{\theta}{2} D_m + x\right)^{p/2}} \right] dx \\
&\quad = \delta_{\sup}^{2q} \delta_m^{-p} \left(\int_0^{D_m} q x^{q-1} \left[\frac{D_m}{\left(\frac{\theta}{2} D_m + x\right)^{p/2}} \right] dx + \int_{D_m}^\infty q x^{q-1} \left[\frac{D_m}{\left(\frac{\theta}{2} D_m + x\right)^{p/2}} \right] dx \right) \\
&\quad \leq \delta_{\sup}^{2q} \delta_m^{-p} \left(\frac{D_m}{\left(\frac{\theta}{2} D_m\right)^{p/2}} \int_0^{D_m} q x^{q-1} dx + D_m \int_{D_m}^\infty q x^{q-1} \left[\frac{1}{x^{p/2}} \right] dx \right) \\
&\quad = \delta_{\sup}^{2q} \delta_m^{-p} \left(2^{p/2} \theta^{-p/2} D_m^{1-p/2} \int_0^{D_m} q x^{q-1} dx + D_m \int_{D_m}^\infty q x^{q-1-p/2} dx \right) \\
&\quad = \delta_{\sup}^{2q} \delta_m^{-p} \left(2^{p/2} \theta^{-p/2} D_m^{1-p/2} \left[D_m^q \right] + D_m \left[\frac{q}{p/2 - q} D_m^{q-p/2} \right] \right) \\
&\quad = \delta_{\sup}^{2q} \delta_m^{-p} \left(2^{p/2} \theta^{-p/2} D_m^{1-p/2+q} + D_m^{1-p/2+q} \left[\frac{q}{p/2 - q} \right] \right) \\
&\quad = \delta_{\sup}^{2q} \delta_m^{-p} \left(D_m^{-(p/2-1-q)} \left[2^{p/2} \theta^{-p/2} + \frac{q}{p/2 - q} \right] \right). \qquad (6.5)
\end{aligned}
$$

Inequality (6.5) enables to conclude that (4.4) holds assuming (6.3).

We now turn to the proof of (6.3). Recall that, we identify the function $g : \mathcal{X} \to \mathbb{R}^k$ with vectors $\left(g\left(x_1\right) \dots g\left(x_N\right)\right)^\top \in \mathbb{R}^{Nk}$ and we denote by $\langle a, b \rangle_N = \frac{1}{N} \sum_{i=1}^N a_i^\top b_i$ the inner product of $\mathbb{R}^{Nk}$ associated to the norm $\|.\|_N$, where $a = \left(a_1 \dots a_N\right)^\top$ and $b = \left(b_1 \dots b_N\right)^\top$ with $a_i, b_i \in \mathbb{R}^k$ for all $i = 1, \dots, N$. For each $m \in \mathcal{M}$ we denote by $\mathbf{P}_m$ the orthogonal projector onto the linear space $\left\{\left(g\left(x_1\right) \dots g\left(x_N\right)\right)^\top : g \in \mathcal{L}_m\right\} \subset \mathbb{R}^{Nk}$. This linear space is also denoted by $\mathcal{L}_m$. From now on, the subscript m denotes any minimizer of the function $m' \to \|\mathbf{f} - \mathbf{P}_{m'}\mathbf{f}\|_N^2 + pen\left(m'\right)$, $m' \in \mathcal{M}_N$. For any $\mathbf{g} \in \mathbb{R}^{Nk}$ we define the least-squares loss function by

$$\gamma_N\left(\mathbf{g}\right) = \|\mathbf{y} - \mathbf{g}\|_N^2$$

Using the definition of γ_N we have that for all $\mathbf{g} \in \mathbb{R}^{Nk}$,

$$\gamma_N\left(\mathbf{g}\right) = \|\mathbf{f} + \ \varepsilon - \mathbf{g}\|_N^2 .$$

Then we derive that

$$\|\mathbf{f} - \mathbf{g}\|_N^2 = \gamma_N\left(\mathbf{f}\right) + 2\left\langle \mathbf{f} - \mathbf{y},\ \varepsilon \right\rangle_N + \|\ \varepsilon\|_N^2$$

and therefore

$$\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 - \|\mathbf{f}-\mathbf{P}_m\mathbf{f}\|_N^2 = \gamma_N\left(\widetilde{\mathbf{f}}\right) - \gamma_N\left(\mathbf{P}_m\mathbf{f}\right) + 2\left\langle \widetilde{\mathbf{f}}-\mathbf{P}_m\mathbf{f},\ \varepsilon\right\rangle_N. \quad (6.6)$$

By the definition of $\widetilde{\mathbf{f}}$, we know that

$$\gamma_N\left(\widetilde{\mathbf{f}}\right) + pen\left(\widehat{m}\right) \leq \gamma_N\left(\mathbf{g}\right) + pen\left(m\right)$$

for all $m \in \mathcal{M}$ and for all $\mathbf{g} \in \mathcal{L}_m$. Then

$$\gamma_N\left(\widetilde{\mathbf{f}}\right) - \gamma_N\left(\mathbf{P}_m\mathbf{f}\right) \leq pen\left(m\right) - pen\left(\widehat{m}\right). \quad (6.7)$$

So we get from (6.6) and (6.7) that

$$\begin{aligned}\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 \leq{}& \|\mathbf{f}-\mathbf{P}_m\mathbf{f}\|_N^2 + pen\left(m\right) - pen\left(\widehat{m}\right)\\ &+ 2\left\langle \mathbf{f}-\mathbf{P}_m\mathbf{f},\ \varepsilon\right\rangle_N + 2\left\langle \mathbf{P}_{\widehat{m}}\mathbf{f}-\mathbf{f},\ \varepsilon\right\rangle_N + 2\left\langle \widetilde{\mathbf{f}}-\mathbf{P}_{\widehat{m}}\mathbf{f},\ \varepsilon\right\rangle_N. \quad (6.8)\end{aligned}$$

In the following we set for each $m' \in \mathcal{M}$,

$$\begin{aligned}\mathcal{B}_{m'} &= \left\{\mathbf{g} \in \mathcal{L}_{m'} : \|\mathbf{g}\|_N \leq 1\right\},\\ G_{m'} &= \sup_{t\in\mathcal{B}_{m'}} \left\langle \mathbf{g},\ \varepsilon\right\rangle_N = \left\|\mathbf{P}_{m'}\ \varepsilon\right\|_N,\\ \mathbf{u}_{m'} &= \begin{cases} \frac{\mathbf{P}_{m'}\mathbf{f}-\mathbf{f}}{\|\mathbf{P}_{m'}\mathbf{f}-\mathbf{f}\|_N} & \text{if } \|\mathbf{P}_{m'}\mathbf{f}-\mathbf{f}\|_N \neq 0\\ 0 & \text{otherwise.}\end{cases}\end{aligned}$$

Since $\widetilde{\mathbf{f}} = \mathbf{P}_{\widehat{m}}\ \mathbf{f} + \mathbf{P}_{\widehat{m}}\ \varepsilon$, (6.8) gives

$$\begin{aligned}\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 \leq{}& \|\mathbf{f}-\mathbf{P}_m\mathbf{f}\|_N^2 + pen\left(m\right) - pen\left(\widehat{m}\right)\\ &+ 2\left\|\mathbf{f}-\mathbf{P}_m\mathbf{f}\right\|_N\left|\left\langle \mathbf{u}_m,\ \varepsilon\right\rangle_N\right| + 2\left\|\mathbf{f}-\mathbf{P}_{\widehat{m}}\mathbf{f}\right\|_N\left|\left\langle \mathbf{u}_{\widehat{m}},\ \varepsilon\right\rangle_N\right| + 2G_{\widehat{m}}^2. \quad (6.9)\end{aligned}$$

Using repeatedly the following elementary inequality that holds for all positive numbers α, x, z

$$2xz \leq \alpha x^2 + \frac{1}{\alpha}z^2 \quad (6.10)$$

we get for any $m' \in \mathcal{M}$

$$2\left\|\mathbf{f}-\mathbf{P}_{m'}\mathbf{f}\right\|_N\left|\left\langle \mathbf{u}_{m'},\ \varepsilon\right\rangle_N\right| \leq \alpha\left\|\mathbf{f}-\mathbf{P}_{m'}\mathbf{f}\right\|_N^2 + \frac{1}{\alpha}\left|\left\langle \mathbf{u}_{m'},\ \varepsilon\right\rangle_N\right|^2. \quad (6.11)$$

By Pythagoras Theorem we have

$$\begin{aligned}\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 &= \|\mathbf{f}-\mathbf{P}_{\widehat{m}}\mathbf{f}\|_N^2 + \left\|\mathbf{P}_{\widehat{m}}\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2\\ &= \|\mathbf{f}-\mathbf{P}_{\widehat{m}}\mathbf{f}\|_N^2 + G_{\widehat{m}}^2. \quad (6.12)\end{aligned}$$

We derive from (6.9) and (6.11) that for any $\alpha > 0$:

$$\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 \leq \|\mathbf{f}-\mathbf{P}_m\mathbf{f}\|_N^2 + \alpha\|\mathbf{f}-\mathbf{P}_m\mathbf{f}\|_N^2 + \frac{1}{\alpha}\langle \mathbf{u}_m,\ \varepsilon\rangle_N^2 + \alpha\|\mathbf{f}-\mathbf{P}_{\widehat{m}}\mathbf{f}\|_N^2 + \frac{1}{\alpha}\langle \mathbf{u}_{\widehat{m}},\ \varepsilon\rangle_N^2 + 2G_{\widehat{m}}^2 + pen(m) - pen(\widehat{m}).$$

Now taking into account that by equation (6.12) $\|\mathbf{f}-\mathbf{P}_{\widehat{m}}\mathbf{f}\|_N^2 = \left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 - G_{\widehat{m}}^2$ the above inequality is equivalent to:

$$(1-\alpha)\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 \leq (1+\alpha)\|\mathbf{f}-\mathbf{P}_m\mathbf{f}\|_N^2 + \frac{1}{\alpha}\langle \mathbf{u}_m,\varepsilon\rangle_N^2 + \frac{1}{\alpha}\langle \mathbf{u}_{\widehat{m}},\varepsilon\rangle_N^2 + (2-\alpha)G_{\widehat{m}}^2 + pen(m) - pen(\widehat{m}). \quad (6.13)$$

We choose $\alpha = \frac{2}{2+\eta} \in]0,1[$, but for sake of simplicity we keep using the notation α. Let $\widetilde{p}_1$ and $\widetilde{p}_2$ be two functions depending on η mapping $\mathcal{M}$ into $\mathbb{R}_+$. They will be specified later to satisfy

$$pen(m') \geq (2-\alpha)\widetilde{p}_1(m') + \frac{1}{\alpha}\widetilde{p}_2(m') \ \ \forall (m') \in \mathcal{M}. \quad (6.14)$$

Since $\frac{1}{\alpha}\widetilde{p}_2(m') \leq pen(m')$ and $1+\alpha \leq 2$, we get from (6.13) and (6.14) that

$$\begin{aligned}(1-\alpha)\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 &\leq (1+\alpha)\|\mathbf{f}-\mathbf{P}_m\mathbf{f}\|_N^2 + pen(m) \\ &\quad + \frac{1}{\alpha}\widetilde{p}_2(m) + (2-\alpha)\left(G_{\widehat{m}}^2 - \widetilde{p}_1(\widehat{m})\right) \\ &\quad + \frac{1}{\alpha}\left(\langle \mathbf{u}_{\widehat{m}},\varepsilon\rangle_N^2 - \widetilde{p}_2(\widehat{m})\right) + \frac{1}{\alpha}\left(\langle \mathbf{u}_m,\varepsilon\rangle_N^2 - \widetilde{p}_2(m)\right) \\ &\leq 2\left(\|\mathbf{f}-\mathbf{P}_m\mathbf{f}\|_N^2 + pen(m)\right) + (2-\alpha)\left(G_{\widehat{m}}^2 - \widetilde{p}_1(\widehat{m})\right) \\ &\quad + \frac{1}{\alpha}\left(\langle \mathbf{u}_{\widehat{m}},\varepsilon\rangle_N^2 - \widetilde{p}_2(\widehat{m})\right) + \frac{1}{\alpha}\left(\langle \mathbf{u}_m,\varepsilon\rangle_N^2 - \widetilde{p}_2(m)\right). \end{aligned} \quad (6.15)$$

As $\frac{2}{1-\alpha} = 2 + \frac{4}{\eta}$ we obtain that

$$\begin{aligned}&(1-\alpha)\mathcal{H}(\mathbf{f}) \\ &= \left\{(1-\alpha)\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 - (1-\alpha)\left(2+\frac{4}{\eta}\right)\inf_{m'\in\mathcal{M}}\left(\|\mathbf{f}-\mathbf{P}_{m'}\mathbf{f}\|_N^2 + pen(m')\right)\right\}_+ \\ &= \left\{(1-\alpha)\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 - 2\left(\|\mathbf{f}-\mathbf{P}_m\mathbf{f}\|_N^2 + 2pen(m)\right)\right\}_+ \\ &\leq \left\{(2-\alpha)\left(G_{\widehat{m}}^2 - \widetilde{p}_1(\widehat{m})\right) + \frac{1}{\alpha}\left(\langle \mathbf{u}_{\widehat{m}},\varepsilon\rangle_N^2 - \widetilde{p}_2(\widehat{m})\right)\right. \\ &\quad \left. + \frac{1}{\alpha}\left(\langle \mathbf{u}_m,\varepsilon\rangle_N^2 - \widetilde{p}_2(m)\right)\right\}_+ \end{aligned}$$

using that m minimizes the function $\|\mathbf{f}-\mathbf{P}_{m'}\|^2+pen\,(m')$ and (6.15).

For any $x>0$,

$$\begin{aligned}\mathbb{P}\left((1-\alpha)\,\mathcal{H}\,(\mathbf{f})\geq\frac{x\delta_m^2}{N}\right)&\leq\mathbb{P}\left(\exists m'\in\mathcal{M}:(2-\alpha)\left(G_{m'}^2-\widetilde{p}_1\,(m')\right)\geq\frac{x\delta_{m'}^2}{3N}\right)\\&\quad+\mathbb{P}\left(\exists m'\in\mathcal{M}:\frac{1}{\alpha}\left(\langle\mathbf{u}_{m'},\varepsilon\rangle_N^2-\widetilde{p}_2\,(m')\right)\geq\frac{x\delta_{m'}^2}{3N}\right)\\&\leq\sum_{m'\in\mathcal{M}}\mathbb{P}\left((2-\alpha)\left(\|\mathbf{P}_{m'}\varepsilon\|_N^2-\widetilde{p}_1\,(m')\right)\geq\frac{x\delta_{m'}^2}{3N}\right)\\&\quad+\sum_{m'\in\mathcal{M}}\mathbb{P}\left(\frac{1}{\alpha}\left(\langle\mathbf{u}_{m'},\varepsilon\rangle_N^2-\widetilde{p}_2\,(m')\right)\geq\frac{x\delta_{m'}^2}{3N}\right)\\&:=\sum_{m'\in\mathcal{M}}P_{1,m'}\,(x)+\sum_{m'\in\mathcal{M}}P_{2,m'}\,(x)\,.\end{aligned}\tag{6.16}$$

We first bound $P_{2,m'}\,(x)$. Let t be some positive number,

$$\mathbb{P}\left(|\langle\mathbf{u}_{m'},\varepsilon\rangle_N|\geq t\right)\leq t^{-p}\mathbb{E}\left(|\langle\mathbf{u}_{m'},\varepsilon\rangle_N|^p\right).\tag{6.17}$$

Since $\langle\mathbf{u}_{m'},\varepsilon\rangle_N=\frac{1}{N}\sum_{i=1}^N\langle\mathbf{u}_{im'},\varepsilon_i\rangle$ with ε_i i.i.d. and with zero mean, then by Rosenthal's inequality we know that for some constant $c\,(p)$ that depends on p only

$$\begin{aligned}c^{-1}\,(p)\,N^p\mathbb{E}\,|\langle\mathbf{u}_{m'},\varepsilon\rangle_N|^p&\leq\sum_{i=1}^N\mathbb{E}\,|\langle\mathbf{u}_{im'},\varepsilon_i\rangle|^p+\left(\sum_{i=1}^N\mathbb{E}\left(\langle\mathbf{u}_{im'},\varepsilon_i\rangle^2\right)\right)^{\frac{p}{2}}\\&\leq\sum_{i=1}^N\mathbb{E}\,\|\mathbf{u}_{im'}\|^p\,\|\varepsilon_i\|^p+\left(\sum_{i=1}^N\mathbb{E}\,\|\mathbf{u}_{im'}\|^2\,\|\varepsilon_i\|^2\right)^{\frac{p}{2}}\\&=\mathbb{E}\,\|\varepsilon_1\|^p\sum_{i=1}^N\|\mathbf{u}_{im'}\|^p+\left(\mathbb{E}\,\|\varepsilon_1\|^2\right)^{\frac{p}{2}}\left(\sum_{i=1}^N\|\mathbf{u}_{im'}\|^2\right)^{\frac{p}{2}}.\end{aligned}\tag{6.18}$$

Since $p\geq2$, $\left(\mathbb{E}\,\|\varepsilon_1\|^2\right)^{\frac{1}{2}}\leq\left(\mathbb{E}\,\|\varepsilon_1\|^p\right)^{\frac{1}{p}}$ and

$$\left(\mathbb{E}\,\|\varepsilon_1\|^2\right)^{\frac{p}{2}}\leq\mathbb{E}\,\|\varepsilon_1\|^p.\tag{6.19}$$

Using also that by definition $\|\mathbf{u}_{m'}\|_N^2=\frac{1}{N}\sum_{i=1}^N\|\mathbf{u}_{im'}\|^2=1$, then $\frac{\|\mathbf{u}_{im'}\|^2}{N}\leq1$ and therefore $\frac{\|\mathbf{u}_{im'}\|}{N^{\frac{1}{2}}}\leq1$. Thus

$$\sum_{i=1}^N\|\mathbf{u}_{im'}\|^p=N^{\frac{p}{2}}\sum_{i=1}^N\left(\frac{\|\mathbf{u}_{im'}\|}{N^{\frac{1}{2}}}\right)^p\leq N^{\frac{p}{2}}\sum_{i=1}^N\left(\frac{\|\mathbf{u}_{im'}\|}{N^{\frac{1}{2}}}\right)^2=N^{\frac{p}{2}}\,\|\mathbf{u}_{m'}\|_N^2=N^{\frac{p}{2}}.\tag{6.20}$$

We deduce from (6.18), (6.19) and (6.20) that

$$c^{-1}(p) N^p \mathbb{E} |\langle \mathbf{u}_{m'}, \varepsilon \rangle_N|^p \leq \mathbb{E} \|\varepsilon_1\|^p N^{\frac{p}{2}} + \mathbb{E} \|\varepsilon_1\|^p N^{\frac{p}{2}}.$$

Then for some constant $c'(p)$ that only depends on p

$$\mathbb{E} |\langle \mathbf{u}_{m'}, \varepsilon \rangle_N|^p \leq c'(p) \mathbb{E} \|\varepsilon_1\|^p N^{-\frac{p}{2}}.$$

By this last inequality and (6.17) we get that

$$\mathbb{P}(|\langle \mathbf{u}_{m'}, \varepsilon \rangle_N| \geq t) \leq c'(p) \mathbb{E} \|\varepsilon_1\|^p N^{-\frac{p}{2}} t^{-p}. \tag{6.21}$$

Let υ be some positive number depending on η only to be chosen later. We take t such that $Nt^2 = \min\left(\upsilon, \frac{\alpha}{3}\right)(L_{m'} D_{m'} + x)\delta_{m'}^2$ and set $N\widetilde{p}_2(m') = \upsilon L_{m'} D_{m'} \delta_{m'}^2$. We get

$$\begin{aligned}
P_{2,m'}(x) &= \mathbb{P}\left(\frac{1}{\alpha}\left(\langle \mathbf{u}_{m'}, \varepsilon \rangle_N^2 - \widetilde{p}_2(m')\right) \geq \frac{x\delta_{m'}^2}{3N}\right) \\
&= \mathbb{P}\left(N \langle \mathbf{u}_{m'}, \varepsilon \rangle_N^2 \geq N\widetilde{p}_2(m') + \alpha \frac{\delta_{m'}^2}{3} x\right) \\
&= \mathbb{P}\left(N \langle \mathbf{u}_{m'}, \varepsilon \rangle_N^2 \geq \upsilon L_{m'} D_{m'} \delta_{m'}^2 + \alpha \frac{\delta_{m'}^2}{3} x\right) \\
&\leq \mathbb{P}\left(|\langle \mathbf{u}_{m'}, \varepsilon \rangle_N| \geq N^{-\frac{1}{2}} \sqrt{\min\left(\upsilon, \frac{\alpha}{3}\right)} \sqrt{(L_{m'} D_{m'} + x)} \delta_{m'}\right) \\
&\leq c'(p) \mathbb{E} \|\varepsilon_1\|^p N^{-\frac{p}{2}} \frac{N^{\frac{p}{2}}}{\left(\min\left(\upsilon, \frac{\alpha}{3}\right)\right)^{\frac{p}{2}} (L_{m'} D_{m'} + x)^{\frac{p}{2}} \delta_{m'}^p} \\
&= c''(p, \eta) \frac{\mathbb{E} \|\varepsilon_1\|^p}{\delta_{m'}^p} \frac{1}{(L_{m'} D_{m'} + x)^{\frac{p}{2}}}.
\end{aligned} \tag{6.22}$$

The last inequality holds using (6.21).

We now bound $P_{1,m'}(x)$ for those $m' \in \mathcal{M}$ such that $D_{m'} \geq 1$. By using our version of Corollary 5.1 in Baraud with $\widetilde{A} = \mathbf{P}_{m'}$, $\mathrm{Tr}(\widetilde{A}) = D_{m'}$ and $\rho(\widetilde{A}) = 1$, we obtain from (4.5) that for any positive $x_{m'}$

$$\mathbb{P}\left(N \|\mathbf{P}_{m'} \varepsilon\|_N^2 \geq \delta_{m'}^2 D_{m'} + 2\delta_{m'}^2 \sqrt{D_{m'} x_{m'}} + \delta_{m'}^2 x_{m'}\right) \leq C(p) \frac{\mathbb{E} \|\varepsilon_1\|^p}{\delta_{m'}^p} D_{m'} x_{m'}^{-\frac{p}{2}}. \tag{6.23}$$

Since for any $\beta > 0$, $2\sqrt{D_{m'} x_{m'}} \leq \beta D_{m'} + \beta^{-1} x_{m'}$ then (6.23) imply that

$$\mathbb{P}\left(N \|\mathbf{P}_{m'} \varepsilon\|_N^2 \geq (1+\beta) D_{m'} \delta_{m'}^2 + \left(1 + \beta^{-1}\right) x_{m'} \delta_{m'}^2\right) \leq C(p) \frac{\mathbb{E} \|\varepsilon_1\|^p}{\delta_{m'}^p} D_{m'} x_{m'}^{-\frac{p}{2}}. \tag{6.24}$$

Now for some number β depending on η only to be chosen later, we take

$$x_{m'} = \left(1 + \beta^{-1}\right) \min\left(\upsilon, \frac{(2-\alpha)^{-1}}{3}\right) (L_{m'} D_{m'} + x)$$

and $N\widetilde{p}_1(m') = \upsilon L_{m'} D_{m'} \delta^2_{m'} + (1+\beta) D_{m'} \delta^2_{m'}$. By (6.24) this gives

$$\begin{aligned}
P_{1,m'}(x) &= \mathbb{P}\left(\|\mathbf{P}_{m'}\varepsilon\|_N^2 - \widetilde{p}_1(m') \geq \frac{(2-\alpha)^{-1} x \delta^2_{m'}}{3N} \right) \\
&= \mathbb{P}\left(N\|\mathbf{P}_{m'}\varepsilon\|_N^2 \geq \upsilon L_{m'} D_{m'} \delta^2_{m'} + (1+\beta) D_{m'} \delta^2_{m'} + \frac{(2-\alpha)^{-1}}{3} x \delta^2_{m'} \right) \\
&\leq \mathbb{P}\left(N\|\mathbf{P}_{m'}\varepsilon\|_N^2 \geq (1+\beta) D_{m'} \delta^2_{m'} + \left(1+\beta^{-1}\right) x_{m'} \delta^2_{m'} \right) \\
&\leq c(p) \frac{\mathbb{E}\|\varepsilon_1\|^p}{\delta^p_{m'}} D_{m'} x_{m'}^{-\frac{p}{2}} \leq c'(p,\eta) \frac{\mathbb{E}\|\varepsilon_1\|^p}{\delta^p_{m'}} \frac{D_{m'}}{(L_{m'} D_{m'} + x)^{\frac{p}{2}}}. \qquad (6.25)
\end{aligned}$$

Gathering (6.22), (6.25) and (6.16) we get that

$$\begin{aligned}
\mathbb{P}\left(\mathcal{H}(\mathbf{f}) \geq \frac{x\delta^2_{m'}}{N(1-\alpha)} \right) &\leq \sum_{m' \in \mathcal{M}} P_{1,m'}(x) + \sum_{m' \in \mathcal{M}} P_{2,m'}(x) \\
&\leq \sum_{m' \in \mathcal{M}} c'(p,\eta) \frac{\mathbb{E}\|\varepsilon_1\|^p}{\delta^p_{m'}} \frac{D_{m'}}{(L_{m'} D_{m'} + x)^{\frac{p}{2}}} \\
&\quad + \sum_{m' \in \mathcal{M}} c''(p,\eta) \frac{\mathbb{E}\|\varepsilon_1\|^p}{\delta^p_{m'}} \frac{1}{(L_{m'} D_{m'} + x)^{\frac{p}{2}}}.
\end{aligned}$$

Since $\frac{1}{(1-\alpha)} = \left(1 + 2\eta^{-1}\right)$, then (6.3) holds:

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{H}(\mathbf{f}) \geq \left(1+2\eta^{-1}\right) \frac{x\delta^2_{m'}}{N} \right) \\
&\quad \leq \sum_{m' \in \mathcal{M}} \frac{\mathbb{E}\|\varepsilon_1\|^p}{\delta^p_{m'} (L_{m'} D_{m'} + x)^{\frac{p}{2}}} \max(D_{m'}, 1) \left(c'(p,\eta) + c''(p,\eta) \right) \\
&\quad = c(p,\eta) \frac{\mathbb{E}\|\varepsilon_1\|^p}{\delta^p_{m'}} \sum_{m' \in \mathcal{M}} \frac{D_{m'} \vee 1}{(L_{m'} D_{m'} + x)^{\frac{p}{2}}}.
\end{aligned}$$

It remains to choose β and δ for (6.14) to hold (we recall that $\alpha = \frac{2}{2+\eta}$). This is the case if $(2-\alpha)(1+\beta) = 1+\eta$ and $\left(2 - \alpha + \alpha^{-1}\right)\delta = 1$, therefore we take $\beta = \frac{\eta}{2}$ and $\delta = \left[1 + \frac{\eta}{2} + 2\frac{(1+\eta)}{(2+\eta)}\right]^{-1}$. □

6.3. Proof of the concentration inequality

Proof of Proposition (4.3)

Proof. Since $\widetilde{A}$ is nonnegative and symmetric there exists $A \in \mathbb{R}^{Nk \times Nk} \setminus \{\mathbf{0}\}$

such that $\widetilde{A} = A^\top A$. Then

$$\zeta^2(\varepsilon) = \varepsilon^\top \widetilde{A}\varepsilon = (A\varepsilon)^\top A\varepsilon = \|A\varepsilon\|^2 = \left[\sup_{\|u\|\leq 1} \langle A\varepsilon, \mathbf{u}\rangle\right]^2$$
$$= \left[\sup_{\|\mathbf{u}\|\leq 1} \left\langle \varepsilon, A^\top \mathbf{u}\right\rangle\right]^2 = \left[\sup_{\|\mathbf{u}\|\leq 1} \sum_{i=1}^N \left\langle \varepsilon_i, \left(A^\top \mathbf{u}\right)_i\right\rangle\right]^2$$
$$= \left[\sup_{\|\mathbf{u}\|\leq 1} \sum_{i=1}^N \left\langle \varepsilon_i, A_i^\top \mathbf{u}\right\rangle\right]^2 = \left[\sup_{\|\mathbf{u}\|\leq 1} \sum_{i=1}^N \sum_{j=1}^k \varepsilon_{ij}\left(A_i^\top \mathbf{u}\right)_j\right]^2$$

with $A = (A_1 \mid \ldots \mid A_N)$, where A_i is a $(Nk) \times k$ matrix.

Now take $\mathcal{G} = \{g_{\mathbf{u}} : g_{\mathbf{u}}(\mathbf{x}) = \sum_{i=1}^N \left\langle \mathbf{x}_i, A_i^\top \mathbf{u}\right\rangle = \sum_{i=1}^N \left\langle B_i\mathbf{x}, B_i A^\top \mathbf{u}\right\rangle,\ \mathbf{u}, \mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\top \in \mathbb{R}^{(Nk)},\ \|\mathbf{u}\| \leq 1\}$. Let $M_i = [\mathbf{0}, \ldots, \mathbf{0}, I_k, \mathbf{0}, \ldots, \mathbf{0}]^\top \in \mathbb{R}^{(Nk)\times(Nk)}$, where I_k is the i-th block of M_i, $B_i = [0, \ldots, 0, I_k, 0, \ldots 0] \in \mathbb{R}^{(Nk)\times(Nk)}$, $\varepsilon_i = B_i\,\varepsilon$ and $M_i\,\varepsilon = [\mathbf{0}, \ldots, \mathbf{0}, \varepsilon_i, \mathbf{0}, \ldots, \mathbf{0}]^\top$.

Then

$$\zeta(\varepsilon) = \sup_{\|\mathbf{u}\|\leq 1} \sum_{i=1}^N g_{\mathbf{u}}(M_i\varepsilon).$$

Now take $\mathbf{U}_i = M_i\,\varepsilon$, $\varepsilon \in \mathbb{R}^{(Nk)}$. Then for each positive number t and $p > 0$

$$\begin{aligned}
\mathbb{P}(\zeta(\varepsilon) \geq \mathbb{E}(\zeta(\varepsilon)) + t) &\leq \mathbb{P}(|\zeta(\varepsilon) - \mathbb{E}(\zeta(\varepsilon))| > t) \\
&\leq t^{-p}\mathbb{E}(|\zeta(\varepsilon) - \mathbb{E}(\zeta(\varepsilon))|^p) \text{ by Markov inequality} \\
&\leq c(p)\,t^{-p}\left\{\mathbb{E}\left(\max_{i=1,\ldots,N}\sup_{\|\mathbf{u}\|\leq 1}\left|\left\langle \varepsilon_i, A_i^\top \mathbf{u}\right\rangle\right|^p\right)\right. \\
&\quad \left. + \left[\mathbb{E}\left(\sup_{\|\mathbf{u}\|\leq 1}\sum_{i=1}^N \left(\left\langle \varepsilon_i, A_i^\top \mathbf{u}\right\rangle\right)^2\right)\right]^{p/2}\right\} \\
&= c(p)\,t^{-p}\left(\mathbb{E}_1 + \mathbb{E}_2^{p/2}\right). \qquad (6.26)
\end{aligned}$$

We start by bounding $\mathbb{E}_1$. For all $\mathbf{u}$ such that $\|\mathbf{u}\| \leq 1$ and $i \in \{1, \ldots, N\}$,

$$\left\|A_i^\top \mathbf{u}\right\|^2 \leq \left\|A^\top \mathbf{u}\right\|^2 \leq \rho^2(A),$$

where $\rho(M) = \sup_{x\neq 0}\frac{\|Mx\|}{\|x\|}$ for all matrix M. For $p \geq 2$ we have that $\left\|A_i^\top \mathbf{u}\right\|^p \leq \rho^{p-2}(A)\left\|A_i^\top \mathbf{u}\right\|^2$, then

$$\left|\left\langle \varepsilon_i, A_i^\top \mathbf{u}\right\rangle\right|^p \leq \left[\|\varepsilon_i\|\left\|A_i^\top \mathbf{u}\right\|\right]^p \leq \rho^{p-2}(A)\|\varepsilon_i\|^p\left\|A_i^\top \mathbf{u}\right\|^2.$$

Therefore

$$\mathbb{E}_1 \leq \rho^{p-2}(A)\,\mathbb{E}\left(\sup_{\|\mathbf{u}\|=1}\sum_{i=1}^N \|\varepsilon_i\|^p\left\|A_i^\top \mathbf{u}\right\|^2\right).$$

Since $\|\mathbf{u}\| \leq 1$, $\forall i = 1, \ldots, N$

$$\left\|A_i^\top \mathbf{u}\right\|^2 = \mathbf{u}^\top A_i A_i^\top \mathbf{u} \leq \rho\left(A_i A_i^\top\right) \leq \mathrm{Tr}\left(A_i A_i^\top\right),$$

then

$$\sum_{i=1}^N \left\|A_i^\top \mathbf{u}\right\|^2 \leq \sum_{i=1}^N \mathrm{Tr}\left(A_i A_i^\top\right) = \mathrm{Tr}\left(\sum_{i=1}^N A_i A_i^\top\right) = \mathrm{Tr}\left(\widetilde{A}\right).$$

Thus,

$$\mathbb{E}_1 \leq \rho^{p-2}(A)\,\mathrm{Tr}\left(\widetilde{A}\right)\mathbb{E}\left(\|\varepsilon_1\|^p\right). \tag{6.27}$$

We now bound $\mathbb{E}_2$ via a truncation argument. Since for all $\mathbf{u}$ such that $\|\mathbf{u}\| \leq 1$, $\left\|A^\top \mathbf{u}\right\|^2 \leq \rho^2(A)$, for any positive number c to be specified later we have that

$$\begin{aligned}
\mathbb{E}_2 &\leq \mathbb{E}\left(\sup_{\|\mathbf{u}\|\leq 1}\sum_{i=1}^N \|\varepsilon_i\|^2 \left\|A_i^\top \mathbf{u}\right\|^2 1_{\{\|\varepsilon_i\|\leq c\}}\right) \\
&\quad + \mathbb{E}\left(\sup_{\|\mathbf{u}\|\leq 1}\sum_{i=1}^N \|\varepsilon_i\|^2 \left\|A_i^\top \mathbf{u}\right\|^2 1_{\{\|\varepsilon_i\|> c\}}\right) \\
&\leq \mathbb{E}\left(c^2 \sup_{\|\mathbf{u}\|\leq 1}\sum_{i=1}^N \left\|A_i^\top \mathbf{u}\right\|^2 1_{\{\|\varepsilon_i\|\leq c\}}\right) \\
&\quad + \mathbb{E}\left(\sup_{\|\mathbf{u}\|\leq 1}\sum_{i=1}^N \|\varepsilon_i\|^2 \left\|A_i^\top \mathbf{u}\right\|^2 1_{\{\|\varepsilon_i\|> c\}}\right) \\
&\leq c^2\rho^2(A) + c^{2-p}\mathbb{E}\left(\sup_{\|\mathbf{u}\|\leq 1}\sum_{i=1}^N \|A_i\mathbf{u}\|^2 \|\varepsilon_i\|^p\right) \\
&\leq c^2\rho^2(A) + c^{2-p}\mathbb{E}\left(\|\varepsilon_1\|^p\right)\mathrm{Tr}\left(\widetilde{A}\right)
\end{aligned} \tag{6.28}$$

using the bound obtained for $\mathbb{E}_1$. It remains to take $c^p = \mathbb{E}\left(\|\varepsilon_1\|^p\right)\mathrm{Tr}\left(\widetilde{A}\right)/\rho^2(A)$ to get that:

$$\mathbb{E}_2 \leq c^2\rho^2(A) + c^2\rho^2(A) = 2c^2\rho^2(A),$$

therefore

$$\mathbb{E}_2^{p/2} \leq 2^{p/2} c^p \rho^p(A), \tag{6.29}$$

which implies that

$$2^{-p/2}\mathbb{E}_2^{p/2} \leq \mathbb{E}\left(\|\varepsilon_1\|^p\right)\mathrm{Tr}\left(\widetilde{A}\right)\rho^{p-2}(A).$$

We straightforwardly derive from (6.26) that

$$\mathbb{P}\left(\zeta^2(\varepsilon) \geq \left[\mathbb{E}\left(\zeta(\varepsilon)\right)\right]^2 + 2\mathbb{E}\left(\zeta(\varepsilon)\right)t + t^2\right) \leq c(p)\,t^{-p}\left(\mathbb{E}_1 + \mathbb{E}_2^{p/2}\right).$$

Since $\left[\mathbb{E}\left(\zeta\left(\varepsilon\right)\right)\right]^2 \leq \mathbb{E}\left(\zeta^2\left(\varepsilon\right)\right)$, (6.27) and (6.29) imply that

$$\begin{aligned}
&\mathbb{P}\left(\zeta^2\left(\varepsilon\right) \geq \mathbb{E}\left(\zeta^2\left(\varepsilon\right)\right) + 2\sqrt{\mathbb{E}\left(\zeta^2\left(\varepsilon\right)\right)t^2} + t^2\right) \\
&\quad \leq c\left(p\right)t^{-p}\left(\mathbb{E}_1 + \mathbb{E}_2^{p/2}\right) \\
&\quad \leq c\left(p\right)t^{-p}\left(\rho^{p-2}\left(A\right)\mathrm{Tr}\left(\widetilde{A}\right)\mathbb{E}\left(\left\|\varepsilon_1\right\|^p\right) + 2^{p/2}\mathbb{E}\left(\left\|\varepsilon_1\right\|^p\right)\mathrm{Tr}\left(\widetilde{A}\right)\rho^{p-2}\left(A\right)\right) \\
&\quad \leq c'\left(p\right)t^{-p}\rho^{p-2}\left(A\right)\mathrm{Tr}\left(\widetilde{A}\right)\mathbb{E}\left(\left\|\varepsilon_1\right\|^p\right), \qquad (6.30)
\end{aligned}$$

for all $t > 0$. Moreover

$$\begin{aligned}
\mathbb{E}\left(\zeta^2\left(\varepsilon\right)\right) &= \mathbb{E}\left(\varepsilon^\top\widetilde{A}\varepsilon\right) = \mathbb{E}\left(\mathrm{Tr}\left(\varepsilon^\top\widetilde{A}\varepsilon\right)\right) = \mathbb{E}\left(\mathrm{Tr}\left(\widetilde{A}\varepsilon\varepsilon^\top\right)\right) \\
&= \mathrm{Tr}\left(\widetilde{A}\mathbb{E}\left(\varepsilon\varepsilon^\top\right)\right) = \mathrm{Tr}\left(\widetilde{A}\left(I_N \otimes \Phi\right)\right) = \delta^2\mathrm{Tr}\left(\widetilde{A}\right) \qquad (6.31)
\end{aligned}$$

Using (6.31), take $t^2 = \rho\left(\widetilde{A}\right)\delta^2 x > 0$ in (6.30) to get that

$$\begin{aligned}
&\mathbb{P}\left(\zeta^2\left(\varepsilon\right) \geq \delta^2\mathrm{Tr}\left(\widetilde{A}\right) + 2\sqrt{\delta^2\mathrm{Tr}\left(\widetilde{A}\right)\rho\left(\widetilde{A}\right)\delta^2 x} + \rho\left(\widetilde{A}\right)\delta^2 x\right) \\
&\quad \leq c'\left(p\right)\rho^{-p/2}\left(\widetilde{A}\right)\delta^{-p/2}x^{-p/2}\rho^{p-2}\left(A\right)\mathrm{Tr}\left(\widetilde{A}\right)\mathbb{E}\left(\left\|\varepsilon_1\right\|^p\right).
\end{aligned}$$

Since $\rho\left(\widetilde{A}\right) = \rho^2\left(A\right)$ (with the Euclidean norm) the desired result follows:

$$\begin{aligned}
&\mathbb{P}\left(\zeta^2\left(\varepsilon\right) \geq \delta^2\mathrm{Tr}\left(\widetilde{A}\right) + 2\delta^2\sqrt{\rho\left(\widetilde{A}\right)\mathrm{Tr}\left(\widetilde{A}\right)x} + \delta^2\rho\left(\widetilde{A}\right)x\right) \\
&\quad \leq c'\left(p\right)\frac{\mathbb{E}\left\|\varepsilon_1\right\|^p}{\delta^p}\frac{\mathrm{Tr}\left(\widetilde{A}\right)}{\rho\left(\widetilde{A}\right)x^{p/2}}. \qquad (6.32)
\end{aligned}$$

□

References

[1] Adler, Robert J. *An introduction to continuity, extrema, and related topics for general Gaussian processes*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 12. Institute of Mathematical Statistics, Hayward, CA, 1990. MR1088478

[2] Baraud, Yannick. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493, 2000. MR1777129

[3] Baraud, Yannick. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic), 2002. MR1918295

[4] Bickel, P.J. and Levina, E. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36:199–227, 2008. MR2387969

[5] Diaz-Frances E., Biscay, R. J. and Rodriguez, L. M. Cross-validation of covariance structures using the frobenius matrix distance as a discrepancy function. *Journal of Statistical Computation and Simulation*, 1997.

[6] Biscay, R., Jimenez, J.C. and Gonzalez, A. Smooth approximation of nonnegative definite kernels. In *Approximation and optimization in the Caribbean, II (Havana, 1993)*, volume 8 of *Approx. Optim.*, pages 114–128. Lang, Frankfurt am Main, 1995. MR1368168

[7] Comte, Fabienne. Adaptive estimation of the spectrum of a stationary Gaussian sequence. *Bernoulli*, 7(2):267–298, 2001. MR1828506

[8] Cressie, Noel A.C. *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. MR1127423

[9] Perrin, O., Elogne, S.N. and Thomas-Agnan, C. Non parametric estimation of smooth stationary covariance functions by interpolation methods. *Phd*, 2003.

[10] Engl, H., Hanke, M. and Neubauer, A. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996. MR1408680

[11] Fan, Y., Fan, J. and Lv, J. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147:186–197, 2008. MR2472991

[12] Gendre, Xavier. Simultaneous estimation of the mean and the variance in heteroscedastic Gaussian regression. *Electron. J. Stat.*, 2:1345–1372, 2008. MR2471290

[13] Journel, A.G. Kriging in terms of projections. *J. Internat. Assoc. Mathematical Geol.*, 9(6):563–586, 1977. MR0456314

[14] Kollo, Tõnu and von Rosen, Dietrich. *Advanced multivariate statistics with matrices*, volume 579 of *Mathematics and Its Applications (New York)*. Springer, Dordrecht, 2005. MR2162145

[15] Levina, Elizaveta, Rothman, Adam and Ji Zhu. Sparse estimation of large covariance matrices via a nested Lasso penalty. *Ann. Appl. Stat.*, 2(1):245–263, 2008. MR2415602

[16] Loubes, J.-M. and Ludena, C. Adaptive complexity regularization for inverse problems. *Electronic Journal Of Statistics*, 2:661–677, 2008. MR2426106

[17] Loubes, J.-M. and Ludena, C. Penalized estimators for non linear inverse problems. *ESAIM Probab. Statist.*, 14:173–191, 2010.

[18] Lütkepohl, H. *Handbook of matrices*. John Wiley & Sons Ltd., Chichester, 1996. MR1433592

[19] Nychka D.W., Matsuo, T. and Paul, D. Nonstationary covariance modeling for incomplete data: smoothed Monte-Carlo approach. *preprint*, 2008.

[20] Ramsey J.O. and Silverman B.W. *Functional Data Analysis*. Springer: NY, 2005. MR2168993

[21] Raj Rao, N., Mingo, James A., Speicher, Roland and Edelman, Alan. Statistical eigen-inference from large Wishart matrices. *Ann. Statist.*, 36(6):2850–2885, 2008. MR2485015

[22] Schäfer, Juliane and Strimmer, Korbinian. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, 4:Art. 32, 28 pp. (electronic), 2005. MR2183942

[23] Stein, Michael L. *Interpolation of spatial data. Some theory for kriging.* Springer Series in Statistics. New York, NY: Springer. xvii, 247 p., 1999. MR1697409