Group Lasso estimation of high-dimensional covariance matrices

Jérémie Bigot[1,2], Rolando J. Biscay[4], Jean-Michel Loubes[1] and Lilian Muñiz-Alvarez[1,3]

IMT, Université Paul Sabatier, Toulouse, France[1]
Center for Mathematical Modelling, Universidad de Chile, Santiago, Chile[2]
Facultad de Matemática y Computación, Universidad de La Habana, Cuba[3]
DEUV-CIMFAV, Facultad de Ciencias, Universidad de Valparaiso, Chile[4]

October 24, 2011

Abstract

In this paper, we consider the Group Lasso estimator of the covariance matrix of a stochastic process corrupted by an additive noise. We propose to estimate the covariance matrix in a high-dimensional setting under the assumption that the process has a sparse representation in a large dictionary of basis functions. Using a matrix regression model, we propose a new methodology for high-dimensional covariance matrix estimation based on empirical contrast regularization by a group Lasso penalty. Using such a penalty, the method selects a sparse set of basis functions in the dictionary used to approximate the process, leading to an approximation of the covariance matrix into a low dimensional space. Consistency of the estimator is studied in Frobenius and operator norms and an application to sparse PCA is proposed.

Keywords: Group Lasso; ℓ^1 penalty; high-dimensional covariance estimation; basis expansion; sparsity; oracle inequality; sparse PCA.
Subject Class. MSC-2000 : 62G05, 62H25

Acknowledgments: this work was supported in part by Egide, under the Program of Eiffel excellency Phd grants, as well as by the BDI CNRS grant. J. Bigot would like to thank the Center for Mathematical Modeling and the CNRS for financial support and excellent hospitality while visiting Santiago where part of this work was carried out.

1 Introduction

Let $\mathbb{T}$ be some subset of $\mathbb{R}^p$, $p \in \mathbb{N}$, and let $X = (X(t))_{t\in\mathbb{T}}$ be a stochastic process with values in $\mathbb{R}$. Assume that X has zero mean $\mathbb{E}(X(t)) = 0$ for all $t \in \mathbb{T}$, and finite covariance $\sigma(s,t) = \mathbb{E}(X(s)X(t))$ for all $s,t \in \mathbb{T}$. Let $t_1,\ldots,t_n$ be fixed points in $\mathbb{T}$ (deterministic design), $X_1,...,X_N$ independent copies of the process X, and suppose that we observe the noisy processes

$$\widetilde{X}_i(t_j) = X_i(t_j) + \mathcal{E}_i(t_j) \text{ for } i = 1,...,N,\ j = 1,...,n, \tag{1.1}$$

where $\mathcal{E}_1, ..., \mathcal{E}_N$ are independent copies of a second order Gaussian process $\mathcal{E}$ with zero mean and independent of X, which represent an additive source of noise in the measurements. Based on the noisy observations (1.1), an important problem in statistics is to construct an estimator of the covariance matrix $\mathbf{\Sigma} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^\top\right)$ of the process X at the design points, where $\mathbf{X} = (X(t_1), ..., X(t_n))^\top$. This problem is a fundamental issue in many applications, ranging from geostatistics, financial series or epidemiology for instance (see [Stein, 1999], [Journel, 1977] or [Cressie, 1993, Wikle and Cressie, 1999] for general references and applications). Estimating such a covariance matrix has also important applications in dimension reduction by principal component analysis (PCA) or classification by linear or quadratic discriminant analysis (LDA and QDA).

In [Bigot et al., 2010], using N independent copies of the process X, we have proposed to construct an estimator of the covariance matrix $\mathbf{\Sigma}$ by expanding the process X into a dictionary of basis functions. The method in [Bigot et al., 2010] is based on model selection techniques by empirical contrast minimization in a suitable matrix regression model. This new approach to covariance estimation is well adapted to the case of low-dimensional covariance estimation when the number of replicates N of the process is larger than the number of observations points n. However, many application areas are currently dealing with the problem of estimating a covariance matrix when the number of observations at hand is small when compared to the number of parameters to estimate. Examples include biomedical imaging, proteomic/genomic data, signal processing in neurosciences and many others. This issue corresponds to the problem of covariance estimation for high-dimensional data. This problem is challenging since, in a high-dimensional setting (when $n >> N$ or $n \sim N$), it is well known that the sample covariance matrices

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i\mathbf{X}_i^\top \in \mathbb{R}^{n\times n}, \text{ where } \mathbf{X}_i = (X_i(t_1), ..., X_i(t_n))^\top, i = 1, \ldots, N$$

and

$$\widetilde{\mathbf{S}} = \frac{1}{N}\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top \in \mathbb{R}^{n\times n}, \text{ where } \widetilde{\mathbf{X}}_i = \left(\widetilde{X}_i(t_1), ..., \widetilde{X}_i(t_n)\right)^\top, i = 1, \ldots, N$$

behave poorly, and are not consistent estimators of $\mathbf{\Sigma}$. For example, suppose that the $\mathbf{X}_i$'s are independent and identically distributed (i.i.d.) random vectors in $\mathbb{R}^n$ drawn from a multivariate Gaussian distribution. Then, when $\frac{n}{N} \to c > 0$ as $n, N \to +\infty$, neither the eigenvalues nor the eigenvectors of the sample covariance matrix $\mathbf{S}$ are consistent estimators of the eigenvalues and eigenvectors of $\mathbf{\Sigma}$ (see [Johnstone, 2001]). This topic has thus recently received a lot of attention in the statistical literature. To achieve consistency, recently developed methods for high-dimensional covariance estimation impose sparsity restrictions on the matrix $\mathbf{\Sigma}$. Such restrictions imply that the true (but unknown) dimension of the model is much lower than the number $\frac{n(n+1)}{2}$ of parameters of an unconstrained covariance matrix. Under various sparsity assumptions, different regularizing methods of the empirical covariance matrix have been proposed. Estimators based on thresholding or banding the entries of the empirical covariance matrix have been studied in [Bickel and Levina, 2008a] and [Bickel and Levina, 2008b]. Thresholding the components of the empirical covariance matrix has also been proposed by [El Karoui, 2008] and the consistency

of such estimates is studied using tools from random matrix theory. [Fan et al., 2008] impose sparsity on the covariance via a factor model which is appropriate in financial applications. [Levina et al., 2008] and [Rothman et al., 2008] propose regularization techniques with a Lasso penalty to estimate the covariance matrix or its inverse. More general penalties have been studied in [Lam and Fan, 2009]. Another approach is to impose sparsity on the eigenvectors of the covariance matrix which leads to sparse PCA. [Zou et al., 2006] use a Lasso penalty to achieve sparse representation in PCA, [d'Aspremont et al., 2008] study properties of sparse principal components by convex programming, while [Johnstone and Lu, 2009] propose a PCA regularization by expanding the empirical eigenvectors in a sparse basis and then apply a thresholding step.

In this paper, we propose to estimate $\mathbf{\Sigma}$ in a high-dimensional setting by using the assumption that the process X has a sparse representation in a large dictionary of basis functions. Using a matrix regression model as in [Bigot et al., 2010], we propose a new methodology for high-dimensional covariance matrix estimation based on empirical contrast regularization by a group Lasso penalty. Using such a penalty, the method selects a sparse set of basis functions in the dictionary used to approximate the process X. This leads to an approximation of the covariance matrix $\mathbf{\Sigma}$ into a low dimensional space, and thus to a new method of dimension reduction for high-dimensional data. Group Lasso estimators have been studied in the standard linear model and in multiple kernel learning to impose a group-sparsity structure on the parameters to recover (see [Nardi and Rinaldo, 2008], [Bach, 2008] and references therein). However, to the best of our knowledge, it has not been used for the estimation of covariance matrices using a functional approximation of the process X.

The rest of the paper is organized as follows. In Section 2, we describe a matrix regression model for covariance estimation, and we define our estimator by group Lasso regularization. The consistency of such a procedure is investigated in Section 3 using oracle inequalities and a non-asymptotic point of view by holding fixed the number of replicates N and observation points n. Consistency of the estimator is studied in Frobenius and operator norms. Various results existing in matrix theory show that convergence in operator norm implies convergence of the eigenvectors and eigenvalues (e.g. through the use of the $\sin(\theta)$ theorems in [Davis and Kahan, 1970]). Consistency in operator norm is thus well suited for PCA applications. Numerical experiments are given in Section 4, and an application to sparse PCA is proposed. A technical Appendix contains all the proofs.

2 Model and definition of the estimator

To impose sparsity restrictions on the covariance matrix $\mathbf{\Sigma}$, our approach is based on an approximation of the process in a finite dictionary of (not necessarily orthogonal) basis functions $g_m : \mathbb{T} \to \mathbb{R}$ for $m = 1, ..., M$. Suppose that

$$X(t) \approx \sum_{m=1}^{M} a_m g_m(t), \tag{2.1}$$

where a_m, $m = 1, ..., M$ are real valued random variables, and that for each trajectory X_i

$$X_i(t_j) \approx \sum_{m=1}^{M} a_{i,m} g_m(t_j). \tag{2.2}$$

The notation $\approx$ means that the process X can be well approximated into the dictionary. A precise meaning of this will be discussed later on. Then (2.2) can be written in matrix notation as:

$$\mathbf{X}_i \approx \mathbf{G}\mathbf{a}_i, \ i = 1, ..., N \tag{2.3}$$

where $\mathbf{G}$ is the $n \times M$ matrix with entries

$$\mathbf{G}_{jm} = g_m(t_j) \text{ for } 1 \leq j \leq n \text{ and } 1 \leq m \leq M,$$

and $\mathbf{a}_i$ is the $M \times 1$ random vector of components $a_{i,m}$, with $1 \leq m \leq M$.

Recall that we want to estimate the covariance matrix $\mathbf{\Sigma} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^\top\right)$ from the noisy observations (1.1). Since $\mathbf{X} \approx \mathbf{G}\mathbf{a}$ with $\mathbf{a} = (a_m)_{1 \leq m \leq M}$ with a_m as in (2.1), it follows that

$$\mathbf{\Sigma} \approx \mathbb{E}\left(\mathbf{G}\mathbf{a}(\mathbf{G}\mathbf{a})^\top\right) = \mathbb{E}\left(\mathbf{G}\mathbf{a}\mathbf{a}^\top\mathbf{G}^\top\right) = \mathbf{G}\mathbf{\Psi}^*\mathbf{G}^\top \text{ with } \mathbf{\Psi}^* = \mathbb{E}\left(\mathbf{a}\mathbf{a}^\top\right).$$

Given the noisy observations $\widetilde{\mathbf{X}}_i$ as in (1.1) with $i = 1, ..., N$, consider the following matrix regression model

$$\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top = \mathbf{\Sigma} + \mathbf{U}_i + \mathbf{W}_i \ i = 1, \ldots, N, \tag{2.4}$$

where $\mathbf{U}_i = \mathbf{X}_i\mathbf{X}_i^\top - \mathbf{\Sigma}$ are i.i.d centered matrix errors, and

$$\mathbf{W}_i = \mathcal{E}_i\mathcal{E}_i^\top \in \mathbb{R}^{n \times n} \text{ where } \mathcal{E}_i = (\mathcal{E}_i(t_1), ..., \mathcal{E}_i(t_n))^\top, i = 1, \ldots, N.$$

The size M of the dictionary can be very large, but it is expected that the process X has a sparse expansion in this basis, meaning that, in approximation (2.1), many of the random coefficients a_m are close to zero. We are interested in obtaining an estimate of the covariance $\mathbf{\Sigma}$ in the form $\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top$ such that $\widehat{\mathbf{\Psi}}$ is a symmetric $M \times M$ matrix with many zero rows (and so, by symmetry, many corresponding zero columns). Note that setting the k-th row of $\widehat{\mathbf{\Psi}}$ to $\mathbf{0} \in \mathbb{R}^M$ means to remove the function g_k from the set of basis functions $(g_m)_{1 \leq m \leq M}$ in the function expansion associated to $\mathbf{G}$.

Let us now explain how to select a sparse set of rows/columns in the matrix $\widehat{\mathbf{\Psi}}$. For this, we use a group Lasso approach to threshold some rows/columns of $\widehat{\mathbf{\Psi}}$ which corresponds to removing some basis functions in the approximation of the process X. For two $p \times p$ matrices $\mathbf{A}, \mathbf{B}$ define the inner product $\langle \mathbf{A}, \mathbf{B} \rangle_F := tr\left(\mathbf{A}^\top\mathbf{B}\right)$ and the associated Frobenius norm $\|\mathbf{A}\|_F^2 := tr\left(\mathbf{A}^\top\mathbf{A}\right)$. Let $\mathcal{S}_M$ denote the set of $M \times M$ symmetric matrices with real entries. We define the group Lasso estimator of the covariance matrix $\mathbf{\Sigma}$ by

$$\widehat{\mathbf{\Sigma}}_\lambda = \mathbf{G}\widehat{\mathbf{\Psi}}_\lambda\mathbf{G}^\top \in \mathbb{R}^{n \times n}, \tag{2.5}$$

where $\widehat{\mathbf{\Psi}}_\lambda$ is the solution of the following optimization problem:

$$\widehat{\mathbf{\Psi}}_\lambda = \underset{\mathbf{\Psi} \in \mathcal{S}_M}{\operatorname{argmin}} \left\{ \frac{1}{N}\sum_{i=1}^{N} \left\| \widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top \right\|_F^2 + 2\lambda \sum_{k=1}^{M} \gamma_k \sqrt{\sum_{m=1}^{M} \Psi_{mk}^2} \right\}, \tag{2.6}$$

where $\mathbf{\Psi} = (\Psi_{mk})_{1\leq m,k\leq M} \in \mathbb{R}^{M\times M}$, λ is a positive number and γ_k are some weights whose values will be discuss later on. In (2.6), the penalty term imposes to give preference to solutions with components $\mathbf{\Psi}_k = \mathbf{0}$, where $(\mathbf{\Psi}_k)_{1\leq k\leq M}$ denotes the columns of $\mathbf{\Psi}$. Recall that $\widetilde{\mathbf{S}} = \frac{1}{N}\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top$ denotes the sample covariance matrix from the noisy observations (1.1). It can be checked that minimizing the criterion (2.6) is equivalent to

$$\widehat{\mathbf{\Psi}}_\lambda = \underset{\mathbf{\Psi}\in\mathcal{S}_M}{\operatorname{argmin}} \left\{ \left\| \widetilde{\mathbf{S}} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top \right\|_F^2 + 2\lambda \sum_{k=1}^{M} \gamma_k \sqrt{\sum_{m=1}^{M} \Psi_{mk}^2} \right\}. \tag{2.7}$$

Thus $\widehat{\mathbf{\Psi}}_\lambda \in \mathbb{R}^{M\times M}$ can be interpreted as a group Lasso estimator of $\mathbf{\Sigma}$ in the following matrix regression model

$$\widetilde{\mathbf{S}} = \mathbf{\Sigma} + \mathbf{U} + \mathbf{W} \approx \mathbf{G}\mathbf{\Psi}^*\mathbf{G}^\top + \mathbf{U} + \mathbf{W}, \tag{2.8}$$

where $\mathbf{U} \in \mathbb{R}^{n\times n}$ is a centered error matrix given by $\mathbf{U} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{U}_i$ and $\mathbf{W} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{W}_i$. In the above regression model (2.8), there are two errors terms of a different nature. The term $\mathbf{W}$ corresponds to the additive Gaussian errors $\mathcal{E}_1, ..., \mathcal{E}_N$ in model (1.1), while the term $\mathbf{U} = \mathbf{S} - \mathbf{\Sigma}$ represents the difference between the (unobserved) sample covariance matrix $\mathbf{S}$ and the matrix $\mathbf{\Sigma}$ that we want to estimate.

This approach can be interpreted as a thresholding procedure of the entries of an empirical matrix. To see this, consider the simple case where $M = n$ and the basis functions and observations points are chosen such that the matrix $\mathbf{G}$ is orthogonal. Let $\mathbf{Y} = \mathbf{G}^\top\widetilde{\mathbf{S}}\mathbf{G}$ be a transformation of the empirical covariance matrix $\widetilde{\mathbf{S}}$. In the orthogonal case, the following proposition shows that the group Lasso estimator $\widehat{\mathbf{\Psi}}_\lambda$ defined by (2.7) consists in thresholding the columns/rows of $\mathbf{Y}$ whose ℓ_2-norm is too small, and in multiplying the other columns/rows by weights between 0 and 1. Hence, the group Lasso estimate (2.7) can be interpreted as covariance estimation by soft-thresholding the columns/rows of $\mathbf{Y}$.

Proposition 1 *Suppose that $M = n$ and that $\mathbf{G}^\top\mathbf{G} = \mathbf{I}_n$ where $\mathbf{I}_n$ denotes the identity matrix of size $n \times n$. Let $\mathbf{Y} = \mathbf{G}^\top\widetilde{\mathbf{S}}\mathbf{G}$. Then, the group Lasso estimator $\widehat{\mathbf{\Psi}}_\lambda$ defined by (2.7) is the $n \times n$ symmetric matrix whose entries are given by*

$$\left(\widehat{\mathbf{\Psi}}_\lambda\right)_{mk} = \begin{cases} 0 & \text{if} \quad \sqrt{\sum_{j=1}^{M} \mathbf{Y}_{jk}^2} \leq \lambda\gamma_k, \\ Y_{mk}\left(1 - \frac{\lambda\gamma_k}{\sqrt{\sum_{j=1}^{M} \mathbf{Y}_{mk}^2}}\right) & \text{if} \quad \sqrt{\sum_{j=1}^{M} \mathbf{Y}_{jk}^2} > \lambda\gamma_k, \end{cases} \tag{2.9}$$

for $1 \leq k, m \leq M$.

3 Consistency of the group Lasso estimator

3.1 Notations and main assumptions

Let us begin by some definitions. For a symmetric $p \times p$ matrix $\mathbf{A}$ with real entries, $\rho_{min}(\mathbf{A})$ denotes the smallest eigenvalue of $\mathbf{A}$, and $\rho_{max}(\mathbf{A})$ denotes the largest eigenvalue of $\mathbf{A}$. For

$\beta \in \mathbb{R}^q$, $\|\beta\|_{\ell_2}$ denotes the usual Euclidean norm of β. For $p \times q$ matrix $\mathbf{A}$ with real entries, $\|\mathbf{A}\|_2 = \sup_{\beta \in \mathbb{R}^q,\, \beta \neq 0} \frac{\|\mathbf{A}\beta\|_{\ell_2}}{\|\beta\|_{\ell_2}}$ denotes the operator norm of $\mathbf{A}$. Recall that if $\mathbf{A}$ is a non negative definite matrix with $p = q$ then $\|\mathbf{A}\|_2 = \rho_{max}(\mathbf{A})$.

Let $\mathbf{\Psi} \in \mathcal{S}_M$ and β a vector in $\mathbb{R}^M$. For a subset $J \subset \{1, \ldots, M\}$ of indices of cardinality $|J|$, then β_J is the vector in $\mathbb{R}^M$ that has the same coordinates as β on J and zeros coordinates on the complement J^c of J. The $n \times |J|$ matrix obtained by removing the columns of $\mathbf{G}$ whose indices are not in J is denoted by $\mathbf{G}_J$. The sparsity of $\mathbf{\Psi}$ is defined as its number of non-zero columns (and thus by symmetry non-zero rows) namely

Definition 1 *For $\mathbf{\Psi} \in \mathcal{S}_M$, the sparsity of $\mathbf{\Psi}$ is*

$$\mathcal{M}(\mathbf{\Psi}) = \#\{k : \mathbf{\Psi}_k \neq \mathbf{0}\}.$$

Then, let us introduce the following quantities that control the minimal eigenvalues of submatrices of small size extracted from the matrix $\mathbf{G}^\top \mathbf{G}$, and the correlations between the columns of $\mathbf{G}$:

Definition 2 *Let $0 < s \leq M$. Then,*

$$\rho_{\min}(s) := \inf_{\substack{J \subset \{1,\ldots,M\} \\ |J| \leq s}} \left(\frac{\beta_J^\top \mathbf{G}^\top \mathbf{G} \beta_J}{\|\beta_J\|_{\ell_2}^2} \right) = \inf_{\substack{J \subset \{1,\ldots,M\} \\ |J| \leq s}} \rho_{\min}\left(\mathbf{G}_J^\top \mathbf{G}_J \right).$$

Definition 3 *The mutual coherence $\theta(\mathbf{G})$ of the columns $\mathbf{G}_k$, $k = 1, \ldots, M$ of $\mathbf{G}$ is defined as*

$$\theta(\mathbf{G}) := \max\left\{ \left| \mathbf{G}_{k'}^\top \mathbf{G}_k \right|,\ k \neq k',\ 1 \leq k, k' \leq M \right\},$$

and let

$$\mathbf{G}_{\max}^2 := \max\left\{ \|\mathbf{G}_k\|_{\ell_2}^2,\ 1 \leq k \leq M \right\}.$$

To derive oracle inequalities showing the consistency of the group Lasso estimator $\widehat{\mathbf{\Psi}}_\lambda$ the correlations between the columns of $\mathbf{G}$ (measured by $\theta(\mathbf{G})$) should not be too large when compared to the minimal eigenvalues of small matrices extracted from $\mathbf{G}^\top \mathbf{G}$, which is formulated in the following assumption:

Assumption 1 *Let $c_0 > 0$ be some constant and $0 < s \leq M$. Then*

$$\theta(\mathbf{G}) < \frac{\rho_{\min}(s)^2}{c_0 \rho_{\max}(\mathbf{G}^\top \mathbf{G}) s}.$$

Assumption 1 is inspired by recent results in [Bickel et al., 2009] on the consistency of Lasso estimators in the standard nonparametric regression model using a large dictionary of basis functions. In [Bickel et al., 2009], a general condition called *restricted eigenvalue assumption* is introduced to control the minimal eigenvalues of the Gram matrix associated to the dictionary over sets of sparse vectors. In the setting of nonparametric regression, a condition similar to

Assumption 1 is given in [Bickel et al., 2009] as an example for which the restricted eigenvalue assumption holds.

Let us give some examples for which Assumption 1 is satisfied. If $M \leq n$ and the design points are chosen such that the columns of the matrix $\mathbf{G}$ are orthonormal vectors in $\mathbb{R}^n$, then for any $0 < s \leq M$ one has that $\rho_{\min}(s) = 1$ and $\theta(\mathbf{G}) = 0$ and thus Assumption 1 holds for any value of c_0 and s.

Now, suppose that the columns of $\mathbf{G}$ are normalized to one, i.e $\|\mathbf{G}_k\|_{\ell_2} = 1$, $k = 1, \ldots, M$ implying that $\mathbf{G}_{\max} = 1$. Let $\beta \in \mathbb{R}^M$. Then, for any $J \subset \{1, \ldots, M\}$ with $|J| \leq s \leq \min(n, M)$

$$\beta_J^\top \mathbf{G}^\top \mathbf{G} \beta_J \geq \|\beta_J\|_{\ell_2}^2 - \theta(\mathbf{G}) s \|\beta_J\|_{\ell_2}^2,$$

which implies that

$$\rho_{\min}(s) \geq 1 - \theta(\mathbf{G}) s.$$

Therefore, if $(1 - \theta(\mathbf{G})(s-1))^2 > c_0 \theta(\mathbf{G}) \rho_{\max}(\mathbf{G}^\top \mathbf{G}) s$, then Assumption 1 is satisfied.

Let us now specify the law of the stochastic process X. For this, recall that for a real-valued random variable Z, the ψ_α Orlicz norm of Z is

$$\|Z\|_{\psi_\alpha} := \inf \left\{ C > 0 \, ; \, \mathbb{E} \exp\left(\frac{|Z|^\alpha}{C^\alpha} \right) \leq 2 \right\}.$$

Such Orlicz norms are useful to characterize the tail behavior of random variables. Indeed, if $\|Z\|_{\psi_\alpha} < +\infty$ then this is equivalent to assuming that there exists two constants $K_1, K_2 > 0$ such that for all $x > 0$

$$\mathbb{P}\left(|Z| \geq x \right) \leq K_1 \exp\left(-\frac{x^\alpha}{K_2^\alpha} \right),$$

(see e.g. [Mendelson and Pajor, 2006] for more details on Orlicz norms of random variables) . Therefore, if $\|Z\|_{\psi_2} < +\infty$ then Z is said to have a sub-Gaussian behavior and if $\|Z\|_{\psi_1} < +\infty$ then Z is said to have a sub-Exponential behavior. In the next sections, oracle inequalities for the group Lasso estimator will be derived under the following assumption on X:

Assumption 2 *The random vector* $\mathbf{X} = \left(X\left(t_1\right), ..., X\left(t_n\right) \right)^\top \in \mathbb{R}^n$ *is such that*

(A1) *There exists* $\rho\left(\mathbf{\Sigma}\right) > 0$ *such that, for all vector* $\beta \in \mathbb{R}^n$ *with* $\|\beta\|_{\ell_2} = 1$, *then* $\left(\mathbb{E} |\mathbf{X}^\top \beta|^4 \right)^{1/4} < \rho\left(\mathbf{\Sigma}\right)$.

(A2) *Set* $Z = \|\mathbf{X}\|_{\ell_2}$. *There exists* $\alpha \geq 1$ *such that* $\|Z\|_{\psi_\alpha} < +\infty$.

Note that **(A1)** implies that $\|\mathbf{\Sigma}\|_2 \leq \rho\left(\mathbf{\Sigma}\right)^2$. Indeed, one has that

$$\begin{aligned} \|\mathbf{\Sigma}\|_2 = \rho_{\max}(\mathbf{\Sigma}) &= \sup_{\beta \in \mathbb{R}^n, \, \|\beta\|_{\ell_2} = 1} \beta^\top \mathbf{\Sigma} \beta = \sup_{\beta \in \mathbb{R}^n, \, \|\beta\|_{\ell_2} = 1} \mathbb{E} \beta^\top \mathbf{X} \mathbf{X}^\top \beta \\ &= \sup_{\beta \in \mathbb{R}^n, \, \|\beta\|_{\ell_2} = 1} \mathbb{E} |\beta^\top \mathbf{X}|^2 \leq \sup_{\beta \in \mathbb{R}^n, \, \|\beta\|_{\ell_2} = 1} \sqrt{\mathbb{E} |\beta^\top \mathbf{X}|^4} \leq \rho^2\left(\mathbf{\Sigma}\right). \end{aligned}$$

When X is a Gaussian process, it follows that for any $\beta \in \mathbb{R}^n$ with $\|\beta\|_{\ell_2} = 1$ then $\left(\mathbb{E}|\mathbf{X}^\top \beta|^4\right)^{1/4} = 3^{1/4}\left(\beta^\top \mathbf{\Sigma}\beta\right)^{1/2}$ since $\mathbf{X}^\top\beta \sim N(0, \beta^\top\mathbf{\Sigma}\beta)$. Therefore, under the assumption that X is a Gaussian process, Assumption **(A1)** holds with $\rho(\mathbf{\Sigma}) = 3^{1/4}\|\mathbf{\Sigma}\|_2^{1/2}$.

Assumption **(A2)** requires that $\|Z\|_{\psi_\alpha} < +\infty$, where $Z = \|\mathbf{X}\|_{\ell_2}$. The following proposition provides some examples where such an assumption holds.

Proposition 2 *Let $Z = \|\mathbf{X}\|_{\ell_2} = \left(\sum_{i=1}^n |X(t_i)|^2\right)^{1/2}$. Then*

- *If X is a Gaussian process*
$$\|Z\|_{\psi_2} < \sqrt{8/3}\sqrt{tr(\mathbf{\Sigma})}.$$
- *If the random process X is such that $\|Z\|_{\psi_2} < +\infty$, and there exists a constant C_1 such that $\|\mathbf{\Sigma}_{ii}^{-1/2}|X(t_i)|\|_{\psi_2} \leq C_1$ for all $i = 1, \ldots, n$, then*
$$\|Z\|_{\psi_2} < C_1\sqrt{tr(\mathbf{\Sigma})}.$$
- *If X is a bounded process, meaning that there exists a constant $R > 0$ such that for all $t \in \mathbb{T}$, $|X(t)| \leq R$, then for any $\alpha \geq 1$,*
$$\|Z\|_{\psi_\alpha} \leq \sqrt{n}R(\log 2)^{-1/\alpha}.$$

Assumption 2 will be used to control the deviation in operator norm between the sample covariance matrix $\mathbf{S}$ and the true covariance matrix $\mathbf{\Sigma}$ in the sense of the following proposition whose proof follows from Theorem 2.1 in [Mendelson and Pajor, 2006].

Proposition 3 *Let $X_1, ..., X_N$ be independent copies of the stochastic process X, let $Z = \|\mathbf{X}\|_{\ell_2}$ and $\mathbf{X}_i = (X_i(t_1), ..., X_i(t_n))^\top$ for $i = 1, \ldots, N$. Recall that $\mathbf{S} = \frac{1}{N}\sum_{i=1}^N \mathbf{X}_i\mathbf{X}_i^\top$ and $\mathbf{\Sigma} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^\top\right)$. Suppose that X satisfies Assumption 2. Let $d = \min(n, N)$. Then, there exists a universal constant $\delta_* > 0$ such that for all $x > 0$*

$$\mathbb{P}\left(\left\|\mathbf{S} - \mathbf{\Sigma}\right\|_2 \geqslant \tau_{d,N,n}x\right) \leqslant \exp\left(-(\delta_*^{-1}x)^{\frac{\alpha}{2+\alpha}}\right), \tag{3.1}$$

where $\tau_{N,n} = \max(A_{N,n}^2, B_{N,n})$, with

$$A_{N,n} = \|Z\|_{\psi_\alpha}\frac{\sqrt{\log d}(\log N)^{1/\alpha}}{\sqrt{N}} \text{ and } B_{N,n} = \frac{\rho^2(\mathbf{\Sigma})}{\sqrt{N}} + \|\mathbf{\Sigma}\|_2^{1/2}A_{N,n}.$$

Let us briefly comment Proposition 3 in some specific cases. If X is Gaussian, then Proposition 2 implies that $A_{N,n} \leq A_{N,n,1}$, where

$$A_{N,n,1} = \sqrt{8/3}\sqrt{tr(\mathbf{\Sigma})}\frac{\sqrt{\log d}(\log N)^{1/\alpha}}{\sqrt{N}} \leq \sqrt{8/3}\,\|\mathbf{\Sigma}\|_2^{1/2}\sqrt{\frac{n}{N}}\sqrt{\log d}(\log N)^{1/\alpha}, \tag{3.2}$$

and in this case inequality (3.1) becomes

$$\mathbb{P}\left(\left\|\mathbf{S}-\mathbf{\Sigma}\right\|_2 \geqslant \max\left(A_{N,n,1}^2, B_{N,n,1}\right) x\right) \leqslant \exp\left(-(\delta_*^{-1}x)^{\frac{\alpha}{2+\alpha}}\right) \tag{3.3}$$

for all $x>0$, where $B_{N,n,1}=\frac{\rho^2(\mathbf{\Sigma})}{\sqrt{N}}+\|\mathbf{\Sigma}\|_2^{1/2}A_{N,n,1}$.

If X is a bounded process by some constant $R>0$, then using Proposition 2 and by letting $\alpha\to+\infty$, Proposition 3 implies that for all $x>0$,

$$\mathbb{P}\left(\left\|\mathbf{S}-\mathbf{\Sigma}\right\|_2 \geqslant \max\left(A_{N,n,2}^2, B_{N,n,2}\right) x\right) \leqslant \exp\left(-\delta_*^{-1}x\right), \tag{3.4}$$

where

$$A_{N,n,2}=R\sqrt{\frac{n}{N}}\sqrt{\log d} \text{ and } B_{N,n,2}=\frac{\rho^2\left(\mathbf{\Sigma}\right)}{\sqrt{N}}+\|\mathbf{\Sigma}\|_2^{1/2}A_{N,n,2}. \tag{3.5}$$

Contrary to the low-dimensional case ($n<<N$), in a high-dimensional setting when $n>>N$ or when n and N are of the same magnitude ($\frac{n}{N}\to c>0$ as $n,N\to+\infty$), inequalities (3.3) and (3.4) cannot be used to conclude that the norm $\left\|\mathbf{S}-\mathbf{\Sigma}\right\|_2$ concentrates around zero. Actually, it is well known that the sample covariance $\mathbf{S}$ is a bad estimator of $\mathbf{\Sigma}$ in a high-dimensional setting, and that without any further restriction on the structure of the covariance matrix $\mathbf{\Sigma}$, then $\mathbf{S}$ cannot be a consistent estimator. However, we would like to point out that Proposition 3 relates the quality of $\mathbf{S}$ to the "true dimensionality" of the vector $\mathbf{X}=\left(X\left(t_1\right),...,X\left(t_n\right)\right)^\top\in\mathbb{R}^n$ that is measured by the quantity $\|Z\|_{\psi_\alpha}$ with $Z=\|\mathbf{X}\|_{\ell_2}$. Indeed, if X is a low-dimensional Gaussian process such that $tr(\mathbf{\Sigma})=1$ then Proposition 3 and inequality (3.2) imply that

$$\mathbb{P}\left(\left\|\mathbf{S}-\mathbf{\Sigma}\right\|_2 \geqslant \max\left(A_N^2, B_N\right) x\right) \leqslant \exp\left(-(\delta_*^{-1}x)^{\frac{1}{2}}\right) \tag{3.6}$$

for all $x>0$, where $A_N=\sqrt{8/3}\frac{\sqrt{\log N}(\log N)^{1/\alpha}}{\sqrt{N}}$ and $B_N=\frac{\rho^2(\mathbf{\Sigma})}{\sqrt{N}}+\|\mathbf{\Sigma}\|_2^{1/2}A_N$. Hence, inequality (3.6) shows that, under an assumption of low-dimensionality of the process X, the deviation in operator norm between $\mathbf{S}$ and $\mathbf{\Sigma}$ depends on the ratio $\frac{1}{N}$ and not on $\frac{n}{N}$, and thus the quality of $\mathbf{S}$ as an estimator of $\mathbf{\Sigma}$ is much better in such settings.

More generally, another assumption of low-dimensionality for the process X is to suppose that it has a sparse representation in a dictionary of basis functions, which may also improve the quality of $\mathbf{S}$ as an estimator of $\mathbf{\Sigma}$. To see this, consider the simplest case $X=X^0$, where the process X^0 has a sparse representation in the basis $(g_m)_{1\leq m\leq M}$ given by

$$X^0(t)=\sum_{m\in J^*} a_m g_m(t),\ t\in\mathbb{T}, \tag{3.7}$$

where $J^*\subset\{1,\ldots,M\}$ is a subset of indices of cardinality $|J^*|=s_*$ and a_m, $m\in J^*$ are random coefficients (possibly correlated). Under such an assumption, the following proposition holds.

Proposition 4 *Suppose that $X=X^0$ with X^0 defined by (3.7) with $s_*\leq\min(n,M)$. Assume that X satisfies Assumption 2 and that the matrix $\mathbf{G}_{J^*}^\top\mathbf{G}_{J^*}$ is invertible, where $\mathbf{G}_{J^*}$ denotes the*

$n \times |J^*|$ *matrix obtained by removing the columns of* $\mathbf{G}$ *whose indices are not in* J^**. Then, there exists a universal constant* $\delta_* > 0$ *such that for all* $x > 0$*,*

$$\mathbb{P}\left(\left\|\mathbf{S} - \mathbf{\Sigma}\right\|_2 \geqslant \tilde{\tau}_{N,s_*} x\right) \leqslant \exp\left(-(\delta_*^{-1} x)^{\frac{\alpha}{2+\alpha}}\right), \tag{3.8}$$

where $\tilde{\tau}_{N,s_*} = \max(\tilde{A}^2_{N,s_*}, \tilde{B}_{N,s_*})$*, with*

$$\tilde{A}_{N,s_*} = \rho^{1/2}_{\max}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right) \|\tilde{Z}\|_{\psi_\alpha} \frac{\sqrt{\log d^*}(\log N)^{1/\alpha}}{\sqrt{N}},$$

and

$$\tilde{B}_{N,s_*} = \left(\frac{\rho_{\max}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)}{\rho_{\min}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)}\right)\frac{\rho^2\left(\mathbf{\Sigma}\right)}{\sqrt{N}} + \left(\frac{\rho_{\max}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)}{\rho_{\min}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)}\right)^{1/2} \|\mathbf{\Sigma}\|_2^{1/2}\, \tilde{A}_{d^*,N,s_*},$$

with $d^* = \min(N, s_*)$ *and* $\tilde{Z} = \|\mathbf{a}_{J^*}\|_{\ell_2}$*, where* $\mathbf{a}_{J^*} = (\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*})^{-1}\mathbf{G}^\top_{J^*}\mathbf{X} \in \mathbb{R}^{s_*}$*.*

Using Proposition 2 and Proposition 4 it follows that

- If $X = X^0$ is a Gaussian process then

$$\tilde{A}_{N,s_*} \leq \sqrt{8/3}\left(\frac{\rho_{\max}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)}{\rho_{\min}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)}\right)^{1/2} \|\mathbf{\Sigma}\|_2^{1/2}\sqrt{\frac{s_*}{N}}\sqrt{\log d^*}(\log N)^{1/\alpha} \tag{3.9}$$

- If $X = X^0$ is such that the random variables a_m are bounded by for some constant $R > 0$, then

$$\tilde{A}_{N,s_*} \leq R\|g\|_\infty\sqrt{\frac{s_*}{N}}\sqrt{\log d^*} \tag{3.10}$$

with $\|g\|_\infty = \max_{1\leq m\leq M}\|g_m\|_\infty$ where $\|g_m\|_\infty = \sup_{t\in\mathcal{T}}|g_m(t)|$.

Therefore, let us compare the bounds (3.9) and (3.10) with the inequalities (3.2) and (3.5). It follows that, in the case $X = X^0$, if the sparsity s_* of X in the dictionary is small compared to the number of time points n then the deviation between $\mathbf{S}$ and $\mathbf{\Sigma}$ is much smaller than in the general case without any assumption on the structure of $\mathbf{\Sigma}$. Obviously, the gain also depends on the control of the ratio $\frac{\rho_{\max}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)}{\rho_{\min}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)}$. Note that in the case of an orthonormal design ($M = n$ and $\mathbf{G}^\top\mathbf{G} = \mathbf{I}_n$) then $\rho_{\max}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right) = \rho_{\min}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right) = 1$ for any J^*, and thus the gain in operator norm between $\mathbf{S}$ and $\mathbf{\Sigma}$ clearly depends on the size of $\frac{s_*}{N}$ compared to $\frac{n}{N}$. Supposing that $X = X^0$ also implies that the operator norm of the error term $\mathbf{U}$ in the matrix regression model (2.8) is controlled by the ratio $\frac{s_*}{N}$ instead of the ratio $\frac{n}{N}$ when no assumptions are made on the structure of $\mathbf{\Sigma}$. This means that if X has a sparse representation in the dictionary then the error term $\mathbf{U}$ becomes smaller.

3.2 An oracle inequality for the Frobenius norm

Consistency is first studied for the normalized Frobenius norm $\frac{1}{n}\|\mathbf{A}\|_F^2$ for an $n \times n$ matrix $\mathbf{A}$. The following theorem provides an oracle inequality for the group Lasso estimator $\widehat{\mathbf{\Sigma}}_\lambda = \mathbf{G}\widehat{\mathbf{\Psi}}_\lambda \mathbf{G}^\top$.

Theorem 1 *Assume that X satisfies Assumption 2. Let $\epsilon > 0$ and $1 \leq s \leq \min(n, M)$. Suppose that Assumption 1 holds with $c_0 = 3 + 4/\epsilon$, and that the covariance matrix $\mathbf{\Sigma}_{noise} = \mathbb{E}(\mathbf{W}_1)$ of the noise is positive-definite. Consider the group Lasso estimator $\widehat{\mathbf{\Sigma}}_\lambda$ defined by (2.5) with the choices*

$$\gamma_k = 2\|\mathbf{G}_k\|_{\ell_2}\sqrt{\rho_{\max}(\mathbf{G}\mathbf{G}^\top)},$$

and

$$\lambda = \|\mathbf{\Sigma}_{noise}\|_2 \left(1 + \sqrt{\frac{n}{N}} + \sqrt{\frac{2\delta \log M}{N}}\right)^2 \text{ for some constant } \delta > 1.$$

Then, with probability at least $1 - M^{1-\delta}$ one has that

$$\begin{aligned} \frac{1}{n}\left\|\widehat{\mathbf{\Sigma}}_\lambda - \mathbf{\Sigma}\right\|_F^2 \quad \leq \quad (1+\epsilon) \inf_{\substack{\mathbf{\Psi} \in \mathcal{S}_M \\ \mathcal{M}(\mathbf{\Psi}) \leq s}} & \left(\frac{4}{n}\left\|\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top - \mathbf{\Sigma}\right\|_F^2 + \frac{8}{n}\|\mathbf{S} - \mathbf{\Sigma}\|_F^2 \right. \\ & \left. + C(\epsilon)\frac{\mathbf{G}_{\max}^2 \rho_{\max}(\mathbf{G}^\top\mathbf{G})}{\kappa_{s,c_0}^2}\|\mathbf{\Sigma}_{noise}\|_2^2 \left(1 + \sqrt{\frac{n}{N}} + \sqrt{\frac{2\delta\log M}{N}}\right)^4 \frac{\mathcal{M}(\mathbf{\Psi})}{n}\right), \end{aligned} \tag{3.11}$$

where $\kappa_{s,c_0}^2 = \rho_{\min}(s)^2 - c_0\theta(\mathbf{G})\rho_{\max}(\mathbf{G}^\top\mathbf{G})s$, and $C(\epsilon) = 8\frac{\epsilon}{1+\epsilon}(1 + 2/\epsilon)^2$.

The first term $\frac{1}{n}\left\|\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top - \mathbf{\Sigma}\right\|_F^2$ in inequality (3.11) is the bias of the estimator $\widehat{\mathbf{\Sigma}}_\lambda$. It reflects the quality of the approximation of $\mathbf{\Sigma}$ by the set of matrices of the form $\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top$, with $\mathbf{\Psi} \in \mathcal{S}_M$ and $\mathcal{M}(\mathbf{\Psi}) \leq s$. As an example, suppose that $X = X^0$, where the process X^0 has a sparse representation in the basis $(g_m)_{1\leq m\leq M}$ given by

$$X^0(t) = \sum_{m \in J^*} a_m g_m(t),\ t \in \mathbb{T},$$

where $J^* \subset \{1, \ldots, M\}$ is a subset of indices of cardinality $|J^*| = s_* \leq s$ and $a_m, m \in J^*$ are random coefficients. Then, in this case, since $s_* \leq s$ the bias term in (3.11) is equal to zero.

The second term $\frac{1}{n}\|\mathbf{S} - \mathbf{\Sigma}\|_F^2$ in (3.11) is a variance term as the empirical covariance matrix $\mathbf{S}$ is an unbiased estimator of $\mathbf{\Sigma}$. Using the inequality $\frac{1}{n}\|A\|_F^2 \leq \|A\|_2^2$ that holds for any $n \times n$ matrix A, it follows that $\frac{1}{n}\|\mathbf{S} - \mathbf{\Sigma}\|_F^2 \leq \|\mathbf{S} - \mathbf{\Sigma}\|_2^2$. Therefore, under the assumption that X has a sparse representation in the dictionary (e.g. when $X = X_0$ as above) then the variance term $\frac{1}{n}\|\mathbf{S} - \mathbf{\Sigma}\|_F^2$ is controlled by the ratio $\frac{s_*}{N} \leq \frac{s}{N}$ (see Proposition 4) instead of the ratio $\frac{n}{N}$ without any assumption on the structure of $\mathbf{\Sigma}$.

The third term in (3.11) is also a variance term due to the noise in the measurements (1.1). If there exists a constant $c > 0$ independent of n and N such that $\frac{n}{N} \leq c$ then the decay of this

third variance term is essentially controlled by the ratio $\frac{\mathcal{M}(\mathbf{\Psi})}{n} \leq \frac{s}{n}$. Therefore, if $\mathcal{M}(\mathbf{\Psi}) \leq s$ with sparsity s much smaller than n then the variance of the group Lasso estimator $\widehat{\mathbf{\Sigma}}_\lambda$ is smaller than the variance of $\widetilde{\mathbf{S}}$. This shows some of the improvements achieved by regularization (2.7) of the empirical covariance matrix $\widetilde{\mathbf{S}}$ with a group Lasso penalty.

An important assumption of Theorem 1 is that the covariance matrix of the noise $\mathbf{\Sigma}_{noise} = \mathbb{E}(\mathbf{W}_1)$ is positive definite. This restriction is clearly necessary as illustrated by the following example: suppose that the contaminating process $\mathcal{E}(t) = \zeta g_1(t)$ with $\zeta \sim N(0, \sigma_1^2)$, implying that $\mathbf{\Sigma}_{noise} = \sigma_1^2 \mathbf{g}_1 \mathbf{g}_1^\top$ with $\mathbf{g}_1 = (g_1(t_1), \ldots, g_1(t_n))^\top$ has $n-1$ eigenvalues equal to zero. Now, suppose that $X(t) = a_2 g_2(t)$ with $a_2 \sim N(0, \sigma_2^2)$. If $\sigma_1 > \sigma_2$ then the group LASSO regularization alone cannot get rid of the additive error term without eliminating first the right component g_2. Hence, in such settings, group LASSO regularization does not yield to a consistent estimation of $\mathbf{\Sigma} = \sigma_2^2 \mathbf{g}_2 \mathbf{g}_2^\top$ with $\mathbf{g}_2 = (g_2(t_1), \ldots, g_2(t_n))^\top$.

3.3 An oracle inequality for the operator norm

The "normalized" Frobenius norm $\frac{1}{n}\left\|\widehat{\mathbf{\Sigma}}_\lambda - \mathbf{\Sigma}\right\|_F^2$, i.e the average of the eigenvalues of $\left(\widehat{\mathbf{\Sigma}}_\lambda - \mathbf{\Sigma}\right)^2$, can be viewed as a reasonable proxy for the operator norm $\left\|\widehat{\mathbf{\Sigma}}_\lambda - \mathbf{\Sigma}\right\|_2^2$ (maximum eigenvalue of $\left(\widehat{\mathbf{\Sigma}}_\lambda - \mathbf{\Sigma}\right)^2$). It is thus expected that the results of Theorem 1 imply that the group Lasso estimator $\widehat{\mathbf{\Sigma}}_\lambda$ is a good estimator of $\mathbf{\Sigma}$ in operator norm. Let us recall that controlling the operator norm enables to study the convergence of the eigenvectors and eigenvalues of $\widehat{\mathbf{\Sigma}}_\lambda$ by controlling of the angles between the eigenspaces of a population and a sample covariance matrix through the use of the $\sin(\theta)$ theorems in [Davis and Kahan, 1970].

Now, let us consider the case where X consists in noisy observations of the process X^0 (3.7) meaning that

$$\widetilde{X}(t_j) = X^0(t_j) + \mathcal{E}(t_j), \ j = 1, \ldots, n, \tag{3.12}$$

where $\mathcal{E}$ is a second order Gaussian process $\mathcal{E}$ with zero mean and independent of X^0. In this case, one has that

$$\mathbf{\Sigma} = \mathbf{G}\mathbf{\Psi}^*\mathbf{G}^\top, \text{ where } \mathbf{\Psi}^* = \mathbb{E}\left(\mathbf{a}\mathbf{a}^\top\right),$$

where $\mathbf{a}$ is the random vector of $\mathbb{R}^M$ with $\mathbf{a}_m = a_m$ for $m \in J^*$ and $\mathbf{a}_m = 0$ for $m \notin J^*$. Therefore, using Theorem 1 by replacing s by $s^* = |J^*|$, since $\mathbf{\Psi}^* \in \{\mathbf{\Psi} \in \mathcal{S}_M : M(\mathbf{\Psi}) \leq s_*\}$, one can derive the following corrollary:

Corollary 1 *Suppose that the observations $\widetilde{X}_i(t_j)$ with $i = 1, ..., N$ and $j = 1, \ldots, n$ are i.i.d random variables from model (3.12) and that the conditions of Theorem 1 are satisfied with $1 \leq s = s_* \leq \min(n, M)$. Then, with probability at least $1 - M^{1-\delta}$ one has that*

$$\frac{1}{n}\left\|\widehat{\mathbf{\Sigma}}_\lambda - \mathbf{\Sigma}\right\|_F^2 \leq C_0(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}), \tag{3.13}$$

where

$$C_0(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}) = (1+\epsilon)\left(\frac{8}{n}\left\|\mathbf{S} - \mathbf{G}\mathbf{\Psi}^*\mathbf{G}^\top\right\|_F^2 + C(\epsilon)\frac{\mathbf{G}_{\max}^2 \rho_{\max}(\mathbf{G}^\top\mathbf{G})}{\kappa_{s_*, c_0}^2}\lambda^2\frac{s_*}{n}\right).$$

To simplify notations, write $\widehat{\mathbf{\Psi}} = \widehat{\mathbf{\Psi}}_\lambda$, with $\widehat{\mathbf{\Psi}}_\lambda$ given by (2.7). Define $\hat{J}_\lambda \subset \{1, \ldots, M\}$ as

$$\hat{J}_\lambda \equiv \hat{J} := \left\{ k : \frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} > C_1 \left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise} \right) \right\}, \text{ with } \delta_k = \frac{\|\mathbf{G}_k\|_{\ell_2}}{\mathbf{G}_{\max}}, \quad (3.14)$$

and $C_1 \left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise} \right) = C_1$ with

$$C_1 = \max \left(\gamma_{\max}^{-1} n^{-1/2} \frac{1+\epsilon}{\lambda} \left\| \mathbf{S} - \mathbf{G}\mathbf{\Psi}^* \mathbf{G}^\top \right\|_F^2 ; \frac{4\left(1+\epsilon\right)\sqrt{s_*}}{\epsilon \kappa_{s_*, c_0}} \sqrt{C_0 \left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise} \right)} \right). \quad (3.15)$$

with $\gamma_{\max} = 2\mathbf{G}_{\max} \sqrt{\rho_{\max}(\mathbf{G}^\top \mathbf{G})}$. The set of indices $\hat{J}$ is an estimation of the set of active basis functions J^*. Note that such thresholding procedure (3.14) does not lead immediately to a practical way to choose the set $\hat{J}$. Indeed the constant C_1 in (3.14) depends on the a priori unknown sparsity s_* and on the amplitude of the noise in the matrix regression model (2.8) measured by the quantities $\frac{8}{n} \left\| \mathbf{S} - \mathbf{G}\mathbf{\Psi}^* \mathbf{G}^\top \right\|_F^2$ and $\|\mathbf{\Sigma}_{noise}\|_2^2$. Nevertheless, in Section 4 on numerical experiments we give a simple procedure to automatically threshold the ℓ_2-norm of the columns of the matrix $\widehat{\Psi}_\lambda$ that are two small.

Note that to estimate J^* we did not simply take $\hat{J} = \hat{J}_0 := \left\{ k : \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} \neq 0 \right\}$, but rather apply a thresholding step to discard the columns of $\widehat{\mathbf{\Psi}}$ whose ℓ_2-norm are too small. By doing so, we want to stress the fact that to obtain a consistent procedure with respect to the operator norm it is not sufficient to simply take $\hat{J} = \hat{J}_0$. A similar thresholding step is proposed in [Lounici, 2008] and [Lounici et al., 2009] in the standard linear model to select a sparse set of active variables when using regularization by a Lasso or group-Lasso penalty. In the paper ([Lounici, 2008]), the second thresholding step used to estimate the true sparsity pattern depends on a unknown constant that is related to the amplitude of the unknown coefficients to estimate.

Then, the following theorem holds.

Theorem 2 *Under the assumptions of Corollary 1, for any solution of problem (2.7), we have that with probability at least* $1 - M^{1-\delta}$,

$$\max_{1 \leq k \leq M} \frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k^* \right\|_{\ell_2} \leq C_1 \left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise} \right). \quad (3.16)$$

If in addition

$$\min_{k \in J^*} \frac{\delta_k}{\sqrt{n}} \|\mathbf{\Psi}_k^*\|_{\ell_2} > 2C_1 \left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise} \right) \quad (3.17)$$

then with the same probability the set of indices $\hat{J}$, *defined by (3.14), estimates correctly the true set of active basis functions* J^*, *that is* $\hat{J} = J^*$ *with probability at least* $1 - M^{1-\delta}$.

The results of Theorem 2 indicate that if the ℓ_2-norm of the columns of $\mathbf{\Psi}_k^*$ for $k \in J^*$ are sufficiently large with respect to the level of noise in the matrix regression model (2.8) and the sparsity s_*, then $\hat{J}$ is a consistent estimation of the active set of variables. Indeed, if $\mathcal{M}(\mathbf{\Psi}^*) = s_*$, then by symmetry the columns of $\mathbf{\Psi}^*$ such $\mathbf{\Psi}_k^* \neq 0$ have exactly s_* non-zero entries. Hence, the condition (3.17) means that the ℓ_2-norm of $\mathbf{\Psi}_k^* \neq 0$ (normalized by $\frac{\delta_k}{\sqrt{n}}$) has

to be larger than $\frac{4(1+\epsilon)}{\epsilon\kappa_{s_*,c_0}}\sqrt{s_*}\sqrt{C_0}$. A simple condition to satisfy such an assumption is that the amplitude of the s_* non-vanishing entries of $\mathbf{\Psi}_k^* \neq 0$ are larger than $\frac{\sqrt{n}}{\delta_k}\frac{4(1+\epsilon)}{\epsilon\kappa_{s_*,c_0}}\sqrt{C_0}$ which can be interpreted as a kind of measure of the noise in model (2.8). This suggests to take as a final estimator of $\mathbf{\Sigma}$ the following matrix:

$$\widehat{\mathbf{\Sigma}}_{\hat{J}} = \mathbf{G}_{\hat{J}}\widehat{\mathbf{\Psi}}_{\hat{J}}\mathbf{G}_{\hat{J}} \tag{3.18}$$

where $\mathbf{G}_{\hat{J}}$ denotes the $n \times |\hat{J}|$ matrix obtained by removing the columns of $\mathbf{G}$ whose indices are not in $\hat{J}$, and

$$\widehat{\mathbf{\Psi}}_{\hat{J}} = \underset{\mathbf{\Psi}\in\mathcal{S}_{|\hat{J}|}}{\operatorname{argmin}} \left\{ \left\| \widetilde{\mathbf{S}} - \mathbf{G}_{\hat{J}}\mathbf{\Psi}\mathbf{G}_{\hat{J}}^\top \right\|_F^2 \right\},$$

where $\mathcal{S}_{|\hat{J}|}$ denotes the set of $|\hat{J}| \times |\hat{J}|$ symmetric matrices. Note that if $\mathbf{G}_{\hat{J}}^\top\mathbf{G}_{\hat{J}}$ is invertible, then

$$\widehat{\mathbf{\Psi}}_{\hat{J}} = \left(\mathbf{G}_{\hat{J}}^\top\mathbf{G}_{\hat{J}}\right)^{-1}\mathbf{G}_{\hat{J}}^\top\widetilde{\mathbf{S}}\mathbf{G}_{\hat{J}}\left(\mathbf{G}_{\hat{J}}^\top\mathbf{G}_{\hat{J}}\right)^{-1}.$$

Let us recall that if the observations are i.i.d random variables from model (3.12) then

$$\mathbf{\Sigma} = \mathbf{G}\mathbf{\Psi}^*\mathbf{G}^\top,$$

where $\mathbf{\Psi}^* = \mathbb{E}\left(\mathbf{a}\mathbf{a}^\top\right)$, and $\mathbf{a}$ is the random vector of $\mathbb{R}^M$ with $\mathbf{a}_m = a_m$ for $m \in J^*$ and $\mathbf{a}_m = 0$ for $m \notin J^*$. Then, define the random vector $\mathbf{a}_{J^*} \in \mathbb{R}^{J^*}$ whose coordinates are the random coefficients a_m for $m \in J^*$. Let $\mathbf{\Psi}_{J^*} = \mathbb{E}\left(\mathbf{a}_{J^*}\mathbf{a}_{J^*}^\top\right)$ and denote by $\mathbf{G}_{J^*}$ the $n \times |J^*|$ matrix obtained by removing the columns of $\mathbf{G}$ whose indices are not in J^*. Note that $\mathbf{\Sigma} = \mathbf{G}_{J^*}\mathbf{\Psi}_{J^*}\mathbf{G}_{J^*}^\top$.

Assuming that $\mathbf{G}_{J^*}^\top\mathbf{G}_{J^*}$ is invertible, define the matrix

$$\mathbf{\Sigma}_{J^*} = \mathbf{\Sigma} + \mathbf{G}_{J^*}(\mathbf{G}_{J^*}^\top\mathbf{G}_{J^*})^{-1}\mathbf{G}_{J^*}^\top\mathbf{\Sigma}_{noise}\mathbf{G}_{J^*}\left(\mathbf{G}_{J^*}^\top\mathbf{G}_{J^*}\right)^{-1}\mathbf{G}_{J^*}^\top. \tag{3.19}$$

Then, the following theorem gives a control of deviation between $\widehat{\mathbf{\Sigma}}_{\hat{J}}$ and $\mathbf{\Sigma}_{J^*}$ in operator norm.

Theorem 3 *Suppose that the observations are i.i.d random variables from model (3.12) and that the conditions of Theorem 1 are satisfied with* $1 \leq s = s_* \leq \min(n, M)$. *Suppose that* $\mathbf{G}_{J^*}^\top\mathbf{G}_{J^*}$ *is an invertible matrix, and that*

$$\min_{k\in J^*}\frac{\delta_k}{\sqrt{n}}\left\|\mathbf{\Psi}_k^*\right\|_{\ell_2} > 2C_1\left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}\right),$$

where $C_1\left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}\right)$ *is the constant defined in (3.15). Let* $\mathbf{Y} = \left(\mathbf{G}_{J^*}^\top\mathbf{G}_{J^*}\right)^{-1}\mathbf{G}_{J^*}^\top\widetilde{\mathbf{X}}$ *and* $\tilde{Z} = \|\mathbf{Y}\|_{\ell_2}$. *Let* $\rho\left(\mathbf{\Sigma}_{noise}\right) = \left(\sup_{\beta\in\mathbb{R}^n,\|\beta\|_{\ell_2}=1}\mathbb{E}|\mathcal{E}^\top\beta|^4\right)^{1/4}$ *where* $\mathcal{E} = \left(\mathcal{E}\left(t_1\right), ..., \mathcal{E}\left(t_n\right)\right)^\top$. *Then, with probability at least* $1 - M^{1-\delta} - M^{-\left(\frac{\delta_\star}{\delta_*}\right)^{\frac{\alpha}{2+\alpha}}}$, *with* $\delta > 1$ *and* $\delta_\star > \delta_*$ *one has that*

$$\left\|\widehat{\mathbf{\Sigma}}_{\hat{J}} - \mathbf{\Sigma}_{J^*}\right\|_2 \leq \rho_{\max}\left(\mathbf{G}_{J^*}^\top\mathbf{G}_{J^*}\right)\tilde{\tau}_{N,s_*}\delta_\star\left(\log(M)\right)^{\frac{2+\alpha}{\alpha}}, \tag{3.20}$$

where $\tilde{\tau}_{N,s_*} = \max(\tilde{A}^2_{N,s_*}, \tilde{B}_{N,s_*})$, *with* $\tilde{A}_{N,s_*} = \|\tilde{Z}\|_{\psi_\alpha} \frac{\sqrt{\log d^*}(\log N)^{1/\alpha}}{\sqrt{N}}$, $\tilde{B}_{N,s_*} = \frac{\tilde{\rho}^2(\mathbf{\Sigma},\mathbf{\Sigma}_{noise})\rho^{-1}_{\min}(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*})}{\sqrt{N}} + \left(\|\mathbf{\Psi}_{J^*}\|_2 + \rho^{-1}_{\min}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)\|\mathbf{\Sigma}_{noise}\|_2\right)^{1/2}\tilde{A}_{N,s_*}$, *where* $d^* = \min(N, s_*)$ *and* $\tilde{\rho}(\mathbf{\Sigma}, \mathbf{\Sigma}_{noise}) = 8^{1/4}\left(\rho^4(\mathbf{\Sigma}) + \rho^4(\mathbf{\Sigma}_{noise})\right)^{1/4}$.

First note that the above theorem gives a deviation in operator norm from $\widehat{\mathbf{\Sigma}}_{\hat{J}}$ to the matrix $\mathbf{\Sigma}_{J^*}$ (3.19) which is not equal to the true covariance $\mathbf{\Sigma}$ of X at the design points. Indeed, even if we know the true sparsity set J^*, the additive noise in the measurements in model (1.1) complicates the estimation of $\mathbf{\Sigma}$ in operator norm. However, although $\mathbf{\Sigma}_{J^*} \neq \mathbf{\Sigma}$, they can have the same eigenvectors if the structure of the additive noise matrix term in (3.19) is not too complex. As an example, consider the case of an additive white noise, for which $\mathbf{\Sigma}_{noise} = \sigma^2\mathbf{I}_n$ where σ is the level of noise and $\mathbf{I}_n$ the $n \times n$ identity matrix. Under such an assumption, if we further suppose for simplicity that $(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*})^{-1} = \mathbf{I}_{s_*}$, then $\mathbf{\Sigma}_{J^*} = \mathbf{\Sigma} + \sigma^2\mathbf{G}_{J^*}(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*})^{-1}\mathbf{G}^\top_{J^*} = \mathbf{\Sigma} + \sigma^2\mathbf{I}_n$ and clearly $\mathbf{\Sigma}_{J^*}$ and $\mathbf{\Sigma}$ have the same eigenvectors. Therefore, the eigenvectors of $\widehat{\mathbf{\Sigma}}_{\hat{J}}$ can be used as estimators of the eigenvectors of $\mathbf{\Sigma}$ which is suitable for the sparse PCA application described in the next section on numerical experiments.

Let us illustrate the implications of Theorem 3 on a simple example. If X is Gaussian, the random vector $\mathbf{Y} = \left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)^{-1}\mathbf{G}^\top_{J^*}(\mathbf{X} + \mathcal{E})$ is also Gaussian and Proposition 2 can be used to prove that

$$\begin{aligned}\|\tilde{Z}\|_{\psi_2} &\leq \sqrt{8/3}\sqrt{tr\left(\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)^{-1}\mathbf{G}^\top_{J^*}(\mathbf{\Sigma} + \mathbf{\Sigma}_{noise})\mathbf{G}_{J^*}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)^{-1}\right)} \\ &\leq \sqrt{8/3}\|\mathbf{\Sigma} + \mathbf{\Sigma}_{noise}\|_2^{1/2}\rho^{-1/2}_{\min}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)\sqrt{s_*}.\end{aligned}$$

Then Theorem 3 implies that with high probability

$$\left\|\widehat{\mathbf{\Sigma}}_{\hat{J}} - \mathbf{\Sigma}_{J^*}\right\|_2 \leq \rho_{\max}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)\tilde{\tau}_{N,s_*,1}\delta\left(\log(M)\right)^{\frac{2+\alpha}{\alpha}},$$

where $\tilde{\tau}_{N,s_*,1} = \max(\tilde{A}^2_{N,s_*,1}, \tilde{B}_{N,s_*,1})$, with

$$\tilde{A}_{N,s_*,1} = \sqrt{8/3}\|\mathbf{\Sigma} + \mathbf{\Sigma}_{noise}\|_2^{1/2}\rho^{-1/2}_{\min}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)\sqrt{\log d^*}(\log N)^{1/\alpha}\sqrt{\frac{s_*}{N}}$$

and

$$\tilde{B}_{N,s_*,1} = \frac{\tilde{\rho}^2(\mathbf{\Sigma}, \mathbf{\Sigma}_{noise})\rho^{-1}_{\min}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)}{\sqrt{N}} + \left(\|\mathbf{\Psi}_{J^*}\|_2 + \rho^{-1}_{\min}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)\|\mathbf{\Sigma}_{noise}\|_2\right)^{1/2}\tilde{A}_{N,s_*,1}.$$

Therefore, in the Gaussian case (but also under other assumptions for X such as those in Proposition 2) the above equations show that the operator norm $\left\|\widehat{\mathbf{\Sigma}}_{\hat{J}} - \mathbf{\Sigma}_{J^*}\right\|_2^2$ depends on the ratio $\frac{s_*}{N}$. Recall that $\|\mathbf{S} - \mathbf{\Sigma}\|_2^2$ depends on the ratio $\frac{n}{N}$. Thus, using $\widehat{\mathbf{\Sigma}}_{\hat{J}}$ clearly yields significant improvements if s_* is small compared to n.

To summarize our results let us finally consider the case of an orthogonal design. Combining Theorems 1, 2 and 3 one arrives at the following corrolary:

Corollary 2 *Suppose that the observations are i.i.d random variables from model (3.12). Suppose that $M = n$ and that $\mathbf{G}^\top \mathbf{G} = \mathbf{I}_n$ (orthogonal design) and that X^0 satisfies Assumption 2. Let $\epsilon > 0$ and $1 \leq s_* \leq \min(n, M)$. Consider the group Lasso estimator $\widehat{\mathbf{\Sigma}}_\lambda$ defined by (2.5) with the choices*

$$\gamma_k = 2, k = 1, \ldots, n \text{ and } \lambda = \|\mathbf{\Sigma}_{noise}\|_2 \left(1 + \sqrt{\frac{n}{N}} + \sqrt{\frac{2\delta \log M}{N}}\right)^2 \text{ for some constant } \delta > 1.$$

Suppose that

$$\min_{k \in J^*} \|\mathbf{\Psi}_k^*\|_{\ell_2} > 2n^{1/2} \tilde{C}_1 \left(\sigma, n, s_*, N, \delta\right), \tag{3.21}$$

where $\tilde{C}_1 \left(\sigma, n, s, N, \delta\right) = \frac{4(1+\epsilon)\sqrt{s_*}}{\epsilon} \sqrt{\tilde{C}_0 \left(\sigma, n, s_*, N, \delta\right)}$ *and*

$$\tilde{C}_0 \left(\sigma, n, s_*, N, \delta\right) = (1+\epsilon) \left(\frac{8}{n} \left\| \mathbf{S} - \mathbf{G}\mathbf{\Psi}^* \mathbf{G}^\top \right\|_F^2 + C(\epsilon) \|\mathbf{\Sigma}_{noise}\|_2^2 \left(1 + \sqrt{\frac{n}{N}} + \sqrt{\frac{2\delta \log M}{N}}\right)^4 \frac{s_*}{n} \right).$$

Take $\hat{J} := \left\{ k : \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} > n^{1/2} \tilde{C}_1 \left(\sigma, n, s, N, \delta\right) \right\}$. *Let* $\mathbf{Y} = \mathbf{G}_{J^*}^\top \widetilde{\mathbf{X}}$ *and* $\tilde{Z} = \|\mathbf{Y}\|_{\ell_2}$. *Then, with probability at least* $1 - M^{1-\delta} - M^{-\left(\frac{\delta_\star}{\delta_*}\right)^{\frac{\alpha}{2+\alpha}}}$, *with* $\delta > 1$ *and* $\delta_\star > \delta_*$ *one has that*

$$\left\| \widehat{\mathbf{\Sigma}}_{\hat{J}} - \mathbf{\Sigma}_{J^*} \right\|_2 \leq \tilde{\tau}_{N, s_*} \delta_\star \left(\log(M)\right)^{\frac{2+\alpha}{\alpha}}, \tag{3.22}$$

where $\tilde{\tau}_{N,s_*} = \max(\tilde{A}_{N,s_*}^2, \tilde{B}_{N,s_*})$, *with* $\tilde{A}_{N,s_*} = \|\tilde{Z}\|_{\psi_\alpha} \frac{\sqrt{\log d^*} (\log N)^{1/\alpha}}{\sqrt{N}}$ *and* $\tilde{B}_{N,s_*} = \frac{\tilde{\rho}^2(\mathbf{\Sigma}, \mathbf{\Sigma}_{noise})}{\sqrt{N}} + \left(\|\mathbf{\Psi}_{J^*}\|_2 + \|\mathbf{\Sigma}_{noise}\|_2\right)^{1/2} \tilde{A}_{N,s_*}$.

3.4 Comparison with the standard Lasso

In this work, we chose a Group Lasso estimation procedure rather than a standard Lasso. As a matter of fact, for covariance estimation in our setting, the group structure enables to impose a constraint on the number of non zero columns of the matrix Ψ and not on the single entries of the matrix Ψ. This corresponds to the natural assumption of obtaining a sparse representation of the process $X(t)$ in the basis given by the functions g_m's and replacing its dimension by its sparsity. Alternatively, the standard Lasso in our setting would be the estimator defined by

$$\widehat{\mathbf{\Psi}}_L = \underset{\mathbf{\Psi} \in \mathcal{S}_M}{\operatorname{argmin}} \left\{ \left\| \widetilde{\mathbf{S}} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top \right\|_F^2 + 2\lambda \sum_{k=1}^M \sum_{m=1}^M \gamma_{mk} |\Psi_{mk}| \right\},$$

where $\lambda \geq 0$ is a regularization parameters and the γ_{mk}'s are positive weights. This procedure leads to the following Lasso estimator of the covariance matrix $\mathbf{\Sigma}$

$$\widehat{\mathbf{\Sigma}}_L = \mathbf{G}\widehat{\mathbf{\Psi}}_L \mathbf{G}^\top \in \mathbb{R}^{n \times n}. \tag{3.23}$$

In the orthogonal case (i.e. $M = n$ and $\mathbf{G}^\top \mathbf{G} = \mathbf{I}_n$), this gives rise to the estimator $\widehat{\mathbf{\Psi}}_L$ obtained by soft thresholding individually each entry Y_{mk} of the matrix $\mathbf{Y} = \mathbf{G}^\top \widetilde{\mathbf{S}} \mathbf{G}$ with the thresholds

$\lambda\gamma_{mk}$. Proposition 5 (see below) allows a simple comparison of the statistical performances of the group Lasso estimator $\widehat{\mathbf{\Sigma}}_L$ with those of the standard Lasso estimator $\widehat{\mathbf{\Sigma}}_\lambda$ in terms of upper bounds for the Frobenius norm. To simplify the discussion, we only consider the orthogonal case and the simple model

$$\widetilde{X}(t_j) = X^0(t_j) + \mathcal{E}(t_j),\ j = 1, \ldots, n, \tag{3.24}$$

where the process X^0 is defined in (3.7). The statement of the result for the group Lasso is an immediate consequence of Theorem 1, while the proof to obtain the upper bound for the standard Lasso is an immediate adaptation of the arguments in the proof of Theorem 1.

Proposition 5 *Assume that X satisfies model* (3.24) *and that the covariance matrix $\mathbf{\Sigma}_{noise} = \mathbb{E}(\mathbf{W}_1)$ of the noise is positive-definite. Consider the group Lasso estimator $\widehat{\mathbf{\Sigma}}_\lambda$ and the standard Lasso estimator $\widehat{\mathbf{\Sigma}}_L$ with the choices*

$$\gamma_k = 2,\ \gamma_{mk} = 2,\ \lambda = \|\mathbf{\Sigma}_{noise}\|_2 \left(2 + \sqrt{\frac{2\delta \log M}{N}}\right)^2 \text{ for some constant } \delta > 1.$$

Then, there exist two positive constants C_1, C_2 not depending on n, N, s_* such that with probability at least $1 - M^{1-\delta}$ one has that

$$\frac{1}{n}\left\|\widehat{\mathbf{\Sigma}}_\lambda - \mathbf{\Sigma}\right\|_F^2 \leq \frac{C_1}{n}\|\mathbf{S} - \mathbf{\Sigma}\|_F^2 + C_2\|\mathbf{\Sigma}_{noise}\|_2^2 \left(2 + \sqrt{\frac{2\delta \log n}{N}}\right)^4 \frac{s_*}{n},$$

and

$$\frac{1}{n}\left\|\widehat{\mathbf{\Sigma}}_L - \mathbf{\Sigma}\right\|_F^2 \leq \frac{C_1}{n}\|\mathbf{S} - \mathbf{\Sigma}\|_F^2 + C_2\|\mathbf{\Sigma}_{noise}\|_2^2 \left(2 + \sqrt{\frac{2\delta \log n}{N}}\right)^4 \frac{s_*^2}{n}.$$

Proposition 5 illustrates the advantages of the Group Lasso over the standard Lasso. Indeed, the second term in the upper bound for the group Lasso is much smaller (of the order $\frac{s_*}{n}$) than the second term in the upper bound for the standard Lasso (of the order $\frac{s_*^2}{n}$). This comes from the fact that the sparsity prior of the Group Lasso is on the number of vanishing columns of the matrix Ψ, while the sparsity prior of the standard Lasso only controls the number of non-zero entries of Ψ. However, to really demonstrate the benefits of our method when compared to the performances of the standard Lasso, it is required to also derive lower bounds. This issue is a difficult task which has been considered in few papers and that is beyond the scope of this paper. For recent work in this direction, we refer to [Huang and Zhang, 2010] for regression models or [Lounici et al., 2011] and [Lounici et al., 2009] for linear regression and multi-task learning.

However, the analysis in [Huang and Zhang, 2010, Lounici et al., 2011] of Group Lasso regularization is carried out the setting of multiple regression models where the parameters to estimate are vectors and with error terms that are centered. Therefore, the results in [Huang and Zhang, 2010, Lounici et al., 2011] cannot be applied to the matrix regression model (2.4) since, in our setting, the parameter to estimate is the matrix $\mathbf{\Sigma}$ and the error terms $\mathbf{U}_i + \mathbf{W}_i$ in (2.4) are not centered.

4 Numerical experiments and an application to sparse PCA

In this section we present some simulated examples to illustrate the practical behaviour of the covariance matrix estimator by group Lasso regularization proposed in this paper. In particular, we show its performances with an application to sparse Principal Components Analysis (PCA). In the numerical experiments, we use the explicit estimator described in Proposition 1 in the case $M = n$ and an orthogonal design matrix $\mathbf{G}$, and also the estimator proposed in the more general situation when $n < M$. The programs for our simulations were implemented using the MATLAB programming environment.

4.1 Description of the estimating procedure and the data

We consider a noisy stochastic processes $\widetilde{X}$ on $\mathbb{T} = [0, 1]$ with values in $\mathbb{R}$ observed at fixed location points $t_1, ..., t_n$ in $[0, 1]$, generated according to

$$\widetilde{X}(t_j) = X^0(t_j) + \sigma\epsilon_j,\ j = 1, \ldots, n, \tag{4.1}$$

where $\sigma > 0$ is the level of noise, $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. standard Gaussian variables, and X^0 is a random process independent of the ϵ_j's. For the process X^0 we consider two simple models. The first one is given by

$$X^0(t) = af(t), \tag{4.2}$$

where a is a Gaussian random coefficient such that $\mathbb{E}a = 0$, $\mathbb{E}a^2 = \gamma^2$, and $f : [0, 1] \to \mathbb{R}$ is an unknown function. The second model for X^0 is

$$X^0(t) = a_1 f_1(t) + a_2 f_2(t), \tag{4.3}$$

where a_1 and a_2 are independent Gaussian variables such that $\mathbb{E}a_1 = \mathbb{E}a_2 = 0$, $\mathbb{E}a_1^2 = \gamma_1^2$, $\mathbb{E}a_2^2 = \gamma_2^2$ (with $\gamma_1 > \gamma_2$), and $f_1, f_2 : [0, 1] \to \mathbb{R}$ are unknown functions. The simulated data consists in a sample of N independent observations of the process $\widetilde{X}$ at the points $t_1, ..., t_n$, which are generated according to (4.1). Therefore, throughout the numerical experiments, one has that

$$\mathbf{\Sigma}_{noise} = \sigma^2 \mathbf{I}_n.$$

In model (4.2), the covariance matrix $\mathbf{\Sigma}$ of the process X^0 at the locations points is given by $\mathbf{\Sigma} = \gamma^2 \mathbf{F}\mathbf{F}^\top$, where by definition

$$\mathbf{F} = \left(f\left(t_1\right), ..., f\left(t_1\right)\right)^\top \in \mathbb{R}^n.$$

Note that the largest eigenvalue of $\mathbf{\Sigma}$ is $\gamma^2 \|\mathbf{F}\|_{\ell_2}^2$ with corresponding eigenvector $\mathbf{F}$. We suppose that the signal f has some sparse representation in a large dictionary of basis functions of size M, given by $\{g_m,\, m = 1, \ldots, M\}$, meaning that $f\left(t\right) = \sum_{m=1}^M \beta_m g_m\left(t\right)$, with $J^* = \{m, \beta_m \neq 0\}$ of small cardinality s_*. Then, the process X^0 can be written as $X^0(t) = \sum_{m=1}^M a\beta_m g_m\left(t\right)$, and thus $\mathbf{\Sigma} = \gamma^2 \mathbf{G}\mathbf{\Psi}_{J^*}\mathbf{G}^\top$, where $\mathbf{\Psi}_{J^*}$ is an $M \times M$ matrix with entries equal to $\beta_m\beta_{m'}$ for $1 \leq m, m' \leq M$.

Similarly, in model (4.3), the covariance matrix $\mathbf{\Sigma}$ of the process X^0 at the locations points is given by $\mathbf{\Sigma} = \gamma_1^2 \mathbf{F}_1 \mathbf{F}_1^\top + \gamma_2^2 \mathbf{F}_2 \mathbf{F}_2^\top$, where by definition

$$\mathbf{F}_1 = (f_1(t_1), ..., f(t_1))^\top \in \mathbb{R}^n \text{ and } \mathbf{F}_2 = (f_2(t_1), ..., f(t_1))^\top \in \mathbb{R}^n.$$

In the following simulations, the functions f_1 and f_2 are chosen such that $\mathbf{F}_1$ and $\mathbf{F}_2$ are orthogonal vectors in $\mathbb{R}^n$ with $\|\mathbf{F}_1\|_{\ell_2} = 1$ and $\|\mathbf{F}_2\|_{\ell_2} = 1$. Under such an assumption and since $\gamma_1 > \gamma_2$, the largest eigenvalue of $\mathbf{\Sigma}$ is γ_1^2 with corresponding eigenvector $\mathbf{F}_1$, and the second largest eigenvalue of $\mathbf{\Sigma}$ is γ_2^2 with corresponding eigenvector $\mathbf{F}_2$. We suppose that the signals f_1 and f_2 have some sparse representations in a large dictionary of basis functions of size M, given by $f_1(t) = \sum_{m=1}^M \beta_m^1 g_m(t)$, and $f_2(t) = \sum_{m=1}^M \beta_m^2 g_m(t)$. Then, the process X^0 can be written as $X^0(t) = \sum_{m=1}^M (a_1 \beta_m^1 + a_2 \beta_m^2) g_m(t)$ and thus $\mathbf{\Sigma} = \mathbf{G}(\gamma_1^2 \mathbf{\Psi}^1 + \gamma_2^2 \mathbf{\Psi}^2)\mathbf{G}^\top$, where $\mathbf{\Psi}^1, \mathbf{\Psi}^2$ are $M \times M$ matrix with entries equal to $\beta_m^1(\beta_m^1)'$ and $\beta_m^2(\beta_m^2)'$ for $1 \leq m, m' \leq M$ respectively.

In models (4.2) and (4.3), we aim at estimating either $\mathbf{F}$ or $\mathbf{F}_1, \mathbf{F}_2$ by the eigenvectors corresponding to the largest eigenvalues of the matrix $\widehat{\mathbf{\Sigma}}_{\hat{J}}$ defined in (3.18), in a high-dimensional setting with $n > N$ and by using different type of dictionaries. The idea behind this is that $\widehat{\mathbf{\Sigma}}_{\hat{J}}$ is a consistent estimator of $\mathbf{\Sigma}_{J^*}$ (see its definition in 3.19) in operator norm. Although the matrices $\mathbf{\Sigma}_{J^*}$ and $\mathbf{\Sigma}$ may have different eigenvectors (depending on the design points and chosen dictionary), the examples below show the eigenvectors of $\widehat{\mathbf{\Sigma}}_{\hat{J}}$ can be used as estimators of the eigenvectors of $\mathbf{\Sigma}$.

The estimator $\widehat{\mathbf{\Sigma}}_{\hat{J}}$ of the covariance matrix $\mathbf{\Sigma}$ is computed as follows. Once the dictionary has been chosen, we compute the covariance group Lasso (CGL) estimator $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}} = \mathbf{G}\widehat{\mathbf{\Psi}}_{\widehat{\lambda}}\mathbf{G}^\top$, where $\widehat{\mathbf{\Psi}}_{\widehat{\lambda}}$ is defined in (2.7). We use a completely data-driven choice for the regularizarion parameter λ, given by $\widehat{\lambda} = \|\widehat{\mathbf{\Sigma}_{noise}}\|_2 \left(1 + \sqrt{\frac{n}{N}} + \sqrt{\frac{2\delta \log M}{N}}\right)^2$, where $\|\widehat{\mathbf{\Sigma}_{noise}}\|_2 = \widehat{\sigma}^2$ is the median absolute deviation (MAD) estimator of σ^2 used in standard wavelet denoising (see e.g. [Antoniadis et al., 2001]) and $\delta = 1.1$. Hence, the method to compute $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$ is fully data-driven. Furthermore, we will show in the examples below that replacing λ by $\widehat{\lambda}$ into the penalized criterion yields a very good practical performance of the covariance estimation procedure.

As a final step, one needs to compute the estimator $\widehat{\mathbf{\Sigma}}_{\hat{J}}$ of $\mathbf{\Sigma}$, as in (3.18). For this, we need to have an idea of the true sparsity s_*, since $\hat{J}$ defined in (3.14) depends on s_* and also on unknown upper bounds on the level of noise in the matrix regression model (2.8) . A similar problem arises in the selection of a sparse set of active variables when using regularization by a Lasso penalty in the standard linear model. As an example, recall that in [Lounici, 2008], a second thresholding step is aso used to estimate the true sparsity pattern. However, the suggested thresholding procedure in [Lounici, 2008] also depends on a priori unknown quantities (such as the amplitude of the coefficients to estimate). To overcome this drawback in our case, we can define the final covariance group Lasso (FCGL) estimator as the matrix

$$\widehat{\mathbf{\Sigma}}_{\hat{J}} = \mathbf{G}_{\hat{J}} \widehat{\mathbf{\Psi}}_{\hat{J}} \mathbf{G}_{\hat{J}}^\top, \tag{4.4}$$

with $\hat{J} = \hat{J}_\epsilon = \left\{k : \left\|\widehat{\mathbf{\Psi}}_k\right\|_{\ell_2} > \varepsilon\right\}$, where ε is a positive constant. To select an appropriate value of ϵ, one can plot the cardinality of $\hat{J}_\epsilon$ as a function of ϵ, and then use an L-curve criterion to

only keep in $\hat{J}$ the indices of the columns of $\widehat{\mathbf{\Psi}}_{\hat{\lambda}}$ with a significant value in ℓ_2-norm. This choice for $\hat{J}$ is sufficient for numerical purposes.

In the simulations, to measure the accuracy of the estimation procedure, we also use the empirical average of the Frobenius and operator norm of the estimators $\widehat{\mathbf{\Sigma}}_{\hat{\lambda}}$ and $\widehat{\mathbf{\Sigma}}_{\hat{J}}$ with respect to the true covariance matrix $\mathbf{\Sigma}$ defined by $EAFN = \frac{1}{P}\sum_{p=1}^{P}\left\|\widehat{\mathbf{\Sigma}}^p_{\hat{\lambda}} - \mathbf{\Sigma}\right\|_F$ and $EAON = \frac{1}{P}\sum_{p=1}^{P}\left\|\widehat{\mathbf{\Sigma}}^p_{\hat{J}} - \mathbf{\Sigma}\right\|_2$ respectively, over a number P of iterations, where $\widehat{\mathbf{\Sigma}}^p_{\hat{\lambda}}$ and $\widehat{\mathbf{\Sigma}}^p_{\hat{J}}$ are the CGL and FCGL estimators of $\mathbf{\Sigma}$, respectively, obtained at the p-th iteration. We also compute the empirical average of the operator norm of the estimator $\widehat{\mathbf{\Sigma}}_{\hat{J}}$ with respect to the matrix $\mathbf{\Sigma}_{J^*}$, defined by $EAON^* = \frac{1}{P}\sum_{p=1}^{P}\left\|\widehat{\mathbf{\Sigma}}^p_{\hat{J}} - \mathbf{\Sigma}_{J^*}\right\|_2$.

4.2 Model (4.2) - case of an orthonormal design (with $n = M$)

First, the size of the dictionary M as well as the basis functions $\{g_m, m = 1, ..., M\}$ have to be specified. In model (4.2), we will use for the test function f the signals HeaviSine and Blocks (see e.g. [Antoniadis et al., 2001] for a definition), and the Symmlet 8 and Haar wavelet basis for the HeaviSine and Blocks signals respectively, which are implemented in the Matlab's open-source library WaveLab (see e.g. [Antoniadis et al., 2001] for further references on wavelet methods in nonparametric statistics). Then, we took $n = M$ and the location points $t_1, ..., t_n$ are given by the equidistant grid of points $t_j = \frac{j}{M}$, $j = 1, \ldots, M$ such that the design matrix $\mathbf{G}$ (using either the Symmlet 8 or the Haar basis) is orthogonal.

Figures 1, 2, and 3 present the results obtained for a particular simulated sample of size $N = 25$ according to (4.1), with $n = M = 256$, $\sigma = 0.015$, $\gamma = 0.5$ and with f being either the function HeaviSine or the function Blocks. It can be observed in Figures 1(a) and 1(b) that, as expected in this high dimensional setting ($N < n$), the empirical eigenvector of $\widetilde{\mathbf{S}}$ associated to its largest empirical eigenvalue does not lead to a consistent estimator of $\mathbf{F}$.

The CGL estimator $\widehat{\mathbf{\Sigma}}_{\hat{\lambda}}$ is computed directly from Proposition 1. In Figures 2(a) and 2(b), we display the eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\hat{\lambda}}$ as an estimator of $\mathbf{F}$. Note that this estimator behaves poorly. The estimation considerably improves by taking the FCGL estimator $\widehat{\mathbf{\Sigma}}_{\hat{J}}$ defined in (4.4). Figures 3(a) and 3(b) illustrate the very good performance of the eigenvector associated to the largest eigenvalue of the matrix $\widehat{\mathbf{\Sigma}}_{\hat{J}}$ as an estimator of $\mathbf{F}$.

It is clear that the estimators $\widehat{\mathbf{\Sigma}}_{\hat{\lambda}}$ and $\widehat{\mathbf{\Sigma}}_{\hat{J}}$ are random matrices that depend on the observed sample. Tables 1(a) and 1(b) show the values of $EAFN$, $EAON$ and $EAON^*$ corresponding to $P = 100$ simulated samples of different sizes N and different values of the level of noise σ. It can be observed that for both signals the empirical averages $EAFN$, $EAON$ and $EAON^*$ behaves similarly, being the values of $EAON$ smaller than its corresponding values of $EAFN$ as expected. Observing each table separately we can remark that, for N fixed, when the level of noise σ increases then the values of $EAFN$, $EAON$ and $EAON^*$ also increase. By simple inspection of the values of $EAFN$, $EAON$ and $EAON^*$ in the same position at Tables 1(a) and 1(b) we can check that, for σ fixed, when the number of replicates N increases then the values of $EAFN$, $EAON$ and $EAON^*$ decrease in all cases. We can also observe how the difference

between $EAON$ and $EAON^*$ is bigger as the level of noise increases.

Table 1(a). Values of $EAFN$, $EAON$ and $EAON^*$ corresponding to signals HeaviSine and Blocks for $M = n = 256$, $N = 25$.

Signal	σ	0.005	0.01	0.05	0.1	0.5	1
HeaviSine	$EAFN$	0.0634	0.0634	0.2199	0.2500	0.2500	0.2500
HeaviSine	$EAON$	0.0619	0.0569	0.1932	0.2500	0.2500	0.2500
HeaviSine	$EAON^*$	0.0619	0.0569	0.1943	0.2600	0.5000	1.2500
Blocks	$EAFN$	0.0553	0.0681	0.2247	0.2500	0.2500	0.2500
Blocks	$EAON$	0.0531	0.0541	0.2083	0.2500	0.2500	0.2500
Blocks	$EAON^*$	0.0531	0.0541	0.2107	0.2600	0.5000	1.2500

Table 1(b). Values of $EAFN$, $EAON$ and $EAON^*$ corresponding to signals HeaviSine and Blocks for $M = n = 256$, $N = 40$.

Signal	σ	0.005	0.01	0.05	0.1	0.5	1
HeaviSine	$EAFN$	0.0501	0.0524	0.1849	0.2499	0.2500	0.2500
HeaviSine	$EAON$	0.0496	0.0480	0.1354	0.2496	0.2500	0.2500
HeaviSine	$EAON^*$	0.0496	0.0480	0.1366	0.2596	0.5000	1.2500
Blocks	$EAFN$	0.0485	0.0494	0.2014	0.2500	0.2500	0.2500
Blocks	$EAON$	0.0483	0.0429	0.1871	0.2500	0.2500	0.2500
Blocks	$EAON^*$	0.0483	0.0429	0.1893	0.2600	0.5000	1.2500

4.3 Model (4.3) - the case $M = 2n$ by mixing two orthonormal basis

Consider now the setting of model (4.3) with $\gamma_1 = 0.5$, $\gamma_2 = 0.2$, $\sigma = 0.045$, $N = 25$ and an equidistant grid of design points $t_1, ..., t_n$ given by $t_j = \frac{j}{n}$, $j = 1, \ldots, n$ with $n = 128$. For the signals f_1 and f_2 we took the test functions displayed in Figure 4(a) and 4(b). Obviously, the signal f_1 has a sparse representation in a Haar basis while the signal f_2 has a sparse representation in a Fourier basis. Thus, this suggests to construct a dictionary by mixing two orthonormal basis. More precisely, we construct a $n \times n$ orthogonal matrix $\mathbf{G}^1$ using the Haar basis and a $n \times n$ orthogonal matrix $\mathbf{G}^2$ using a Fourier basis (cosine and sine at various frequencies) at the design points. Then, we form the $n \times M$ design matrix $\mathbf{G} = [\mathbf{G}^1\ \mathbf{G}^2]$ with $M = 2n$. The CGL estimator $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$ is computed by the minimization procedure (2.7) using the Matlab package *minConf* of [Schmidt et al., 2008].

In Figures 5(a) and 5(b), we display the eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$ as an estimator of $\mathbf{F}_1$, and the eigenvector associated to the second largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$ as an estimator of $\mathbf{F}_2$. Note that these estimators behaves poorly. The estimation considerably improves by taking the FCGL estimator $\widehat{\mathbf{\Sigma}}_{\hat{J}}$ defined in (4.4). Figures 6(a) and 6(b) illustrate the very good performance of the eigenvectors associated to the largest eigenvalue and second largest eigenvalue of the matrix $\widehat{\mathbf{\Sigma}}_{\hat{J}}$ as estimators of $\mathbf{F}_1$ and $\mathbf{F}_2$.

Finally, to illustrate the benefits of mixing two orthonormal basis, we also display in Figures 7 and 8 the estimation of $\mathbf{F}_1$ and $\mathbf{F}_2$ when computing the matrix $\widehat{\mathbf{\Sigma}}_{\hat{J}}$ by using either only the Haar basis (i.e. $\mathbf{G} = \mathbf{G}^1$ and $M = n$) or only the Fourier basis (i.e. $\mathbf{G} = \mathbf{G}^1$ and $M = n$). The results are clearly much worse and not satisfactory.

4.4 Model (4.2) - case of non equispaced design points such that $n < M$

Let us now return to the setting of model (4.2). The test functions f are either the signal HeaviSine and or the signal Blocks. We also use the Symmlet 8 and Haar wavelet basis for the HeaviSine and Blocks functions respectively. However, we now choose to take a setting where the number of design points n is smaller than the size M of the dictionary. Taking $n < M$, the location points are given by a subset $\{t_1, ..., t_n\} \subset \{\frac{k}{M} : k = 1, ..., M\}$ of size n, such that the design matrix $\mathbf{G}$ is an $n \times M$ matrix (using either the Symmlet 8 and Haar basis). For a fixed value of n, the subset $\{t_1, ..., t_n\}$ is chosen by taking the first n points obtained from a random permutation of the elements of the set $\{\frac{1}{M}, \frac{2}{M}, ..., 1\}$. Figures 9 and 10 present the results obtained for a particular simulated sample of size $N = 25$ according to (4.1), with $n = 90$, $M = 128$, $\sigma = 0.02$, $\gamma = 0.5$ and with f being either the function HeaviSine or the function Blocks. It can be observed in Figures 9(a) and 9(b) that, as expected in this high dimensional setting ($N < n$), the empirical eigenvector of $\widetilde{\mathbf{S}}$ associated to its largest empirical eigenvalue are noisy versions of $\mathbf{F}$. As explained previously, the CGL estimator $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$ is computed by the minimization procedure (2.7) using the Matlab package *minConf* of [Schmidt et al., 2008]. In Figures 10(a) and 10(b) is shown the eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$ as an estimator of $\mathbf{F}$. Note that this estimator is quite noisy. Again, the eigenvector associated to the largest eigenvalue of the matrix $\widehat{\mathbf{\Sigma}}_{\widehat{f}}$ defined in (4.4) is much a better estimator of $\mathbf{F}$. This is illustrated in Figures 11(a) and 11(b). To compare the accuracy of the estimators for different simulated samples, we compute the values of $EAFN$, $EAON$ and $EAON^*$ with fixed values of $\sigma = 0.05$, $M = 128$, $N = 40$, $P = 50$ for different values of the number of design points n. For all the values of n considered, the design points $t_1, ..., t_n$ are selected as the first n points obtained from the same random permutation of the elements of the set $\{\frac{1}{M}, \frac{2}{M}, ..., 1\}$. The chosen subset $\{t_1, ..., t_n\}$ is used for all the P iterations needed in the computation of the empirical averages (fixed design over the iterations). Figure 12 shows the values of $EAFN$, $EAON$ and $EAON^*$ obtained for each value of n for both signals HeaviSine and Blocks. It can be observed that the values of the empirical averages $EAON$ and $EAON^*$ are much smaller than its corresponding values of $EAFN$ as expected. We can remark that, when n increases, the values of $EAFN$, $EAON$ and $EAON^*$ first increase and then decrease, and the change of monotony occurs when $n > N$. Note that the case $n = M = 128$ is included in these results.

Orthonormal case - Model (4.2)

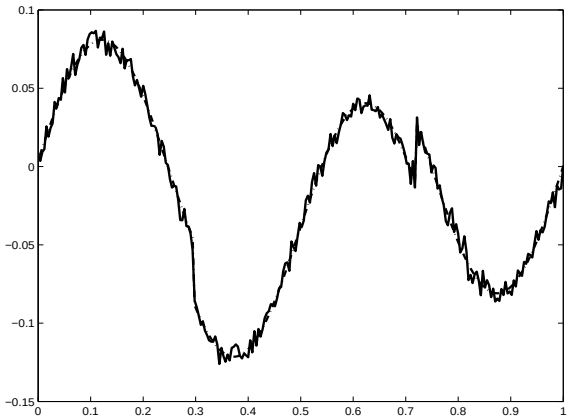

Figure 1(a). Signal HeaviSine and Eigenvector associated to the largest eigenvalue of $\widetilde{\mathbf{S}}$

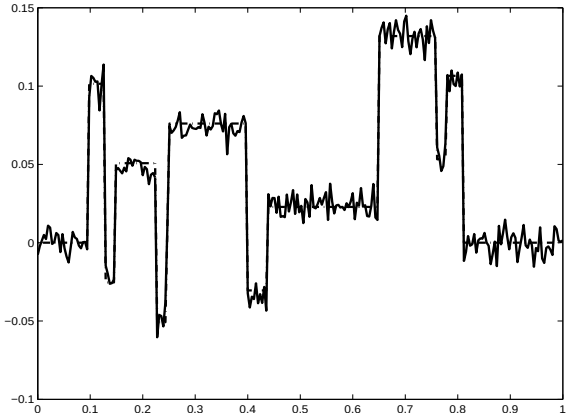

Figure 1(b). Signal Blocks and Eigenvector associated to the largest eigenvalue of $\widetilde{\mathbf{S}}$

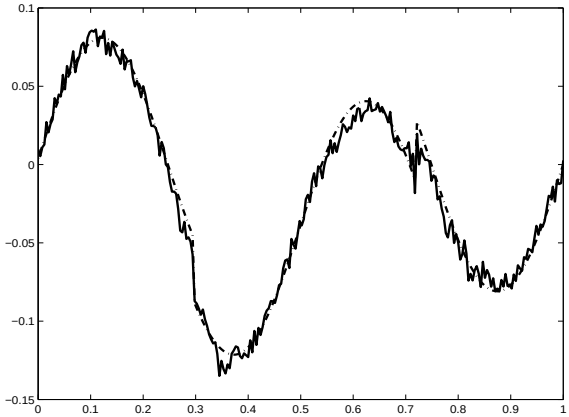

Figure 2(a). Signal HeaviSine and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$

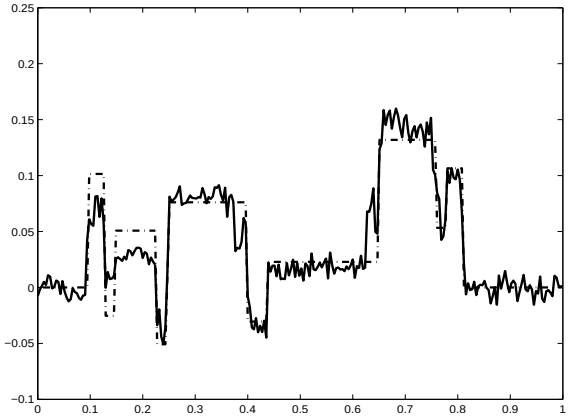

Figure 2(b). Signal Blocks and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$

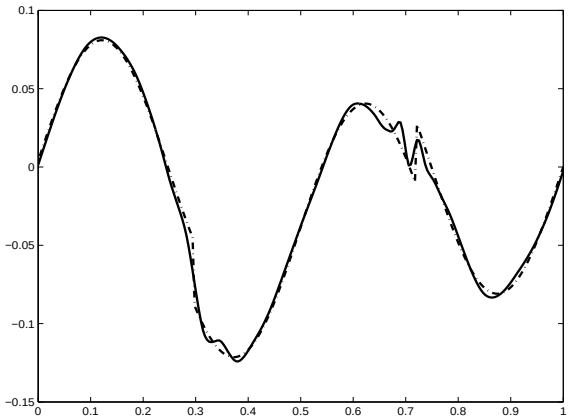

Figure 3(a). Signal HeaviSine and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$

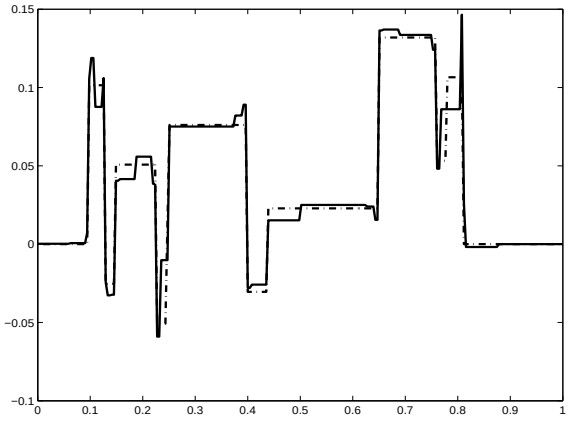

Figure 3(b). Signal Blocks and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$

Case $M = 2n$ (Haar + Fourier basis)

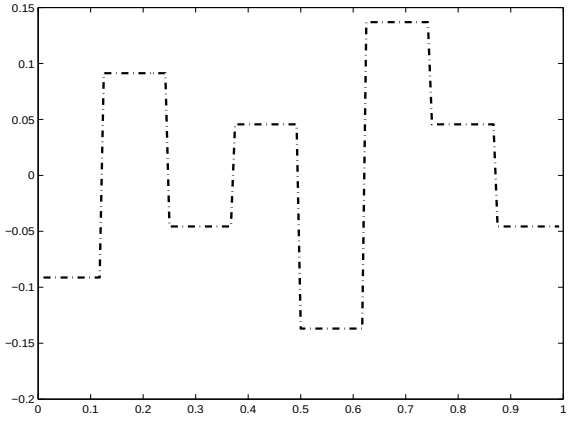

Figure 4(a). Signal F_1

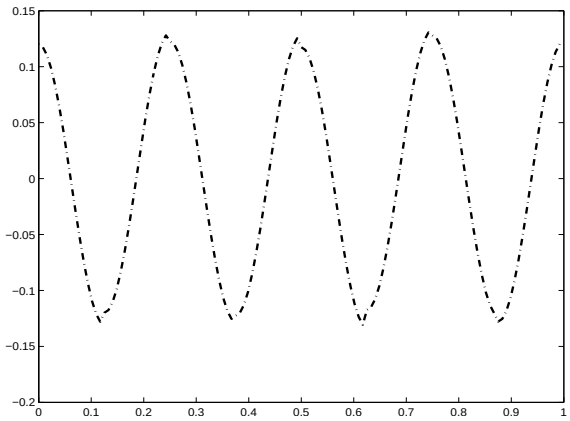

Figure 4(b). Signal F_2

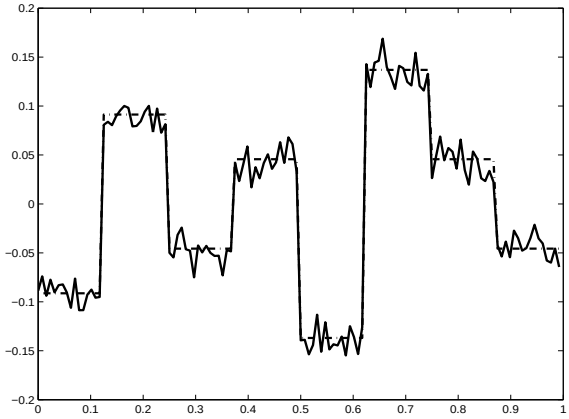

Figure 5(a). Signal F_1 and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$

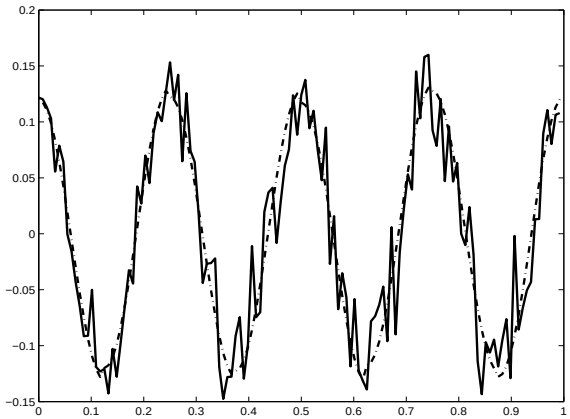

Figure 5(b). Signal F_2 and Eigenvector associated to the second largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$

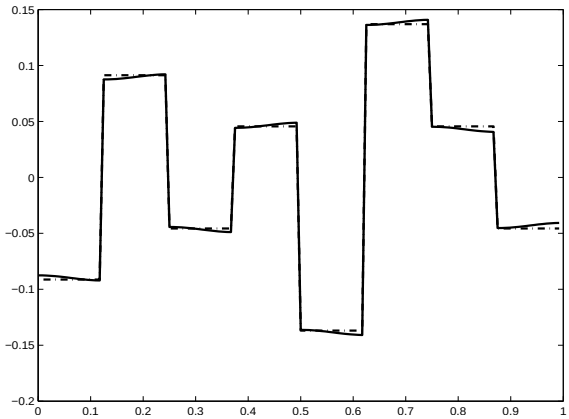

Figure 6(b). Signal F_1 and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{f}}$ with $\mathbf{G} = [\mathbf{G}^1\ \mathbf{G}^2]$

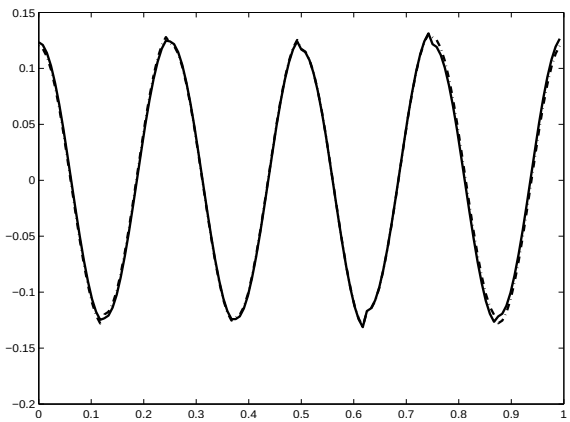

Figure 6(b). Signal F_2 and Eigenvector associated to the second largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{f}}$ with $\mathbf{G} = [\mathbf{G}^1\ \mathbf{G}^2]$

Orthonormal case $M = n$ (Haar)

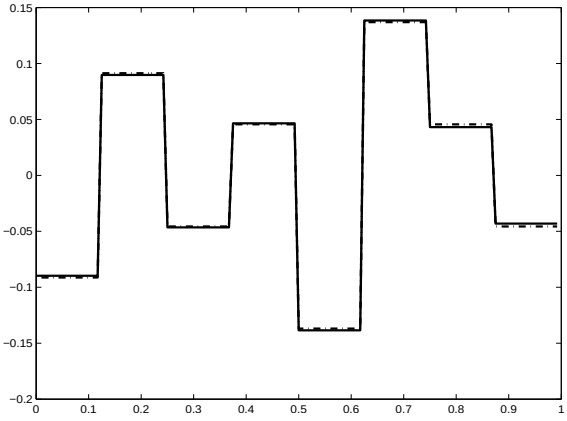

Figure 7(a). Signal F_1 and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ with $\mathbf{G} = \mathbf{G}^1$

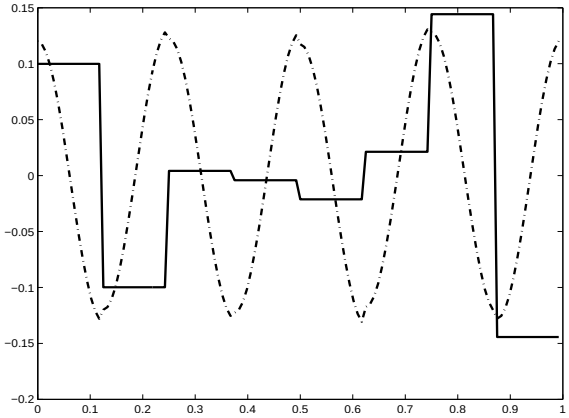

Figure 7(b). Signal F_2 and Eigenvector associated to the second largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ with $\mathbf{G} = \mathbf{G}^1$

Orthonormal case $M = n$ (Fourier)

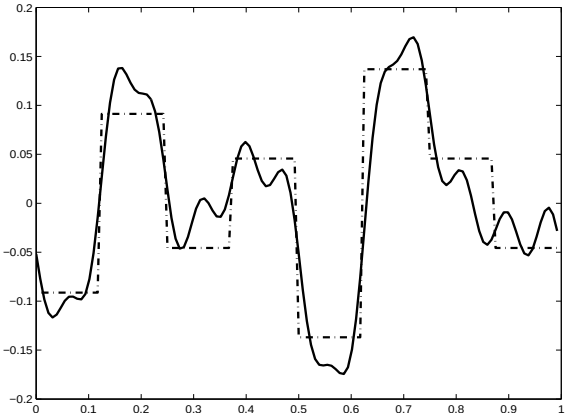

Figure 8(a). Signal F_1 and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ with $\mathbf{G} = \mathbf{G}^2$

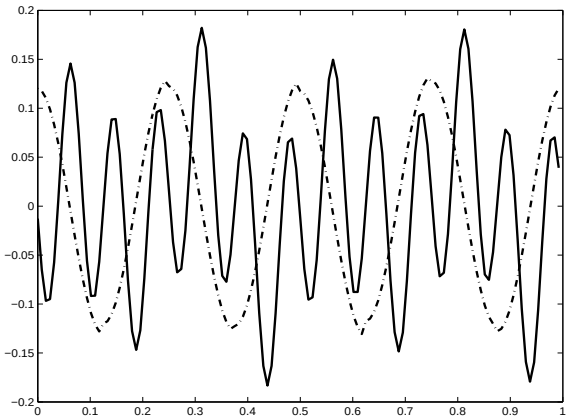

Figure 8(b). Signal F_2 and Eigenvector associated to the second largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ with $\mathbf{G} = \mathbf{G}^2$

Non equi-spaced points with $n < M$

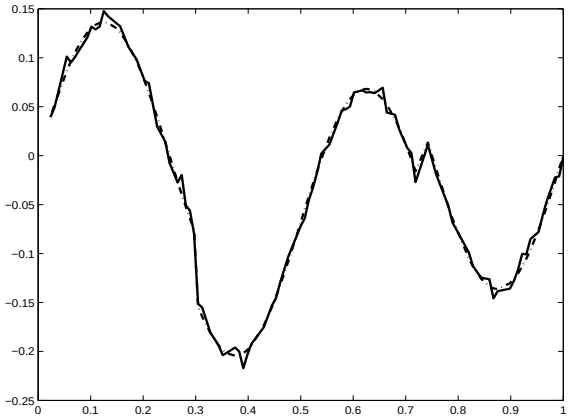

Figure 9(a). Signal HeaviSine and Eigenvector associated to the largest eigenvalue of $\widetilde{\mathbf{S}}$

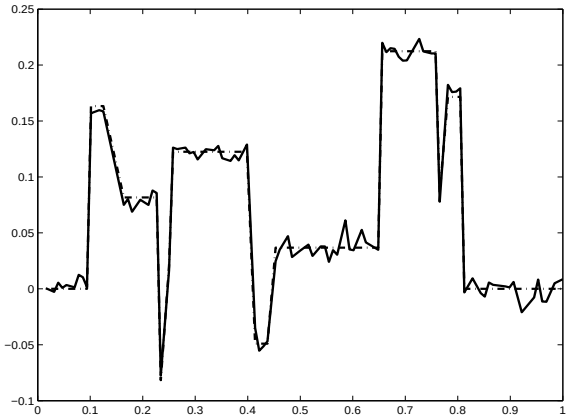

Figure 9(b). Signal Blocks and Eigenvector associated to the largest eigenvalue of $\widetilde{\mathbf{S}}$

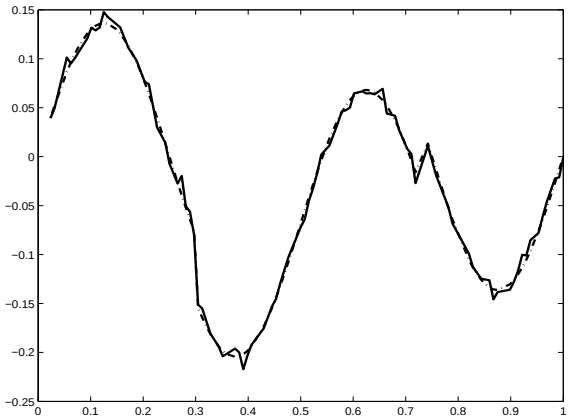

Figure 10(a). Signal HeaviSine and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$

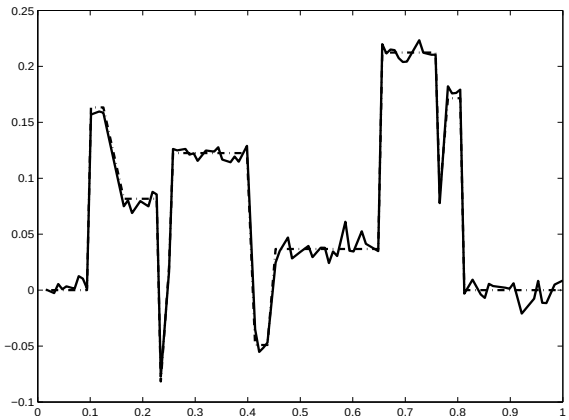

Figure 10(b). Signal Blocks and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$

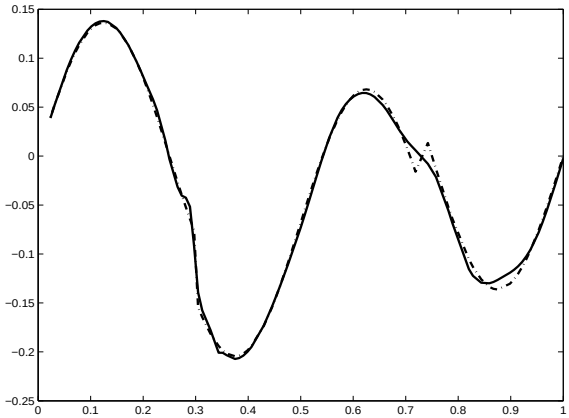

Figure 11(a). Signal HeaviSine and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$

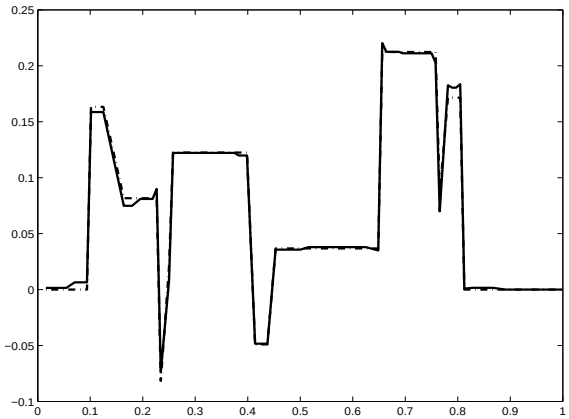

Figure 11(b). Signal Blocks and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$

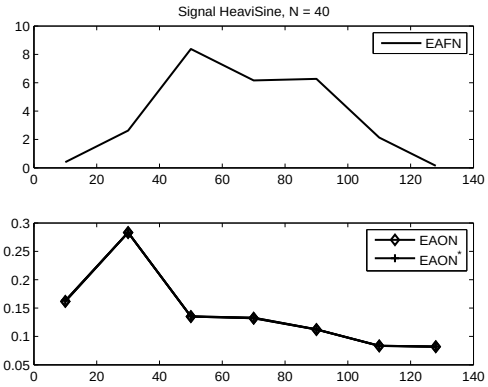

Figure 12(a). Values of $EAFN$, $EAON$ and $EAON^*$ for Signal HeaviSine as a function of n

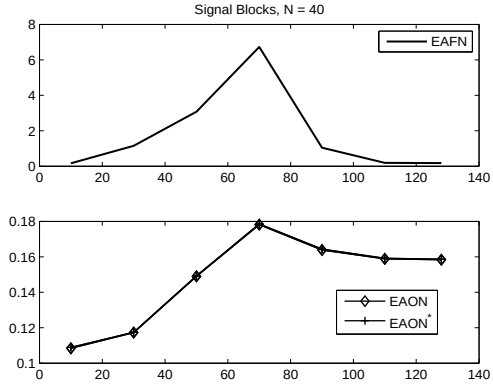

Figure 12(b). Values of $EAFN$, $EAON$ and $EAON^*$ for Signal Blocks as a function of n

A

A.1 Notations

First let us introduce some notations and properties that will be used throughout this Appendix. The vectorization of a $p \times q$ matrix $\mathbf{A} = (a_{ij})_{1\leq i\leq p, 1\leq j\leq q}$ is the $pq \times 1$ column vector denoted by $vec\,(\mathbf{A})$, obtain by stacking the columns of the matrix $\mathbf{A}$ on top of one another. That is $vec(\mathbf{A}) = [a_{11}, ..., a_{p1}, a_{12}, ..., a_{p2}, ..., a_{1q}, ..., a_{pq}]^\top$. If $\mathbf{A} = (a_{ij})_{1\leq i\leq k, 1\leq j\leq n}$ is a $k \times n$ matrix and $\mathbf{B} = (b_{ij})_{1\leq i\leq p, 1\leq j\leq q}$ is a $p \times q$ matrix, then the Kronecker product of the two matrices, denoted by $\mathbf{A} \otimes \mathbf{B}$, is the $kp \times nq$ block matrix

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

In what follows, we repeatedly use the fact that the Frobenius norm is invariant by the vec operation meaning that

$$\|\mathbf{A}\|_F^2 = \|vec\,(\mathbf{A})\,\|_{\ell_2}^2, \tag{A.1}$$

and the properties that

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

and

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}, \tag{A.3}$$

provided the above matrix products are compatible.

A.2 Proof of Proposition 1

Lemma 1 *Let $\widehat{\mathbf{\Psi}} = \widehat{\mathbf{\Psi}}_\lambda$ denotes the solution of (2.7). Then, for $k = 1, \ldots, M$*

$$\begin{aligned} \left[(\mathbf{G}\otimes\mathbf{G})^\top\left(vec(\widetilde{\mathbf{S}}) - (\mathbf{G}\otimes\mathbf{G})vec(\widehat{\mathbf{\Psi}})\right)\right]^k &= \lambda\gamma_k \frac{\widehat{\mathbf{\Psi}}_k}{\|\widehat{\mathbf{\Psi}}_k\|_{\ell_2}} \quad \text{if} \quad \mathbf{\Psi}_k \neq 0 \\ \left\|\left[(\mathbf{G}\otimes\mathbf{G})^\top\left(vec(\widetilde{\mathbf{S}}) - (\mathbf{G}\otimes\mathbf{G})vec(\widehat{\mathbf{\Psi}})\right)\right]^k\right\|_{\ell_2} &\leq \lambda\gamma_k \quad \text{if} \quad \widehat{\mathbf{\Psi}}_k = 0 \end{aligned}$$

where $\widehat{\mathbf{\Psi}}_k$ denotes the k-th column of the matrix $\widehat{\mathbf{\Psi}}$ and the notation $[\beta]^k$ denotes the vector $(\beta_{k,m})_{m=1,\ldots,M}$ in $\mathbb{R}^M$ for a vector $\beta = (\beta_{k,m})_{k,m=1,\ldots,M} \in \mathbb{R}^{M^2}$.

Proof of Lemma 1 For $\mathbf{\Psi} \in \mathbb{R}^{M\times M}$ define

$$L(\mathbf{\Psi}) = \left\|\widetilde{\mathbf{S}} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 = \left\|vec(\widetilde{\mathbf{S}}) - (\mathbf{G}\otimes\mathbf{G})vec(\mathbf{\Psi})\right\|_{\ell_2}^2,$$

and remark that $\widehat{\mathbf{\Psi}}$ is the solution of the convex optimization problem

$$\widehat{\mathbf{\Psi}} = \underset{\mathbf{\Psi}\in\mathcal{S}_M}{\operatorname{argmin}} \left\{ L(\mathbf{\Psi}) + 2\lambda \sum_{k=1}^M \gamma_k \sqrt{\sum_{m=1}^M \Psi_{mk}^2} \right\}.$$

It follows from standard arguments in convex analysis (see e.g. [Boyd and Vandenberghe, 2004]), that $\widehat{\mathbf{\Psi}}$ is a solution of the above minimization problem if and only if

$$-\nabla L(\widehat{\mathbf{\Psi}}) \in 2\lambda\partial\left(\sum_{k=1}^M \gamma_k \sqrt{\sum_{m=1}^M \hat{\Psi}_{mk}^2}\right)$$

where $\nabla L(\widehat{\mathbf{\Psi}})$ denotes the gradient of L at $\widehat{\mathbf{\Psi}}$ and ∂ denotes the subdifferential given by

$$\partial\left(\sum_{k=1}^M \gamma_k \sqrt{\sum_{m=1}^M \Psi_{mk}^2}\right) = \left\{\mathbf{\Theta}\in\mathbb{R}^{M\times M} : \mathbf{\Theta}_k = \gamma_k \frac{\mathbf{\Psi}_k}{\|\mathbf{\Psi}_k\|_{\ell_2}} \text{ if } \mathbf{\Psi}_k \neq 0, \|\mathbf{\Theta}_k\|_{\ell_2} \leq \gamma_k \text{ if } \mathbf{\Psi}_k = 0\right\}$$

where $\mathbf{\Theta}_k$ denotes the k-th column of $\mathbf{\Theta} \in \mathbb{R}^{M\times M}$ which completes the proof. □

Now, let $\mathbf{\Psi} \in \mathcal{S}_M$ with $M = n$ and suppose that $\mathbf{G}^\top\mathbf{G} = \mathbf{I}_n$. Let $\mathbf{Y} = (\mathbf{Y}_{mk})_{1\leq m,k\leq M} = \mathbf{G}^\top\widetilde{\mathbf{S}}\mathbf{G}$ and remark that $vec(\mathbf{Y}) = (\mathbf{G}\otimes\mathbf{G})^\top vec(\widetilde{\mathbf{S}})$. Then, by using Lemma 1 and the fact that $\mathbf{G}^\top\mathbf{G} = \mathbf{I}_n$ implies that $(\mathbf{G}\otimes\mathbf{G})^\top(\mathbf{G}\otimes\mathbf{G}) = \mathbf{I}_{n^2}$, it follows that $\widehat{\mathbf{\Psi}} = \widehat{\mathbf{\Psi}}_\lambda$ satisfies for $k = 1, \ldots, M$ the following equations

$$\widehat{\mathbf{\Psi}}_k\left(1 + \frac{\lambda\gamma_k}{\sqrt{\sum_{m=1}^M \widehat{\Psi}_{mk}^2}}\right) = \mathbf{Y}_k \text{ for all } \widehat{\mathbf{\Psi}}_k \neq 0,$$

and

$$\sqrt{\sum_{m=1}^{M} \mathbf{Y}_{mk}^2} \leq \lambda\gamma_k \text{ for all } \widehat{\mathbf{\Psi}}_k = 0.$$

where $\widehat{\mathbf{\Psi}}_k = (\widehat{\Psi}_{mk})_{1\leq m\leq M} \in \mathbb{R}^M$ and $\mathbf{Y}_k = (\mathbf{Y}_{mk})_{1\leq m\leq M} \in \mathbb{R}^M$, which implies that the solution is given by

$$\widehat{\Psi}_{mk} = \begin{cases} 0 & \text{if} \quad \sqrt{\sum_{m=1}^{M} \mathbf{Y}_{mk}^2} \leq \lambda\gamma_k \\ Y_{mk}\left(1 - \frac{\lambda\gamma_k}{\sqrt{\sum_{j=1}^{M} \mathbf{Y}_{jk}^2}}\right) & \text{if} \quad \sqrt{\sum_{m=1}^{M} \mathbf{Y}_{mk}^2} > \lambda\gamma_k \end{cases}$$

which completes the proof of Proposition 1. □

A.3 Proof of Proposition 2

First suppose that X is Gaussian. Then, remark that for $Z = \|\mathbf{X}\|_{\ell_2}$, one has that $\|Z\|_{\psi_2} < +\infty$ which implies that $\|Z\|_{\psi_2} = \|Z^2\|_{\psi_1}^{1/2}$. Since $Z^2 = \sum_{i=1}^{n} |X(t_i)|^2$ it follows that

$$\|Z^2\|_{\psi_1} \leq \sum_{i=1}^{n} \|Z_i^2\|_{\psi_1} = \sum_{i=1}^{n} \|Z_i\|_{\psi_2}^2 = \sum_{i=1}^{n} \mathbf{\Sigma}_{ii} \|\mathbf{\Sigma}_{ii}^{-1/2} Z_i\|_{\psi_2}^2,$$

where $Z_i = X(t_i), i = 1, \ldots, n$ and $\mathbf{\Sigma}_{ii}$ denotes the ith diagonal element of $\mathbf{\Sigma}$. Then, the result follows by noticing that $\|Y\|_{\psi_2} \leq \sqrt{8/3}$ if $Y \sim N(0, 1)$. The proof for the case where X is such that $\|Z\|_{\psi_2} < +\infty$ and there exists a constant C_1 such that $\|\mathbf{\Sigma}_{ii}^{-1/2} Z_i\|_{\psi_2} \leq C_1$ for all $i = 1, \ldots, n$ follows from the same arguments.

Now, consider the case where X is a bounded process. Since there exists a constant $R > 0$ such that for all $t \in \mathbb{T}$, $|X(t)| \leq R$, it follows that for $Z = \|\mathbf{X}\|_{\ell_2}$ then $Z \leq \sqrt{n}R$ which implies that for any $\alpha \geq 1$, $\|Z\|_{\psi_\alpha} \leq \sqrt{n}R(\log 2)^{-1/\alpha}$, (by definition of the norm $\|Z\|_{\psi_\alpha}$) which completes the proof of Proposition 2. □

A.4 Proof of Proposition 4

Under the assumption that $X = X^0$, it follows that $\mathbf{\Sigma} = \mathbf{G}\mathbf{\Psi}^*\mathbf{G}^\top$ with $\mathbf{\Psi}^* = \mathbb{E}\left(\mathbf{a}\mathbf{a}^\top\right)$, where $\mathbf{a}$ is the random vector of $\mathbb{R}^M$ with $\mathbf{a}_m = a_m$ for $m \in J^*$ and $\mathbf{a}_m = 0$ for $m \notin J^*$. Then, define the random vector $\mathbf{a}_{J^*} \in \mathbb{R}^{J^*}$ whose coordinates are the random coefficients a_m for $m \in J^*$. Let $\mathbf{\Psi}_{J^*} = \mathbb{E}\left(\mathbf{a}_{J^*}\mathbf{a}_{J^*}^\top\right)$. Note that $\mathbf{\Sigma} = \mathbf{G}_{J^*}\mathbf{\Psi}_{J^*}\mathbf{G}_{J^*}^\top$ and $\mathbf{S} = \mathbf{G}_{J^*}\widehat{\mathbf{\Psi}}_{J^*}\mathbf{G}_{J^*}^\top$, with $\widehat{\mathbf{\Psi}}_{J^*} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{a}_{J^*}^i (\mathbf{a}_{J^*}^i)^\top$, where $\mathbf{a}_{J^*}^i \in \mathbb{R}^{J^*}$ denotes the random vector whose coordinates are the random coefficients a_m^i for $m \in J^*$ such that $X_i(t) = \sum_{m\in J^*} a_m^i g_m(t),\ t \in \mathbb{T}$.

Therefore, $\widehat{\mathbf{\Psi}}_{J^*}$ is a sample covariance matrix of size $s_* \times s_*$ and we can control its deviation in operator norm from $\widehat{\mathbf{\Psi}}_{J^*}$ by using Proposition 3. For this we simply have to verify conditions similar to **(A1)** and **(A2)** in Assumption 2 for the random vector $\mathbf{a}_{J^*} = (\mathbf{G}_{J^*}^\top\mathbf{G}_{J^*})^{-1}\mathbf{G}_{J^*}^\top\mathbf{X} \in \mathbb{R}^{s_*}$. First, let $\beta \in \mathbb{R}^{s_*}$ with $\|\beta\|_{\ell_2} = 1$. Then, remark that $\mathbf{a}_{J^*}^\top\beta = \mathbf{X}^\top\tilde{\beta}$ with $\tilde{\beta} = \mathbf{G}_{J^*}\left(\mathbf{G}_{J^*}^\top\mathbf{G}_{J^*}\right)^{-1}\beta$. Since $\|\tilde{\beta}\|_{\ell_2} \leq \left(\rho_{\min}\left(\mathbf{G}_{J^*}^\top\mathbf{G}_{J^*}\right)\right)^{-1/2}$ and using that X satisfies Assumption 2 it follows that

$$\left(\mathbb{E}|\mathbf{a}_{J^*}^\top\beta|^4\right)^{1/4} \leq \rho\left(\mathbf{\Sigma}\right)\rho_{\min}^{-1/2}\left(\mathbf{G}_{J^*}^\top\mathbf{G}_{J^*}\right). \tag{A.4}$$

Now let $\tilde{Z} = \|\mathbf{a}_{J^*}\|_{\ell_2} \leq \rho_{\min}^{-1/2}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\|\mathbf{X}\|_{\ell_2}$. Given our assumptions on X it follows that there exists $\alpha \geq 1$ such that

$$\|\tilde{Z}\|_{\psi_\alpha} \leq \rho_{\min}^{-1/2}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\|Z\|_{\psi_\alpha} < +\infty, \tag{A.5}$$

where $Z = \|\mathbf{X}\|_{\ell_2}$. Hence, using the relations (A.4) and (A.5), and Proposition 3 (with $\mathbf{a}_{J^*}$ instead of $\mathbf{X}$), it follows that there exists a universal constant $\delta_* > 0$ such that for all $x > 0$,

$$\mathbb{P}\left(\left\|\widehat{\mathbf{\Psi}}_{J^*} - \mathbf{\Psi}_{J^*}\right\|_2 \geqslant \tilde{\tau}_{d^*,N,s_*,1}x\right) \leqslant \exp\left(-(\delta_*^{-1}x)^{\frac{\alpha}{2+\alpha}}\right),$$

where $\tilde{\tau}_{d^*,N,s_*,1} = \max(\tilde{A}^2_{d^*,N,s_*,1}, \tilde{B}_{d^*,N,s_*,1})$, with $\tilde{A}_{d^*,N,s_*,1} = \|\tilde{Z}\|_{\psi_\alpha}\frac{\sqrt{\log d^*}(\log N)^{1/\alpha}}{\sqrt{N}}$, $\tilde{B}_{d^*,N,s_*,1} = \frac{\rho^2(\mathbf{\Sigma})\rho_{\min}^{-1}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}{\sqrt{N}} + \|\mathbf{\Psi}_{J^*}\|_2^{1/2}\tilde{A}_{d^*,N,s_*,1}$ and $d^* = \min(N, s_*)$. Then, using the inequality $\|\mathbf{S} - \mathbf{\Sigma}\|_2 \leq \rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\|\widehat{\mathbf{\Psi}}_{J^*} - \mathbf{\Psi}_{J^*}\|_2$, it follows that

$$\begin{aligned}
& \mathbb{P}\left(\|\mathbf{S} - \mathbf{\Sigma}\|_2 \geq \rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\tilde{\tau}_{d^*,N,s_*,1}x\right) \\
\leq\ & \mathbb{P}\left(\rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\left\|\widehat{\mathbf{\Psi}}_{J^*} - \mathbf{\Psi}_{J^*}\right\|_2 \geqslant \rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\tilde{\tau}_{d^*,N,s_*,1}x\right) \\
=\ & \mathbb{P}\left(\left\|\widehat{\mathbf{\Psi}}_{J^*} - \mathbf{\Psi}_{J^*}\right\|_2 \geqslant \tilde{\tau}_{d^*,N,s_*,1}x\right) \\
\leqslant\ & \exp\left(-(\delta_*^{-1}x)^{\frac{\alpha}{2+\alpha}}\right).
\end{aligned}$$

Hence, the result follows with

$$\begin{aligned}
\tilde{\tau}_{N,s_*} &= \rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\tilde{\tau}_{d^*,N,s_*,1} \\
&= \max(\rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\tilde{A}^2_{d^*,N,s_*,1}, \rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\tilde{B}_{d^*,N,s_*,1}) \\
&= \max(\tilde{A}^2_{d^*,N,s_*}, \tilde{B}_{d^*,N,s_*}),
\end{aligned}$$

where $\tilde{A}_{d^*,N,s_*} = \rho_{\max}^{1/2}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\|\tilde{Z}\|_{\psi_\alpha}\frac{\sqrt{\log d^*}(\log N)^{1/\alpha}}{\sqrt{N}}$ and, using the inequality

$$\|\mathbf{\Psi}_{J^*}\|_2 = \left\|\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)^{-1}\mathbf{G}_{J^*}^\top \mathbf{\Sigma}\mathbf{G}_{J^*}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)^{-1}\right\|_2 \leq \rho_{\min}^{-1}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\|\mathbf{\Sigma}\|_2,$$

$\tilde{B}_{d^*,N,s_*} = \left(\frac{\rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}{\rho_{\min}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}\right)\frac{\rho^2(\mathbf{\Sigma})}{\sqrt{N}} + \left(\frac{\rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}{\rho_{\min}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}\right)^{1/2}\|\mathbf{\Sigma}\|_2^{1/2}\tilde{A}_{d^*,N,s_*}$.

A.5 Proof of Theorem 1

Let us first prove the following lemmas.

Lemma 2 *Let $\mathcal{E}_1, ..., \mathcal{E}_N$ be independent copies of a second order Gaussian process $\mathcal{E}$ with zero mean. Let $\mathbf{W} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{W}_i$ with*

$$\mathbf{W}_i = \mathcal{E}_i\mathcal{E}_i^\top \in \mathbb{R}^{n\times n} \text{ and } \mathcal{E}_i = \left(\mathcal{E}_i(t_1), ..., \mathcal{E}_i(t_n)\right)^\top,\ i = 1, \ldots, N.$$

Suppose that $\mathbf{\Sigma}_{noise} = \mathbb{E}(\mathbf{W}_1)$ *is positive-definite. For* $1 \leq k \leq M$*, let* η_k *be the* k*-th column of the matrix* $\mathbf{G}^\top \mathbf{W} \mathbf{G}$*. Then, for any* $x > 0$*,*

$$\mathbb{P}\left(\|\eta_k\|_{\ell_2} \geq \|\mathbf{G}_k\|_{\ell_2} \sqrt{\rho_{\max}(\mathbf{G}\mathbf{G}^\top)} \|\mathbf{\Sigma}_{noise}\|_2 \left(1 + \sqrt{\frac{n}{N}} + \sqrt{\frac{2x}{N}}\right)^2 \right) \leq \exp(-x).$$

Proof of Lemma 2: by definition one has that $\|\eta_k\|_{\ell_2}^2 = \mathbf{G}_k^\top \mathbf{W}\mathbf{G}\mathbf{G}^\top \mathbf{W}\mathbf{G}_k$ where $\mathbf{G}_k$ denotes the k-th column of $\mathbf{G}$. Hence

$$\|\eta_k\|_{\ell_2}^2 \leq \|\mathbf{G}_k\|_{\ell_2}^2 \rho_{\max}(\mathbf{G}\mathbf{G}^\top) \|\mathbf{W}\|_2^2. \tag{A.6}$$

Using the assumption that $\mathbf{\Sigma}_{noise}$ is positive-definite define the random vectors $Z_i = \mathbf{\Sigma}_{noise}^{-1/2} \mathcal{E}_i, i = 1, \ldots, n$. Note that the Z_i's are i.i.d. Gaussian vectors in $\mathbb{R}^n$ with zero mean and covariance matrix the identity. Then, define the $N \times n$ matrix

$$\Gamma = \frac{1}{\sqrt{N}} \begin{pmatrix} Z_1^\top \\ \vdots \\ Z_N^\top \end{pmatrix}.$$

Since Γ is a matrix with i.i.d. entries following a Gaussian distribution with zero mean and variance $1/N$, it follows from the arguments in the proof of Theorem II.13 in [Davidson and Szarek, 2001] that for any $x > 0$

$$\mathbb{P}\left(\|\mathbf{\Gamma}^\top \mathbf{\Gamma}\|_2 \geq \left(1 + \sqrt{\frac{n}{N}} + \sqrt{\frac{2x}{N}}\right)^2 \right) \leq \exp(-x). \tag{A.7}$$

Now, since $\mathbf{W} = \mathbf{\Sigma}_{noise}^{1/2} \Gamma^\top \Gamma \mathbf{\Sigma}_{noise}^{1/2}$ it follows that $\|\mathbf{W}\|_2 \leq \|\mathbf{\Sigma}_{noise}\|_2 \|\Gamma^\top \Gamma\|_2$. Hence, inequality (A.7) implies that for any $x > 0$

$$\mathbb{P}\left(\|\mathbf{W}\|_2 \geq \|\mathbf{\Sigma}_{noise}\|_2 \left(1 + \sqrt{\frac{n}{N}} + \sqrt{\frac{2x}{N}}\right)^2 \right) \leq \exp(-x),$$

and the result finally follows from inequality (A.6). □

Lemma 3 *Let* $1 \leq s \leq \min(n, M)$ *and suppose that Assumption 1 holds for some* $c_0 > 0$*. Let* $J \subset \{1, \ldots, M\}$ *be a subset of indices of cardinality* $|J| \leq s$*. Let* $\mathbf{\Delta} \in \mathcal{S}_M$ *and suppose that*

$$\sum_{k \in J^c} \|\mathbf{\Delta}_k\|_{\ell_2} \leq c_0 \sum_{k \in J} \|\mathbf{\Delta}_k\|_{\ell_2},$$

where $\mathbf{\Delta}_k$ *denotes the* k*-th column of* $\mathbf{\Delta}$*. Let*

$$\kappa_{s,c_0} = \left(\rho_{\min}(s)^2 - c_0 \theta(\mathbf{G}) \rho_{\max}(\mathbf{G}^\top \mathbf{G}) s \right)^{1/2}.$$

Then,

$$\left\|\mathbf{G}\boldsymbol{\Delta}\mathbf{G}^\top\right\|_F^2 \geq \kappa_{s,c_0}^2 \left\|\boldsymbol{\Delta}_J\right\|_F^2,$$

where $\boldsymbol{\Delta}_J$ denotes the $M \times M$ matrix obtained by setting to zero the rows and columns of $\boldsymbol{\Delta}$ whose indices are not in J.

Proof of Lemma 3: first let us introduce some notations. For $\boldsymbol{\Delta} \in \mathcal{S}_M$ and $J \subset \{1, \ldots, M\}$, then $\boldsymbol{\Delta}_{J^c}$ denotes the $M \times M$ matrix obtained by setting to zero the rows and columns of $\boldsymbol{\Delta}$ whose indices are not in the complementary J^c of J. Now, remark that

$$\begin{aligned}\left\|\mathbf{G}\boldsymbol{\Delta}\mathbf{G}^\top\right\|_F^2 &= \left\|\mathbf{G}\boldsymbol{\Delta}_J\mathbf{G}^\top\right\|_F^2 + \left\|\mathbf{G}\boldsymbol{\Delta}_{J^c}\mathbf{G}^\top\right\|_F^2 + 2tr\left(\mathbf{G}\boldsymbol{\Delta}_J\mathbf{G}^\top\mathbf{G}\boldsymbol{\Delta}_{J^c}\mathbf{G}^\top\right) \\ &\geq \left\|\mathbf{G}\boldsymbol{\Delta}_J\mathbf{G}^\top\right\|_F^2 + 2tr\left(\mathbf{G}\boldsymbol{\Delta}_J\mathbf{G}^\top\mathbf{G}\boldsymbol{\Delta}_{J^c}\mathbf{G}^\top\right). \qquad \text{(A.8)}\end{aligned}$$

Let $\mathbf{A} = \mathbf{G}\boldsymbol{\Delta}_J\mathbf{G}^\top$ and $\mathbf{B} = \mathbf{G}\boldsymbol{\Delta}_{J^c}\mathbf{G}^\top$. Using that $tr\left(\mathbf{A}^\top\mathbf{B}\right) = vec(\mathbf{A})^\top vec(\mathbf{B})$ and the properties (A.1) and (A.3) it follows that

$$tr\left(\mathbf{G}\boldsymbol{\Delta}_J\mathbf{G}^\top\mathbf{G}\boldsymbol{\Delta}_{J^c}\mathbf{G}^\top\right) = vec(\boldsymbol{\Delta}_J)^\top\left(\mathbf{G}^\top\mathbf{G}\otimes\mathbf{G}^\top\mathbf{G}\right)vec(\boldsymbol{\Delta}_{J^c}). \qquad \text{(A.9)}$$

Let $\mathbf{C} = \mathbf{G}^\top\mathbf{G}\otimes\mathbf{G}^\top\mathbf{G}$ and note that $\mathbf{C}$ is a $M^2 \times M^2$ matrix whose elements can be written in the form of $M \times M$ block matrices given by

$$\mathbf{C}_{ij} = (\mathbf{G}^\top\mathbf{G})_{ij}\mathbf{G}^\top\mathbf{G}, \text{ for } 1 \leq i, j \leq M.$$

Now, write the $M^2 \times 1$ vectors $vec(\boldsymbol{\Delta}_J)$ and $vec(\boldsymbol{\Delta}_{J^c})$ in the form of block vectors as $vec(\boldsymbol{\Delta}_J) = [(\boldsymbol{\Delta}_J)_i^\top]_{1\leq i\leq M}^\top$ and $vec(\boldsymbol{\Delta}_{J^c}) = [(\boldsymbol{\Delta}_{J^c})_j^\top]_{1\leq j\leq M}^\top$, where $(\boldsymbol{\Delta}_J)_i \in \mathbb{R}^M$ $(\boldsymbol{\Delta}_{J^c})_j \in \mathbb{R}^M$ for $1 \leq i, j \leq M$. Using (A.9) it follows that

$$\begin{aligned}tr\left(\mathbf{G}\boldsymbol{\Delta}_J\mathbf{G}^\top\mathbf{G}\boldsymbol{\Delta}_{J^c}\mathbf{G}^\top\right) &= \sum_{1\leq i,j\leq M}(\boldsymbol{\Delta}_J)_i^\top\mathbf{C}_{ij}(\boldsymbol{\Delta}_{J^c})_j \\ &= \sum_{i\in J}\sum_{j\in J^c}(\mathbf{G}^\top\mathbf{G})_{ij}(\boldsymbol{\Delta}_J)_i^\top\mathbf{G}^\top\mathbf{G}(\boldsymbol{\Delta}_{J^c})_j.\end{aligned}$$

Now, using that $\left|(\mathbf{G}^\top\mathbf{G})_{ij}\right| \leq \theta(\mathbf{G})$ for $i \neq j$ and that

$$\left|(\boldsymbol{\Delta}_J)_i^\top\mathbf{G}^\top\mathbf{G}(\boldsymbol{\Delta}_{J^c})_j\right| \leq \|\mathbf{G}(\boldsymbol{\Delta}_J)_i\|_{\ell_2}\|\mathbf{G}(\boldsymbol{\Delta}_{J^c})_j\|_{\ell_2} \leq \rho_{\max}(\mathbf{G}^\top\mathbf{G})\|(\boldsymbol{\Delta}_J)_i\|_{\ell_2}\|(\boldsymbol{\Delta}_{J^c})_j\|_{\ell_2},$$

it follows that

$$tr\left(\mathbf{G}\boldsymbol{\Delta}_\mathbf{J}\mathbf{G}^\top\mathbf{G}\boldsymbol{\Delta}_{\mathbf{J^c}}\mathbf{G}^\top\right) \geq -\theta(\mathbf{G})\rho_{\max}(\mathbf{G}^\top\mathbf{G})\left(\sum_{i\in J}\|(\boldsymbol{\Delta}_J)_i\|_{\ell_2}\right)\left(\sum_{j\in J^c}\|(\boldsymbol{\Delta}_{J^c})_j\|_{\ell_2}\right).$$

Now, using the assumption that $\sum_{k\in J^c}\|\boldsymbol{\Delta}_k\|_{\ell_2} \leq c_0\sum_{k\in J}\|\boldsymbol{\Delta}_k\|_{\ell_2}$ it follows that

$$\begin{aligned}tr\left(\mathbf{G}\boldsymbol{\Delta}_\mathbf{J}\mathbf{G}^\top\mathbf{G}\boldsymbol{\Delta}_{\mathbf{J^c}}\mathbf{G}^\top\right) &\geq -c_0\theta(\mathbf{G})\rho_{\max}(\mathbf{G}^\top\mathbf{G})\left(\sum_{i\in J}\|(\boldsymbol{\Delta}_J)_i\|_{\ell_2}\right)^2 \\ &\geq -c_0\theta(\mathbf{G})\rho_{\max}(\mathbf{G}^\top\mathbf{G})s\left\|\boldsymbol{\Delta}_J\right\|_F^2, \qquad \text{(A.10)}\end{aligned}$$

where, for the inequality, we have used the properties that for the positive reals $c_i = \|(\mathbf{\Delta}_J)_i\|_{\ell_2}$, $i \in J$ then $\left(\sum_{i\in J} c_i\right)^2 \leq |J| \sum_{i\in J} c_i^2 \leq s \sum_{i\in J} c_i^2$ and that $\sum_{i\in J} \|(\mathbf{\Delta}_J)_i\|_{\ell_2}^2 = \|\mathbf{\Delta}_J\|_F^2$.

Using the properties (A.1) and (A.2) remark that

$$\begin{aligned} \left\|\mathbf{G}\mathbf{\Delta_J}\mathbf{G}^\top\right\|_F^2 &= \|\mathbf{G}_J \otimes \mathbf{G}_J \, vec(\tilde{\mathbf{\Delta}}_J)\|_{\ell_2}^2 \\ &\geq \rho_{\min}\left(\mathbf{G}_J \otimes \mathbf{G}_J\right) \|vec(\tilde{\mathbf{\Delta}}_J)\|_{\ell_2}^2 \\ &\geq \rho_{\min}(s)^2 \, \|\mathbf{\Delta}_J\|_F^2 , \end{aligned} \tag{A.11}$$

where $vec(\tilde{\mathbf{\Delta}}_J) = [(\mathbf{\Delta}_J)_i^\top]_{i\in J}^\top$. Therefore, combining inequalities (A.8), (A.10) and (A.11) it follows that

$$\left\|\mathbf{G}\mathbf{\Delta}\mathbf{G}^\top\right\|_F^2 \geq \left(\rho_{\min}(s)^2 - c_0\theta(\mathbf{G})\rho_{\max}(\mathbf{G}^\top\mathbf{G})s\right) \|\mathbf{\Delta}_J\|_F^2 ,$$

which completes the proof of Lemma 3. □

Let us now proceed to the proof of Theorem 1. Part of the proof is inspired by results in [Bickel et al., 2009]. Let $s \leq \min(n, M)$ and $\mathbf{\Psi} \in \mathcal{S}_M$ with $\mathcal{M}(\mathbf{\Psi}) \leq s$. Let $J = \{k \, ; \mathbf{\Psi}_k \neq 0\}$. To simplify the notations, write $\widehat{\mathbf{\Psi}} = \widehat{\mathbf{\Psi}}_\lambda$. By definition of $\widehat{\mathbf{\Sigma}}_\lambda = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top$ one has that

$$\left\|\widetilde{\mathbf{S}} - \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2 + 2\lambda \sum_{k=1}^M \gamma_k \|\widehat{\mathbf{\Psi}}_k\|_{\ell_2} \leq \left\|\widetilde{\mathbf{S}} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 + 2\lambda \sum_{k=1}^M \gamma_k \|\mathbf{\Psi}_k\|_{\ell_2}. \tag{A.12}$$

Using the scalar product associated to the Frobenius norm $\langle A, B\rangle_F = tr\left(A^\top B\right)$ then

$$\begin{aligned} \left\|\widetilde{\mathbf{S}} - \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2 &= \left\|\mathbf{S} + \mathbf{W} - \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2 \\ &= \|\mathbf{W}\|_F^2 + \left\|\mathbf{S} - \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2 + 2\left\langle \mathbf{W}, \mathbf{S} - \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\rangle_F . \end{aligned} \tag{A.13}$$

Putting (A.13) in (A.12) we get

$$\begin{aligned} \left\|\mathbf{S} - \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2 + 2\lambda \sum_{k=1}^M \gamma_k \|\widehat{\mathbf{\Psi}}_k\|_{\ell_2} \leq{}& \left\|\mathbf{S} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 + 2\left\langle \mathbf{W}, \mathbf{G}\left(\widehat{\mathbf{\Psi}} - \mathbf{\Psi}\right)\mathbf{G}^\top\right\rangle_F \\ &+ 2\lambda \sum_{k=1}^M \gamma_k \|\mathbf{\Psi}_k\|_{\ell_2}. \end{aligned}$$

For $k = 1, \ldots, M$ define the $M \times M$ matrix $\mathbf{A}_k$ with all columns equal to zero except the k-th which is equal to $\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k$. Then, remark that

$$\begin{aligned} \left\langle \mathbf{W}, \mathbf{G}\left(\widehat{\mathbf{\Psi}} - \mathbf{\Psi}\right)\mathbf{G}^\top\right\rangle_F &= \sum_{k=1}^M \left\langle \mathbf{W}, \mathbf{G}\mathbf{A}_k\mathbf{G}^\top\right\rangle_F = \sum_{k=1}^M \left\langle \mathbf{G}^\top\mathbf{W}\mathbf{G}, \mathbf{A}_k\right\rangle_F = \sum_{k=1}^M \eta_k^\top(\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k) \\ &\leq \sum_{k=1}^M \|\eta_k\|_{\ell_2} \|\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k\|_{\ell_2}, \end{aligned}$$

where η_k is the k-th column of the matrix $\mathbf{G}^\top\mathbf{W}\mathbf{G}$. Define the event

$$\mathcal{A} = \bigcap_{k=1}^{M} \left\{2\|\eta_k\|_{\ell_2} \leq \lambda\gamma_k\right\}. \tag{A.14}$$

Then, the choices

$$\gamma_k = 2\|\mathbf{G}_k\|_{\ell_2}\sqrt{\rho_{\max}(\mathbf{G}\mathbf{G}^\top)},\ \lambda = \|\mathbf{\Sigma}_{noise}\|_2\left(1+\sqrt{\frac{n}{N}}+\sqrt{\frac{2\delta\log M}{N}}\right)^2,$$

and Lemma 2 imply that the probability of the complementary event $\mathcal{A}^c$ satisfies

$$\mathbb{P}\left(\mathcal{A}^c\right) \leq \sum_{k=1}^{M}\mathbb{P}\left(2\|\eta_k\|_{\ell_2} > \lambda\gamma_k\right) \leq M^{1-\delta}.$$

Then, on the event $\mathcal{A}$ one has that

$$\begin{aligned}\left\|\mathbf{S}-\mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2 \quad &\leq \quad \left\|\mathbf{S}-\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 + \lambda\sum_{k=1}^{M}\ \gamma_k\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2} \\ &\quad +2\lambda\sum_{k=1}^{M}\gamma_k\left(\|\mathbf{\Psi}_k\|_{\ell_2}-\|\widehat{\mathbf{\Psi}}_k\|_{\ell_2}\right).\end{aligned}$$

Adding the term $\lambda\sum_{k=1}^{M}\ \gamma_k\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2}$ to both sides of the above inequality yields on the event $\mathcal{A}$

$$\begin{aligned}\left\|\mathbf{S}-\mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2 + \lambda\sum_{k=1}^{M}\ \gamma_k\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2} \quad &\leq \quad \left\|\mathbf{S}-\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 \\ &\quad +2\lambda\sum_{k=1}^{M}\gamma_k\left(\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2}+\|\mathbf{\Psi}_k\|_{\ell_2}-\|\widehat{\mathbf{\Psi}}_k\|_{\ell_2}\right).\end{aligned}$$

Now, remark that for all $k \notin J$, then $\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2}+\|\mathbf{\Psi}_k\|_{\ell_2}-\|\widehat{\mathbf{\Psi}}_k\|_{\ell_2}=0$, which implies that on the event $\mathcal{A}$

$$\begin{aligned}\left\|\mathbf{S}-\mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2 + \lambda\sum_{k=1}^{M}\ \gamma_k\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2} \quad &\leq \quad \left\|\mathbf{S}-\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 \qquad (A.15)\\ &\quad +4\lambda\sum_{k\in J}\gamma_k\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2} \\ &\leq \quad \left\|\mathbf{S}-\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 \qquad (A.16)\\ &\quad +4\lambda\sqrt{\mathcal{M}(\mathbf{\Psi})}\sqrt{\sum_{k\in J}\gamma_k^2\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2}^2}.\end{aligned}$$

where for the last inequality we have used the property that for the positive reals $c_k = \gamma_k \|\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k\|_{\ell_2},\ k \in J$ then $\left(\sum_{k\in J} c_k\right)^2 \leq \mathcal{M}(\mathbf{\Psi}) \sum_{k\in J} c_k^2$.

Let $\epsilon > 0$ and define the event

$$\mathcal{A}_1 = \left\{ 4\lambda \sum_{k\in J} \gamma_k \|\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k\|_{\ell_2} > \epsilon \left\|\mathbf{S} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 \right\}. \tag{A.17}$$

Note that on the event $\mathcal{A} \cap \mathcal{A}_1^c$ then the result of the theorem trivially follows from inequality (A.15). Now consider the event $\mathcal{A} \cap \mathcal{A}_1$ (all the following inequalities hold on this event). Using (A.15) one has that

$$\lambda \sum_{k=1}^{M} \gamma_k \|\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k\|_{\ell_2} \leq 4(1 + 1/\epsilon)\lambda \sum_{k\in J} \gamma_k \|\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k\|_{\ell_2}. \tag{A.18}$$

Therefore, on $\mathcal{A} \cap \mathcal{A}_1$

$$\sum_{k\notin J} \gamma_k \|\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k\|_{\ell_2} \leq (3 + 4/\epsilon) \sum_{k\in J} \gamma_k \|\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k\|_{\ell_2}.$$

Let $\mathbf{\Delta}$ be the $M \times M$ symmetric matrix with columns equal to $\mathbf{\Delta}_k = \gamma_k \left(\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k\right), k = 1, \ldots, M$, and $c_0 = 3 + 4/\epsilon$. Then, the above inequality means that $\sum_{k\in J^c} \|\mathbf{\Delta}_k\|_{\ell_2} \leq c_0 \sum_{k\in J} \|\mathbf{\Delta}_k\|_{\ell_2}$ and thus Assumption 1 and Lemma 3 imply that

$$\kappa_{s,c_0}^2 \sum_{k\in J} \gamma_k^2 \|\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k\|_{\ell_2}^2 \leq \left\|\mathbf{G}\mathbf{\Delta}\mathbf{G}^\top\right\|_F^2 \leq 4\mathbf{G}_{\max}^2 \rho_{\max}(\mathbf{G}^\top\mathbf{G}) \left\|\mathbf{G}(\widehat{\mathbf{\Psi}} - \mathbf{\Psi})\mathbf{G}^\top\right\|_F^2. \tag{A.19}$$

Let $\gamma_{\max}^2 = 4\mathbf{G}_{\max}^2 \rho_{\max}(\mathbf{G}^\top\mathbf{G})$. Combining the above inequality with (A.16) yields

$$\begin{aligned}
\left\|\mathbf{S} - \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2 &\leq \left\|\mathbf{S} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 + 4\lambda\kappa_{s,c_0}^{-1}\gamma_{\max}\sqrt{\mathcal{M}(\mathbf{\Psi})} \left\|\mathbf{G}(\widehat{\mathbf{\Psi}} - \mathbf{\Psi})\mathbf{G}^\top\right\|_F \\
&\leq \left\|\mathbf{S} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 + 4\lambda\kappa_{s,c_0}^{-1}\gamma_{\max}\sqrt{\mathcal{M}(\mathbf{\Psi})} \left(\left\|\mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top - \mathbf{S}\right\|_F \right. \\
&\quad \left. + \left\|\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top - \mathbf{S}\right\|_F\right)
\end{aligned}$$

Now, arguing as in [Bickel et al., 2009], a decoupling argument using the inequality $2xy \leq bx^2 + b^{-1}y^2$ with $b > 1$, $x = 2\lambda\kappa_{s,c_0}^{-1}\gamma_{\max}\sqrt{\mathcal{M}(\mathbf{\Psi})}$ and y being either $\left\|\mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top - \mathbf{S}\right\|_F$ or $\left\|\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top - \mathbf{S}\right\|_F$ yields the inequality

$$\left\|\mathbf{S} - \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2 \leq \left(\frac{b+1}{b-1}\right) \left\|\mathbf{S} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 + \frac{8b^2\gamma_{\max}^2}{(b-1)\kappa_{s,c_0}^2}\lambda^2\mathcal{M}(\mathbf{\Psi}). \tag{A.20}$$

Then, taking $b = 1 + 2/\epsilon$ and using the inequalities $\left\|\mathbf{\Sigma} - \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2 \leq 2\|\mathbf{S} - \mathbf{\Sigma}\|_F^2 + 2\left\|\mathbf{S} - \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2$ and $\left\|\mathbf{S} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 \leq 2\|\mathbf{S} - \mathbf{\Sigma}\|_F^2 + 2\left\|\mathbf{\Sigma} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2$ completes the proof of Theorem 1. □

A.6 Proof of Theorem 2

Part of the proof is inspired by the approach followed in [Lounici, 2008] and [Lounici et al., 2009]. Note first that

$$\max_{1\leq k\leq M}\gamma_k\left\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k^*\right\|_{\ell_2}\leq\sum_{k=1}^{M}\gamma_k\left\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k^*\right\|_{\ell_2}.$$

Since $\mathbf{\Psi}^*\in\{\mathbf{\Psi}\in\mathcal{S}_M : M(\mathbf{\Psi})\leq s_*\}$, we can use some results from the proof of Theorem (1). On the event $\mathcal{A}\cap\mathcal{A}_1$, with $\mathcal{A}$ defined by (A.14) and $\mathcal{A}_1$ defined by (A.17), inequality (A.18) implies that

$$\begin{aligned}\sum_{k=1}^{M}\gamma_k\left\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k^*\right\|_{\ell_2} &\leq 4\left(1+\frac{1}{\epsilon}\right)\sum_{k\in J^*}\gamma_k\left\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k^*\right\|_{\ell_2}\\ &\leq 4\left(1+\frac{1}{\epsilon}\right)\sqrt{s_*}\sqrt{\sum_{k\in J^*}\gamma_k^2\left\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k^*\right\|_{\ell_2}^2}.\end{aligned}$$

Let $\mathbf{\Delta}^*$ be the $M\times M$ symmetric matrix with columns equal to $\mathbf{\Delta}_k^*=\gamma_k\left(\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k^*\right)$, $k=1,\ldots,M$, let $\gamma_{\max}=2\mathbf{G}_{\max}\sqrt{\rho_{\max}(\mathbf{G}^\top\mathbf{G})}$ and $c_0=3+4/\epsilon$. Then, the above inequality and (A.19) imply that on the event $\mathcal{A}\cap\mathcal{A}_1$

$$\begin{aligned}\sum_{k=1}^{M}\gamma_k\left\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k^*\right\|_{\ell_2} &\leq \frac{4\left(1+\frac{1}{\epsilon}\right)\sqrt{s_*}}{\kappa_{s_*,c_0}}\left\|\mathbf{G}\mathbf{\Delta}^*\mathbf{G}^\top\right\|_F\leq\frac{4\left(1+\frac{1}{\epsilon}\right)\sqrt{s_*}}{\kappa_{s_*,c_0}}\gamma_{\max}\left\|\mathbf{G}\left(\widehat{\mathbf{\Psi}}-\mathbf{\Psi}^*\right)\mathbf{G}^\top\right\|_F\\ &= \frac{4\left(1+\epsilon\right)\sqrt{s_*}}{\epsilon\kappa_{s_*,c_0}}\gamma_{\max}\left\|\widehat{\mathbf{\Sigma}}_\lambda-\mathbf{\Sigma}\right\|_F\\ &\leq \frac{4\left(1+\epsilon\right)\sqrt{s_*}}{\epsilon\kappa_{s_*,c_0}}\gamma_{\max}\sqrt{n}\sqrt{C_0\left(n,M,N,s_*,\mathbf{S},\mathbf{\Psi}^*,\mathbf{G},\mathbf{\Sigma}_{noise}\right)},\end{aligned}$$

Then, using (A.15) one has that on the event $\mathcal{A}\cap\mathcal{A}_1^c$

$$\sum_{k=1}^{M}\gamma_k\left\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k^*\right\|_{\ell_2}\leq\frac{1+\epsilon}{\lambda}\left\|\mathbf{S}-\mathbf{G}\mathbf{\Psi}^*\mathbf{G}^\top\right\|_F^2.$$

Therefore, by definition of C_1, the previous inequalities imply that on the event $\mathcal{A}$ (of probability $1-M^{1-\delta}$)

$$\sum_{k=1}^{M}\frac{\|\mathbf{G}_k\|_{\ell_2}}{\sqrt{n}\mathbf{G}_{\max}}\left\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k^*\right\|_{\ell_2}\leq C_1\left(n,M,N,s_*,\mathbf{S},\mathbf{\Psi}^*,\mathbf{G},\mathbf{\Sigma}_{noise}\right). \tag{A.21}$$

Hence $\max_{1\leq k\leq M}\frac{\delta_k}{\sqrt{n}}\left\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k^*\right\|_{\ell_2}\leq C_1\left(\sigma,n,M,N,s_*,\mathbf{G},\mathbf{\Sigma}_{noise}\right)$ with probability at least $1-M^{1-\delta}$, which proves the first assertion of Theorem 2.

Then, to prove that $\hat{J} = J^*$ we use that $\frac{\delta_k}{\sqrt{n}} \left| \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} - \|\mathbf{\Psi}_k^*\|_{\ell_2} \right| \leq \frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k^* \right\|_{\ell_2}$ for all $k = 1, \ldots, M$. Then, by (A.21)

$$\left| \frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} - \frac{\delta_k}{\sqrt{n}} \|\mathbf{\Psi}_k^*\|_{\ell_2} \right| \leq C_1 \left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise} \right),$$

which is equivalent to

$$-C_1 \left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise} \right) \leq \frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} - \frac{\delta_k}{\sqrt{n}} \|\mathbf{\Psi}_k^*\|_{\ell_2} \leq C_1 \left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise} \right). \tag{A.22}$$

If $k \in \hat{J}$ then $\frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} > C_1 \left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise} \right)$. Inequality $\frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} - \frac{\delta_k}{\sqrt{n}} \|\mathbf{\Psi}_k^*\|_{\ell_2} \leq C_1 \left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise} \right)$ from (A.22) imply that $\frac{\delta_k}{\sqrt{n}} \|\mathbf{\Psi}_k^*\|_{\ell_2} \geq \frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} - C_1 \left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise} \right) > 0$, where the last inequality is obtained using that $k \in \hat{J}$. Hence $\|\mathbf{\Psi}_k^*\|_{\ell_2} > 0$ and therefore $k \in J^*$. If $k \in J^*$ then $\|\mathbf{\Psi}_k^*\|_{\ell_2} \neq 0$. Inequality $-C_1 \left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise} \right) \leq \frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} - \frac{\delta_k}{\sqrt{n}} \|\mathbf{\Psi}_k^*\|_{\ell_2}$ from (A.22) imply that $\frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} + C_1 \left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise} \right) \geq \frac{\delta_k}{\sqrt{n}} \|\mathbf{\Psi}_k^*\|_{\ell_2} > 2C_1 \left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise} \right)$, where the last inequality is obtained using Assumption (3.17) on $\frac{\delta_k}{\sqrt{n}} \|\mathbf{\Psi}_k^*\|_{\ell_2}$. Hence $\frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} > 2C_1 \left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise} \right) - C_1 \left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise} \right) = C_1 \left(n, M, N, s_*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise} \right)$ and therefore $k \in \hat{J}$. This completes the proof of Theorem 2. □

A.7 Proof of Theorem 3

Under the assumptions of Theorem 3, we have shown in the proof of Theorem 2 that $\hat{J} = J^*$ on the event $\mathcal{A}$ defined by (A.14). Therefore, under the assumptions of Theorem 3 it can be checked that on the event $\mathcal{A}$ (of probability $1 - M^{1-\delta}$)

$$\widehat{\mathbf{\Sigma}}_{\hat{J}} = \widehat{\mathbf{\Sigma}}_{J^*} = \mathbf{G}_{J^*} \widehat{\mathbf{\Psi}}_{J^*} \mathbf{G}_{J^*}^\top,$$

with

$$\widehat{\mathbf{\Psi}}_{J^*} = \left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*} \right)^{-1} \mathbf{G}_{J^*}^\top \widetilde{\mathbf{S}} \mathbf{G}_{J^*} \left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*} \right)^{-1}.$$

Now, from the definition (3.19) of $\mathbf{\Sigma}_{J^*}$ it follows that on the event $\mathcal{A}$

$$\left\| \widehat{\mathbf{\Sigma}}_{\hat{J}} - \mathbf{\Sigma}_{J^*} \right\|_2 \leq \rho_{\max} \left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*} \right) \left\| \widehat{\mathbf{\Psi}}_{J^*} - \Lambda_{J^*} \right\|_2 \tag{A.23}$$

where $\Lambda_{J^*} = \mathbf{\Psi}_{J^*} + (\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*})^{-1} \mathbf{G}_{J^*}^\top \mathbf{\Sigma}_{noise} \mathbf{G}_{J^*} \left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*} \right)^{-1}$. Let $\mathbf{Y}_i = \left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*} \right)^{-1} \mathbf{G}_{J^*}^\top \widetilde{\mathbf{X}}_i$ for $i = 1, \ldots, N$ and remark that

$$\widehat{\mathbf{\Psi}}_{J^*} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{Y}_i \mathbf{Y}_i^\top \text{ with } \mathbb{E} \widehat{\mathbf{\Psi}}_{J^*} = \Lambda_{J^*}.$$

Therefore, $\widehat{\mathbf{\Psi}}_{J^*}$ is a sample covariance matrix of size $s_* \times s_*$ and we can control its deviation in operator norm from Λ_{J^*} by using Proposition 3. For this we simply have to verify conditions similar to **(A1)** and **(A2)** in Assumption 2 for the random vector $\mathbf{Y} = \left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)^{-1} \mathbf{G}_{J^*}^\top \widetilde{\mathbf{X}} \in \mathbb{R}^{s_*}$. First, let $\beta \in \mathbb{R}^{s_*}$ with $\|\beta\|_{\ell_2} = 1$. Then, remark that $\mathbf{Y}^\top \beta = \widetilde{\mathbf{X}}^\top \tilde{\beta}$ with $\tilde{\beta} = \mathbf{G}_{J^*} \left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)^{-1} \beta$. Since $\|\widetilde{\beta}\|_{\ell_2} \leq \left(\rho_{\min}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\right)^{-1/2}$ it follows that

$$\left(\mathbb{E}|\mathbf{Y}^\top \beta|^4\right)^{1/4} \leq \tilde{\rho}(\mathbf{\Sigma}, \mathbf{\Sigma}_{noise}) \rho_{\min}^{-1/2}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right), \tag{A.24}$$

where $\tilde{\rho}(\mathbf{\Sigma}, \mathbf{\Sigma}_{noise}) = 8^{1/4}\left(\rho^4(\mathbf{\Sigma}) + \rho^4(\mathbf{\Sigma}_{noise})\right)^{1/4}$. Now let $\tilde{Z} = \|\mathbf{Y}\|_{\ell_2} \leq \rho_{\min}^{-1/2}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right) \|\widetilde{\mathbf{X}}\|_{\ell_2}$. Given our assumptions on the process $\widetilde{X} = X + \mathcal{E}$ it follows that there exists $\alpha \geq 1$ such that

$$\|\tilde{Z}\|_{\psi_\alpha} \leq \rho_{\min}^{-1/2}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\left(\|Z\|_{\psi_\alpha} + \|W\|_{\psi_\alpha}\right) < +\infty, \tag{A.25}$$

where $Z = \|\mathbf{X}\|_{\ell_2}$ and $W = \|\mathcal{E}\|_{\ell_2}$, with $\mathbf{X} = \left(X(t_1), ..., X(t_n)\right)^\top$ and $\mathcal{E} = \left(\mathcal{E}(t_1), ..., \mathcal{E}(t_n)\right)^\top$. Finally, remark that

$$\|\Lambda_{J^*}\|_2 \leq \|\mathbf{\Psi}_{J^*}\|_2 + \rho_{\min}^{-1}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right) \|\mathbf{\Sigma}_{noise}\|_2. \tag{A.26}$$

Hence, using the relations (A.24) and (A.25), the bound (A.26) and Proposition 3 (with $\mathbf{Y}$ instead of $\mathbf{X}$), it follows that there exists a universal constant $\delta_* > 0$ such that for all $x > 0$,

$$\mathbb{P}\left(\left\|\widehat{\mathbf{\Psi}}_{J^*} - \Lambda_{J^*}\right\|_2 \geqslant \tilde{\tau}_{N,s_*} x\right) \leqslant \exp\left(-(\delta_*^{-1} x)^{\frac{\alpha}{2+\alpha}}\right), \tag{A.27}$$

where $\tilde{\tau}_{N,s_*} = \max(\tilde{A}_{N,s_*}^2, \tilde{B}_{N,s_*})$, with $\tilde{A}_{N,s_*} = \|\tilde{Z}\|_{\psi_\alpha} \frac{\sqrt{\log d^*}(\log N)^{1/\alpha}}{\sqrt{N}}$ and $\tilde{B}_{N,s_*} = \frac{\tilde{\rho}^2(\mathbf{\Sigma}, \mathbf{\Sigma}_{noise}) \rho_{\min}^{-1}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}{\sqrt{N}} + \left(\|\mathbf{\Psi}_{J^*}\|_2 + \rho_{\min}^{-1}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right) \|\mathbf{\Sigma}_{noise}\|_2\right)^{1/2} \tilde{A}_{N,s_*}$, with $d^* = \min(N, s_*)$. Then, define the event

$$\mathcal{B} = \left\|\widehat{\mathbf{\Psi}}_{J^*} - \Lambda_{J^*}\right\|_2 \leqslant \tilde{\tau}_{N,s_*} \delta_\star \left(\log(M)\right)^{\frac{2+\alpha}{\alpha}},$$

and note that, for $x = \delta_\star \left(\log(M)\right)^{\frac{2+\alpha}{\alpha}}$ with $\delta_\star > \delta_*$, inequality (A.27) implies that $\mathbb{P}(\mathcal{B}) \geq 1 - M^{-\left(\frac{\delta_\star}{\delta_*}\right)^{\frac{\alpha}{2+\alpha}}}$. Therefore, on the event $\mathcal{A} \cap \mathcal{B}$ (of probability at least $1 - M^{1-\delta} - M^{-\left(\frac{\delta_\star}{\delta_*}\right)^{\frac{\alpha}{2+\alpha}}}$), using inequality (A.23) and the fact that $\hat{J} = J^*$ one obtains

$$\left\|\widehat{\mathbf{\Sigma}}_{\hat{J}} - \mathbf{\Sigma}_{J^*}\right\|_2 \leq \rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right) \tilde{\tau}_{N,s_*} \delta_\star \left(\log(M)\right)^{\frac{2+\alpha}{\alpha}},$$

which completes the proof of Theorem 3. □

References

[Antoniadis et al., 2001] Antoniadis, A., Bigot, J., and Sapatinas, T. (2001). Wavelet estimators in nonparametric regression: A comparative simulation study. *Journal of Statistical Software*, 6(6):1–83.

[Bach, 2008] Bach, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225.

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

[Bickel and Levina, 2008b] Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227.

[Bickel et al., 2009] Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732.

[Bigot et al., 2010] Bigot, J., Biscay, R. J., Loubes, J.-M., and Muñiz Alvarez, L. (2010). Nonparametric estimation of covariance functions by model selection. *Electronic Journal of Statistics*, 4:822–855.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press, Cambridge.

[Cressie, 1993] Cressie, N. A. C. (1993). *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York.

[d'Aspremont et al., 2008] d'Aspremont, A., Bach, F., and El Ghaoui, L. (2008). Optimal solutions for sparse principal component analysis. *J. Mach. Learn. Res.*, 9:1269–1294.

[Davidson and Szarek, 2001] Davidson, K. R. and Szarek, S. J. (2001). Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces, Vol. I*, pages 317–366. North-Holland, Amsterdam.

[Davis and Kahan, 1970] Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7:1–46.

[El Karoui, 2008] El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.*, 36(6):2717–2756.

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

[Huang and Zhang, 2010] Huang, J. and Zhang, T. (2010). The benefit of group sparsity. *Ann. Statist.*, 38(4):1978–2004.

[Johnstone, 2001] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, 29(2):295–327.

[Johnstone and Lu, 2009] Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693.

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

[Lam and Fan, 2009] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.*, 37(6B):4254–4278.

[Levina et al., 2008] Levina, E., Rothman, A., and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *Ann. Appl. Stat.*, 2(1):245–263.

[Lounici, 2008] Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.*, 2:90–102.

[Lounici et al., 2009] Lounici, K., Pontil, M., Tsybakov, A. B., and van de Geer, S. (2009). Taking advantage of sparsity in multi-task learning. *COLT*.

[Lounici et al., 2011] Lounici, K., Pontil, M., Tsybakov, A. B., and van de Geer, S. (2011). Oracle Inequalities and Optimal Inference under Group Sparsity. *Ann. Statist.*, to be published.

[Mendelson and Pajor, 2006] Mendelson, S. and Pajor, A. (2006). On singular values of matrices with independent rows. *Bernoulli*, 12(5):761–773.

[Nardi and Rinaldo, 2008] Nardi, Y. and Rinaldo, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Stat.*, 2:605–633.

[Rothman et al., 2008] Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.*, 2:494–515.

[Schmidt et al., 2008] Schmidt, M., Murphy, K., Fung, G., and Rosales, R. (2008). Structure learning in random fields for heart motion abnormality detection (addendum). *CVPR08*.

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

[Wikle and Cressie, 1999] Wikle, C. K. and Cressie, N. (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika*, 86(4):815–829.

[Zou et al., 2006] Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.*, 15(2):265–286.