Some contributions to large precision matrix estimation

Samuel Balmand

UNIVERSITÉ PARIS-EST

École Doctorale Mathématiques et STIC

Laboratoire MATIS de l'IGN

Thèse

Présentée pour l'obtention du grade de Docteur
de l'Université Paris-Est
par

Samuel BALMAND

Quelques contributions à l'estimation de grandes matrices de précision

---

Spécialité : Mathématiques

Dirigée par **Arnak DALALYAN et Marc PIERROT-DESEILLIGNY**

Soutenue le 27 juin 2016 devant un jury composé de :

M. Pierre **ALQUIER**	Professeur, CREST, ENSAE	Examinateur
Mme Cristina **BUTUCEA**	Professeur, LAMA, UPEM	Examinatrice
M. Julien **CHIQUET**	Chargé de recherche HDR, MIA Paris, INRA	Rapporteur
M. Arnak **DALALYAN**	Professeur, CREST, ENSAE	Directeur
M. Erwan **LE PENNEC**	Professeur, CMAP, École Polytechnique	Rapporteur
M. Marc **PIERROT-DESEILLIGNY**	Directeur de recherche, MATIS, IGN	Co-directeur

Thèse effectuée au sein du **Laboratoire MATIS (IGN)**

de l'Université Paris-Est
73, avenue de Paris
94165 Saint-Mandé cedex
France

ÉCOLE NATIONALE
DES SCIENCES
GÉOGRAPHIQUES

à l'**ENSG**

6 et 8, avenue Blaise Pascal
Cité Descartes - Champs-sur-Marne
77455 Marne la Vallée Cedex 2
France

Résumé

Sous l'hypothèse gaussienne, la relation entre indépendance conditionnelle et parcimonie permet de justifier la construction d'estimateurs de l'inverse de la matrice de covariance – également appelée matrice de précision – à partir d'approches régularisées. Cette thèse, motivée à l'origine par la problématique de classification d'images, vise à développer une méthode d'estimation de la matrice de précision en grande dimension, lorsque le nombre n d'observations est petit devant la dimension p du modèle. Notre approche repose essentiellement sur les liens qu'entretiennent la matrice de précision et le modèle de régression linéaire. Elle consiste à estimer la matrice de précision en deux temps. Les éléments non diagonaux sont tout d'abord estimés en considérant p problèmes de minimisation du type racine carrée des moindres carrés pénalisés par la norme ℓ_1. Les éléments diagonaux sont ensuite obtenus à partir du résultat de l'étape précédente, par analyse résiduelle ou maximum de vraisemblance. Nous comparons ces différents estimateurs des termes diagonaux en fonction de leur risque d'estimation. De plus, nous proposons un nouvel estimateur, conçu de sorte à tenir compte de la possible contamination des données par des *outliers*, grâce à l'ajout d'un terme de régularisation en norme mixte ℓ_2/ℓ_1. L'analyse non-asymptotique de la convergence de notre estimateur souligne la pertinence de notre méthode.

Mots-clefs :

matrice de précision, parcimonie, grande dimension, modèles graphiques gaussiens, modèle de régression linéaire, estimation robuste, minimisation convexe, vitesse de convergence, analyse non-asymptotique.

Some contributions to large precision matrix estimation

Abstract

Under the Gaussian assumption, the relationship between conditional independence and sparsity allows to justify the construction of estimators of the inverse of the covariance matrix – also called precision matrix – from regularized approaches. This thesis, originally motivated by the problem of image classification, aims at developing a method to estimate the precision matrix in high dimension, that is when the sample size n is small compared to the dimension p of the model. Our approach relies basically on the connection of the precision matrix to the linear regression model. It consists of estimating the precision matrix in two steps. The off-diagonal elements are first estimated by solving p minimization problems of the type ℓ_1-penalized square-root of least-squares. The diagonal entries are then obtained from the result of the previous step, by residual analysis of likelihood maximization. This various estimators of the diagonal entries are compared in terms of estimation risk. Moreover, we propose a new estimator, designed to consider the possible contamination of data by outliers, thanks to the addition of a ℓ_2/ℓ_1 mixed norm regularization term. The nonasymptotic analysis of the consistency of our estimator points out the relevance of our method.

Keywords:

precision matrix, sparsity, high dimension, Gaussian graphical models, linear regression model, robust estimation, convex minimization, convergence rate, nonasymptotic analysis.

Remerciements

Mes remerciements vont en premier lieu à mes directeurs de thèse. À Arnak Dalalyan auprès de qui j'ai énormément appris, pour ses conseils, son aide, sa façon claire d'exposer des idées complexes, sa hauteur de vue, ses relectures attentives et sa confiance. Je prends pleinement la mesure de tout ce que ce travail de recherche lui doit. À Marc Pierrot-Deseilligny pour son accueil au sein du Lga, pour ses conseils, pour avoir fait le lien avec les autres laboratoires de l'Ign et avoir fait en sorte que les questions administratives ne soient jamais un problème.

Je remercie les membres du jury, pour m'avoir fait l'honneur d'assister à la soutenance de cette thèse. Parmi eux, je remercie les rapporteurs, Julien Chiquet et Erwan Le Pennec qui se sont penchés avec bienveillance sur mon travail. Je remercie chaleureusement Pierre Alquier et Cristina Butucea pour avoir accepté de faire partie du jury.

Je remercie l'Ensg qui m'a recruté sur un poste d'enseignant-chercheur. J'ai toujours pris beaucoup de plaisir à y donner des cours, même s'il a parfois été difficile d'éviter que les activités d'enseignement ne prennent le pas sur celles de recherche. Mes remerciements vont également au laboratoire Matis de l'Ign, pour m'avoir accueilli suite à la disparition du Lga à l'Ensg.

Je n'oublie pas mes collègues à l'Ensg, tout particulièrement Cécile, Cédric, Laurent, Olivier, Philippe et Vincent, pour leur accueil, les cafés et les repas partagés, avec une mention spéciale pour l'équipe réseau qui m'a permis d'accéder aux serveurs de calculs de l'école. Plus généralement, j'adresse mes remerciements à l'ensemble du personnel de l'école qui a, au quotidien, su créer un environnement de travail agréable. Par ailleurs, je pense à mes anciens collègues à l'Ensai, en particulier à Daniel, François, Laurence et Myriam, qui ont beaucoup contribué à mon intérêt pour la recherche et l'enseignement.

Enfin, j'associe ma famille et mes amis à ces remerciements, pour leurs encouragements, et Marie, pour tout le reste, son soutien et son écoute patiente.

Contents

List of Figures

List of Tables

List of Algorithms

Abbreviations

Clime	constrained ℓ_1-minimization for inverse matrix estimation ... 14–16, 37, 76, 104, 131
i.i.d.	independently and identically distributed .. 5, 11, 20, 21, 25, 26, 28, 33, 36, 41, 44, 45, 57, 58, 74, 116, 123–125
LARS	least angle regression 23
Lasso	least absolute shrinkage and selection operator 10, 12, 14, 15, 17–23, 25, 37
MLE	maximum likelihood estimator 11, 57, 58, 110
MSE	mean squared error 18
MST	minimum spanning tree 55, 56
OLS	ordinary least squares ..... xi, 15, 17, 18, 48, 50, 54, 57, 59–61
PCA	principal component analysis 4, 5, 9
PML	penalized maximum likelihood 29, 35, 36, 48–50, 52–54, 56, 57, 59–64, 134–142
RML	relaxed maximum likelihood ... 29, 36, 42, 46–50, 52–54, 57, 59–64, 134–142
RSS	residual sum of squares 18
RV	residual variance 29, 36, 39, 42, 46, 48, 50, 52–54, 57, 59–64, 134–142
SIFT	scale-invariant feature transform 3
SML	symmetry-enforced maximum likelihood .. 29, 35, 36, 47–50, 52–55, 59–64, 134–142
SOCP	second-order cone program 23, 53, 145
SPT	shortest path tree 55
SVM	support vector machine 4, 111

Introduction

RÉSUMÉ. Dans le contexte de la vision par ordinateur, la capacité à identifier le contenu d'une image est un enjeu majeur. Intéressons-nous par exemple au problème de la classification d'images qui requiert une représentation de l'image adaptée. Or les caractéristiques d'une images ont une dimension particulièrement élevée (un grand nombre de variables). En comparaison, le nombre d'observations dont on dispose est souvent relativement petit. Cette situation a plusieurs conséquences. Au premier lieu desquelles, certaines techniques standard de classification sont applicables en théorie, mais en pratique ont des coûts de calculs en temps et en espace prohibitifs. Pour être mises en œuvre efficacement, la classification doit alors être précédée d'une étape visant à réduire la dimension. Cette dernière peut cependant avoir pour conséquence une perte d'information, qui peut se révéler préjudiciable du point de vue de l'objectif de classification. Il est donc souhaitable de se passer de cette réduction de la dimension et de développer des techniques de classification qui soient naturellement adaptées aux données de grande dimension. Considérons par exemple la méthode de classification qui repose sur l'hypothèse bayésienne naïve. Elle nécessite d'être en mesure de calculer les densités de probabilité d'un descripteur, conditionnellement à chaque classe. Nous proposons de faire l'hypothèse que les p caractéristiques d'une image dans une classe donnée sont distribuées selon une loi de probabilité dont les paramètres doivent être estimés. Dans le cadre de la classification bayésienne et sous l'hypothèse gaussienne, il est ainsi nécessaire d'estimer pour chaque classe l'inverse de la matrice de covariance, ou matrice de précision, soit $p(p+1)/2$ paramètres. En dimension élevée, obtenir une estimation fiable pour les paramètres du classifieur n'est pas possible sans hypothèses supplémentaires.

Notre principal apport est la construction et la justification théorique d'un estimateur de cette matrice de précision en grande dimension, sous l'hypothèse

d'indépendance conditionnelle des variables. En particulier, nous nous sommes penchés sur la manière d'améliorer l'estimation des termes diagonaux de cette matrice et avons proposé une approche permettant de tenir compte de la présence potentielle d'observations aberrantes ou extrêmes. Dans ce chapitre introductif, nous justifierons l'approche parcimonieuse, en lien avec l'hypothèse d'indépendance conditionnelle. Nous ferons une revue de l'état de l'art sur l'estimation de matrices de précision en grande dimension. Notre approche étant basée sur le modèle de régression linéaire, nous reviendrons sur les développements récents des techniques de régularisation dans ce domaine. Ceci nous amènera à présenter les propriétés de régularité utilisées pour démontrer la convergence des estimateurs obtenus.

Contents

Problem statement and motivation In the field of computer vision, the ability to identify what an image represents remains a major challenge. In the particular problem of image classification, the objective is to determine to which category a new image belongs. To this end, we need to make the best use of the information contained in the training sample that gathers images with known labels. The first difficulty is to build an image representation adapted to the classification goal. There is a considerable amount of research about building such an effective representation, using the local or the global features of the image. Resting upon the recent developments on this subject, we can use image descriptors such as the local scale-invariant feature transform (SIFT) descriptor [Lowe, 2004], the global GIST [Oliva and Torralba, 2006] descriptor or variations thereof. The other difficulty is about the classification procedure itself. It is yet deeply related to

the chosen image representation. The features of an image have indeed an especially high dimension, that is a great number of variables. It is not uncommon to have to consider representations involving hundred of variables, even thousands in the case of bitmap representation. In comparison, the number of available observations is often quite small. This situation has many consequences. First of all, many classification procedures are theoretically workable, but have in practice prohibitive computational costs, requiring a huge amount of time and space. That is the case, for instance, of standard approaches such as the k-nearest neighbors method or support vector machine (SVM). To be applied efficiently, the classification should be preceded by a step to reduce the dimension. The most widely used techniques include linear methods such as principal component analysis (PCA) and nonlinear such as product quantization [Jegou et al., 2011]. This reduction of the dimension is based on the assumption that the covariates are (highly) correlated and that their number can be reduced without losing information. However, in practice, the reduction of the dimension may involve a loss of information. Incidentally, it also results in lessening the noise, but may be harmful in view of the performance of the classification.

In the context of Bayesian classification and under the Gaussian assumption, we need to estimate the inverse of the covariance matrix, thus $p(p+1)/2$ parameters. In high dimension, providing an accurate estimation for the parameters of this classifier is not possible without additional assumptions. In this background, reducing the dimensionality of data can be considered, but is not an appropriate solution. Indeed, aside from leading to information loss, the reduction of the dimension is likely to be on a collision course with the Gaussian premise. In addition, assuming that the initial covariates are correlated is debatable in high dimension [Zhang et al., 2014]. Another assumption, more tenable, would be to suppose that the initial covariates are pairwise independent, conditionally to the other. The assumption that the variables are highly correlated is of a completely different nature than that of conditional independence of the variables. Intuitively, the former assumption boils down to suppose that the information contained in each observation is redundant, hence the possibility of reducing the dimension of the covariate space without without losing information. In contrast, the second assumption consists in supposing that the additional information provided by two different variables is very often non-redundant. It is thus only necessary to estimate some of the partial correlations of the model, the others being zero. Recalling that the precision matrix can be associated to the graph of dependence relationships between variables, the first assumption results in reducing the number of nodes, but also in considering that each pair of nodes may be linked by an edge. The second assumption signifies that only a quite small number of edges, related to the nonzero entries of the precision matrix, have to be identified. These assumptions are not incompatible, it is however not obvious whether both can be met together. The hypothesis of parsimony has indeed fewer chances to be fulfilled after dimension reduction. Recent work focuses on this question [Han et al., 2014]. It concludes that there is no

theoretical result which guarantees that the sparsity structure of the precision matrix corresponding to unobserved latent variables (for instance the principal components in the case of PCA) inherits from the structure of the precision matrix corresponding to initial variables. Nevertheless, when the structure of the precision matrix (thus of the underlying graph) is particular (modular for instance [Celik et al., 2014]) and under certain conditions (among those the Gaussian assumption), the precision matrix of observed variables can be obtained from that of latent variables.

The method we present is based on naive Bayes assumption. In this case, the probability density functions associated with the image descriptors, conditionally to each class, are needed. This approach has already been implemented in high dimension in [Behmo et al., 2010], through the nonparametric estimation of these densities of probability. We rather assume that the features of an image in a given class are drawn from a probability distribution whose parameters have to be estimated. More precisely, we suppose that the features are Gaussian distributed, thus the rule of classification involves the mean and the precision matrix of the distribution of all the features of each class. Let us state the decision rule in this setting. We want to determine what class C, from a finite set of available classes, an unlabeled image I belongs to. The image is represented by the descriptors $\mathbf{d}^I = \{\mathbf{d}_1^I, \ldots, \mathbf{d}_{n^I}^I\}$. The maximum a posteriori (MAP) rule of classification runs as follows: the image is associated with the class for which the conditional probability is the largest, that is

$$\widehat{C}_I = \operatorname*{argmax}_C p(C|I).$$

We suppose that the prior probability $p(C)$ is the same regardless of the class (uniform prior), thus applying the theorem of Bayes entails that $p(C|I) \propto p(I|C) = p(\mathbf{d}_1^I, \ldots, \mathbf{d}_{n^I}^I|C)$. Then the classifier is none other than the maximum likelihood classifier. Under the naive Bayes hypothesis that the descriptors of an image are independently and identically distributed (i.i.d.) given its class, it implies that the decision rule is

$$\widehat{C}_I = \operatorname*{argmax}_C \prod_{i=1}^{n^I} p(\mathbf{d}_i^I|C) = \operatorname*{argmax}_C \sum_{i=1}^{n^I} \log p(\mathbf{d}_i^I|C). \tag{0.1}$$

Assuming that the descriptors are normally distributed conditionally to the class C, we obtain

$$\widehat{C}_I = \arg\min_C \left\{ \Big(\sum_{i=1}^{n^I} (\mathbf{d}_i^I - \boldsymbol{\mu}^C)^\top \boldsymbol{\Omega}^C (\mathbf{d}_i^I - \boldsymbol{\mu}^C) \Big) - \frac{n^I}{2} \log\det(\boldsymbol{\Omega}^C) \right\}. \tag{0.2}$$

To get a classifier workable in practice, the densities $p(\mathbf{d}_i^I|C)$ have to be estimated from training data. It is therefore necessary to be able to estimate the mean $\boldsymbol{\mu}^C$ and the pre-

cision matrix $\mathbf{\Omega}^C$ for all classes of interest. In the high dimensional setting, the standard statistical techniques fail to estimate accurately the precision matrix. Our main contribution is the construction and theoretical analysis of an estimator of the precision matrix in high dimension under the assumption of conditional independence. In particular, we study how to improve the estimation of the diagonal entries of this matrix and propose a new approach taking into account the possible presence of outliers.

Computer vision is far from being the only field of application of sparse high dimensional methods of estimation. Problems that required a consistent estimation of the precision matrix arise in many other areas. Of these, mention may be made of genomics, econometrics, signal processing or meteorology. In genomics, the development of DNA microarrays demands techniques for interpreting large-scale datasets of gene expression. For instance, Kishino and Waddell [2000] proposed to model the relationships between genes using the partial correlations of the expressions of all pairs of genes. This work on gene expression data has been extended to the high dimensional framework by Schäfer and Strimmer [2005] or Cai et al. [2013] among others.
In econometrics, solutions to portfolio optimizations problems may involve the estimation of a precision matrix [Markowitz, 1952; Brandt, 2010]. The portfolio selection problem aims to identify the best possible combination of p assets whose expectation is denoted by $\boldsymbol{\mu}$ and covariance matrix by $\mathbf{\Sigma}$. Considering a vector of weights $\mathbf{w}$, the quadratic risk of the corresponding portfolio is given by $\mathbf{w}^\top \mathbf{\Sigma} \mathbf{w}$ and its expected return by $\mathbf{w}^\top \boldsymbol{\mu}$. In Markowitz mean-variance analysis, selecting an efficient portfolio amounts to solve the optimization problem

$$\text{maximize} \quad \mathbf{w}^\top \boldsymbol{\mu} / \sqrt{\mathbf{w}^\top \mathbf{\Sigma} \mathbf{w}}\,; \quad \mathbf{w} \in \mathbb{R}^p.$$

The solutions take the form $\mathbf{w} = \lambda \mathbf{\Sigma}^{-1} \boldsymbol{\mu}$, where $\lambda > 0$. For instance, if the quadratic risk is constrained to stay below a level R, we get $\lambda = \sqrt{R / \mathbf{w}^\top \mathbf{\Sigma}^{-1} \mathbf{w}}$, where $\mathbf{w}$ and $\mathbf{\Sigma}^{-1}$ must generally be estimated. El Karoui [2010] analyzed this problem in high dimension, but estimated the precision matrix by the inverse of the covariance matrix and noticed that it may lead to poor estimations when the considered matrices are not well conditioned.

All the proofs of the results presented in this introductory chapter are postponed in Appendix A.

0.1 Notation

For an unknown parameter θ we note θ^* its true value. As usual, $\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{\Sigma}^*)$ is the Gaussian distribution in $\mathbb{R}^p$ with mean $\boldsymbol{\mu}^*$ and covariance matrix $\mathbf{\Sigma}^*$. The corresponding precision matrix (the inverse of the covariance matrix, also known as concentration matrix) is denoted by $\mathbf{\Omega}^*$. The expectation of a random variable X is denoted by $\mathbf{E}(X)$ and its variance by $\mathbf{V}ar(X)$. The covariance between two random variables X and Y is expressed

by the notation $\mathbf{C}ov(X, Y)$. We denote by $\mathbf{1}_n$ the vector from $\mathbb{R}^n$ with all the entries equal to 1 and by $\mathbf{I}_n$ the $n \times n$ identity matrix. We write $\mathbb{1}$ for the indicator function, which is equal to 1 if the considered condition is satisfied and 0 otherwise. The cardinality of a set S is denoted by $|S|$. In what follows, $[p] := \{1, \ldots, p\}$ is the set of positive integers from 1 to p. For $j \in [p]$, the complement of the singleton $\{j\}$ in $[p]$ is denoted by j^c. For a vector $\mathbf{v} \in \mathbb{R}^p$, $\mathbf{D}_\mathbf{v}$ stands for the $p \times p$ diagonal matrix satisfying $(\mathbf{D}_\mathbf{v})_j = \mathbf{v}_j$ for every $j \in [p]$. The transpose of the matrix $\mathbf{M}$ is denoted by $\mathbf{M}^\top$. If this matrix is square, we note $\det(\mathbf{M})$ its determinant and $\mathrm{trace}(\mathbf{M})$ its trace. Furthermore, $\mathbf{M}^\dagger$ stands for its Moore-Penrose pseudo-inverse. The diagonal matrix obtained by zeroing all the off-diagonal entries of the $p \times p$ matrix $\mathbf{M}$ is denoted either by $\mathrm{diag}(\mathbf{M})$ or by $\mathrm{diag}(\{m_{jj}\}_{j\in[p]})$ to emphasize that the off-diagonal entries might be unknown or estimated separately. In addition, $\mathbf{M} \succ 0$ means that $\mathbf{M}$ is positive definite and $\mathbf{M} \succcurlyeq 0$ that the matrix is positive semidefinite. For a $n \times p$ matrix $\mathbf{M}$, the vector of the elements of the kth row (resp. the jth column) whose indexes are given by the subset J of $[p]$ (resp. K of $[n]$) is denoted by $\mathbf{M}_{k,J}$ (resp. $\mathbf{M}_{K,j}$). In particular, the vector made of all the elements of the jth column of the matrix $\mathbf{M}$ at the exception of the element of the kth row is given by $\mathbf{M}_{k^c,j}$. Moreover, the whole kth row (resp. jth column) of $\mathbf{M}$ is denoted by $\mathbf{M}_{k,\bullet}$ (resp. $\mathbf{M}_{\bullet,j}$). As is customary, we define the ℓ_q-norm of a vector $\mathbf{v} \in \mathbb{R}^p$ by $\|\mathbf{v}\|_q = \left\{\sum_{j=1}^p |\mathbf{v}_j|^q\right\}^{1/q}$, for $q > 0$. We denote $\max_{j\in[p]} |\mathbf{v}|$ by $\|\mathbf{v}\|_\infty$. We use the following notation for the (pseudo-)norms of matrices: if $q_1, q_2 > 0$, then

$$\|\mathbf{M}\|_{q_1,q_2} = \left\{\sum_{i=1}^n \|\mathbf{M}_{i,\bullet}\|_{q_1}^{q_2}\right\}^{1/q_2}.$$

With this notation, $\|\mathbf{M}\|_{2,2}$ and $\|\mathbf{M}\|_{1,1}$ are the Frobenius and the element-wise ℓ_1-norm of $\mathbf{M}$, respectively. Among other particular cases, for $q_1 = 1$, taking the limit when tends to infinity, $\|\mathbf{M}\|_{1,\infty}$ is the maximum absolute row sum norm defined by $\max_{i\in[n]} \|\mathbf{M}_{i,\bullet}\|_1$. In the same way, $\|\mathbf{M}\|_{\infty,1} = \max_{j\in[p]} \|\mathbf{M}_{\bullet,j}\|_1$ corresponds to the maximum absolute column sum norm (also known as ℓ_1-matrix norm) and $\|\mathbf{M}\|_{\infty,\infty} = \max_{(i,j)\in[n]\times[p]} |\mathbf{M}_{i,j}|$ to the max norm. In addition, we define $\sigma_{\max}(\mathbf{M})$ and $\sigma_{\min}(\mathbf{M})$, respectively, as the largest and the smallest singular values of the matrix $\mathbf{M}$. Its spectral norm is then denoted by $\|\mathbf{M}\|_2 = \sigma_{\max}(\mathbf{M})$. The context will serve to dispel any ambiguity with regard to the Euclidean vector norm. The sample covariance matrix of the data points $\{\mathbf{X}_{k,\bullet}\}_{k\in[n]}$ is defined by

$$\mathbf{S}_n = \frac{1}{n}\sum_{k=1}^n (\mathbf{X}_{k,\bullet}^\top - \widehat{\boldsymbol{\mu}})(\mathbf{X}_{k,\bullet}^\top - \widehat{\boldsymbol{\mu}})^\top = \frac{1}{n}(\mathbf{X} - \mathbf{1}_n\widehat{\boldsymbol{\mu}}^\top)^\top(\mathbf{X} - \mathbf{1}_n\widehat{\boldsymbol{\mu}}^\top),$$

where $\widehat{\boldsymbol{\mu}}$ is either the sample mean $\frac{1}{n}(\mathbf{1}_n^\top\mathbf{X})^\top$ (when the mean $\boldsymbol{\mu}^*$ is unknown) or the theoretical mean $\boldsymbol{\mu}^*$ (when it is considered as known).

0.2 Sparsity assumption

When we consider n observations drawn from a p-variate distribution, estimating the precision matrix implies to calculate $p(p+1)/2$ unknowns. If the number of unknowns exceeds the sample size, getting a statistically accurate estimation of $\mathbf{\Omega}^*$ is impossible. But it is possible under additional structural assumptions, for instance if we assume that most of its entries are actually equal to zero. This principle of parsimony was first—to the best of our knowledge— introduced and justified by Dempster [1972] to estimate the covariance matrix in moderate dimension for multivariate Gaussian distributions. It is based on the relationship between the entries of the precision matrix and the partial correlations. Recall that partial correlations corresponds to the correlations between two variables once removed the linear dependencies with all other variables. If we consider a Gaussian random vector X, any two random variables in X, for instance X_i and X_j, are independent conditionally to the other variables if and only if their partial correlation $\psi_{i,j}$ is zero. This result is a direct consequence of the following proposition.

Proposition 0.2.1. *Let X and Y be two random variables and Z be a random vector such that $(X, Y, Z^\top)^\top$ has a Gaussian distribution with the precision matrix*

$$\mathbf{\Omega} = \begin{pmatrix} \mathbf{\Omega}_{XX} & \mathbf{\Omega}_{XY} & \mathbf{\Omega}_{XZ} \\ \mathbf{\Omega}_{YX} & \mathbf{\Omega}_{YY} & \mathbf{\Omega}_{YZ} \\ \mathbf{\Omega}_{ZX} & \mathbf{\Omega}_{ZY} & \mathbf{\Omega}_{ZZ} \end{pmatrix}.$$

The following claims are equivalent:

i) X and Y are independent conditionally to Z,

ii) the partial correlation between X and Y is zero,

iii) the entry $\mathbf{\Omega}_{XY}$ is zero.

In the Gaussian framework, the precision matrix also exactly describes the conditional dependencies between pairs of variables given the values of all the other variables. The precision matrix is therefore often used for constructing a graph $\mathscr{G}^*$ of relationships between the p variables [Whittaker, 1990; Lauritzen, 1996]. Each node of this undirected graph corresponds to a variable and two nodes j and j' are connected by an edge if $\omega^*_{jj'} \neq 0$. In others words, the precision matrix may be understood as the adjacency matrix of the undirected graph $\mathscr{G}^*$. In recent work, Liu et al. [2009] extended the fact that the support of the precision matrix matches to the edges of the graph $\mathscr{G}^*$ to non-Gaussian distributions called nonparanormal [Liu et al., 2009]. For their part, Loh and Wainwright [2013] established the connection between some particular graphs and the precision matrix of discrete random variables.

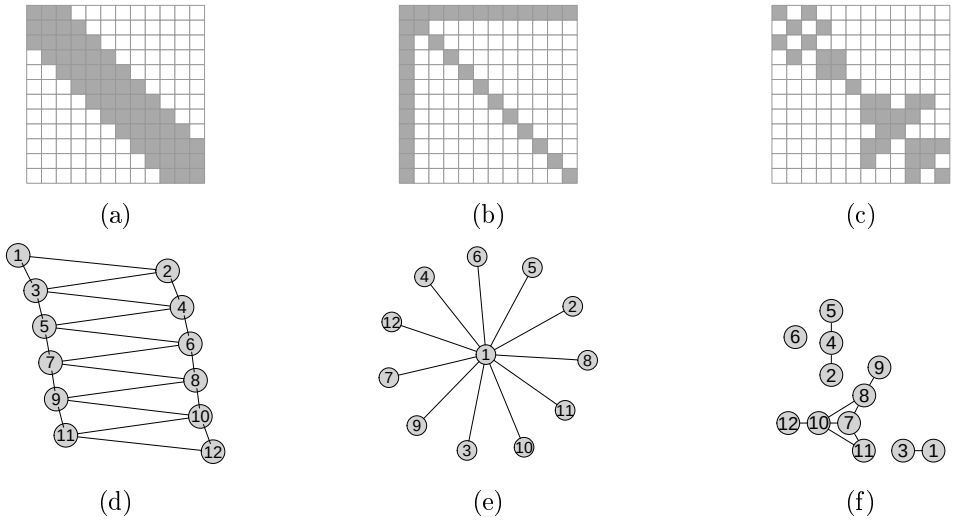


Figure 0.1: Sparsity patterns and corresponding graphs below.

The estimation of this graph, that is the identification of nonzero entries of the precision matrix, and the comprehensive estimation of the precision matrix are challenging recent statistical problems that arise considerable interest. In many applications, for instance those that have motivated this work, the sample size n tends to be much smaller than the dimension p of the model. To be able to estimate consistently the precision matrix or simply to recover its structure, we adopt the usual—since the works of Meinshausen and Bühlmann [2006] and Yuan and Lin [2007]—sparsity assumption for the graph $\mathscr{G}^*$. This assumption may mean that the total number of edges in $\mathscr{G}^*$ is small as compared to the number $p(p-1)/2$ of all possible pairs of variables. In this work, we rather assume that that the maximal degree of the nodes is much smaller than p. Hereafter, we say that a matrix is s-sparse if it has at most s nonzero entries per row/column.

Some problems require the estimation of the covariance matrix rather than the precision matrix, among the most noticeable, PCA. In high dimension, the sparsity of the covariance matrix is a common assumption [Bickel and Levina, 2008; El Karoui, 2008; Lam and Fan, 2009; Cai and Zhou, 2012]. Whereas a zero in the precision matrix corresponds to conditional independence, a zero in the covariance matrix signifies marginal independence in the Gaussian setting. As the conditional dependencies, the marginal dependencies can be represented by a graphical model (bidirected in this case) [Cox and Wermuth, 1993]. We recall that two random variables can be marginally independent, but dependent conditionally to a third variable. In other words, a zero in the covariance matrix does not implies a zero at the same position in the precision matrix. The reverse is also true: conditional independence does not imply marginal independence. Nevertheless, as noticed in [Cox and Wermuth, 1996], if the graph $\mathscr{G}^*$ is only composed of completely connected components (disjoint cliques), then the sparsity patterns of the precision and of the covariance matrices coincide.

0.3 Parsimonious precision matrix estimation

In recent years, the problem of precision matrix estimation under sparsity constraints received a lot of attention. We quickly present here the main estimators that have been proposed. Recall that unless the ratio p/n is very small[1], the inverse of the sample covariance matrix $\mathbf{S}_n$ is a very poor estimator of the precision matrix $\mathbf{\Omega}^*$. The developments that have occurred recently in sparse precision matrices estimation consider generally ℓ_1-penalized off-diagonal elements. The research works of Yuan [2010]; Cai et al. [2011]; Sun and Zhang [2013] and Ren et al. [2015] are among the most significant recent advances in statistical analysis of this approach. Put simply, the reason for considering ℓ_1-regularization is that it can be understood as a relaxed but convex form of an ℓ_0-regularization. A more complete justification is provided in Section 0.4.1.

In the field of graphical models, two distinct but related problems are naturally considered when proposing an estimator. Both selection consistency and estimation consistency are indeed relevant questions. On the one side, a satisfactory estimator should recover the edge structure of the graph, that is the sparsity pattern of the precision matrix. This question and the more involved sign consistency – considering that not only the zeros but also the signs of the entries should be correctly identified – received specific attention in many recent papers. They have in particular been studied by Meinshausen and Bühlmann [2006] whose objective is to identify the connected components of the graph $\mathscr{G}^*$. For their part, Zhao and Yu [2006]; Zou [2006]; Lounici [2008] and Meinshausen and Yu [2009] focused on variable selection using the least absolute shrinkage and selection operator (Lasso) in a linear regression model, and Lam and Fan [2009] or Ravikumar et al. [2011] analyzed the graphical Lasso. Selection consistency is obtained under the assumption that two variables belonging to distinct connected components of the graph $\mathscr{G}^*$ are not strongly correlated. It means that when the partial correlation between two variables is zero, then the magnitude of the (regular) correlation between these variables should be low. This assumption is called neighborhood stability in [Meinshausen and Bühlmann, 2006], irrepresentability in [Zhao and Yu, 2006] and termed mutual (in)coherence in [Bunea et al., 2007b] in reference to [Donoho et al., 2006]. On the other hand, when the estimation consistency is considered, one looks for an estimator $\widehat{\mathbf{\Omega}}$ converging to the true precision matrix with the fastest possible rate. This question is of main interest in our work and is developed in the rest of this section. Before getting into the specifics, we recall that it is easier for an estimator to achieve estimation consistency as it relies on weaker conditions than selection consistency.

[1] When p/n is very small, the covariance matrix can be estimated consistently under the spectral norm [Koltchinskii and Lounici, 2014]. That implies a similar result for the precision matrix. Indeed, when $p < n$, considering a data matrix with sub-Gaussian rows and using Vershynin [2012b, Theorem 5.39], $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_2$ has a convergence rate of order $\sigma_{\max}(\mathbf{\Sigma}^*)\sqrt{p}/\sqrt{n}$, with probability close to one. In the same way, $\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_2$ has a convergence rate of order $\rho(\mathbf{\Sigma}^*)\sqrt{p}/\sqrt{n}$, with probability close to one, $\rho(\mathbf{\Sigma}^*) = \sigma_{\max}(\mathbf{\Sigma}^*)/\sigma_{\min}(\mathbf{\Sigma}^*)$ being the condition number of $\mathbf{\Sigma}^*$.

Indeed, the regularity properties (see Section 0.5) can be deduced irrepresentable conditions [van de Geer and Bühlmann, 2009]. In some settings, these assumptions can be checked using concentration inequalities [Rudelson and Zhou, 2013; Dobriban and Fan, 2016].

We define here the rate of convergence of an estimator which is an essential concept in this work. An estimator $\widehat{\theta}_n$ converges over a set $\mathcal{M}$ under the risk associated with the loss function $l(\cdot)$ at the rate r_n if there exists a constant $C > 0$ such that, for any θ^* belonging to a set $\mathcal{M}$, for any sample size n

$$\mathbf{E}(l(\widehat{\theta}_n, \theta^*)) \leq C\, r_n.$$

In addition, we define the minimax risk over a set $\mathcal{M}$ as $\inf_{\widehat{\theta}_n} \sup_{\theta^* \in \mathcal{M}} \mathbf{E}\big(l(\widehat{\theta}_n, \theta^*)\big)$, taking the infimum over all possible estimators $\widehat{\theta}_n$. Then, the minimax convergence rate is the rate r_n for which there exist constants $C_1, C_2 > 0$, such that for all n

$$C_1\, r_n \leq \inf_{\widehat{\theta}_n} \sup_{\theta^* \in \mathcal{M}} \mathbf{E}\big(l(\widehat{\theta}_n, \theta^*)\big) \leq C_2\, r_n.$$

Any estimator $\widehat{\theta}_n$ that achieves the minimax convergence rate is called minimax. With these definitions, considering $\delta \in (0, 1)$ and any $\theta^* \in \mathcal{M}$, a rate r_n of an estimator $\widehat{\theta}_n$, that satisfies

$$\mathbf{P}\big(l(\widehat{\theta}_n, \theta^*) \leq C\, r_n\big) \geq 1 - \delta,$$

where $C > 0$, provides (by integration) a nonasymptotic upper bound for the minimax rate under $l(\cdot)$ loss function.

0.3.1 ℓ_1-penalized maximum likelihood estimation

Among the alternatives for estimating a sparse precision matrix, in the first place, we mention methods based on minimization of ℓ_1-penalized Gaussian likelihood (graphical Lasso methods), that have been studied in [Banerjee et al., 2008; d'Aspremont et al., 2008; Friedman et al., 2008]. Let us recall the rationale behind this estimator. Considering a $n \times p$ random matrix $\mathbf{X}$ with i.i.d. Gaussian rows with mean $\boldsymbol{\mu}^*$ and covariance $\boldsymbol{\Sigma}^*$, we set $\mathbf{S}_{n,\boldsymbol{\mu}} = (\mathbf{X} - \mathbf{1}_n \boldsymbol{\mu}^\top)^\top (\mathbf{X} - \mathbf{1}_n \boldsymbol{\mu}^\top)/n$ and denote by $\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Omega}, \mathbf{X})$ the Gaussian maximum likelihood function. We estimate the parameters of the distribution using the maximum likelihood estimator (MLE), that is

$$\{\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Omega}}\} = \underset{\substack{\boldsymbol{\Omega} \succ 0 \\ \boldsymbol{\mu} \in \mathbb{R}^p}}{\operatorname{argmax}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Omega}, \mathbf{X}) = \underset{\substack{\boldsymbol{\Omega} \succ 0 \\ \boldsymbol{\mu} \in \mathbb{R}^p}}{\operatorname{argmax}} \Big\{ \log\det(\boldsymbol{\Omega}) - \ \operatorname{trace}\big(\mathbf{S}_{n,\boldsymbol{\mu}} \boldsymbol{\Omega}\big) \Big\},$$

taking the maximum over all $p \times p$ matrices $\boldsymbol{\Omega}$. We can minimize separately with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$. It entails that $\boldsymbol{\mu}^*$ is estimated by the vector of sample means. Re-injecting

the closed-form expression $\boldsymbol{\mu} = \mathbf{X}^\top \mathbf{1}_n / n$ into the optimization problem leads to

$$\widehat{\boldsymbol{\Omega}} = \underset{\substack{\boldsymbol{\Omega} \in \mathbb{R}^{p \times p} \\ \boldsymbol{\Omega} \succ 0}}{\operatorname{argmax}} \Big\{ \log \det(\boldsymbol{\Omega}) - \operatorname{trace}\big(\mathbf{S}_n \boldsymbol{\Omega}\big) \Big\}.$$

To obtain a sparse estimate, we add a regularization term and define the estimator of the precision matrix as a solution to the problem

$$\widehat{\boldsymbol{\Omega}} = \underset{\substack{\boldsymbol{\Omega} \in \mathbb{R}^{p \times p} \\ \boldsymbol{\Omega} \succ 0}}{\arg\min} \Big\{ -\log \det(\boldsymbol{\Omega}) + \operatorname{trace}\big(\mathbf{S}_n \boldsymbol{\Omega}\big) + \lambda \|\boldsymbol{\omega}^{\text{off}}\|_1 \Big\}, \tag{0.3}$$

where $\boldsymbol{\omega}^{\text{off}}$ is the vector composed of the all the off-diagonal entries of $\boldsymbol{\Omega}$ and $\lambda > 0$ is a tuning parameter. This estimator has also been studied by Yuan and Lin [2007] and Rothman et al. [2008] in the case when $\mathbf{X}$ actually stems from a Gaussian distribution. As noticed in [Banerjee et al., 2008], the constraint $\boldsymbol{\Omega} \succ 0$ can be removed as already implied by the log determinant function. In [Banerjee et al., 2008; Friedman et al., 2008], $\|\boldsymbol{\omega}^{\text{off}}\|_1$, is replaced by $\|\boldsymbol{\Omega}\|_{1,1}$. Penalizing the diagonal elements seems surprising, but Ambroise et al. [2009] highlighted that in practice it leads the graphical Lasso algorithm to pick a definite positive estimator for $\boldsymbol{\Omega}^*$.

In [Lam and Fan, 2009], the graphical Lasso is generalized by considering non-convex functions to penalize the off-diagonal elements. The authors also show that the graphical Lasso actually recovers the sparsity structure of the precision matrix when the true matrix is sparse enough and when the sample size is large enough. Nevertheless, their assumptions are too severe to be workable in practice and have been alleviated by Ravikumar et al. [2011]. For instance, in this last paper, it is shown that the sample size only needs to be of order $s^2 \log p$ to recover the support of the precision matrix in the case of sub-Gaussian distributions. Besides, the same condition on n is sufficient to get a convergence rate of order $\sqrt{(S+p)(\log p)/n}$ in Frobenius norm, S being the number of nonzero off-diagonal elements in $\boldsymbol{\Omega}^*$. This rate is identical to these established by Rothman et al. [2008] and Lam and Fan [2009].

As regards the practical resolution of problem (0.3), noting the convexity of this problem d'Aspremont et al. [2008] have proposed to cast the primal problem in the form of a semidefinite program that can be solved using Nesterov's first order method [Nesterov, 2005]. They solved the dual problem by blockwise coordinate descent and called the resulting complete algorithm Covsel. In contrast, Yuan and Lin [2007] have chosen to use an interior point algorithm [Vandenberghe et al., 1998] to solve the same dual problem. However, their algorithm is applicable only if p is small. Starting from the findings of d'Aspremont et al. [2008], Friedman et al. [2008] have proposed to update recursively each column of the covariance matrix by solving the Lasso. The estimate of the precision matrix is obtained as a by-product using blockwise matrix inversion. This faster algorithm

is termed graphical Lasso. A different algorithm, with the same computational complexity of order $O(p^3)$, but based on the Cholesky decomposition of the precision matrix, has been proposed by Rothman et al. [2008].

0.3.2 A linear regression model

Another family of procedures pioneered by Meinshausen and Bühlmann [2006] relies on estimating the entries of $\mathbf{\Omega}^*$ by applying a regularized method for solving linear regression problems. This is the basis upon which our contributions are built. In this section, we provide basic justification for this approach and present existing work. A theoretical rationale for sparse regression is given in Section 0.4.

0.3.2.1 From regression coefficients to precision matrix

Next result rests on two prominent theorems, stated in Appendix A : the theorems on normal correlations [Marsaglia, 1964] and on block matrix inversion. It provides the background that supports this familly of procedures.

Proposition 0.3.1. *Let $(Y^\top, X^\top)^\top$ be a multivariate normal random vector with expectation $\boldsymbol{\mu} = (\boldsymbol{\mu}_Y^\top, \boldsymbol{\mu}_X^\top)^\top$ and covariance*

$$\mathbf{\Sigma} = \mathbf{\Omega}^{-1} = \begin{pmatrix} \mathbf{\Omega}_{YY} & \mathbf{\Omega}_{YX} \\ \mathbf{\Omega}_{XY} & \mathbf{\Omega}_{XX} \end{pmatrix}^{-1}.$$

We suppose that the relationship between the vectors Y and X is described by the linear regression model $Y = B^\top X + \epsilon$, where ϵ follows a zero-mean Gaussian distribution $\mathcal{N}(0, \Phi)$ and is independent of X. It holds that

$$\Phi = \mathbf{\Omega}_{YY}^{-1}, \qquad B = -\mathbf{\Omega}_{XY}\mathbf{\Omega}_{YY}^{-1}.$$

The precision matrix is therefore closely related to the problem of regression of one feature on all the others. Indeed, if we consider the linear regression models defined by taking alternately each column $\mathbf{X}_{\bullet,j}$ as the response and the others $\mathbf{X}_{\bullet,j^c}$ as covariates, we get

$$\mathbf{X}_{\bullet,j} = \mathbf{c}_j^* \mathbf{1}_n - \mathbf{X}_{\bullet,j^c}\mathbf{B}_{j^c,j}^* + \boldsymbol{\phi}_j^* \boldsymbol{\epsilon}_{\bullet,j}, \tag{0.4}$$

with $\mathbf{B}^*$ the $p \times p$ matrix of regression coefficients, two vectors $\mathbf{c}^*, \boldsymbol{\phi}^* \in \mathbb{R}^p$ and where $\boldsymbol{\epsilon}_{\bullet,j}$ is drawn from $\mathcal{N}_n(0, \mathbf{I}_n)$ and is independent of $\mathbf{X}_{\bullet,j^c}$. According to Theorem A.1.1 on normal correlations, the regression coefficients $\mathbf{B}_{j^c,j}^* \in \mathbb{R}^{p-1}$ and the standard deviation $\phi_j^* \in \mathbb{R}$ of residuals can be expressed in terms of the elements of the precision matrix $\mathbf{\Omega}^*$

as follows:

$$\mathbf{B}^*_{ij} = \omega^*_{ij}/\omega^*_{jj}, \qquad \boldsymbol{\phi}^*_j = (\omega^*_{jj})^{-1/2}, \tag{0.5}$$

whereas $\mathbf{c}^*_j = \boldsymbol{\mu}^*_j + (\boldsymbol{\mu}^*_{j^c})^\top \mathbf{B}^*_{j^c,j} = (\boldsymbol{\mu}^*)^\top \mathbf{B}^*_{\bullet,j}$. With this notation, the precision matrix can be written as $\boldsymbol{\Omega}^* = \mathbf{B}^* \mathbf{D}^{-2}_{\boldsymbol{\phi}^*}$.

0.3.2.2 ℓ_1-penalized linear regression

The method proposed by Yuan [2010] shares common framework with the method developed by Meinshausen and Bühlmann [2006] which is aimed to identify the nonzero entries of the precision matrix. Whereas variable selection relies on the Lasso in the approach of Meinshausen and Bühlmann [2006], the estimation of the precision matrix is based on the Dantzig selector [Candes and Tao, 2007] in the work of [Yuan, 2010]. His method consists in solving p linear problems of the form,

$$\text{minimize} \quad \|\boldsymbol{\beta}\|_1 \quad \text{subject to} \quad \|\mathbf{X}^\top_{\bullet,j^c}(\mathbf{X}_{\bullet,j} - \mathbf{X}_{\bullet,j^c}\boldsymbol{\beta})\|_\infty \leq \lambda\,; \quad \boldsymbol{\beta} \in \mathbb{R}^{p-1}, \tag{0.6}$$

to estimate $\mathbf{B}^*_{j^c,j}$, the columns of the matrix $\mathbf{B}^*$, excluding the diagonal entries, all equal to 1. The diagonal elements of $\boldsymbol{\Omega}^*$ are then estimated by the variances of regression residuals, using relations (0.5). Last, the resulting estimate is symmetrized by taking the closest symmetric matrix in ℓ_1-matrix norm. Fast convergence rates have been established for this procedure in ℓ_1-matrix norm and operator norm. These rates rely on assuming that $\boldsymbol{\Omega}^*$ is an s-sparse positive definite matrix with bounded eigenvalues and that $\|\boldsymbol{\Omega}^*\|_{\infty,1}$ is bounded. We note that these assumptions are of the same nature as those used to prove convergence rates for the Clime estimator (detailed below) and refer to a uniformity class of matrices initially defined in [Bickel and Levina, 2008] to estimate the covariance matrix in high dimension. However, despite the fact that most results in regression [Candes and Tao, 2007; Bickel et al., 2009] advocate for choosing the tuning parameter proportional to the noise level, this procedure does not follow this recommendation and considers the tuning parameter as dependent on the diagonal entries.

Sun and Zhang [2013] suggest to use the scaled Lasso [Sun and Zhang, 2012] (also known as square-root Lasso [Belloni et al., 2011]) instead of the Lasso to first estimate the off-diagonal elements of the matrix $\mathbf{B}^*$ of the coefficients of regression. This method has the advantage over the precedent of taking into account the fact that conditional variances are potentially heterogeneous. The square-root Lasso is defined as the solution to

$$\text{minimize} \quad \|\mathbf{X}_{\bullet,j} - \mathbf{X}_{\bullet,j^c}\boldsymbol{\beta}\|_2 + \lambda\|\boldsymbol{\beta}\|_1\,; \quad \boldsymbol{\beta} \in \mathbb{R}^{p-1}, \tag{0.7}$$

where $\lambda > 0$. In this setting, $\widehat{\boldsymbol{\beta}}$ estimates $-\mathbf{B}^*_{j^c,j}$. By replacing the sum of squared residuals by its square root, the problem becomes scale independent and the tuning parameter λ can

be chosen independently of the diagonal entries of $\mathbf{\Omega}^*$. As in [Yuan, 2010], the variances of regression residuals are used to estimate the diagonal entries of the precision matrix and the procedure ends with a symmetrization step. Sun and Zhang [2013] studied the theoretical properties of this estimate and also examined the properties of the OLS estimator obtained after variable selection by the scaled Lasso. Similar results are detailed in Section 0.6.1.

Following the ideas of Sun and Zhang [2013], Ren et al. [2015] propose an estimator of the individual entries of the precision matrix and study its asymptotic behavior. They show in particular that considering independent Gaussian observations, the elements $\widehat{\omega}_{ij}$ are asymptotically normal. Their method is analogous to the scaled Lasso with the sole exception that they consider a multivariate multiple linear regression model, composed of two response variables corresponding to a particular entry of the precision matrix and of $p-2$ explanatory variables that are related to the other dimensions. Instead of taking an interest in the slope of this model, they rather look into the noise level that only depends on ω^*_{ii}, ω^*_{jj} and ω^*_{ij} when i, j corresponds to the two response variables. They establish the minimax risk of estimating the individual entries over a set of sparse enough matrices and show that their estimator achieves an optimal convergence rate under the sparsity condition $s = O(n/\log p)$. Chen et al. [2015b] enlarge upon this approach on inference for low dimensional parameters. In the same spirit, Jankova et al. [2015] establish asymptotic properties in the sub-Gaussian setting for the graphical Lasso estimator.

As another alternative, the constrained ℓ_1-minimization for inverse matrix estimation (Clime) method [Cai et al., 2011], which is the matrix version of the Dantzig selector [Candes and Tao, 2007], solves the problem

$$\text{minimize} \quad \|\mathbf{\Omega}\|_1 \quad \text{subject to} \quad \|\mathbf{S}_n\mathbf{\Omega} - \mathbf{I}_p\|_\infty \leq \lambda\,; \quad \mathbf{\Omega} \in \mathbb{R}^{p\times p}, \tag{0.8}$$

where $\lambda > 0$ is a tuning parameter. For boosting running times, the semi-definite positiveness and symmetry constraints are relaxed in a way the solution can be obtained by solving p independent linear problems. The final estimator $\widehat{\mathbf{\Omega}}$ is obtained by symmetrizing the solution $\widetilde{\mathbf{\Omega}}$ of

$$\widetilde{\mathbf{\Omega}}_{\bullet,j} = \Big\{ \underset{\boldsymbol{\omega}\in\mathbb{R}^p}{\arg\min} \|\boldsymbol{\omega}\|_1 \quad \text{subject to} \quad \|\mathbf{S}_n\boldsymbol{\omega} - (\mathbf{I}_p)_{\bullet,j}\|_\infty \leq \lambda \Big\}. \tag{0.9}$$

Despite not being explicitly based on linear regression, this method is not very different in conception from the procedure of Yuan [2010] (see Appendix A). Soon afterward, the AClime (Adaptive Clime) method [Cai et al., 2012] has been proposed to address the same problem that have motivated the use of the square-root Lasso in place of the Lasso in [Sun and Zhang, 2013]. The penalization parameter of the constrained ℓ_1-minimization for inverse matrix estimation (Clime) procedure is indeed not adapted when the diagonal entries are heterogeneous. The new approach proceeds in two steps. Each of the diagonal

entries is first estimated using an adaptive version of the Dantzig selector. Then, the estimation of the off-diagonal elements relies on a modified version of the Clime in which the previously estimated diagonal entries are injected. In each step, the threshold used in the Dantzig selector depends on the entry being computed. We highlight that in spite of the diagonal elements are first estimated to take into consideration the possible heterogeneity of diagonal entries, it is not done in connection with the coefficients of regression as in [Yuan, 2010] or [Sun and Zhang, 2013]. Cai et al. [2012] first establish minimax lower bounds for the estimation of the precision matrix using a technique developed in [Cai and Zhou, 2012]. They then show that the convergence rate of their estimator is optimal in the sense that it could not be improved by any other estimator of the precision matrix that belongs to the same uniformity class of matrices that is considered in [Yuan, 2010] and [Cai et al., 2011].

As for algorithmic aspects, all the approaches based on linear regression afford the advantage of being related to an optimization problem which can be solved considering p independent smaller problems. These topics are discussed in next Section 0.4. Let us simply make one comment about the similarities between the graphical Lasso of Friedman et al. [2008] and the scaled Lasso of Sun and Zhang [2013] that rely both on ℓ_1-penalization. These methods differ essentially by two points : the cost function has a squared loss term in [Friedman et al., 2008] whereas not in [Sun and Zhang, 2013], moreover this loss term is modified at each iteration in the algorithm of Friedman et al. [2008], not in [Sun and Zhang, 2013]. More precisely, Sun and Zhang [2013] consider the loss term $\|\mathbf{X}_{\bullet,j} - \mathbf{X}_{\bullet,j^c}\boldsymbol{\beta}\|_2$ whereas Friedman et al. [2008] consider $\|\mathbf{S}_{j^c,j^c}^{-1/2}\mathbf{S}_{j^c,j} - \mathbf{S}_{j^c,j^c}^{1/2}\boldsymbol{\beta}\|_2^2$, where $\mathbf{S}$ estimates $\boldsymbol{\Sigma}^*$ and is updated during the estimation procedure.

0.4 Advances in sparse linear regression

Regularization procedures have initially been designed for variable selection to prevent overfitting, thus to improve the generalization/prediction performance of the model and its interpretability. For this purpose, they are an alternative to classic procedures such as stepwise regression (forward/backward selection) that sequentially removes/adds variables from/to the model, or widespread criteria such as Mallows's C_p [Mallows, 1973], Akaike information criterion (AIC) [Akaike, 1974] or Bayesian information criterion (BIC) [Schwarz, 1978]. In this section, we present some prominent methods for regularized linear regression upon which our work relies. In particular, we provide guarantees for the estimation consistency of the square-root Lasso.

0.4.1 Sparse least squares regression

In this section, we consider the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{y} \in \mathbb{R}^n$ is the response variable, $\mathbf{X}$ the $n \times p$ design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ the vector of the coefficients of regression and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ the vector of error terms. We assume that $\boldsymbol{\epsilon} \sim \mathcal{N}_n(0, \phi^2\mathbf{I}_n)$, but other distributions of errors can be considered. A usual way to estimate the coefficients of a linear regression model is to solve the least squares regression problem. When the dimension p is smaller than the sample size n, if $\mathbf{X}$ is full rank, the OLS estimator is unique and given by $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. The OLS estimator is known to be unbiased and to have the lowest variance among all linear unbiased estimators (BLUE) according to the Gauss-Markov Theorem. However, when $\mathbf{X}$ is not full rank, which is always the case when $n < p$, the solution to the least squares regression problem is not unique. In this case, when the rank is not too low, a common alternative is to use the Moore-Penrose pseudo-inverse. The OLS estimate is then $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^\dagger\mathbf{X}^\top\mathbf{y}$. In high dimension, the rank of $\mathbf{X}$ is bounded by the sample size and the least squares regression problem is very ill-posed. Among all approaches developed to gain in interpretability and in prediction the reformulation of the least squares problem by adding a regularization term is very appealing. In particular, assuming that not all the explanatory variables are needed to fit the linear model leads to consider that only a subset of the coefficients of regression are nonzero and have to be estimated. The most natural penalization that produces a sparse $\boldsymbol{\beta}$ estimate uses ℓ_0-norm. Indeed, $\|\boldsymbol{\beta}\|_0$ corresponds to the number of nonzero elements of $\boldsymbol{\beta}$. Foster and George [1994] show that Mallows's C_p, AIC and BIC are special cases of a selection procedure that minimizes least squares with ℓ_0-norm regularization. These criteria can indeed be written on the form

$$\widehat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\phi^2\|\boldsymbol{\beta}\|_0 \right\},$$

with specific values of the parameter $\lambda > 0$ and known ϕ^2. Foster and George [1994] also extended that procedure to consider the case when ϕ^2 is unknown. Nevertheless, penalizing by the ℓ_0-norm of $\boldsymbol{\beta}$ leads to a non-convex optimization problem. Besides, we recall that regularizing with ℓ_q-norms leads to sparse solutions if $q \leq 1$, but only when $q \geq 1$ the objective function is convex. Choosing $q = 1$ is thus the only configuration that likely results in a sparse estimate while solving a convex optimization problem. The Lasso of Tibshirani [1996] is defined as

$$\widehat{\boldsymbol{\beta}}^{\text{Lasso}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{Lasso}$$

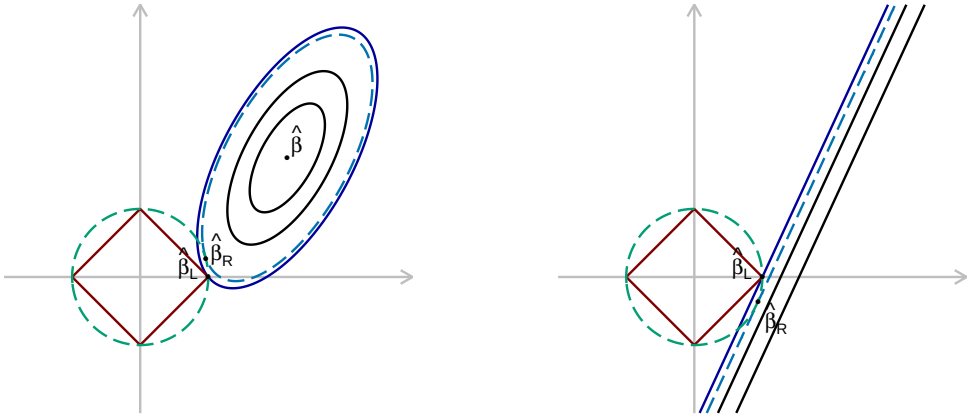
where $\lambda > 0$ is a tunning parameter that controls sparsity.

Figure 0.2 illustrates the addition to OLS of ℓ_1 or ℓ_2 penalization on the vector of coefficients. The latter refers to the ridge regression [Hoerl and Kennard, 1970] – also known as

Tikhonov regularization method – that corresponds to

$$\widehat{\boldsymbol{\beta}}^{\text{Ridge}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \right\}. \tag{Ridge}$$

In this formulation, the penalization parameter $\lambda > 0$ encourages the coefficients of regression to be shrunk towards zero as in the Lasso. However, while ℓ_1 penalization produces an estimate where some coefficients are actually zero, the solution to the least squares problem with ℓ_2 penalization is not sparse. Note that Eq. (Ridge) has an explicit solution $\widehat{\boldsymbol{\beta}}^{\text{Ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y}$. This estimator is biased but has the advantage over the OLS estimator of having a smaller variance. It hence may have a lower mean squared error (MSE) than the OLS estimator.



(a) (b)

Figure 0.2: Effects of ℓ_1 and ℓ_2 regularizations on OLS. The plots are drawn for $p = 2$, where $n \geq p$ (on the left) and $n = 1 < p$ (on the right). Each ellipse/straight line corresponds to a value of the residual sum of squares (RSS). The OLS estimate is denoted by $\widehat{\boldsymbol{\beta}}$. The Lasso estimate $\widehat{\boldsymbol{\beta}}^{\text{L}}$ is at the intersection of the solid ellipse/straight line and the solid square representing the ℓ_1-ball. The ridge regression estimate $\widehat{\boldsymbol{\beta}}^{\text{R}}$ is at the intersection of the dashed ellipse/straight line and the dashed circle representing the ℓ_2-ball.

Note however that while the ℓ_0-penalty is completely scale free, the ℓ_1-penalty is strongly scale dependent. Put differently, the ℓ_0-norm of two candidate estimates of the coefficients of regression can be equal, while their ℓ_1-norms are significantly different. The ability to estimate properly the scale of the coefficients of regression using the Lasso is determined by the choice of the tuning parameter λ.

As the estimator of the ridge regression, the Lasso is biased. In practice, this shrinkage bias is often reduced by using the Lasso for variable selection, and then computing the OLS estimate on selected variables only.

The theoretical properties of the Lasso procedure for estimation have been analyzed in

depth in [Bickel et al., 2009]. When the covariance matrix satisfies appropriate regularity conditions, the Lasso solution converges towards $\boldsymbol{\beta}^*$ at the rate $s\sqrt{\log(p)/n}$ in ℓ_1-vector norm and at the rate $\sqrt{s\log(p)/n}$ in ℓ_2-vector norm. It is worth to recall that the Lasso and the Dantzig selector (0.6) [Candes and Tao, 2007] are closely linked, not only for linear or nonparametric regression models [Bickel et al., 2009], but also for density models [Bunea et al., 2007a; Bertin et al., 2011].

0.4.2 Square-root Lasso

Ignoring that conditional variances $\mathbf{V}ar(\mathbf{X}_{k,j}|\mathbf{X}_{k,i}, i \neq j) = 1/\omega^*_{jj}$ are potentially heterogeneous may lead to scale errors in estimation and then results in a complete failure of the model selection. That is the reason why we choose to use the square-root Lasso that does not depend on the variance of the errors to estimate the entries of the matrix of the coefficients of regression.

0.4.2.1 Formulation

We consider the square-root Lasso estimator, introduced by Belloni et al. [2011], which is indeed the same estimator as the scaled Lasso developed by Sun and Zhang [2012]. The two estimators are defined as follows:

$$\widehat{\boldsymbol{\beta}}^{\sqrt{\text{Lasso}}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \Big\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \lambda\|\boldsymbol{\beta}\|_1 \Big\}, \qquad \text{(square-root Lasso)}$$

$$\{\widehat{\boldsymbol{\beta}}^{\text{scL}}, \widehat{\phi}^{\text{scL}}\} \in \underset{\substack{\boldsymbol{\beta}\in\mathbb{R}^p \\ \phi\in]0,+\infty[}}{\arg\min} \Big\{ \frac{1}{2\phi\sqrt{n}}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\sqrt{n}}{2}\phi + \lambda\|\boldsymbol{\beta}\|_1 \Big\}. \qquad \text{(scaled Lasso)}$$

The square-root Lasso is derived from Lasso by replacing the sum of squared residuals by its square root in the cost function. The problem thus becomes scale independent and the tuning parameter $\lambda > 0$ can be chosen independently of the noise variance of ϕ^{*2}.

Proposition 0.4.1. *The vector $\widehat{\boldsymbol{\beta}}^{\text{scL}}$ is a solution of the optimization problem (square-root Lasso) and, conversely, the pair $(\widehat{\boldsymbol{\beta}}^{\sqrt{\text{Lasso}}}, \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{\sqrt{\text{Lasso}}}\|_2/\sqrt{n})$ is a solution of (scaled Lasso).*

Using this result, we denote indistinctly hereafter both estimators by $\widehat{\boldsymbol{\beta}}$. We note that the slope and the variance of the error are interconnected . The equivalence of the scaled Lasso and the square-root Lasso is obtained when the variance of the error is estimated by the residual variance, that is

$$\widehat{\phi} = \frac{1}{\sqrt{n}}\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|_2.$$

0.4.2.2 Risk bounds on estimation error of the coefficients of regression

As the square-root Lasso has been designed to estimate the coefficients of regression in high dimension, assuming that most of these coefficients are indeed zero, it is important to check that this estimator actually converges towards $\boldsymbol{\beta}^*$. The risk bounds of the estimator are obtained assuming mainly that the sample size n is large enough and that the matrix $\mathbf{X}$ has some regularity properties. In this paragraph, we mainly take up the results of Belloni et al. [2011], adapting them to rely on the more general sensibilities assumptions introduced in [Gautier and Tsybakov, 2011] and [Gautier and Tsybakov, 2013], rather than on the restricted eigenvalues assumptions essentially developed in [Bickel et al., 2009; van de Geer, 2007]. It is worth pointing out that the risk bounds we present are of the same nature as those shown for the Lasso by Bickel et al. [2009] in their seminal paper and obtained under similar conditions.

As the risk bounds of the error of estimation are connected with the bounds of $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$, they are also naturally related to the eigenvalues of $\widehat{\boldsymbol{\Sigma}}$. The sample covariance matrix is nonnegative definite, but is always singular when $p > n$. In that case, its smallest eigenvalue is zero and of multiplicity at least $p-n$. However, if $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ belongs to a restricted set (a cone of the type of that defined in Eq. (0.12)), which is true when the regularization parameter λ is suitably lower bounded, then the ability to bound the error of estimation depends on much weaker regularity properties than the nonzero-ness of the smallest eigenvalue. Checking these properties is nontrivial, but they are fulfilled with high probability in many settings. The key argument is that under proper conditions – if the true covariance matrix is well estimated by the sample covariance matrix – the regularity properties satisfied by the population covariance matrix are also satisfied by the sample covariance matrix. In the (sub-)Gaussian setting, one easily see that such conditions are satisfied when n is large enough.

The following propositions provide finite-sample bounds on the rate of convergence—for ℓ_1 estimation loss and for ℓ_2 prediction loss—of the square-root Lasso estimator when the observations are drawn from a Gaussian distribution.

Proposition 0.4.2. *Set $s = |supp(\boldsymbol{\beta}^*)|$, $1 \leq s \leq p$. Let $\mathbf{X}$ be a $n \times p$ random matrix with i.i.d. centered Gaussian rows whose covariance matrix has unit diagonal entries and satisfies the ℓ_1-sensitivity property $\kappa^*(s, 2, 1) > 0$. Let us consider $\alpha \in (0, 1)$, $\delta \in (0, 1)$ and choose*

$$\lambda = 9\Big(\log \frac{6p}{\delta}\Big)^{1/2}.$$

Let d be a positive constant. We assume that the sample size n satisfies

$$n \geq \Big(12 \log(3/\delta)\Big) \vee \Big(4\lambda^2/\kappa^*(s, 2, 1)\Big) \vee \Big(ds^2 \log(1/\alpha)\Big).$$

Set $A = 32/\kappa^*(s,2,1)$ *and* $B = 14\Big(\kappa^*(s,2,1)\Big)^{-1/2}$.
Then, the solution $\widehat{\boldsymbol{\beta}}$ *of problem* (square-root Lasso) *satisfies*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_q \leq \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq A\frac{\lambda\phi^*}{\sqrt{n}}, \quad \textit{for } q \geq 1, \quad \textit{and} \quad \|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \leq B\lambda\phi^*, \tag{0.10}$$

with probability at least $1 - \delta - \alpha$.

We state and prove these results only in the case of Gaussian design, but they can be extended to a more general class of true covariance matrices, as they are related to Theorem 0.5.2. The bounds presented in Proposition 0.4.2 are not sharp, in the sense that the constants are not optimized. Please note that the sensitivity condition (see Section 0.5 below) bears on $\boldsymbol{\Sigma}^*$, not on the sample covariance matrix. The sparsity level of $\boldsymbol{\beta}^*$ does not appear explicitly in the bounds as we consider the ℓ_1-sensitivity property rather than the restricted eigenvalue condition. However, in view of Proposition 0.5.1, these bounds can be formulated using the restricted eigenvalue or the compatibility conditions to obtain that $\widehat{\boldsymbol{\beta}}$ converges in ℓ_1-vector norm at most at a rate of order $s\sqrt{\log(p)/n}$, hence of the same order as that of the Lasso [Bickel et al., 2009]. As noticed in [Belloni et al., 2011], as opposed to the Lasso, this bound is obtained without needing the variance of the error to be known, but under the additional assumptions $n \geq \lambda^2/\big((\kappa^*(s,2,1) - \iota)\rho\big)$ and $n \geq ds^2 \log 1/\alpha$ on the sample size. We highlight that the bound established in Proposition 0.4.2 for the error of estimation measured in ℓ_2-norm can be improved under a slightly stronger regularity property. As that of Proposition 0.4.2, the claim of Proposition 0.4.3 can be generalized to other settings. Indeed, Theorem A.1.5 used to prove this proposition has been extended to non-Gaussian design (see for instance [Rudelson and Zhou, 2013]).

Proposition 0.4.3. *Set* $s = |supp(\boldsymbol{\beta}^*)|$, $1 \leq s \leq p$. *Let* $\mathbf{X}$ *be a* $n \times p$ *random matrix with i.i.d. centered Gaussian rows whose covariance matrix has unit diagonal entries and satisfies the restricted eigenvalue property* $\bar{\kappa}^{*RE}(s,2) > 0$. *Let us consider* $\alpha \in (0,1)$, $\delta \in (0,1)$ *and choose*

$$\lambda = 9\Big(\log\frac{6p}{\delta}\Big)^{1/2}.$$

Let us consider the universal constants $a, b, d > 0$. *We assume that the sample size* n *satisfies*

$$n \geq \Big(12\log(3/\delta)\Big) \vee \Big(as\lambda^2/\bar{\kappa}^{*RE}(s,2)\Big) \vee \Big(1/d\log(b/\alpha)\Big).$$

Set $C = 128\Big(1 + 2\sqrt{s/n}\Big)/\bar{\kappa}^{*RE}(s,2)$.
Then, the solution $\widehat{\boldsymbol{\beta}}$ *of problem* (square-root Lasso) *satisfies*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq C\sqrt{s}\frac{\lambda\phi^*}{\sqrt{n}}, \tag{0.11}$$

with probability at least $1 - \delta - \alpha$.

0.4.3 Group Lasso

We complete this review of procedures that involve regularizing the vector $\boldsymbol{\beta}$ of coefficients of regression, by mentioning the group Lasso, considered in [Yuan and Lin, 2006; Meier et al., 2008]. Instead of considering a coefficient-by-coefficient ℓ_1-penalty, this method consists in applying such a penalization on groups of coefficients. This amounts to suppose that the vector $\boldsymbol{\beta}$ is sparse with a known underlying group structure. Whereas the Lasso procedure tends to propose an estimate for which the entries related to certain explanatory variables are zero, the group Lasso estimates $\boldsymbol{\beta}$ such that either all the coefficients of a group are zero, or all of them are nonzero.

We consider the same linear model as above and suppose that the vector of regression coefficients is expressed as the concatenation of G groups $\boldsymbol{\beta} = (\boldsymbol{\beta}_{|1}^\top, \ldots, \boldsymbol{\beta}_{|G}^\top)^\top$. The group Lasso estimator is then defined by

$$\widehat{\boldsymbol{\beta}}^{\text{grpLasso}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{g=1}^{G} w_g \|\boldsymbol{\beta}_{|g}\|_2 \right\}, \qquad \text{(group Lasso)}$$

where $w_g > 0, g \in [G]$ are the weights associated to the coefficients of each group. The replacement of the ℓ_1-norm by the mixed ℓ_2/ℓ_1-norm promotes group sparsity. Without going into too much detail, note that asymptotic consistency of the group Lasso has been studied by Bach [2008] while Lounici et al. [2011] proved nonasymptotic oracle inequalities. Further developments of this approach have been proposed. Among others, we mention the estimators introduced by Friedman et al. [2010a], Chiquet et al. [2012] and Obozinski et al. [2011]. The first method (called sparse group Lasso, see also [Sprechmann et al., 2011]) aims to enforce sparsity within groups, by adding an ℓ_1-regularization term to the group Lasso. In the second (termed cooperative Lasso), the ℓ_2-norm that is part of the mixed ℓ_2/ℓ_1-norm is replaced by the sum of the ℓ_2-norm of the positive and negative parts of $\boldsymbol{\beta}$ to ensure sign-coherence. Last, Obozinski et al. [2011] developed the multivariate group Lasso that extends the group Lasso to the multivariate regression model.

The approach we developed to take into account the potential presence of outliers in the Gaussian graphical model is based on an assumption of structured sparsity of the same nature, while pursuing a substantially different purpose. Indeed, when estimating the precision matrix in presence of outliers, the groups on which relates the sparsity assumption do not form a partition of the set of variables, but are composed of all values taken by variables for a single observation.

0.4.4 Optimization algorithms for penalized regression

In this section, we review briefly the algorithms that can be implemented to solve a penalized regression problem, in particular the Lasso and the square-root Lasso. We begin with

two important remarks. First, all the regularized problems that we consider in this section (Lasso, ridge, square-root Lasso and group Lasso) have convex objective functions. Each of these primal problems can therefore be associated with a dual problem, hence allowing to form a stopping criterion using duality gap [d'Aspremont et al., 2008]. Second, as all these problems depend on a regularization parameter λ, the computed solution also depends on the chosen value for this parameter.

Active-set algorithms: taking into consideration that only a subset of the parameters are nonzero, these algorithms produce a regularization path for all the values of the tuning parameter, by updating this subset. For instance, the least angle regression (LARS) has been proposed in [Efron et al., 2004] to solve the Lasso.

Coordinate descent algorithms: these algorithms update successively each coordinate of the solution [Friedman et al., 2007, 2010b]. We note that this approach has been efficiently randomized [Shalev-Shwartz and Tewari, 2011; Nesterov, 2012; Richtárik and Takávecc, 2014].

First-order methods: among these methods, the subgradient descent algorithm is a very general-purpose optimization algorithm that can be used, but converges slowly. More efficient proximal gradient methods have been proposed, for instance Nesterov's accelerated gradient descent [Nesterov, 2007] or iterative shrinkage-thresholding algorithms like FISTA [Beck and Teboulle, 2009]. In addition, noticing that the Lasso and the square-root Lasso can be cast as a second-order cone program (SOCP), they can for instance be solved using the first-order operator splitting method developed by O'Donoghue et al. [2013] that scales well with large problems (implemented in the Splitting Conic Solver).

Interior-point methods: despite not being the fastest to solve such optimization problems, these methods are worth to be considered as being implemented in most of the standard solvers. This approach has been specialized to cope better with regularized regression problems, for instance by Kim et al. [2007] for the Lasso.

We refer the reader to [Bach et al., 2012] for a more exhaustive review of optimization algorithms for regularized problems. Last but not least, we cite recent developments on safe rules to detect the coefficients that are surely zero and to exclude the corresponding variables before optimization [El Ghaoui et al., 2012; Fercoq et al., 2015]. This techniques, well adapted to the coordinate descent algorithm among others, reduce the required computational time.

0.5 Regularity properties of the design matrices

0.5.1 Review of regularity properties

In the field of high dimensional estimation, and in particular in sparse regression, controlling the error of the estimate relies on some regularity conditions that have to be satisfied by the design matrix $\mathbf{X}$. The challenge is of course to provide analytical bounds based on the weakest condition, that is, the most likely to be satisfied.

All the following conditions are defined for an observed design matrix $\mathbf{X}$. Later, we will assume that the regularity property relates to a random matrix, or by a slight abuse of terminology, to the corresponding Gram matrix $\mathbf{X}^\top\mathbf{X}/n$ which is the sample covariance matrix when the observations are centered.

For a subset J of $[p]$ and $c > 0$, we introduce the cone

$$\mathscr{C}_J(c) \triangleq \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}_{J^c}\|_1 \leq c\|\boldsymbol{\delta}_J\|_1\}. \tag{0.12}$$

As in [Gautier and Tsybakov, 2013], for $s \in [p]$ and $c > 0$, for $q \in \mathbb{N}_*$, the ℓ_q-sensitivity is defined by

$$\kappa(s, c, q) \triangleq \min_{\substack{J\subset[p],\\ |J|=s}} \quad \min_{\substack{\boldsymbol{\delta}\in\mathscr{C}_J(c),\\ \boldsymbol{\delta}\neq 0}} \quad \frac{1}{n}\frac{\|\mathbf{X}^\top\mathbf{X}\boldsymbol{\delta}\|_\infty}{\|\boldsymbol{\delta}_J\|_q}. \tag{0.13}$$

The assumption $\kappa(s, c, q) > 0$ is of the same nature as, but more general than the restricted eigenvalue property. Recall that the restricted eigenvalue defined as

$$\kappa^{RE}(s, c) \triangleq \min_{\substack{J\subset[p],\\ |J|=s}} \quad \min_{\substack{\boldsymbol{\delta}\in\mathscr{C}_J(c),\\ \boldsymbol{\delta}\neq 0}} \quad \frac{1}{n}\frac{\|\mathbf{X}\boldsymbol{\delta}\|_2^2}{\|\boldsymbol{\delta}_J\|_2^2}, \tag{0.14}$$

has been introduced by van de Geer [2007]; Bickel et al. [2009] to obtain oracle inequalities for the Lasso and the Dantzig selector. The restricted eigenvalue condition $\kappa^{RE}(s, c) > 0$ can be understood as the requirement that $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$ has to be "positive definite" on the restricted set of vectors δ satisfying $\|\delta_{J^c}\|_1 \leq c\|\delta_J\|_1$. For the sake of completeness, we note that similar conditions have also been considered by Zhang and Huang [2008] under the name of sparse Riesz condition and by Meinshausen and Yu [2009] as sparse eigenvalues condition. We also recall that the restricted isometry property, first introduced in [Candes and Tao, 2005], has been used in [Candes and Tao, 2007] to prove nonasymptotic oracle inequalities for the Dantzig selector. In addition, we define

$$\bar{\kappa}^{RE}(s, c) \triangleq \min_{\substack{J\subset[p],\\ |J|=s}} \quad \min_{\substack{\boldsymbol{\delta}\in\mathscr{C}_J(c),\\ \boldsymbol{\delta}\neq 0}} \quad \frac{1}{n}\frac{\|\mathbf{X}\boldsymbol{\delta}\|_2^2}{\|\boldsymbol{\delta}\|_2^2}. \tag{0.15}$$

The related condition $\bar{\kappa}^{RE}(s, c) > 0$ implies $\kappa^{RE}(s, c) > 0$. This condition is useful to

establish a fast convergence rate under Euclidean norm for the Lasso or the square-root Lasso.

We also define the compatibility constant (used for example in Belloni et al. [2011]) as

$$\kappa^C(s,c) \triangleq \min_{\substack{J\subset[p],\\ |J|=s}} \quad \min_{\substack{\boldsymbol{\delta}\in\mathscr{C}_J(c),\\ \boldsymbol{\delta}\neq 0}} \quad \frac{1}{n}\frac{|J|\|\mathbf{X}\boldsymbol{\delta}\|_2^2}{\|\boldsymbol{\delta}_J\|_1^2}. \tag{0.16}$$

Since $\|\mathbf{u}_J\|_1^2 \leq |J|\|\mathbf{u}_J\|_2^2$ for any $\mathbf{u}\in\mathbb{R}^p$, the condition $\kappa^C(s,c)>0$ is a relaxed version of restricted eigenvalue properties.

Note that all these properties are indeed defined with regard to the matrix $\mathbf{X}$ which is generally clearly identifiable from the context and thus does not appear in notation. The following proposition clarifies the link between the four precedent assumptions [Gautier and Tsybakov, 2013, Lemma 4.1].

Proposition 0.5.1. *Let us consider a deterministic matrix* $\mathbf{X}$. *For any* $s\in[p]$ *and* $c>0$, *it holds that*

$$\kappa(s,c,1) \geq s^{-1}(1+c)^{-1}\kappa^C(s,c) \geq s^{-1}(1+c)^{-1}\kappa^{RE}(s,c). \tag{0.17}$$

0.5.2 Checking the sensitivity property

This section is devoted to the verification of regularity conditions that have to be satisfied by the sample covariance matrix to ensure the fast convergence of ℓ_l-regularized procedures, such as Lasso, Dantzig selector or square-root Lasso. Significant effort has gone into understanding under which conditions the restricted eigenvalue property holds for the sample covariance matrix. We refer to [Rudelson and Zhou, 2013] for results for a general class of design matrices $\mathbf{X}$ having independent sub-Gaussian rows, but not necessarily i.i.d., and to [van de Geer and Bühlmann, 2009] and [Raskutti et al., 2010] for earlier results. We are interested in the conditions under which the fact that the population covariance matrix $\boldsymbol{\Sigma}^*$ has a sensitivity property involves that the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ satisfies a sensitivity property as well.

We define the ℓ_q-sensitivity on the true covariance matrix, for $q\in\mathbb{N}_*$ by

$$\kappa^*(s,c,q) \triangleq \min_{\substack{J\subset[p],\\ |J|=s}} \quad \min_{\substack{\boldsymbol{\delta}\in\mathscr{C}_J(c),\\ \|\boldsymbol{\delta}_J\|_q=1}} \quad \|\boldsymbol{\Sigma}^*\boldsymbol{\delta}\|_\infty. \tag{0.18}$$

As its counterpart for the sample covariance matrix, the ℓ_q-sensitivity property $\kappa^*(s,c,q)>0$ is implied by the restricted eigenvalue property $\bar{\kappa}^{*RE}(s,c)>0$, where

$$\bar{\kappa}^{*RE}(s,c) \triangleq \min_{\substack{J\subset[p],\\ |J|=s}} \quad \min_{\substack{\boldsymbol{\delta}\in\mathscr{C}_J(c),\\ \|\boldsymbol{\delta}\|_2=1}} \quad \|\boldsymbol{\Sigma}^{*1/2}\boldsymbol{\delta}\|_2^2. \tag{0.19}$$

We point up that if the true covariance matrix is positive definite, then these regularity conditions hold.

We simply state here sufficient conditions for the ℓ_1-sensitivity property to hold for a random matrix $\mathbf{X}$ having certain characteristics (for instance with i.i.d. sub-Gaussian rows). To the best of our knowledge, the following theorem, due to Dobriban and Fan [2016, Theorem 2], gives the most general conditions that ensure that the ℓ_q-sensitivity property holds with high probability for a random matrix $\mathbf{X}$.

Theorem 0.5.2 (Dobriban and Fan [2016]). *Let $\mathbf{X}$ be a $n \times p$ random matrix having zero-mean rows with a covariance matrix $\mathbf{\Sigma}^*$ satisfying the ℓ_q-sensitivity property $\kappa^*(s,c,q) > 0$. Let $\kappa(s,c,q)$ be the ℓ_q-sensitivity of the empirical covariance matrix $\widehat{\mathbf{\Sigma}}$. For any $a > 0$ and $\iota \in \big(0, \kappa^*(s,c,q)\big)$, there exists a constant d (that depends on ι and a but is independent of n, p and s) such that*

i) if $\mathbf{X}$ has i.i.d. sub-Gaussian rows and $n \geq ds^2 \log(2p^2)$, then $\kappa(s,c,q) > \kappa^*(s,c,q) - \iota > 0$ holds with probability at least $1 - (2p^2)^{-a}$,

ii) if there exists a real number b such that $\|\mathbf{X}\|_{\infty,\infty} \leq b$ and $n \geq ds^2 \log(2p^2)$, then $\kappa(s,c,q) > \kappa^*(s,c,q) - \iota > 0$ holds with probability at least $1 - (2p^2)^{-a}$,

iii) if there exists a real number b and a positive integer r such that $\mathbb{E}(|\mathbf{X}_{i,j}|^{4r}) < b$, for any $i \in [n], j \in [p]$, and if $n^{1-a/r} \geq ds^2 p^{2/r}$, then $\kappa(s,c,q) > \kappa^*(s,c,q) - \iota > 0$ holds with probability at least $1 - n^{-a}$.

Note that the ℓ_q-sensitivity property we use is slightly different from the one introduced in [Gautier and Tsybakov, 2013] in which the Gram matrix $\mathbf{X}^\top \mathbf{X}/n$ is normalized to have unit diagonal entries. Our ℓ_q-sensitivity property differs also a bit from the one defined in [Dobriban and Fan, 2016], mainly because the sparsity level s does not appear in our formulation. These differences are not decisive in view of the sufficient conditions established in Proposition 0.4.2. In addition, note that this theorem was initially proven in the more general setting of instrumental variables regression.

We refer the reader to the paper [Dobriban and Fan, 2016] for a complete proof of this theorem. In the sub-Gaussian setting, this proof is based on a Bernstein-type inequality due to Vershynin [2012b, Corollary 5.17] which is used to show that $\|\mathbf{\Sigma}^* - \widehat{\mathbf{\Sigma}}\|_{\infty,\infty}$ is bounded by $O\big((\log(2p^2)/n)^{1/2}\big)$ with high probability. This result is of the same nature as a known bound for the estimation error of $\widehat{\mathbf{\Sigma}}$ measured using the spectral matrix norm [Vershynin, 2012a, Proposition 2.1]. In order to highlight the importance of the control of the error of estimation of the covariance matrix, we give an argument in the case of $q = 1$.

Lemma 0.5.3. *If $\kappa^*(s,c,1)$ is the ℓ_1-sensitivity of $\mathbf{\Sigma}^*$ and $\kappa(s,c,1)$ is that of $\widehat{\mathbf{\Sigma}}$, then*

$$\kappa(s,c,1) \geq \kappa^*(s,c,1) - (c+1)\|\mathbf{\Sigma}^* - \widehat{\mathbf{\Sigma}}\|_{\infty,\infty}.$$

Last, but not least, Dobriban and Fan [2016] have pointed out that checking that the sensitivity property holds is NP-hard, as is checking the restricted isometry property [Tillmann and Pfetsch, 2014], despite the former is much less severe.

0.6 Contributions on high dimensional precision matrix estimation

0.6.1 Background

In the Gaussian setting, by the theorem on normal correlations, a precision matrix representation arises from a linear regression model. Reciprocally, a hidden regression model emerges from a given precision matrix.
Let us consider the $n \times p$ random matrix $\mathbf{X}$ whose rows are independently drawn from a Gaussian distribution with mean $\boldsymbol{\mu}^*$ and covariance $\boldsymbol{\Sigma}^*$. Assuming that this distribution is nondegenerate, the diagonal entries of the inverse $\boldsymbol{\Omega}^*$ of the covariance matrix are bounded away from zero. We denote $\mathrm{diag}(\{\omega^*_{jj}\}_{j\in[p]})$ by $\mathbf{D}^*$ and by $\mathbf{B}^*$ the matrix $\left(\omega^*_{ij}/\omega^*_{jj}\right)_{i\in[p],j\in[p]}$. It thus holds that $\boldsymbol{\Omega}^* = \mathbf{B}^*\mathbf{D}^*$.

As we assume that $\mathbf{X}_{i,\bullet} \sim \mathcal{N}_p(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, it implies that $\left(\mathbf{X}_{i,\bullet} - (\boldsymbol{\mu}^*)^\top\right)\boldsymbol{\Omega}^*_{\bullet,j} \sim \mathcal{N}(0, \omega^*_{jj})$. Setting $\boldsymbol{\phi}^*_j = (\omega^*_{jj})^{-1/2}$, it follows that $\left(\mathbf{X}_{i,\bullet} - (\boldsymbol{\mu}^*)^\top\right)\mathbf{B}^*_{\bullet,j} \sim \mathcal{N}\left(0, (\boldsymbol{\phi}^*_j)^2\right)$. Then, as the observations are independent, it entails that there exists $\boldsymbol{\epsilon}_{\bullet,j} \sim \mathcal{N}_n(0, \mathbf{I}_n)$ such that

$$\left(\mathbf{X} - \mathbf{1}_n(\boldsymbol{\mu}^*)^\top\right)\mathbf{B}^*_{\bullet,j} = \boldsymbol{\phi}^*_j\boldsymbol{\epsilon}_{\bullet,j}.$$

If we denote $(\mathbf{B}^*)^\top\boldsymbol{\mu}^*$ by $\mathbf{c}^*$, as by definition $\mathbf{B}^*$ has unit diagonal elements, we end with the linear regression model

$$\mathbf{X}_{\bullet,j} = \mathbf{1}_n\mathbf{c}^*_j - \mathbf{X}_{\bullet,j^c}\mathbf{B}^*_{j^c,j} + \boldsymbol{\phi}^*_j\boldsymbol{\epsilon}_{\bullet,j}.$$

As in [Sun and Zhang, 2013], we estimate the precision matrix by first solving the optimization problem

$$\widehat{\mathbf{B}} = \underset{\substack{\mathbf{B}\in\mathbb{R}^{p\times p}\\ \mathbf{B}_{jj}=1}}{\arg\min}\ \min_{\mathbf{c}\in\mathbb{R}^p}\left\{\|(\mathbf{X}\mathbf{B} - \mathbf{1}_n\mathbf{c}^\top)^\top\|_{2,1} + \lambda\|\mathbf{B}\|_{1,1}\right\}, \tag{0.20}$$

for a given tuning parameter $\lambda \geq 0$, and then estimating the variances of the errors by the residual variances

$$\widehat{\omega}_{jj} = \left(\frac{1}{n}\|(\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top)\mathbf{X}\widehat{\mathbf{B}}_{\bullet,j}\|_2^2\right)^{-1}; \qquad \widehat{\boldsymbol{\Omega}} = \widehat{\mathbf{B}}\cdot\mathrm{diag}(\{\widehat{\omega}_{jj}\}_{j\in[p]}). \tag{0.21}$$

We remark that Problem (0.20) is equivalent to solving p independent problems of square-

root Lasso. To simplify the formulation of the estimator, we assume that $\boldsymbol{\mu}^* = 0$ (thus $\mathbf{c}^* = 0$). The residuals are thus $\widehat{\boldsymbol{\epsilon}}_{\bullet,j} = \mathbf{X}\widehat{\mathbf{B}}_{\bullet,j}$. The proposed estimator of the precision matrix is then given by

$$\widehat{\mathbf{B}}_{j^c,j} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p-1}}{\arg\min}\Big\{\|\mathbf{X}_{\bullet,j} + \mathbf{X}_{\bullet,j^c}\boldsymbol{\beta}\|_2 + \lambda\|\boldsymbol{\beta}\|_1\Big\}; \qquad \widehat{\mathbf{B}}_{j,j} = 1, \tag{0.22}$$
$$\widehat{\omega}_{jj} = n\|\widehat{\boldsymbol{\epsilon}}_{\bullet,j}\|_2^{-2}; \qquad \widehat{\boldsymbol{\Omega}} = \widehat{\mathbf{B}}\cdot\mathrm{diag}(\{\widehat{\omega}_{jj}\}_{j\in[p]}).$$

This estimator is accurate, it is not only asymptotically consistent, but has also an optimal convergence rate for a certain class of true precision matrices [Cai et al., 2012, Theorem 5]. Next propositions provide finite sample risk bounds on precision matrix estimation. We show for instance that the estimator given by (0.22) reaches a rate of convergence of order $s\sqrt{\log(p)/n}$ in ℓ_1-matrix norm with high probability. In Frobenius norm, the convergence rate is of order $\sqrt{sp\log(p)/n}$, hence comparable with the rate established in [Rothman et al., 2008] for the graphical Lasso estimator, but obtained under slightly different assumptions.

Proposition 0.6.1. *We assume that the maximal number of nonzero entries in a column of $\boldsymbol{\Omega}^*$ is $s \in [p]$. Let $\mathbf{X}$ be a $n \times p$ random matrix with i.i.d. centered Gaussian rows whose covariance matrix has unit diagonal entries and satisfies the ℓ_1-sensitivity property $\kappa^*(s,2,1) > 0$. Let us consider $\alpha \in (0,1)$, $\delta \in (0,1)$ and choose*

$$\lambda = 6\Big(\log\frac{8p^2}{\delta}\Big)^{1/2}.$$

Let d be a positive constant. We assume that the sample size n satisfies

$$n \geq \Big(16\log(8p/\delta)\Big) \vee \Big(4\lambda^2/\kappa^*(s,2,1)\Big) \vee \Big(ds^2\log(1/\alpha)\Big).$$

We set $A = 128/\kappa^*(s,2,1)$.
Then, the solution $\widehat{\boldsymbol{\Omega}}$ of problem (0.22) satisfies the following inequalities

$$\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\|_{\infty,1} \leq \frac{1}{\sqrt{n}}\lambda\big(\max_j \omega^*_{jj}\big)^{1/2}\Big(A + \frac{2}{3}s\big(\max_j \omega^*_{jj}\big)^{1/2}\Big), \tag{0.23}$$
$$\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\|_{1,1} \leq \frac{p}{\sqrt{n}}\lambda\big(\max_j \omega^*_{jj}\big)\Big(A\big(\min_j \omega^*_{jj}\big)^{-1/2} + \frac{2}{3}s\Big), \tag{0.24}$$

with probability at least $1 - \delta - \alpha$.

In the next proposition, we establish a convergence rate for the estimator of the precision matrix in Frobenius norm under slightly stronger assumptions.

Proposition 0.6.2. *We assume that the maximal number of nonzero entries in a column of $\boldsymbol{\Omega}^*$ is $s \in [p]$. Let $\mathbf{X}$ be a $n \times p$ random matrix with i.i.d. centered Gaussian rows whose*

covariance matrix has unit diagonal entries and satisfies the restricted eigenvalue property $\bar{\kappa}^{*RE}(s,2) > 0$. *Let us consider* $\alpha \in (0,1)$, $\delta \in (0,1)$ *and choose*

$$\lambda = 6\Big(\log\frac{8p^2}{\delta}\Big)^{1/2}.$$

For universal constants $a, b, d > 0$, *we assume that the sample size* n *satisfies*

$$n \geq \Big(16\log(8p/\delta)\Big) \vee \Big(as\lambda^2/\bar{\kappa}^{*RE}(s,2)\Big) \vee \Big(1/d\log(b/\alpha)\Big).$$

We set $C = 512\Big(1+2\sqrt{s/n}\Big)/\bar{\kappa}^{*RE}(s,2)$.
Then, the solution $\widehat{\mathbf{\Omega}}$ *of problem* (0.22) *satisfies the following inequality*

$$\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{2,2} \leq \frac{\sqrt{p}}{\sqrt{n}}\lambda\sigma_{\max}(\mathbf{\Omega}^*)\Big(\sqrt{s}C\big(\min_j \omega^*_{jj}\big)^{-1/2} + \frac{2}{3}\Big), \tag{0.25}$$

with probability at least $1-\delta-\alpha$.

0.6.2 Estimation of diagonal elements

Whether in the method of Yuan [2010] or in that of Sun and Zhang [2013], the estimation of the precision matrix rests on the presumptive relation $\widehat{\mathbf{\Omega}} = \widehat{\mathbf{B}} \cdot \mathrm{diag}(\{\widehat{\omega}_{jj}\}_{j\in[p]})$. The diagonal entries, that is the conditional variances which correspond to the inverses of the variances of errors in the linear model, and the coefficients of regression are estimated in two steps. Finally, the previous relation is used to compute the off-diagonal elements. In [Yuan, 2010] as in [Sun and Zhang, 2013], the diagonal entries of the precision matrix are naturally estimated using the variance of the regression residuals starting from the estimation of the coefficients of regression. Thereby, the error of estimation made on $\mathbf{B}^*$ has an impact on the estimates of the diagonal entries. However, the appositeness of residual variance does not seem to have ever been questioned. In Chapter 1, we compare several different estimators of the diagonal elements on the basis of their quadratic risks in the oracle case when $\mathbf{B}^*$ is known. The considered estimators are residual variance (RV), squared average absolute deviation (AD), relaxed maximum likelihood (RML), symmetry-enforced maximum likelihood (SML) and penalized maximum likelihood (PML). We show that the usual natural choice of residual variance is appropriate in practice, but that the maximum likelihood based estimators are better as long as the error of estimation on $\mathbf{B}^*$ remains small. We propose algorithms for estimation and provide an empirical evaluation of the performance of the various options.

0.6.3 Robust estimation

The question of the estimation of the precision matrix in presence of outliers has not been considered until quite recently. We extend the existing work on sparse precision matrix estimation to a situation where the Gaussian distribution of interest is not directly observed, but in which the available data is corrupted by additive outliers. We propose to take the possible presence of outliers into account by adding a convex penalization in the cost function. We consider both the moderate dimensional case, where the cost function is composed of a data fidelity term plus the regularization term corresponding to outliers, and the high dimensional case, where we further add a regularization term to promote sparsity. We establish optimal finite sample bounds for the error of estimation—measured using ℓ_1/ℓ_1, Frobenius and mixed ℓ_2/ℓ_1 norms—of the matrix representing outliers when the dimension is of smaller order than the sample size. In this case, we also provide a convergence rate in Frobenius norm for the estimator of the precision matrix. In the high dimensional case, we show that for an appropriate choice of the tunning parameter, when the sample size is large enough, when a matrix compatibility condition is satisfied by the data matrix and when the diagonal entries of the precision matrix are lower bounded, then both estimators of the matrix of the coefficients of regression and of the matrix corresponding to outliers converge at a fast rate. Indeed, if we assume that $\mathbf{\Omega}^*$ is s-sparse and that there are at most t outlying observations, then $\widehat{\mathbf{B}}$ converges towards $\mathbf{B}^*$ with a rate of order $(sp+t)\sqrt{\log(np)/(n-t)}$ in ℓ_1/ℓ_1 norm. This rate is optimal – up to logarithmic factors – if t is small compared to n. It therefore leads to an optimal convergence rate (of same order) for the estimator of the precision matrix. Note that in absence of outliers, Proposition 0.6.1 states a convergence rate of order $sp\sqrt{\log(p)/n}$. This rate is of the same order as the former when $t = 0$. As mentioned above, our results in high dimension rest on a particular matrix compatibility condition. It remains to assert that this condition is met with high probability for a wide class of data matrices $\mathbf{X}$.

0.7 Manuscript organization

This manuscript consists of two chapters whose content coincides with the contributions that we have described in the previous section. In this way, Chapter 1 concerns the estimation of the diagonal entries and Chapter 2 the estimation in presence of outliers of the precision matrix. These results have been pre-published on http://arxiv.org/. The proofs of the claimed results are included in each chapter. Some theoretical, resp. experimental, results that complete Chapter 1 are given in Appendix A, resp. Appendix B. As a last point, an overview of the implementation of the estimators that we have introduced and analyzed is provided in Appendix C. The code has been made publicly available as an R package [R Core Team, 2016].

Chapter 1

Estimation of the diagonal elements of a sparse precision matrix

The main part of this chapter is taken from the article *On estimation of the diagonal elements of a sparse precision matrix* [Balmand and Dalalyan, 2016], which has been published in the Electronic Journal of Statistics (EJS) on May 2016.

RÉSUMÉ. Dans ce chapitre, nous présentons différents estimateurs des éléments diagonaux de l'inverse de la matrice de covariance, également appelée matrice de précision, d'un échantillon de réalisations de vecteurs aléatoires indépendants et identiquement distribués. Nous nous intéressons principalement au cas de vecteurs de grande dimension, dont la matrice de précision est creuse. Il est désormais clair que lorsque que la distribution sous-jacente est gaussienne, chacune des colonnes de la matrice de précision peut être estimée indépendamment des autres, par la résolution d'un problème de régression linéaire sous contraintes de parcimonie. Cette approche conduit à une stratégie d'estimation de la matrice de précision efficace sur le plan calculatoire. Dans un premier temps, les vecteurs des coefficients de régression sont estimés, ils sont ensuite utilisés pour estimer les termes diagonaux de la matrice de précision et dans un dernier temps, les estimateurs des étapes précédentes sont combinés pour obtenir ceux des éléments non diagonaux. Alors que l'étape consistant à estimer les vecteurs des coefficients de régression a été l'objet de nombreux travaux ces dix dernières années, la manière d'en tirer des estimateurs statistiquement précis des termes diagonaux a suscité moins

d'intérêt. Ce chapitre a pour objectif de combler ces lacunes en présentant quatre estimateurs – qui nous semblent les plus naturels – des éléments diagonaux de la matrice de précision et de les évaluer dans le détail du point de vue empirique. Les estimateurs que nous considérons sont la variance résiduelle, l'estimateur du maximum de vraisemblance (EMV) en relâchant les contraintes de symétrie sur la matrice de précision, l'EMV en imposant les contraintes de symétrie, ainsi que l'EMV pénalisé. Nous montrons, à la fois théoriquement et empiriquement, que l'EMV en imposant les contraintes de symétrie a la plus faible erreur d'estimation, lorsque les vecteurs des coefficients de régression mentionnés plus haut sont estimés sans erreur. Néanmoins, dans des conditions plus réalistes, lorsque les vecteurs des coefficients de régression sont estimés grâce à une méthode de calcul efficace, favorisant l'émergence d'une solution parcimonieuse, les performances des estimateurs considérés deviennent relativement proches avec toutefois une légère supériorité de l'estimateur de la variance résiduelle.

Contents

1.1 Introduction

We consider the problem of precision matrix estimation that has been extensively studied in recent years partly because of its tight relation with the graphical models. More precisely, assuming that we observe p features on n individuals, an interesting object to display is the graph of associations between the features, especially when the number of features is large. The associations may be of different type: linear correlations, partial correlations, measures of independence and so on. A measure of association between the features, which is particularly relevant for Gaussian [Lauritzen, 1996] and, more generally, nonparanormal distributions [Liu et al., 2009; Lafferty et al., 2012] is the partial correlation. This leads to a Gaussian graphical model in which two nodes are connected by an edge if the partial correlation between the features corresponding to these two nodes is nonzero, which is equivalent to the nonzeroness of the corresponding entry of the precision matrix [Lauritzen, 1996, Proposition 5.2]. The graph constructed in such a way relies on the population precision matrix, which is not available in practice. Therefore, an important statistical problem is to infer this graph from n i.i.d. observations of the p-dimensional feature-vector. In view of the aforementioned connection with the precision matrix, the estimated graph may be deduced from the estimated precision matrix by comparing its entries with a suitably chosen threshold.

Another important problem for which the precision matrix estimation is relevant[1] is the linear [Fisher, 1936] or quadratic discriminant analysis [Anderson, 2003]. Indeed, the de-

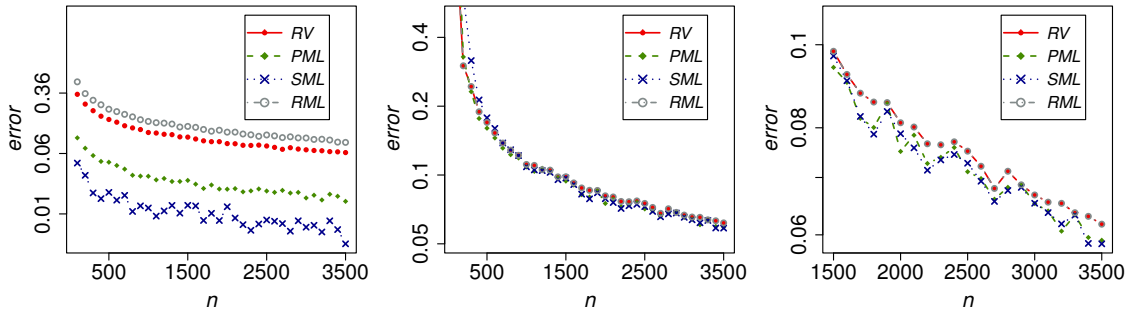
[1] In the case of linear discriminant analysis for binary classification, a simpler approach consisting in replacing the sparsity of the precision matrix by the sparsity of the product of the latter with the difference of the class means has been proposed and studied by Cai and Liu [2011].

cision boundary in the binary or multi-class classification problem—under the assumption that the conditional distributions of the features given the class are Gaussian—is defined in terms of the precision matrix. In order to infer this decision boundary from data, it is therefore relevant to start with estimating the precision matrix. The simplest way of estimating the latter is by inverting the sample covariance matrix or, if the inverse does not exist, by computing the pseudo-inverse of the sample covariance matrix. However, when the dimension p is such that the number of unknown parameters $p(p+1)$ is comparable to or larger than the sample-size n, the (pseudo-)inversion of the sample covariance matrix leads to very poor results. To circumvent this shortcoming, additional assumptions on the precision matrix should be imposed which should preferably be realistic, interpretable and lead to statistically and computationally efficient estimation procedures. The sparsity of the precision matrix offers a convenient setting in which these criteria are met.

To present in a more concrete fashion the content of the present work, let $\mathbf{X}$ be a $n \times p$ random matrix representing the values of p variables observed on n individuals. Assume that the rows of the matrix $\mathbf{X}$ are independent and Gaussian with mean $\boldsymbol{\mu}^*$ and covariance matrix $\boldsymbol{\Sigma}^*$. The inverse of $\boldsymbol{\Sigma}^*$, called the precision matrix and denoted by $\boldsymbol{\Omega}^* = (\omega_{ij}^*)$, is an object of central interest since—as mentioned earlier—it encodes the conditional dependencies between pairs of variables given the values of all the other variables. Based on the precision matrix, the graph $\mathscr{G}^*$ of relationships between the p variables is constructed as follows: each node of the graph represents a variable and two nodes i and j are connected by an edge if and only if $\omega_{ij}^* \neq 0$. Estimating this graph from a sample of size n represented by the rows of $\mathbf{X}$ is a challenging statistical problem that has attracted a lot of attention in the past decade. In a frequently encountered situation of the dimension p comparable to or even larger than n, a commonly used assumption is the sparsity of the graph $\mathscr{G}^*$. Namely, it is assumed that the maximal degree of the nodes of $\mathscr{G}^*$ is much smaller than p (see, for instance, [Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007] for early references).

Most approaches of estimating sparse precision matrices that gained popularity in recent years rely on weighted ℓ_1-penalization of the off-diagonal elements of the precision matrix; recent contributions on the statistical aspects of this approach can be found in [Yuan, 2010; Cai et al., 2011; Cai et al., 2012; Sun and Zhang, 2013] and the references therein. The rationale behind this approach is that the weighted ℓ_1-penalty can be viewed as a convexified version of the ℓ_0-penalty, the latter being understood as the number of nonzero elements. The convexity of the penalty in conjunction with the convexity of the data fidelity term leads to estimators that can be efficiently computed by convex programming [Banerjee et al., 2008; Friedman et al., 2008].

To further improve the computational complexity, it is possible to split the problem of estimating p^2 entries of the precision matrix into p independent problems of estimating the p-dimensional columns of it [Meinshausen and Bühlmann, 2006]. To this end, the matrix $\boldsymbol{\Omega}^*$ is written as $\mathbf{B}^*\mathbf{D}^*$, where $\mathbf{D}^*$ is a diagonal matrix while $\mathbf{B}^*$ is a $p \times p$ matrix with



$\widehat{\mathbf{B}}$ coincides with $\mathbf{B}^*$ $\widehat{\mathbf{B}}$ estimated by SqRL+OLS: zoom

$\widehat{\mathbf{B}}$ estimated by SqRL+OLS

Figure 1.1: The average ℓ_2-error (computed from 50 independent trials) of the four estimators considered in this work as a function of the sample size. The plots concern Model 2 described in Section 1.4.1 and dimension $p = 60$. One can observe, in particular, that when $\mathbf{B}^*$ is estimated without error (left panel), the estimators SML and PML improve on the residual variance and relaxed maximum likelihood estimators.

all diagonal entries equal to one. Each columns of the matrix $\mathbf{B}^*$ can be estimated by regressing one column of the data matrix $\mathbf{X}$ on all the remaining columns. In the context of high dimensionality and sparse precision matrix, this can be performed by sparsity favoring methods [Bühlmann and van de Geer, 2011] such as the Lasso [Tibshirani, 1996], the Dantzig selector [Candes and Tao, 2007], the square-root Lasso [Belloni et al., 2011], etc. A crucial observation at this stage is that the sparsity patterns, that is, the locations of nonzero entries, of the matrices $\mathbf{B}^*$ and $\mathbf{\Omega}^*$ coincide. In particular, the degree of the j-th node in the graph $\mathscr{G}^*$ is equal to the number of nonzero entries of the j-th column of $\mathbf{B}^*$, for every $j = 1, \ldots, p$.

Once the columns of $\mathbf{B}^*$ successfully estimated, one needs to estimate the diagonal matrix $\mathbf{D}^*$, the diagonal entries of which coincide with those of the precision matrix $\mathbf{\Omega}^*$. This step is necessary for recovering the precision matrix (both diagonal and off-diagonal entries) but it is also important for constructing the graph[2] of conditional dependencies. Of course, the latter can be estimated by thresholding the entries of the estimator of $\mathbf{B}^*$ without resorting to an estimator of $\mathbf{D}^*$, but the choice of the threshold is in this case a difficult issue deprived of clear statistical interpretation. In contrast with this, if along with an estimator of $\mathbf{B}^*$, an estimator of $\mathbf{D}^*$ is available, then one may straightforwardly estimate the partial correlations and threshold them to infer the graph of conditional dependencies. In this case, the threshold has a more clear statistical meaning since the partial correlations are in absolute value bounded by one.

[2]We put an emphasize on this last point since we did not find it in the literature.

It follows from the above discussion that the problem of estimating the matrix $\mathbf{D}^*$ built from the diagonal entries of the precision matrix is an important ingredient of the estimation of the precision matrix and the graph of conditional dependencies between the features. The purpose of the present work is to propose several natural estimators of $\mathbf{D}^*$ and to study their statistical properties, essentially from an empirical point of view. Combining standard arguments, we present four estimators, termed residual variance (RV), relaxed maximum likelihood (RML), symmetry-enforced maximum likelihood (SML) and penalized maximum likelihood (PML). The first one, residual variance, is the most commonly used estimator when the matrix $\mathbf{B}^*$ is estimated column-wise by a sparse linear regression approach briefly mentioned in the foregoing discussion. The other three methods considered in this chapter are based on the principle of likelihood maximization under various approaches for handling the prior information. In order to give the reader a foretaste of the content of next sections, we present in Figure 1.1 the accuracy of the four methods of estimating the diagonal elements of the precision matrix on a synthetic data-set. More details are given in Section 1.4.1.

1.2 Preliminaries on precision matrix estimation

This section recalls some preliminary material on sparse precision matrix estimation. In this chapter, and only in it, unlike in the rest of the manuscript, we consider $\phi_j^* = 1/\omega_{jj}^*$ (in place of $\phi_j^{*2} = 1/\omega_{jj}^*$), for any $j \in [p]$, to simplify notations.

Throughout the chapter we will present estimators of the diagonal elements of the precision matrix in the case of a general multidimensional Gaussian distribution, but in all theoretical developments we will assume that the marginals of $\mathbf{X}$ are standard Gaussian distributions, that is, $\boldsymbol{\mu}^* = 0$ and $\boldsymbol{\Sigma}_{jj}^* = 1$ for every $j \in [p]$. This assumption is reasonable, since we are concerned with the problems in which the sample size is large enough to consistently estimate the individual means and the individual variances of the variables. So, one can always center the variables by the sample mean and divide by the sample standard deviation to get close to the assumption[3] that random variables $\mathbf{X}_{1,j}, \ldots, \mathbf{X}_{n,j}$ are i.i.d. $\mathcal{N}(0,1)$ for every j.

Let us recall that the precision matrix is closely related to the problem of regression of one feature on all the others. Indeed, there exists a $p \times p$ matrix $\mathbf{B}^*$ and two vectors $\mathbf{c}^*, \boldsymbol{\phi}^* \in \mathbb{R}^p$ such that

$$\mathbf{X}_{\bullet,j} = c_j^* \mathbf{1}_n - \mathbf{X}_{\bullet,j^c} \mathbf{B}_{j^c,j}^* + \phi_j^{*1/2} \boldsymbol{\xi}_j, \tag{1.1}$$

where $\boldsymbol{\xi}_j$ is drawn from $\mathcal{N}_n(0, \mathbf{I}_n)$ and is independent of $\mathbf{X}_{\bullet,j^c}$. According to the theorem on normal correlations [Marsaglia, 1964], the regression coefficients $\mathbf{B}_{j^c,j}^* \in \mathbb{R}^{p-1}$ and the variance $\phi_j^* \in \mathbb{R}$ of residuals can be expressed in terms of the elements of the precision

[3] Unless expressly stated otherwise, in the whole chapter, $1 \leq i, j \leq p$ and $1 \leq k \leq n$.

matrix $\mathbf{\Omega}^*$ as follows:

$$\mathbf{B}^*_{ij} = \omega^*_{ij}/\omega^*_{jj}, \qquad \phi^*_j = 1/\omega^*_{jj}, \tag{1.2}$$

whereas $c^*_j = \mu^*_j + (\boldsymbol{\mu}^*_{j^c})^\top \mathbf{B}^*_{j^c,j} = (\boldsymbol{\mu}^*)^\top \mathbf{B}^*_{\bullet,j}$. If we assume that $\boldsymbol{\mu}^* = 0$ then $c^*_j = 0$ for any j. With these notation, the precision matrix can be written as $\mathbf{\Omega}^* = \mathbf{B}^* \mathbf{D}^{-1}_{\boldsymbol{\phi}^*}$.

Several state-of-the-art methods for estimating sparse precision matrices proceed in two steps [Meinshausen and Bühlmann, 2006; Cai et al., 2011; Liu and Wang, 2012; Sun and Zhang, 2013]. The first step consists in estimating the matrix $\mathbf{B}^*$ and the vector $\boldsymbol{\phi}^*$ by solving the sparse linear regression problems (1.1) for each j, while in the second step an estimator of the matrix $\mathbf{\Omega}^*$ is inferred from the estimators of $\mathbf{B}^*$ and $\boldsymbol{\phi}^*$ using relations (1.2). The goal of the present work is to explore both theoretically and empirically different possible strategies for this second step.

The square-root Lasso is perhaps the method of estimating the matrix $\mathbf{B}^*$ that offers the best trade-off between the computational and the statistical complexities. It can be redefined as follows: the square-root Lasso estimates the matrix $\mathbf{B}^*$ by solving the convex optimization problem

$$\widehat{\mathbf{B}} = \operatorname*{arg\,min}_{\substack{\mathbf{B}\in\mathbb{R}^{p\times p} \\ \mathbf{B}_{jj}=1}} \min_{\mathbf{c}\in\mathbb{R}^p} \Big\{ \|\mathbf{X}\mathbf{B} - \mathbf{1}_n \mathbf{c}^\top\|_{2,1} + \lambda \|\mathbf{B}\|_{1,1} \Big\}, \tag{1.3}$$

where the first min is over all matrices $\mathbf{B}$ having all their diagonal entries equal to 1. The tuning parameter $\lambda > 0$ corresponds to the penalty level. The purpose of the penalization is indeed to get a precision matrix estimate which fits the sparsity assumption. As the penalty of a matrix $\mathbf{B}$ is its $\|\cdot\|_{1,1}$ norm, the resulting precision matrix estimator is expected to be sparse in the sense that its overall number of nonzero elements should be small. In addition, one can check that computing a solution to problem (1.3) is equivalent to computing each column of $\widehat{\mathbf{B}}$ separately (and independently) by solving the optimization problem

$$\widehat{\mathbf{B}}_{\bullet,j} = \operatorname*{arg\,min}_{\substack{\boldsymbol{\beta}\in\mathbb{R}^p \\ \boldsymbol{\beta}_j=1}} \min_{c_j\in\mathbb{R}} \Big\{ \|\mathbf{X}\boldsymbol{\beta}_j - c_j \mathbf{1}_n\|_2 + \lambda \|\boldsymbol{\beta}\|_1 \Big\}, \quad j \in [p]. \tag{1.4}$$

In addition to being efficiently computable even for large p, this estimator has the following appealing property that makes it preferable, for instance, to the column-wise Lasso [Meinshausen and Bühlmann, 2006] and the Clime [Cai et al., 2011]. The choice of the parameter λ in (1.3-1.4) is scale free: it can be chosen independently of the noise variance in linear regression (1.1). This fact has been first established by Belloni et al. [2011] and then further investigated in [Sun and Zhang, 2012; Belloni et al., 2014a]. In the context of precision matrix estimation, this method has been explored[4] by Sun and Zhang [2013].

[4] Although Sun and Zhang [2012, 2013] refer to this method as the scaled Lasso, we prefer to use the original term square-root Lasso coined by Belloni et al. [2011] in order to avoid any possible confusion with the earlier method of Städler et al. [2010a,b], for which the term "scaled Lasso" has been already employed.

1.3 Four estimators of the variance of noise in sparse regression

As mentioned earlier, the aim of this work is to compare different estimators of the vector $\boldsymbol{\phi}^*$ based on an initial estimator of the matrix $\mathbf{B}^*$. Clearly, the error of the estimation of $\mathbf{B}^*$ impacts the error of the estimation of $\boldsymbol{\phi}^*$ and, therefore, the latter is not easy to assess in full generality. In order to gain some insight on the behavior of various natural estimators, in theoretical results we will consider the ideal situation where the matrix $\mathbf{B}^*$ is estimated without error.

A brief note on the choice of the estimators that we call natural: the first rests on the fact that ϕ_j^* is the variance of the error in the regression model (1.1), thus is usually estimated by the residual variance. The other three estimators are obtained by maximizing the likelihood of $\mathbf{X}$ whose rows are assumed to be independent and Gaussian, under strong or relaxed constraints of symmetry.

1.3.1 Residual variance estimator

In view of the regression equation presented in (1.1), a standard and natural method[5]—used, in particular, by the square-root Lasso[6] of Sun and Zhang [2013]—to deduce estimators $\widehat{\boldsymbol{\phi}}$ and $\widehat{\boldsymbol{\Omega}}$ from an estimator $\widehat{\mathbf{B}}$ is to set

$$\widehat{\phi}_j = \frac{1}{n}\|(\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top)\mathbf{X}\widehat{\mathbf{B}}_{\bullet,j}\|_2^2; \qquad \widehat{\boldsymbol{\Omega}} = \widehat{\mathbf{B}} \cdot \mathbf{D}_{\widehat{\boldsymbol{\phi}}}^{-1}. \tag{1.5}$$

Note that the matrix $(\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top)$ present in this expression is the orthogonal projector in $\mathbb{R}^n$ onto the orthogonal complement of the linear subspace $\mathrm{Span}(\mathbf{1}_n)$ of all constant vectors. The multiplication by this matrix annihilates the intercept c_j^* in (1.1) and is a standard way of reducing the affine regression to the linear regression. In what follows, we refer to $\widehat{\boldsymbol{\phi}}$ defined by (1.5) as the residual variance estimator and denote it by $\widehat{\boldsymbol{\phi}}^{\mathrm{RV}}$. Using the sample covariance matrix $\mathbf{S}_n$, the residual variance estimator of $\boldsymbol{\phi}^*$ can be written as

$$\widehat{\phi}_j^{\mathrm{RV}} = \widehat{\mathbf{B}}_{\bullet,j}^\top \mathbf{S}_n \widehat{\mathbf{B}}_{\bullet,j}.$$

Note also that if we consider the linear regression model (1.1) conditionally to $\mathbf{X}_{\bullet,j^c}$, then the residual variance estimator of ϕ_j^* coincides with the maximum likelihood estimator.

Proposition 1.3.1. *If $\widehat{\mathbf{B}}_{\bullet,j}$ estimates $\mathbf{B}^*_{\bullet,j}$ without error, then the residual variance estimator of ϕ_j^* has a quadratic risk equal to $\frac{2}{n}\phi_j^{*2}$, that is*

$$\mathbf{E}[(\widehat{\phi}_j^{\mathrm{RV}} - \phi_j^*)^2] = \frac{2\phi_j^{*2}}{n}.$$

[5] This kind of estimators have recently been the subject of a simulation study by Reid et al. [2016] in the context of Lasso regression.

[6] See footnote 4.

Furthermore, for every $t > 0$, the following bound on the tails of the maximal error holds true:

$$\mathbf{P}\Big(\max_{j\in[p]} \frac{|\widehat{\phi}_j^{\mathrm{RV}} - \phi_j^*|}{\phi_j^*} > 2\Big(\frac{t+\log p}{n}\Big)^{1/2} + 2\frac{t+\log p}{n}\Big) \leq 2\,\mathrm{e}^{-t}\,.$$

Proof. Using equation (1.1) and the assumption $\widehat{\mathbf{B}}_{\bullet,j} = \mathbf{B}^*_{\bullet,j}$, we get

$$\widehat{\phi}_j^{\mathrm{RV}} = \frac{1}{n}\|\mathbf{X}\mathbf{B}^*_{\bullet,j}\|_2^2 = \frac{\phi_j^*}{n}\|\boldsymbol{\xi}_j\|_2^2. \tag{1.6}$$

Since $\boldsymbol{\xi}_j$ is a standard Gaussian vector, the random variable $\zeta = \|\boldsymbol{\xi}_j\|_2^2$ is drawn from a χ^2_n distribution. This implies that $\mathbf{E}(\zeta) = n$ and $\mathbf{V}ar(\zeta) = 2n$. Therefore,

$$\mathbf{E}[(\widehat{\phi}_j^{\mathrm{RV}} - \phi_j^*)^2] = \mathbf{E}\Big[\Big(\frac{\phi_j^*\zeta}{n} - \phi_j^*\Big)^2\Big] = \frac{\phi_j^{*2}}{n^2}\Big(\mathbf{V}ar(\zeta) + \big(\mathbf{E}(\zeta) - n\big)^2\Big) = \frac{2\phi_j^{*2}}{n}.$$

This completes the proof of the first claim. To prove the second claim, we set $z = t + \log p$ and use the union bound to get

$$\begin{aligned}\mathbf{P}\Big(\max_{j\in[p]} \frac{|\widehat{\phi}_j^{\mathrm{RV}} - \phi_j^*|}{\phi_j^*} > 2\Big(\frac{z}{n}\Big)^{1/2} + 2\frac{z}{n}\Big) &\leq p \max_{j\in[p]} \mathbf{P}\Big(\frac{|\widehat{\phi}_j^{\mathrm{RV}} - \phi_j^*|}{\phi_j^*} > 2\Big(\frac{z}{n}\Big)^{1/2} + 2\frac{z}{n}\Big) \\ &= p\mathbf{P}\Big(|\zeta - n| > 2\sqrt{zn} + 2z\Big).\end{aligned}$$

The second claim follows from the tail bound of the χ^2 distribution established, for instance, in [Laurent and Massart, 2000, Lemma 1]. □

Note that in this result, the case of known means $\mu_j^* = 0$ is considered. The case of unknown μ_j can be handled similarly, the estimation bias is then ϕ_j^*/n and the resulting mean squared error is $(2n-1)\phi_j^{*2}/n^2$. One may observe that, as expected, the rate of convergence of the quadratic risk is the usual parametric rate $1/n$ and that the asymptotic variance is $2\phi_j^{*2}$.

As a complement, we also propose to estimate the variance of the error of the model using a different measure of dispersion than empirical variance. We consider the squared average absolute deviation (AD), properly normalized. It has the distinction of being less sensitive to outliers than variance. We thus set

$$\widehat{\phi}_j^{\mathrm{AD}} = \frac{\pi}{2n^2}\|(\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top)\mathbf{X}\widehat{\mathbf{B}}_{\bullet,j}\|_1^2. \tag{1.7}$$

Next proposition establishes that the AD estimator[7] is a little less accurate with respect to the quadratic risk than the RV estimator.

[7] In fact, we study four plus one estimators of $\boldsymbol{\phi}^*$. The AD estimator is comparable in nature with the RV estimator and is presented in view of Chapter 2. We observe that it performs almost as well as the RV estimator in our experimental settings.

Proposition 1.3.2. *If $\widehat{\mathbf{B}}_{\bullet,j}$ estimates $\mathbf{B}^*_{\bullet,j}$ without error, then the quadratic risk of the AD estimator of ϕ^*_j satisfies*

$$\mathbf{E}[(\widehat{\phi}^{\mathrm{AD}}_j - \phi^*_j)^2] = \frac{\phi^{*2}_j}{n}\Big(5\big(\frac{\pi}{2}-1\big)+a_n\Big); \qquad \frac{3}{4n}+\frac{2}{n^2} \le a_n \le \frac{4}{5n}+\frac{3}{n^2}.$$

Moreover, there exists an universal constant $C>0$, such that for any $t>0$ that satisfies $(t+\log p)b\pi \le n$, the following bound on the tails of the maximal error holds true:

$$\mathbf{P}\Big(\max_{j\in[p]} \frac{|\widehat{\phi}^{\mathrm{AD}}_j - \phi^*_j|}{\phi^*_j} > 2\sqrt{C}\Big(\frac{t+\log p}{n}\Big)^{1/2} + C\frac{t+\log p}{n}\Big) \le 2\,\mathrm{e}^{-t}\,.$$

Proof. By equation (1.1), as we assume that $\mu^*_j = 0$ and that $\widehat{\mathbf{B}}_{\bullet,j} = \mathbf{B}^*_{\bullet,j}$ holds, the AD estimator is given by

$$\widehat{\phi}^{\mathrm{AD}}_j = \frac{\pi}{2n^2}\phi^*_j\|\boldsymbol{\xi}_j\|_1^2.$$

It holds that

$$\|\boldsymbol{\xi}_j\|_1^2 = \Big(\sum_{i=1}^n |(\boldsymbol{\xi}_j)_i|\Big)^2 = \|\boldsymbol{\xi}_j\|_2^2 + \sum_{\substack{i,k=1\\ i\neq k}}^n |(\boldsymbol{\xi}_j)_i||(\boldsymbol{\xi}_j)_k|.$$

As $\boldsymbol{\xi}_j \sim \mathcal{N}_n(0,\mathbf{I}_n)$, then $\|\boldsymbol{\xi}_j\|_2^2$ follows a χ^2 distribution with n degrees of freedom and each of the independent entries of $|\boldsymbol{\xi}_j|$ is drawn from a half-normal distribution, thus having an expectation equal to $\sqrt{2}/\sqrt{\pi}$. We therefore arrive at

$$\mathbf{E}[\widehat{\phi}^{\mathrm{AD}}_j] = \frac{\pi}{2n^2}\phi^*_j\big(n+\frac{2}{\pi}(n^2-n)\big) = \phi^*_j\Big(1+\frac{\pi-2}{2n}\Big).$$

Note that we only have to multiply $\widehat{\phi}^{\mathrm{AD}}_j$ by $2n/(2n+\pi-2)$ to make it unbiased.

To obtain the quadratic risk, we need to develop the expression $\|\boldsymbol{\xi}_j\|_1^4$. To make a long story short, we end with[8]

$$\|\boldsymbol{\xi}_j\|_1^4 = \sum_{i=1}^n (\boldsymbol{\xi}_j)_i^4 + \sum_{\substack{i,k=1\\ i\neq k}}^n (\boldsymbol{\xi}_j)_i^2(\boldsymbol{\xi}_j)_k^2 + 6\sum_{\substack{i,k,h=1\\ i\neq k\neq h}}^n (\boldsymbol{\xi}_j)_h^2|(\boldsymbol{\xi}_j)_i||(\boldsymbol{\xi}_j)_k| + 4\sum_{\substack{i,k=1\\ i\neq k}}^n |(\boldsymbol{\xi}_j)_i|^3|(\boldsymbol{\xi}_j)_k|$$
$$+ \sum_{\substack{i,k,h,l=1\\ i\neq k\neq h\neq l}}^n |(\boldsymbol{\xi}_j)_i||(\boldsymbol{\xi}_j)_k||(\boldsymbol{\xi}_j)_h||(\boldsymbol{\xi}_j)_l|.$$

Then, using as above that the vector $\boldsymbol{\xi}_j$ is standard Gaussian, we have $\mathbf{E}[|(\boldsymbol{\xi}_j)_i|] = \sqrt{2}/\sqrt{\pi}$,

[8] $i \neq k \neq h \neq l$ means that the integers i,k,h and l are all different in each term of the summation.

$\mathbf{E}[(\boldsymbol{\xi}_j)_i^2] = 1$, $\mathbf{E}[|(\boldsymbol{\xi}_j)_i|^3] = 2\sqrt{2}/\sqrt{\pi}$ and $\mathbf{E}[(\boldsymbol{\xi}_j)_i^4] = 3$. We finally get that

$$\begin{aligned}\mathbf{E}[(\widehat{\phi}_j^{\mathrm{AD}} - \phi_j^*)^2] &= \mathbf{E}[(\widehat{\phi}_j^{\mathrm{AD}})^2 - 2\widehat{\phi}_j^{\mathrm{AD}}\phi_j^* + \phi_j^{*2}] \\ &= \frac{\phi_j^{*2}}{n}\Big(5\big(\frac{\pi}{2} - 1\big) + \frac{1}{n}\big(\frac{\pi^2}{4} - 5\pi + 14\big) + \frac{1}{n^2}\big(\frac{\pi^2}{2} + 2\pi - 9\big)\Big).\end{aligned}$$

Next, for the second assertion of the proposition, let us set $z = t + \log p$, we get

$$\begin{aligned}\mathcal{P} &= \mathbf{P}\Big(\frac{|\widehat{\phi}_j^{\mathrm{AD}} - \phi_j^*|}{\phi_j^*} > 2\sqrt{b\pi}\Big(\frac{z}{n}\Big)^{1/2} + b\pi\frac{z}{n}\Big) = \mathbf{P}\Big(\Big|\frac{\pi}{2n^2}\|\boldsymbol{\xi}_j\|_1^2 - 1\Big| > 2\Big(b\pi\frac{z}{n}\Big)^{1/2} + b\pi\frac{z}{n}\Big) \\ &= \mathbf{P}\Big(\frac{\pi}{2n^2}\|\boldsymbol{\xi}_j\|_1^2 - 1 > 2\Big(b\pi\frac{z}{n}\Big)^{1/2} + b\pi\frac{z}{n}\Big) + \mathbf{P}\Big(-\frac{\pi}{2n^2}\|\boldsymbol{\xi}_j\|_1^2 + 1 > 2\Big(b\pi\frac{z}{n}\Big)^{1/2} + b\pi\frac{z}{n}\Big).\end{aligned}$$

As $zb\pi \le n$, the second term of the right-hand side of the above equation is bounded by $\mathbf{P}\Big(\|\boldsymbol{\xi}_j\|_1/n - (2/\pi)^{1/2} < (2bz/n)^{1/2}\Big)$. The first term is equal to $\mathbf{P}\Big(\|\boldsymbol{\xi}_j\|_1/n - (2/\pi)^{1/2} > (2bz/n)^{1/2}\Big)$. Then, using the Hoeffding bounds (see for instance [Vershynin, 2012b, Proposition 5.10]) for the i.i.d. sub-Gaussian variables $|(\boldsymbol{\xi}_j)_i|$, all having the same expectation $(2/\pi)^{1/2}$ and sub-Gaussian parameter[9] b, it follows that $\mathcal{P} \le 2\,\mathrm{e}^{-z}$. We take $C = b\pi$ and as in the proof of Proposition 1.3.1, we use the union bound to conclude the proof. □

1.3.2 Relaxed maximum likelihood estimator

One could expect that the global maximum likelihood estimator of $\boldsymbol{\phi}^*$ would be better than the maximum of the conditional likelihood, since it is well known that under proper regularity conditions, the quadratic risk of the maximum likelihood estimator is the smallest, at least asymptotically. Since the vectors $\mathbf{X}_{k,\bullet} \sim \mathcal{N}_p(\boldsymbol{\mu}^*, \boldsymbol{\Omega}^{*-1})$ are independent, the log-likelihood is given by (up to irrelevant additive terms independent of the unknown parameters $\boldsymbol{\mu}^*$ and $\boldsymbol{\Omega}^*$)

$$\mathcal{L}(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Omega}) = \frac{n}{2}\log\det(\boldsymbol{\Omega}) - \frac{1}{2}\sum_{k=1}^{n}(\mathbf{X}_{k,\bullet} - \boldsymbol{\mu}^\top)\boldsymbol{\Omega}(\mathbf{X}_{k,\bullet} - \boldsymbol{\mu}^\top)^\top. \tag{1.8}$$

Maximizing the log-likelihood with respect to $\boldsymbol{\mu} \in \mathbb{R}^p$ leads to

$$\max_{\boldsymbol{\mu}\in\mathbb{R}^p} \mathcal{L}(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Omega}) = \frac{n}{2}\Big(\log\det(\boldsymbol{\Omega}) - \operatorname{trace}\big[\mathbf{S}_n\boldsymbol{\Omega}\big]\Big). \tag{1.9}$$

Recall now that in view of (1.2), we have $\boldsymbol{\Omega}^* = \mathbf{B}^*\mathbf{D}_{\boldsymbol{\phi}^*}^{-1}$. Therefore, the profiled log-likelihood (with respect to $\boldsymbol{\mu}$) of $\mathbf{X}$ given the parameters $\mathbf{B}$ and $\boldsymbol{\phi}$ is

$$\max_{\boldsymbol{\mu}\in\mathbb{R}^p} \mathcal{L}(\mathbf{X}|\boldsymbol{\mu}, \mathbf{B}, \boldsymbol{\phi}) = \frac{n}{2}\Big(\log\det(\mathbf{B}) - \sum_{j=1}^{p}\big\{\log(\phi_j) + (\mathbf{S}_n\mathbf{B})_{jj}\phi_j^{-1}\big\}\Big). \tag{1.10}$$

[9] In the sense that $\mathbf{E}\big(\mathrm{e}^{x\left(|(\boldsymbol{\xi}_j)_i| - (2/\pi)^{1/2}\right)}\big) \le \mathrm{e}^{bx^2/2}$, for any $x \in \mathbb{R}$.

For a given $\mathbf{B}$, this profiled log-likelihood is a decomposable function of $\boldsymbol{\phi}$ and, therefore, can be easily maximized with respect to $\boldsymbol{\phi}$. This leads to

$$\arg\max_{\boldsymbol{\phi}\in\mathbb{R}_+^p}\max_{\boldsymbol{\mu}\in\mathbb{R}^p}\mathcal{L}(\mathbf{X}|\boldsymbol{\mu},\mathbf{B},\boldsymbol{\phi}) = \big((\mathbf{S}_n\mathbf{B})_{jj}\vee 0\big)_{j\in[p]}. \tag{1.11}$$

Thus, when an estimator $\widehat{\mathbf{B}}$ of $\mathbf{B}^*$ is available, one possible approach for estimating $\boldsymbol{\phi}^*$ is to set

$$\widehat{\phi}_j^{\mathrm{RML}} = (\mathbf{S}_n\widehat{\mathbf{B}})_{jj}\vee 0, \qquad j\in[p]. \tag{1.12}$$

We call this estimator relaxed maximum likelihood (RML) estimator. It will be clear a little bit later why it is called relaxed. The analysis of the risk of the RML estimator is more involved than that of the RV estimator considered in the previous section. This is due to the truncation at the level 0. For this reason, the next result does not provide the precise value of the risk, but just an inequality which is sufficient for our purposes.

Proposition 1.3.3. *If $\widehat{\mathbf{B}}$ estimates $\mathbf{B}^*_{\bullet,j}$ without error, then the risk of the RML estimator of ϕ_j^* satisfies $\mathbf{E}[(\widehat{\phi}_j^{\mathrm{RML}}-\phi_j^*)^2]\geq\frac{1}{n}\big(\phi_j^{*2}+\phi_j^*\boldsymbol{\Sigma}^*_{jj}-O(n^{-1/2})\big)$.*

Before providing the proof of this result, let us present a brief discussion. Note that in view of (1.1), $\boldsymbol{\Sigma}^*_{jj}$ is always not smaller than ϕ_j^* (see Proposition A.2.1 in Appendix A). Furthermore, $\boldsymbol{\Sigma}^*_{jj}>\phi_j^*$ if $\mathbf{B}^*_{j^c,j}$ has at least one nonzero entry. Therefore, the last proposition, combined with Proposition 1.3.1, establishes that the residual variance estimator has an asymptotic variance which is smaller (and, in many cases, strictly smaller) than the asymptotic variance of the maximum likelihood estimator. At a first sight, this is very surprising and seems to be in contradiction with the well established theory [Ibragimov and Has′minskiĭ, 1981; Le Cam and Yang, 2000] of asymptotic efficiency of the maximum likelihood estimator for regular models. Our explanation of this inefficiency of $\widehat{\phi}_j^{\mathrm{RML}}$ is that it is not really the maximum likelihood estimator. It maximizes the likelihood, certainly, but not over the correct set of parameters. Indeed, when we defined the RML estimator we neglected an important property of the vector $\boldsymbol{\phi}^*$: the fact that $\mathbf{B}^*\mathbf{D}_{\boldsymbol{\phi}^*}^{-1}=\mathbf{D}_{\boldsymbol{\phi}^*}^{-1}\mathbf{B}^{*\top}$ (this follows from the symmetry of $\boldsymbol{\Omega}^*$). Ignoring this constraint allowed us to get a tractable optimization problem but caused the loss of the (asymptotic) efficiency of the estimator. This also explains why we call $\widehat{\boldsymbol{\phi}}^{\mathrm{RML}}$ *relaxed* maximum likelihood estimator.

Proof of Proposition 1.3.3. Since $\boldsymbol{\mu}^*$ is assumed to be known and equal to zero, according to (1.1), we have

$$(\mathbf{S}_n\widehat{\mathbf{B}})_{jj}=\frac{1}{n}\mathbf{X}_{\bullet,j}^\top\mathbf{X}\mathbf{B}^*_{\bullet,j}=\frac{\phi_j^{*1/2}}{n}\mathbf{X}_{\bullet,j}^\top\boldsymbol{\xi}_j=\frac{\phi_j^{*1/2}}{n}\big(-\mathbf{X}_{\bullet,j^c}\mathbf{B}^*_{j^c,j}+\phi_j^{*1/2}\,\boldsymbol{\xi}_j\big)^\top\boldsymbol{\xi}_j.$$

Denoting $\eta_1=-\boldsymbol{\xi}_j^\top\mathbf{X}_{\bullet,j^c}\mathbf{B}^*_{j^c,j}$, we get $(\mathbf{S}_n\widehat{\mathbf{B}})_{jj}=\frac{1}{n}(\phi_j^{*1/2}\,\eta_1+\phi_j^*\,\|\boldsymbol{\xi}_j\|_2^2)$. Furthermore, it follows from (1.1) that $\mathbf{E}[(\mathbf{X}_{k,j^c}\mathbf{B}^*_{j^c,j})^2]=\boldsymbol{\Sigma}^*_{jj}-\phi_j^*$ for each k. Since, in addition, for

different ks the random variables $\mathbf{X}_{k,j^c}\mathbf{B}^*_{j^c,j}$ are independent, centered and Gaussian, we get that—in view of the independence of $\boldsymbol{\xi}_j$ and $\mathbf{X}_{\bullet,j^c}$—the conditional distribution of η_1 given $\boldsymbol{\xi}_j$ is Gaussian with zero mean and variance $\|\boldsymbol{\xi}_j\|_2^2(\boldsymbol{\Sigma}^*_{jj}-\phi^*_j)$. Hence, the random variable $\eta=\eta_1/(\|\boldsymbol{\xi}_j\|_2(\boldsymbol{\Sigma}^*_{jj}-\phi^*_j)^{1/2})$ is standard Gaussian, independent of $\|\boldsymbol{\xi}_j\|_2^2$ and

$$(\mathbf{S}_n\widehat{\mathbf{B}})_{jj}=\frac{\sqrt{\phi^*_j\|\boldsymbol{\xi}_j\|_2^2(\boldsymbol{\Sigma}^*_{jj}-\phi^*_j)}}{n}\,\eta+\frac{\phi^*_j}{n}\,\|\boldsymbol{\xi}_j\|_2^2.$$

This relation readily implies that $\mathbf{E}[(\mathbf{S}_n\widehat{\mathbf{B}})_{jj}]=\phi^*_j$ and

$$\mathbf{E}[((\mathbf{S}_n\widehat{\mathbf{B}})_{jj}-\phi^*_j)^2]=\mathbf{V}ar[(\mathbf{S}_n\widehat{\mathbf{B}})_{jj}]=\frac{\boldsymbol{\Sigma}^*_{jj}\phi^*_j+\phi^{*2}_j}{n}.$$

Furthermore, for the fourth moment, we have

$$\begin{aligned}\mathbf{E}[((\mathbf{S}_n\widehat{\mathbf{B}})_{jj}-\phi^*_j)^4]&\le\frac{8\phi^{*2}_j(\boldsymbol{\Sigma}^*_{jj}-\phi^*_j)^2}{n^4}\mathbf{E}[\|\boldsymbol{\xi}_j\|_2^4]\mathbf{E}[\eta^4]+\frac{8\phi^{*4}_j}{n^4}\,\mathbf{E}[(\|\boldsymbol{\xi}_j\|_2^2-n)^4]\\&\le\frac{72\phi^{*2}_j(\boldsymbol{\Sigma}^*_{jj}-\phi^*_j)^2}{n^2}+\frac{8\phi^{*4}_j}{n^4}(60n+12n^2).\end{aligned}$$

To analyze the truncated estimator, we set $\zeta=(\mathbf{S}_n\widehat{\mathbf{B}})_{jj}$. Then $\widehat{\phi}^{\mathrm{RML}}_j=\zeta\cdot\mathbb{1}(\zeta>0)$ and hence,

$$\begin{aligned}\mathbf{E}[(\widehat{\phi}^{\mathrm{RML}}_j-\phi^*_j)^2]&=\mathbf{E}[(\zeta-\phi^*_j)^2\mathbb{1}(\zeta>0)]+\phi^{*2}_j\mathbf{P}(\zeta\le0)\\&=\mathbf{E}[(\zeta-\phi^*_j)^2]-\mathbf{E}[(\zeta-\phi^*_j)^2\mathbb{1}(\zeta\le0)]+\phi^{*2}_j\mathbf{P}(\zeta\le0)\\&\ge\mathbf{E}[(\zeta-\phi^*_j)^2]-\mathbf{E}[(\zeta-\phi^*_j)^4]^{1/2}\mathbf{P}(\zeta\le0)^{1/2}.\end{aligned}$$

We have already computed the first expectation in the right-hand side, as well as upper bounded the second one. Let us show that the probability of the event $\zeta\le0$ goes to zero as n increases to ∞. This follows from the Tchebychev inequality, since $\mathbf{P}(\zeta\le0)=\mathbf{P}(\phi^*_j-\zeta\ge\phi^*_j)\le\mathbf{V}ar[\zeta]/\phi^{*2}_j=O(1/n)$. This completes the proof of the proposition. □

1.3.3 MLE taking into account the symmetry constraints

As we have seen in previous sections, the relaxed maximum likelihood estimator is sub-optimal; in particular, it is less accurate than the residual variance estimator. To check that this lack of efficiency is indeed due to the relaxation of the symmetry constraints, we propose here to analyze the constrained maximum likelihood estimator in the following idealized set-up. We will consider, as in Propositions 1.3.1 and 1.3.3, that $\widehat{\mathbf{B}}$ estimates $\mathbf{B}^*$ without error, and that[10] there is a column $\mathbf{B}^*_{\bullet,i}$ in $\mathbf{B}^*$ such that all the elements of $\mathbf{B}^*_{\bullet,i}$ are different from zero. Without loss of generality, we suppose that $i=1$ and, consequently,

[10] This assumption will be relaxed later in this subsection.

for every $j \in [p]$, we have $\mathbf{B}^*_{j1} \neq 0$ which is equivalent to $\omega^*_{j1} \neq 0$. Therefore, the symmetry constraint $\mathbf{B}^*\mathbf{D}^{-1}_{\boldsymbol{\phi}^*} = \boldsymbol{\Omega}^* = \boldsymbol{\Omega}^{*\top} = \mathbf{D}^{-1}_{\boldsymbol{\phi}^*}\mathbf{B}^{*\top}$ implies that $\mathbf{D}_{\boldsymbol{\phi}^*}\mathbf{B}^* = \mathbf{B}^{*\top}\mathbf{D}_{\boldsymbol{\phi}^*}$ and, in particular, that

$$\mathbf{B}^*_{1j}\phi^*_1 = \mathbf{B}^*_{j1}\phi^*_j, \qquad \forall j \in [p].$$

This relation entails that in the case of known matrix $\mathbf{B}^*$ and unknown vector $\boldsymbol{\phi}^*$, only the first entry of $\boldsymbol{\phi}^*$ needs to be estimated, all the remaining entries can be computed using the first one by the formula $\widehat{\phi}_j = (\mathbf{B}^*_{1j}/\mathbf{B}^*_{j1})\widehat{\phi}_1$.

Proposition 1.3.4. *Under the assumption that the rows of $\mathbf{X}$ are i.i.d. Gaussian vectors with precision matrix $\boldsymbol{\Omega}^* = \mathbf{B}^*\mathbf{D}^{-1}_{\boldsymbol{\phi}^*}$, the maximum likelihood estimator of $\boldsymbol{\phi}^*$ is defined by*

$$\widehat{\phi}^{\mathrm{SML}}_j = \frac{1}{p}(\mathbf{B}^*_{1j}/\mathbf{B}^*_{j1})\ \mathrm{trace}\big(\mathbf{S}_n\mathbf{B}^*\mathbf{D}_{\mathbf{B}^*_{\bullet,1}}\mathbf{D}^{-1}_{\mathbf{B}^*_{1,\bullet}}\big). \tag{1.13}$$

The quadratic risk of this estimator is given by

$$\mathbf{E}[(\widehat{\phi}^{\mathrm{SML}}_j - \phi^*_j)^2] = \frac{2}{np}\,\phi^{*2}_j. \tag{1.14}$$

Furthermore, for every $t > 0$, the following bound on the tails of the maximal error holds true:

$$\mathbf{P}\bigg(\max_{j\in[p]} \frac{|\widehat{\phi}^{\mathrm{SML}}_j - \phi^*_j|}{\phi^*_j} > 2\Big(\frac{t+\log p}{np}\Big)^{1/2} + 2\frac{t+\log p}{np}\bigg) \leq 2\,\mathrm{e}^{-t}\,.$$

Proof. To ease notation, we denote by $\mathbf{D}^*$ the diagonal matrix whose jth element is $\mathbf{B}^*_{j,1}/\mathbf{B}^*_{1,j}$. Then, applying (1.10) for a given $\mathbf{B}^*$, the profiled Gaussian log-likelihood can be written as

$$\max_{\boldsymbol{\mu}\in\mathbb{R}^p} \mathcal{L}(\mathbf{X}|\boldsymbol{\mu}, \mathbf{B}^*, \boldsymbol{\phi}) = \frac{n}{2}\log\det(\mathbf{B}^*) - \frac{n}{2}\sum_{j=1}^{p}\big\{\log(\phi_j) + (\mathbf{S}_n\mathbf{B}^*)_{jj}\phi_j^{-1}\big\}.$$

The goal is to maximize the right-hand side over all the vectors $\boldsymbol{\phi} \in \mathbb{R}^p$ such that $\mathbf{B}^*\mathbf{D}^{-1}_{\boldsymbol{\phi}}$ is a valid precision matrix.

Let us first check that under the conditions of the proposition, for $\mathbf{B}^*\mathbf{D}^{-1}_{\boldsymbol{\phi}}$ to be a valid precision matrix it is necessary and sufficient that $\phi_1 > 0$ and $\phi_j = (\mathbf{B}^*_{1j}/\mathbf{B}^*_{j1})\phi_1$ for every $j \in [p]$. The necessary part follows from that fact that a precision matrix is symmetric and positive-semidefinite, which entails that $(\mathbf{B}^*\mathbf{D}^{-1}_{\boldsymbol{\phi}})_{1j} = (\mathbf{B}^*\mathbf{D}^{-1}_{\boldsymbol{\phi}})_{j1}$ and $(\mathbf{B}^*\mathbf{D}^{-1}_{\boldsymbol{\phi}})_{jj} = \boldsymbol{\phi}_j^{-1} > 0$. Therefore, $\phi_j = (\mathbf{B}^*_{1j}/\mathbf{B}^*_{j1})\phi_1$ and $\phi_1 > 0$. To check the sufficient part, we remark that if $\boldsymbol{\phi}$ satisfies $\phi_j = (\mathbf{B}^*_{1j}/\mathbf{B}^*_{j1})\phi_1$ with $\phi_1 > 0$, then $\mathbf{B}^*\mathbf{D}^{-1}_{\boldsymbol{\phi}} = (\phi^*_1/\phi_1)\mathbf{B}^*\mathbf{D}^{-1}_{\boldsymbol{\phi}^*} = (\phi^*_1/\phi_1)\boldsymbol{\Omega}^*$. This implies that $\mathbf{B}^*\mathbf{D}^{-1}_{\boldsymbol{\phi}}$ is symmetric and positive-semidefinite, hence a valid precision matrix.

The maximum likelihood estimator $\widehat{\boldsymbol{\phi}}^{\mathrm{SML}}$ is thus given by

$$\widehat{\boldsymbol{\phi}}^{\mathrm{SML}} \in \arg \min_{\substack{\boldsymbol{\phi}\in\mathbb{R}_+^p \\ \phi_j=(\mathbf{B}^*_{1j}/\mathbf{B}^*_{j1})\phi_1}} \sum_{j=1}^{p} \big\{ \log(\phi_j) + (\mathbf{S}_n\mathbf{B}^*)_{jj}\phi_j^{-1} \big\},$$

which leads to $\widehat{\phi}_1^{\mathrm{SML}} \in \arg\min_{\phi_1>0} \big\{ p\log(\phi_1) + \phi_1^{-1} \sum_j (\mathbf{S}_n\mathbf{B}^*)_{jj}\mathbf{B}^*_{j1}/\mathbf{B}^*_{1j} \big\}$. The cost function of the last minimization problem is convex, since $(\mathbf{S}_n\mathbf{B}^*)_{jj}\mathbf{B}^*_{j1}/\mathbf{B}^*_{1j} = (\mathbf{S}_n\mathbf{B}^*)_{jj}\phi_1^*/\phi_j^* = \phi_1^*(\mathbf{S}_n\boldsymbol{\Omega}^*)_{jj}$ and hence

$$\sum_j (\mathbf{S}_n\mathbf{B}^*)_{jj}\mathbf{B}^*_{j1}/\mathbf{B}^*_{1j} = \phi_1^* \operatorname{trace}(\mathbf{S}_n\boldsymbol{\Omega}^*) = \phi_1^* \operatorname{trace}(\boldsymbol{\Omega}^{*1/2}\mathbf{S}_n\boldsymbol{\Omega}^{*1/2}) \geq 0.$$

The aforementioned cost function is continuously differentiable and convex, its minimum is attained at the point where the derivative vanishes, which provides $\widehat{\phi}_1^{\mathrm{SML}} = \frac{1}{p}\sum_j (\mathbf{S}_n\mathbf{B}^*)_{jj}\mathbf{B}^*_{j1}/\mathbf{B}^*_{1j}$. Combining with the relation $\widehat{\phi}_j^{\mathrm{SML}} = (\mathbf{B}^*_{1j}/\mathbf{B}^*_{j1})\widehat{\phi}_1^{\mathrm{SML}}$, this leads to (1.13).

To check (1.14), we start by noting that

$$\widehat{\phi}_j^{\mathrm{SML}} = \frac{1}{p}\phi_j^* \operatorname{trace}(\mathbf{S}_n\boldsymbol{\Omega}^*) = \frac{1}{np}\phi_j^* \operatorname{trace}(\mathbf{X}^\top\mathbf{X}\boldsymbol{\Sigma}^{*-1}).$$

Using the well-known commutativity property of the trace operator and setting $\mathbf{Y} = \boldsymbol{\Sigma}^{*-1/2}\mathbf{X}^\top$, we get $\operatorname{trace}(\mathbf{X}^\top\mathbf{X}\boldsymbol{\Sigma}^{*-1}) = \operatorname{trace}(\mathbf{Y}^\top\mathbf{Y})$. Since $\mathbf{X}$ has i.i.d. rows drawn from a $\mathcal{N}_p(0,\boldsymbol{\Sigma}^*)$ distribution, $\mathbf{Y}$ has i.i.d. columns drawn from $\mathcal{N}_p(0,\mathbf{I}_p)$ distribution. Hence, the random variable $\operatorname{trace}(\mathbf{Y}^\top\mathbf{Y}) = \sum_{j\in[p],k\in[n]} \mathbf{Y}_{jk}^2$ is distributed according to χ^2_{np} distribution. This readily implies that $\widehat{\phi}_j^{\mathrm{SML}}$ is an unbiased estimator of ϕ_j^* and, therefore, its quadratic risk coincides with its variance and is given by (1.14).

The proof of the last claim of the proposition is very similar to that of the second claim of Proposition 1.3.1. □

Assuming that there exists $i \in [p]$ such that for any $j \in [p]$, $\omega^*_{ij} \neq 0$, put differently that the i-th node of the graph $\mathscr{G}^*$ is connected by an edge to any other node is quite restrictive. Among other implications, it entails that the graph $\mathscr{G}^*$ is connected which might be a strong assumption. It is therefore useful to adapt what precedes to the case where the graph $\mathscr{G}^*$ has more than one connected component. The rest of this subsection is devoted to the description of this adaptation.

We note $\mathscr{C}$ the set of the connected components of the graph $\mathscr{G}^*$. Each connected component $c \in \mathscr{C}$ is a subset of vertices of $\mathscr{G}^*$ whose cardinality is denoted by p_c. Clearly, the sum of p_c over all $c \in \mathscr{C}$ equals p. For two vertices i and j, we will write $i \sim_{\mathscr{G}^*} j$ for indicating that they belong to the same connected component. Thus, each connected component is a class of equivalence with respect to the relation $\sim_{\mathscr{G}^*}$. Let $i \sim_{\mathscr{G}^*} j$ be two vertices from $c \in \mathscr{C}$ and let C_{ji} be a path connecting these two vertices, that is, C_{ji} is a

sequence of q distinct vertices $\{v_1, \dots, v_q\}$ such that $v_1 = j$, $v_q = i$, $q \leq p_c$ and each pair (v_h, v_{h+1}) is connected by an edge in $\mathscr{G}^*$. Recall that the symmetry of the precision matrix $\mathbf{\Omega}^* = \mathbf{B}^* \mathbf{D}_{\boldsymbol{\phi}^*}^{-1}$ implies that $\mathbf{B}^*_{v_h, v_{h+1}} \phi^*_{v_h} = \mathbf{B}^*_{v_{h+1}, v_h} \phi^*_{v_{h+1}}$ for every $h \in [q-1]$. This readily yields

$$\phi^*_j = \phi^*_i \prod_{1 \leq h < q} \big(\mathbf{B}^*_{v_{h+1}, v_h} / \mathbf{B}^*_{v_h, v_{h+1}} \big).$$

To ease notation, we introduce the $p \times p$ diagonal matrix $\mathbf{\Delta}^*_j$ the diagonal entries of which are defined by

$$(\mathbf{\Delta}^*_j)_{ii} = \mathbb{1}(i \sim_{\mathscr{G}^*} j) \prod_{1 \leq h < q} \big(\mathbf{B}^*_{v_{h+1}, v_h} / \mathbf{B}^*_{v_h, v_{h+1}} \big), \tag{1.15}$$

where $\{v_1, \dots, v_q\} = C_{ji}$ is any path connecting j to i in $\mathscr{G}^*$. With this notation, $\phi^*_j = (\mathbf{\Delta}^*_j)_{ii} \phi^*_i$. One can reproduce the arguments of the proof of Proposition 1.3.4 to check that the maximum likelihood estimator of $\boldsymbol{\phi}^*$, if $\mathbf{B}^*$ is known (and therefore so is $\mathbf{\Delta}^*_j$), is defined by

$$\widehat{\phi}^{\mathrm{SML}}_j = \frac{1}{p_c} \operatorname{trace}\big(\mathbf{\Delta}^*_j \mathbf{S}_n \mathbf{B}^* \big), \tag{1.16}$$

for j belonging to the connected component c.

Comparing the results of Propositions 1.3.1, 1.3.3 and 1.3.4, we observe that the RV estimator outperforms the RML estimator, but—at least in the case where there is a column in $\mathbf{B}^*$ which has only nonzero entries—they are both dominated by the maximum likelihood estimator that takes advantage of the symmetry constraints. Furthermore, using the same type of arguments as those of Proposition 1.3.4, one can check that if the vertex j of the graph $\mathscr{G}^*$ belongs to a connected component of cardinal p_c then the risk of the MLE in the ideal case of known $\mathbf{B}^*$ is equal to $\frac{2}{np_c} \phi^{*2}_j$. This shows that in the ideal case the MLE systematically outperforms the widely used residual variance estimator, and the gain in the risk may be huge for vertices belonging to large connected components. On the other extreme, all the three estimators discussed in the previous section coincide when the matrix $\mathbf{B}^*$ is diagonal.

In order to apply equation (1.16) for estimating $\boldsymbol{\phi}^*$ when an estimator $\widehat{\mathbf{B}}$ of $\mathbf{B}^*$ is available, we need to construct an estimator $\widehat{\mathscr{G}}$ of the graph $\mathscr{G}^*$. We propose here an original approach for deriving $\widehat{\mathscr{G}}$ from $\widehat{\mathbf{B}}$. It is based on the observation that $\mathbf{B}^*_{ij} \mathbf{B}^*_{ji} = {\omega^*_{ij}}^2 / (\omega^*_{ii} \omega^*_{jj})^2$, the square of the partial correlation between the i-th and j-th variables. As mentioned earlier, this quantity is always between 0 and 1 and provides a convenient rule of selection for the edges to keep in the graph. More precisely, we connect i to j if the estimated squared partial correlation $\widehat{\mathbf{B}}_{ij} \widehat{\mathbf{B}}_{ji}$ is larger than a prescribed threshold $t \in (0, 1)$. In our implementation, we chose (somewhat arbitrarily) the threshold $t = 0.01 \wedge n^{-1/2}$.

Note that when $\mathbf{B}^*$ is replaced by an estimator, the right-hand side of (1.16) is not necessarily invariant with respect to the choice of the path connecting i to j. Therefore, even when $\widehat{\mathbf{B}}$ and $\widehat{\mathscr{G}}$ are fixed, if $\widehat{\mathscr{G}}$ contains cycles there are many ways of estimating

$\boldsymbol{\phi}^*$ based on (1.16) depending on how the paths are chosen. We have tried two possible approaches: the minimum spanning tree and the shortest path tree based on the following weight function[11] defined on the edges:

$$\mathbf{W}_{ij} = \begin{cases} \exp\big(-\widehat{\mathbf{B}}_{ij}\widehat{\mathbf{B}}_{ji}\big)\mathbb{1}(\widehat{\mathbf{B}}_{ij}\widehat{\mathbf{B}}_{ji} > t), & \text{for } i \neq j, \\ 0, & \text{otherwise.} \end{cases}$$

Combining these ingredients, we get the algorithm summarized in Algorithm 1.1.

Algorithm 1.1: Estimator $\widehat{\boldsymbol{\phi}}^{\mathrm{SML}}$ based on shortest path trees or minimum spanning trees

Input: matrices $\mathbf{X}$ and $\widehat{\mathbf{B}}$, threshold t.
Output: vector $\widehat{\boldsymbol{\phi}}^{\mathrm{SML}}$.
1: compute the matrix of weights $\mathbf{W}$.
2: initialize k to 1.
repeat
  3: choose the node with the largest degree as root.
  4: compute the shortest path tree (or the minimum spanning tree) $\mathscr{T}_k$ from the chosen root.
  5: estimate $\widehat{\boldsymbol{\phi}}^{\mathrm{SML}}$'s elements related to $\mathscr{T}_k$ using Eq. (1.16).
  6: remove all the nodes of the tree $\mathscr{T}_k$ from the initial graph.
  7: increment k.
until *graph is empty*

The rationale behind the foregoing definition of the weights and the use of the minimum spanning tree or shortest path tree algorithm is to favor the paths that are short and contain edges corresponding to large (in absolute value) partial correlations. The aim is to reduce the risk of propagating the estimation error of $\widehat{\mathbf{B}}$. We have implemented both versions of the algorithm and have observed that the version using the minimum spanning tree leads to better results. More details on the implementation and computational complexity are given in the next section.

1.3.4 Penalized maximum likelihood estimation

We have seen that enforcing symmetry constraints is beneficial when the matrix $\widehat{\mathbf{B}}$ has a small error, but raises intricate issues related to the graph estimation and, more importantly, path selection in the graph. A workaround to this issue is to replace the hard constraints by a penalty term that measures the degree of violation of the constraints. This provides an intermediate solution between the SML and the RML. More precisely, we

[11] A weight equal to zero corresponds to the absence of edge.

p	30			60			90		
n	200	800	2000	200	800	2000	200	800	2000
$\mathbf{B}^*$ estimated by square-root Lasso									
RV	**0.883 (.077)**	**0.399 (.036)**	**0.224 (.016)**	**1.425 (.075)**	**0.649 (.030)**	**0.374 (.022)**	**1.849 (.085)**	**0.853 (.029)**	**0.495 (.019)**
RML	1.356 (.079)	0.786 (.040)	0.532 (.017)	2.114 (.082)	1.234 (.032)	0.841 (.022)	2.705 (.090)	1.590 (.029)	1.086 (.019)
SML	1.476 (.098)	0.805 (.040)	0.548 (.018)	2.388 (.164)	1.250 (.032)	0.852 (.021)	3.104 (.188)	1.608 (.028)	1.096 (.020)
PML	1.371 (.079)	0.792 (.041)	0.539 (.017)	2.134 (.078)	1.236 (.032)	0.846 (.021)	2.728 (.091)	1.593 (.030)	1.089 (.019)
$\mathbf{B}^*$ estimated by square-root Lasso followed by OLS									
RV	**0.726 (.079)**	**0.340 (.045)**	**0.241 (.016)**	**1.088 (.076)**	**0.616 (.051)**	**0.354 (.020)**	**1.365 (.080)**	**0.854 (.046)**	**0.443 (.018)**
RML	**0.726 (.079)**	**0.340 (.045)**	**0.241 (.016)**	**1.088 (.076)**	**0.616 (.051)**	**0.354 (.020)**	**1.365 (.080)**	**0.854 (.046)**	**0.443 (.018)**
SML	0.807 (.082)	0.440 (.058)	0.280 (.018)	1.193 (.088)	0.793 (.066)	0.381 (.018)	1.557 (.170)	1.116 (.089)	0.468 (.018)
PML	0.737 (.074)	0.419 (.051)	0.302 (.018)	1.095 (.071)	0.722 (.052)	0.405 (.019)	1.377 (.081)	0.984 (.044)	0.494 (.019)
$\mathbf{B}^*$ is estimated without error									
RV	0.263 (.034)	0.132 (.017)	0.081 (.012)	0.370 (.038)	0.179 (.017)	0.115 (.008)	0.455 (.038)	0.222 (.019)	0.143 (.012)
RML	0.322 (.042)	0.165 (.018)	0.104 (.013)	0.463 (.038)	0.227 (.022)	0.144 (.011)	0.562 (.040)	0.280 (.020)	0.178 (.015)
SML	**0.043 (.030)**	**0.024 (.018)**	**0.010 (.010)**	**0.042 (.030)**	**0.018 (.014)**	**0.011 (.009)**	**0.042 (.037)**	**0.015 (.013)**	**0.010 (.007)**
PML	0.079 (.025)	0.043 (.015)	0.023 (.007)	0.107 (.028)	0.049 (.012)	0.030 (.007)	0.128 (.027)	0.059 (.011)	0.039 (.007)

Table 1.1: Performance of the estimators of diagonal elements of the precision matrix in Model 1. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

propose a penalized maximum likelihood (PML) estimator of $\boldsymbol{\phi}^*$ defined by

$$\widehat{\boldsymbol{\phi}}^{\mathrm{PML}} \in \operatorname*{arg\,min}_{\boldsymbol{\phi}\in(0,1]^p} \left\{ \sum_{j=1}^{p} \{\log(\phi_j) + (\mathbf{S}_n\widehat{\mathbf{B}})_{j,j}\phi_j^{-1}\} + \kappa \sum_{\substack{i<j \\ \widehat{\mathbf{B}}_{ji}\widehat{\mathbf{B}}_{ij}>t}} \frac{(\widehat{\mathbf{B}}_{ji}\phi_i^{-1} - \widehat{\mathbf{B}}_{ij}\phi_j^{-1})^2}{\widehat{\mathbf{B}}_{ij}^2 + \widehat{\mathbf{B}}_{ji}^2} \right\}, \quad (1.17)$$

where $\kappa > 0$ is a tuning parameter responsible for the trade-off between the likelihood and the constraint violation. The choice $\kappa = \infty$ corresponds to enforcing the symmetry constraints: its main shortcoming is that the feasible set might very well be empty. On the other extreme, when $\kappa = 0$, the PML coincides with the RML. The PML estimator also coincides with the previous ones if $\widehat{\mathbf{B}}$ is known to be diagonal.

Note that the parameter t appearing in the penalty term of the PML plays the same role as the one used in the SML. The definition of the feasible set in the above optimization problem is justified by the fact that we assume all the individual variances of the features to be equal to one. In other terms, the assumption $\mathbf{V}ar(\mathbf{X}_{1,j}) = 1$ in (1.1) implies that $\phi_j^* \leq 1$. Making the change of variable $\mathbf{v} = (1/\phi_j)_{j\in[p]}$, the optimization problem of Eq. (1.17) becomes convex with the feasible set $\mathbf{v} \in [1 + \infty)^p$ and the objective function:

$$f(\mathbf{v}) = \sum_{j=1}^{p} \big\{ -\log(v_j) + (\mathbf{S}_n \widehat{\mathbf{B}})_{j,j} v_j \big\} + \kappa \sum_{\substack{i<j \\ \widehat{\mathbf{B}}_{ji}\widehat{\mathbf{B}}_{ij} > t}} \frac{(\widehat{\mathbf{B}}_{ji} v_i - \widehat{\mathbf{B}}_{ij} v_j)^2}{\widehat{\mathbf{B}}_{ij}^2 + \widehat{\mathbf{B}}_{ji}^2}. \tag{1.18}$$

Furthermore, if we restrict the feasible set to $\mathbf{v} \in \mathcal{V} = [1, n^{1/2}]^p$, the problem becomes strongly convex. In addition, on this restricted feasible set the gradient of the objective function is Lipschitz-continuous.

It is possible to use the standard steepest gradient descent algorithm with a fixed step-size for efficiently approximating the solution $\widehat{\boldsymbol{\phi}}^{\mathrm{PML}}$. Indeed, in the optimization problem (1.18), if ∇f is Lipschitz-continuous with constant $L < \infty$ and strongly convex with constant $l > 0$, the gradient descent algorithm with a constant step-size $t = 2/(l+L)$ converges at a linear rate (see Nesterov [2004] for a detailed proof). Note that the convergence rate depends on L/l which is an upper bound on the condition number of the Hessian matrix $\nabla^2 f(\mathbf{v})$; this ratio should not be too high for the algorithm to converge fast. Unfortunately, the values of l and L that we manage to obtain in our problem are far too loose. That is why we resort to a steepest descent algorithm with adaptive step-size and scaled descent direction $-\nabla f(\mathbf{v}_h)/\|\nabla f(\mathbf{v}_h)\|_2$. More details on the implementation are provided in Section 1.4.2.

1.4 Experimental evaluation

In this section, we describe the experimental set-up and report the results of the numerical experiments performed on synthetic data-sets. We also provide detailed explanation of the implementation used for the symmetry-enforced and the penalized maximum-likelihood estimators. A companion R package called DESP (for Diagonal Elements of Sparse Precision-Matrices estimation) is created and uploaded on CRAN[12].

[12] http://cran.r-project.org/web/packages/DESP/index.html

p	30			60			90		
n	200	800	2000	200	800	2000	200	800	2000
$\mathbf{B}^*$ estimated by square-root Lasso									
RV	**0.400 (.059)**	**0.125 (.020)**	**0.070 (.009)**	**0.632 (.047)**	**0.174 (.023)**	**0.094 (.011)**	**0.821 (.051)**	**0.215 (.020)**	**0.113 (.012)**
RML	1.048 (.061)	0.508 (.020)	0.320 (.015)	1.644 (.048)	0.780 (.023)	0.491 (.014)	2.120 (.053)	0.997 (.023)	0.626 (.014)
SML	1.334 (.221)	0.539 (.028)	0.340 (.018)	2.246 (.277)	0.824 (.039)	0.520 (.023)	3.243 (.516)	1.047 (.034)	0.653 (.020)
PML	1.130 (.068)	0.530 (.020)	0.333 (.016)	1.790 (.049)	0.813 (.026)	0.508 (.016)	2.311 (.054)	1.036 (.024)	0.645 (.015)
$\mathbf{B}^*$ estimated by square-root Lasso followed by OLS									
RV	**0.247 (.053)**	0.101 (.015)	0.065 (.009)	**0.322 (.057)**	0.129 (.019)	0.081 (.007)	**0.381 (.061)**	0.150 (.017)	0.095 (.009)
RML	**0.247 (.053)**	0.101 (.015)	0.065 (.009)	**0.322 (.057)**	0.129 (.019)	0.081 (.007)	**0.381 (.061)**	0.150 (.017)	0.095 (.009)
SML	0.329 (.107)	**0.096 (.016)**	0.065 (.010)	0.622 (.299)	0.129 (.021)	**0.076 (.010)**	0.882 (.501)	0.147 (.020)	0.090 (.011)
PML	**0.247 (.068)**	0.098 (.017)	**0.064 (.011)**	0.337 (.075)	**0.125 (.021)**	0.077 (.009)	0.441 (.101)	**0.142 (.017)**	**0.089 (.011)**
$\mathbf{B}^*$ is estimated without error									
RV	0.204 (.032)	0.101 (.015)	0.065 (.008)	0.258 (.033)	0.129 (.019)	0.081 (.007)	0.300 (.030)	0.149 (.015)	0.095 (.009)
RML	0.280 (.038)	0.136 (.017)	0.086 (.011)	0.354 (.032)	0.177 (.019)	0.113 (.010)	0.429 (.038)	0.214 (.021)	0.135 (.012)
SML	**0.033 (.022)**	**0.012 (.008)**	**0.008 (.007)**	**0.024 (.017)**	**0.012 (.009)**	**0.008 (.006)**	**0.027 (.019)**	**0.011 (.008)**	**0.007 (.006)**
PML	0.065 (.021)	0.027 ((.011)	0.019 (.006)	0.065 (.020)	0.031 (.009)	0.021 (.006)	0.073 (.022)	0.035 (.010)	0.023 (.006)

Table 1.2: Performance of the estimators of diagonal elements of the precision matrix in Model 2. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

1.4.1 Experiments on synthetic datasets

We conducted a comprehensive experimental evaluation of the accuracy of different estimates of diagonal elements of the precision matrix. In order to cover as many situations as possible, we used in experiments our six different forms of precision matrices along with various values for n and p. In each configuration, we considered several methods of

estimating the matrix $\mathbf{B}^*$.

Let us first describe in a precise manner the precision matrices used in our experiments. It is worthwhile to underline here that all the precision matrices are normalized in such a way that all the diagonal entries of the corresponding covariance matrix $\mathbf{\Sigma}^* = (\mathbf{\Omega}^*)^{-1}$ are equal to one. To this end, we first define a $p \times p$ positive semidefinite matrix $\mathbf{A}$ and then set $\mathbf{\Omega}^* = (\mathrm{diag}(\mathbf{A}^{-1}))^{\frac{1}{2}}\mathbf{A}(\mathrm{diag}(\mathbf{A}^{-1}))^{\frac{1}{2}}$. The matrices $\mathbf{A}$ used in the six models for which the experiments are carried out are defined as follows.

Model 1: $\mathbf{A}$ is a Toeplitz matrix with the entries $\mathbf{A}_{ij} = 0.6^{|i-j|}$ for any $i, j \in [p]$.

Model 2: We start by defining a $p \times p$ pentadiagonal matrix with the entries

$$\bar{\mathbf{A}}_{ij} = \begin{cases} 1 & , \text{ for } |i-j| = 0, \\ -1/3 & , \text{ for } |i-j| = 1, \\ -1/10 & , \text{ for } |i-j| = 2, \\ 0 & , \text{ otherwise.} \end{cases}$$

Then, we denote by $\mathbf{A}$ the matrix with the entries $\mathbf{A}_{ij} = (\bar{\mathbf{A}}^{-1})_{ij}\mathbb{1}(|i-j| \leq 2)$. One can check that the matrix $\mathbf{A}$ defined in such a way is positive semidefinite.

Model 3: We set $\mathbf{A}_{ij} = 0$ for all the off-diagonal entries that are neither on the first row nor on the first column of $\mathbf{A}$. The diagonal entries of $\mathbf{A}$ are

$$\mathbf{A}_{11} = p, \qquad \mathbf{A}_{ii} = 2, \quad \text{for any} \quad i \in \{2, \ldots, p\},$$

whereas the off-diagonal entries located either on the first row or on the first column are $\mathbf{A}_{1i} = \mathbf{A}_{i1} = \sqrt{2}$ for $i \in \{2, \ldots, p\}$.

Model 4: We introduce the integer $k = \lceil \sqrt{p} \rceil$ and define a sparse $k \times k$ matrix $\bar{\mathbf{A}}$ so that its only nonzero elements are $\bar{\mathbf{A}}_{11} = k$ and, for any $i \in [2; k]$, $\bar{\mathbf{A}}_{ii} = 2k$ and $\bar{\mathbf{A}}_{1i} = \bar{\mathbf{A}}_{i1} = \sqrt{2}$. Then, we set

$$\mathbf{A} = \begin{pmatrix} \bar{\mathbf{A}} & 0 \\ 0 & \mathbf{I}_{p-k} \end{pmatrix}.$$

Model 5: We introduce $k = \lceil \sqrt{p} \rceil$ and define a sparse $k \times k$ matrix $\bar{\mathbf{A}}$ so that its only nonzero elements are $\bar{\mathbf{A}}_{11} = 50$ and, for any $i \in [2; k]$, $\bar{\mathbf{A}}_{ii} = 5$ and $\bar{\mathbf{A}}_{1i} = \bar{\mathbf{A}}_{i1} = 5/2$. Then, similarly to previous model, we set

$$\mathbf{A} = \begin{pmatrix} \bar{\mathbf{A}} & 0 \\ 0 & \mathbf{I}_{p-k} \end{pmatrix}.$$

Model 6: We set $k = 6$, $p' = k\lceil p/k \rceil$ and define the $k \times k$ matrix $\bar{\mathbf{A}}$ as in model 5 above.

Then, we build the $p' \times p'$ block-diagonal matrix $\mathbf{A}$ by

$$\mathbf{A} = \underbrace{\begin{pmatrix} \bar{\mathbf{A}} & & 0 \\ & \ddots & \\ 0 & & \bar{\mathbf{A}} \end{pmatrix}}_{\lceil p/k \rceil - \text{times}}.$$

Note that, in general, the resulting precision matrix in this model is not of size $p \times p$ but of size $p' \times p'$ with $p' = 6\lceil p/6 \rceil$. However, since in the experiments reported in this section p is always a multiple of 6, we have $p = p'$.

In this experimental evaluation, we compare the performance of the following four estimators—introduced in previous sections—of the diagonal elements of the precision matrix:

- RV corresponds to the residual variance estimator defined in Section 1.3.1.
- RML corresponds to the relaxed maximum likelihood estimator described by equation (1.12).
- SML corresponds to the symmetry-enforced maximum likelihood estimator described in Algorithm 1.1.
- PML corresponds to the penalized maximum likelihood estimator described by equation (1.17).

Note that all these algorithms need an estimator of the matrix $\mathbf{B}^*$ to produce an estimator of the diagonal entries of the precision matrix. We conducted experiments in three different scenarios. The first scenario is when the matrix $\mathbf{B}^*$ is estimated column-by-column by the square-root Lasso, using the penalization parameter $\lambda = \sqrt{2 \log p}$. This value for λ is commonly called the universal choice and has proved to lead to optimal theoretical results and fairly good empirical results [Dalalyan and Chen, 2012; Sun and Zhang, 2012; Dalalyan et al., 2013]. The second scenario is when the matrix $\mathbf{B}^*$ is estimated column-by-column by the ordinary least squares estimator applied to the covariates that correspond to nonzero entries of the square-root Lasso estimator[13] with the aforementioned value of λ. Finally, the third scenario is an unrealistic one; it corresponds to the case of a known matrix $\mathbf{B}^*$. This scenario is included in the experimental evaluation in order to check the consistency between the theoretical and the empirical results as well as in order to better understand how the error in estimating $\mathbf{B}^*$ impacts the quality of estimation of the diagonal entries of the precision matrix.

Thus, each configuration of our empirical study corresponds to choosing

- a model out of 6 models described above

[13] A discussion on the strengths and weaknesses of this estimator can be found in [Belloni and Chernozhukov, 2013; Lederer, 2014].

- a dimension $p \in \{30, 60, 90\}$
- a sample size $n \in \{200, 800, 2000\}$
- a method of estimating $\mathbf{B}^*$.

In each configuration, we computed the estimators RV, RML, SML and PML for 50 independent datasets. Using these $R = 50$ replications, we estimate the expected risk of estimating $\boldsymbol{\phi}^*$, $\mathbf{E}(\|\boldsymbol{\phi}^* - \widehat{\boldsymbol{\phi}}\|_2)$, by the average $\frac{1}{R}\sum_{r=1}^{R} \|\boldsymbol{\phi}^* - \widehat{\boldsymbol{\phi}}_{(r)}\|_2$. In Tables 1-6, we report these averages along with the standard deviations of the errors measured by ℓ_2-vector norm. All the experiments were conducted in R [R Core Team, 2016], using the Mosek solver (see Andersen and Andersen [2000]) for computing the square-root Lasso estimator by second-order cone programming. We note that other general-purpose solvers like Gurobi Gurobi Optimization [2015] or SCS O'Donoghue et al. [2013] produce comparable results. Besides, in terms of computational efficiency, we recall that using coordinate descent to obtain the square-root Lasso estimates is better[14].

The results when the dimension is smaller than the sample size are reported in Tables 1.1-1.6. In Tables 1.7-1.12, we present experimental measures of performance obtained for a sample size $n = 50$ for dimensions $p = 30, 60, 90$ and with $\kappa = \frac{1}{3}\sqrt{\log p}$ for the PML estimation. For the last two values of p, the dimension is larger than the sample size.

In the ideal case when $\mathbf{B}^*$ is estimated without error (by itself), the empirical results reflect perfectly the theoretical results of the previous sections. The comparison of the performance of the estimators indicates that the maximum likelihood estimators SML and PML are preferable to the residual variance estimator. The maximum likelihood estimator considering symmetry constraints outperforms all the other estimators. However, in practice when $\widehat{\mathbf{B}}$ is obtained by the square-root Lasso without any refinement, $\widehat{\boldsymbol{\phi}}^{\mathrm{RV}}$ outperforms all the other estimators in the vast majority of configurations. Some exceptions can be observed in models 5 and 6 (see the top part of Tables 1.5 and 1.6, where RV is slightly worse than the other procedures for small sample sizes ($n = 200$), or Tables 1.11 and 1.12). It should be, however, acknowledged that the difference of the quality between the estimators in these cases is not large enough to advocate for using RML, SML or PML.

It is interesting to observe what happens when an additional step of estimation of $\mathbf{B}^*$ using the ordinary least squares on the sparsity pattern provided by the square-root Lasso is performed. The impact of this step is not the same in all the models under consideration. In particular, the quality of estimation is mostly improved for all the four estimators in models 1 and 2. Furthermore, thanks to this variable selection step, the maximum-likelihood-type estimators perform nearly as well as the residual variance estimator RV. In model 3, the variable selection step deteriorates the quality of estimation in most configurations, whereas

[14] In addition, the coordinate descent algorithm tends to produce estimated coordinates that are exactly zero, while the SOCP solutions are in general only approximately zero.

p	30			60			90		
n	200	800	2000	200	800	2000	200	800	2000
$\mathbf{B}^*$ estimated by square-root Lasso									
RV	**0.273 (.042)**	**0.138 (.016)**	**0.084 (.010)**	**0.402 (.036)**	**0.194 (.014)**	**0.123 (.012)**	**0.524 (.037)**	**0.243 (.017)**	**0.150 (.011)**
RML	0.509 (.062)	0.272 (.022)	0.173 (.018)	0.722 (.061)	0.395 (.026)	0.261 (.013)	0.880 (.069)	0.496 (.028)	0.321 (.014)
SML	1.080 (.132)	0.678 (.095)	0.375 (.045)	1.276 (.146)	0.802 (.075)	0.641 (.050)	1.235 (.137)	0.651 (.052)	0.454 (.029)
PML	0.509 (.062)	0.272 (.021)	0.173 (.017)	0.722 (.061)	0.395 (.026)	0.261 (.013)	0.880 (.069)	0.496 (.028)	0.322 (.014)
$\mathbf{B}^*$ estimated by square-root Lasso followed by OLS									
RV	**0.792 (.192)**	**0.144 (.051)**	**0.084 (.010)**	**2.251 (.203)**	**1.857 (.161)**	**0.943 (.221)**	**3.261 (.184)**	**3.815 (.157)**	**3.689 (.120)**
RML	**0.792 (.192)**	**0.144 (.051)**	**0.084 (.010)**	**2.251 (.203)**	**1.857 (.161)**	**0.943 (.221)**	**3.261 (.184)**	**3.815 (.157)**	**3.689 (.120)**
SML	1.211 (.131)	0.610 (.106)	0.336 (.057)	2.515 (.189)	1.956 (.143)	1.095 (.194)	3.415 (.175)	3.832 (.152)	3.700 (.118)
PML	0.879 (.175)	0.150 (.051)	**0.084 (.011)**	2.366 (.207)	**1.857 (.160)**	**0.943 (.221)**	3.342 (.176)	3.816 (.157)	**3.689 (.120)**
$\mathbf{B}^*$ is estimated without error									
RV	0.267 (.041)	0.138 (.016)	0.084 (.010)	0.380 (.036)	0.192 (.014)	0.122 (.011)	0.476 (.033)	0.237 (.018)	0.148 (.011)
RML	0.330 (.046)	0.163 (.016)	0.104 (.013)	0.469 (.044)	0.229 (.023)	0.151 (.013)	0.584 (.048)	0.289 (.020)	0.178 (.014)
SML	**0.042 (.035)**	**0.019 (.013)**	**0.012 (.009)**	**0.044 (.033)**	**0.021 (.015)**	**0.011 (.007)**	**0.048 (.041)**	**0.021 (.017)**	**0.011 (.010)**
PML	0.330 (.046)	0.163 (.016)	0.104 (.013)	0.470 (.044)	0.229 (.023)	0.151 (.012)	0.584 (.048)	0.289 (.020)	0.178 (.014)

Table 1.3: Performance of the estimators of diagonal elements of the precision matrix in Model 3. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

in models 4-6 this step has almost no consequence on the estimation accuracy.

The graphics of Figure 1.1 are drawn for Model 2 with $p = 60$. The left plot corresponds to the estimation error—measured by ℓ_2-vector norm—as a function of the sample size in the scenario $\widehat{\mathbf{B}} = \mathbf{B}^*$, whereas the central plot corresponds to the same error when $\mathbf{B}^*$ is estimated by the OLS on the sparsity pattern furnished by the square-root Lasso. The

right plot is just a zoom on the center plot. These plots illustrate the convergence to zero of the error of estimation for the estimators considered in this chapter. The speed of convergence in these empirical results, as expected, is nearly $n^{-1/2}$ for fixed dimension p.

1.4.2 Details on the implementation

Symmetry-enforced maximum likelihood. As we explained earlier, the product structure of the term $\mathbf{\Delta}_j^*$ in (1.15) may cause the amplification of the estimation error when passing from $\widehat{\mathbf{B}}$ to $\widehat{\boldsymbol{\phi}}$. In order to reduce as much as possible this phenomenon, we suggested to choose the path $\mathscr{C}$ by minimizing its length. In addition, the fact that some entries of $\mathbf{B}^*$ appear in the denominator of $\mathbf{\Delta}_j^*$, make it unsuitable to include in $\mathscr{C}$ edges corresponding to small values of $\widehat{\mathbf{B}}_{ij}$. The combination of these two arguments suggests to define edge weights as decreasing functions of $\widehat{\mathbf{B}}_{ij}$ and to look for paths that somehow minimize the overall weight defined as the sum of the weights of the edges contained in $\mathscr{C}$.

The two versions of the SML algorithm that have been implemented and tested in this work make use of the minimum spanning tree (MST) and the shortest path tree (SPT) in the step of determining the way of computation the elements of $\widehat{\boldsymbol{\phi}}$ belonging to a connected component $\mathscr{C}$ of the graph $\widehat{\mathscr{G}}$. A MST of $\mathscr{C}$ is a tree that spans $\mathscr{C}$ and has the smallest total weight among all the spanning trees of $\mathscr{C}$. The shortest path tree having a given node r as a root is a spanning tree $\mathscr{T}$ of $\mathscr{C}$ such that for any node $j \in \mathscr{C}$ the weight of the path from j to r in $\mathscr{T}$ is the smallest among the weights of all possible paths from j to r in $\mathscr{C}$.

We have used the Kruskal [Kruskal, 1956] algorithm for finding the MST and the Jarnik-Prim-Dijkstra algorithm [Jarník, 1930; Prim, 1957; Dijkstra, 1959] for the SPT. The worst-case computational complexities of the construction of these trees are the following [Cormen et al., 2009]. When the graph $\mathscr{G}$ has p nodes and q edges, the Kruskal algorithm runs in $O(q \log p)$ time. Its output is a set of MSTs per connected component. The version of the SML based on the shortest path tree requires $O(p+q)$ operations to find the connected components. In a connected component having p_c nodes and q_c edges, the node of largest degree can be obtained in $O(q_c)$ operations, while the computational complexity of finding the shortest paths from a node to all the others is $O(q_c \log(p_c))$. Therefore, determining a shortest path tree per connected component has a complexity of $O(p + q\log(p))$, or $O(sp\log(p))$ where s is the maximal degree of a node of $\widehat{\mathscr{G}}$. Thus, the computational complexities of the two versions of the SML estimator are comparable and, at most, of the order $O(sp\log(p))$.

In our experiments, we have also tried[15] a third version consisting in computing the shortest path trees from every node of a connected component and then choosing the one with the minimal overall weight, rather than first choosing the root as the node having largest

[15] We used the package RBGL of R [Carey et al., 2015] for various algorithms related to weighted graphs.

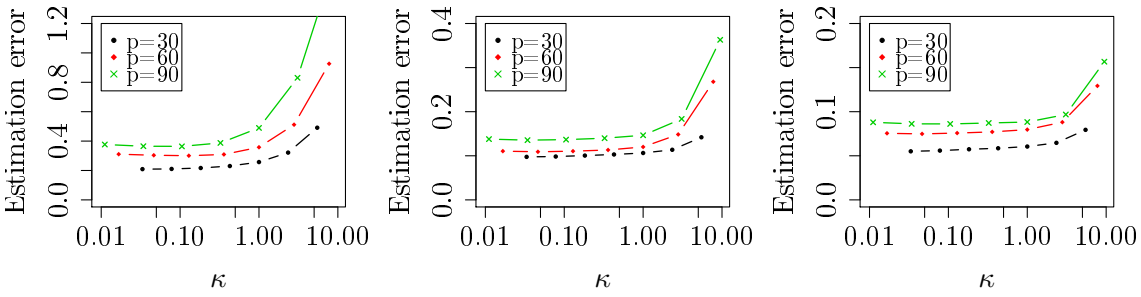


Figure 1.2: The estimation error of the PML as a function of κ. The plots are obtained for the synthetic experiment of Model 2 with various values of p and for $n = 200$ (left), $n = 800$ (middle) and $n = 2000$. Please note that the limits of the y-axis are not the same in the three plots and that the x-axis is presented in logarithmic scale.

degree. Several other variants have been tested as well, but the simplest version based on choosing the MST has lead to the best empirical results.

Penalized maximum likelihood. As mentioned earlier, the PML estimator is computed by solving the optimization problem (1.18). We implement a steepest descent algorithm with adaptive step-size and scaled descent direction $-\nabla f(\mathbf{v}_h)/\|\nabla f(\mathbf{v}_h)\|_2$. At each iteration, one common adaptation for every coordinate of the descent direction is performed. If the objective function increases, the current iteration is done again with a halved step-size. On the opposite, if the objective function decreases, the step-size is increased by a constant factor for the next iteration.

Mathematically speaking, the update operations for our gradient descent algorithm are

$$\mathbf{v}_0 = \mathbf{1}, \quad \mathbf{v}_{h+1} = \mathbf{v}_h + t_h \mathbf{u}_h, \qquad h = 0, 1, 2, \ldots, \tag{1.19}$$

where the descent direction is $\mathbf{u}_h = -\nabla f(\mathbf{v}_h)/\|\nabla f(\mathbf{v}_h)\|_2$ and t_h is the step-size. Thanks to the convexity, the convergence of this algorithm is guaranteed for any starting point $\mathbf{v}_0$. The step-size is updated at each iteration according to the following rule:

$$t_{h+1} = \begin{cases} 1.2 \times t_h, & \text{for } f(\mathbf{v}_{h+1}) < f(\mathbf{v}_h), \\ 0.5 \times t_h, & \text{otherwise.} \end{cases}$$

The multiplicative factors we use for adaptive step-size are those propose by Riedmiller and Braun [1992] for the Rprop algorithm. We stop iterating when the gradient magnitude measured in the ℓ_2-norm is below a certain level (10^{-5} in our experiments) or when the limit of 5000 iterations is attained.

For the choice of the tuning parameter κ, we did a cross-validation by choosing a geometric grid over the values of κ ranging from $1/p$ to $\sqrt{p}$. The results, for Models 2 and 4, are

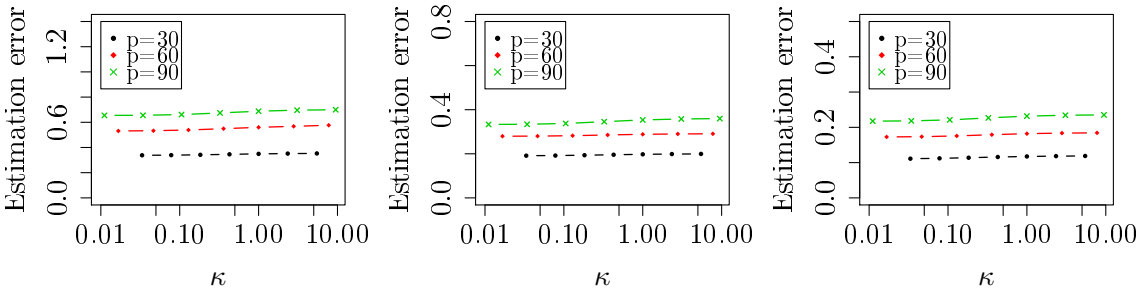


Figure 1.3: The estimation error of the PML as a function of κ. The plots are obtained for the synthetic experiment of Model 4 with various values of p and for $n = 200$ (left), $n = 800$ (middle) and $n = 2000$. Please note that the limits of the y-axis are not the same in the three plots and that the x-axis is presented in logarithmic scale.

plotted in Fig. 1.2 and 1.3, respectively. We can clearly see that there is a large interval of values of κ for which the error is nearly minimal. Based on this observation, we chose $\kappa = \frac{1}{3}\sqrt{\log p}$ for all the numerical experiments reported in Tables 1.1-1.6.

We also check the performance of the PML estimator with a tuning parameter $\kappa = 0.05$. The related results are reported in Tables B.1-B.6 and Tables B.7-B.12 of Appendix B. With such a small value for κ in comparison of the precedent one, the symmetry constraints on the precision matrix are less strong. Therefore, the resulting estimates are closer to those obtained with the RML estimator. As the true covariance matrix is symmetric, choosing a small value for κ is indeed a good strategy only when the off-diagonal entries are well estimated. In our experiments, with a smaller κ, the estimates accuracies are improved for the Models 1 and 2, when $\mathbf{B}^*$ is estimated by the square-root Lasso followed by OLS. For these models, the performance is often better for the PML estimator than for the RV estimator. In counterpart, the performance of the PML estimator is not as good with a smaller κ for the same models when $\mathbf{B}^*$ is known.

1.5 Conclusion

This chapter introduces three estimators of the diagonal entries of a sparse precision matrix when n i.i.d. copies of a Gaussian vector with this precision matrix are observed. The properties of these estimators are discussed and compared with those of the commonly used residual variance estimator. At a theoretical level, an interesting finding is that the naive maximum likelihood estimator (MLE) that does not take into account the symmetry constraints has a significantly larger risk than the residual variance estimator and, hence, is not optimal even asymptotically. The symmetry-enforced MLE and the penalized MLE circumvent this drawback and are shown in all numerical experiments to outperform the

residual variance estimator when the matrix $\mathbf{B}^*$ is known. Similar but unreported results are obtained when the estimators of the diagonal entries use a noisy matrix $\widehat{\mathbf{B}} = \mathbf{B}^* + \mathbf{\Xi}$, provided the noise matrix $\mathbf{\Xi}$ has i.i.d. Gaussian entries with zero mean and small variance. However, in a more realistic situation when $\mathbf{B}^*$ is estimated by the square-root Lasso or by the ordinary least squares conducted over the submodel selected by the square-root Lasso, the accuracies of the four estimators of the diagonal entries become comparable with a slight advantage for the residual variance estimator.

We would like also to mention the introduction of a novel and simple method of estimating partial correlations and of symmetrizing the precision matrix estimator derived from the nonsymmetric matrix $\widehat{\mathbf{B}}$. It is based on the observation that the square of the partial correlation between i-th and j-th variables is equal to $\mathbf{B}^*_{ij}\mathbf{B}^*_{ji}$.

In the future, it would be interesting to look for an estimator of $\mathbf{B}^*$ which is more accurate than the square-root Lasso and could hopefully—in combination with the symmetry-enforced MLE or the penalized MLE—lead to better precision matrix estimate than the one obtained by the association of the square-root Lasso and the residual variance estimator. Another appealing avenue for future research is the investigation of the case when the matrix $\mathbf{X}$ is observed with an error. Recent papers [Rosenbaum and Tsybakov, 2013; Belloni et al., 2014b] may provide valuable guidance for accomplishing this task. Next chapter addresses this last question.

p	30			60			90		
n	200	800	2000	200	800	2000	200	800	2000
$\mathbf{B}^*$ estimated by square-root Lasso									
RV	**0.372 (.066)**	**0.184 (.036)**	**0.113 (.023)**	**0.526 (.066)**	**0.269 (.035)**	**0.161 (.024)**	**0.655 (.076)**	**0.327 (.046)**	**0.206 (.025)**
RML	0.419 (.067)	0.212 (.033)	0.134 (.020)	0.583 (.065)	0.301 (.033)	0.183 (.023)	0.722 (.074)	0.361 (.045)	0.228 (.024)
SML	0.468 (.076)	0.228 (.033)	0.144 (.020)	0.664 (.079)	0.334 (.032)	0.201 (.024)	0.843 (.095)	0.405 (.042)	0.252 (.024)
PML	0.450 (.070)	0.224 (.032)	0.142 (.020)	0.622 (.069)	0.326 (.032)	0.198 (.024)	0.763 (.073)	0.394 (.042)	0.247 (.023)
$\mathbf{B}^*$ estimated by square-root Lasso followed by OLS									
RV	**0.368 (.065)**	**0.182 (.036)**	**0.113 (.023)**	**0.516 (.064)**	**0.267 (.035)**	**0.160 (.024)**	**0.641 (.075)**	**0.324 (.046)**	**0.205 (.025)**
RML	**0.368 (.065)**	**0.182 (.036)**	**0.113 (.023)**	**0.516) (.064)**	**0.267 (.035)**	**0.160 (.024)**	**0.641 (.075)**	**0.324 (.046)**	**0.205 (.025)**
SML	0.392 (.069)	0.191 (.037)	0.118 (.025)	0.558 (.078)	0.286 (.033)	0.173 (.025)	0.712 (.084)	0.351 (.043)	0.220 (.025)
PML	0.383 (.067)	0.188 (.037)	0.116 (.024)	0.539 (.067)	0.280 (.033)	0.169 (.024)	0.680 (.077)	0.343 (.043)	0.215 (.025)
$\mathbf{B}^*$ is estimated without error									
RV	0.366 (.066)	0.182 (.036)	0.113 (.023)	0.515 (.065)	0.267 (.035)	0.160 (.024)	0.640 (.074)	0.324 (.046)	0.204 (.025)
RML	0.374 (.065)	0.187 (.035)	0.116 (.023)	0.524 (.066)	0.271 (.035)	0.163 (.024)	0.649 (.073)	0.330 (.046)	0.208 (.025)
SML	**0.352 (.065)**	**0.173 (.039)**	**0.108 (.024)**	**0.500 (.066)**	**0.259 (.036)**	**0.156 (.025)**	**0.624 (.074)**	**0.316 (.046)**	**0.199 (.025)**
PML	0.353 (.065)	0.174 (.039)	0.109 (.024)	**0.500 (.066)**	**0.259 (.036)**	**0.156 (.025)**	0.625 (.074)	0.317 (.046)	0.200 (.025)

Table 1.4: Performance of the estimators of diagonal elements of the precision matrix in Model 4. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

p	30			60			90		
n	200	800	2000	200	800	2000	200	800	2000
$\mathbf{B}^*$ estimated by square-root Lasso									
RV	0.384 (.077)	**0.202 (.029)**	**0.125 (.023)**	0.543 (.060)	**0.279 (.039)**	**0.185 (.024)**	0.701 (.064)	**0.342 (.040)**	**0.222 (.021)**
RML	**0.380 (.076)**	0.206 (.027)	0.128 (.023)	**0.539 (.060)**	0.287 (.040)	0.190 (.025)	**0.697 (.064)**	0.352 (.041)	0.230 (.021)
SML	**0.380 (.076)**	0.205 (.029)	0.131 (.024)	**0.539 (.060)**	0.290 (.042)	0.194 (.024)	**0.697 (.064)**	0.353 (.041)	0.233 (.024)
PML	**0.380 (.076)**	0.206 (.027)	0.128 (.023)	**0.539 (.060)**	0.287 (.040)	0.190 (.025)	**0.697 (.064)**	0.352 (.041)	0.230 (.021)
$\mathbf{B}^*$ estimated by square-root Lasso followed by OLS									
RV	**0.379 (.076)**	**0.209 (.029)**	**0.130 (.025)**	**0.534 (.061)**	**0.295 (.040)**	**0.194 (.027)**	**0.693 (.064)**	**0.367 (.044)**	**0.235 (.025)**
RML	**0.379 (.076)**	**0.209 (.029)**	**0.130 (.025)**	**0.534 (.061)**	**0.295 (.040)**	**0.194 (.027)**	**0.693 (.064)**	**0.367 (.044)**	**0.235 (.025)**
SML	**0.379 (.076)**	**0.209 (.031)**	0.134 (.026)	**0.534 (.061)**	0.297 (.041)	0.199 (.027)	**0.693 (.064)**	0.368 (.043)	0.241 (.027)
PML	**0.379 (.076)**	**0.209 (.029)**	**0.130 (.025)**	**0.534 (.061)**	**0.295 (.040)**	**0.194 (.027)**	**0.693 (.063)**	**0.367 (.043)**	0.236 (.025)
$\mathbf{B}^*$ is estimated without error									
RV	0.384 (.075)	0.201 (.030)	0.125 (.022)	0.530 (.060)	0.275 (.038)	0.184 (.023)	0.686 (.066)	0.339 (.040)	0.221 (.022)
RML	0.383 (.076)	0.201 (.029)	0.126 (.022)	0.531 (.061)	0.277 (.037)	0.184 (.023)	0.687 (.066)	0.339 (.040)	0.221 (.022)
SML	**0.347 (.078)**	**0.180 (.032)**	**0.112 (.024)**	**0.498 (.061)**	**0.257 (.042)**	**0.170 (.025)**	**0.647 (.067)**	**0.319 (.042)**	**0.206 (.023)**
PML	0.383 (.076)	0.201 (.029)	0.126 (.022)	0.531 (.061)	0.277 (.037)	0.184 (.023)	0.687 (.066)	0.339 (.040)	0.221 (.022)

Table 1.5: Performance of the estimators of diagonal elements of the precision matrix in Model 5. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

p	30			60			90		
n	200	800	2000	200	800	2000	200	800	2000
$\mathbf{B}^*$ estimated by square-root Lasso									
RV	0.383 (.059)	**0.207 (.031)**	**0.140 (.018)**	0.534 (.054)	**0.310 (.031)**	**0.205 (.017)**	0.651 (.057)	**0.374 (.034)**	**0.255 (.018)**
RML	**0.378 (.058)**	0.223 (.030)	0.157 (.020)	**0.531 (.052)**	0.335 (.033)	0.236 (.020)	**0.648 (.055)**	0.408 (.036)	0.299 (.019)
SML	**0.378 (.058)**	0.229 (.030)	0.169 (.022)	**0.531 (.052)**	0.339 (.036)	0.249 (.021)	0.649 (.055)	0.410 (.036)	0.312 (.019)
PML	**0.378 (.058)**	0.223 (.030)	0.157 (.020)	**0.531 (.052)**	0.335 (.033)	0.236 (.020)	**0.648 (.055)**	0.408 (.036)	0.299 (.019)
$\mathbf{B}^*$ estimated by square-root Lasso followed by OLS									
RV	**0.383 (.058)**	**0.245 (.030)**	**0.170 (.019)**	**0.534 (.053)**	**0.373 (.030)**	**0.262 (.024)**	**0.649 (.053)**	**0.453 (.033)**	**0.341 (.025)**
RML	**0.383 (.058)**	**0.245 (.030)**	**0.170 (.019)**	**0.534 (.053)**	**0.373 (.030)**	**0.262 (.024)**	**0.649 (.053)**	**0.453 (.033)**	**0.341 (.025)**
SML	**0.383 (.058)**	0.251 (.027)	0.186 (.022)	**0.534 (.053)**	0.375 (.030)	0.281 (.023)	**0.649 (.053)**	0.454 (.033)	0.357 (.026)
PML	0.385 (.057)	**0.245 (.029)**	**0.170 (.019)**	**0.534 (.053)**	**0.373 (.030)**	**0.262 (.024)**	0.650 (.053)	**0.453 (.033)**	**0.341 (.025)**
$\mathbf{B}^*$ is estimated without error									
RV	0.408 (.068)	0.210 (.030)	0.141 (.018)	0.569 (.068)	0.309 (.031)	0.205 (.018)	0.697 (.063)	0.370 (.030)	0.251 (.018)
RML	0.411 (.070)	0.212 (.030)	0.142 (.019)	0.578 (.067)	0.313 (.033)	0.208 (.018)	0.702 (.064)	0.372 (.031)	0.254 (.018)
SML	**0.182 (.057)**	**0.097 (.023)**	**0.061 (.020)**	**0.277 (.064)**	**0.142 (.033)**	**0.094 (.022)**	**0.311 (.073)**	**0.178 (.030)**	**0.110 (.019)**
PML	0.411 (.070)	0.212 (.030)	0.142 (.019)	0.578 (.067)	0.313 (.033)	0.208 (.018)	0.702 (.064)	0.372 (.031)	0.254 (.018)

Table 1.6: Performance of the estimators of diagonal elements of the precision matrix in Model 6. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

	with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}}$			with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}+OLS}$			with $\mathbf{B}^*$		
p	30	60	90	30	60	90	30	60	90
RV	**1.701**	**2.703**	**3.449**	1.424	2.365	3.073	0.542	0.725	0.901
	(.149)	**(.173)**	**(.147)**	(.190)	(.213)	(.192)	(.079)	(.070)	(.084)
RML	2.027	3.082	3.840	1.424	2.365	3.073	0.667	0.904	1.109
	(.122)	(.136)	(.129)	(.190)	(.213)	(.192)	(.094)	(.092)	(.080)
SML	2.084	3.098	3.846	1.400	2.342	3.049	**0.086**	**0.096**	**0.115**
	(.127)	(.128)	(.123)	(.199)	(.221)	(.192)	**(.058)**	**(.052)**	**(.075)**
PML	2.028	3.080	3.839	**1.391**	**2.338**	**3.046**	0.172	0.203	0.253
	(.120)	(.133)	(.129)	**(.202)**	**(.222)**	**(.192)**	(.047)	(.042)	(.060)

Table 1.7: Performance of the estimators of diagonal elements of the precision matrix in Model 1 for $n = 50$. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

	with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}}$			with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}+OLS}$			with $\mathbf{B}^*$		
p	30	60	90	30	60	90	30	60	90
RV	**1.675**	**2.810**	**3.772**	**1.585**	**2.622**	**3.419**	0.399	0.504	0.588
	(.142)	**(.190)**	**(.203)**	**(.191)**	**(.232)**	**(.221)**	(.058)	(.057)	(.049)
RML	2.237	3.591	4.713	**1.585**	**2.622**	**3.419**	0.525	0.684	0.842
	(.140)	(.153)	(.189)	**(.191)**	**(.232)**	**(.221)**	(.084)	(.070)	(.081)
SML	2.534	3.844	4.910	1.937	2.914	3.668	**0.074**	**0.072**	**0.076**
	(.160)	(.140)	(.156)	(.212)	(.189)	(.220)	**(.054)**	**(.051)**	**(.054)**
PML	2.265	3.612	4.728	1.751	2.759	3.527	0.124	0.143	0.151
	(.136)	(.151)	(.183)	(.170)	(.190)	(.213)	(.049)	(.043)	(.038)

Table 1.8: Performance of the estimators of diagonal elements of the precision matrix in Model 2 for $n = 50$. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

	with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}}$			with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}+OLS}$			with $\mathbf{B}^*$		
p	30	60	90	30	60	90	30	60	90
RV	**0.596**	**0.880**	**1.141**	**1.044**	**1.617**	**2.009**	0.534	0.757	0.935
	(.077)	**(.073)**	**(.077)**	**(.164)**	**(.191)**	**(.179)**	(.064)	(.073)	(.070)
RML	0.925	1.241	1.494	**1.044**	**1.617**	**2.009**	0.658	0.911	1.137
	(.147)	(.110)	(.129)	**(.164)**	**(.191)**	**(.179)**	(.088)	(.089)	(.102)
SML	1.483	2.105	2.549	1.430	2.023	2.409	**0.099**	**0.104**	**0.103**
	(.214)	(.327)	(.419)	(.196)	(.203)	(.227)	**(.069)**	**(.075)**	**(.074)**
PML	0.926	1.246	1.499	1.165	1.754	2.121	0.659	0.911	1.139
	(.147)	(.111)	(.129)	(.174)	(.175)	(.178)	(.089)	(.091)	(.104)

Table 1.9: Performance of the estimators of diagonal elements of the precision matrix in Model 3 for $n = 50$. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

	with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}}$			with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}+OLS}$			with $\mathbf{B}^*$		
p	30	60	90	30	60	90	30	60	90
RV	**0.769**	**1.095**	**1.416**	**0.765**	**1.113**	**1.456**	0.725	1.042	1.349
	(.153)	**(.142)**	**(.131)**	**(.155)**	**(.146)**	**(.141)**	(.146)	(.142)	(.130)
RML	0.856	1.172	1.517	**0.765**	**1.113**	**1.456**	0.744	1.062	1.365
	(.148)	(.141)	(.135)	**(.155)**	**(.146)**	**(.141)**	(.143)	(.146)	(.128)
SML	0.931	1.281	1.600	0.809	1.195	1.513	**0.693**	**1.015**	**1.318**
	(.145)	(.143)	(.132)	(.165)	(.152)	(.149)	**(.152)**	**(.141)**	**(.131)**
PML	0.876	1.185	1.530	0.788	1.155	1.489	0.695	1.016	1.319
	(.149)	(.136)	(.136)	(.158)	(.146)	(.145)	(.152)	(.141)	(.130)

Table 1.10: Performance of the estimators of diagonal elements of the precision matrix in Model 4 for $n = 50$. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

	with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}}$			with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}+OLS}$			with $\mathbf{B}^*$		
p	30	60	90	30	60	90	30	60	90
RV	0.790	1.130	1.410	**0.786**	**1.139**	**1.423**	0.771	1.099	1.376
	(.130)	(.137)	(.141)	**(.137)**	**(.155)**	**(.148)**	(.134)	(.130)	(.137)
RML	**0.775**	**1.108**	**1.389**	**0.786**	**1.139**	**1.423**	0.772	1.098	1.378
	(.131)	**(.133)**	**(.139)**	**(.137)**	**(.155)**	**(.148)**	(.134)	(.130)	(.139)
SML	**0.775**	1.109	**1.389**	**0.786**	**1.139**	**1.423**	**0.702**	**1.036**	**1.313**
	(.131)	(.133)	**(.139)**	**(.137)**	**(.154)**	**(.148)**	**(.131)**	**(.131)**	**(.133)**
PML	**0.775**	**1.108**	**1.389**	0.787	**1.139**	**1.423**	0.772	1.098	1.378
	(.131)	**(.133)**	**(.139)**	(.138)	**(.155)**	**(.148)**	(.134)	(.130)	(.139)

Table 1.11: Performance of the estimators of diagonal elements of the precision matrix in Model 5 for $n = 50$. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

	with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}}$			with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}+OLS}$			with $\mathbf{B}^*$		
p	30	60	90	30	60	90	30	60	90
RV	0.751	1.036	1.238	**0.763**	**1.050**	**1.257**	0.802	1.109	1.341
	(.108)	(.121)	(.122)	**(.111)**	**(.133)**	**(.131)**	(.102)	(.126)	(.123)
RML	**0.734**	**1.012**	**1.218**	**0.763**	**1.050**	**1.257**	0.810	1.119	1.352
	(.107)	**(.120)**	**(.120)**	**(.111)**	**(.133)**	**(.131)**	(.107)	(.121)	(.121)
SML	**0.734**	**1.012**	**1.218**	**0.763**	**1.050**	**1.257**	**0.372**	**0.522**	**0.640**
	(.107)	**(.120)**	**(.120)**	**(.111)**	**(.133)**	**(.131)**	**(.132)**	**(.128)**	**(.109)**
PML	**0.734**	**1.012**	**1.218**	0.764	1.051	**1.257**	0.810	1.119	1.352
	(.107)	**(.120)**	**(.120)**	(.111)	(.133)	**(.131)**	(.107)	(.121)	(.121)

Table 1.12: Performance of the estimators of diagonal elements of the precision matrix in Model 6 for $n = 50$. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

Chapter 2

Robust estimation of a sparse precision matrix

The main part of this chapter is taken from the preprint *Convex programming approach to robust estimation of a multivariate Gaussian model*, which has been posted on the arXiv repository on December 2015. A version of this preprint is available at http://arxiv.org/abs/1512.04734.

RÉSUMÉ. La distribution gaussienne multivariée est couramment utilisée comme une première approximation de la distribution de données de grande dimension. L'estimation des paramètres de cette distribution sous diverses contraintes est un sujet abondamment étudié en statistique et cette problématique sert souvent de modèle de référence pour tester de nouveaux algorithmes ou de nouveaux cadres théoriques. Dans ce chapitre, nous développons une approche non-asymptotique du problème d'estimation des paramètres d'une distribution gaussienne multivariée lorsque les données sont corrompues par des observations prenant des valeurs extrêmes ou aberrantes (*outliers*). Nous proposons un nouvel estimateur – calculable efficacement en résolvant un problème d'optimisation convexe – qui estime de manière robuste la moyenne et la matrice de covariance de la population, même lorsque l'échantillon contient une proportion significative d'*outliers*. Lorsque l'ordre de grandeur de la dimension p des observations est plus petit que celui de la taille n de l'échantillon, on peut prouver que notre estimateur a une vitesse de convergence optimale à la fois pour la norme ℓ_1 élément par élément, pour la norme de Frobenius et la norme mixte ℓ_2/ℓ_1. De plus, cette optimalité est atteinte grâce à une méthode de type « racine carrée des moindres carrés pénalisés » faisant intervenir un paramètre d'ajustement universel servant à régler l'importance de la pénalisation. Ces résultats sont en partie étendus au cas où p est potentiellement plus grand que n, sous l'hypothèse supplémentaire que l'inverse de la matrice de covariance est creuse.

Contents

2.1 Introduction

In many applications where statistical methodology is employed, multivariate Gaussian distribution plays a central role as a first approximation to the distribution of high-dimensional data. It is mainly motivated by the fact that high dimensional data being sparsely distributed in space, can be reasonably well fitted by an elliptically countered distribution, of which the Gaussian distribution is the most famous representative. Another reason is that in high dimensional inference, sophisticated nonparametric methods suffer from the curse of dimensionality and lead to poor results (both in theory and in practice). For these reasons, recent years have witnessed an increased interest for simple parametric models in the statistical literature, with a particular emphasis on the effects of high dimensionality and the necessity to develop nonasymptotic theoretical guarantees. In this context, Gaussian models play a particular role in relation with the graphical modeling and discriminant analysis, but also because they provide a convenient theoretical framework for showcasing new ideas and testing new algorithms.

Determining the parameters of the Gaussian distribution under various constraints is a widely studied problem in statistics. Recent developments around sparse coding and compressed sensing have opened new lines of research on Gaussian models in which classical estimators such as the ordinary least squares and the empirical covariance matrix are strongly sub-optimal. Novel statistical procedures—often based on convex optimization—have emerged to cope with the aforementioned sub-optimality of traditional techniques. In addition, establishing nonasymptotic theoretical guarantees that highlight the impact of the dimensionality and the level of sparsity has appeared as a primary target of theoretical studies. The present work continues this line of research by developing a nonasymptotic approach to the problem of estimating the parameters of a multivariate Gaussian distribution from a sample of independent and identically distributed observations corrupted by outliers.

We propose an estimator—efficiently computable by solving a convex program—that robustly estimates the population mean and the population (inverse) covariance matrix even when the sample contains a significant proportion of outliers. The estimator is defined as the minimizer of a cost function that combines a data fidelity term with a sparsity-promoting penalization. Following and extending the methodology developed in [Belloni et al., 2011; Sun and Zhang, 2012], the data fidelity term is defined as the mixed ℓ_2/ℓ_1-norm of the residual matrix. The penalty term is proportional to the mixed ℓ_2/ℓ_1 norm of a matrix that models the outliers. Our estimator of the corruption matrix is proved to be rate optimal simultaneously for the entry-wise ℓ_1-norm, the Frobenius norm and the mixed ℓ_2/ℓ_1 norm. Furthermore, this optimality is achieved by a penalized square-root of least squares method with a universal tuning parameter calibrating the magnitude of the penalty.

The results are partly extended to the case where p is potentially larger than n, but the inverse covariance matrix is sparse. In such a situation, we recommend to add to the cost function an additional penalty term that corresponds, to some extent, to a weighted entry-wise ℓ_1 norm of the inverse covariance matrix. The theoretical guarantees established in this case are not as complete and satisfactory as those of low/moderate dimensional case. In particular, the obtained risk bounds are valid in the event that the empirical covariance matrix satisfies a particular type of restricted eigenvalues condition [Bickel et al., 2009]. At this stage, we are not able to theoretically assess the probability of this event. Another open problem is the practical choice of the tuning parameter. We are currently working on these issues and hope to address them in a forthcoming paper.

2.1.1 Mathematical framework

We adopt here the following formalization of the multivariate Gaussian model in presence of outliers. We assume that the outlier-free data $\mathbf{Y}$ consists of n row-vectors independently

drawn from a multivariate Gaussian distribution with mean $\boldsymbol{\mu}^*$ and covariance matrix $\boldsymbol{\Sigma}^*$, hereafter denoted by $\mathcal{N}_p(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$. However, the data $\mathbf{Y}$ is revealed to the Statistician after being corrupted by outliers. So, the Statistician has access to a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ satisfying

$$\mathbf{X} = \mathbf{Y} + \mathbf{E}^*. \tag{2.1}$$

The matrix of errors $\mathbf{E}^*$ has a special structure: most rows of $\mathbf{E}^*$—corresponding to inliers—have only zero entries. We will denote by O the subset of indices from $\{1, \dots, n\}$ corresponding to the outliers and by $I = \{1, \dots, n\} \setminus O$ the subset of inliers. The following two conditions will be assumed throughout the chapter:

(C1) The n rows of the matrix $\mathbf{Y}$ are independent $\mathcal{N}_p(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ random vectors.

(C2) The contamination matrix $\mathbf{E}^*$ is deterministic and, for every $i \in I \subset \{1, \dots, n\}$, the i-th row of $\mathbf{E}^*$ is zero. Furthermore, the rows of $\mathbf{E}^*(\boldsymbol{\Sigma}^*)^{-1/2}$ are bounded in Euclidean norm by $M_{\mathbf{E}}\sqrt{p}$, for some constant $M_{\mathbf{E}}$.

For an introduction to the problem of robust estimation in statistics, we refer the reader to [Hampel et al., 1986; Maronna et al., 2006; Huber and Ronchetti, 2009]. An overview of more recent advances relevant to the present work can be found in [Chen et al., 2015a; Loh and Tan, 2015].

2.1.2 Robust estimator by convex programming

In the situation under investigation in this work, it is assumed that the sample contains some outliers. In other terms, the relation $\mathbf{X}_{i,\bullet} \sim \mathcal{N}_p(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ holds true only for indices i belonging to some subset I of $[n]$. The set I is large, but does not necessarily coincide with the entire set $[n]$. In such a context, our proposal consist in extending the methodology developed in [Sun and Zhang, 2013]. Recall that in the case when no outlier is present in the sample, the square-root Lasso[1] [Sun and Zhang, 2013] estimates the matrix $\boldsymbol{\Omega}^* = (\boldsymbol{\Sigma}^*)^{-1}$ by first solving the optimization problem

$$\widehat{\mathbf{B}} = \operatorname*{arg\,min}_{\mathbf{B}:\mathbf{B}_{jj}=1} \min_{\mathbf{c}\in\mathbb{R}^p} \Big\{ \|(\mathbf{X}\mathbf{B} - \mathbf{1}_n\mathbf{c}^\top)^\top\|_{2,1} + \bar{\lambda}\|\mathbf{B}\|_{1,1} \Big\}, \tag{2.2}$$

for a given tuning parameter $\bar{\lambda} \geq 0$, where the min is over all $p \times p$ matrices $\mathbf{B}$ having all their diagonal entries equal to 1. The second step of the square-root Lasso procedure is to set

$$\widehat{\omega}_{jj} = \Big(\frac{1}{n} \|(\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top)\mathbf{X}\widehat{\mathbf{B}}_{\bullet,j}\|_2^2 \Big)^{-1}; \qquad \widehat{\boldsymbol{\Omega}} = \widehat{\mathbf{B}} \cdot \mathrm{diag}(\{\widehat{\omega}_{jj}\}_{j\in[p]}). \tag{2.3}$$

[1] Referred to as "scaled Lasso" in [Sun and Zhang, 2012, 2013], see footnote 4 of Chapter 1.

In the case of observations corrupted by outliers, we propose to modify the square-root Lasso procedure as follows. Let us denote by $\mathbf{u}_n$ the vector $\mathbf{1}_n/\sqrt{n}$ and by $\mathbf{X}^{(n)}$ the matrix $\mathbf{X}/\sqrt{n}$. This scaling is convenient since it makes the columns of the data matrix to be of a nearly constant Euclidean norm, at least in the case without outliers. We replace step (2.2) by

$$\{\widehat{\mathbf{B}}, \widehat{\mathbf{\Theta}}\} = \underset{\substack{\mathbf{B}:\mathbf{B}_{jj}=1 \\ \mathbf{\Theta}\in\mathbb{R}^{n\times p}}}{\arg\min} \min_{\mathbf{c}\in\mathbb{R}^p} \left\{ \big\|(\mathbf{X}^{(n)}\mathbf{B} - \mathbf{u}_n\mathbf{c}^\top - \mathbf{\Theta})^\top\big\|_{2,1} + \lambda\big(\|\mathbf{\Theta}\|_{2,1} + \gamma\|\mathbf{B}\|_{1,1}\big) \right\}, \tag{2.4}$$

where $\lambda \geq 0$ is a tuning parameter associated with the regularization term promoting robustness and where $\lambda\gamma \geq 0$ corresponds to the tuning parameter whose aim is to encourage sparsity of the matrix $\mathbf{B}$ (or, equivalently, of the corresponding graph). Using the estimators $\{\widehat{\mathbf{B}}, \widehat{\mathbf{\Theta}}\}$, the entries of the precision matrix $\mathbf{\Omega}^*$ are estimated by

$$\widehat{\omega}_{jj} = \frac{2n}{\pi}\|(\mathbf{I}_n - \mathbf{u}_n\mathbf{u}_n^\top)(\mathbf{X}^{(n)}\widehat{\mathbf{B}}_{\bullet,j} - \widehat{\mathbf{\Theta}}_{\bullet,j})\|_1^{-2}; \qquad \widehat{\mathbf{\Omega}} = \widehat{\mathbf{B}} \cdot \mathrm{diag}(\{\widehat{\omega}_{jj}\}_{j\in[p]}). \tag{2.5}$$

The matrix $\mathbf{E}^*$ and the vector $\boldsymbol{\mu}^*$ can be estimated by

$$\widehat{\mathbf{E}} = \sqrt{n}\,\widehat{\mathbf{\Theta}}\widehat{\mathbf{B}}^\dagger \qquad \text{and} \qquad \widehat{\boldsymbol{\mu}} = \frac{1}{n}(\mathbf{X} - \widehat{\mathbf{E}})^\top\mathbf{1}_n. \tag{2.6}$$

It is important to stress right away that the robust estimation procedure described by equations (2.4)-(2.6) can be efficiently realized in practice even for large dimensions p. Indeed, the first step boils down to solving a convex program, that can be cast into a second-order cone program, whereas the two last steps involve only simple operations with matrices and vectors.

To explain the rationale behind this estimator, let us recall the following well-known result concerning multivariate Gaussian distribution. If we denote $\mathbf{B}^* = \mathbf{\Omega}^*\mathrm{diag}(\mathbf{\Omega}^*)^{-1}$, then we have

$$\big(\mathbf{Y} - \mathbf{1}_n(\boldsymbol{\mu}^*)^\top\big)\mathbf{B}^*_{\bullet,j} = \phi_j^*\,\boldsymbol{\epsilon}_{\bullet,j},$$

where $\boldsymbol{\epsilon}_{\bullet,j} \sim \mathcal{N}_n(0, \mathbf{I}_n)$ is a random vector independent of $\mathbf{Y}_{\bullet,j^c}$ and $\phi_j^* = (\omega_{jj}^*)^{-1/2}$. Combining this relation with (2.1) and using the notations $\mathbf{\Theta}^* = \mathbf{E}^*\mathbf{B}^*/\sqrt{n} \in \mathbb{R}^{n\times p}$ and $\mathbf{c}^* = (\mathbf{B}^*)^\top\boldsymbol{\mu}^*$, we get

$$\mathbf{X}^{(n)}\mathbf{B}^*_{\bullet,j} = c_j^*\mathbf{u}_n + \mathbf{\Theta}^*_{\bullet,j} + \frac{\phi_j^*}{\sqrt{n}}\,\boldsymbol{\epsilon}_{\bullet,j}, \qquad \forall j \in [p]. \tag{2.7}$$

Furthermore, the matrix $\mathbf{\Theta}^*$ inherits the row-sparse structure of the matrix $\mathbf{E}$ whereas the matrix $\mathbf{B}^*$ has exactly the same sparsity pattern as the precision matrix $\mathbf{\Omega}^*$. This suggests to recover the triplet $(\mathbf{c}^*, \mathbf{B}^*, \mathbf{\Theta}^*)$ by minimizing a penalized loss where the penalty imposed

on $\mathbf{\Theta}$ promotes the row-sparsity, while the penalty imposed on $\mathbf{B}$ favors sparse matrices without any particular structure of the sparsity pattern. It is well known in the literature on group sparsity (see [Lounici et al., 2011] and the references therein) that the mixed ℓ_2/ℓ_1-norm penalty $\|\cdot\|_{2,1}$ is well suited for taking advantage of the row-sparsity while preserving the convexity of the penalty. A more standard application of the Lasso to our setting would suggest to use the residual sum of squares $\left\|\mathbf{X}^{(n)}\mathbf{B} - \mathbf{u}_n\mathbf{c}^\top - \mathbf{\Theta}\right\|_{2,2}^2$ as the data fidelity term, instead of the mixed ℓ_2/ℓ_1-norm written in (2.4). However, similarly the to the square-root Lasso [Belloni et al., 2011], and as shown in the results of the next sections, the latter has the advantage of making the tuning parameter λ scale free. It allows us to define a universal value of λ that does not depend on the noise levels ϕ_j^* in Eq. (2.7) and, nevertheless, leads to rate optimal risk bounds.

Note that during the past ten years several authors proposed to employ convex penalty based approaches to robust estimation in various settings, see for instance [Candès and Randall, 2008; Dalalyan and Chen, 2012; Dalalyan and Keriven, 2012; Nguyen and Tran, 2013]. The problems considered in these papers concern the estimation of a vector parameter and do not directly carry over the problem under investigation in the present work.

From the theoretical point of view, analyzing statistical properties of the estimators $\widehat{\mathbf{\Theta}}$, $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{\Omega}}$ turns out to be a challenging task. Indeed, despite the obvious similarity of problem (2.4) to its vector regression counterpart [Belloni et al., 2011; Sun and Zhang, 2012], optimization problem (2.4) contains an important difference: the objective function is not decomposable with respect to neither rows nor columns of the matrix $\mathbf{\Theta}$. In fact, the objective is the sum of two terms, the first being decomposable with respect to the columns of $\mathbf{\Theta}$ and non-decomposable with respect to the rows, while the second is decomposable with respect to the rows but non-decomposable with respect to the columns. As shown in the theorems stated below as well as in their proofs, we succeeded in overcoming this difficulty by means of nontrivial combinations of elementary arguments. We believe that some of the tricks used in the proofs may be useful in other problems where the objective function happens to be non-decomposable.

The rest of this chapter is organized as follows. Having already introduced the proposed method for robust estimation of a sparse precision matrix, we present our main theoretical findings in Section 2.2. A discussion on the advantages and limitations of the obtained results as compared to previous work on robust estimation, as well as extensions to high dimensional setting, are included in Section 2.3. Technical proofs are postponed to Section 2.4. Algorithmic aspects related to the implementation of our method are presented in Section 2.5 and some promising numerical results are reported in Section 2.6.

2.2 Moderate dimensional case: theoretical results

In order to ease notation and to avoid some technicalities that may blur the main ideas, we assume that $\boldsymbol{\mu}^* = 0$ which implies that $\mathbf{c} = 0$, see Eq. (2.7), and we do not need to minimize with respect to $\mathbf{c}$ in (2.10). We introduce the (unnormalized) residuals $\boldsymbol{\xi}_{\bullet,j} = \boldsymbol{\phi}_j^* \, \boldsymbol{\epsilon}_{\bullet,j}/\sqrt{n}$, so that the following relation holds:

$$\mathbf{X}^{(n)}\mathbf{B}^* = \boldsymbol{\Theta}^* + \boldsymbol{\xi}. \tag{2.8}$$

For a better understanding of the assumptions that are needed to establish a tight upper bound on the error of estimation of the matrix $\mathbf{B}^*$ of coefficients and the matrix $\boldsymbol{\Theta}^*$ corresponding to the outliers, we start by analyzing the problem of robust estimation with p is of smaller order than n, and no sparsity assumption on $\boldsymbol{\Omega}^*$ is made. We call this setting the moderate dimensional case, since we allow the dimension to go to infinity with the sample size, provided that the ratio p/n remains small[2]. In such a situation there is no longer need to penalize nonsparse matrices $\mathbf{B}$ in the optimization problem. We work with the estimator

$$\{\widehat{\mathbf{B}}, \widehat{\boldsymbol{\Theta}}\} = \underset{\substack{\mathbf{B}\in\mathbb{R}^{p\times p}\\ \mathbf{B}_{jj}=1}}{\arg\min} \min_{\boldsymbol{\Theta}\in\mathbb{R}^{n\times p}} \left\{ \left\| (\mathbf{X}^{(n)}\mathbf{B} - \boldsymbol{\Theta})^\top \right\|_{2,1} + \lambda \|\boldsymbol{\Theta}\|_{2,1} \right\}. \tag{2.9}$$

For a given matrix $\boldsymbol{\Theta}$, the minimum with respect to $\mathbf{B}$ in the foregoing optimization problem is a solution to the convex program

$$\widehat{\mathbf{B}}(\boldsymbol{\Theta}) = \underset{\substack{\mathbf{B}\in\mathbb{R}^{p\times p}\\ \mathbf{B}_{jj}=1}}{\arg\min} \left\{ \sum_{j=1}^{p} \left\| \mathbf{X}^{(n)}_{\bullet,j} - \boldsymbol{\Theta}_{\bullet,j} + \mathbf{X}^{(n)}_{\bullet,j^c}\mathbf{B}_{j^c,j} \right\|_2 \right\}, \tag{2.10}$$

which decomposes into p independent ordinary least squares problems. A solution of the latter is provided by the formula

$$\mathbf{X}^{(n)}_{\bullet,j^c}\widehat{\mathbf{B}}_{j^c,j}(\boldsymbol{\Theta}) = -\boldsymbol{\Pi}^{j^c}(\mathbf{X}^{(n)}_{\bullet,j} - \boldsymbol{\Theta}_{\bullet,j}) \quad \text{and} \quad \widehat{\mathbf{B}}_{jj}(\boldsymbol{\Theta}) = 1, \tag{2.11}$$

where the notation $\boldsymbol{\Pi}^{j^c}$ is used for the orthogonal projector in $\mathbb{R}^n$ onto the subspace spanned by the columns of $\mathbf{X}^{(n)}_{\bullet,j^c}$. Let us introduce now the matrices $\mathbf{Z}^j = \mathbf{I}_n - \boldsymbol{\Pi}^{j^c}$ that are orthogonal projectors onto the orthogonal complement of the linear subspace of $\mathbb{R}^n$ spanned by the columns of $\mathbf{X}_{\bullet,j^c}$ (or, equivalently, by $\mathbf{X}^{(n)}_{\bullet,j^c}$). Using this notation and

[2] This is different from the "low dimensional case" in which p is assumed fixed when n goes to infinity, so that the quantities depending only on p are treated as constants.

replacing expression (2.11) in problem (2.9), we arrive at

$$\widehat{\mathbf{\Theta}} = \underset{\mathbf{\Theta}\in\mathbb{R}^{n\times p}}{\arg\min} \left\{ \sum_{j=1}^{p} \left\| \mathbf{Z}^j (\mathbf{X}^{(n)}_{\bullet,j} - \mathbf{\Theta}_{\bullet,j}) \right\|_2 + \lambda \|\mathbf{\Theta}\|_{2,1} \right\}. \tag{2.12}$$

In what follows, we rely on formulae (2.12) and (2.11) both for computing and analyzing the estimator provided by Eq. (2.9). Our first result concerns the quality of estimating the outlier matrix $\mathbf{\Theta}^*$.

Theorem 2.2.1. *Let assumptions* **(C1)** *and* **(C2)** *be satisfied. Let* $\delta \in (0,1)$ *such that* $n \geq |O| + 8p + 16\log(4/\delta)$ *and choose*

$$\lambda = 6\left(\frac{p\log(2np/\delta)}{n}\right)^{1/2}. \tag{2.13}$$

If $40|O|p(13\log(2np/\delta) + 2(1+M_{\mathbf{E}})^2) \leq n - |O|$, *then with probability at least* $1-3\delta$,

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{1,1} \leq 3C_0 \max_j (\omega^*_{jj})^{-1/2} |O| p \left(\frac{\log(2np/\delta)}{n}\right)^{1/2}, \tag{2.14}$$

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{2,1} \leq 3C_0 \max_j (\omega^*_{jj})^{-1/2} |O| \left(\frac{p\log(2np/\delta)}{n}\right)^{1/2}, \tag{2.15}$$

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{2,2} \leq C_0 \max_j (\omega^*_{jj})^{-1/2} \left(\frac{|O|p\log(2np/\delta)}{n}\right)^{1/2}. \tag{2.16}$$

Here C_0 *is an universal constant smaller than* 4224.

Several comments are in order. First of all, let us stress that the obtained guarantees are nonasymptotic: it is not required that the sample size n or another quantity tend to infinity for this result to be true. To the best of our knowledge, this is the first[3] nonasymptotic result in robust estimation of a multivariate Gaussian model. Second, the value of the tuning parameter proposed by this result is scale free, that is it does not depend on the magnitude of the unknown parameters of the model. Third, one can show that the right-hand side expressions in Eq. (2.14)-(2.16) are minimax optimal up to logarithmic terms. Thus, the same estimator of $\mathbf{\Theta}^*$ is provably optimal for the three aforementioned norms. This remarkable property is due to the particular form of the penalty used in the estimation procedure.

Let us switch now to results describing statistical properties of the estimator $\widehat{\mathbf{\Omega}}$ of the precision matrix. Unfortunately, mathematical formulae we obtained as risk bounds for $\widehat{\mathbf{\Omega}}$ are not as compact and elegant as those of the last theorem. Therefore, to improve their legibility, we opted for presenting the results in a more asymptotic form. Namely,

[3] When this work was in preparation, the preprint [Loh and Tan, 2015] has been posted on arXiv that contains nonasymptotic results for another robust estimator of a multivariate Gaussian model. Detailed comparison of the results therein with the ours is provided below in the discussion on the previous work.

we replace the condition $40|O|p(13\log(2np/\delta)+2(1+M_{\mathbf{E}})^2)\leq n-|O|$ by the following one $|O|p\log n\leq c_0 n$, for some sufficiently small constant $c_0>0$, and we do not provide explicit constants.

Theorem 2.2.2. *Let assumptions* **(C1)** *and* **(C2)** *be satisfied and let* λ *be as in* (2.13). *Then there exists universal constants* $C, c_0>0$ *and* $n_0\in\mathbb{N}$ *such that for* $n\geq n_0$ *and* $|O|p\log n\leq c_0 n$, *the inequality*

$$\|\widehat{\mathbf{\Omega}}-\mathbf{\Omega}^*\|_{2,2}\leq C\frac{\sigma_{\max}(\mathbf{\Omega}^*)^2}{\sigma_{\min}(\mathbf{\Omega}^*)}\left\{M_{\mathbf{E}}\frac{|O|p\log n}{n}+\left(\frac{p^2\log n}{n}\right)^{1/2}\right\} \tag{2.17}$$

holds true with probability at least $1-5/n$.

This result tells us that in an asymptotic setting when all the three parameters n, p and $|O|$ are allowed to tend to infinity but so that $|O|p=o(n/\log n)$, the rate of convergence of the estimator $\widehat{\mathbf{\Omega}}$, measured in the Frobenius norm, is $p(\frac{|O|}{n}+\frac{1}{n^{1/2}})$. This rate contains two components, $p/n^{1/2}$ and $p|O|/n$, that have clear explanation. The rate $p/n^{1/2}$ comes from the fact that we are estimating p^2 entries of the matrix $\mathbf{\Omega}^*$ based on n observations. This term is unavoidable if no additional assumption (such as the sparsity) is made; it is the minimax rate of convergence in the outlier-free set-up. The second term, $p|O|/n$, originates from the fact that the outlier matrix has $p|O|$ nonzero entries which need to be somehow estimated for making it possible to estimate the model parameters. So, this term of the risk reflects the deterioration caused by the presence of outliers.

2.3 Discussion and extensions to high dimension

Our bounds versus those of always zero estimator Given that the matrix $\mathbf{\Theta}^*$ is defined as $\mathbf{E}^*$ divided by $\sqrt{n}$, one may wonder what is the advantage of our results as compared to the risk bound of the trivial estimator $\widehat{\mathbf{\Theta}}^0$ all the entries of which are 0. Clearly, the square of the error of this estimator measured in Frobenius norm is of the order $M_{\mathbf{E}}^2|O|p/n$. One may erroneously think that this bound is of the same order as the one we obtained above for the convex programming based estimator. In contrast with this, the risk bound of our estimator—although requires $M_{\mathbf{E}}^2|O|p/n$ to be bounded by some small constant—does not depend on $M_{\mathbf{E}}$. For instance, if $M_{\mathbf{E}}=\frac{1}{12}(\frac{n}{|O|p})^{1/2}$, the trivial estimator will have a constant risk whereas the estimator $\widehat{\mathbf{\Theta}}$ will be consistent and rate optimal provided that $|O|p\log(n+p)=o(n)$.

Another important advantage of our estimator—inherent to its definition and reflected in the obtained risk bounds—is that its squared error is proportional to the quantity $\max_{j\in p}(\phi_j^*)^2$, where $(\phi_j^*)^2$ represents the conditional variance of the j-th variable given all the others. In situations where the variables contain strong correlations, these conditional variances are significantly smaller than the marginal variances of the variables.

What happens if some outliers have very large norms ? The risk bound established for our estimator requires the constant $M_{\mathbf{E}}$, measuring the order of magnitude of the Euclidean norm of the outliers, to be not too large. This is not an artifact of our mathematical arguments, but an inherent limitation of our method. We did some experiments on simulated data that confirmed that when $M_{\mathbf{E}}$ is large, our estimator behaves poorly. However, we believe that this is not a serious limitation, since one can always pre-process the data by removing the observations that have atypically large Euclidean norm.

Lower bounds It is possible to establish lower bounds that show that the rates of convergence of the risk bounds that appear in Theorem 2.2.1 are optimal up to logarithmic factors. Indeed, one can show that there exists a constant $c > 0$ such that

$$\inf_{\bar{\mathbf{\Theta}}_n} \sup_{(\mathbf{\Omega}^*, \mathbf{\Theta}^*)} \mathbf{E}\big[\|\bar{\mathbf{\Theta}}_n - \mathbf{\Theta}^*\|_{q,q'}\big] \geq c\left(\frac{p^{2/q}|O|^{2/q'}}{n}\right)^{1/2}, \qquad (q, q') \in \{(1,1); (1,2); (2,2)\}, \tag{2.18}$$

where the inf is over all possible estimators $\bar{\mathbf{\Theta}}_n$ while the sup is over all matrices $\mathbf{\Omega}^*, \mathbf{\Theta}^*$ such that $\mathbf{E}^* = \sqrt{n}\mathbf{\Theta}^* \mathrm{diag}(\mathbf{\Omega}^*)(\mathbf{\Omega}^*)^{-1}$ satisfies condition **(C2)**. This lower bound can be proved by lower bounding the sup over all possible precision matrices by the corresponding expression for the identity precision matrix $\mathbf{\Omega}^* = \mathbf{I}_p$. In this case, $\mathbf{E}^* = \sqrt{n}\,\mathbf{\Theta}^*$ and we observe $\mathbf{X}^{(n)} = \mathbf{\Theta}^* + n^{-1/2}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a $n \times p$ matrix with i.i.d. standard Gaussian entries. If we further lower bound the sup over all $|O|$-(row)sparse matrices $\mathbf{\Theta}^*$ by the sup over matrices whose rows $|O|+1, \ldots, n$ vanish, we get a simple Gaussian mean estimation problem for the entries θ^*_{ij} with $i = 1, \ldots, |O|$ and $j = 1, \ldots, p$, under the condition $\max_{i,j} |\theta^*_{ij}| \leq n^{-1/2} M_{\mathbf{E}}$. It is well known that in this problem the individual entries θ^*_{ij} can not be estimated at a rate faster than $n^{-1/2}$. This yields the result for $q = q' = 1$. The corresponding upper bounds for $(q, q') = (2, 1)$ and $(q, q') = (2, 2)$ readily follow from that of $(q, q') = (1, 1)$ by a simple application of the Cauchy-Schwarz inequality. Furthermore, very recently, the cases $(q, q') = (2, 1)$ and $(q, q') = (2, 2)$ have been thoroughly studied by Klopp and Tsybakov [2015]. In particular, lower bounds including logarithmic terms have been established that prove that our estimator is minimax rate optimal when $p/|O|$ is of the order n^r for some $r \in (0, 1)$.

ϵ-contamination model and minimax sub-optimality The estimator proposed in this work can be applied in the context of ϵ-contamination model often used in statistics for quantifying the performance of robust estimators. It corresponds to assuming that each of n rows of the data matrix $\mathbf{X}$ is given by $\mathbf{X}_i = (1 - \epsilon_i)\mathbf{Y}_i + \epsilon_i \mathbf{E}_i$, where $\epsilon_i \in \{0, 1\}$ is a Bernoulli random variable with $\mathbf{P}(\epsilon_i = 1) = \epsilon$, $\mathbf{Y}_i \sim \mathcal{N}(\boldsymbol{\mu}^*, \mathbf{\Sigma}^*)$ is as before and $\mathbf{E}_i$ is randomly drawn from a distribution Q. The random variables ϵ_i, $\mathbf{Y}_i$ and $\mathbf{E}_i$ are independent and, perhaps the main difference with the model we considered above is that

all $\mathbf{E}_i$'s are drawn from the same distribution Q. One may wonder whether our procedure is minimax optimal in this ϵ-contamination model.

As proved in Theorems 3.1 and 3.2 of [Chen et al., 2015a], the minimax rate for estimating the covariance matrix $\mathbf{\Sigma}^*$ in the squared operator norm is $\frac{p}{n} + \epsilon^2$. In our notations, the role of ϵ is played by $|O|/n$. Therefore, the aforementioned result from [Chen et al., 2015a] suggests that one can estimate the precision matrix in the squared Frobenius norm with the rate $p(\frac{p}{n} + \epsilon^2) = p(\frac{p}{n} + \frac{|O|^2}{n^2})$, where the factor p comes from the fact that the square of the Frobenius norm is upper bounded by p-times the operator norm. Recall that the rate provided by the upper bound of Theorem 2.2.2 is $p(\frac{p}{n} + \frac{|O|^2 p}{n^2})$.

Therefore, the rate obtained by a direct application of Theorem 2.2.2 is sub-optimal in the minimax sense for the ϵ-contamination model (when both the dimension and the number of outliers tend to infinity with the sample size so that $|O|^2/n$ tends to infinity). However, under Huber contamination model, if we take mild assumptions on the distribution Q in addition to condition **(C2)**, the bounds stated in Lemma 2.4.9 can provably be tightened. Then, optimal rates for the estimation of $\mathbf{\Omega}^*$ can be obtained, up to logarithmic factors. It is still an open question whether the rate $p(\frac{p}{n} + \frac{|O|^2 p}{n^2})$ is minimax optimal over the set $\mathcal{M}(\underline{\tau}, \overline{\tau}, M_{\mathbf{E}})$ of matrices $(\mathbf{\Sigma}^*, \mathbf{E}^*)$ such that $\underline{\tau} \leq \sigma_{\min}(\mathbf{\Sigma}^*) \leq \sigma_{\max}(\mathbf{\Sigma}^*) \leq \overline{\tau}$ and $\mathbf{E}^*$ satisfies condition **(C2)**. Theorem 2.2.2 establishes that $p(\frac{p}{n} + \frac{|O|^2 p}{n^2})$ is an upper bound for the minimax rate, but the question of getting matching lower bound remains open.

Extensions to the case of large p In the case of large p, most ingredients of the proof used in moderate dimensional case remain valid after a suitable adaptation. Perhaps the most important difference is in the definition of the dimension-reduction cone. In order to present it, let $\mathcal{J} = \{J_j : j \in [p]\}$ be a collection of p subsets of $[p]$–supports of each row of the precision matrix–for which we use the notation $|\mathcal{J}| = \sum_{j=1}^p |J_j|$. By a slight abuse of notation, we will write $\mathcal{J}^c$ for the collection $\{J_j^c : j \in [p]\}$ and, for every $p \times p$ matrix $\mathbf{A}$, we define $\mathbf{A}_{\mathcal{J}}$ as the matrix obtained from $\mathbf{A}$ by zeroing all the elements $\mathbf{A}_{i,j}$ such that $i \notin J_j$. Let O be the subset of $[n]$ corresponding to the outliers. $\boldsymbol{\xi}_{\mathcal{O}}$ is obtained by zeroing all the rows $\boldsymbol{\xi}_{i,\bullet}$ such that $i \in O$. We define the dimension reduction cone

$$\mathscr{C}_{\mathcal{J},O}(c,\gamma) \triangleq \Big\{ \mathbf{\Delta} \in \mathbb{R}^{(p+n)\times p} : \gamma \|\mathbf{\Delta}^{\mathbf{B}}_{\mathcal{J}^c}\|_{1,1} + \|\mathbf{\Delta}^{\mathbf{\Theta}}_{O^c,\bullet}\|_{2,1} \leq c\big(\gamma \|\mathbf{\Delta}^{\mathbf{B}}_{\mathcal{J}}\|_{1,1} + \|\mathbf{\Delta}^{\mathbf{\Theta}}_{O,\bullet}\|_{2,1}\big) \Big\},$$

for $c > 1$ and $\gamma > 0$, where $\mathbf{\Delta}^{\mathbf{B}} = \mathbf{\Delta}_{1:p,\bullet}$ and $\mathbf{\Delta}^{\mathbf{\Theta}} = \mathbf{\Delta}_{(p+1):(p+n),\bullet}$. For a constant $\kappa > 0$, let us introduce the matrix $\mathbf{M} = [\mathbf{X}^{(n)}; -\mathbf{I}_n]$ and the event

$$\mathcal{E}_\kappa = \left\{ \|\mathbf{M}\mathbf{\Delta}\|_F^2 \geq \kappa \left(\frac{\|\mathbf{\Delta}^{\mathbf{B}}_{\mathcal{J}}\|_{1,1}^2}{|\mathcal{J}|} \right) \bigvee \left(\frac{\|\mathbf{\Delta}^{\mathbf{\Theta}}_{O,\bullet}\|_{2,1}^2}{|O|} \right) \quad \text{for all} \quad \mathbf{\Delta} \in \mathscr{C}_{\mathcal{J},O}(2,1) \right\}. \tag{2.19}$$

This event corresponds to the situations where the matrix $\mathbf{M}$ satisfies the (matrix) compatibility condition. To simplify the statement of the result, we assume that all the diagonal

entries of the covariance matrix $\mathbf{\Sigma}^*$ are equal to one. Note that this assumption can be approached by dividing the columns of $\mathbf{X}$ by the corresponding robust estimators of their standard deviation.

Theorem 2.3.1. *Let $\mathcal{J}$ and O be such that $\mathbf{B}^*_{\mathcal{J}^c} = 0$ and $\mathbf{\Theta}^*_{O^c,\bullet} = 0$. Choose $\gamma = 1$ and $\delta \in (0,1)$ such that $n \geq |O| + 16\log(2p/\delta)$ and choose*

$$\lambda = 6\left(\frac{\log(2np/\delta)}{n - |O|}\right)^{1/2}. \tag{2.20}$$

If $4\lambda(|\mathcal{J}|^{1/2} + |O|^{1/2}) < \kappa^{1/2}$ holds, then there exists an event $\mathcal{E}_0$ of probability at least $1 - 2\delta$ such that in $\mathcal{E}_\kappa \cap \mathcal{E}_0$, we have

$$\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_{1,1} + \|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* - \boldsymbol{\xi}_{\mathcal{O}}\|_{2,1} \leq \frac{C_1}{\kappa} \max_{j\in[p]}(\omega^*_{jj})^{-1/2}\big(|\mathcal{J}| + |O|\big)\left(\frac{\log(2np/\delta)}{n - |O|}\right)^{1/2} \tag{2.21}$$

with $C_1 \leq 900$.

The proof of this theorem follows the same scheme as the one of Theorem 2.2.1, it is given in Section 2.4.4. We will not comment this result too much because we find it incomplete at this stage. Indeed, the main conclusion of the theorem is formulated as a risk bound that holds in an event close to $\mathcal{E}_\kappa$. Unfortunately, we are not able now to provide a theoretical evaluation of $\mathbf{P}(\mathcal{E}_\kappa)$. We believe however that this probability is close to one, since the matrix $\mathbf{M}$ is composed of two matrices $\mathbf{X}^{(n)}$ and $-\mathbf{I}_n$ that have weakly correlated columns and each of these matrices satisfy the restricted eigenvalues condition. We hope that we will be able to make this rigorous in near future. Note also that this result tells us that one gets the optimal rate (up to logarithmic factors) of estimating $\mathbf{B}^*$ in ℓ_1-norm if the number of outliers is at most of the same order as the sparsity of the precision matrix.

Other related works In recent years, several methodological contributions have been made to the problem of robust estimation in multivariate Gaussian models under various kinds of contamination models. For instance, Wang and Lin [2014] have proposed a group-Lasso type strategy in the context of errors-in-variables with a pre-specified group structure on the set of covariates whereas Hirose and Fujisawa [2015] have introduced the method γ-Lasso, a robust sparse estimation procedure of the inverse covariance matrix based on the γ-divergence. Under cell-wise contamination model, Öllerer and Croux [2015] and Tarr et al. [2016] proposed to estimate the precision matrix by using either the graphical Lasso [d'Aspremont et al., 2008; Friedman et al., 2008] or the Clime estimator [Cai et al., 2011] in conjunction with a robust estimator of the covariance matrix. While [Tarr et al., 2016] have mainly focused on the methodological aspects, [Öllerer and Croux, 2015] carried out a breakdown analysis. Risk bounds on the statistical error of this procedure have been

established by Loh and Tan [2015]. They have shown that the element-wise squared error when estimating the precision matrix $\mathbf{\Omega}^*$ is of the order $\|\mathbf{\Omega}^*\|_{1,\infty}^2 \left(\frac{p}{n} + \frac{|O|^2}{n^2}\right)$. This result is particularly appealing for very sparse precision matrices having small $\ell_{1,\infty}$ norm. However, in moderate dimensional situations where the precision matrix is not necessarily sparse, the term $\|\mathbf{\Omega}^*\|_{1,\infty}^2$ is generally proportional to $p\sigma_{\max}(\mathbf{\Omega}^*)^2$ and the resulting upper bound is very likely to be sub-optimal. If we apply this result for assessing the quality of estimation in the squared Frobenius norm, we get an upper bound of the order $p^2\left(\frac{p}{n} + \frac{|O|^2}{n^2}\right)$, whereas our result provides an upper bound of the order $p\left(\frac{p}{n} + \frac{|O|^2 p}{n^2}\right)$. Furthermore, the results in [Loh and Tan, 2015] require the tuning parameter λ to be larger than an expression that involves the proportion of the outliers and the $\ell_{1,\infty}$ norm of the matrix $\mathbf{\Omega}^*$. This quantities are rarely available in practice and their estimation is often a hard problem. Finally, in the context of robust estimation of large matrices, let us also mention the recent work [Klopp et al., 2014], proposing a robust method of matrix completion and establishing sharp risk bounds on its statistical error.

2.4 Technical results and proofs

This section contains the proofs of all the mathematical claims of the chapter. The section is split into four parts. The first part contains the proof of Theorem 2.2.1, up to some technical lemmas characterizing the order of magnitude of the stochastic terms. The proof of Theorem 2.2.2 is presented in the second part, while the third part contains the aforementioned lemmas on the tail behavior of random quantities appearing in the proofs. The fourth and last part contains the proof of Theorem 2.3.1.

To ease notation, we define the projection matrix $\mathbf{Z} = \mathbf{I}_n - \mathbf{X}\big(\mathbf{X}^\top\mathbf{X}\big)^\dagger\mathbf{X}^\top$.

2.4.1 Risk bounds for outlier estimation

In this subsection, we provide a proof of Theorem 2.2.1, which contains perhaps the most original mathematical arguments of this work. Prior to diving into low-level technical arguments, let us provide a high-level overview of the proof. We can split it into four steps as follows:

Step 1: We check that if

$$\lambda \geq 3 \max_{i\in[n]} \left(\sum_{j\in[p]} \frac{(\mathbf{Z}^j_{i,\bullet}\boldsymbol{\epsilon}_{\bullet,j})^2}{\|\mathbf{Z}^j\boldsymbol{\epsilon}_{\bullet,j}\|_2^2} \right)^{1/2} \tag{2.22}$$

then the vector $\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$ belongs to the dimension-reduction cone

$$\|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{O^c,\bullet}\|_{2,1} \leq 2\|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{O,\bullet}\|_{2,1}. \tag{2.23}$$

Step 2: Using the Karush-Kuhn-Tucker conditions, we establish the bound

$$\|\mathbf{Z}\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,2}^2 \leq \frac{14\lambda}{3}\|\boldsymbol{\xi}^\top\|_{2,\infty}\|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,1} + \big(\lambda\|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,1}\big)^2 \tag{2.24}$$

for λ satisfying (2.22).

Step 3: Combining the two previous steps and using notation $\alpha := \|\mathbf{I}_n - \mathbf{Z}\|_{\infty,\infty}$, we obtain

$$\|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,2} \leq 140\lambda\|\boldsymbol{\xi}^\top\|_{2,\infty}|O|^{1/2} \qquad \text{and} \qquad \|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,1} \leq 520\lambda\|\boldsymbol{\xi}^\top\|_{2,\infty}|O|, \tag{2.25}$$

provided that $|O|(\lambda^2 + \alpha) < 1/10$.

Step 4: We conclude by establishing deterministic bounds on the random variables that appear in expressions (2.22) and (2.25), as well as on α.

The proofs of Steps 1 and 4 are, up to some additional technicalities, similar to those for the square-root Lasso. Steps 2 and 3 contain more original ingredients. The detailed proofs of all these steps are given below.

For every $c > 0$ and $O \subset [n]$, we define the cone

$$\mathscr{C}_O(c) \triangleq \Big\{\mathbf{\Delta} \in \mathbb{R}^{n\times p} : \|\mathbf{\Delta}^{\mathbf{\Theta}}_{O^c,\bullet}\|_{2,1} \leq c\|\mathbf{\Delta}^{\mathbf{\Theta}}_{O,\bullet}\|_{2,1}\Big\}.$$

Lemma 2.4.1. *If, for some constant $c > 1$, the penalty level λ satisfies the condition*

$$\lambda \geq \frac{c+1}{c-1}\max_{i\in[n]}\left(\sum_{j\in[p]}\frac{(\mathbf{Z}^j_{i,\bullet}\boldsymbol{\epsilon}_{\bullet,j})^2}{\|\mathbf{Z}^j\boldsymbol{\epsilon}_{\bullet,j}\|_2^2}\right)^{1/2}, \tag{2.26}$$

then the matrix $\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}$ belongs to the cone $\mathscr{C}_O(c)$.

Proof. The definition of $\widehat{\mathbf{\Theta}}$ by optimization problem (2.12) immediately leads to

$$\lambda\big(\|\widehat{\mathbf{\Theta}}\|_{2,1} - \|\mathbf{\Theta}^*\|_{2,1}\big) \leq \sum_{j\in[p]}\big(\|\mathbf{Z}^j(\mathbf{X}^{(n)}_{\bullet,j} - \mathbf{\Theta}^*_{\bullet,j})\|_2 - \|\mathbf{Z}^j(\mathbf{X}^{(n)}_{\bullet,j} - \widehat{\mathbf{\Theta}}_{\bullet,j})\|_2\big). \tag{2.27}$$

We use the inequality $\|a\|_2 - \|b\|_2 \leq (a-b)^\top a/\|a\|_2$ which ensues from the Cauchy-Schwarz inequality and is true for any pair of vectors (a, b), here with $a = \mathbf{Z}^j(\mathbf{X}^{(n)}_{\bullet,j} - \mathbf{\Theta}^*_{\bullet,j})$ and

$b = \mathbf{Z}^j(\mathbf{X}^{(n)}_{\bullet,j} - \widehat{\boldsymbol{\Theta}}_{\bullet,j})$. Clearly, we have $a - b = \mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}$ and $a = \mathbf{Z}^j \boldsymbol{\xi}_{\bullet,j}$. Hence, we obtain

$$\big\|\mathbf{Z}^j(\mathbf{X}^{(n)}_{\bullet,j} - \boldsymbol{\Theta}^*_{\bullet,j})\big\|_2 - \big\|\mathbf{Z}^j(\mathbf{X}^{(n)}_{\bullet,j} - \widehat{\boldsymbol{\Theta}}_{\bullet,j})\big\|_2 \leq (\mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j})^\top \frac{\mathbf{Z}^j \boldsymbol{\xi}_{\bullet,j}}{\big\|\mathbf{Z}^j \boldsymbol{\xi}_{\bullet,j}\big\|_2} = \sum_{i=1}^n \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{i,j} \frac{\mathbf{Z}^j_{i,\bullet} \boldsymbol{\xi}_{\bullet,j}}{\big\|\mathbf{Z}^j \boldsymbol{\xi}_{\bullet,j}\big\|_2}.$$

Then summing on $j \in [p]$ and applying the Cauchy-Schwarz inequality, we get

$$\sum_{j\in[p]} \|\mathbf{Z}^j(\mathbf{X}^{(n)}_{\bullet,j} - \boldsymbol{\Theta}^*_{\bullet,j})\|_2 - \|\mathbf{Z}^j(\mathbf{X}^{(n)}_{\bullet,j} - \widehat{\boldsymbol{\Theta}}_{\bullet,j})\|_2 \leq \sum_{i=1}^n \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{i,\bullet}\|_2 \Big(\sum_{j=1}^p \frac{(\mathbf{Z}^j_{i,\bullet} \boldsymbol{\xi}_{\bullet,j})^2}{\big\|\mathbf{Z}^j \boldsymbol{\xi}_{\bullet,j}\big\|_2^2} \Big)^{\frac{1}{2}}.$$

This inequality, in conjunction with Eq. (2.27) and the obvious inequality $\|\widehat{\boldsymbol{\Theta}}\|_{2,1} - \|\boldsymbol{\Theta}^*\|_{2,1} \geq \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{O^c,\bullet}\|_{2,1} - \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{O,\bullet}\|_{2,1}$ leads to

$$\begin{aligned}\lambda\big(\|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{O^c,\bullet}\|_{2,1} - \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{O,\bullet}\|_{2,1}\big) &\leq \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{2,1} \max_{i\in[n]} \Big(\sum_{j=1}^p \frac{(\mathbf{Z}^j_{i,\bullet} \boldsymbol{\xi}_{\bullet,j})^2}{\big\|\mathbf{Z}^j \boldsymbol{\xi}_{\bullet,j}\big\|_2^2} \Big)^{\frac{1}{2}} \\ &\leq \lambda \frac{c-1}{c+1} \Big(\|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{O,\bullet}\|_{2,1} + \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{O^c,\bullet}\|_{2,1} \Big),\end{aligned}$$

where the last line follows from condition (2.26). In conclusion, we get $\|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{O^c,\bullet}\|_{2,1} \leq c\|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{O,\bullet}\|_{2,1}$, which coincides with the claim of the lemma. □

The second step will be split into several lemmas, whereas the final conclusion is presented below in Lemma 2.4.6.

Lemma 2.4.2. *Let us introduce the vectors $\widehat{\boldsymbol{\xi}}_{\bullet,j} = \mathbf{Z}^j(\mathbf{X}^{(n)}_{\bullet,j} - \widehat{\boldsymbol{\Theta}}_{\bullet,j})$, $j \in [p]$. There exists a $n \times p$ matrix $\mathbf{V}$ such that*

$$\|\mathbf{V}_{i,\bullet}\|_2 \leq 1, \qquad \mathbf{V}^\top_{i,\bullet} \widehat{\boldsymbol{\Theta}}_{i,\bullet} = \|\widehat{\boldsymbol{\Theta}}_{i,\bullet}\|_2, \qquad \forall\, i \in [n], \tag{2.28}$$

and, for every $j \in [p]$, the following relation holds

$$\|\mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}\|_2^2 = \boldsymbol{\xi}^\top_{\bullet,j} \mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} - \lambda \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2 \mathbf{V}^\top_{\bullet,j} \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}. \tag{2.29}$$

Proof. Let us first consider the case $\widehat{\boldsymbol{\xi}}_{\bullet,j} \neq 0$. It is helpful to introduce the functions $g_1(\boldsymbol{\Theta}) = \sum_{j=1}^p \big\|\mathbf{Z}^j(\mathbf{X}^{(n)}_{\bullet,j} - \boldsymbol{\Theta}_{\bullet,j})\big\|_2$ and $g_2(\boldsymbol{\Theta}) = \sum_{i=1}^n \|\boldsymbol{\Theta}_{i,\bullet}\|_2$. The Karush-Kuhn-Tucker conditions imply that there exist two matrices $\mathbf{U}$ and $\mathbf{V}$ in $\mathbb{R}^{n\times p}$ satisfying $\mathbf{U} \in \partial_{\boldsymbol{\Theta}} g_1(\widehat{\boldsymbol{\Theta}})$, $\mathbf{V} \in \partial_{\boldsymbol{\Theta}} g_2(\widehat{\boldsymbol{\Theta}})$ and $\mathbf{U} + \lambda \mathbf{V} = 0$. For every $j \in [p]$, let $\mathbf{u}_j$ and $\mathbf{v}_j$ be the jth column of $\mathbf{U}$ and $\mathbf{V}$, respectively, so that $\mathbf{u}_j + \lambda \mathbf{v}_j = 0$ for every $j \in [p]$. With the assumption that $\|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2 > 0$, $\mathbf{u}_j$ is a differential and $\mathbf{u}_j = (\mathbf{Z}^{j\top} \mathbf{Z}^j \widehat{\boldsymbol{\Theta}}_{\bullet,j} - \mathbf{Z}^{j\top} \mathbf{X}^{(n)}_{\bullet,j})/\|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2$. Thus $\mathbf{Z}^{j\top} \mathbf{Z}^j = \mathbf{Z}^j$ leads to $\mathbf{u}_j = \mathbf{Z}^j(\widehat{\boldsymbol{\Theta}}_{\bullet,j} - \mathbf{X}^{(n)}_{\bullet,j})/\|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2$. Hence, we deduce that $\mathbf{Z}^j \widehat{\boldsymbol{\Theta}}_{\bullet,j} - \mathbf{Z}^j \mathbf{X}^{(n)}_{\bullet,j} + \lambda \mathbf{v}_j \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2 = 0$. Furthermore, as $\mathbf{X}^{(n)}_{\bullet,j} = -\mathbf{X}^{(n)}_{\bullet,j^c} \mathbf{B}^*_{j^c,j} + \boldsymbol{\Theta}^*_{\bullet,j} + \boldsymbol{\xi}_{\bullet,j}$ and $\mathbf{Z}^j$ is the projector onto the subspace

orthogonal to $\mathbf{X}^{(n)}_{\bullet,j^c}$, it follows that

$$\mathbf{Z}^j \mathbf{X}^{(n)}_{\bullet,j} = \mathbf{Z}^j \boldsymbol{\Theta}^*_{\bullet,j} + \mathbf{Z}^j \boldsymbol{\xi}_{\bullet,j}. \tag{2.30}$$

This yields $\mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} - \mathbf{Z}^j \boldsymbol{\xi}_{\bullet,j} + \lambda \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2 \mathbf{v}_j = 0$ where $\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*$. Finally, taking the scalar product of both sides with $\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}$, we get

$$(\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j})^\top \mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} - \boldsymbol{\xi}^\top_{\bullet,j} \mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} + \lambda \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2 \mathbf{v}_j^\top \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} = 0.$$

Since $\mathbf{v}_j = \mathbf{V}_{\bullet,j}$, this completes the proof of (2.29). To check relation (2.28), it suffices to remark that $\mathbf{V}_{i,\bullet}$ belongs to the sub-differential of the Euclidean norm $\|\boldsymbol{\Theta}_{i,\bullet}\|_2$ evaluated at $\widehat{\boldsymbol{\Theta}}$.

Let us now consider the case $\widehat{\boldsymbol{\xi}}_{\bullet,j} = 0$. This can be equivalently written as $\mathbf{Z}^j(\mathbf{X}^{(n)}_{\bullet,j} - \widehat{\boldsymbol{\Theta}}_{\bullet,j}) = 0$. In view of Eq. (2.30), we get $\mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} = \mathbf{Z}^j \boldsymbol{\xi}_{\bullet,j}$. Taking the scalar product of both sides with $\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}$ and using the fact that $\mathbf{Z}^j$ is idempotent, we get relation (2.29). □

Lemma 2.4.3. *Let R, A, B be arbitrary real numbers satisfying the inequality $R^2 \leq A + BR$. Then, the inequality $R^2 \leq 2A + B^2$ holds true.*

Proof. The inequality $R^2 \leq A + BR$ is equivalent to $(2R - B)^2 \leq 4A + B^2$. This entails that $|2R - B| \leq \sqrt{4A + B^2}$ and, therefore, $2R \leq B + \sqrt{4A + B^2}$. We get the desired result by taking the square of both sides and using the inequality $(a + b)^2 \leq 2a^2 + 2b^2$. □

Lemma 2.4.4. *Equation (2.29) implies that*

$$\|\mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}\|_2^2 \leq 2\boldsymbol{\xi}^\top_{\bullet,j} \mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} - 2\lambda \|\mathbf{Z}^j \boldsymbol{\xi}_{\bullet,j}\|_2 \mathbf{V}^\top_{\bullet,j} \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} + (\lambda \mathbf{V}^\top_{\bullet,j} \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j})^2.$$

Proof. According to Eq. (2.30), we have $\mathbf{Z}^j(\mathbf{X}^{(n)}_{\bullet,j} - \boldsymbol{\Theta}^*_{\bullet,j}) = \mathbf{Z}^j \boldsymbol{\xi}_{\bullet,j}$. Therefore, from the definition of the estimated residuals $\widehat{\boldsymbol{\xi}}_{\bullet,j}$ we infer that $\mathbf{Z}^j \boldsymbol{\xi}_{\bullet,j} - \widehat{\boldsymbol{\xi}}_{\bullet,j} = \mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}$, which implies the inequality

$$\big| \|\mathbf{Z}^j \boldsymbol{\xi}_{\bullet,j}\|_2 - \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2 \big| \leq \|\mathbf{Z}^j \boldsymbol{\xi}_{\bullet,j} - \widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2 = \|\mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}\|_2.$$

Combining this bound with equation (2.29) of Lemma 2.4.2, we obtain

$$\begin{aligned} \|\mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}\|_2^2 &= \boldsymbol{\xi}^\top_{\bullet,j} \mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} - \lambda \|\mathbf{Z}^j \boldsymbol{\xi}_{\bullet,j}\|_2 \mathbf{V}^\top_{\bullet,j} \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} + \lambda (\|\mathbf{Z}^j \boldsymbol{\xi}_{\bullet,j}\|_2 - \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2) \mathbf{V}^\top_{\bullet,j} \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} \\ &\leq \boldsymbol{\xi}^\top_{\bullet,j} \mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} - \lambda \|\mathbf{Z}^j \boldsymbol{\xi}_{\bullet,j}\|_2 \mathbf{V}^\top_{\bullet,j} \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} + \lambda |\mathbf{V}^\top_{\bullet,j} \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}| \cdot \|\mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}\|_2. \end{aligned}$$

We conclude using Lemma 2.4.3 with $R = \|\mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}\|_2$. □

Lemma 2.4.5. *Assuming that* $\lambda \geq \frac{c+1}{c-1} \max_{i\in[n]} \Big(\sum_{j\in[p]} \frac{(\mathbf{Z}^j_{i,\bullet}\boldsymbol{\epsilon}_{\bullet,j})^2}{\|\mathbf{Z}^j\boldsymbol{\epsilon}_{\bullet,j}\|_2^2} \Big)^{1/2}$, *it holds*

$$\sum_{j=1}^{p} \boldsymbol{\xi}_{\bullet,j}^\top \mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} \leq \lambda \frac{c-1}{c+1} \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{2,1} \max_{j\in[p]} \|\boldsymbol{\xi}_{\bullet,j}\|_2.$$

Proof. We have

$$\sum_{j=1}^{p} \boldsymbol{\xi}_{\bullet,j}^\top \mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} = \sum_{i=1}^{n}\sum_{j=1}^{p} (\mathbf{Z}^j\boldsymbol{\xi}_{\bullet,j})_i \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{i,j} \leq \max_{j\in[p]} \|\mathbf{Z}^j\boldsymbol{\xi}_{\bullet,j}\|_2 \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{|(\mathbf{Z}^j\boldsymbol{\xi}_{\bullet,j})_i|}{\|\mathbf{Z}^j\boldsymbol{\xi}_{\bullet,j}\|_2} |\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{i,j}|.$$

Thus by the Cauchy-Schwarz inequality and the assumption of the lemma,

$$\begin{aligned}\sum_{j=1}^{p} \boldsymbol{\xi}_{\bullet,j}^\top \mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} &\leq \max_{j\in[p]} \|\mathbf{Z}^j\boldsymbol{\xi}_{\bullet,j}\|_2 \sum_{i=1}^{n} \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{i,\bullet}\|_2 \Big(\sum_{j\in[p]} \frac{(\mathbf{Z}^j_{i,\bullet}\boldsymbol{\xi}_{\bullet,j})^2}{\|\mathbf{Z}^j\boldsymbol{\xi}_{\bullet,j}\|_2^2} \Big)^{1/2} \\ &\leq \lambda \frac{c-1}{c+1} \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{2,1} \max_{j\in[p]} \|\mathbf{Z}^j\boldsymbol{\xi}_{\bullet,j}\|_2.\end{aligned}$$

Moreover, as the operator norm associated with the Euclidean norm is the spectral norm, it holds that $\|\mathbf{Z}^j\boldsymbol{\xi}_{\bullet,j}\|_2 \leq \|\mathbf{Z}^j\|_2 \|\boldsymbol{\xi}_{\bullet,j}\|_2$. Then, as $\mathbf{Z}^j$ is a projection matrix, $\|\mathbf{Z}^j\|_2 = 1$ and $\|\mathbf{Z}^j\boldsymbol{\xi}_{\bullet,j}\|_2 \leq \|\boldsymbol{\xi}_{\bullet,j}\|_2$. The claimed result follows. □

Lemma 2.4.6. *If conditions* (2.26) *and* (2.29) *hold, then*

$$\sum_{j=1}^{p} \|\mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}\|_2^2 \leq 2\lambda \|\boldsymbol{\xi}^\top\|_{2,\infty} \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{2,1} \Big(\frac{c-1}{c+1} + 2 \Big) + \Big(\lambda \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{2,1} \Big)^2. \tag{2.31}$$

Proof. We first note that $\|\mathbf{V}_{i,\bullet}\|_2 \leq 1$ yields

$$\sum_{j=1}^{p} |\mathbf{V}_{\bullet,j}^\top \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}| \leq \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{2,1} \quad \text{and} \quad \sum_{j=1}^{p} (\lambda \mathbf{V}_{\bullet,j}^\top \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j})^2 \leq (\lambda \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{2,1})^2.$$

Thus, using relation (2.29) and Lemma 2.4.4, we arrive at

$$\begin{aligned}\sum_{j=1}^{p} \|\mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}\|_2^2 &\leq 2\sum_{j=1}^{p} \boldsymbol{\xi}_{\bullet,j}^\top \mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} + 2\lambda \sum_{j=1}^{p} \|\mathbf{Z}^j\boldsymbol{\xi}_{\bullet,j}\|_2 |\mathbf{V}_{\bullet,j}^\top \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}| + \sum_{j=1}^{p} (\lambda \mathbf{V}_{\bullet,j}^\top \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j})^2 \\ &\leq 2\sum_{j=1}^{p} \boldsymbol{\xi}_{\bullet,j}^\top \mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} + 2\lambda \max_{j\in[p]} \|\mathbf{Z}^j\boldsymbol{\xi}_{\bullet,j}\|_2 \sum_{j=1}^{p} |\mathbf{V}_{\bullet,j}^\top \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}| + (\lambda \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{2,1})^2 \\ &\leq 2\sum_{j=1}^{p} \boldsymbol{\xi}_{\bullet,j}^\top \mathbf{Z}^j \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} + 2\lambda \max_{j\in[p]} \|\boldsymbol{\xi}_{\bullet,j}\|_2 \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{2,1} + (\lambda \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{2,1})^2.\end{aligned}$$

The combination of the latter with Lemma 2.4.5 implies inequality (2.31). □

Note that $\mathbf{Z}$ and $\mathbf{Z}^j$ are two orthogonal projection matrices on nested subspaces of dimensions $n-p$ and $n-p+1$, respectively. Hence, for any $j \in [p]$, $\|\mathbf{Z}\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}\|_2 \leq \|\mathbf{Z}^j\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j}\|_2$. Using this inequality to lower bound the left-hand side of Eq. (2.31) and choosing $c=2$, we get inequality (2.24) of Step 2. We are now in a position to carry out Step 3.

Proposition 2.4.7. *If the penalty level λ satisfies the condition (2.22) and $|O|(\lambda^2+\alpha) < 1/10$, then*

$$\|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{2,2} \leq 140\lambda\|\boldsymbol{\xi}^\top\|_{2,\infty}|O|^{1/2} \qquad \text{and} \qquad p^{-1/2}\|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{1,1} \leq \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{2,1} \leq 520\lambda\|\boldsymbol{\xi}^\top\|_{2,\infty}|O|, \tag{2.32}$$

where $\alpha := \|\mathbf{I}_n - \mathbf{Z}\|_{\infty,\infty}$.

Proof. In the few lines that follow, we write $\mathbf{X}$ instead of $\mathbf{X}^{(n)}$ and $\widehat{\boldsymbol{\Delta}}$ instead of $\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}$. Simple algebra yields

$$\|(\mathbf{I}_n - \mathbf{Z})\widehat{\boldsymbol{\Delta}}\|_{2,2}^2 = \operatorname{trace}\big((\mathbf{I}_n - \mathbf{Z})\widehat{\boldsymbol{\Delta}}((\mathbf{I}_n - \mathbf{Z})\widehat{\boldsymbol{\Delta}})^\top\big).$$

Using the facts that $\operatorname{trace}(\mathbf{AB}) = \operatorname{trace}(\mathbf{BA})$ (whenever the matrix products are well defined), $\operatorname{trace}(\mathbf{AB}) \leq \|\mathbf{A}\|_{\infty,\infty}\|\mathbf{B}\|_{1,1}$ and $\|\mathbf{A}\mathbf{A}^\top\|_{q,q} \leq \|\mathbf{A}\|_{2,q}^2$, for any $q \in [1,\infty]$, (the last one is a simple consequence of the Cauchy-Schwarz inequality) we get

$$\|(\mathbf{I}_n - \mathbf{Z})\widehat{\boldsymbol{\Delta}}\|_{2,2}^2 = \operatorname{trace}\big((\mathbf{I}_n - \mathbf{Z})\widehat{\boldsymbol{\Delta}}\widehat{\boldsymbol{\Delta}}^\top\big) \leq \|\mathbf{I}_n - \mathbf{Z}\|_{\infty,\infty} \cdot \|\widehat{\boldsymbol{\Delta}}\widehat{\boldsymbol{\Delta}}^\top\|_{1,1} \leq \|\mathbf{I}_n - \mathbf{Z}\|_{\infty,\infty} \cdot \|\widehat{\boldsymbol{\Delta}}\|_{2,1}^2.$$

Adding the last inequality to Eq. (2.24) of Step 2 and using the Pythagorean theorem, we get

$$\|\widehat{\boldsymbol{\Delta}}\|_{2,2}^2 \leq \frac{14\lambda}{3}\|\boldsymbol{\xi}^\top\|_{2,\infty}\|\widehat{\boldsymbol{\Delta}}\|_{2,1} + (\lambda^2+\alpha)\|\widehat{\boldsymbol{\Delta}}\|_{2,1}^2. \tag{2.33}$$

Since according to Step 1 we have $\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}} \in \mathscr{C}_O(c)$, we infer that

$$\|\widehat{\boldsymbol{\Delta}}\|_{2,2}^2 \leq 14\lambda\|\boldsymbol{\xi}^\top\|_{2,\infty}\|\widehat{\boldsymbol{\Delta}}_{O,\bullet}\|_{2,1} + 9(\lambda^2+\alpha)\|\widehat{\boldsymbol{\Delta}}_{O,\bullet}\|_{2,1}^2.$$

Finally, using the Cauchy-Schwarz inequality, we have $\|\widehat{\boldsymbol{\Delta}}_{O,\bullet}\|_{2,1}^2 \leq |O| \cdot \|\widehat{\boldsymbol{\Delta}}_{O,\bullet}\|_{2,2}^2$, which leads to

$$\|\widehat{\boldsymbol{\Delta}}\|_{2,2}^2 \leq 14\lambda\|\boldsymbol{\xi}^\top\|_{2,\infty}|O|^{1/2}\|\widehat{\boldsymbol{\Delta}}_{O,\bullet}\|_{2,2} + 9|O|(\lambda^2+\alpha)\|\widehat{\boldsymbol{\Delta}}_{O,\bullet}\|_{2,2}^2.$$

Since the last norm in the right-hand side is bounded from above by $\|\widehat{\boldsymbol{\Delta}}\|_{2,2}$, we get

$$\|\widehat{\boldsymbol{\Delta}}\|_{2,2}^2 \leq 14\lambda\|\boldsymbol{\xi}^\top\|_{2,\infty}|O|^{1/2}\|\widehat{\boldsymbol{\Delta}}\|_{2,2} + 9|O|(\lambda^2+\alpha)\|\widehat{\boldsymbol{\Delta}}\|_{2,2}^2.$$

This implies that either $\|\widehat{\mathbf{\Delta}}\|_{2,2} = 0$ or

$$\|\widehat{\mathbf{\Delta}}\|_{2,2} \leq \frac{14\lambda\|\boldsymbol{\xi}^\top\|_{2,\infty}|O|^{1/2}}{1 - 9|O|(\lambda^2 + \alpha)}, \tag{2.34}$$

provided that the denominator of the last expression is positive. Note that under the same condition, one can bound the norm $\|\widehat{\mathbf{\Delta}}\|_{2,1}$ as follows:

$$\|\widehat{\mathbf{\Delta}}\|_{2,1} \leq 3\|\widehat{\mathbf{\Delta}}_{O,\bullet}\|_{2,1} \leq 3|O|^{1/2}\|\widehat{\mathbf{\Delta}}_{O,\bullet}\|_{2,2} \leq 3|O|^{1/2}\|\widehat{\mathbf{\Delta}}\|_{2,2} \leq \frac{52\lambda\|\boldsymbol{\xi}^\top\|_{2,\infty}|O|}{1 - 9|O|(\lambda^2 + \alpha)}. \tag{2.35}$$

This completes the proof. □

The details of Step 4 are postponed to Subsection 2.4.3. Let us just stress here that if for a $\delta \in (0,1)$ we define the event $\mathcal{E}$ as the one in which the following inequalities are satisfied:

$$\begin{aligned}
\max_{i\in[n],j\in[p]} |\mathbf{Z}^j_{i,\bullet}\boldsymbol{\epsilon}_{\bullet,j}| &\leq \sqrt{2\log(2np/\delta)} \\
\min_{j\in[p]} \|\mathbf{Z}^j\boldsymbol{\epsilon}_{\bullet,j}\|_2^2 &\geq n - p + 1 - 2\sqrt{(n-p+1)\log(2p/\delta)} \geq n/2 \\
\sigma_{\min}(\mathbf{X}(\mathbf{\Omega}^*)^{1/2}) &\geq \sqrt{(n-|O|)/4} \\
\|\mathbf{I}_n - \mathbf{Z}\|_{\infty,\infty} &\leq \frac{8(1+M_{\mathbf{E}})^2 p + 16\log(2n/\delta)}{n - |O|} \\
\|\boldsymbol{\epsilon}^\top\|_{2,\infty} &\leq \sqrt{n} + \sqrt{2\log(p/\delta)} \leq \sqrt{n}\,(1 + 2^{-3/2}).
\end{aligned}$$

According to Eq. (2.42), Lemma 2.4.12 and Lemma 2.4.13 below, as well as the union bound, we have $\mathbf{P}(\mathcal{E}) \geq 1 - 3\delta$. Furthermore, combining the above upper bound on $\alpha = \|\mathbf{I}_n - \mathbf{Z}\|_{\infty,\infty}$ with the condition of the theorem, we get that $|O|(\lambda^2 + \alpha) \leq 1/10$ in $\mathcal{E}$. Thus, Proposition 2.4.7 implies the claim of Theorem 2.2.1.

2.4.2 Bounds on estimation error of the precision matrix

Let us denote by $\widehat{\mathbf{D}}$ and $\mathbf{D}^*$ the $p \times p$ diagonal matrices with $\widehat{\mathbf{D}}_{jj} = \widehat{\omega}_{jj}$ and $\mathbf{D}^*_{jj} = \omega^*_{jj}$, respectively. We know that $\widehat{\mathbf{\Omega}} = \widehat{\mathbf{B}}\widehat{\mathbf{D}}$ and $\mathbf{\Omega}^* = \mathbf{B}^*\mathbf{D}^*$. Hence, an upper bound on the error of estimation of $\mathbf{\Omega}^*$ can be readily inferred from bounds on the estimation error of $\mathbf{B}^*$ and $\mathbf{D}^*$. Indeed,

$$\begin{aligned}
\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{2,2} &\leq \|(\widehat{\mathbf{B}} - \mathbf{B}^*)\widehat{\mathbf{D}}\|_{2,2} + \|\mathbf{B}^*(\widehat{\mathbf{D}} - \mathbf{D}^*)\|_{2,2} \\
&\leq \|\widehat{\mathbf{B}} - \mathbf{B}^*\|_{2,2} \max_j \widehat{\omega}_{jj} + \sigma_{\max}(\mathbf{\Omega}^*)\|\widehat{\mathbf{D}}(\mathbf{D}^*)^{-1} - \mathbf{I}_p\|_{2,2}.
\end{aligned} \tag{2.36}$$

To formulate the corresponding result, let us define the condition number $\rho^* \geq 1$ by $(\rho^*)^2 = \sigma_{\max}(\mathbf{\Omega}^*)/\sigma_{\min}(\mathbf{\Omega}^*)$. Throughout this proof, we use C as a generic notation for a

universal constant, whose value may change at each appearance.

Lemma 2.4.8. *It holds that*

$$\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_{2,2} \le \sigma_{\min}(\mathbf{X}^{(n)})^{-1}\big(\alpha^{1/2}\|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,1} + p^{1/2}\max_j \|(\mathbf{I}_n - \mathbf{Z}^j)\boldsymbol{\xi}_{\bullet,j}\|_2\big). \tag{2.37}$$

In addition, if $(|O|p) = o(n/\log n)$*, there exists an absolute constant* $C > 0$ *such that for sufficiently large values of* n *the inequality*

$$\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_{2,2} \le C\rho^*\left\{M_{\mathbf{E}}\frac{|O|p\log n}{n} + \left(\frac{p^2\log n}{n}\right)^{1/2}\right\} \tag{2.38}$$

holds with probability larger than $1 - (5/n)$.

Proof. To ease notation, throughout this proof we write $\mathbf{\Omega}$ and ω_{jj} instead of $\mathbf{\Omega}^*$ and ω^*_{jj}, respectively. One can check that $\mathbf{X}^{(n)}(\widehat{\mathbf{B}}_{\bullet,j} - \mathbf{B}^*_{\bullet,j}) = \mathbf{X}^{(n)}_{\bullet,j^c}(\widehat{\mathbf{B}}_{j^c,j} - \mathbf{B}^*_{j^c,j}) = (\mathbf{I}_n - \mathbf{Z}^j)(\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{\bullet,j} - \boldsymbol{\xi}_{\bullet,j})$ for every $j \in [p]$. Therefore, by the triangle inequality, we get $\|\mathbf{X}^{(n)}(\widehat{\mathbf{B}} - \mathbf{B}^*)\|_{2,2} \le \|(\mathbf{I}_n - \mathbf{Z})\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,2} + p^{1/2}\max_j \|(\mathbf{I}_n - \mathbf{Z}^j)\boldsymbol{\xi}_{\bullet,j}\|_2$. We have already used in the previous section the inequality $\|(\mathbf{I}_n - \mathbf{Z})\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,2} \le \alpha^{1/2}\|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,1}$. This yields

$$\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_{2,2} \le \sigma_{\min}(\mathbf{X}^{(n)})^{-1}\big(\alpha^{1/2}\|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,1} + p^{1/2}\max_j \|(\mathbf{I}_n - \mathbf{Z}^j)\boldsymbol{\xi}_{\bullet,j}\|_2\big).$$

Combining inequality $\sigma_{\min}(\mathbf{X}^{(n)}) \ge \sigma_{\min}(\mathbf{X}^{(n)}\mathbf{\Omega}^{1/2})\sigma_{\min}(\mathbf{\Omega}^{-1/2}) = \sigma_{\min}(\mathbf{X}^{(n)}\mathbf{\Omega}^{1/2})\sigma_{\max}(\mathbf{\Omega})^{-1/2}$ with the last claim of Lemma 2.4.12 (with $\delta = 1/n$), for n sufficiently large, we get that the inequality $\sigma_{\min}(\mathbf{X}^{(n)}) \ge C\sigma_{\max}(\mathbf{\Omega})^{-1/2}$ holds with probability at least $1 - 1/n$. Similarly, using Theorem 2.2.1 with $\delta = 1/n$ we check that for n large enough, with probability at least $1 - 3/n$, we have $\|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,1} \le 3C_0(\max_j \omega_{jj}^{-1/2})|O|(\frac{p\log n}{n})^{1/2}$. In order to evaluate the term $\|(\mathbf{I}_n - \mathbf{Z}^j)\boldsymbol{\xi}_{\bullet,j}\|_2$, we note that its square is drawn from the scaled khi-square distribution $(n\omega_{jj})^{-1}\chi^2_{p-1}$. Therefore, applying the same argument as in Lemma 2.4.13, we check that with probability at least $1 - 1/n$,

$$\max_{j\in[p]} \|(\mathbf{I}_n - \mathbf{Z}^j)\boldsymbol{\xi}_{\bullet,j}\|_2 \le \max_j (n\omega_{jj})^{-1/2}(\sqrt{p-1} + \sqrt{2\log(pn)}) \le 3\max_j \omega_{jj}^{-1/2}\Big(\frac{p\log n}{n}\Big)^{1/2}.$$

In addition, it is clear that $\max_j \omega_{jj}^{-1/2} = (\min_j \omega_{jj})^{-1/2} \le \sigma_{\min}(\mathbf{\Omega})^{-1/2}$. Putting all these bounds together, we obtain the claimed result. □

Lemma 2.4.9. *If* $|O|p = o(n/\log n)$ *then there exists a universal constant* C *such that for* n *large enough, the inequalities*

$$\max_j \frac{\widehat{\omega}_{jj}}{\omega^*_{jj}} \le C, \qquad \|\widehat{\mathbf{D}}(\mathbf{D}^*)^{-1} - \mathbf{I}_p\|_{2,2} \le C\Big\{\rho^* M_{\mathbf{E}}\frac{|O|p\log n}{n} + \Big(\frac{p\log n}{n}\Big)^{1/2}\Big\} \tag{2.39}$$

hold with probability larger than $1 - 4/n$.

Proof. To ease notation, we write ω_{jj} instead of ω^*_{jj} and C_ω for $\max_j \omega_{jj}^{1/2}$. Let us consider the first term in the right-hand side of the above inequality. Recall that the diagonal entries ω_{jj} are estimated by

$$\widehat{\omega}_{jj} = \frac{2n}{\pi \|\mathbf{Z}^j(\mathbf{X}^{(n)}_{\bullet,j} - \widehat{\mathbf{\Theta}}_{\bullet,j})\|_1^2} = \frac{2n}{\pi \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_1^2}.$$

This implies that

$$\begin{aligned}
\Big|\Big(\frac{\omega_{jj}}{\widehat{\omega}_{jj}}\Big)^{\frac{1}{2}} - 1\Big| &= \Big|\Big(\frac{\pi\omega_{jj}}{2n}\Big)^{\frac{1}{2}} \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_1 - 1\Big| \\
&\leq \Big(\frac{\pi\omega_{jj}}{2n}\Big)^{\frac{1}{2}} \big|\|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_1 - \|\boldsymbol{\xi}_{\bullet,j}\|_1\big| + \Big|\Big(\frac{\pi\omega_{jj}}{2n}\Big)^{\frac{1}{2}} \|\boldsymbol{\xi}_{\bullet,j}\|_1 - 1\Big| \\
&\leq \Big(\frac{\pi\omega_{jj}}{2n}\Big)^{\frac{1}{2}} \big(\|\widehat{\boldsymbol{\xi}}_{\bullet,j} - \mathbf{Z}^j\boldsymbol{\xi}_{\bullet,j}\|_1 + \|(\mathbf{I}_n - \mathbf{Z}^j)\boldsymbol{\xi}_{\bullet,j}\|_1\big) + \Big|\Big(\frac{\pi\omega_{jj}}{2n}\Big)^{\frac{1}{2}} \|\boldsymbol{\xi}_{\bullet,j}\|_1 - 1\Big|.
\end{aligned} \tag{2.40}$$

The first term above can be bounded using Theorem 2.2.1 since $\|\widehat{\boldsymbol{\xi}}_{\bullet,j} - \mathbf{Z}^j\boldsymbol{\xi}_{\bullet,j}\|_1 = \|\mathbf{Z}^j\widehat{\boldsymbol{\Delta}}^{\mathbf{\Theta}}_{\bullet,j}\|_1$ and

$$\begin{aligned}
\|\widehat{\boldsymbol{\xi}}_{\bullet,j} - \mathbf{Z}^j\boldsymbol{\xi}_{\bullet,j}\|_1 &\leq \|\widehat{\boldsymbol{\Delta}}^{\mathbf{\Theta}}_{\bullet,j}\|_1 + \|(\mathbf{I}_n - \mathbf{Z}^j)\widehat{\boldsymbol{\Delta}}^{\mathbf{\Theta}}_{\bullet,j}\|_1 \\
&\leq \|\widehat{\boldsymbol{\Delta}}^{\mathbf{\Theta}}_{\bullet,j}\|_1 + \sqrt{n}\,\|(\mathbf{I}_n - \mathbf{Z}^j)\widehat{\boldsymbol{\Delta}}^{\mathbf{\Theta}}_{\bullet,j}\|_2 \\
&\leq \|\widehat{\boldsymbol{\Delta}}^{\mathbf{\Theta}}_{\bullet,j}\|_1 + \sqrt{n}\,\|(\mathbf{I}_n - \mathbf{Z})\widehat{\boldsymbol{\Delta}}^{\mathbf{\Theta}}_{\bullet,j}\|_2.
\end{aligned} \tag{2.41}$$

Note that the result of Theorem 2.2.1 applies to this matrix as well. For the second term of the right-hand side of (2.40), we can use the Cauchy-Schwarz inequality in conjunction with the fact that $n\omega_{jj}\,\|(\mathbf{I}_n - \mathbf{Z}^j)\boldsymbol{\xi}_{\bullet,j}\|_2^2$ is a khi-square random variable with $p-1$ degrees of freedom degrees of freedom and apply Lemma 1 of Laurent and Massart [2000]. The third term of the right-hand side of (2.40) can be bounded using the Hoeffding bounds (see for instance [Vershynin, 2012b, Proposition 5.10]). Let us denote by $b > 0$ the constant such that for any $x \in \mathbb{R}$, the sub-Gaussian random variable $|\boldsymbol{\epsilon}_{i,j}|$—whose expectation is $\sqrt{2/\pi}$—satisfies $\mathbf{E}\big(\mathrm{e}^{x(|\boldsymbol{\epsilon}_{i,j}| - \sqrt{2/\pi})}\big) \leq \mathrm{e}^{x^2 b/2}$. Thus, for any $t > 0$, each of the following bounds

$$\frac{1}{n}\|\boldsymbol{\epsilon}_{\bullet,j}\|_1 - \Big(\frac{2}{\pi}\Big)^{1/2} \leq \Big(\frac{2tb}{n}\Big)^{1/2} \quad ; \quad -\frac{1}{n}\|\boldsymbol{\epsilon}_{\bullet,j}\|_1 + \Big(\frac{2}{\pi}\Big)^{1/2} \leq \Big(\frac{2tb}{n}\Big)^{1/2}$$

holds with probability at least $1 - \mathrm{e}^{-t}$. Then, with probability at least $1 - 2\,\mathrm{e}^{-t}$, it holds that

$$\Big|\Big(\frac{\pi\omega_{jj}}{2n}\Big)^{\frac{1}{2}} \|\boldsymbol{\xi}_{\bullet,j}\|_1 - 1\Big| \leq \Big(\frac{t\pi b}{n}\Big)^{1/2}.$$

By the Minkowski inequality, this readily yields that for n large enough, the inequality

$$\|(\mathbf{D}^*)^{1/2}\widehat{\mathbf{D}}^{-1/2} - \mathbf{I}_p\|_{2,2}^2 = \Big\{ \sum_{j\in[p]} \Big| \Big(\frac{\omega_{jj}}{\widehat{\omega}_{jj}}\Big)^{1/2} - 1\Big|^2 \Big\}$$
$$\le C\Big(\frac{C_\omega^2}{n}\sum_{j=1}^p \|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{\bullet,j}\|_1^2 + C_\omega^2\|(\mathbf{I}_n - \mathbf{Z})\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,2}^2 + \frac{p\log n}{n}\Big).$$

holds with probability at least $1 - 2/n$. One can show that $\sum_{j=1}^p \|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{\bullet,j}\|_1^2 \le \|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,1}^2$ and $\|(\mathbf{I}_n - \mathbf{Z})\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,2}^2 \le \alpha\|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,1}^2$ (see the proof of Prop. 2.4.7). Combining with Theorem 2.2.1 and Eq. (2.43), this yields

$$\|(\mathbf{D}^*)^{1/2}\widehat{\mathbf{D}}^{-1/2} - \mathbf{I}_p\|_{2,2}^2 \le C\Big((\rho^*)^2 M_{\mathbf{E}}^2 \frac{|O|^2 p^2 \log n}{n^2} + \frac{p\log n}{n}\Big).$$

On the other hand, on the same event, we have

$$\big\|\widehat{\boldsymbol{\xi}}_{\bullet,j}\big\|_1 \ge \big\|\mathbf{Z}^j\boldsymbol{\xi}_{\bullet,j}\big\|_1 - \|\mathbf{Z}^j\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{\bullet,j}\|_1 \ge \big\|\boldsymbol{\xi}_{\bullet,j}\big\|_1 - \big\|(\mathbf{I}_n - \mathbf{Z}^j)\boldsymbol{\xi}_{\bullet,j}\big\|_1 - \sqrt{n}\|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,2}$$
$$\ge \sqrt{n}\Big(\frac{C}{\omega_{jj}^{1/2}} - \|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,2}\Big).$$

Therefore, for n large enough, as we assume that $|O|p = o(n/\log n)$, with probability at least $1 - 4/n$ we have $\big\|\widehat{\boldsymbol{\xi}}_{\bullet,j}\big\|_2 \ge \frac{Cn^{1/2}}{2\omega_{jj}^{1/2}}$ for all $j \in [p]$ and hence $\max_j \widehat{\omega}_{jj}/\omega_{jj} \le C$.

For the second claim of the lemma, we use the inequalities

$$\|\widehat{\mathbf{D}}(\mathbf{D}^*)^{-1} - \mathbf{I}_p\|_{2,2} \le 2\max_j \frac{\widehat{\omega}_{jj} \vee \omega_{jj}}{\omega_{jj}} \big\|(\mathbf{D}^*)^{1/2}\widehat{\mathbf{D}}^{-1/2} - \mathbf{I}_p\big\|_{2,2}$$
$$\le C\big(\max_j \omega_{jj}^{1/2}\|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,2} + (p\log n/n)^{1/2}\big)$$
$$\le C\Big\{\rho^* M_{\mathbf{E}} \frac{|O|p\log n}{n} + \Big(\frac{p\log n}{n}\Big)^{1/2}\Big\}.$$

This completes the proof of the lemma. □

The claim of Theorem 2.2.2 readily follows from Lemmas 2.4.8 and 2.4.9, in conjunction with (2.36).

2.4.3 Probabilistic bounds

This section is devoted to the establishing nonasymptotic bounds on the stochastic terms encountered during the evaluation of the estimation error.

Lemma 2.4.10. *For any $\delta \in (0,1)$, the inequality*

$$\max_{i\in[n]} \max_{j\in[p]} \frac{(\mathbf{Z}^j_{i,\bullet}\boldsymbol{\epsilon}_{\bullet,j})^2}{\|\mathbf{Z}^j\boldsymbol{\epsilon}_{\bullet,j}\|_2^2} \leq \frac{2\log(2np/\delta)}{n-p+1-2((n-p+1)\log(2p/\delta))^{1/2}}$$

holds with probability at least $1-\delta$. Furthermore, if $n \geq 8p + 16\log(4/\delta)$ then

$$\max_{i\in[n]} \max_{j\in[p]} \frac{(\mathbf{Z}^j_{i,\bullet}\boldsymbol{\epsilon}_{\bullet,j})^2}{\|\mathbf{Z}^j\boldsymbol{\epsilon}_{\bullet,j}\|_2^2} \leq \frac{4\log(2np/\delta)}{n}$$

holds with probability at least $1-\delta$.

Proof. Let us introduce the following random variables

$$N_{ij} := \mathbf{Z}^j_{i,\bullet}\boldsymbol{\epsilon}_{\bullet,j} \quad \text{and} \quad D_j := \|\mathbf{Z}^j\boldsymbol{\epsilon}_{\bullet,j}\|_2^2.$$

The random vector $\boldsymbol{\epsilon}_{\bullet,j}$ being Gaussian and independent of $\mathbf{X}_{\bullet,j^c}$, we infer that conditionally to $\mathbf{Z}^j$, the random variable N_{ij} is drawn from a zero mean Gaussian distribution. Furthermore, its conditional variance given $\mathbf{Z}^j$ equals $\mathbf{Z}^j_{i,\bullet}(\mathbf{Z}^j_{i,\bullet})^\top = \mathbf{Z}^j_{i,i}$ and, therefore is less than or equal to 1. (Here, we have used the fact that $\mathbf{Z}^j$ is symmetric, idempotent and that all the entries of a projection matrix are in absolute value smaller than or equal to 1.) This implies that for any $\delta > 0$, it holds that $\mathbf{P}\big(\max_{i\in[n],j\in[p]} |N_{ij}| > \sqrt{2\log(2np/\delta)}\big) \leq \delta/2$.

We know that $\mathbf{Z}^j$ is an orthogonal projection matrix onto a subspace of dimension $\operatorname{rank}(\mathbf{Z}^j)$. We recall that the square of the Euclidean norm of the orthogonal projection in a subspace of dimension k of a standard Gaussian random vector is a χ^2 random variable with k degrees of freedom. It entails that, conditionally to $\mathbf{Z}^j$, D_j has a χ^2 distribution with $\operatorname{rank}(\mathbf{Z}^j)$ degrees of freedom. Therefore, noticing that $\operatorname{rank}(\mathbf{Z}^j) \geq n - \operatorname{rank}(\mathbf{X}_{\bullet,j^c}) = n-p+1$ almost surely and using a prominent result on tail bounds for the χ^2 distribution (see Lemma 1 of Laurent and Massart [2000]), we get, for every $\delta \in (0,1)$

$$\mathbf{P}\big(\min_{j\in[p]} D_j \leq n-p+1-2\sqrt{(n-p+1)\log(2p/\delta)}\big) \leq \delta/2.$$

Thus, on an event of probability at least $1-\delta$, we have

$$\max_{\substack{i\in[n]\\ j\in[p]}} |N_{ij}| \leq \sqrt{2\log(2np/\delta)} \qquad \text{and} \qquad \min_{j\in[p]} D_j \geq n-p+1-2\sqrt{(n-p+1)\log(2p/\delta)}. \tag{2.42}$$

This readily entails the first claim of the lemma. The second claim follows from the first

one. Indeed, $n \geq 8p + 16\log(4/\delta)$ implies that $3p + 8\log(4/\delta) \leq 0.5n - p$ and, hence,

$$\begin{aligned}
16(n-p+1)\log(2p/\delta) &\leq \big(0.5(n-p+1) + 8\log(2p/\delta)\big)^2 \\
&\leq \big(0.5n - p + 1 + 0.5p + 8\log(p/2) + 8\log(4/\delta)\big)^2 \\
&\leq \big(0.5n - p + 1 + 3p + 8\log(4/\delta)\big)^2 \\
&\leq \big(n - 2p + 1\big)^2.
\end{aligned}$$

This yields $n - p + 1 - 2((n-p+1)\log(2p/\delta))^{1/2} \geq n/2$. □

The element-wise ℓ_∞-norm of the orthogonal projection matrix $\mathbf{I}_n - \mathbf{Z}$ also appears in the upper bounds of the estimation error. Lemma 2.4.12 below provides a sharp tail bound for this norm. Before showing this result, let us provide a useful technical lemma that relies essentially on a lower bound for the smallest singular value of a Gaussian matrix.

Lemma 2.4.11. *If $\mathbf{X}$ is an $n \times p$ random matrix satisfying conditions (C1) and (C2) with $\mathbf{\Sigma}^* = \mathbf{I}_p$, then for every $\delta \in (0,1)$, with probability at least $1 - \delta$, it holds that*

$$\sigma_{\min}(\mathbf{X}) \geq \sqrt{n - |O|} - \sqrt{p} - \sqrt{2\log(2/\delta)}.$$

Proof. To begin, we note that the matrix $\mathbf{X}^\top\mathbf{X}$ can be split into two parts, by summing the terms derived from inliers ($I \subset [n]$) on one hand and those derived from outliers ($O \subset [n]$) on the other hand,

$$\mathbf{X}^\top\mathbf{X} = \sum_{i\in[n]} \mathbf{X}_{i,\bullet}^\top \mathbf{X}_{i,\bullet} = \sum_{i\in I} \mathbf{X}_{i,\bullet}^\top \mathbf{X}_{i,\bullet} + \sum_{i\in O} \mathbf{X}_{i,\bullet}^\top \mathbf{X}_{i,\bullet} = \mathbf{X}_{I,\bullet}^\top \mathbf{X}_{I,\bullet} + \mathbf{X}_{O,\bullet}^\top \mathbf{X}_{O,\bullet}.$$

As the matrix $\mathbf{X}_{O,\bullet}^\top \mathbf{X}_{O,\bullet}$ is always nonnegative definite and $\mathbf{X}^\top\mathbf{X} = \mathbf{X}_{O,\bullet}^\top \mathbf{X}_{O,\bullet} + \mathbf{X}_{I,\bullet}^\top \mathbf{X}_{I,\bullet}$, we infer that $\sigma_{\min}(\mathbf{X}^\top\mathbf{X}) \geq \sigma_{\min}(\mathbf{X}_{I,\bullet}^\top \mathbf{X}_{I,\bullet})$. We can therefore deduce that

$$\sigma_{\min}(\mathbf{X}) = \sigma_{\min}(\mathbf{X}^\top\mathbf{X})^{1/2} \geq \sigma_{\min}(\mathbf{X}_{I,\bullet}^\top \mathbf{X}_{I,\bullet})^{1/2} = \sigma_{\min}(\mathbf{X}_{I,\bullet}).$$

Given that $\mathbf{X}_{I,\bullet}$ is a matrix whose rows are independent Gaussian vectors with zero-mean and identity covariance, as shown in [Vershynin, 2012b, Corollary 5.35], for every $t \geq 0$, it holds that

$$\sigma_{\min}(\mathbf{X}_{I,\bullet}) \geq \sqrt{|I|} - \sqrt{p} - t.$$

with probability at least $1 - 2\,\mathrm{e}^{-t^2/2}$. Taking $t = \sqrt{2\log(2/\delta)}$, the claim of the lemma follows. □

Lemma 2.4.12. *If $\mathbf{X} = \mathbf{Y} + \mathbf{E}^*$ is an $n \times p$ random matrix with $\mathbf{Y}$ and $\mathbf{E}^*$ satisfying*

assumptions **(C1)** *and* **(C2)** *with* $\boldsymbol{\mu}^* = 0$*, then for any* $\delta \in (0,1)$*, the inequality*

$$\|\mathbf{I}_n - \mathbf{Z}\|_{\infty,\infty} \leq \left(\frac{(1+M_{\mathbf{E}})\sqrt{p} + \sqrt{2\log(2n/\delta)}}{\sqrt{n-|O|} - \sqrt{p} - \sqrt{2\log(4/\delta)}} \right)^2,$$

holds with probability at least $1-\delta$*. Furthermore, if* $n \geq |O| + 8p + 16\log(4/\delta)$*, then with probability at least* $1-\delta$*,*

$$\|\mathbf{I}_n - \mathbf{Z}\|_{\infty,\infty} \leq \frac{8(1+M_{\mathbf{E}})^2 p + 16\log(2n/\delta)}{n-|O|}. \tag{2.43}$$

and $\sigma_{\min}(\mathbf{X}(\boldsymbol{\Omega}^*)^{1/2}) \geq \sqrt{(n-|O|)/4}$*.*

Proof. We denote by $\{\mathbf{e}_i\}_{i\in[n]} \subset \mathbb{R}^n$ the vectors of the canonical basis. All the components of the vector $\mathbf{e}_i \in \mathbb{R}^n$ are equal to zero with the exception of the i-th entry which is equal to one. With this notation, and using the fact that all the off-diagonal entries of a symmetric positive semi-definite matrix are dominated by the largest diagonal entry, we have

$$\|\mathbf{I}_n - \mathbf{Z}\|_{\infty,\infty} = \max_{i\in[n]} \mathbf{e}_i^\top (\mathbf{I}_n - \mathbf{Z})\mathbf{e}_i.$$

We also denote $\mathbf{X}(\boldsymbol{\Sigma}^*)^{-1/2}$ by $\widetilde{\mathbf{X}}$ and, similarly, $\mathbf{Y}(\boldsymbol{\Sigma}^*)^{-1/2}$ by $\widetilde{\mathbf{Y}}$. It follows that for any $i \in [n]$

$$\mathbf{e}_i^\top (\mathbf{I}_n - \mathbf{Z})\mathbf{e}_i = \mathbf{e}_i^\top \widetilde{\mathbf{X}} \big(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}\big)^\dagger \widetilde{\mathbf{X}}^\top \mathbf{e}_i \leq \|\widetilde{\mathbf{X}}_{i,\bullet}\|_2^2 \sigma_{\max}\big((\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^\dagger\big).$$

where the last inequality is a direct consequence of the fact that the spectral norm is the matrix norm induced by the Euclidean norm. We may now bound each term of the right side of the previous inequality. First, by assumption, it holds that

$$\|\widetilde{\mathbf{X}}_{i,\bullet}\|_2 = \|\widetilde{\mathbf{Y}}_{i,\bullet} + \mathbf{E}^*_{i,\bullet}(\boldsymbol{\Sigma}^*)^{-1/2}\|_2 \leq \|\widetilde{\mathbf{Y}}_{i,\bullet}\|_2 + \|\mathbf{E}^*_{i,\bullet}(\boldsymbol{\Sigma}^*)^{-1/2}\|_2 \leq \|\widetilde{\mathbf{Y}}_{i,\bullet}\|_2 + M_{\mathbf{E}}\sqrt{p}.$$

As $\widetilde{\mathbf{Y}}_{i,\bullet} \sim \mathcal{N}_p(0, \mathbf{I}_p)$, the random variable $\|\widetilde{\mathbf{Y}}_{i,\bullet}\|_2^2$ has a χ^2 distribution with p degrees of freedom. Applying [Laurent and Massart, 2000, Lemma 1] and combining it with the union bound, for any $\delta \in (0,1)$, we get that

$$\max_{i\in[n]} \|\widetilde{\mathbf{Y}}_{i,\bullet}\|_2 \leq \sqrt{p} + \sqrt{2\log(2n/\delta)},$$

with probability at least $1-\delta/2$. We complete the proof by bounding $\sigma_{\max}((\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^\dagger) = \sigma_{\min}(\widetilde{\mathbf{X}})^{-2}$. By Lemma 2.4.11, for every $\delta \in (0,1)$, it holds that

$$\sigma_{\max}((\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^\dagger) \leq (\sqrt{|I|} - \sqrt{p} - \sqrt{2\log(4/\delta)})^{-2}$$

with probability at least $1-\delta/2$. By bringing together what was written above, with probability at least $1-\delta$, we have

$$\max_{i\in[n]} \mathbf{e}_i^\top (\mathbf{I}_n - \mathbf{Z})\mathbf{e}_i \leq \left(\frac{(1+M_{\mathbf{E}})\sqrt{p} + \sqrt{2\log(2n/\delta)}}{\sqrt{|I|} - \sqrt{p} - \sqrt{2\log(4/\delta)}} \right)^2 .$$

This yields the first claim of the lemma. To derive the second claim from the first one, it suffices to upper bound the numerator using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ and to lower bound the denominator by using that $\sqrt{p} + \sqrt{2\log(4/\delta)} \leq \sqrt{2p + 4\log(4/\delta)} \leq \frac{1}{2}\sqrt{n - |O|}$. □

Lemma 2.4.13. *For any $\delta \in (0,1)$, the following inequality*

$$\|\boldsymbol{\epsilon}^\top\|_{2,\infty} \leq \sqrt{n} + \sqrt{2\log(p/\delta)}, \tag{2.44}$$

holds with probability at least $1-\delta$.

Proof. We recall that $\|\boldsymbol{\epsilon}^\top\|_{2,\infty} = \max_{j\in[p]} \|\boldsymbol{\epsilon}_{\bullet,j}\|_2$. As we have already mentioned just after equation (2.7), the vector $\boldsymbol{\epsilon}_{\bullet,j}$ is drawn from the Gaussian $\mathcal{N}_n(0, \mathbf{I}_n)$ distribution. Therefore, $\|\boldsymbol{\epsilon}_{\bullet,j}\|_2^2$ is a χ^2 random variable with n degrees of freedom. Thus, using [Laurent and Massart, 2000, Lemma 1] in combination with the union bound, it holds that

$$\|\boldsymbol{\epsilon}^\top\|_{2,\infty}^2 \leq n + 2\sqrt{n\log(p/\delta)} + 2\log(p/\delta) \leq (\sqrt{n} + \sqrt{2\log(p/\delta)})^2,$$

with probability at least $1-\delta$. □

2.4.4 Proofs in high dimension

In this section, we provide the proof of the risk bound in the high dimensional case, when the estimator is obtained by solving the optimization problem in (2.4). We define $\mathcal{O} = O \times [p]$ and, by a slight abuse of notation, $\mathcal{O}^c = O^c \times [p]$. We denote by $\boldsymbol{\xi}_{\mathcal{O}}$, resp. $\boldsymbol{\xi}_{\mathcal{O}^c}$, the matrix obtained by zeroing all the rows $\boldsymbol{\xi}_{i,\bullet}$ such that $i \in O$, resp. $i \in O^c$. We set $\bar{\boldsymbol{\Theta}}^* = \boldsymbol{\Theta}^* + \boldsymbol{\xi}_{\mathcal{O}}$ and $\bar{\boldsymbol{\xi}} = \boldsymbol{\xi}_{\mathcal{O}^c}$. We further define $\widehat{\boldsymbol{\Delta}}^{\mathbf{B}} = \widehat{\mathbf{B}} - \mathbf{B}^*$, $\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Theta}} - \bar{\boldsymbol{\Theta}}^*$, $\widehat{\boldsymbol{\Delta}} = \begin{bmatrix} \widehat{\boldsymbol{\Delta}}^{\mathbf{B}} \\ \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}} \end{bmatrix} \in \mathbb{R}^{(p+n)\times p}$ and $\widehat{\boldsymbol{\xi}} = \mathbf{X}^{(n)}\widehat{\mathbf{B}} - \widehat{\boldsymbol{\Theta}}$. Since $\mathbf{M} = [\mathbf{X}^{(n)}; -\mathbf{I}_n]$, the estimator $(\widehat{\mathbf{B}}, \widehat{\boldsymbol{\Theta}})$ is defined as the minimizer of the cost function

$$F(\mathbf{B}, \boldsymbol{\Theta}) = \left\| \left(\mathbf{M} \begin{bmatrix} \mathbf{B} \\ \boldsymbol{\Theta} \end{bmatrix} \right)^\top \right\|_{2,1} + \lambda\big(\|\boldsymbol{\Theta}\|_{2,1} + \gamma\|\mathbf{B}\|_{1,1}\big).$$

Recall that $\mathcal{J}$ and O are such that $\mathbf{B}^*_{\mathcal{J}^c} = 0$ and $\boldsymbol{\Theta}^*_{O^c,\bullet} = 0$. This sets are interpreted as the supports of $\mathbf{B}^*$ and $\boldsymbol{\Theta}^*$. The set $\mathcal{J}$ corresponds to the sparsity pattern and O to the outliers. Throughout this section, we adopt the convention that $0/0 = 0$.

Proposition 2.4.14. *If, for some constant $c > 1$, the penalty levels λ and γ satisfy the conditions*

$$\lambda\gamma \geq \frac{c+1}{c-1} \max_{j\in[p]} \frac{\|\mathbf{X}_{I,j^c}^{(n)\top} \boldsymbol{\epsilon}_{I,j}\|_\infty}{\|\boldsymbol{\epsilon}_{I,j}\|_2} \quad \textit{and} \quad \lambda \geq \frac{c+1}{c-1} \max_{i\in[n]} \Big(\sum_{j\in[p]} \frac{\epsilon_{ij}^2}{\|\boldsymbol{\epsilon}_{I,j}\|_2^2} \Big)^{1/2}, \tag{2.45}$$

then the matrix $\widehat{\boldsymbol{\Delta}}$ belongs to the cone $\mathscr{C}_{\mathcal{J},O}(c,\gamma)$.

Proof. Let us define $\widehat{\boldsymbol{\xi}}$ as the $n \times p$ matrix of estimated residuals: $\widehat{\boldsymbol{\xi}} = \mathbf{X}^{(n)}\widehat{\mathbf{B}} - \widehat{\boldsymbol{\Theta}}$. By definition of $\widehat{\mathbf{B}}$ and $\widehat{\boldsymbol{\Theta}}$, we obtain the inequality

$$\|(\mathbf{X}^{(n)}\widehat{\mathbf{B}} - \widehat{\boldsymbol{\Theta}})^\top\|_{2,1} + \lambda\big(\gamma\|\widehat{\mathbf{B}}\|_{1,1} + \|\widehat{\boldsymbol{\Theta}}\|_{2,1}\big) \leq \|(\mathbf{X}^{(n)}\mathbf{B}^* - \bar{\boldsymbol{\Theta}}^*)^\top\|_{2,1} + \lambda\big(\gamma\|\mathbf{B}^*\|_{1,1} + \|\bar{\boldsymbol{\Theta}}^*\|_{2,1}\big),$$

that can be equivalently written as

$$\|\widehat{\boldsymbol{\xi}}^\top\|_{2,1} + \lambda\gamma\|\widehat{\mathbf{B}}\|_{1,1} + \lambda\|\widehat{\boldsymbol{\Theta}}\|_{2,1} \leq \|\bar{\boldsymbol{\xi}}^\top\|_{2,1} + \lambda\gamma\|\mathbf{B}^*\|_{1,1} + \lambda\|\bar{\boldsymbol{\Theta}}^*\|_{2,1},$$

or as

$$\lambda\gamma(\|\widehat{\mathbf{B}}\|_{1,1} - \|\mathbf{B}^*\|_{1,1}) + \lambda(\|\widehat{\boldsymbol{\Theta}}\|_{2,1} - \|\bar{\boldsymbol{\Theta}}^*\|_{2,1}) \leq \sum_{j\in[p]} (\|\bar{\boldsymbol{\xi}}_{\bullet,j}\|_2 - \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2). \tag{2.46}$$

In view of the inequality $\|a\|_2 - \|b\|_2 \leq (a-b)^\top a/\|a\|_2$, which holds for every pair of vectors (a,b) and is a simple consequence of the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\|\bar{\boldsymbol{\xi}}_{\bullet,j}\|_2 - \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2 &\leq (\boldsymbol{\xi}_{I,j} - \widehat{\boldsymbol{\xi}}_{I,j})^\top \frac{\boldsymbol{\xi}_{I,j}}{\|\boldsymbol{\xi}_{I,j}\|_2} \\
&= (\boldsymbol{\xi}_{I,j} - \widehat{\boldsymbol{\xi}}_{I,j})^\top \frac{\boldsymbol{\epsilon}_{I,j}}{\|\boldsymbol{\epsilon}_{I,j}\|_2} \\
&= (-\mathbf{X}_{I,\bullet}^{(n)}\widehat{\boldsymbol{\Delta}}_{\bullet,j}^{\mathbf{B}} + \widehat{\boldsymbol{\Delta}}_{I,j}^{\boldsymbol{\Theta}})^\top \frac{\boldsymbol{\epsilon}_{I,j}}{\|\boldsymbol{\epsilon}_{I,j}\|_2}.
\end{aligned}$$

Summing these inequalities over all $j \in [p]$ and applying the duality inequalities we infer that

$$\begin{aligned}
\sum_{j\in[p]} (\|\boldsymbol{\xi}_{\bullet,j}\|_2 - \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2) &\leq -\sum_{j\in[p]} (\mathbf{X}_{I,\bullet}^{(n)}\widehat{\boldsymbol{\Delta}}_{\bullet,j}^{\mathbf{B}})^\top \frac{\boldsymbol{\epsilon}_{I,j}}{\|\boldsymbol{\epsilon}_{I,j}\|_2} + \sum_{i\in I}\sum_{j\in[p]} \widehat{\boldsymbol{\Delta}}_{i,j}^{\boldsymbol{\Theta}} \frac{\boldsymbol{\epsilon}_{i,j}}{\|\boldsymbol{\epsilon}_{I,j}\|_2} \\
&\leq \sum_{j\in[p]} \|\widehat{\boldsymbol{\Delta}}_{\bullet,j}^{\mathbf{B}}\|_1 \frac{\|\mathbf{X}_{I,j^c}^{(n)\top} \boldsymbol{\epsilon}_{I,j}\|_\infty}{\|\boldsymbol{\epsilon}_{I,j}\|_2} + \sum_{i\in[n]} \|\widehat{\boldsymbol{\Delta}}_{i,\bullet}^{\boldsymbol{\Theta}}\|_2 \Big(\sum_{j\in[p]} \frac{\epsilon_{ij}^2}{\|\boldsymbol{\epsilon}_{I,j}\|_2^2} \Big)^{1/2}.
\end{aligned}$$

When condition (2.45) is satisfied, the last inequality yields

$$\sum_{j\in[p]}(\|\bar{\boldsymbol{\xi}}_{\bullet,j}\|_2 - \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2) \le \left(\frac{c-1}{c+1}\right)\left(\lambda\gamma\sum_{j\in[p]}\|\widehat{\boldsymbol{\Delta}}^{\mathbf{B}}_{\bullet,j}\|_1 + \lambda\sum_{i\in[n]}\|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{i,\bullet}\|_2\right)$$
$$= \lambda\left(\frac{c-1}{c+1}\right)(\gamma\|\widehat{\boldsymbol{\Delta}}^{\mathbf{B}}\|_{1,1} + \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{2,1}).$$

This inequality, in conjunction with Eq. (2.46), implies that

$$\gamma(\|\widehat{\mathbf{B}}\|_{1,1} - \|\mathbf{B}^*\|_{1,1}) + (\|\widehat{\boldsymbol{\Theta}}\|_{2,1} - \|\bar{\boldsymbol{\Theta}}^*\|_{2,1}) \le \left(\frac{c-1}{c+1}\right)(\gamma\|\widehat{\boldsymbol{\Delta}}^{\mathbf{B}}\|_{1,1} + \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{2,1}). \quad (2.47)$$

On the other hand, using the triangle inequality and the fact that $\mathbf{B}^*_{\mathcal{J}^c} = \boldsymbol{\Theta}^*_{O^c,\bullet} = 0$, we get

$$\begin{aligned}
\|\widehat{\mathbf{B}}\|_{1,1} - \|\mathbf{B}^*\|_{1,1} &= \|\widehat{\mathbf{B}}_{\mathcal{J}^c}\|_{1,1} + \|\widehat{\mathbf{B}}_{\mathcal{J}}\|_{1,1} - \|\mathbf{B}^*_{\mathcal{J}}\|_{1,1} \\
&\ge \|\widehat{\boldsymbol{\Delta}}^{\mathbf{B}}_{\mathcal{J}^c}\|_{1,1} - \|\widehat{\boldsymbol{\Delta}}^{\mathbf{B}}_{\mathcal{J}}\|_{1,1}, \\
\|\widehat{\boldsymbol{\Theta}}\|_{2,1} - \|\bar{\boldsymbol{\Theta}}^*\|_{2,1} &= \|\widehat{\boldsymbol{\Theta}}_{O^c,\bullet}\|_{2,1} + \|\widehat{\boldsymbol{\Theta}}_{O,\bullet}\|_{2,1} - \|\bar{\boldsymbol{\Theta}}^*_{O,\bullet}\|_{2,1} \\
&\ge \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{O^c,\bullet}\|_{2,1} - \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{O,\bullet}\|_{2,1}.
\end{aligned}$$

The combination of these bounds with Eq. (2.47) leads to

$$\|\widehat{\boldsymbol{\Delta}}^{\mathbf{B}}_{\mathcal{J}^c}\|_{1,1} + \gamma^{-1}\|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{O^c,\bullet}\|_{2,1} \le c\big(\|\widehat{\boldsymbol{\Delta}}^{\mathbf{B}}_{\mathcal{J}}\|_{1,1} + \gamma^{-1}\|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{O,\bullet}\|_{2,1}\big),$$

which completes the proof of the proposition. □

The following lemmas prepare the proof of Theorem 2.3.1. Lemma 2.4.15 presents an inequality obtained by writing the KKT conditions for the cost function F.

Lemma 2.4.15. *There exists a $n \times p$ matrix $\mathbf{V}$ is such that*

$$\|\mathbf{V}_{i,\bullet}\|_2 \le 1, \qquad \mathbf{V}_{i,\bullet}^\top \widehat{\boldsymbol{\Theta}}_{i,\bullet} = \|\widehat{\boldsymbol{\Theta}}_{i,\bullet}\|_2, \qquad \forall\, i \in [n] \quad (2.48)$$

and, for every $j \in [p]$ such that $\widehat{\boldsymbol{\xi}}_{\bullet,j} \neq 0$, the following inequality holds

$$\|\mathbf{M}\widehat{\boldsymbol{\Delta}}_{\bullet,j}\|_2^2 \le -\bar{\boldsymbol{\xi}}_{\bullet,j}^\top \mathbf{M}\widehat{\boldsymbol{\Delta}}_{\bullet,j} - \lambda\|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2 \mathbf{V}_{\bullet,j}^\top \widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{\bullet,j} + \lambda\gamma\|\widehat{\xi}_{\bullet,j}\|_2(\|\mathbf{B}^*_{j^c,j}\|_1 - \|\widehat{\mathbf{B}}_{j^c,j}\|_1). \quad (2.49)$$

Proof. Recall that the estimator $(\widehat{\mathbf{B}}, \widehat{\boldsymbol{\Theta}})$ minimizes the cost function

$$F(\mathbf{B}, \boldsymbol{\Theta}) = \sum_{j=1}^{p}\|\mathbf{X}^{(n)}_{\bullet,j} + \mathbf{X}^{(n)}_{\bullet,j^c}\mathbf{B}_{j^c,j} - \boldsymbol{\Theta}_{\bullet,j}\|_2 + \lambda\gamma\sum_{j=1}^{p}\|\mathbf{B}_{\bullet,j}\|_1 + \lambda\sum_{i=1}^{n}\|\boldsymbol{\Theta}_{i,\bullet}\|_2. \quad (2.50)$$

According to the KKT conditions, this convex function is minimized at $(\widehat{\mathbf{B}}, \widehat{\boldsymbol{\Theta}})$ if and only

if the zero vector belongs to the sub-differential of F at $(\widehat{\mathbf{B}}, \widehat{\mathbf{\Theta}})$, denoted by $\partial F(\widehat{\mathbf{B}}, \widehat{\mathbf{\Theta}})$. This entails, in particular, that for every $j \in [p]$, $\mathbf{0}_{p-1+n} \in \partial_{(\mathbf{B}_{j^c,j}, \mathbf{\Theta}_{\bullet,j})} F(\widehat{\mathbf{B}}, \widehat{\mathbf{\Theta}})$. In other terms, there exist vectors $\mathbf{u}_j \in \partial_{(\mathbf{B}_{j^c,j}, \mathbf{\Theta}_{\bullet,j})} \|\mathbf{X}^{(n)}_{\bullet,j} + \mathbf{X}^{(n)}_{\bullet,j^c} \widehat{\mathbf{B}}_{j^c,j} - \widehat{\mathbf{\Theta}}_{\bullet,j}\|_2$, $\mathbf{w}_j \in \partial_{(\mathbf{B}_{j^c,j}, \mathbf{\Theta}_{\bullet,j})} \|\widehat{\mathbf{B}}_{\bullet,j}\|_1$ and $\mathbf{v}_j \in \partial_{(\mathbf{B}_{j^c,j}, \mathbf{\Theta}_{\bullet,j})} \sum_{i=1}^n \|\widehat{\mathbf{\Theta}}_{i,\bullet}\|_2$ such that $\mathbf{u}_j + \lambda\gamma \mathbf{w}_j + \lambda \mathbf{v}_j = 0$. Since we assume that $\|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2 > 0$, the first partial sub-differential out of three appearing in the previous sentence is actually a differential and thus $\mathbf{u}_j = [\mathbf{X}^{(n)}_{\bullet,j^c}; -\mathbf{I}_n]^\top (\mathbf{X}^{(n)} \widehat{\mathbf{B}}_{\bullet,j} - \widehat{\mathbf{\Theta}}_{\bullet,j}) / \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2$. After a multiplication by $\|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2$, we get

$$[\mathbf{X}^{(n)}_{\bullet,j^c}; -\mathbf{I}_n]^\top (\mathbf{X}^{(n)} \widehat{\mathbf{B}}_{\bullet,j} - \widehat{\mathbf{\Theta}}_{\bullet,j}) = -\lambda\gamma \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2 \mathbf{w}_j - \lambda \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2 \mathbf{v}_j.$$

This equation (combined with relation (2.8)) can be equivalently written as

$$[\mathbf{X}^{(n)}_{\bullet,j^c}; -\mathbf{I}_n]^\top \mathbf{M} \widehat{\mathbf{\Delta}}_{\bullet,j} = -[\mathbf{X}^{(n)}_{\bullet,j^c}; -\mathbf{I}_n]^\top \bar{\boldsymbol{\xi}}_{\bullet,j} - \lambda\gamma \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2 \mathbf{w}_j - \lambda \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2 \mathbf{v}_j.$$

We take the scalar product of the both sides of this relation with the vector $\widehat{\mathbf{\Delta}}_{j^c,j}$ and, using the fact that $\widehat{\mathbf{\Delta}}_{j,j} = 0$, we obtain

$$\|\mathbf{M} \widehat{\mathbf{\Delta}}_{\bullet,j}\|_2^2 = -\widehat{\mathbf{\Delta}}^\top_{\bullet,j} \mathbf{M}^\top \bar{\boldsymbol{\xi}}_{\bullet,j} - \lambda\gamma \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2 \widehat{\mathbf{\Delta}}^\top_{\bullet,j} \mathbf{w}_j - \lambda \|\widehat{\boldsymbol{\xi}}_{\bullet,j}\|_2 \widehat{\mathbf{\Delta}}^\top_{\bullet,j} \mathbf{v}_j.$$

The desired inequality follows by setting $\mathbf{V} = [(\mathbf{v}_1)_{p:(p-1+n)}, \ldots, (\mathbf{v}_p)_{p:(p-1+n)}]$ and by using the following simple properties of the sub-differentials of the ℓ_1 and ℓ_2-norms:

$$\begin{aligned}
(\mathbf{w}_j)_l &= 0, && \forall l \geq p, \\
|(\mathbf{w}_j)_l| &\leq 1, && \forall l \in [p-1], \\
(\mathbf{w}_j)^\top_{1:(p-1)} \widehat{\mathbf{B}}_{j^c,j} &= \|\widehat{\mathbf{B}}_{j^c,j}\|_1, \\
(\mathbf{v}_j)_l &= 0, && \forall l \in [p-1], \\
(\mathbf{v}_j)_{p-1+i} &= \frac{\widehat{\mathbf{\Theta}}_{i,j}}{\|\widehat{\mathbf{\Theta}}_{i,\bullet}\|_2}, && \begin{cases} i \in [n], \\ \|\widehat{\mathbf{\Theta}}_{i,\bullet}\|_2 > 0, \end{cases} \\
|(\mathbf{v}_j)_{p-1+i}| &\leq \frac{|\boldsymbol{\theta}_j|}{\|\boldsymbol{\theta}\|_2}, && \begin{cases} i \in [n], \\ \|\widehat{\mathbf{\Theta}}_{i,\bullet}\|_2 = 0, \\ \forall \boldsymbol{\theta} \in \mathbb{R}^p, \|\boldsymbol{\theta}\|_2 > 0. \end{cases}
\end{aligned}$$

Indeed, the first three relations imply that $-\widehat{\mathbf{\Delta}}^\top_{\bullet,j} \mathbf{w}_j \leq \|\mathbf{B}^*_{\bullet,j}\|_1 - \|\widehat{\mathbf{B}}_{\bullet,j}\|_1$ while the three last relations yield $\widehat{\mathbf{\Delta}}^\top_{\bullet,j} \mathbf{v}_j = \mathbf{V}^\top_{\bullet,j} \widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{\bullet,j}$ along with $\|\mathbf{V}_{i,\bullet}\|_2 \leq 1$ and $\mathbf{V}_{i,\bullet} \widehat{\mathbf{\Theta}}^\top_{i,\bullet} = \|\widehat{\mathbf{\Theta}}_{i,\bullet}\|_2$. □

Lemma 2.4.16. *If inequality (2.49) is true, then*

$$\begin{aligned}
\|\mathbf{M} \widehat{\mathbf{\Delta}}_{\bullet,j}\|_2^2 \leq{}& -2\bar{\boldsymbol{\xi}}^\top_{\bullet,j} \mathbf{M} \widehat{\mathbf{\Delta}}_{\bullet,j} - 2\lambda \|\bar{\boldsymbol{\xi}}_{\bullet,j}\|_2 \mathbf{V}^\top_{\bullet,j} \widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{\bullet,j} + 2\lambda\gamma \|\bar{\boldsymbol{\xi}}_{\bullet,j}\|_2 \big(\|\mathbf{B}^*_{j^c,j}\|_1 - \|\widehat{\mathbf{B}}_{j^c,j}\|_1\big) \\
&+ \lambda^2 (\gamma \|\widehat{\mathbf{\Delta}}^{\mathbf{B}}_{\bullet,j}\|_1 + |\mathbf{V}^\top_{\bullet,j} \widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{\bullet,j}|)^2. \qquad (2.51)
\end{aligned}$$

Proof. This is a direct consequence of Lemma 2.4.3 (with $R = \|\mathbf{M}\widehat{\mathbf{\Delta}}_{\bullet,j}\|_2$) and the fact that $|\|\widehat{\xi}_{\bullet,j}\|_2 - \|\bar{\xi}_{\bullet,j}\|_2| \leq \|\mathbf{M}\widehat{\mathbf{\Delta}}_{\bullet,j}\|_2$. □

Lemma 2.4.17. *If inequality (2.51) is true and if the penalty levels λ and γ satisfy conditions (2.45) for some constant $c > 1$, then*

$$\|\mathbf{M}\widehat{\mathbf{\Delta}}\|_F^2 \leq 4\lambda c\|\boldsymbol{\xi}_{I,\bullet}^\top\|_{2,\infty}\big(\gamma\|\widehat{\mathbf{\Delta}}^{\mathbf{B}}_{\mathcal{J}}\|_{1,1} + \|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{O,\bullet}\|_{2,1}\big) + \lambda^2(1+c)^2\big(\gamma\|\widehat{\mathbf{\Delta}}^{\mathbf{B}}_{\mathcal{J}}\|_{1,1} + \|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{O,\bullet}\|_{2,1}\big)^2. \tag{2.52}$$

Proof. We begin by noting that for a $n \times p$ matrix $\mathbf{V}$ that satisfies $\|\mathbf{V}_{i,\bullet}\|_2 \leq 1$ for any i belonging to $[n]$, the Cauchy-Schwarz inequality yields that

$$\sum_{j=1}^p |\mathbf{V}_{\bullet,j}^\top \widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{\bullet,j}| \leq \sum_{i=1}^n \sum_{j=1}^p |\mathbf{V}_{i,j}\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{i,j}| \leq \sum_{i=1}^n \|\mathbf{V}_{i,\bullet}\|_2 \|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{i,\bullet}\|_2 \leq \|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,1}. \tag{2.53}$$

We also deduce

$$\begin{aligned}\sum_{j=1}^p \big(\gamma\|\widehat{\mathbf{\Delta}}^{\mathbf{B}}_{\bullet,j}\|_1 + |\mathbf{V}_{\bullet,j}^\top \widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{\bullet,j}|\big)^2 &\leq \Big(\sum_{j=1}^p \gamma\|\widehat{\mathbf{\Delta}}^{\mathbf{B}}_{\bullet,j}\|_1 + |\mathbf{V}_{\bullet,j}^\top \widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{\bullet,j}|\Big)^2 \\ &\leq \big(\gamma\|\widehat{\mathbf{\Delta}}^{\mathbf{B}}\|_{1,1} + \|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,1}\big)^2. \end{aligned} \tag{2.54}$$

Besides, it holds

$$\begin{aligned} -\sum_{j=1}^p \bar{\boldsymbol{\xi}}_{\bullet,j}^\top \mathbf{M}\widehat{\mathbf{\Delta}}_{\bullet,j} &= \sum_{j=1}^p \big(\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{\bullet,j} - \mathbf{X}^{(n)}\widehat{\mathbf{\Delta}}^{\mathbf{B}}_{\bullet,j}\big)^\top \bar{\boldsymbol{\xi}}_{\bullet,j} \\ &= \sum_{j=1}^p \|\bar{\boldsymbol{\xi}}_{\bullet,j}\|_2 \Big(\sum_{i\in I} \widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{i,j}\frac{\boldsymbol{\epsilon}_{i,j}}{\|\boldsymbol{\epsilon}_{\bullet,j}\|_2} - \widehat{\mathbf{\Delta}}^{\mathbf{B}}_{\bullet,j}{}^\top \mathbf{X}^{(n)\top}_{I,\bullet}\frac{\boldsymbol{\epsilon}_{I,j}}{\|\boldsymbol{\epsilon}_{I,j}\|_2}\Big) \\ &\leq (\max_{j\in[p]}\|\boldsymbol{\xi}_{I,j}\|_2)\Big(\sum_{i\in I}\sum_{j=1}^p \frac{|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{i,j}\boldsymbol{\epsilon}_{i,j}|}{\|\boldsymbol{\epsilon}_{I,j}\|_2} + \sum_{j=1}^p \frac{|\widehat{\mathbf{\Delta}}^{\mathbf{B}}_{\bullet,j}{}^\top \mathbf{X}^{(n)\top}_{I,\bullet}\boldsymbol{\epsilon}_{I,j}|}{\|\boldsymbol{\epsilon}_{I,j}\|_2}\Big), \end{aligned}$$

thus, by the duality inequality $|\widehat{\mathbf{\Delta}}^{\mathbf{B}}_{\bullet,j}{}^\top \mathbf{X}^{(n)\top}_{I,\bullet}\boldsymbol{\epsilon}_{I,j}| \leq \|\widehat{\mathbf{\Delta}}^{\mathbf{B}}_{\bullet,j}\|_1 \|\mathbf{X}^{(n)\top}_{I,\bullet}\boldsymbol{\epsilon}_{I,j}\|_\infty$ and the Cauchy-Schwarz inequality, and as the penalty levels satisfy conditions (2.45), we find

$$\begin{aligned} -\sum_{j=1}^p \bar{\boldsymbol{\xi}}_{\bullet,j}^\top \mathbf{M}\widehat{\mathbf{\Delta}}_{\bullet,j} &\leq \|\boldsymbol{\xi}_{I,\bullet}^\top\|_{2,\infty}\Big(\sum_{i\in I}\|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}_{i,\bullet}\|_2\Big(\sum_{j=1}^p \frac{\boldsymbol{\epsilon}_{i,j}^2}{\|\boldsymbol{\epsilon}_{I,j}\|_2^2}\Big)^{\frac{1}{2}} + \sum_{j=1}^p \|\widehat{\mathbf{\Delta}}^{\mathbf{B}}_{\bullet,j}\|_1 \frac{\|\mathbf{X}^{(n)\top}_{I,\bullet}\boldsymbol{\epsilon}_{I,j}\|_\infty}{\|\boldsymbol{\epsilon}_{I,j}\|_2}\Big) \\ &\leq \lambda\frac{c-1}{c+1}\big(\gamma\|\widehat{\mathbf{\Delta}}^{\mathbf{B}}\|_{1,1} + \|\widehat{\mathbf{\Delta}}^{\mathbf{\Theta}}\|_{2,1}\big)\|\boldsymbol{\xi}_{I,\bullet}^\top\|_{2,\infty}. \end{aligned} \tag{2.55}$$

From inequality (2.51), we get

$$\|\mathbf{M}\widehat{\boldsymbol{\Delta}}_{\bullet,j}\|_2^2 \leq -2\bar{\boldsymbol{\xi}}_{\bullet,j}^\top \mathbf{M}\widehat{\boldsymbol{\Delta}}_{\bullet,j} - 2\lambda\|\bar{\boldsymbol{\xi}}_{\bullet,j}\|_2\Big(\mathbf{V}_{\bullet,j}^\top \widehat{\boldsymbol{\Delta}}_{\bullet,j}^{\boldsymbol{\Theta}} + \gamma\big(\|\widehat{\mathbf{B}}_{j^c,j}\|_1 - \|\mathbf{B}^*_{j^c,j}\|_1\big)\Big) + \lambda^2\big(\gamma\|\widehat{\boldsymbol{\Delta}}_{\bullet,j}^{\mathbf{B}}\|_1 + |\mathbf{V}_{\bullet,j}^\top \widehat{\boldsymbol{\Delta}}_{\bullet,j}^{\boldsymbol{\Theta}}|\big)^2,$$

for every $j \in [p]$. Then, summing over all j and using the triangle inequality, we have

$$\|\mathbf{M}\widehat{\boldsymbol{\Delta}}\|_F^2 \leq -2\sum_{j=1}^p \bar{\boldsymbol{\xi}}_{\bullet,j}^\top \mathbf{M}\widehat{\boldsymbol{\Delta}}_{\bullet,j} + 2\lambda\|\boldsymbol{\xi}_{I,j}\|_2\Big(|\mathbf{V}_{\bullet,j}^\top \widehat{\boldsymbol{\Delta}}_{\bullet,j}^{\boldsymbol{\Theta}}| + \gamma\|\widehat{\boldsymbol{\Delta}}_{\bullet,j}^{\mathbf{B}}\|_1\Big) + \lambda^2\big(\gamma\|\widehat{\boldsymbol{\Delta}}_{\bullet,j}^{\mathbf{B}}\|_1 + |\mathbf{V}_{\bullet,j}^\top \widehat{\boldsymbol{\Delta}}_{\bullet,j}^{\boldsymbol{\Theta}}|\big)^2.$$

Combining the latter with equations (2.53), (2.54) and (2.55), we arrive at

$$\|\mathbf{M}\widehat{\boldsymbol{\Delta}}\|_F^2 \leq \lambda\frac{4c}{c+1}\big(\gamma\|\widehat{\boldsymbol{\Delta}}^{\mathbf{B}}\|_{1,1} + \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{2,1}\big)\|\boldsymbol{\xi}_{I,\bullet}^\top\|_{2,\infty} + \lambda^2\big(\gamma\|\widehat{\boldsymbol{\Delta}}^{\mathbf{B}}\|_{1,1} + \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{2,1}\big)^2,$$

we finally apply Proposition 2.4.14 that gives inequality (2.52). □

Finally, next proposition states the risk bound in the high dimensional settings. Eq. (2.58) indeed corresponds to the claimed inequality (2.21) of Theorem 2.3.1.

Proposition 2.4.18. *Choose $\gamma = 1$ and $\delta \in (0,1)$ such that $n \geq |O| + 16\log(2p/\delta)$ and choose*

$$\lambda = 6\left(\frac{\log(2np/\delta)}{n - |O|}\right)^{1/2}. \tag{2.56}$$

Then

i) with probability at least $1-\delta$, the penalty levels λ and γ satisfy conditions (2.45) for some constant $c = 2$.

ii) If $4\lambda(|\mathcal{J}|^{1/2} + |O|^{1/2}) < \kappa^{1/2}$ holds, then there exists an event $\mathcal{E}_0$ of probability at least $1-2\delta$ such that in[4] $\mathcal{E}_\kappa \cap \mathcal{E}_0$, we have

$$\|\mathbf{M}\widehat{\boldsymbol{\Delta}}\|_{2,2} \leq \frac{C_2}{\sqrt{\kappa}}\max_{j\in[p]}(\omega^*_{jj})^{-1/2}\big(|\mathcal{J}|^{1/2} + |O|^{1/2}\big)\left(\frac{\log(2np/\delta)}{n-|O|}\right)^{1/2}, \tag{2.57}$$

$$\|\widehat{\boldsymbol{\Delta}}^{\mathbf{B}}\|_{1,1} + \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{2,1} \leq \frac{12C_2}{\kappa}\max_{j\in[p]}(\omega^*_{jj})^{-1/2}\big(|\mathcal{J}| + |O|\big)\left(\frac{\log(2np/\delta)}{n-|O|}\right)^{1/2} \tag{2.58}$$

with $C_2 \leq 75$.

Proof. Claim i) of the proposition is obtained by standard arguments relying on tail bounds for Gaussian and χ^2 distributions and the union bound. These arguments are similar to those presented in Section 2.4.3 and, therefore, are skipped.

[4] Recall that $\mathcal{E}_\kappa$ is the event defined by (2.19)

Analogously, using Lemma 2.4.13, we find that with probability at least $1-\delta$, we have $\|\boldsymbol{\xi}_{I,\bullet}\|_{2,\infty} \leq (1+2^{-3/2})\max_j(\omega^*_{jj})^{-1/2}$. We denote by $\mathcal{E}_0$ the intersection of this event with the one of claim i). By the union bound, we have $\mathbf{P}(\mathcal{E}_0) \geq 1-2\delta$. In the rest of this proof, we place ourselves in the event $\mathcal{E}_0 \cap \mathcal{E}_\kappa$. By the compatibility assumption (event $\mathcal{E}_\kappa$), we have

$$\|\widehat{\boldsymbol{\Delta}}^{\mathbf{B}}_{\mathcal{J}}\|_{1,1} \leq \frac{|\mathcal{J}|^{1/2}}{\kappa^{1/2}}\|\mathbf{M}\widehat{\boldsymbol{\Delta}}\|_{2,2} \quad \text{and} \quad \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{O,\bullet}\|_{2,1} \leq \frac{|O|^{1/2}}{\kappa^{1/2}}\|\mathbf{M}\widehat{\boldsymbol{\Delta}}\|_{2,2}. \tag{2.59}$$

Since in the event $\mathcal{E}_0$ the conditions of Lemma 2.4.17 are met, inequality (2.52) readily implies inequality (2.57). On the other hand, we know from Proposition 2.4.14 that $\widehat{\boldsymbol{\Delta}}$ belongs to the dimension reduction cone $\mathscr{C}_{\mathcal{J},O}(2,1)$. Therefore,

$$\|\widehat{\boldsymbol{\Delta}}^{\mathbf{B}}\|_{1,1} + \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}\|_{2,1} \leq 3(\|\widehat{\boldsymbol{\Delta}}^{\mathbf{B}}_{\mathcal{J}}\|_{1,1} + \|\widehat{\boldsymbol{\Delta}}^{\boldsymbol{\Theta}}_{O,\bullet}\|_{2,1}) \leq \frac{3(|\mathcal{J}| \vee |O|)^{1/2}}{\kappa^{1/2}}\|\mathbf{M}\widehat{\boldsymbol{\Delta}}\|_{2,2}.$$

Using the upper bound on $\|\mathbf{M}\widehat{\boldsymbol{\Delta}}\|_{2,2}$ provided by (2.57), we immediately obtain bound (2.58). □

2.5 Algorithmic aspects

In this section, we propose an algorithm that efficiently computes the estimator (2.4) of the precision matrix in the presence of outliers. First, we develop a method that addresses this issue in the moderate dimensional case.

2.5.1 Algorithm in the moderate dimensional case

We start here by reformulating optimization problem (2.9). For this purpose, we remark that for any $j \in [p]$,

$$\min_{\mathbf{t}_j \in \mathbb{R}} \frac{1}{2}\Big(\mathbf{t}_j + \frac{\|(\mathbf{X}^{(n)}\mathbf{B} - \boldsymbol{\Theta})_{\bullet,j}\|_2^2}{\mathbf{t}_j}\Big) = \|(\mathbf{X}^{(n)}\mathbf{B} - \boldsymbol{\Theta})_{\bullet,j}\|_2, \tag{2.60}$$

where the minimum is obtained for $\mathbf{t}_j = \|(\mathbf{X}^{(n)}\mathbf{B} - \boldsymbol{\Theta})_{\bullet,j}\|_2$. We then rewrite the optimization problem using this trick. To this end, we denote the $p \times p$ diagonal matrix whose jth entry is equal to $\mathbf{t}_j^{-1/2}$ by $\mathbf{D}_{\mathbf{t}}^{-1/2}$ and introduce the function

$$g(\mathbf{B}, \boldsymbol{\Theta}, \mathbf{t}) := \frac{1}{2}\sum_{i=1}^{n}\big\|(\mathbf{X}^{(n)}\mathbf{B} - \boldsymbol{\Theta})_{i,\bullet}\mathbf{D}_{\mathbf{t}}^{-1/2}\big\|_2^2 + \lambda\|\boldsymbol{\Theta}_{i,\bullet}\|_2.$$

Problem (2.9) is therefore equivalent to

$$\{\widehat{\mathbf{B}}, \widehat{\mathbf{\Theta}}, \widehat{\mathbf{t}}\} = \underset{\substack{\mathbf{B}\in\mathbb{R}^{p\times p}\\ \mathbf{B}_{jj}=1}}{\arg\min} \min_{\mathbf{\Theta}\in\mathbb{R}^{n\times p}} \min_{\substack{\mathbf{t}\in\mathbb{R}^p\\ \mathbf{t}_j\neq 0}} \left\{ g(\mathbf{B}, \mathbf{\Theta}, \mathbf{t}) + \sum_{j=1}^{p} \frac{1}{2}\mathbf{t}_j \right\}. \tag{2.61}$$

This convex problem can be solved by minimizing separately with respect to the parameters $\mathbf{B}, \mathbf{\Theta}$ and $\mathbf{t}$. We already have computed $\widehat{\mathbf{B}}$ given $\mathbf{\Theta}$ (see Eq. (2.11)) and for $\widehat{\mathbf{t}}$ given $\mathbf{\Theta}$ and $\mathbf{B}$. We further look for an explicit expression $\widehat{\mathbf{\Theta}}$ given $\mathbf{B}$ and $\mathbf{t}$. For the sake of clarity, we consider in the following that the function g depends only on $\mathbf{\Theta}$ and simply note $g(\mathbf{\Theta})$. This function is decomposable and in the sense that $g(\mathbf{\Theta}) = \sum_{i\in[n]} g_i(\mathbf{\Theta}_{i,\bullet})$ with

$$g_i(\boldsymbol{\theta}) := \frac{1}{2}\left\|(\mathbf{X}_{i,\bullet}^{(n)}\mathbf{B} - \boldsymbol{\theta})\mathbf{D}_{\mathbf{t}}^{-1/2}\right\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2.$$

In addition, let $\mathbf{u}_i \in \partial_{\boldsymbol{\theta}} g_i(\widehat{\mathbf{\Theta}}_{i,\bullet})$ be the ith row of $\mathbf{U}$. The Karush-Kuhn-Tucker conditions for problem (2.61), given $\mathbf{B}$ and $\mathbf{t}$, entail that the vector $\mathbf{u}_i$ satisfies $\mathbf{u}_i = 0$. It follows that either $\|\widehat{\mathbf{\Theta}}_{i,\bullet}\|_2 = 0$ or

$$-\mathbf{X}_{i,\bullet}^{(n)}\mathbf{B}\mathbf{D}_{\mathbf{t}}^{-1} + \widehat{\mathbf{\Theta}}_{i,\bullet}\mathbf{D}_{\mathbf{t}}^{-1} + \lambda\frac{\widehat{\mathbf{\Theta}}_{i,\bullet}}{\|\widehat{\mathbf{\Theta}}_{i,\bullet}\|_2} = 0. \tag{2.62}$$

The first alternative corresponds to $\widehat{\mathbf{\Theta}}_{i,\bullet} = 0$. In this case we must have $0 \in \partial_{\boldsymbol{\theta}} g_i(0)$, that is $0 \in \{-\mathbf{X}_{i,\bullet}^{(n)}\mathbf{B}\mathbf{D}_{\mathbf{t}}^{-1} + \lambda\boldsymbol{\theta} | \boldsymbol{\theta} \in \mathbb{R}^p, \|\boldsymbol{\theta}\|_2 \leq 1\}$. It implies that there exists $\boldsymbol{\theta}$ such that $-\mathbf{X}_{i,\bullet}^{(n)}\mathbf{B}\mathbf{D}_{\mathbf{t}}^{-1} + \lambda\boldsymbol{\theta} = 0$, hence that $\|\mathbf{X}_{i,\bullet}^{(n)}\mathbf{B}\mathbf{D}_{\mathbf{t}}^{-1}\|_2 \leq \lambda$. Otherwise, when $\|\mathbf{X}_{i,\bullet}^{(n)}\mathbf{B}\mathbf{D}_{\mathbf{t}}^{-1}\|_2 > \lambda$, we deduce from Eq. (2.62) that

$$\widehat{\mathbf{\Theta}}_{i,\bullet} = \|\widehat{\mathbf{\Theta}}_{i,\bullet}\|_2 \mathbf{X}_{i,\bullet}^{(n)}\mathbf{B}\mathbf{D}_{\mathbf{v}_i}^{-1} \tag{2.63}$$

holds for any $i \in [n]$, where $\mathbf{v}_i = \|\widehat{\mathbf{\Theta}}_{i,\bullet}\|_2 \mathbf{1}_p + \lambda\mathbf{t}$. In this case, we have an explicit solution for the parameter $\mathbf{\Theta}$ of the optimization problem up to a multiplicative constant. To determine this constant, let us introduce a function f defined for any $x \in \mathbb{R}$ by

$$f(x) = \|\mathbf{X}_{i,\bullet}^{(n)}\mathbf{B}\mathbf{D}_{x\mathbf{1}_p+\lambda\mathbf{t}}^{-1}\|_2^2.$$

Taking the Euclidean norm of both sides of Eq. (2.63) we obtain $f(\|\widehat{\mathbf{\Theta}}_{i,\bullet}\|_2) = 1$. We note that the function f is decreasing on $(0, +\infty)$. In particular, $\lim_{x\to+\infty} f(x) = 0$ and $\|\mathbf{X}_{i,\bullet}^{(n)}\mathbf{B}\mathbf{D}_{\mathbf{t}}^{-1}\|_2 > \lambda$ entails that $f(0) > 1$. Moreover, putting f in an expanded form,

$$f(x) = \sum_{j=1}^{p} \left(\frac{\mathbf{X}_{i,\bullet}^{(n)}\mathbf{B}_{\bullet,j}}{x + \lambda\mathbf{t}_j}\right)^2,$$

we get a tractable expression for its derivative f',

$$f'(x) = -2\|\mathbf{X}_{i,\bullet}^{(n)}\mathbf{B}\mathbf{D}_{x\mathbf{1}_p+\lambda\mathbf{t}}^{-3/2}\|_2^2.$$

Starting from $x = 0$, we can apply the method of Newton to obtain a numerical solution to the equation $f(x) = 1$. At each iteration h of the method of Newton, the current value x^h of x is updated in the following way

$$x^{h+1} = x^h + \frac{\|\mathbf{X}_{i,\bullet}^{(n)}\mathbf{B}\mathbf{D}_{x^h\mathbf{1}_p+\lambda\mathbf{t}}^{-1}\|_2^2 - 1}{2\|\mathbf{X}_{i,\bullet}^{(n)}\mathbf{B}\mathbf{D}_{x^h\mathbf{1}_p+\lambda\mathbf{t}}^{-3/2}\|_2^2},$$

until $|f(x^h)-1| < \epsilon$, where ϵ is a threshold depending on the desired accuracy, for example $\epsilon = 10^{-6}$. In conclusion, we propose to solve the optimization problem (2.9) by updating successively the three parameters $\mathbf{B}, \mathbf{t}$ and $\mathbf{\Theta}$. We also introduce a new parameter $\mathbf{x}$ that represents the Euclidean norm of the rows of $\mathbf{\Theta}$. Each component $\mathbf{x}_i$ is estimated as explained above. Note that to compute $\mathbf{x}$, we only need $\mathbf{B}$ and $\mathbf{t}$, not $\mathbf{\Theta}$. We denote the estimated values of the parameters at step k by $\mathbf{B}^k$, $\mathbf{\Theta}^k$, $\mathbf{t}^k$, $\mathbf{x}^k$ and $\mathbf{v}^k$. At each iteration, we perform the following update operations

$$\begin{aligned}
&\mathbf{B}_{j^c,j}^k = -\big(\mathbf{X}_{\bullet,j^c}^{(n)\top}\mathbf{X}_{\bullet,j^c}^{(n)}\big)^\dagger\mathbf{X}_{\bullet,j^c}^{(n)\top}(\mathbf{X}_{\bullet,j}^{(n)} - \mathbf{\Theta}_{\bullet,j}^{k-1}) \quad \text{and} \quad \mathbf{B}_{j,j}^k = 1, \text{ for any } j,\\
&\mathbf{t}_j^k = \|(\mathbf{X}^{(n)}\mathbf{B}^k - \mathbf{\Theta}^{k-1})_{\bullet,j}\|_2, \text{ for any } j,\\
&\text{update } \mathbf{x}_i^k, \text{ for any } i, \text{ by the method of Newton},\\
&\mathbf{v}_i^k = \mathbf{x}_i^k\mathbf{1}_p + \lambda\mathbf{t}^k, \text{ for any } i,\\
&\mathbf{\Theta}_{i,\bullet}^k = \mathbf{x}_i^k\mathbf{X}_{i,\bullet}^{(n)}\mathbf{B}^k\mathbf{D}_{\mathbf{v}_i^k}^{-1}, \text{ for any } i.
\end{aligned}$$

Thus, our optimization problem can be solved in a computationally efficient way by repeating the operations stated above. The summarized resolution procedure is described in Algorithm 2.1.

As the objective function is convex, our greedy procedure causes the successive estimates to move nearer to the minimum at each iteration, after a certain step. The stopping criterion is defined using the ratio of the variation of the cost function. The objective function is considered to be stabilized when the difference between the value of the objective function at the precedent step and at the current one, divided by its value at the current step, is below a given threshold. To be precise, denoting the objective function by $F(\mathbf{B}, \mathbf{\Theta}) = g(\mathbf{B}, \mathbf{\Theta}, \mathbf{t}) + \sum_{j=1}^p \mathbf{t}_j/2$, and the threshold by T, the stopping criterion at the step k is given by

$$\left|\frac{F(\mathbf{B}^{k-1}, \mathbf{\Theta}^{k-1}) - F(\mathbf{B}^k, \mathbf{\Theta}^k)}{F(\mathbf{B}^k, \mathbf{\Theta}^k)}\right| < T.$$

Algorithm 2.1: Estimation of $(\mathbf{B}, \mathbf{\Theta})$ by solving optimization problem (2.9)

Input: matrix $\mathbf{X}$ of observations.
penalty level λ.

Output: $\widehat{\mathbf{B}}$ the estimate of the matrix of regression coefficients.
$\widehat{\mathbf{\Theta}}$ the estimate of the matrix of outliers.

1: initialize $\mathbf{\Theta}$ to the null matrix.

repeat

2: for the current value of $\mathbf{\Theta}$, compute $\mathbf{B}_{\bullet,j}$ for each $j \in [p]$.

3: for the current values of $\mathbf{\Theta}$ and $\mathbf{B}$, update $\mathbf{t}_j$ for each $j \in [p]$.

4: for the current value of $\mathbf{\Theta}$, compute $\mathbf{x}_i$ for each $i \in [n]$.

5: for the current values of $\mathbf{B}$ and $\mathbf{t}$, compute a new estimate for each row of $\mathbf{\Theta}$.

until *the change in the value of the objective function falls below a certain threshold*

We further note that no outlier is detected when the penalty level λ is larger than

$$\max_{i\in[n]} \left(\sum_{j\in[p]} \frac{(\mathbf{Z}^j_{i,\bullet}\mathbf{X}^{(n)}_{\bullet,j})^2}{\|\mathbf{Z}^j\mathbf{X}^{(n)}_{\bullet,j}\|_2^2} \right)^{1/2}. \tag{2.64}$$

Indeed, to have $\widehat{\mathbf{\Theta}} = 0$, the null vector must belong to the sub-differential of the objective function evaluated at zero. We already show that for any $i \in [n]$ it entails that $\lambda \geq \|\mathbf{X}^{(n)}_{i,\bullet}\mathbf{B}\mathbf{D}^{-1}_{\mathbf{t}}\|_2$. Using that $\mathbf{t}_j = \|(\mathbf{X}^{(n)}\mathbf{B} - \mathbf{\Theta})_{\bullet,j}\|_2$, we arrive at

$$\lambda \geq \max_{i\in[n]} \left(\sum_{j\in[p]} \frac{(\mathbf{Z}^j_{i,\bullet}\mathbf{X}^{(n)}_{\bullet,j} + \mathbf{\Pi}^{j^c}\mathbf{\Theta}_{\bullet,j})^2}{\|\mathbf{Z}^j\mathbf{X}^{(n)}_{\bullet,j} + \mathbf{Z}^j\mathbf{\Theta}_{\bullet,j}\|_2^2} \right)^{1/2}.$$

Thus, in $\mathbf{\Theta} = 0$, we end with condition (2.64).

2.5.2 Algorithm in the high dimensional case

When $p \gg n$, we consider optimization problem (2.4). This problem can be solved in the same way as in moderate dimensional case using alternating minimization techniques. The only difference is that the successive estimations of the columns of the matrix $\mathbf{B}^*$ are no longer obtained by least squares optimization, but by the resolution of a square-root Lasso problem. Indeed, for a given $\mathbf{\Theta}$ and assuming that $\mathbf{c} = 0$, the problem (2.4) reduces to

$$\widehat{\mathbf{B}} = \underset{\substack{\mathbf{B}\in\mathbb{R}^{p\times p}\\ \mathbf{B}_{jj}=1}}{\arg\min} \left\{ \left\|(\mathbf{X}^{(n)}\mathbf{B} - \mathbf{\Theta})^\top\right\|_{2,1} + \lambda\gamma\|\mathbf{B}\|_{1,1} \right\}. \tag{2.65}$$

The latter is then equivalent solving p independent square-root Lasso problems. Indeed for each $j \in [p]$, we have

$$\widehat{\mathbf{B}}_{j^c,j} = \underset{\mathbf{b}\in\mathbb{R}^{p-1}}{\arg\min} \left\{ \left\| \mathbf{X}^{(n)}_{\bullet,j^c}\mathbf{b} + \mathbf{X}^{(n)}_{\bullet,j} - \mathbf{\Theta}_{\bullet,j} \right\|_2 + \lambda\gamma\|\mathbf{b}\|_1 \right\}. \tag{2.66}$$

Each of these square-root Lasso problems can be solved by a standard convex optimizer such that SCS [O'Donoghue et al., 2013], Gurobi [Gurobi Optimization, 2015] or Mosek [Andersen and Andersen, 2000], or, as already noticed in Chapter 1, using the coordinate descent algorithm. Ultimately, it is sufficient to adapt the first step of the exterior loop of Algorithm 2.1 according to what precedes. The objective function should also be modified accordingly: $F(\mathbf{B}, \mathbf{\Theta}) = g(\mathbf{B}, \mathbf{\Theta}, \mathbf{t}) + \sum_{j=1}^{p} \mathbf{t}_j/2 + \lambda\gamma\|\mathbf{B}\|_{1,1}$.

2.6 Empirical evaluation

In this section, we report the results of some numerical experiments performed on synthetic data. The main goal of this part is to demonstrate the potential of the method based on Eq. (2.4) and (2.5). To this end, we have considered several scenarios and in each of them compared our method with several other competitors. In order to provide a fair comparison independent of the delicate question of choosing the tuning parameter, the results of all the methods are reported for the oracle values of the tuning parameters chosen from a grid by minimizing the distance to the true precision matrix. We have used the coordinate descent algorithm for solving the convex optimization problem of Eq. (2.4).

2.6.1 Structure of the precision matrix

Let us first describe the precision matrices used in our experiments. It is worthwhile to underline here that all the precision matrices are normalized in such a way that all the diagonal entries of the corresponding covariance matrix $\mathbf{\Sigma}^* = (\mathbf{\Omega}^*)^{-1}$ are equal to one. To this end, we first define a $p \times p$ positive semidefinite matrix $\mathbf{A}$ and then set $\mathbf{\Omega}^* = (\mathrm{diag}(\mathbf{A}^{-1}))^{\frac{1}{2}}\mathbf{A}(\mathrm{diag}(\mathbf{A}^{-1}))^{\frac{1}{2}}$. The matrices $\mathbf{A}$ used in the five models for which the experiments are carried out are defined as follows.

Model 0: $\mathbf{A}$ is the identity matrix.

Model 1: $\mathbf{A}$ is a Toeplitz matrix with the entries $\mathbf{A}_{ij} = 0.6^{|i-j|}$ for any $i, j \in [p]$.

Model 2: We start by defining a $p \times p$ pentadiagonal matrix with the entries

$$\bar{\mathbf{A}}_{ij} = \begin{cases} 1 & , \text{ for } |i-j| = 0, \\ -1/3 & , \text{ for } |i-j| = 1, \\ -1/10 & , \text{ for } |i-j| = 2, \\ 0 & , \text{ otherwise.} \end{cases}$$

Then, we denote by $\mathbf{A}$ the matrix with the entries $\mathbf{A}_{ij} = (\bar{\mathbf{A}}^{-1})_{ij} \mathbb{1}(|i-j| \leq 2)$. One can check that the matrix $\mathbf{A}$ defined in such a way is positive semidefinite.

Model 3: We set $\mathbf{A}_{ij} = 0$ for all the off-diagonal entries that are neither on the first row nor on the first column of $\mathbf{A}$. The diagonal entries of $\mathbf{A}$ are

$$\mathbf{A}_{11} = p, \qquad \mathbf{A}_{ii} = 2, \quad \text{for any} \quad i \in \{2, \ldots, p\},$$

whereas the off-diagonal entries located either on the first row or on the first column are $\mathbf{A}_{1i} = \mathbf{A}_{i1} = \sqrt{2}$ for $i \in \{2, \ldots, p\}$.

Model 4: The diagonal entries of $\mathbf{A}$ are all equal to 1. Besides, we set $\mathbf{A}_{ij} = 0.5$ for any $i \neq j$.

2.6.2 Contamination scheme and measure of quality

The positions of outliers were chosen by a simple random sampling without replacement. The proportion of outliers, $\epsilon = |O|/n$, used in our experiments varies between 0% and 30%. The entries of the rows of $\mathbf{X}$ corresponding to outliers were drawn randomly from a standard Gaussian distribution and independently of one another. The rows of $\mathbf{X}$ corresponding to inliers are drawn from a zero mean Gaussian distribution with the precision matrix specified by one of the foregoing models. Note that the magnitude of the individual entries of outliers are similar to those of the inliers, which makes the outliers particularly hard to detect.

We measure the distance between the true precision matrix of a multivariate normal distribution and its estimator using the distance induced by the Frobenius norm. Recall that our method does not guarantee the positive definiteness of the estimate of the precision matrix. When the estimate is not positive definite, one can always get a valid precision matrix from $\widehat{\mathbf{\Omega}}$. A number of methods have been proposed in the literature for adjusting a matrix such that it is positive definite. In practice, replacing $\widehat{\mathbf{\Omega}}$ by the positive definite matrix obtained by the approach of Higham [2002], seems to be a good choice as it does not significantly affect the norm-induced distance between the true precision matrix and its estimate.

We also measure the ability to recover the structure of the precision matrix. To this end, we compute the false positive – the proportion of zero entries of the precision matrix that are

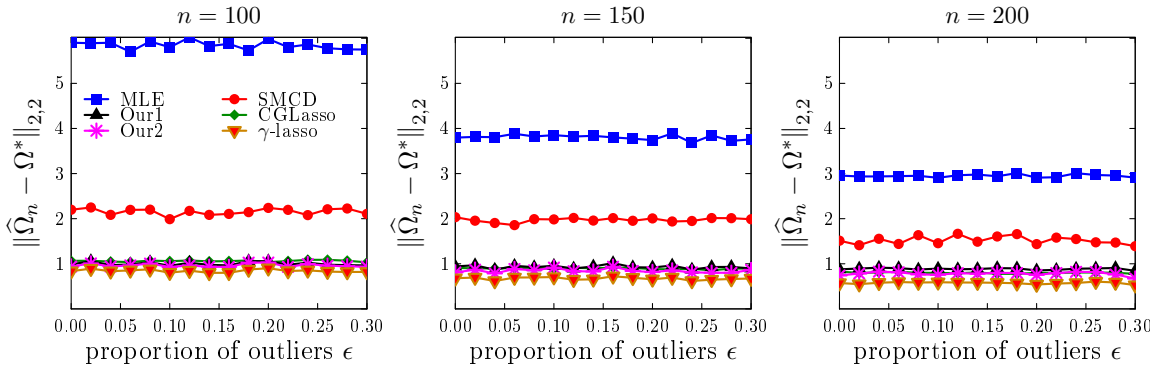


Figure 2.1: The average error (measured in Frobenius norm) of estimating $\mathbf{\Omega}^*$ in Model 0 for p = 30, when ϵ is between 0% and 30%. Each point is the average of 50 replications.

estimated to be nonzero – and false negative rates – the proportion of nonzero off-diagonal entries of the precision matrix that are estimated to be zero – up to a given threshold. In other terms, if we refer to the underlying graph of relationships between variables, the false positive, resp. false negative, rate corresponds to the proportion of wrongly placed edges, resp. wrongly removed. Denoting by $\mathbf{\Psi} = (\text{diag}(\mathbf{\Omega}))^{-1/2}\mathbf{\Omega}(\text{diag}(\mathbf{\Omega}))^{-1/2}$ the matrix of partial correlations associated to the precision matrix $\mathbf{\Omega}$ and by $t > 0$ the chosen threshold, we define

$$\text{FP}_{\widehat{\mathbf{\Psi}}}(t) = |\{j \neq j' : \psi^*_{jj'} = 0, |\widehat{\psi}_{jj'}| > t\}|/|\{(jj') : \psi^*_{jj'} = 0\}|,$$
$$\text{FN}_{\widehat{\mathbf{\Psi}}}(t) = |\{j \neq j' : \psi^*_{jj'} \neq 0, |\widehat{\psi}_{jj'}| < t\}|/|\{j \neq j' : \psi^*_{jj'} \neq 0\}|.$$

model	0		1		2		3		4	
Error type	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
SMCD	0.973		0.989	0.013	0.987	**0.000**	0.990	0.003		0.016
MLE	0.990		0.994	**0.007**	0.992	**0.000**	0.989	0.003		**0.000**
CGLASSO	0.126		0.922	0.070	0.897	**0.000**	0.908	0.072		0.023
Our	0.150		**0.699**	0.215	**0.379**	**0.000**	**0.420**	0.145		0.005
γ-LASSO	**0.000**		0.952	0.029	0.941	**0.000**	0.954	**0.002**		0.002

Table 2.1: Sparsity pattern recovery : false positive and false negative rates (the smaller, the better) of the estimators of the precision matrix for $p = 30$, $n = 200$, when ϵ is equal to 10%, and with a threshold equal to 10^{-3}. The number of replications in each case is $R = 50$.

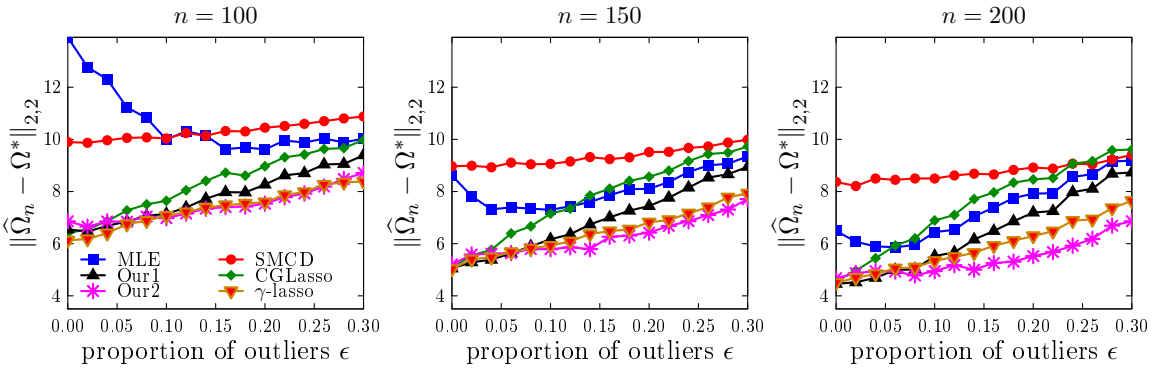


Figure 2.2: The average error (measured in Frobenius norm) of estimating $\mathbf{\Omega}^*$ in Model 1 for p = 30, when ϵ is between 0% and 30%. Each point is the average of 50 replications.

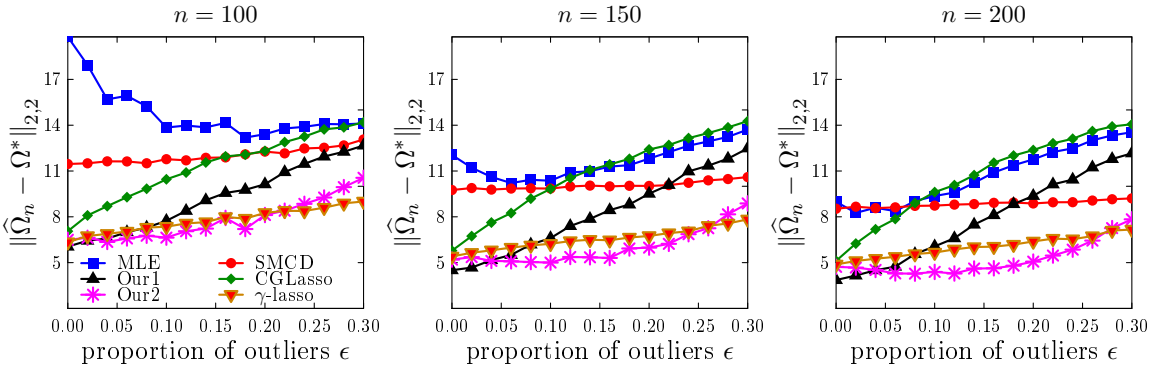


Figure 2.3: The average error (measured in Frobenius norm) of estimating $\mathbf{\Omega}^*$ in Model 2 for p = 30, when ϵ is between 0% and 30%. Each point is the average of 50 replications.

2.6.3 Precision matrix estimators

We have compared our method to four other estimators of the precision matrix. The first and the most naive estimator, referred to as the MLE, consists in computing the (pseudo-)inverse of the empirical covariance matrix.

The second estimator is the inverse of a robust covariance estimate computed by the minimum covariance determinant (MCD) method introduced in [Rousseeuw, 1984]. We have used a shrinkage coefficient coming from the improvement of the Ledoit-Wolf shrinkage [Ledoit and Wolf, 2004], developed by Chen et al. [2010] for multivariate Gaussian distributions. We therefore refined the MCD estimator using the covariance Oracle Shrinkage Approximating (OAS) estimator. In the following, we refer to it as SMCD. We also did experiments estimating the covariance matrix by the minimum volume ellipsoid (MVE) estimator [Rousseeuw, 1985] and by the scaled Kendall's tau estimator [Chen et al., 2015a]. The results obtained for the latter estimators are not reported as they showed no improvement over the SMCD.

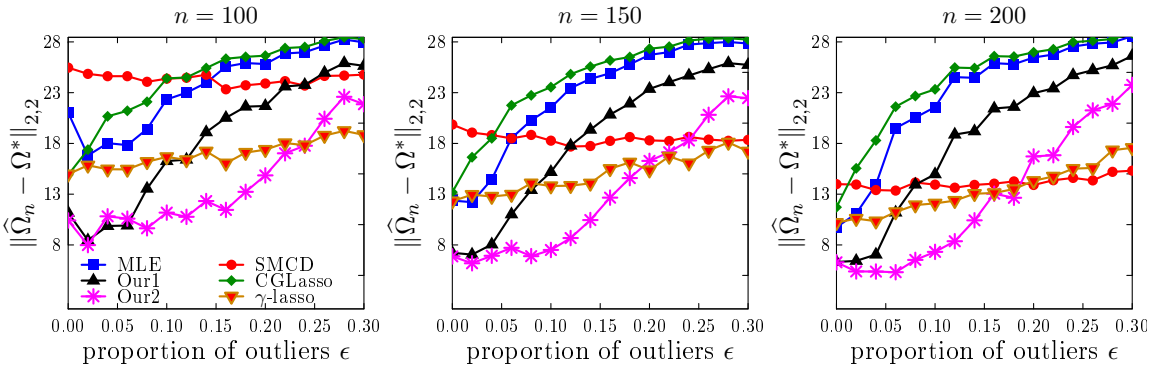


Figure 2.4: The average error (measured in Frobenius norm) of estimating $\mathbf{\Omega}^*$ in Model 3 for p = 30, when ϵ is between 0% and 30%. Each point is the average of 50 replications.

The third estimator of the precision matrix is obtained by solving an optimization problem whose cost function depends on a robust estimate of the covariance matrix. Two versions of this approach are particularly interesting: the maximum log-likelihood with ℓ_1-penalization known as graphical Lasso [Banerjee et al., 2008; d'Aspremont et al., 2008; Friedman et al., 2008] and the constrained ℓ_1-minimization for inverse matrix estimation (Clime) of Cai et al. [2011]. Robust versions of these estimators have been proposed by Öllerer and Croux [2015] and Tarr et al. [2016] and further investigated by Loh and Tan [2015]. In this approach, robust estimates of the covariance matrix are plugged-in the graphical Lasso or Clime estimators. In our experiments, the quality of these two versions were comparable. Therefore, we report only the results for the version based on the graphical Lasso. In [Öllerer and Croux, 2015], the authors proposed an enhancement that simplifies the estimator and reduces the computational cost, by estimating aside the variances and the correlations. Following their work, we choose to estimate the correlations by the robust Gaussian rank correlation [Boudt et al., 2012] and adopt their implementation choices. In particular, as a robust measure of scale, we used the Q_n estimator of Rousseeuw and Croux [1993] that is an alternative to the median absolute deviation (MAD). To sum up, we implemented the correlation based precision matrix estimator obtained by plugging-in the covariance matrix estimate based on pairwise correlations in the graphical Lasso[5] (hereinafter referred to as CGLASSO).

The fourth estimator used in our experiments is the γ-LASSO proposed by Hirose and Fujisawa [2015]. The crux of the method is the replacement of the penalized negative log-likelihood function by the penalized negative γ-likelihood function [Fujisawa and Eguchi, 2008; Cichocki and Amari, 2010]. We used the R package rsggm developed by Hirose and Fujisawa [2015].

Finally we considered two version of our approach, referred to as Our1 and Our2. The

[5]We used the implementation of the graphical Lasso of the R package huge.

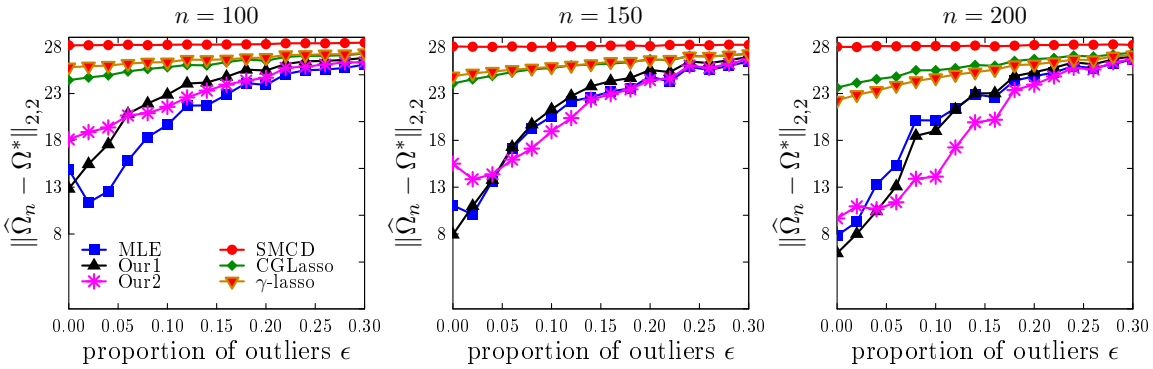


Figure 2.5: The average error (measured in Frobenius norm) of estimating $\mathbf{\Omega}^*$ in Model 4 for p = 30, when ϵ is between 0% and 30%. Each point is the average of 50 replications.

first version merely provided by (2.4) and (2.5), while the second version consists in re-estimating the precision matrix after removing the observations classified as outliers[6]. In the latter, we did not look after outliers, but still estimated a sparse precision matrix.

2.6.4 Results

The results of our experiments are depicted in Figures 2.1-2.1. In all the experiments, the the dimension p is equal to 30 and the contamination rate, denoted by ϵ, is between 0% and 30%. The results show that our procedure is competitive with the state-of-the-art robust estimators of the precision matrix, even when the proportion of outliers is high.

One may observe that the step of re-estimation of the precision matrix after the removal of the observations classified as outliers reduces the error of estimation in all the considered situations. We would also like to mention that the γ-Lasso, which has a highly competitive statistical accuracy is defined as the minimizer of a non-convex cost function. Furthermore, there is no theoretical guarantee ensuring the convergence of the algorithm or controlling its statistical error.

In addition, we compare the methods in terms of computational time. As shown in Table 2.2, the execution time of our method is the lowest. The results obtained for other models are comparable, thus not reported in the manuscript.

Figures 2.6-2.10 present the performance of the estimators when the sample size n is small in comparison with the dimension p. The MLE results are not reported as far worse than the others. The SMCD is essentially interesting in low-dimension, partly due to its huge relative computational cost when p grows, but also because it becomes less robust when p increases (and is not defined for $p \geq n$).

[6]The R package DESP has been extended to cope with outlying observations.

p	n	SMCD	MLE	CGLASSO	Our	γ-LASSO
10	100	0.212	0.356	0.517	**0.002**	0.014
	150	0.307	0.356	0.434	**0.002**	0.015
	200	0.398	0.356	0.433	**0.002**	0.016
30	100	0.766	0.357	0.445	**0.007**	0.028
	150	1.295	0.360	0.449	**0.009**	0.046
	200	1.711	0.360	0.468	**0.017**	0.060
60	100	2.131	0.465	0.458	**0.021**	0.058
	150	3.080	0.395	0.477	**0.065**	0.172
	200	5.593	0.382	0.511	**0.065**	0.277

Table 2.2: Average computation times of the different methods for Model 2, when ϵ is equal to 10%. The number of replications in each case is $R = 50$. The times are measured in seconds and the experiments are done using a 64-bit computer with an Intel i7 quad-core processor and 8GB memory.

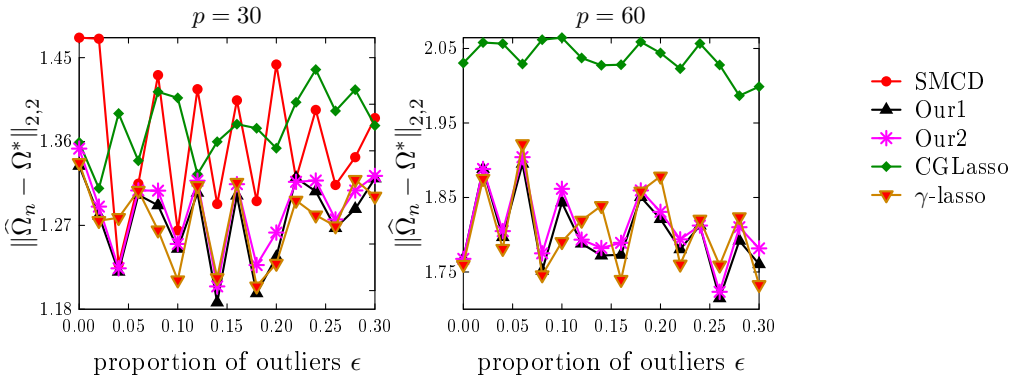


Figure 2.6: The average error (measured in Frobenius norm) of estimating $\mathbf{\Omega}^*$ in Model 0 for n = 50, when ϵ is between 0% and 30%. Each point is the average of 50 replications. Note that the scale of the y-axis is not the same in both plots.

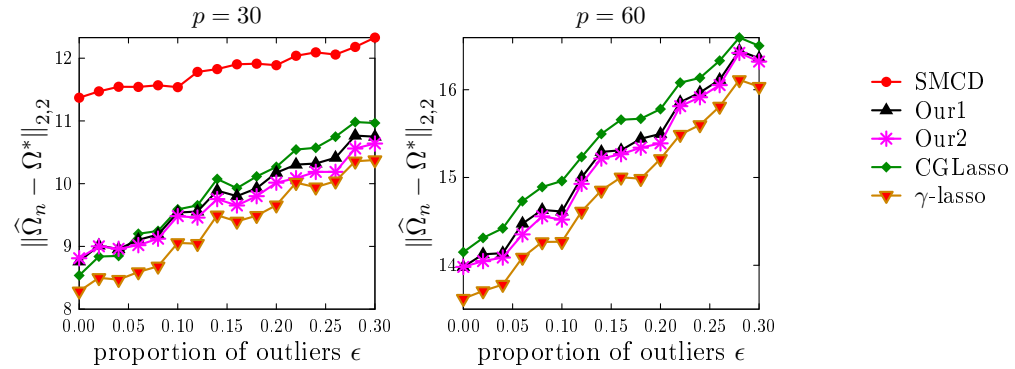


Figure 2.7: The average error (measured in Frobenius norm) of estimating $\mathbf{\Omega}^*$ in Model 1 for n = 50, when ϵ is between 0% and 30%. Each point is the average of 50 replications. Note that the scale of the y-axis is not the same in both plots.

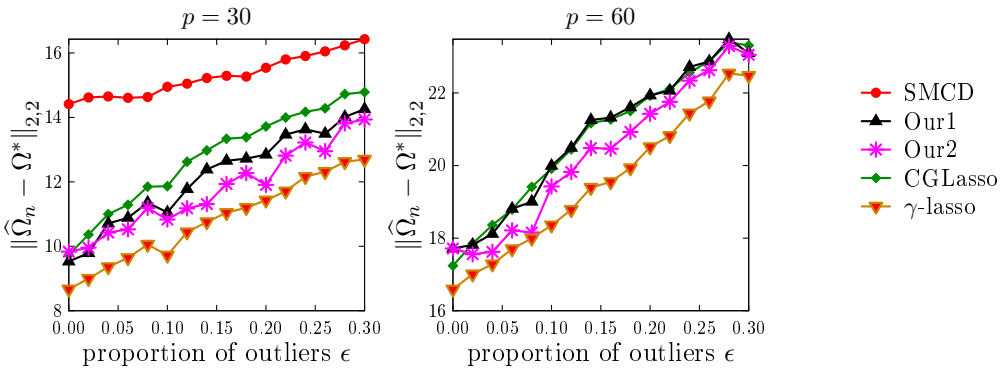


Figure 2.8: The average error (measured in Frobenius norm) of estimating $\mathbf{\Omega}^*$ in Model 2 for n = 50, when ϵ is between 0% and 30%. Each point is the average of 50 replications. Note that the scale of the y-axis is not the same in both plots.

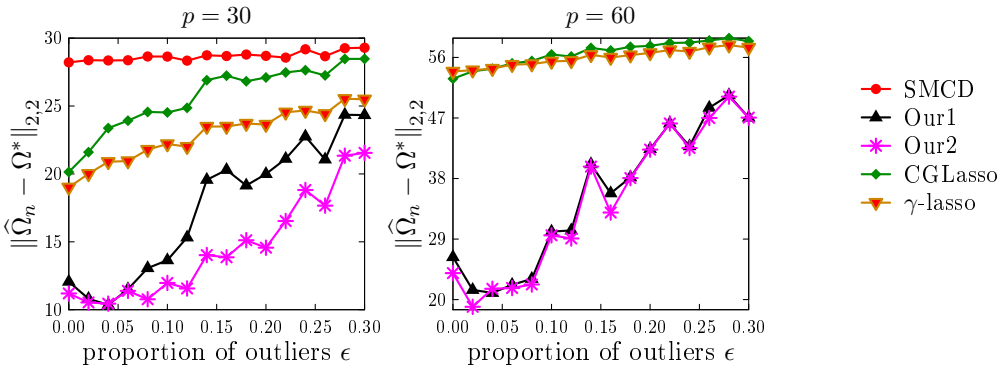


Figure 2.9: The average error (measured in Frobenius norm) of estimating $\mathbf{\Omega}^*$ in Model 3 for n = 50, when ϵ is between 0% and 30%. Each point is the average of 50 replications. Note that the scale of the y-axis is not the same in both plots.

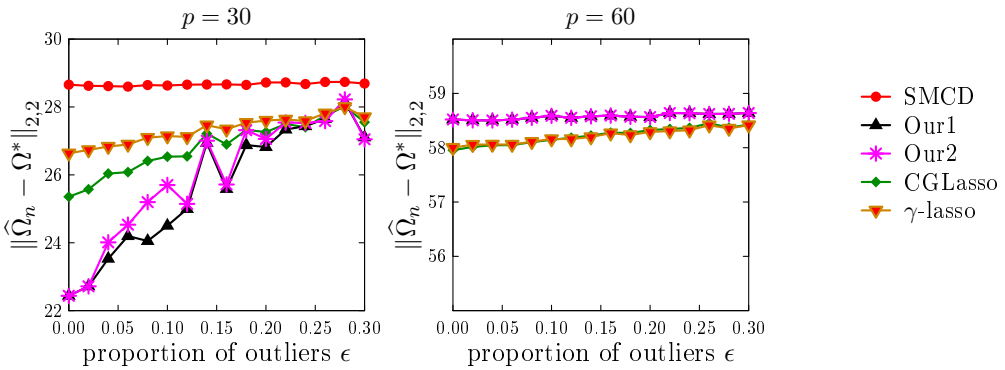


Figure 2.10: The average error (measured in Frobenius norm) of estimating $\mathbf{\Omega}^*$ in Model 4 for n = 50, when ϵ is between 0% and 30%. Each point is the average of 50 replications. Note that the scale of the y-axis is not the same in both plots.

2.7 Perspectives

Calibrating the tuning parameters λ and γ of problem (2.4) is not an easy question. In our experiments, we choose the values of these parameters from a grid. These are indeed oracle values, not accessible in real situations. In Theorems 2.2.1, 2.2.2 and 2.3.1, we provide theoretical values for λ that are suitable to prove estimation consistency. As for the penalization parameter that promotes sparsity, theoretical concerns advocate for choosing $\gamma = \sqrt{2 \log p} \,/\, \lambda$. This choice is guided by the connection between our method and the square-root Lasso. Indeed, when the observations are not corrupted by outliers, $\mathbf{\Theta} = 0$ and problem (2.4) simplifies to problem (2.2), with $\bar{\lambda} = \sqrt{2 \log p}$. This value is known as the universal choice for the square-root Lasso, it gives optimal theoretical results [Dalalyan and Chen, 2012; Sun and Zhang, 2012; Dalalyan et al., 2013]. As we noticed in Section 2.1.2, the square-root Lasso is designed to tackle the problem of heterogeneous variances of the noise terms in the regression model. However, this method is not fully adaptive, and common techniques like cross-validation, bootstrap or criteria like AIC or BIC can still be useful to select the best tuning parameters. From a computational point of view, adjusting and implementing the screening techniques developed in [Fercoq et al., 2015; Ndiaye et al., 2015] should allow to consider more candidate values for the tuning parameters without impeding too much on execution time. Note that Lederer and Müller [2014] proposed an original approach for variable selection in the linear regression framework, that has the advantage of being tuning-free. However, since their estimator is formulated as a non-convex optimization problem, their algorithm carries no assurance that it converges. To the best of our knowledge, a completely data-driven procedure for regularized regression still not exists. Again referring to the tuning parameters, in our method, the penalty level is the same for all the p underlying square-root Lasso problems (2.66). There is no reason in general to expect similarities in the sparsity pattern across rows/columns of $\mathbf{\Omega}^*$. In particular, the number of nonzero entries is usually not the same over rows/columns. Therefore, replacing $\gamma \|\mathbf{B}\|_{1,1}$ by $\sum_{j \in [p]} \gamma_j \|\mathbf{B}_{\bullet,j}\|_1$ may be interesting, but we end with the issue of selecting a vector of tuning parameters instead of a scalar.

Conclusion

Résumé. Pour conclure, l'estimateur robuste de la matrice de précision que nous avons construit en partant du modèle de régression linéaire offre des garanties théoriques satisfaisantes. D'une part, on a étudié plusieurs estimateurs des éléments diagonaux dont on peut évaluer le risque d'estimation. D'autre part, l'analyse de l'estimateur robuste nous a permis d'obtenir des vitesses de convergence optimales au sens minimax pour l'estimateur de la matrice des erreurs et celui de la matrice des coefficients de régression. Cependant, certaines questions n'ont pas encore de réponse. Il reste par exemple à déterminer si l'estimateur de la matrice de précision est optimal et à prouver des bornes probabilistes pour cet estimateur robuste.

The estimation of large precision matrices is a rather recent topic of interest. It is indeed connected with the growth of data collection that has generated new needs. As we pinpointed in the Introduction, a wide range of problems have a solution that can be formulated simply as soon as an accurate estimator of the precision matrix is available. However, the framework of analysis, parsimony in our case, the tools and techniques involved, like regularization, partly preexisted this recent concern.

The close ties between the question of estimating the precision matrix and the linear regression model led us to analyze apart the estimators of diagonal entries and those of off-diagonal ones. This estimation in two steps is not merely an artifact of the chosen approach. These two components of the precision matrix cover in fact different meanings. Thus, whereas the off-diagonal elements correspond to a certain type of relationships between the variables, the diagonal entries are for their part more directly associated to the scale of the matrix, that is the magnitude of its individual entries.

The estimators studied in Chapter 1 refer to a simpler model than the one considered in Chapter 2. Nevertheless, the obtained results concerning the diagonal elements still hold for this last model. As we noticed in Chapter 2, the robust estimator provided by the optimization problem (2.4) allows to estimate consistently—on a particular event—the matrix of the coefficients of regression. Indeed, it converges towards $\mathbf{B}^*$ with an optimal convergence rate in element-wise ℓ_1-norm, provided that the number of outliers is not too large compared to the number of nonzero elements of the precision matrix. We can thus plug this estimator $\widehat{\mathbf{B}}$ into the estimators proposed in Chapter 1. Only the estimator based on average absolute deviation has been examined from a theoretical point of view in Chapter 2. Despite this, it should be noted that it does not mean that getting comparable results while considering the residual variance estimator or the MLE is impossible.

Constructing new estimators and observing that they perform well in some experimental settings is interesting, but not enough. Analyzing theoretically an estimator allows to understand under what conditions it is accurate and efficient, and what are its inherent limitations. This comprehension is necessary to use appropriately this estimator and to be able to improve it afterwards. Therefore, even if some of our assumptions could appear somewhat restrictive, we are convinced that our results constitute an important step towards estimating large precision matrices. Besides, all our hypotheses have not the same strength. For instance, although focus is restricted to Gaussian distributions, most of the results we proved also hold true in the sub-Gaussian setting. We recall that our analysis of the estimators relies on recent developments in probability theory and that some results are decisive. The results established in [Laurent and Massart, 2000] or [Vershynin, 2012b], for instance, bring major arguments in many proofs. As another example, we have already mentioned that assuming that the outliers have bounded norms might not be too severe as excluding rough outliers is not sticky. On the other hand, the results of Chapter 1 rest on the basic assumption that $\mathbf{B}^*$ is known without error. Providing the quadratic risk of

the considered estimators as a function of the error of estimation made on $\mathbf{B}^*$ would be undoubtedly more valuable but is also much more complicated.

As for the future developments, we recall that some important issues remain unsettled. First, whether or not the convergence rate established in Theorem 2.2.2 is minimax optimal over a suitably conditioned class of matrices and when outliers have bounded norms still is an open question. Second, the risk bound stated in Theorem 2.3.1 depends on an event whose probability is expected, but yet unproven, to be close to one. In addition, extending our framework to consider noisy data might be interesting. As such, the approach of Belloni et al. [2014b]; Rosenbaum and Tsybakov [2013] is an attractive lead to follow, by adding a random error term to the observed data matrix.

We have already cited recent works dealing as well with robust precision matrix estimation. The approach based on the graphical Lasso coupled with a robust estimator of the covariance matrix has received the more attention at the moment [Loh and Tan, 2015; Öllerer and Croux, 2015; Tarr et al., 2016]. The other method that relies on γ-divergence seems to be promising, at least empirically. However, the theoretical analysis of these approaches should be further explored. This would help to understand the extent to which the resulting estimators differ and what are their respective advantages.

Finally, we expect to achieve satisfactory results when using this estimator in a naive Bayes classifier for image classification purpose. In this context, we indeed have the conviction that this approach could be a relevant alternative to kernel classifiers like the standard SVM.

Appendix A

Supplementary proofs

RÉSUMÉ. Afin que ce manuscrit soit complet, cette annexe regroupe les preuves de résultats énoncés dans le chapitre introductif et dans le chapitre 1. En particulier, nous exposons la manière dont peuvent être obtenues des bornes supérieures de la vitesse de convergence de l'estimateur *square-root Lasso* en norme ℓ_1 et en norme euclidienne. Nous justifions également les bornes non asymptotiques du risque d'estimation de la matrice de précision lorsque cette dernière est estimée en résolvant p problèmes indépendants de type *square-root Lasso*.

A.1 Proofs of Introduction

A.1.1 Conditional independence and precision matrix structure under Gaussian assumption

We first state two well-known results from linear algebra.

Theorem A.1.1 (normal correlations [Marsaglia, 1964]). *Let Y and X be two random vectors such that $(Y^\top, X^\top)^\top$ is distributed according to $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with*

$$\boldsymbol{\mu} = \mathbf{E}\begin{pmatrix} Y \\ X \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_X \end{pmatrix}, \quad \boldsymbol{\Sigma} = \mathbf{C}ov\left(\begin{pmatrix} Y \\ X \end{pmatrix}, \begin{pmatrix} Y \\ X \end{pmatrix}\right) = \begin{pmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{pmatrix}.$$

Then the conditional expectation and covariance of Y given X are

$$\begin{aligned} \mathbf{E}(Y|X) &= \boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^\dagger(X - \boldsymbol{\mu}_X), \\ \mathbf{C}ov(Y, Y|X) &= \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^\dagger\boldsymbol{\Sigma}_{XY}. \end{aligned}$$

Proof. For the proof, see for example [Liptser and Shiryaev, 2013]. □

Theorem A.1.2 (block matrix inversion). *Let $\mathbf{M}$ be a $p \times p$ matrix written by blocks, such that $\mathbf{M} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix}$ then*

$$\mathbf{M}^\dagger = \begin{pmatrix} (\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^\dagger\mathbf{M}_{21})^\dagger & -(\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^\dagger\mathbf{M}_{21})^\dagger\mathbf{M}_{12}\mathbf{M}_{22}^\dagger \\ -(\mathbf{M}_{22} - \mathbf{M}_{21}\mathbf{M}_{11}^\dagger\mathbf{M}_{12})^\dagger\mathbf{M}_{21}\mathbf{M}_{11}^\dagger & (\mathbf{M}_{22} - \mathbf{M}_{21}\mathbf{M}_{11}^\dagger\mathbf{M}_{12})^\dagger \end{pmatrix}.$$

Proof. This result is based on the resolution of a system of equations obtained from $\mathbf{M}$ and its inverse written as block matrices. □

The partial correlations of a Gaussian random vector are zero if and only if the two related components are independent conditionally to all other components.

Proof of Proposition 0.2.1 We consider the covariance matrix written by blocks,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XZ} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YZ} \\ \boldsymbol{\Sigma}_{ZX} & \boldsymbol{\Sigma}_{ZY} & \boldsymbol{\Sigma}_{ZZ} \end{pmatrix}.$$

The successive application of Theorem A.1.1 on normal correlations and then of Theorem

A.1.2 on block matrix inversion implies that

$$\mathbf{C}ov\left(\begin{pmatrix}X\\Y\end{pmatrix},\begin{pmatrix}X\\Y\end{pmatrix}\Big|Z\right)=\begin{pmatrix}\mathbf{\Sigma}_{XX}&\mathbf{\Sigma}_{XY}\\\mathbf{\Sigma}_{YX}&\mathbf{\Sigma}_{YY}\end{pmatrix}-\begin{pmatrix}\mathbf{\Sigma}_{XZ}\\\mathbf{\Sigma}_{YZ}\end{pmatrix}\mathbf{\Sigma}_{ZZ}^{-1}\begin{pmatrix}\mathbf{\Sigma}_{XZ}&\mathbf{\Sigma}_{YZ}\end{pmatrix}$$
$$=\begin{pmatrix}\mathbf{\Omega}_{XX}&\mathbf{\Omega}_{XY}\\\mathbf{\Omega}_{YX}&\mathbf{\Omega}_{YY}\end{pmatrix}.$$

Furthermore, under the normal hypothesis, X and Y are independent conditionally to Z is equivalent to $\mathbf{C}ov(X,Y|Z)=0$. Thus the conditional independence condition is equivalent to the fact that $\mathbf{C}ov\Big(\binom{X}{Y},\binom{X}{Y}\Big|Z\Big)$ is diagonal. The latter means that $\mathbf{\Omega}_{XY}=0$. As the corresponding partial correlation is $\mathbf{\Psi}_{XY}=-\mathbf{\Omega}_{XY}\big(\mathbf{\Omega}_{XX}\mathbf{\Omega}_{YY}\big)^{-1/2}$, that concludes the proof. □

Under the Gaussian assumption the coefficients of regression and the variance of the error of a linear regression model can be formulated using the precision matrix of both dependent and explanatory variables.

Proof of Proposition 0.3.1 We begin with a first formulation of the conditional covariance of Y, $\mathbf{C}ov(Y,Y|X)=B^\top\mathbf{C}ov(X,X|X)B+2B^\top\mathbf{C}ov(X,\epsilon|X)+\mathbf{C}ov(\epsilon,\epsilon|X)$. As $\mathbb{E}(X|X)=X$, we get $\mathbf{C}ov(X,X|X)=0$. We also have $\mathbf{C}ov(X,\epsilon|X)=0$ and $\mathbf{C}ov(\epsilon,\epsilon|X)=\mathbf{C}ov(\epsilon,\epsilon)=\Phi$ using that X independent of ϵ. Besides, the combination of the results of the theorem on normal correlations and the theorem on blockwise matrix inversion leads to $\mathbf{C}ov(Y,Y|X)=\mathbf{\Omega}_{YY}^{-1}$. That implies $\Phi=\mathbf{\Omega}_{YY}^{-1}$.
To obtain the expression of the matrix B, we note that $\mathbf{\Sigma}_{YX}=\mathbf{C}ov(Y,X)=B^\top\mathbf{C}ov(X,X)+\mathbf{C}ov(\epsilon,X)=B^\top\mathbf{\Sigma}_{XX}$. We deduce that $B^\top=\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}$. Moreover, applying again the theorem on blockwise matrix inversion yields to

$$\mathbf{\Omega}_{YX}=-(\mathbf{\Sigma}_{YY}-\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY})^{-1}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}=-\mathbf{\Omega}_{YY}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}=-\mathbf{\Omega}_{YY}B^\top,$$

that completes the proof. □

A.1.2 Risk bounds of the square-root Lasso

The square-root Lasso and the scaled Lasso procedures are equivalent.

Proof of Proposition 0.4.1 To convince oneself of the equivalence of these estimators, one just has to maximize the scaled Lasso over ϕ. Thus, $\widehat{\phi}^{\mathrm{scL}}$ satisfies $-\frac{1}{2\widehat{\phi}^{\mathrm{scL}\,2}\sqrt{n}}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\frac{\sqrt{n}}{2}=0$, then $\widehat{\phi}^{\mathrm{scL}\,2}=\frac{1}{n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2$. It remains to replace ϕ by its estimator in the scaled Lasso optimization problem to end the proof. □

Before getting into the proofs of our main results, let us state an intermediate restricted eigenvalue condition. Considering $H \in [p]$ and introducing the set

$$\mathscr{H}_{H,J} \triangleq \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}_{H^c \setminus J}\|_\infty \leq \min_{j \in H} |\boldsymbol{\delta}_j|\},$$

for $m \geq s \geq 1$ and $s + m \leq p$, we define

$$\kappa^{RE}(s,c,m) \triangleq \min_{\substack{J,H \subset [p],\\ J \cap H = \varnothing,\\ |J|=s, |H|=m}} \min_{\substack{\boldsymbol{\delta} \in \mathscr{C}_J(c) \cap \mathscr{H}_{H,J},\\ \boldsymbol{\delta} \neq 0}} \frac{1}{n} \frac{\|\mathbf{X}\boldsymbol{\delta}\|_2^2}{\|\boldsymbol{\delta}_{J \cup H}\|_2^2}. \tag{A.1}$$

Note that in the expression above, the condition $\boldsymbol{\delta} \in \mathscr{H}_{H,J}$ can be dropped from the second min as this minimum is anyway reached when this condition is satisfied. This variant of the restricted eigenvalue has been set up in [Bickel et al., 2009] and studied in [van de Geer and Bühlmann, 2009]. The related condition $\kappa^{RE}(s,c,m) > 0$ is obviously stronger than $\kappa^{RE}(s,c) > 0$, but weaker than $\bar{\kappa}^{RE}(s,c) > 0$. This condition is one of the main ingredients of the proof of the second claim of Proposition A.1.8.

In all this section, we simply denote by $\widehat{\boldsymbol{\beta}}$ the square-root Lasso estimator of the coefficients of regression $\boldsymbol{\beta}^*$. For the convenience of the reader, a complete set of proofs can be found below. The Propositions 0.4.2 and 0.4.3 correspond to the Propositions A.1.3 and A.1.4 in the special case where $\rho = 1/2$ and $\iota = \kappa^*(s,2,1)/2$.

Proposition A.1.3. *Set $s = |supp(\boldsymbol{\beta}^*)|$, $1 \leq s \leq p$. Let $\mathbf{X}$ be a $n \times p$ random matrix with i.i.d. centered Gaussian rows whose covariance matrix has unit diagonal entries and satisfies the ℓ_1-sensitivity property $\kappa^*(s,2,1) > 0$. Let us consider $\alpha \in (0,1)$, $\delta \in (0,1)$ and choose*

$$\lambda = 9\Big(\log \frac{6p}{\delta}\Big)^{1/2}.$$

Let $\iota > 0$ and $\rho < 1$ be some (arbitrary) constants. We assume that the sample size n satisfies

$$n \geq \Big(12\log(3/\delta)\Big) \vee \Big(\lambda^2/\big((\kappa^*(s,2,1) - \iota)\rho\big)\Big) \vee \Big(ds^2\log(1/\alpha)\Big),$$

where d is a constant depending on ι.
Set $A = 8\Big((\kappa^*(s,2,1) - \iota)\big(1-\rho\big)\Big)^{-1}$ and $B = 8\Big(\sqrt{3}\sqrt{(\kappa^*(s,2,1) - \iota)}\big(1-\rho\big)\Big)^{-1}$.
Then, the solution $\widehat{\boldsymbol{\beta}}$ of problem (square-root Lasso) satisfies

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_q \leq \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq A\frac{\lambda\phi^*}{\sqrt{n}}, \quad \text{for } q \geq 1, \quad \text{and} \quad \|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \leq B\lambda\phi^*, \tag{A.2}$$

with probability at least $1 - \delta - \alpha$.

Proposition A.1.4. *Set $s = |supp(\boldsymbol{\beta}^*)|$, $1 \leq s \leq p$. Let $\mathbf{X}$ be a $n \times p$ random matrix with i.i.d. centered Gaussian rows whose covariance matrix has unit diagonal entries and satisfies the restricted eigenvalue property $\bar{\kappa}^{*RE}(s,2) > 0$. Let us consider $\alpha \in (0,1)$,*

$\delta \in (0,1)$ and choose

$$\lambda = 9\Big(\log\frac{6p}{\delta}\Big)^{1/2}.$$

Let us consider the universal constants $a, b, d > 0$ and an (arbitrary) constant $\rho < 1$. We assume that the sample size n satisfies

$$n \geq \Big(12\log(3/\delta)\Big) \vee \Big(24s\lambda^2/\big(\bar{\kappa}^{*RE}(s,2)\rho\big)\Big) \vee \Big(1/d\log(b/\alpha)\Big) \vee \Big(as\log(p)/\bar{\kappa}^{*RE}(s,2)\Big).$$

Set $C = 64\Big(1+2\sqrt{s/n}\Big)\Big(\bar{\kappa}^{*RE}(s,2)\big(1-\rho\big)\Big)^{-1}$.
Then, the solution $\widehat{\boldsymbol{\beta}}$ of problem (square-root Lasso) satisfies

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq C\sqrt{s}\frac{\lambda\phi^*}{\sqrt{n}}, \tag{A.3}$$

with probability at least $1-\delta-\alpha$.

The proof of Proposition A.1.3 easily follows from the application of Corollary A.1.7 using Lemma A.1.13, along with Theorem 0.5.2. On the other hand, Proposition A.1.4 relies on Theorem A.1.5 below instead of Theorem 0.5.2.

Theorem A.1.5 (Raskutti et al. [2010]). *Let $\mathbf{X}$ be a $n \times p$ random matrix having zero-mean Gaussian rows whose covariance matrix $\boldsymbol{\Sigma}^*$ has unit diagonal entries and satisfies the restricted eigenvalue property $\bar{\kappa}^{*RE}(s,c) > 0$. Let $a, b, d > 0$ be universal constants. If $n > a(1+c)^2 s\log(p)/\bar{\kappa}^{*RE}(s,c)$, then the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ satisfies the restricted eigenvalue property $\bar{\kappa}^{RE}(s,c) = \bar{\kappa}^{*RE}(s,c)/8 > 0$ with probability at least $1 - b\,\mathrm{e}^{-dn}$.*

When the sample covariance matrix satisfies the ℓ_1-sensitivity property, the error of estimation measured in ℓ_1-norm is of the order of $\sqrt{\log(p)/n}$ with high probability.

Theorem A.1.6. *We assume that the diagonal entries of $\mathbf{X}^\top\mathbf{X}/n$ are equal to 1 and that $|supp(\boldsymbol{\beta}^*)| = s$, where $1 \leq s \leq p$. We denote $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ by $\widehat{\boldsymbol{\delta}}$. Let $\delta \in (0,1)$, $c > 1$ and choose*

$$\lambda = \frac{c+1}{c-1}\left(\frac{2\sqrt{n}\log(4p/\delta)}{\sqrt{n} - 2\sqrt{\log 4/\delta}}\right)^{1/2}.$$

If the ℓ_1-sensitivity property $\kappa(s,c,1) > 0$ is satisfied and $\lambda \leq \sqrt{n\kappa(s,c,1)\rho}$ with $\rho < 1$, then, denoting $A = 2c\Big(\kappa(s,c,1)\big(1-\rho\big)\Big)^{-1}$ and $B = 2c\Big(\sqrt{c+1}\sqrt{\kappa(s,c,1)}\big(1-\rho\big)\Big)^{-1}$,

$$\|\widehat{\boldsymbol{\delta}}\|_q \leq \|\widehat{\boldsymbol{\delta}}\|_1 \leq A\lambda\phi^*\frac{\sqrt{n} + \sqrt{2\log 2/\delta}}{n}, \quad \text{for } q \geq 1, \tag{A.4}$$

and

$$\|\mathbf{X}\widehat{\boldsymbol{\delta}}\|_2 \leq B\lambda\phi^*\frac{\sqrt{n} + \sqrt{2\log 2/\delta}}{\sqrt{n}} \tag{A.5}$$

hold with probability at least $1-\delta$.
Moreover, if the stronger restricted eigenvalue condition $\kappa^{RE}(s,c,m)>0$ *is satisfied, for* m *such that* $m \geq s \geq 1$ *and* $s+m \leq p$, *setting* $C = 2c\Big(1+c\sqrt{s/n}\Big)\Big(\kappa^{RE}(s,c,m)\big(1-\rho\big)\Big)^{-1}$, *then*

$$\|\widehat{\boldsymbol{\delta}}\|_2 \leq C\lambda\phi^*\sqrt{s}\frac{\sqrt{n}+\sqrt{2\log 2/\delta}}{n} \tag{A.6}$$

holds with the same probability as above.

For readability purposes, we state the above theorem in the particular case of $c=2$ and with an additional assumption on the sample size. Besides, we replace the assumption $\kappa^{RE}(s,2,m)>0$ by $\bar{\kappa}^{RE}(s,2)>0$ in the second claim.

Corollary A.1.7. *We assume that the diagonal entries of* $\mathbf{X}^\top\mathbf{X}/n$ *are equal to 1 and that* $|supp(\boldsymbol{\beta}^*)| = s$, *where* $1 \leq s \leq p$. *We denote* $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*$ *by* $\widehat{\boldsymbol{\delta}}$. *Let* $\delta \in (0,1)$, $n \geq 16\log 2/\delta$ *and choose*

$$\lambda = 6\Big(\log\frac{2p}{\delta}\Big)^{1/2}.$$

If the ℓ_1*-sensitivity property* $\kappa(s,2,1)>0$ *is satisfied and* $\lambda \leq \sqrt{n\kappa(s,2,1)\rho}$ *with* $\rho<1$, *then, denoting* $A = 8\Big(\kappa(s,2,1)\big(1-\rho\big)\Big)^{-1}$ *and* $B = 8\Big(\sqrt{3}\sqrt{\kappa(s,2,1)}\big(1-\rho\big)\Big)^{-1}$,

$$\|\widehat{\boldsymbol{\delta}}\|_q \leq \|\widehat{\boldsymbol{\delta}}\|_1 \leq A\frac{\lambda\phi^*}{\sqrt{n}}, \quad \textit{for } q \geq 1, \quad \textit{and} \quad \|\mathbf{X}\widehat{\boldsymbol{\delta}}\|_2 \leq B\lambda\phi^* \tag{A.7}$$

hold with probability at least $1-\delta$.
Moreover, if the stronger restricted eigenvalue condition $\bar{\kappa}^{RE}(s,2)>0$ *is satisfied, setting* $C = 8\Big(1+2\sqrt{s/n}\Big)\Big(\bar{\kappa}^{RE}(s,2)\big(1-\rho\big)\Big)^{-1}$, *then*

$$\|\widehat{\boldsymbol{\delta}}\|_2 \leq C\sqrt{s}\frac{\lambda\phi^*}{\sqrt{n}} \tag{A.8}$$

holds with the same probability as above.

Proof. As Theorem A.1.6, this result is based on Proposition A.1.8. Using the result of the Lemma A.1.11, with $c=2$ and $\lambda = 6\big(\log(2p/\delta)\big)^{1/2}$, when n is large enough, it holds that

$$\frac{\|\mathbf{X}^\top\boldsymbol{\epsilon}\|_\infty}{\|\boldsymbol{\epsilon}\|_2} \leq \frac{c-1}{c+1}\lambda,$$

with probability at least $1-\delta/2$. Moreover, if $n \geq 16\log 2/\delta \geq 2\log 2/\delta$, then Lemma A.1.12 implies that

$$\frac{\|\boldsymbol{\epsilon}\|_2}{\sqrt{n}} \leq \frac{2}{\sqrt{n}}$$

holds with probability at least $1-\delta/2$. □

General scheme of the proof of Theorem A.1.6 :

Step 1: We provide bounds for $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$ and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ that rely on $\|\boldsymbol{\epsilon}\|_2$, under the assumptions that the regularization parameter λ is lower bounded according to Eq. (A.17) and that the sample covariance matrix satisfies the ℓ_1-sensitivity property.

Step 2: We prove that λ satisfies condition (A.17) and that $\|\boldsymbol{\epsilon}\|_2$ is of order $O(\sqrt{n})$ with high probability.

We state the following proposition that gives an analytical bound for the error of estimation of the square-root Lasso procedure, assuming that the ℓ_1-sensitivity property is satisfied by the design matrix $\mathbf{X}$.

Proposition A.1.8. *Let us assume that $\boldsymbol{\beta}^*$ is s-sparse, $1 \leq s \leq p$, $\kappa(s,c,1) > 0$ and $\lambda \leq \sqrt{n\kappa(s,c,1)\rho}$ with $\rho < 1$. We denote $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ by $\widehat{\boldsymbol{\delta}}$. If the penalty level λ satisfies condition (A.17) for some constant $c > 1$, on the event $\{\boldsymbol{\epsilon} \neq 0\}$, then*

$$\|\widehat{\boldsymbol{\delta}}\|_q \leq \|\widehat{\boldsymbol{\delta}}\|_1 \leq A\frac{\lambda\phi^*\|\boldsymbol{\epsilon}\|_2}{n}, \quad \text{for } q \geq 1, \quad \text{and} \quad \|\mathbf{X}\widehat{\boldsymbol{\delta}}\|_2 \leq B\frac{\lambda\phi^*\|\boldsymbol{\epsilon}\|_2}{\sqrt{n}}, \tag{A.9}$$

where $A = 2c\Big(\kappa(s,c,1)\big(1-\rho\big)\Big)^{-1}$ and $B = 2c\Big(\sqrt{c+1}\sqrt{\kappa(s,c,1)}\big(1-\rho\big)\Big)^{-1}$.
Moreover, if the stronger restricted eigenvalue condition $\kappa^{RE}(s,c,m) > 0$ is satisfied, for m such that $m \geq s \geq 1$ and $s + m \leq p$, then

$$\|\widehat{\boldsymbol{\delta}}\|_2 \leq C\sqrt{s}\frac{\lambda\phi^*\|\boldsymbol{\epsilon}\|_2}{n}, \tag{A.10}$$

where $C = 2c\Big(1 + c\sqrt{s/n}\Big)\Big(\kappa^{RE}(s,c,m)\big(1-\rho\big)\Big)^{-1}$.

Proof. By definition of ℓ_1-sensitivity (0.13), considering $J = \mathrm{supp}(\boldsymbol{\beta}^*)$, we have

$$\|\widehat{\boldsymbol{\delta}}_J\|_1 \leq \frac{1}{n}\frac{\|\mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\delta}}\|_\infty}{\kappa(s,c,1)}. \tag{A.11}$$

The combination of the latter with Lemma A.1.9 entails that

$$\|\mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\delta}}\|_\infty \leq \lambda\frac{2c}{c+1}\phi^*\|\boldsymbol{\epsilon}\|_2\left(1 - \frac{\lambda^2}{n\kappa(s,c,1)}\right)^{-1}. \tag{A.12}$$

Then, inequalities (A.11) and (A.12) and $\|\widehat{\boldsymbol{\delta}}\|_1 \leq (1+c)\|\widehat{\boldsymbol{\delta}}_J\|_1$ (by Lemma A.1.10) imply immediately the first inequality of (A.9). Next, since it holds that $\|\mathbf{X}\boldsymbol{\delta}\|_2^2 \leq \|\boldsymbol{\delta}\|_1\|\mathbf{X}^\top\mathbf{X}\boldsymbol{\delta}\|_\infty$ and as $\widehat{\boldsymbol{\delta}} \in \mathscr{C}_J(c)$ by Lemma A.1.10, we obtain

$$\|\mathbf{X}\widehat{\boldsymbol{\delta}}\|_2^2 \leq (1+c)\|\widehat{\boldsymbol{\delta}}_J\|_1\|\mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\delta}}\|_\infty,$$

combining the latter with (A.11) gives

$$\|\mathbf{X}\widehat{\boldsymbol{\delta}}\|_2 \leq \Big(\frac{1+c}{n}\Big)^{1/2} \frac{\|\mathbf{X}^\top \mathbf{X}\widehat{\boldsymbol{\delta}}\|_\infty}{\big(\kappa(s,c,1)\big)^{1/2}}.$$

Finally, using inequality (A.12), we arrive at

$$\|\mathbf{X}\widehat{\boldsymbol{\delta}}\|_2 \leq \frac{\lambda\phi^*\|\boldsymbol{\epsilon}\|_2}{\sqrt{n}} \frac{2c/\sqrt{c+1}}{\big(\kappa(s,c,1)\big)^{1/2}\big(1-\lambda^2/(n\kappa(s,c,1))\big)}, \tag{A.13}$$

hence the second part of (A.9).

To prove the second claim of the proposition, we follow the sketch of the last part of the proof of Theorem 7.1 of [Bickel et al., 2009]. As in this paper, we denote the kth largest element in absolute value of a vector $\mathbf{v}$ by $|\mathbf{v}|_{(k)}$. Let H be the subset of $[p]\backslash J$ of size m that corresponds to the indexes of the m largest elements—outside J—of $\widehat{\boldsymbol{\delta}}$ in absolute value. It holds that

$$\|\widehat{\boldsymbol{\delta}}_{(H\cup J)^c}\|_2^2 = \|\widehat{\boldsymbol{\delta}}_{J^c\backslash H}\|_2^2 = \sum_{k=m+1}^{|J^c|} |\widehat{\boldsymbol{\delta}}_{J^c}|_{(k)}^2 \leq \|\widehat{\boldsymbol{\delta}}_{J^c}\|_1^2 \sum_{k=m+1}^{|J^c|} \frac{1}{k^2} \leq \frac{1}{m}\|\widehat{\boldsymbol{\delta}}_{J^c}\|_1^2.$$

Then, we apply Lemma A.1.10 which yields

$$\|\widehat{\boldsymbol{\delta}}_{(H\cup J)^c}\|_2^2 \leq c^2 \frac{1}{m}\|\widehat{\boldsymbol{\delta}}_J\|_1^2 \leq c^2 \frac{s}{m}\|\widehat{\boldsymbol{\delta}}_J\|_2^2 \leq c^2 \frac{s}{m}\|\widehat{\boldsymbol{\delta}}_{J\cup H}\|_2^2.$$

Using that $\|\widehat{\boldsymbol{\delta}}\|_2^2 = \|\widehat{\boldsymbol{\delta}}_{J\cup H}\|_2^2 + \|\widehat{\boldsymbol{\delta}}_{(H\cup J)^c}\|_2^2$, we get that

$$\|\widehat{\boldsymbol{\delta}}\|_2 \leq \Big(1+c\sqrt{s/m}\Big)\|\widehat{\boldsymbol{\delta}}_{J\cup H}\|_2 \leq \Big(1+c\sqrt{s/m}\Big)\frac{1}{\sqrt{n}}\frac{\|\mathbf{X}\widehat{\boldsymbol{\delta}}\|_2}{\sqrt{\kappa^{RE}(s,c,m)}},$$

where the last inequality comes from the restricted eigenvalue condition. To conclude, Eq. (A.13) combined with the inequality $\kappa(s,c,1) \geq s^{-1}(1+c)^{-1}\kappa^{RE}(s,c,m)$ leads to

$$\|\widehat{\boldsymbol{\delta}}\|_2 \leq \frac{\lambda\phi^*\|\boldsymbol{\epsilon}\|_2}{n}\sqrt{s}\Big(1+c\sqrt{s/m}\Big)\frac{2c}{\kappa^{RE}(s,c,m)(1-\rho)}.$$

□

Lemma A.1.9. *If the penalty level λ satisfies condition (A.17) for some constant $c > 1$, then*

$$\|\mathbf{X}^\top \mathbf{X}\widehat{\boldsymbol{\delta}}\|_\infty \leq \lambda\frac{2c}{c+1}\phi^*\|\boldsymbol{\epsilon}\|_2 + \lambda^2\|\widehat{\boldsymbol{\delta}}_J\|_1,$$

where $J = supp(\boldsymbol{\beta}^*)$ and $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$.

Proof. The Karush-Kuhn-Tucker conditions imply that the estimate $\widehat{\boldsymbol{\beta}}$ of the square-root

Lasso satisfies

$$\begin{cases} \|\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^{-1}(\mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}^\top\mathbf{y})+\lambda\mathrm{sgn}(\widehat{\boldsymbol{\beta}})=0 & , \quad \text{if } \|\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2\neq 0 \text{ and } \widehat{\boldsymbol{\beta}}_j\neq 0 \text{ for all } j, \\ \widehat{\boldsymbol{\beta}}=0 & , \quad \text{if } \|\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2=0, \\ \|\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^{-1}(\mathbf{X}_{\bullet,j}^\top\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}_{\bullet,j}^\top\mathbf{y})\in[-\lambda,\lambda] & , \quad \text{if } \widehat{\boldsymbol{\beta}}_j=0. \end{cases}$$

To sum up, it holds that

$$\|\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^{-1}(\mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}^\top\mathbf{y})=-\lambda L, \tag{A.14}$$

where the subgradient L of $\boldsymbol{\beta}\mapsto\|\boldsymbol{\beta}\|_1$ satisfies $L_j\in\begin{cases}\mathrm{sgn}(\widehat{\boldsymbol{\beta}}_j) \text{ if } \widehat{\boldsymbol{\beta}}_j\neq 0\\ [-1,1] \text{ if } \widehat{\boldsymbol{\beta}}_j=0\end{cases}$, for all $j\in[p]$.
In view of Eq. (A.14), we get

$$\|\mathbf{X}^\top(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})\|_\infty\leq\lambda\|\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2.$$

As, by the triangle inequality, we have

$$\|\mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\delta}}\|_\infty=\|\mathbf{X}^\top(\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{y}+\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^*)\|_\infty\leq\|\mathbf{X}^\top(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})\|_\infty+\phi^*\|\mathbf{X}^\top\boldsymbol{\epsilon}\|_\infty,$$

it follows that, under condition (A.17),

$$\|\mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\delta}}\|_\infty\leq\lambda\|\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2+\lambda\frac{c-1}{c+1}\phi^*\|\boldsymbol{\epsilon}\|_2. \tag{A.15}$$

Besides, as Eq. (A.21) implies that

$$\|\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2\leq\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^*\|_2+\lambda\|\widehat{\boldsymbol{\delta}}_J\|_1=\phi^*\|\boldsymbol{\epsilon}\|_2+\lambda\|\widehat{\boldsymbol{\delta}}_J\|_1, \tag{A.16}$$

we end with the claimed result. □

When the tuning parameter λ is large enough, the error of estimation $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*$ belongs to the cone defined by Eq. (0.12).

Lemma A.1.10. *Considering $J=supp(\boldsymbol{\beta}^*)$, if, for some constant $c>1$, the penalty level λ satisfies the condition*

$$\lambda\geq\frac{c+1}{c-1}\frac{\|\mathbf{X}^\top\boldsymbol{\epsilon}\|_\infty}{\|\boldsymbol{\epsilon}\|_2}, \tag{A.17}$$

then $\widehat{\boldsymbol{\delta}}=\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\in\mathscr{C}_J(c)$, on the event $\{\boldsymbol{\epsilon}\neq 0\}$.

Remark. *A similar property is valid for the standard Lasso and the Dantzig selector, respectively with constants $c=3$ and $c=1$ (see Bickel et al. [2009]).*

Proof. This proof is due to Belloni et al. [2011].

We first state the basic inequality

$$\|\boldsymbol{\beta}^*_J\|_1 - \|\widehat{\boldsymbol{\beta}}_J\|_1 \leq \|\widehat{\boldsymbol{\delta}}_J\|_1, \tag{A.18}$$

that comes from triangle inequality. Furthermore by definition of $\widehat{\boldsymbol{\beta}}$, it holds that

$$\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|_2 + \lambda\|\widehat{\boldsymbol{\beta}}\|_1 \leq \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2 + \lambda\|\boldsymbol{\beta}^*\|_1. \tag{A.19}$$

Also, as $\boldsymbol{\beta} \mapsto \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2$ is a convex differentiable function on the event $\{\boldsymbol{\epsilon} \neq 0\}$, we get that

$$\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|_2 - \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2 \geq \nabla(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2)|^\top_{\boldsymbol{\beta}^*}\widehat{\boldsymbol{\delta}}.$$

The gradient of $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2$ in $\boldsymbol{\beta}^*$ satisfies

$$\nabla(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2)|_{\boldsymbol{\beta}^*} = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}^* - \mathbf{y}) = -\|\boldsymbol{\epsilon}\|_2^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}.$$

Thus, with the condition (A.17), the previous inequality leads to

$$\|\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2-\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^*\|_2 \geq -\|\boldsymbol{\epsilon}\|_2^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}\widehat{\boldsymbol{\delta}} \geq \frac{1-c}{1+c}\lambda\|\widehat{\boldsymbol{\delta}}\|_1 = \frac{1-c}{1+c}\lambda(\|\widehat{\boldsymbol{\delta}}_J\|_1+\|\widehat{\boldsymbol{\delta}}_{J^c}\|_1). \tag{A.20}$$

Besides, equations (A.18) and (A.19) imply

$$\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|_2 - \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2 \leq \lambda(\|\widehat{\boldsymbol{\delta}}_J\|_1 - \|\widehat{\boldsymbol{\delta}}_{J^c}\|_1). \tag{A.21}$$

The combination of Eq. (A.20) and Eq. (A.21) entails $\|\widehat{\boldsymbol{\delta}}_{J^c}\|_1 \leq c\|\widehat{\boldsymbol{\delta}}_J\|_1$, that concludes the proof. □

Remark. *As Belloni et al. [2011], we use the gradient of $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2$ in $\boldsymbol{\beta}^*$ to set a lower bound for the penalty level. In other words, we consider the smallest λ that gives the following upper bound for the noise term*

$$\|\mathbf{X}^\top\boldsymbol{\epsilon}\|_\infty \leq \frac{c-1}{c+1}\|\boldsymbol{\epsilon}\|_2\lambda.$$

The same assumption is made in [Sun and Zhang, 2012] and [Sun and Zhang, 2013] to prove theoretical bounds for the scaled Lasso.

Probabilistic bounds The regularization parameter λ satisfies condition (A.17) with high probability.

Lemma A.1.11. *We assume that the diagonal entries of $\mathbf{X}^\top\mathbf{X}/n$ are equal to 1. For*

$\delta \in (0,1)$, if $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_n)$, the inequality

$$\frac{\|\mathbf{X}^\top \boldsymbol{\epsilon}\|_\infty}{\|\boldsymbol{\epsilon}\|_2} \leq \left(\frac{2\sqrt{n}\log(2p/\delta)}{\sqrt{n} - 2\sqrt{\log 2/\delta}} \right)^{1/2}$$

holds with probability at least $1-\delta$. Furthermore, if $n \geq 16 \log 2/\delta$, then

$$\frac{\|\mathbf{X}^\top \boldsymbol{\epsilon}\|_\infty}{\|\boldsymbol{\epsilon}\|_2} \leq 2\Big(\log \frac{2p}{\delta} \Big)^{1/2}$$

holds with the same probability as above.

Proof. As the errors $\boldsymbol{\epsilon}_i, i \in [n]$, follow i.i.d. standard normal distributions and are independent of $\mathbf{X}$, then conditionally to $\mathbf{X}$, the random variable $\mathbf{X}_{\bullet,j}^\top \boldsymbol{\epsilon}$ has a zero mean normal distribution with variance $\big(\mathbf{X}^\top \mathbf{X}\big)_{j,j} = n$. It means that, conditionally to $\mathbf{X}$, $n^{-1/2}\mathbf{X}_{\bullet,j}^\top \boldsymbol{\epsilon}$ is a standard normal random variable. We recall that a standard normal distribution function is lower bounded by $1 - \mathrm{e}^{-t^2/2}/2$ in $t \geq 1$. Therefore, combining this tail bound with the union bound entails that

$$\max_{j \in [p]} \big| n^{-1/2} \mathbf{X}_{\bullet,j}^\top \boldsymbol{\epsilon} \big| \leq \sqrt{2\log(2p/\delta)}$$

holds with probability at least $1 - \delta/2$. It means that $\|\mathbf{X}^\top \boldsymbol{\epsilon}\|_\infty \leq \sqrt{2n\log(2p/\delta)}$ holds with this probability. Moreover, as $\|\boldsymbol{\epsilon}\|_2^2 \sim \mathcal{X}^2(n)$, applying [Laurent and Massart, 2000, Lemma 1], we get that

$$\|\boldsymbol{\epsilon}\|_2 \geq \big(n - 2\sqrt{n \log 2/\delta} \big)^{1/2}$$

holds with probability at least $1 - \delta/2$. This leads to the first claim of the lemma. To obtain the second claim, we remark that $n \geq 16 \log 2/\delta$ implies that $\sqrt{n}/\big(\sqrt{n} - 2\sqrt{\log 2/\delta}\big) \leq 2$, and we substitute this inequality into the first bound of the lemma. □

The ℓ_2-norm of the error has an upper bound of order $\sqrt{n}$ with high probability.

Lemma A.1.12. *For any $\delta \in (0,1)$, the following inequality*

$$\|\boldsymbol{\epsilon}\|_2 \leq \sqrt{n} + \sqrt{2\log 1/\delta} \tag{A.22}$$

holds with probability at least $1-\delta$.

Proof. This lemma is an immediate consequence of the application of [Laurent and Massart, 2000, Lemma 1]. Indeed, as $\boldsymbol{\epsilon} \sim \mathcal{N}(0,1)$, $\|\boldsymbol{\epsilon}\|_2^2$ has a $\mathcal{X}^2$ distribution with n degrees of freedom. The aforementioned lemma implies that

$$\|\boldsymbol{\epsilon}\|_2^2 \leq n + 2\sqrt{n\log 1/\delta} + 2\log 1/\delta \leq \big(\sqrt{n} + \sqrt{2\log 1/\delta} \big)^2 \tag{A.23}$$

holds with probability at least $1-\delta$. □

We provide a tail bound on the diagonal entries of the covariance matrix of a Gaussian i.i.d. sample. The result of the following lemma is used to modify the bounds obtained in Lemma A.1.11 when the assumption that the diagonal entries of $\mathbf{X}^\top\mathbf{X}/n$ are equal to 1 is replaced by the assumption that $\mathbf{X}$ is a Gaussian random matrix whose covariance matrix $\mathbf{\Sigma}^*$ has unit diagonal elements. For instance, under this new assumption, for any $\delta \in (0,1)$, if $n \geq 12\log 3/\delta$, then

$$\frac{\|\mathbf{X}^\top\boldsymbol{\epsilon}\|_\infty}{\|\boldsymbol{\epsilon}\|_2} \leq 2\sqrt{2}\Big(\log\frac{3p}{\delta}\Big)^{1/2}$$

holds with probability at least $1-\delta$.

Lemma A.1.13. *For $\delta \in (0,1)$, if the $n \times p$ matrix $\mathbf{X}$ has independent centered Gaussian rows with covariance matrix $\mathbf{\Sigma}^*$ having unit diagonal, then for any $j \in [p]$,*

$$\big(\mathbf{X}^\top\mathbf{X}\big)_{j,j} \leq n + 2\sqrt{n\log 1/\delta} + 2\log 1/\delta$$

holds with probability at least $1-\delta$.

Proof. If $\mathbf{X}_{i,\bullet} \sim \mathcal{N}(0,\mathbf{\Sigma}^*)$, then $\big(\mathbf{\Sigma}^*_{j,j}\big)^{-1/2}\mathbf{X}_{i,j}$ follows a standard normal distribution. This yields that $\big(\mathbf{\Sigma}^*_{j,j}\big)^{-1/2}\big(\mathbf{X}^\top\mathbf{X}\big)_{j,j}$ is a $\mathcal{X}^2$ random variable with n degrees of freedom. Therefore, as we assume that $\mathbf{\Sigma}^*_{j,j} = 1$, applying [Laurent and Massart, 2000, Lemma 1] leads to

$$\big(\mathbf{X}^\top\mathbf{X}\big)_{j,j} \leq n + 2\sqrt{nt} + 2t,$$

with probability at least $1-\mathrm{e}^{-t}$, for any $t \geq 0$. Taking $t = \log 1/\delta$ yields the claimed bound. □

A.1.3 Risk bounds on precision matrix estimation

Propositions 0.6.1 and 0.6.2 simply correspond to the Propositions A.1.14 and A.1.15 in the special case where $\rho = 1/2$ and $\iota = \kappa^*(s,2,1)/2$.

Proposition A.1.14. *We assume that the maximal number of nonzero entries in a column of $\mathbf{\Omega}^*$ is $s \in [p]$. Let $\mathbf{X}$ be a $n \times p$ random matrix with i.i.d. centered Gaussian rows whose covariance matrix has unit diagonal entries and satisfies the ℓ_1-sensitivity property $\kappa^*(s,2,1) > 0$. Let us consider $\alpha \in (0,1)$, $\delta \in (0,1)$ and choose*

$$\lambda = 6\Big(\log\frac{8p^2}{\delta}\Big)^{1/2}.$$

For any $\iota > 0$ and $\rho < 1$, we assume that the sample size n satisfies

$$n \geq \Big(16\log(8p/\delta)\Big) \vee \Big(\lambda^2/\big((\kappa^*(s,2,1)-\iota)\rho\big)\Big) \vee \Big(ds^2\log(1/\alpha)\Big),$$

where d is a constant depending on ι. We set $A = 32\Big((\kappa^*(s,2,1)-\iota)\big(1-\rho\big)\Big)^{-1}$.
Then, the solution $\widehat{\mathbf{\Omega}}$ of problem (0.22) satisfies the following inequalities

$$\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{\infty,1} \leq \frac{1}{\sqrt{n}}\lambda\big(\max_j \omega^*_{jj}\big)^{1/2}\Big(A + \frac{2}{3}s\big(\max_j \omega^*_{jj}\big)^{1/2}\Big), \tag{A.24}$$

$$\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{1,1} \leq \frac{p}{\sqrt{n}}\lambda\big(\max_j \omega^*_{jj}\big)\Big(A\big(\min_j \omega^*_{jj}\big)^{-1/2} + \frac{2}{3}s\Big), \tag{A.25}$$

with probability at least $1-\delta-\alpha$.

Proposition A.1.15. *We assume that the maximal number of nonzero entries in a column of $\mathbf{\Omega}^*$ is $s \in [p]$. Let $\mathbf{X}$ be a $n \times p$ random matrix with i.i.d. centered Gaussian rows whose covariance matrix has unit diagonal entries and satisfies the restricted eigenvalue property $\bar{\kappa}^{*RE}(s,2) > 0$. Let us consider $\alpha \in (0,1)$, $\delta \in (0,1)$ and choose*

$$\lambda = 6\Big(\log\frac{8p^2}{\delta}\Big)^{1/2}.$$

For universal constants $a, b, d > 0$ and any $\rho < 1$, we assume that the sample size n satisfies

$$n \geq \Big(16\log(8p/\delta)\Big) \vee \Big(24s\lambda^2/\big(\bar{\kappa}^{*RE}(s,2)\rho\big)\Big) \vee \Big(1/d\log(b/\alpha)\Big) \vee \Big(as\log(p)/\bar{\kappa}^{*RE}(s,2)\Big).$$

We set $C = 256\Big(1+2\sqrt{s/n}\Big)\Big(\bar{\kappa}^{*RE}(s,2)\big(1-\rho\big)\Big)^{-1}$.
Then, the solution $\widehat{\mathbf{\Omega}}$ of problem (0.22) satisfies the following inequality

$$\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{2,2} \leq \frac{\sqrt{p}}{\sqrt{n}}\lambda\sigma_{\max}(\mathbf{\Omega}^*)\Big(\sqrt{s}C\big(\min_j \omega^*_{jj}\big)^{-1/2} + \frac{2}{3}\Big), \tag{A.26}$$

with probability at least $1-\delta-\alpha$.

Proposition A.1.14 claims that, in ℓ_∞/ℓ_1-norm, under suitable conditions, the error of estimation of the precision matrix has a convergence rate of order $s\sqrt{\log(p)/n}$ with high probability. This result is deduced from the combination of Theorem A.1.16 below and of Theorem 0.5.2, whereas Proposition A.1.15 ensues from the second claim of Theorem A.1.16 and Theorem A.1.5.

Theorem A.1.16. *We assume that the diagonal entries of $\mathbf{X}^\top\mathbf{X}/n$ are equal to 1 and that the maximal number of nonzero entries in a column of $\mathbf{\Omega}^*$ is $s \in [p]$. Let us consider*

$\delta \in (0,1)$, $n \geq 16 \log 8p/\delta$ and choose

$$\lambda = 6\Big(\log \frac{8p^2}{\delta}\Big)^{1/2}.$$

If the ℓ_1-sensitivity property $\kappa(s,2,1) > 0$ is satisfied and $\lambda \leq \sqrt{n\kappa(s,2,1)\rho}$ with $\rho < 1$, then, denoting $A = 32\Big(\kappa(s,2,1)\big(1-\rho\big)\Big)^{-1}$, each of the following inequalities holds

$$\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{\infty,1} \leq \frac{1}{\sqrt{n}}\lambda\big(\max_j \omega^*_{jj}\big)^{1/2}\Big(A + \frac{2}{3}s\big(\max_j \omega^*_{jj}\big)^{1/2}\Big), \tag{A.27}$$

$$\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{1,1} \leq \frac{p}{\sqrt{n}}\lambda\big(\max_j \omega^*_{jj}\big)\Big(A\big(\min_j \omega^*_{jj}\big)^{-1/2} + \frac{2}{3}s\Big), \tag{A.28}$$

with probability at least $1-\delta$.
Moreover, if the stronger restricted eigenvalue condition $\bar{\kappa}^{RE}(s,2) > 0$ is satisfied, setting $C = 32\Big(1+2\sqrt{s/n}\Big)\Big(\bar{\kappa}^{RE}(s,2)\big(1-\rho\big)\Big)^{-1}$, then

$$\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{2,2} \leq \frac{\sqrt{p}}{\sqrt{n}}\lambda\sigma_{\max}(\mathbf{\Omega}^*)\Big(\sqrt{s}C\big(\min_j \omega^*_{jj}\big)^{-1/2} + \frac{2}{3}\Big) \tag{A.29}$$

holds with the same probability as above.

The proof of Theorem A.1.16 is essentially based on Corollary A.1.7 and on the following lemma.

Lemma A.1.17. *For any $\delta \in (0,1)$,*

$$\max_{j\in[p]} \frac{\widehat{\omega}_{jj}}{\omega^*_{jj}} \leq \Big(1 - 2\sqrt{\frac{\log p/\delta}{n}}\Big)^{-1} \tag{A.30}$$

holds with probability at least $1-\delta$.
Moreover, under appropriate conditions (such that the conclusion of Corollary A.1.7 holds), for any $c > 1$, if $n \geq (2c/(c-1))^2 \log 2p/\delta$, then each of the following inequalities

$$\max_{j\in[p]} \Big|\frac{\widehat{\omega}_{jj}}{\omega^*_{jj}} - 1\Big| \leq 2c\sqrt{\frac{\log 2p/\delta}{n}}, \tag{A.31}$$

$$\|\widehat{\mathbf{D}}(\mathbf{D}^*)^{-1} - \mathbf{I}_p\|_{2,2} \leq 2c\sqrt{\frac{p\log 2p/\delta}{n}}; \quad \|\widehat{\mathbf{D}}(\mathbf{D}^*)^{-1} - \mathbf{I}_p\|_{1,1} \leq 2cp\sqrt{\frac{\log 2p/\delta}{n}}, \tag{A.32}$$

holds with the same probability as above.

Proof. We recall that $\widehat{\omega}_{jj}/\omega^*_{jj} = n/\big(\omega^*_{jj}\|\widehat{\boldsymbol{\epsilon}}_{\bullet,j}\|_2^2\big)$. To prove relation (A.30), we thus only have to get a lower bound for

$$\|\widehat{\boldsymbol{\epsilon}}_{\bullet,j}\|_2^2 = \|\mathbf{X}_{\bullet,j^c}(\widehat{\mathbf{B}}_{j^c,j} - \mathbf{B}^*_{j^c,j})\|_2^2 + \|\boldsymbol{\epsilon}_{\bullet,j}\|_2^2/\omega^*_{jj} \geq \|\boldsymbol{\epsilon}_{\bullet,j}\|_2^2/\omega^*_{jj}.$$

As $\|\boldsymbol{\epsilon}_{\bullet,j}\|_2^2 \sim \mathcal{X}^2(n)$, applying [Laurent and Massart, 2000, Lemma 1] involves that, for any $\delta \in (0,1)$, $\|\boldsymbol{\epsilon}_{\bullet,j}\|_2^2 \geq n - 2\sqrt{n \log 1/\delta}$ holds with probability at least $1-\delta$. It entails that, with this probability,

$$\frac{\widehat{\omega}_{jj}}{\omega^*_{jj}} \leq \left(1 - 2\sqrt{\frac{\log 1/\delta}{n}}\right)^{-1}. \tag{A.33}$$

The claimed bound follows from the union bound.

Next, to obtain the second bound of the lemma, we recall that Corollary A.1.7, under the following conditions : { the diagonal entries of $\mathbf{X}^\top\mathbf{X}/n$ are equal to 1, $n \geq 16 \log 8/\delta$, $\lambda = 6\sqrt{\log 2p/\delta}$, $\lambda \leq \sqrt{n\kappa(s,2,1)\rho}$ for $\rho < 1$, $B = 8\big(\sqrt{3}\sqrt{\kappa(s,2,1)}(1-\rho)\big)^{-1}$ and each column of the precision matrix $\boldsymbol{\Omega}^*$ has at most s nonzero elements }, entails that

$$\|\mathbf{X}_{\bullet,j^c}(\widehat{\mathbf{B}}_{j^c,j} - \mathbf{B}^*_{j^c,j})\|_2^2 \leq B^2\lambda^2/\omega^*_{jj}$$

holds with probability at least $1-\delta/4$.

Furthermore, applying [Laurent and Massart, 2000, Lemma 1] leads to $\|\boldsymbol{\epsilon}_{\bullet,j}\|_2^2 \leq n + 2\sqrt{n \log 4/\delta} + 2\log 4/\delta$, with probability at least $1-\delta/4$. Set $b = B^2\lambda^2/n + 2\sqrt{\log 4/\delta}/\sqrt{n} + 2\log(4/\delta)/n$, we thus arrive at

$$\|\widehat{\boldsymbol{\epsilon}}_{\bullet,j}\|_2^2 \leq n(1+b)/\omega^*_{jj},$$

with probability at least $1-\delta/2$. Along with Eq. (A.33), it implies that each of the following inequalities

$$\frac{\widehat{\omega}_{jj}}{\omega^*_{jj}} - 1 \leq \frac{2\sqrt{\frac{\log 2/\delta}{n}}}{1 - 2\sqrt{\frac{\log 2/\delta}{n}}} \quad \text{and} \quad 1 - \frac{\widehat{\omega}_{jj}}{\omega^*_{jj}} \leq \frac{b}{1+b} \leq \frac{2\sqrt{\frac{\log 2/\delta}{n}}}{1 - 2\sqrt{\frac{\log 2/\delta}{n}}} \tag{A.34}$$

holds with probability at least $1-\delta/2$. Finally, combining the previous inequalities and the union bound, it holds that

$$\max_{j\in[p]} \left|\frac{\widehat{\omega}_{jj}}{\omega^*_{jj}} - 1\right| \leq 2\sqrt{\frac{\log 2p/\delta}{n}}\left(1 - 2\sqrt{\frac{\log 2p/\delta}{n}}\right)^{-1}$$

holds with probability at least $1-\delta$. As for any $c \geq 1$, for all $x \in [0, (c-1)/c]$, $x/(1-x) \leq cx$ holds, we find that

$$\max_{j\in[p]} \left|\frac{\widehat{\omega}_{jj}}{\omega^*_{jj}} - 1\right| \leq 2c\sqrt{\frac{\log 2p/\delta}{n}}$$

holds with probability at least $1-\delta$, when $n \geq 4c^2/(c-1)^2 \log 2p/\delta$, for any $c > 1$.

From inequalities (A.34), we also get that

$$\|\widehat{\mathbf{D}}(\mathbf{D}^*)^{-1} - \mathbf{I}_p\|_{2,2}^2 = \sum_{j\in[p]} \left|\frac{\widehat{\omega}_{jj}}{\omega^*_{jj}} - 1\right|^2 \leq 4p\frac{\log 2p/\delta}{n}\left(1 - 2\sqrt{\frac{\log 2p/\delta}{n}}\right)^{-2},$$

with probability at least $1-\delta$. Then, as above, for any $c \geq 1$, when $n \geq 4c^2/(c-1)^2 \log 2p/\delta$, it holds that

$$\|\widehat{\mathbf{D}}(\mathbf{D}^*)^{-1} - \mathbf{I}_p\|_{2,2} \leq 2c\sqrt{\frac{p\log 2p/\delta}{n}},$$

with probability at least $1-\delta$. We prove the last inequality in the same way. □

We can now prove Theorem A.1.16.

Proof of Theorem A.1.16 As $\mathbf{\Omega}^* = \mathbf{B}^*\mathbf{D}^*$ and $\widehat{\mathbf{\Omega}} = \widehat{\mathbf{B}}\widehat{\mathbf{D}}$, the triangle inequality entails that

$$\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{\infty,1} \leq \|(\widehat{\mathbf{B}} - \mathbf{B}^*)\widehat{\mathbf{D}}\|_{\infty,1} + \|\mathbf{B}^*(\widehat{\mathbf{D}} - \mathbf{D}^*)\|_{\infty,1}.$$

We bound separately the two terms of the right side. First, we note that

$$\begin{aligned}\|(\widehat{\mathbf{B}} - \mathbf{B}^*)\widehat{\mathbf{D}}\|_{\infty,1} &= \max_{j\in[p]} \|(\widehat{\mathbf{B}}_{j^c,j} - \mathbf{B}^*_{j^c,j})\widehat{\omega}_{jj}\|_1 \\ &\leq \Big(\max_{j\in[p]} \|(\widehat{\mathbf{B}}_{j^c,j} - \mathbf{B}^*_{j^c,j})\omega^*_{jj}\|_1\Big)\Big(\max_{j\in[p]} \widehat{\omega}_{jj}/\omega^*_{jj}\Big),\end{aligned}$$

since

$$\|(\widehat{\mathbf{B}}_{j^c,j} - \mathbf{B}^*_{j^c,j})\widehat{\omega}_{jj}\|_1 \leq \|(\widehat{\mathbf{B}}_{j^c,j} - \mathbf{B}^*_{j^c,j})\omega^*_{jj}\|_1\Big(\max_{j\in[p]} \widehat{\omega}_{jj}/\omega^*_{jj}\Big).$$

Set $M = \max_{j\in[p]} \sqrt{\omega^*_{jj}}$. In the conditions of Corollary A.1.7, applying this corollary together with the union bound , taking $\lambda = 6\sqrt{\log(8p^2/\delta)}$, for any $\delta \in (0,1)$, with $n \geq 16\log 8p/\delta$ and $A = 8\Big(\kappa(s,2,1)\big(1-\rho\big)\Big)^{-1}$,

$$\max_{j\in[p]} \|(\widehat{\mathbf{B}}_{j^c,j} - \mathbf{B}^*_{j^c,j})\omega^*_{jj}\|_1 \leq A\frac{\lambda M}{\sqrt{n}}$$

holds with probability at least $1-\delta/4$.
Using Lemma A.1.17, Eq. (A.30), as $\big(1-2\sqrt{\log(4p/\delta)/n}\big)^{-1} \leq \big(1-\lambda/(3\sqrt{n})\big)^{-1}$, we deduce that

$$\|(\widehat{\mathbf{B}} - \mathbf{B}^*)\widehat{\mathbf{D}}\|_{\infty,1} \leq A\frac{\lambda M}{\sqrt{n}}\big(1 - \lambda/(3\sqrt{n})\big)^{-1}$$

holds with probability at least $1-\delta/2$.
Concerning the second term, we start with

$$\|\mathbf{B}^*(\widehat{\mathbf{D}} - \mathbf{D}^*)\|_{\infty,1} = \|\mathbf{\Omega}^*\big(\widehat{\mathbf{D}}(\mathbf{D}^*)^{-1} - \mathbf{I}_p\big)\|_{\infty,1} \leq \|\mathbf{\Omega}^*\|_{\infty,1}\Big(\max_{j\in[p]} |\widehat{\omega}_{jj}/\omega^*_{jj} - 1|\Big).$$

Then, noticing that $\|\mathbf{\Omega}^*\|_{\infty,1} \leq s\big(\max_{j\in[p]}\omega^*_{jj}\big) = sM^2$ and applying Lemma A.1.17, Eq. (A.31), for any $c>1$, when $n \geq (2c/(c-1))^2 \log 4p/\delta$, it holds that

$$\|\mathbf{B}^*(\widehat{\mathbf{D}}-\mathbf{D}^*)\|_{\infty,1} \leq 2csM^2\sqrt{\frac{\log 4p/\delta}{n}},$$

with probability at least $1-\delta/2$. Taking $c=2$ in the previous inequality and bringing it all together, when $n \geq 16\log 8p/\delta$, we end with

$$\|\widehat{\mathbf{\Omega}}-\mathbf{\Omega}^*\|_{\infty,1} \leq \frac{\lambda M}{\sqrt{n}}\bigg(\frac{A}{1-\lambda/(3\sqrt{n})}+\frac{2}{3}sM\bigg),$$

with probability at least $1-\delta$.

The bounds in Frobenius and ℓ_1/ℓ_1 norms are calculated likewise. We begin with applying the triangle inequality that leads to

$$\|\widehat{\mathbf{\Omega}}-\mathbf{\Omega}^*\|_{2,2} \leq \|\widehat{\mathbf{B}}-\mathbf{B}^*\|_{2,2}\bigg(\max_{j\in[p]}\frac{\widehat{\omega}_{jj}}{\omega^*_{jj}}\bigg)\Big(\max_{j\in[p]}\omega^*_{jj}\Big)+\sigma_{\max}(\mathbf{\Omega}^*)\|\widehat{\mathbf{D}}(\mathbf{D}^*)^{-1}-\mathbf{I}_p\|_{2,2}.$$

Under the conditions of Corollary A.1.7, choosing $\lambda = 6\sqrt{\log(8p^2/\delta)}$, for any $\delta\in(0,1)$, with $n \geq 16\log 8p/\delta$ and $C = 8\Big(1+2\sqrt{s/n}\Big)\Big(\bar{\kappa}^{RE}(s,2)\big(1-\rho\big)\Big)^{-1}$, it holds that

$$\|\widehat{\mathbf{B}}-\mathbf{B}^*\|^2_{2,2} = \sum_{j=1}^{p}\|(\widehat{\mathbf{B}}-\mathbf{B}^*)_{\bullet,j}\|_2^2 \leq \frac{ps\lambda^2C^2}{n}\big(\min_j\omega^*_{jj}\big)^{-1},$$

with probability at least $1-\delta/4$. Besides, for any $c>1$, when $n \geq (2c/(c-1))^2\log 4p/\delta$, equations (A.30) and (A.32) of Lemma A.1.17 entail that

$$\max_{j\in[p]}\frac{\widehat{\omega}_{jj}}{\omega^*_{jj}} \leq \big(1-\lambda/(3\sqrt{n})\big)^{-1} \quad \text{and} \quad \|\widehat{\mathbf{D}}(\mathbf{D}^*)^{-1}-\mathbf{I}_p\|_{2,2} \leq 2c\sqrt{\frac{p\log 4p/\delta}{n}}$$

hold respectively with probability at least $1-\delta/4$ and probability at least $1-\delta/2$. We take $c=2$ and recall that $\max_j\omega^*_{jj} \leq \sigma_{\max}(\mathbf{\Omega}^*)$. When $n \geq 16\log 8p/\delta$, we therefore conclude that

$$\|\widehat{\mathbf{\Omega}}-\mathbf{\Omega}^*\|_{2,2} \leq \frac{\sqrt{p}}{\sqrt{n}}\lambda\sigma_{\max}(\mathbf{\Omega}^*)\Big(\frac{\sqrt{s}C\big(\min_j\omega^*_{jj}\big)^{-1/2}}{1-\lambda/(3\sqrt{n})}+\frac{2}{3}\Big)$$

holds with probability at least $1-\delta$. To obtain the remaining bound in element-wise ℓ_1-matrix norm, we simply need to note that

$$\|\mathbf{B}^*(\widehat{\mathbf{D}}-\mathbf{D}^*)\|_{1,1} = \sum_{j=1}^{p}\|\mathbf{\Omega}^*_{j,\bullet}(\widehat{\mathbf{D}}(\mathbf{D}^*)^{-1}-\mathbf{I}_p)\|_1 \leq s\big(\max_j\omega^*_{jj}\big)\|\widehat{\mathbf{D}}(\mathbf{D}^*)^{-1}-\mathbf{I}_p\|_{1,1}.$$

Then, under the same conditions that have been used to get the bound in ℓ_∞/ℓ_1-norm,

similar reasoning leads to

$$\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\|_{1,1} \leq \frac{p}{\sqrt{n}}\lambda\big(\max_j \omega^*_{jj}\big)\Big(\frac{A\big(\min_j \omega^*_{jj}\big)^{-1/2}}{1-\lambda/(3\sqrt{n})} + \frac{2}{3}s\Big),$$

with the same probability as above. We finally simplify the bounds noting that $n \geq 16\log 8p/\delta$ implies that $\big(1-\lambda/(3\sqrt{n})\big)^{-1} \leq \sqrt{2}/(\sqrt{2}-1) \leq 4$. □

A.1.4 Regularity properties

The ℓ_q-sensitivity is a weaker condition than the compatibility condition, which is itself weaker than the restricted eigenvalue condition.

Proof of Proposition 0.5.1 Let J be a subset of $[p]$ such that $|J| = s$ and $\boldsymbol{\delta} \in \mathscr{C}_J(c)$. Using the inequality $\|\mathbf{X}\boldsymbol{\delta}\|_2^2 \leq \|\boldsymbol{\delta}\|_1\|\mathbf{X}^\top\mathbf{X}\boldsymbol{\delta}\|_\infty$ and as in $\mathscr{C}_J(c)$, $\boldsymbol{\delta}$ satisfies $\|\boldsymbol{\delta}\|_1 \leq (1+c)\|\boldsymbol{\delta}_J\|_1$, we get that

$$\frac{1}{n}\frac{\|\mathbf{X}\boldsymbol{\delta}\|_2^2}{\|\boldsymbol{\delta}_J\|_1^2} \leq \frac{1}{n}\frac{\|\boldsymbol{\delta}\|_1\|\mathbf{X}^\top\mathbf{X}\boldsymbol{\delta}\|_\infty}{\|\boldsymbol{\delta}_J\|_1^2} \leq \frac{1}{n}(1+c)\frac{|\mathbf{X}^\top\mathbf{X}\boldsymbol{\delta}\|_\infty}{\|\boldsymbol{\delta}_J\|_1}.$$

Moreover, by the Cauchy-Schwarz inequality it holds $\|\boldsymbol{\delta}_J\|_1 \leq |J|^{\frac{1}{2}}\|\boldsymbol{\delta}_J\|_2$, we thus arrive at

$$\frac{1}{n}\frac{\|\mathbf{X}\boldsymbol{\delta}\|_2^2}{\|\boldsymbol{\delta}_J\|_2^2} \leq \frac{1}{n}|J|\frac{\|\mathbf{X}\boldsymbol{\delta}\|_2^2}{\|\boldsymbol{\delta}_J\|_1^2} \leq \frac{1}{n}|J|(1+c)\frac{|\mathbf{X}^\top\mathbf{X}\boldsymbol{\delta}\|_\infty}{\|\boldsymbol{\delta}_J\|_1}. \tag{A.35}$$

It remains to take the minimum over $\mathscr{C}_J(c)$ and over all possible sets J on each block to obtain the claimed result. □

Controlling the error of estimation of the covariance matrix $\|\boldsymbol{\Sigma}^* - \widehat{\boldsymbol{\Sigma}}\|_{\infty,\infty}$ is crucial to infer the ℓ_1-sensitivity property on the sample covariance matrix from the ℓ_1-sensitivity property on the true covariance matrix.

Proof of Lemma 0.5.3 The proof follows essentially the same sketch as van de Geer and Bühlmann [2009, Lemma 10.1].
Let us consider $\boldsymbol{\delta} \in \mathscr{C}_J(c)$, where $J \subset [p]$ and $|J| = s$. Using the triangle inequality, we observe that

$$|\|\boldsymbol{\Sigma}^*\boldsymbol{\delta}\|_\infty - \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\delta}\|_\infty| = \|(\boldsymbol{\Sigma}^* - \widehat{\boldsymbol{\Sigma}})\boldsymbol{\delta}\|_\infty \leq \|\boldsymbol{\Sigma}^* - \widehat{\boldsymbol{\Sigma}}\|_{\infty,\infty}\|\boldsymbol{\delta}\|_1.$$

Then, as $\|\boldsymbol{\delta}_{J^c}\|_1 \leq c\|\boldsymbol{\delta}_J\|_1$ and $\kappa^*(s,c,1) \leq \|\boldsymbol{\Sigma}^*\boldsymbol{\delta}\|_\infty/\|\boldsymbol{\delta}_J\|_1$, we have $|\|\boldsymbol{\Sigma}^*\boldsymbol{\delta}\|_\infty - \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\delta}\|_\infty| \leq (c+1)\|\boldsymbol{\Sigma}^* - \widehat{\boldsymbol{\Sigma}}\|_{\infty,\infty}\|\boldsymbol{\Sigma}^*\boldsymbol{\delta}\|_\infty/\kappa^*(s,c,1)$. Thus, we get

$$\frac{\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\delta}\|_\infty}{\|\boldsymbol{\Sigma}^*\boldsymbol{\delta}\|_\infty} \geq 1 - \frac{(c+1)\|\boldsymbol{\Sigma}^* - \widehat{\boldsymbol{\Sigma}}\|_{\infty,\infty}}{\kappa^*(s,c,1)},$$

it follows that

$$\frac{\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\delta}\|_\infty}{\|\boldsymbol{\delta}_J\|_1} \geq \frac{\|\boldsymbol{\Sigma}^*\boldsymbol{\delta}\|_\infty}{\|\boldsymbol{\delta}_J\|_1}\Big(1 - \frac{(c+1)\|\boldsymbol{\Sigma}^* - \widehat{\boldsymbol{\Sigma}}\|_{\infty,\infty}}{\kappa^*(s,c,1)}\Big),$$

and it remains to take the minimum of both sides to end the proof. □

A.1.5 Additional notes

Relation between the Clime method and the procedure of Yuan [2010] Considering that $\boldsymbol{\Omega} = \mathbf{BD}$, with $\mathbf{B}_{j,j} = 1$ for any j, the optimization problems (0.9) are equivalent to estimating $\boldsymbol{\Omega}^*_{\bullet,j}$ by $\widetilde{\boldsymbol{\Omega}}_{\bullet,j} = \widehat{\mathbf{B}}_{\bullet,j}\widehat{\boldsymbol{\phi}}_j^{-2}$, where

$$\{-\widehat{\mathbf{B}}_{\bullet,j}, \widehat{\boldsymbol{\phi}}_j\} = \underset{\substack{\boldsymbol{\beta}\in\mathbb{R}^p \\ \boldsymbol{\beta}_j=-1}}{\arg\min} \min_{\phi\in(0,\infty)} \|\boldsymbol{\beta}\|_1\phi^2 \quad \text{subject to} \quad \|\mathbf{S}_n\boldsymbol{\beta} + \phi^2(\mathbf{I}_p)_{\bullet,j}\|_\infty \leq \lambda\phi^2.$$

We further note that the Dantzig-type constraints of these problems are equivalent to

$$\Big(\|(\mathbf{S}_n)_{j^c,j^c}\boldsymbol{\beta}_{j^c} - (\mathbf{S}_n)_{j^c,j}\|_\infty \vee |(\mathbf{S}_n)_{\bullet,j}\boldsymbol{\beta} + \phi^2|\Big) \leq \lambda\phi^2.$$

Under these conditions, replacing the joint minimization according to $\boldsymbol{\beta}$ and ϕ by a two step procedure that consists in computing $-\widehat{\mathbf{B}}_{\bullet,j}$ first, as follows

$$-\widehat{\mathbf{B}}_{\bullet,j} = \underset{\substack{\boldsymbol{\beta}\in\mathbb{R}^p \\ \boldsymbol{\beta}_j=-1}}{\arg\min} \|\boldsymbol{\beta}\|_1 \quad \text{subject to} \quad \|(\mathbf{S}_n)_{j^c,j^c}\boldsymbol{\beta}_{j^c} - (\mathbf{S}_n)_{j^c,j}\|_\infty \leq \bar{\lambda},$$

then setting $\widehat{\boldsymbol{\phi}}_j^2 = \big(\mathbf{S}_n\widehat{\mathbf{B}}\big)_{j,j} \vee 0$ (the relaxed maximum likelihood estimator, see Chapter 1) leads to a new estimator of the precision matrix. This last procedure differs from the one developed in [Yuan, 2010] only through the fact that the diagonal elements of the precision matrix are there estimated by the variances of regression residuals, that is $\widehat{\boldsymbol{\phi}}_j^2 = \widehat{\mathbf{B}}_{\bullet,j}^\top\mathbf{S}_n\widehat{\mathbf{B}}_{\bullet,j}$.

A.2 Proofs of Chapter 1

We state a simple inequality between the diagonal entries of a positive definite matrix and the corresponding diagonal entries of its inverse.

Proposition A.2.1. *For any $p \times p$ positive definite matrix $\boldsymbol{\Sigma}$, whose inverse is denoted by $\boldsymbol{\Omega}$, it holds that*

$$\boldsymbol{\Omega}_{ii} \geq (\boldsymbol{\Sigma}_{ii})^{-1}, \quad \textit{for any } i \in [p].$$

In Chapter 1, we suppose that the covariance matrix $\boldsymbol{\Sigma}^*$ has unit diagonal entries. Therefore, it implies that $\boldsymbol{\Omega}^*_{ii} \geq 1$, for any i.

We recall the following useful lemma, needed to prove Proposition A.2.1.

Lemma A.2.2. *For any symmetric positive definite matrix* $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B}^\top \\ \mathbf{B} & \mathbf{C} \end{pmatrix}$, *it holds that*

$$\det(\mathbf{M}) = \det(\mathbf{A}) \det(\mathbf{C} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^\top) \leq \det(\mathbf{A}) \det(\mathbf{C}).$$

Proof. This result is obtained by writing $\mathbf{M}$ as a product of block triangular matrices:

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{B} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{A}^{-1}\mathbf{B}^\top \\ \mathbf{0} & \mathbf{C} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^\top \end{pmatrix}.$$

Then, we recall that for any eigenvalue λ_v of $\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^\top$ associated to an eigenvector $\mathbf{v}$, it holds that $\mathbf{v}^\top \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^\top \mathbf{v} = \lambda_v \mathbf{v}^\top \mathbf{v}$. Furthermore, as $\mathbf{M}$ is symmetric positive definite, we find that

$$\mathbf{v}^\top \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^\top \mathbf{v} = \mathbf{v}^\top \left[(\mathbf{A}^{-1}\mathbf{B}^\top)^\top ; \mathbf{0} \right] \mathbf{M} \begin{pmatrix} \mathbf{A}^{-1}\mathbf{B}^\top \\ \mathbf{0} \end{pmatrix} \mathbf{v} > 0.$$

We thus deduce that $\lambda_v > 0$ and it follows that $\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^\top$ is symmetric positive definite. We may conclude using that for any symmetric positive definite matrices $\mathbf{N}$, $\mathbf{M}$, $\mathbf{N}^{-1}\mathbf{M}$ is also positive definite, thus there exists a matrix $\mathbf{P}$ of eigenvectors such that $\mathbf{N}^{-1}\mathbf{M} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$, where $\mathbf{D}$ is the diagonal matrix of eigenvalues. Then, $\det(\mathbf{N} + \mathbf{M}) = \det(\mathbf{N}(\mathbf{I} + \mathbf{P}\mathbf{D}\mathbf{P}^{-1})) = \det(\mathbf{N}) \det(\mathbf{P}(\mathbf{I} + \mathbf{D})\mathbf{P}^{-1}) = \det(\mathbf{N}) \det(\mathbf{I} + \mathbf{D}) \geq \det(\mathbf{N})$. It remains to take $\mathbf{N} = \mathbf{C} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^\top$ and $\mathbf{M} = \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^\top$ to end the proof. □

Proof of Proposition A.2.1 For any $i \in [p]$, we start with

$$\mathbf{\Omega}_{ii} = (\mathbf{\Sigma}^{-1})_{ii} = \det(\mathbf{\Sigma})^{-1}(-1)^{i+i} \det(\mathbf{\Sigma}_{-i,-i}) = \det(\mathbf{\Sigma})^{-1} \det(\mathbf{\Sigma}_{-i,-i}),$$

by writing the cofactor. Besides, we remark that

$$\det(\mathbf{\Sigma}) = \det \begin{pmatrix} \mathbf{\Sigma}_{ii} & \mathbf{\Sigma}_{i,-i} \\ \mathbf{\Sigma}_{-i,i} & \mathbf{\Sigma}_{-i,-i} \end{pmatrix} \leq \mathbf{\Sigma}_{ii} \det(\mathbf{\Sigma}_{-i,-i}).$$

The claimed inequality $\mathbf{\Omega}_{ii} \geq (\mathbf{\Sigma}_{ii})^{-1}$ follows. □

Appendix B

Additional experimental results

Résumé. Les résultats expérimentaux présentés dans cette annexe apportent un éclairage complémentaire sur le comportement des estimateurs étudiés dans le Chapitre 1, en particulier en ce qui concerne le choix du paramètre d'ajustement κ et la performance en dimension élevée des estimateurs des éléments diagonaux.

B.1 Experimental results for Chapter 1

B.1.1 For $\kappa = 0.05$ in PML

p	30			60			90		
n	200	800	2000	200	800	2000	200	800	2000
B* estimated by square-root Lasso									
RV	**0.895 (.065)**	**0.402 (.037)**	**0.223 (.019)**	**1.422 (.087)**	**0.657 (.035)**	**0.374 (.021)**	**1.814 (.076)**	**0.852 (.034)**	**0.495 (.023)**
RML	1.374 (.068)	0.789 (.036)	0.532 (.020)	2.113 (.094)	1.243 (.036)	0.840 (.022)	2.668 (.080)	1.590 (.034)	1.087 (.023)
SML	1.500 (.090)	0.807 (.036)	0.546 (.019)	2.397 (.189)	1.260 (.034)	0.851 (.021)	3.056 (.233)	1.608 (.034)	1.096 (.023)
PML	1.375 (.068)	0.789 (.036)	0.533 (.019)	2.113 (.095)	1.243 (.035)	0.840 (.022)	2.668 (.079)	1.590 (.034)	1.087 (.023)
B* estimated by square-root Lasso followed by OLS									
RV	0.741 (.066)	**0.347 (.048)**	**0.242 (.018)**	1.087 (.084)	0.617 (.057)	**0.354 (.020)**	1.331 (.069)	0.855 (.055)	**0.443 (.021)**
RML	0.741 (.066)	**0.347 (.048)**	**0.242 (.018)**	1.087 (.084)	0.617 (.057)	**0.354 (.020)**	1.331 (.069)	0.855 (.055)	**0.443 (.021)**
SML	0.827 (.074)	0.457 (.063)	0.277 (.017)	1.209 (.108)	0.809 (.085)	0.381 (.019)	1.485 (.114)	1.100 (.077)	0.465 (.020)
PML	**0.734 (.067)**	0.348 (.049)	0.246 (.017)	**1.076 (.085)**	**0.614 (.059)**	0.357 (.020)	**1.318 (.069)**	**0.850 (.056)**	0.444 (.020)
B* is estimated without error									
RV	0.263 (.035)	0.132 (.015)	0.082 (.010)	0.373 (.033)	0.182 (.018)	0.117 (.012)	0.451 (.041)	0.220 (.022)	0.142 (.011)
RML	0.322 (.041)	0.160 (.019)	0.102 (.011)	0.468 (.041)	0.226 (.021)	0.146 (.013)	0.553 (.049)	0.281 (.024)	0.175 (.013)
SML	**0.034 (.024)**	**0.018 (.014)**	**0.010 (.008)**	**0.048 (.032)**	**0.017 (.014)**	**0.013 (.010)**	**0.052 (.032)**	**0.017 (.014)**	**0.011 (.009)**
PML	0.190 (.026)	0.093 (.013)	0.059 (.008)	0.269 (.029)	0.128 (.012)	0.084 (.009)	0.314 (.028)	0.156 (.012)	0.099 (.007)

Table B.1: Performance of the estimators of diagonal elements of the precision matrix in Model 1, with $\kappa = 0.05$ for the PML estimation. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

p	30			60			90		
n	200	800	2000	200	800	2000	200	800	2000
$\mathbf{B}^*$ estimated by square-root Lasso									
RV	**0.381 (.048)**	**0.122 (.018)**	**0.068 (.012)**	**0.613 (.053)**	**0.170 (.019)**	**0.092 (.012)**	**0.814 (.052)**	**0.217 (.019)**	**0.111 (.012)**
RML	1.030 (.052)	0.498 (.026)	0.317 (.017)	1.632 (.059)	0.777 (.020)	0.487 (.014)	2.110 (.059)	0.997 (.024)	0.625 (.013)
SML	1.266 (.157)	0.527 (.033)	0.337 (.019)	2.358 (.360)	0.819 (.031)	0.513 (.018)	3.386 (.609)	1.047 (.032)	0.656 (.022)
PML	1.040 (.051)	0.499 (.027)	0.318 (.017)	1.652 (.058)	0.781 (.021)	0.487 (.015)	2.135 (.059)	1.000 (.024)	0.625 (.014)
$\mathbf{B}^*$ estimated by square-root Lasso followed by OLS									
RV	0.241 (.046)	0.107 (.012)	0.063 (.010)	0.309 (.049)	0.125 (.013)	0.081 (.011)	0.378 (.059)	0.153 (.013)	0.094 (.010)
RML	0.241 (.046)	0.107 (.012)	0.063 (.010)	0.309 (.049)	0.125 (.013)	0.081 (.011)	0.378 (.059)	0.153 (.013)	0.094 (.010)
SML	0.343 (.156)	0.105 (.020)	0.065 (.015)	0.551 (.241)	0.128 (.023)	0.080 (.013)	0.907 (.608)	0.148 (.016)	0.089 (.011)
PML	**0.225 (.047)**	**0.100 (.013)**	**0.059 (.011)**	**0.282 (.048)**	**0.115 (.014)**	**0.075 (.011)**	**0.350 (.057)**	**0.140 (.014)**	**0.085 (.010)**
$\mathbf{B}^*$ is estimated without error									
RV	0.202 (.030)	0.106 (.013)	0.063 (.010)	0.255 (.028)	0.124 (.014)	0.081 (.011)	0.304 (.028)	0.152 (.013)	0.094 (.010)
RML	0.265 (.040)	0.139 (.017)	0.083 (.010)	0.351 (.036)	0.179 (.016)	0.114 (.013)	0.432 (.031)	0.214 (.014)	0.132 (.010)
SML	**0.035 (.020)**	**0.017 (.011)**	**0.010 (.007)**	**0.030 (.020)**	**0.012 (.009)**	**0.008 (.007)**	**0.033 (.023)**	**0.014 (.009)**	**0.008 (.005)**
PML	0.138 (.034)	0.072 (.013)	0.043 (.010)	0.156 (.023)	0.080 (.012)	0.052 (.010)	0.189 (.026)	0.095 (.010)	0.057 (.007)

Table B.2: Performance of the estimators of diagonal elements of the precision matrix in Model 2, with $\kappa = 0.05$ for the PML estimation. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

p	30			60			90		
n	200	800	2000	200	800	2000	200	800	2000
$\mathbf{B}^*$ estimated by square-root Lasso									
RV	**0.278 (.034)**	**0.135 (.020)**	**0.082 (.011)**	**0.410 (.033)**	**0.190 (.015)**	**0.123 (.012)**	**0.516 (.040)**	**0.238 (.016)**	**0.152 (.011)**
RML	0.498 (.057)	0.272 (.024)	0.169 (.017)	0.728 (.053)	0.393 (.029)	0.258 (.013)	0.887 (.053)	0.499 (.029)	0.326 (.012)
SML	1.100 (.125)	0.659 (.108)	0.380 (.045)	1.289 (.159)	0.808 (.054)	0.647 (.054)	1.222 (.149)	0.677 (.057)	0.461 (.028)
PML	0.499 (.057)	0.272 (.025)	0.169 (.017)	0.729 (.053)	0.393 (.029)	0.258 (.013)	0.887 (.053)	0.499 (.029)	0.327 (.012)
$\mathbf{B}^*$ estimated by square-root Lasso followed by OLS									
RV	**0.815 (.198)**	**0.169 (.113)**	**0.082 (.011)**	**2.272 (.181)**	**1.824 (.161)**	**0.928 (.177)**	**3.280 (.228)**	**3.823 (.137)**	**3.685 (.120)**
RML	**0.815 (.198)**	**0.169 (.113)**	**0.082 (.011)**	**2.272 (.181)**	**1.824 (.161)**	**0.928 (.177)**	**3.280 (.228)**	**3.823 (.137)**	**3.685 (.120)**
SML	1.277 (.148)	0.592 (.134)	0.343 (.052)	2.518 (.165)	1.933 (.147)	1.086 (.149)	3.440 (.235)	3.840 (.134)	3.696 (.117)
PML	0.838 (.191)	0.171 (.113)	**0.082 (.011)**	2.291 (.179)	**1.824 (.161)**	**0.928 (.177)**	3.295 (.229)	**3.823 (.137)**	**3.685 (.120)**
$\mathbf{B}^*$ is estimated without error									
RV	0.268 (.033)	0.134 (.020)	0.082 (.011)	0.388 (.033)	0.186 (.014)	0.122 (.012)	0.467 (.034)	0.233 (.017)	0.151 (.011)
RML	0.320 (.043)	0.166 (.023)	0.100 (.013)	0.477 (.040)	0.229 (.019)	0.149 (.014)	0.581 (.045)	0.286 (.023)	0.187 (.014)
SML	**0.043 (.031)**	**0.017 (.013)**	**0.014 (.010)**	**0.043 (.031)**	**0.023 (.017)**	**0.010 (.008)**	**0.038 (.030)**	**0.019 (.016)**	**0.010 (.008)**
PML	0.320 (.043)	0.166 (.023)	0.101 (.013)	0.478 (.040)	0.229 (.019)	0.149 (.014)	0.582 (.045)	0.286 (.023)	0.187 (.014)

Table B.3: Performance of the estimators of diagonal elements of the precision matrix in Model 3, with $\kappa = 0.05$ for the PML estimation. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

p	30			60			90		
n	200	800	2000	200	800	2000	200	800	2000
$\mathbf{B}^*$ estimated by square-root Lasso									
RV	**0.372 (.077)**	**0.185 (.034)**	**0.120 (.023)**	**0.548 (.084)**	**0.267 (.034)**	**0.164 (.017)**	**0.666 (.070)**	**0.322 (.037)**	**0.202 (.028)**
RML	0.424 (.072)	0.217 (.032)	0.139 (.019)	0.605 (.081)	0.299 (.034)	0.186 (.018)	0.726 (.067)	0.361 (.034)	0.226 (.026)
SML	0.469 (.086)	0.236 (.033)	0.149 (.019)	0.700 (.098)	0.327 (.034)	0.203 (.019)	0.848 (.104)	0.410 (.043)	0.254 (.027)
PML	0.431 (.073)	0.222 (.032)	0.142 (.019)	0.615 (.081)	0.307 (.034)	0.192 (.019)	0.736 (.066)	0.373 (.034)	0.234 (.026)
$\mathbf{B}^*$ estimated by square-root Lasso followed by OLS									
RV	**0.363 (.077)**	**0.183 (.034)**	**0.120 (.023)**	**0.537 (.084)**	**0.264 (.034)**	**0.163 (.017)**	**0.660 (.076)**	**0.319 (.037)**	**0.201 (.028)**
RML	**0.363 (.077)**	**0.183 (.034)**	**0.120 (.023)**	**0.537 (.084)**	**0.264 (.034)**	**0.163 (.017)**	**0.660 (.076)**	**0.319 (.037)**	**0.201 (.028)**
SML	0.387 (.080)	0.193 (.034)	0.126 (.023)	0.586 (.104)	0.280 (.033)	0.173 (.018)	0.728 (.096)	0.350 (.042)	0.221 (.029)
PML	0.364 (.078)	**0.183 (.034)**	**0.120 (.023)**	0.539 (.085)	0.265 (.034)	0.164 (.017)	0.663 (.076)	0.321 (.037)	0.202 (.028)
$\mathbf{B}^*$ is estimated without error									
RV	0.362 (.077)	0.183 (.034)	0.120 (.023)	0.536 (.084)	0.264 (.034)	0.163 (.017)	0.652 (.070)	0.319 (.037)	0.201 (.028)
RML	0.375 (.077)	0.186 (.034)	0.123 (.023)	0.545 (.085)	0.268 (.033)	0.166 (.018)	0.660 (.068)	0.325 (.036)	0.204 (.027)
SML	**0.346 (.080)**	**0.176 (.035)**	**0.114 (.023)**	**0.520 (.082)**	**0.256 (.034)**	**0.159 (.018)**	**0.636 (.072)**	**0.310 (.038)**	**0.195 (.029)**
PML	0.360 (.079)	0.181 (.034)	0.119 (.023)	0.532 (.083)	0.262 (.034)	0.162 (.018)	0.648 (.070)	0.318 (.037)	0.200 (.028)

Table B.4: Performance of the estimators of diagonal elements of the precision matrix in Model 4, with $\kappa = 0.05$ for the PML estimation. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

p	30			60			90		
n	200	800	2000	200	800	2000	200	800	2000
B* estimated by square-root Lasso									
RV	0.394 (.079)	**0.190 (.040)**	**0.127 (.020)**	0.555 (.070)	**0.285 (.034)**	**0.178 (.021)**	0.678 (.080)	**0.350 (.041)**	**0.213 (.021)**
RML	**0.390 (.078)**	0.193 (.039)	0.129 (.021)	**0.552 (.068)**	0.291 (.035)	0.184 (.022)	**0.675 (.079)**	0.361 (.040)	0.223 (.020)
SML	**0.390 (.078)**	0.195 (.042)	0.135 (.025)	**0.552 (.068)**	0.294 (.038)	0.187 (.023)	**0.675 (.079)**	0.363 (.039)	0.228 (.022)
PML	**0.390 (.078)**	0.193 (.039)	0.129 (.021)	**0.552 (.068)**	0.291 (.035)	0.184 (.022)	**0.675 (.079)**	0.361 (.040)	0.223 (.020)
B* estimated by square-root Lasso followed by OLS									
RV	0.390 (.078)	**0.198 (.040)**	**0.132 (.023)**	**0.548 (.069)**	**0.300 (.037)**	**0.187 (.025)**	**0.673 (.077)**	**0.373 (.038)**	**0.229 (.023)**
RML	0.390 (.078)	**0.198 (.040)**	**0.132 (.023)**	**0.548 (.069)**	**0.300 (.037)**	**0.187 (.025)**	**0.673 (.077)**	**0.373 (.038)**	**0.229 (.023)**
SML	**0.389 (.078)**	0.200 (.040)	0.140 (.024)	**0.548 (.069)**	0.302 (.038)	0.192 (.025)	**0.673 (.077)**	0.376 (.038)	0.238 (.023)
PML	0.390 (.078)	**0.198 (.040)**	**0.132 (.023)**	**0.548 (.069)**	**0.300 (.037)**	**0.187 (.025)**	**0.673 (.077)**	**0.373 (.038)**	**0.229 (.023)**
B* is estimated without error									
RV	0.393 (.075)	0.190 (.041)	0.127 (.020)	0.542 (.070)	0.283 (.034)	0.176 (.021)	0.661 (.079)	0.344 (.041)	0.211 (.022)
RML	0.394 (.076)	0.190 (.041)	0.127 (.020)	0.542 (.071)	0.284 (.034)	0.177 (.021)	0.663 (.078)	0.344 (.041)	0.212 (.022)
SML	**0.360 (.077)**	**0.172 (.040)**	**0.112 (.021)**	**0.511 (.066)**	**0.264 (.038)**	**0.165 (.022)**	**0.626 (.082)**	**0.325 (.041)**	**0.197 (.022)**
PML	0.394 (.076)	0.190 (.041)	0.127 (.020)	0.542 (.071)	0.284 (.034)	0.177 (.021)	0.663 (.078)	0.344 (.041)	0.212 (.022)

Table B.5: Performance of the estimators of diagonal elements of the precision matrix in Model 5, with $\kappa = 0.05$ for the PML estimation. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

p	30			60			90		
n	200	800	2000	200	800	2000	200	800	2000
$\mathbf{B}^*$ estimated by square-root Lasso									
RV	0.369 (.062)	**0.205 (.028)**	**0.144 (.016)**	0.524 (.055)	**0.305 (.025)**	**0.203 (.015)**	0.650 (.053)	**0.377 (.032)**	**0.247 (.018)**
RML	**0.363 (.058)**	0.225 (.029)	0.159 (.017)	**0.521 (.051)**	0.335 (.024)	0.232 (.018)	**0.647 (.051)**	0.413 (.030)	0.288 (.017)
SML	0.364 (.057)	0.226 (.033)	0.174 (.020)	0.522 (.051)	0.341 (.026)	0.244 (.022)	0.648 (.051)	0.418 (.030)	0.304 (.023)
PML	**0.363 (.058)**	0.225 (.029)	0.159 (.017)	**0.521 (.051)**	0.335 (.024)	0.232 (.018)	**0.647 (.051)**	0.413 (.030)	0.288 (.017)
$\mathbf{B}^*$ estimated by square-root Lasso followed by OLS									
RV	**0.366 (.058)**	0.251 (.030)	**0.166 (.021)**	**0.524 (.052)**	**0.367 (.027)**	**0.257 (.025)**	**0.648 (.048)**	**0.458 (.031)**	**0.326 (.024)**
RML	**0.366 (.058)**	0.251 (.030)	**0.166 (.021)**	**0.524 (.052)**	**0.367 (.027)**	**0.257 (.025)**	**0.648 (.048)**	**0.458 (.031)**	**0.326 (.024)**
SML	**0.366 (.058)**	0.252 (.031)	0.186 (.022)	**0.524 (.052)**	0.370 (.025)	0.273 (.023)	**0.648 (.048)**	0.460 (.031)	0.345 (.026)
PML	**0.366 (.058)**	**0.250 (.030)**	**0.166 (.021)**	**0.524 (.052)**	**0.367 (.027)**	**0.257 (.025)**	**0.648 (.048)**	**0.458 (.031)**	**0.326 (.024)**
$\mathbf{B}^*$ is estimated without error									
RV	0.394 (.071)	0.204 (.029)	0.146 (.017)	0.558 (.064)	0.301 (.029)	0.203 (.016)	0.692 (.073)	0.371 (.031)	0.247 (.020)
RML	0.398 (.069)	0.207 (.028)	0.147 (.017)	0.564 (.061)	0.306 (.030)	0.206 (.015)	0.702 (.074)	0.377 (.031)	0.250 (.020)
SML	**0.183 (.063)**	**0.097 (.021)**	**0.067 (.023)**	**0.234 (.061)**	**0.142 (.030)**	**0.091 (.017)**	**0.318 (.055)**	**0.179 (.030)**	**0.112 (.023)**
PML	0.398 (.069)	0.207 (.028)	0.147 (.017)	0.564 (.061)	0.306 (.030)	0.206 (.015)	0.702 (.074)	0.377 (.031)	0.250 (.020)

Table B.6: Performance of the estimators of diagonal elements of the precision matrix in Model 6, with $\kappa = 0.05$ for the PML estimation. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

B.1.2 In high-dimension

In Tables 1.7-1.12, we present experimental measures of performance obtained for a sample size $n = 50$ for dimensions $p = 30, 60, 90$ and with $\kappa = 0.05$ for the PML estimation. For the last two values of p, the dimension is larger than the sample size.

	with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}}$			with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}+OLS}$			with $\mathbf{B}^*$		
p	30	60	90	30	60	90	30	60	90
RV	**1.692**	**2.670**	**3.486**	1.395	2.344	3.096	0.515	0.723	0.872
	(.144)	**(.145)**	**(.154)**	(.174)	(.190)	(.241)	(.079)	(.073)	(.073)
RML	2.031	3.038	3.869	1.395	2.344	3.096	0.654	0.911	1.102
	(.120)	(.126)	(.112)	(.174)	(.190)	(.241)	(.087)	(.080)	(.075)
SML	2.063	3.053	3.877	1.387	**2.324**	**3.077**	**0.082**	**0.114**	**0.099**
	(.116)	(.123)	(.110)	(.181)	**(.195)**	**(.247)**	**(.060)**	**(.074)**	**(.070)**
PML	2.032	3.038	3.869	**1.377**	2.329	3.084	0.387	0.525	0.627
	(.119)	(.125)	(.112)	**(.177)**	(.194)	(.246)	(.057)	(.047)	(.044)

Table B.7: Performance of the estimators of diagonal elements of the precision matrix in Model 1 for $n = 50$, with $\kappa = 0.05$ for the PML estimation. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

	with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}}$			with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}+OLS}$			with $\mathbf{B}^*$		
p	30	60	90	30	60	90	30	60	90
RV	**1.668**	**2.837**	**3.830**	**1.594**	**2.610**	**3.536**	0.397	0.517	0.585
	(.149)	**(.161)**	**(.172)**	**(.219)**	**(.218)**	**(.212)**	(.059)	(.064)	(.055)
RML	2.241	3.644	4.725	**1.594**	**2.610**	**3.536**	0.525	0.708	0.836
	(.164)	(.156)	(.162)	**(.219)**	**(.218)**	**(.212)**	(.080)	(.072)	(.057)
SML	2.590	3.901	4.895	2.038	2.948	3.748	**0.073**	**0.060**	**0.071**
	(.163)	(.167)	(.138)	(.308)	(.199)	(.188)	**(.056)**	**(.045)**	**(.048)**
PML	2.245	3.648	4.728	1.600	2.617	3.540	0.277	0.323	0.366
	(.164)	(.155)	(.161)	(.221)	(.215)	(.211)	(.060)	(.051)	(.046)

Table B.8: Performance of the estimators of diagonal elements of the precision matrix in Model 2 for $n = 50$, with $\kappa = 0.05$ for the PML estimation. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

	with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}}$			with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}+OLS}$			with $\mathbf{B}^*$		
p	30	60	90	30	60	90	30	60	90
RV	**0.587**	**0.879**	**1.159**	**1.044**	**1.634**	**1.980**	0.523	0.751	0.946
	(.073)	**(.077)**	**(.093)**	**(.194)**	**(.169)**	**(.159)**	(.064)	(.068)	(.080)
RML	0.907	1.240	1.499	**1.044**	**1.634**	**1.980**	0.630	0.921	1.168
	(.115)	(.137)	(.117)	**(.194)**	**(.169)**	**(.159)**	(.089)	(.088)	(.092)
SML	1.452	2.018	2.569	1.452	2.002	2.426	**0.081**	**0.112**	**0.106**
	(.169)	(.263)	(.333)	(.190)	(.174)	(.267)	**(.057)**	**(.074)**	**(.083)**
PML	0.907	1.241	1.500	1.071	1.655	2.000	0.630	0.921	1.168
	(.115)	(.137)	(.117)	(.189)	(.168)	(.157)	(.091)	(.088)	(.093)

Table B.9: Performance of the estimators of diagonal elements of the precision matrix in Model 3 for $n = 50$, with $\kappa = 0.05$ for the PML estimation. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

	with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}}$			with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}+OLS}$			with $\mathbf{B}^*$		
p	30	60	90	30	60	90	30	60	90
RV	**0.761**	**1.120**	**1.385**	**0.756**	**1.129**	**1.438**	0.723	1.064	1.320
	(.138)	**(.130)**	**(.149)**	**(.149)**	**(.148)**	**(.165)**	(.136)	(.127)	(.138)
RML	0.851	1.198	1.480	**0.756**	**1.129**	**1.438**	0.743	1.082	1.339
	(.134)	(.135)	(.148)	**(.149)**	**(.148)**	**(.165)**	(.136)	(.128)	(.136)
SML	0.944	1.286	1.577	0.814	1.179	1.512	**0.694**	**1.032**	**1.292**
	(.134)	(.122)	(.140)	(.137)	(.152)	(.156)	**(.139)**	**(.132)**	**(.143)**
PML	0.854	1.200	1.480	0.759	1.130	1.441	0.719	1.055	1.315
	(.133)	(.135)	(.147)	(.148)	(.150)	(.164)	(.136)	(.130)	(.139)

Table B.10: Performance of the estimators of diagonal elements of the precision matrix in Model 4 for $n = 50$, with $\kappa = 0.05$ for the PML estimation. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

	with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}}$			with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}+OLS}$			with $\mathbf{B}^*$		
p	30	60	90	30	60	90	30	60	90
RV	0.802	1.132	1.371	**0.813**	**1.135**	**1.374**	0.785	1.104	1.349
	(.146)	(.146)	(.134)	**(.147)**	**(.151)**	**(.144)**	(.135)	(.143)	(.135)
RML	**0.788**	**1.113**	**1.355**	**0.813**	**1.135**	**1.374**	0.786	1.105	1.350
	(.141)	**(.144)**	**(.133)**	**(.147)**	**(.151)**	**(.144)**	(.137)	(.143)	(.135)
SML	**0.788**	**1.113**	**1.355**	**0.813**	**1.135**	**1.374**	**0.726**	**1.043**	**1.290**
	(.141)	**(.144)**	**(.133)**	**(.147)**	**(.151)**	**(.144)**	**(.133)**	**(.149)**	**(.126)**
PML	**0.788**	**1.113**	**1.355**	**0.813**	**1.135**	**1.374**	0.786	1.105	1.350
	(.141)	**(.144)**	**(.133)**	**(.147)**	**(.151)**	**(.144)**	(.137)	(.143)	(.135)

Table B.11: Performance of the estimators of diagonal elements of the precision matrix in Model 5 for $n = 50$, with $\kappa = 0.05$ for the PML estimation. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

	with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}}$			with $\widehat{\mathbf{B}}^{\sqrt{\text{Lasso}}+OLS}$			with $\mathbf{B}^*$		
p	30	60	90	30	60	90	30	60	90
RV	0.709	1.008	1.223	**0.725**	**1.027**	**1.242**	0.748	1.075	1.322
	(.110)	(.117)	(.118)	**(.117)**	**(.133)**	**(.124)**	(.107)	(.118)	(.120)
RML	**0.691**	**0.986**	**1.203**	**0.725**	**1.027**	**1.242**	0.759	1.087	1.335
	(.111)	**(.115)**	**(.115)**	**(.117)**	**(.133)**	**(.124)**	(.111)	(.116)	(.117)
SML	0.692	0.987	**1.203**	**0.725**	**1.027**	**1.242**	**0.322**	**0.492**	**0.630**
	(.111)	(.115)	**(.115)**	**(.117)**	**(.133)**	**(.124)**	**(.120)**	**(.104)**	**(.133)**
PML	**0.691**	**0.986**	**1.203**	**0.725**	**1.027**	**1.242**	0.759	1.087	1.335
	(.111)	**(.115)**	**(.115)**	**(.117)**	**(.133)**	**(.124)**	(.111)	(.116)	(.117)

Table B.12: Performance of the estimators of diagonal elements of the precision matrix in Model 6 for $n = 50$, with $\kappa = 0.05$ for the PML estimation. The number of replications in each case is $R = 50$. More details on the experimental set-up are presented in Section 1.4.1.

Appendix C

Overview of the DESP package

RÉSUMÉ. Nous présentons ici très succinctement le paquet R DESP. Ce paquet a pour objet l'estimation des paramètres d'une distribution normale multivariée, y compris lorsque la dimension des données est élevée ou en présence d'observations extrêmes ou aberrantes. Une attention particulière est prêtée à l'estimation des éléments diagonaux de la matrice de précision.

C.1 Introduction

C.1.1 Purpose of this package

DESP is an R package[1], designed to estimate efficiently the parameters $\boldsymbol{\Omega}$ and $\boldsymbol{\mu}$ of a Gaussian distribution as developed in this manuscript. Its main characteristics are the ability to deal with data contaminated by outliers, with high-dimensional data and the availability of several estimators of the diagonal elements of the precision matrix.

In this chapter, we zoom in on the function `desp()` which is an interface to most of the features of the package.

C.1.2 Estimators

Let us first introduce the proposed estimator that indeed corresponds to the estimator developed in Chapters 1 and 2 with a slight redefinition of the tuning parameter that promotes sparsity. We consider a possible additive contamination of the data by outliers. We assume that the matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ of observed data satisfies

$$\mathbf{X} = \mathbf{Y} + \mathbf{E}^*, \tag{C.1}$$

where $\mathbf{E}^*$ is the matrix of errors and $\mathbf{Y}$ the outlier-free data matrix. The rows latter are supposed to be independent Gaussian, such that $\mathbf{Y}_{i,\bullet} \sim \mathcal{N}_p(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$. We also assume that most of the rows of the matrix $\mathbf{E}^*$ correspond to inliers, hence are only filled with zeros. The $p \times p$ matrix $\mathbf{B}^*$ corresponds to $\boldsymbol{\Omega}^* \cdot \mathrm{diag}(\{1/\omega^*_{jj}\}_{j\in[p]})$. We introduce the matrix $\boldsymbol{\Theta}^* = \mathbf{E}^*\mathbf{B}^*/\sqrt{n}$ that has the same sparsity pattern as $\mathbf{E}^*$. We denote by $\mathbf{X}^{(n)}$ the matrix $\mathbf{X}/\sqrt{n}$, the estimators of the parameters of the Gaussian distribution as defined as follows:

$$\{\widehat{\mathbf{B}}, \widehat{\boldsymbol{\Theta}}\} = \underset{\substack{\mathbf{B}:\mathbf{B}_{jj}=1\\ \boldsymbol{\Theta}\in\mathbb{R}^{n\times p}}}{\arg\min} \min_{\mathbf{c}\in\mathbb{R}^p} \left\{ \left\|(\mathbf{X}^{(n)}\mathbf{B} - \mathbf{u}_n\mathbf{c}^\top - \boldsymbol{\Theta})^\top\right\|_{2,1} + \lambda\|\boldsymbol{\Theta}\|_{2,1} + \gamma\|\mathbf{B}\|_{1,1} \right\}, \tag{C.2}$$

where $\lambda \geq 0$ and $\gamma \geq 0$ are tuning parameters respectively promoting robustness and sparsity of the matrix $\mathbf{B}$. The precision matrix $\boldsymbol{\Omega}^*$ can be estimated by

$$\widehat{\omega}_{jj} = \frac{2n}{\pi}\|(\mathbf{I}_n - \mathbf{u}_n\mathbf{u}_n^\top)(\mathbf{X}^{(n)}\widehat{\mathbf{B}}_{\bullet,j} - \widehat{\boldsymbol{\Theta}}_{\bullet,j})\|_1^{-2}; \qquad \widehat{\boldsymbol{\Omega}} = \widehat{\mathbf{B}} \cdot \mathrm{diag}(\{\widehat{\omega}_{jj}\}_{j\in[p]}). \tag{C.3}$$

We highlight that the estimator of the diagonal entries of the precision matrix stated above—based on average absolute deviation around the mean—is only one of the alternatives proposed in the DESP package. The other possibilities rest on residual variance or likelihood maximization (relaxed, symmetry-enforced or penalized). The expectation

[1] Licensed under GPL (version 3) and available on CRAN at http://cran.r-project.org/web/packages/DESP/index.html

vector $\boldsymbol{\mu}^*$ can be estimated by

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}(\mathbf{X} - \widehat{\mathbf{E}})^\top \mathbf{1}_n, \quad \text{where} \quad \widehat{\mathbf{E}} = \sqrt{n}\,\widehat{\boldsymbol{\Theta}}\widehat{\mathbf{B}}^\dagger. \tag{C.4}$$

The solutions of the problem (C.2) are obtained iteratively, by optimizing separately with respect to $\mathbf{B}$ and $\boldsymbol{\Theta}$. The details of the algorithm are provided in 2.5.

C.2 Implementation

The convex optimization problem (C.2) is decomposed in p independent sub-problems. When both tuning parameters λ and γ are zero, the solution of problem (C.2) is obtained using ordinary least squares to estimate each column of $\mathbf{B}$. When $\gamma \neq 0$, each of these p problems corresponding to the square-root Lasso can be either cast as a SOCP, or solved using the coordinate descent algorithm. In the first case, we propose to use the splitting conic solver (SCS, [O'Donoghue et al., 2013]) that solves efficiently convex cone problems. We note however that a more efficient solution in terms of computational time is obtained using the coordinate descend algorithm (stochastic or not). The p square-root Lasso problems can be solved in parallel when the OpenMP application programming interface (API)[2] is supported. Most of the linear algebra operations are performed calling BLAS and LAPACK routines[3]. For the details of available options of the function `desp()`, we refer the user to the package reference manual.

Plans for future enhancements include notably the possibility to settle a warm start solution that should lead to significant gain of execution time when using cross-validation techniques to choose the values of the tuning parameters. In addition, we intend to introduce safe rules to restrict the set of possibly nonzero coefficients before optimizing by the square-root Lasso.

C.3 Installation

As available on CRAN[4], this package can be simply installed by entering the following instruction:

```
install.packages("DESP")
```

We recommend to use a compiler that supports OpenMP to allow multithreading.

[2] OpenMP Architecture Review Board, see http://openmp.org.
[3] These API are for instance implemented by OpenBLAS, available at http://www.openblas.net/.
[4] The Comprehensive R Archive Network, https://cran.r-project.org/.

C.4 Example

We estimate the parameters of the distribution of Fisher's iris data [Anderson, 1935] for each of three iris species, assuming that these data are normally distributed.
We first load the package and the data set:

```
library(DESP)
data(iris3)
```

We will use the function `desp.cv()` that relies on the function `desp()` to estimate $\widehat{\boldsymbol{\Omega}}$ and $\widehat{\boldsymbol{\mu}}$, choosing the tuning parameters λ and γ by v-fold cross-validation Geisser [1975]. To define this function, we have introduced the following partition of the sample $S = \bigcup_{i=1}^{v} S_i$. The values of these parameters are selected over a grid such that the risk (the expectation of the loss) is the lowest. In connexion with the regression model, we might consider a quadratic loss function and select:

$$\{\lambda_{\mathrm{vc}}, \gamma_{\mathrm{vc}}\} = \arg\min_{\lambda,\gamma} \frac{1}{v} \sum_{i=1}^{v} \frac{1}{|S_i|} \sum_{k=1}^{|S_i|} \|\mathbf{X}_{k,\bullet} \widehat{\mathbf{B}}_{(i,\lambda,\gamma)} - \widehat{\boldsymbol{\mu}}_{(i,\lambda,\gamma)}^{\top} \widehat{\mathbf{B}}_{(i,\lambda,\gamma)}\|_2^2$$

where $\widehat{\boldsymbol{\mu}}_{(i,\lambda,\gamma)}$ and $\widehat{\mathbf{B}}_{(i,\lambda,\gamma)}$ are the estimates obtained on the training set $S \backslash S_i$. As the chosen loss function is not robust, we use instead

$$\{\lambda_{\mathrm{vc}}, \gamma_{\mathrm{vc}}\} = \arg\min_{\lambda,\gamma} \frac{1}{v} \sum_{i=1}^{v} \frac{1}{|S_i|} \sum_{k=1}^{|S_i|} \|\mathbf{X}_{k,\bullet} \widehat{\mathbf{B}}_{(i,\lambda,\gamma)} - \widehat{\boldsymbol{\mu}}_{(i,\lambda,\gamma)}^{\top} \widehat{\mathbf{B}}_{(i,\lambda,\gamma)}\|_2$$

that is inspired by the ℓ_1 cross-validation procedure [Wang and Scott, 1994].
We choose the estimator based on average absolute deviation around the mean (C.3) to estimate the diagonal entries. Besides, we choose a number of folds equal to 5.

```
settings <- list(diagElem='AD')
v <- 5
```

Then, we call this function on the first 25 observations of each species of iris:

```
set.seed(1)
categories <- colnames(iris3[1,,])
params <- vector(mode="list", length=length(categories))
for(c in 1:length(categories)){
  obs <- 1:25
  lr <- (9/10)^(0:9)
  gr <- (1/sqrt(2))^(0:4) * sqrt(2*log(ncol(iris3[,,c])))
  params[[c]] <- desp.cv(iris3[obs,,c], v=v, lambda.range=lr,
    gamma.range=gr, settings=settings)
}
```

Bibliography

Hirotugu Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, December 1974. MR0423716.

Christophe Ambroise, Julien Chiquet, and Catherine Matias. Inferring sparse gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3:205–238, 2009. MR2495837.

Erling D. Andersen and Knud D. Andersen. The mosek interior point optimizer for linear programming: an implementation of the homogeneous algorithm. In *High Performance Optimization*, pages 197–232. 2000. MR1748773.

Edgar Anderson. The irises of the gaspe peninsula. *Bulletin of the American Iris Society*, 59:2–5, 1935.

Theodore W. Anderson. *An introduction to multivariate statistical analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, third edition, 2003. MR1990662.

Francis R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008. MR2417268.

Francis R. Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4 (1):1–106, January 2012.

Samuel Balmand and Arnak S. Dalalyan. On estimation of the diagonal elements of a sparse precision matrix. *Electronic Journal of Statistics*, 10(1):1551–1579, 2016. MR3507373.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, June 2008. MR2417243.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on imaging sciences*, 2(1):183–202, 2009. MR2486527.

Régis Behmo, Paul Marcombes, Arnak S. Dalalyan, and Véronique Prinet. Towards optimal naive bayes nearest neighbor. In *Proceedings of ECCV 2010 : Part IV*, pages 171–184. Springer-Verlag, 2010.

Alexandre Belloni and Victor Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, May 2013. MR3037163.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root Lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, December 2011. MR2860324.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Pivotal estimation via square-root Lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788, 2014a. MR3210986.

Alexandre Belloni, Mathieu Rosenbaum, and Alexandre B. Tsybakov. An l_1, l_2, l_∞-regularization approach to high-dimensional errors-in-variables models. Technical Report CREST, *ArXiv e-prints*, 1412.7216, 2014b.

Karine Bertin, Erwan Le Pennec, and Vincent Rivoirard. Adaptive dantzig density estimation. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 47(1):43–74, February 2011. MR2779396.

Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, December 2008. MR2485008.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, August 2009. MR2533469.

Kris Boudt, Jonathan Cornelissen, and Christophe Croux. The gaussian rank correlation estimator: robustness properties. *Statistics and Computing*, 22(2):471–483, 2012. MR2865030.

Michael W. Brandt. Portfolio choice problems. In *Handbook of Financial Econometrics: Tools and Techniques*, volume 1 of *Handbooks in Finance*, chapter 5, pages 269–336. North-Holland, San Diego, 2010.

Peter Bühlmann and Sara A. van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. MR2807761.

Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Sparse density estimation with ℓ_1 penalties. In *Learning Theory*, volume 4539 of *Lecture Notes in Computer Science*, pages 530–543. Springer Berlin Heidelberg, 2007a. MR2397610.

Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007b. MR2312149.

Tony T. Cai and Weidong Liu. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566–1577, 2011. MR2896857.

Tony T. Cai and Harrison H. Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389–2420, October 2012. MR3097607.

Tony T. Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494): 594–607, February 2011. MR2847973.

Tony T. Cai, W. Liu, and H. H. Zhou. Estimating Sparse Precision Matrix: Optimal Rates of Convergence and Adaptive Estimation. *ArXiv e-prints*, 1212.2882, December 2012.

Tony T. Cai, Hongzhe Li, Weidong Liu, and Jichun Xie. Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, 100(1):139–156, March 2013. MR3034329.

Emmanuel J. Candès and Paige A. Randall. Highly robust error correction by convex programming. *Information Theory, IEEE Transactions on*, 54(7):2829–2840, 2008. MR2450835.

Emmanuel J. Candes and Terence Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005. MR2243152.

Emmanuel J. Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, December 2007. MR2382644.

Vince Carey, Li Long, and R. Gentleman. *RBGL: An interface to the BOOST graph library*, October 2015. URL `http://www.bioconductor.org`. R package version 1.46.0.

Safiye Celik, Benjamin Logsdon, and Su-In Lee. Efficient dimensionality reduction for high-dimensional network estimation. In *Journal of Machine Learning Research - W & CP 32(2) (ICML 2014)*, pages 1953–1961, 2014.

Mengjie Chen, Chao Gao, and Zhao Ren. Robust Covariance Matrix Estimation via Matrix Depth. *ArXiv e-prints*, 1506.00691, June 2015a.

Mengjie Chen, Zhao Ren, Hongyu Zhao, and Harrison Zhou. Asymptotically normal and efficient estimation of covariate-adjusted gaussian graphical model. *Journal of the American Statistical Association*, 2015b. To appear.

Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero. Shrinkage algorithms for mmse covariance estimation. *Signal Processing, IEEE Transactions on*, 58(10):5016–5029, October 2010. MR2722661.

Julien Chiquet, Yves Grandvalet, and Camille Charbonnier. Sparsity with sign-coherent groups of variables via the cooperative-lasso. *The Annals of Applied Statistics*, 6(2): 795–830, June 2012. MR2976492.

Andrzej Cichocki and Shun-ichi Amari. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010. MR2659408.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to algorithms*. MIT Press, Cambridge, MA, third edition, 2009. MR2572804.

David R. Cox and Nanny Wermuth. Linear dependencies represented by chain graphs. *Statistical Science*, 8(3):204–218, August 1993. MR1243593.

David R. Cox and Nanny Wermuth. *Multivariate dependencies: Models, analysis and interpretation*, volume 67. CRC Press, 1996. MR1456990.

Arnak S. Dalalyan and Yin Chen. Fused sparsity and robust estimation for linear models with unknown variance. In *Advances in Neural Information Processing Systems (NIPS) 25*, pages 1268–1276, 2012.

Arnak S. Dalalyan and Renaud Keriven. Robust estimation for an inverse problem arising in multiview geometry. *Journal of Mathematical Imaging and Vision*, 43(1):10–23, 2012. MR2910870.

Arnak S. Dalalyan, Mohamed Hebiri, Katia Meziani, and Joseph Salmon. Learning heteroscedastic models by convex programming under group sparsity. In *Journal of Machine Learning Research - W & CP 28(3) (ICML 2013)*, pages 379–387, 2013.

Alexandre d'Aspremont, Onureena Banerjee, and Laurent El Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30 (1):56–66, February 2008. MR2399568.

Arthur P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

Edsger W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959. MR0107609.

Edgar Dobriban and Jianqing Fan. Regularity properties for sparse regression. *Communications in Mathematics and Statistics*, 4(1):1–19, 2016. MR3475839.

David L. Donoho, Michael Elad, and Vladimir N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on*, 52(1):6–18, January 2006. MR2237332.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004. MR2060166.

Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination in sparse supervised learning. *Pacific Journal of Optimization*, 8(4):667–698, October 2012.

Noureddine El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, 36(6):2717–2756, December 2008. MR2485011.

Noureddine El Karoui. High-dimensionality effects in the markowitz problem and other quadratic programs with linear constraints: Risk underestimation. *The Annals of Statistics*, 38(6):3487–3566, December 2010. MR2766860.

Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Mind the duality gap: safer rules for the lasso. In *Journal of Machine Learning Research - W & CP 37 (ICML 2015)*, pages 333–342, 2015.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

Dean P. Foster and Edward I. George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4):1947–1975, December 1994. MR1329177.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, December 2007. MR2415737.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, July 2008.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. A note on the group lasso and a sparse group lasso. *ArXiv e-prints*, 1001.0736, January 2010a.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1–22, 2010b.

Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, October 2008. MR2466551.

E. Gautier and Alexandre B. Tsybakov. High-dimensional instrumental variables regression and confidence sets. *ArXiv e-prints*, 1105.2454, May 2011.

E. Gautier and Alexandre B. Tsybakov. Pivotal estimation in high-dimensional regression via linear programming. In *Empirical Inference*, pages 195–204. Springer, Heidelberg, 2013. MR3236866.

Seymour Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328, 1975.

Inc. Gurobi Optimization. Gurobi optimizer reference manual, 2015. URL `http://www.gurobi.com`.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, New York, 1986. MR0829458.

F. Han, H. Qiu, H. Liu, and B. Caffo. On the Impact of Dimension Reduction on Graphical Structures. *ArXiv e-prints*, 1404.7547, April 2014.

Nicholas J. Higham. Computing the nearest correlation matrix—a problem from finance. *IMA Journal of Numerical Analysis*, 22(3):329–343, 2002. MR1918653.

Kei Hirose and Hironori Fujisawa. Robust sparse Gaussian graphical modeling. *ArXiv e-prints*, 1508.05571, August 2015.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Peter J. Huber and Elvezio M. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, second edition, 2009. MR2488795.

Il'dar A. Ibragimov and Rafail Z. Has′minskiĭ. *Statistical estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, New York-Berlin, 1981. MR620321.

Jana Jankova, Sara A. van de Geer, et al. Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9(1):1205–1229, 2015. MR3354336.

Vojtvecech Jarník. *O jistém problému minimálním: (Z dopisu panu O. Borůskovi)*. Práce Moravské pvecrírodovvecedecké spoleveccnosti. Mor. pvecrírodovvecedecká spoleveccnost, 1930.

Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33 (1):117–128, 2011.

Seung-Jean Kim, Kwangmoo Koh, Michael Lustig, Stephen Boyd, and Dimitry Gorinevsky. An interior-point method for large-scale l1-regularized least squares. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):606–617, December 2007.

Hirohisa Kishino and Peter J. Waddell. Correspondence analysis of genes and tissue types and finding genetic links from microarray data. *Genome Informatics*, 11:83–95, 2000.

Olga Klopp and Alexandre B. Tsybakov. Estimation of matrices with row sparsity. *ArXiv e-prints*, 1509.00319, September 2015.

Olga Klopp, Karim Lounici, and Alexandre B. Tsybakov. Robust Matrix Completion. *ArXiv e-prints*, 1412.8132, December 2014.

Vladimir Koltchinskii and Karim Lounici. Concentration Inequalities and Moment Bounds for Sample Covariance Operators. *ArXiv e-prints*, 1405.2468, May 2014.

Joseph B. Kruskal, Jr. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7:48–50, 1956. MR0078686.

John Lafferty, Han Liu, and Larry Wasserman. Sparse nonparametric graphical models. *Statistical Science*, 27(4):519–537, 2012. MR3025132.

Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B):4254–4278, December 2009. MR2572459.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, October 2000. MR1805785.

Steffen L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1996. MR1419991. Oxford Science Publications.

Lucien Le Cam and Grace Lo Yang. *Asymptotics in statistics*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 2000. MR1784901.

Johannes Lederer. Trust, but verify: benefits and pitfalls of least-squares refitting in high dimensions. *ArXiv e-prints*, 1306.0113, 2014.

Johannes Lederer and Christian Müller. Don't Fall for Tuning Parameters: Tuning-Free Variable Selection in High Dimensions With the TREX. *ArXiv e-prints*, 1404.0541, April 2014.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, February 2004. MR2026339.

Robert S. Liptser and Albert N. Shiryaev. *Statistics of Random Processes: I. General Theory*, volume 5. Springer Science & Business Media, 2013. MR1800857.

Han Liu and Lie Wang. TIGER: A Tuning-Insensitive Approach for Optimally Estimating Gaussian Graphical Models. Technical report, *ArXiv e-prints*, 1209.2437, September 2012.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10:2295–2328, 2009. MR2563983.

Po-Ling Loh and Xin Lu Tan. High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *ArXiv e-prints*, 1509.07229, September 2015.

Po-Ling Loh and Martin J. Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *The Annals of Statistics*, 41(6):3022–3049, 12 2013. MR3161456.

Karim Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008. MR2386087.

Karim Lounici, Massimiliano Pontil, Sara A. van de Geer, and Alexandre B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, August 2011. MR2893865.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

Colin L. Mallows. Some comments on C_p. *Technometrics*, 15(4):661–675, 1973.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, March 1952.

Ricardo A. Maronna, Douglas R. Martin, and Victor J. Yohai. *Robust Statistics: Theory and Methods*. John Wiley & Sons, New York, 2006. MR2238141.

George Marsaglia. Conditional means and covariances of normal variables with singular covariance matrix. *Journal of the American Statistical Association*, 59(308):1203–1204, 1964.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1): 53–71, February 2008. MR2412631.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, June 2006. MR2278363.

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, February 2009. MR2488351.

Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparse multi-task and multi-class models. In *Advances in Neural Information Processing Systems (NIPS) 28*, pages 811–819, 2015.

Yurii Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. MR2142598.

Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1, Ser. A):127–152, 2005. MR2166537.

Yurii Nesterov. Gradient methods for minimizing composite objective function. Technical report, UCL, 2007.

Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. MR2968857.

Nam H. Nguyen and Trac D. Tran. Robust Lasso with missing and grossly corrupted observations. *Information Theory, IEEE Transactions on*, 59(4):2036–2058, 2013. MR3043781.

Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1–47, February 2011. MR2797839.

Brendan O'Donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. Operator splitting for conic optimization via homogeneous self-dual embedding. *ArXiv e-prints*, 1312.3039, 2013.

Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

Viktoria Öllerer and Christophe Croux. Robust high-dimensional precision matrix estimation. In Klaus Nordhausen and Sara Taskinen, editors, *Modern nonparametric, robust and multivariate methods*, pages 325–350. Springer International Publishing, 2015.

Robert C. Prim. Shortest connection networks and some generalizations. *Bell System Technology Journal*, 36:1389–1401, 1957.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, August 2010. MR2719855.

Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011. MR2836766.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org`.

Stephen Reid, Robert Tibshirani, and Jerome Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica*, 26:35–67, 2016.

Zhao Ren, Tingni Sun, Cun-Hui Zhang, and Harrison H. Zhou. Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, June 2015. MR3346695.

Peter Richtárik and Martin Takávecc. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144 (1-2, Ser. A):1–38, April 2014. MR3179953.

Martin Riedmiller and Heinrich Braun. Rprop - a fast adaptive learning algorithm. Technical report, Proc. of ISCIS VII), Universitat, 1992.

Mathieu Rosenbaum and Alexandre B. Tsybakov. Improved matrix uncertainty selector. In *From probability to statistics and back: high-dimensional models and processes*, volume 9

of *Inst. Math. Stat. (IMS) Collect.*, pages 276–290. Inst. Math. Statist., Beachwood, OH, 2013. MR3202640.

Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008. MR2417391.

Peter J. Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984. MR0770281.

Peter J. Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, 8:283–297, 1985. MR0851060.

Peter J. Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283, 1993. MR1245360.

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. *Information Theory, IEEE Transactions on*, 59(6):3434–3447, June 2013. MR3061256.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, March 1978. MR0468014.

Juliane Schäfer and Korbinian Strimmer. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005.

Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for l 1-regularized loss minimization. *Journal of Machine Learning Research*, 12:1865–1892, 2011. MR2819020.

Pablo Sprechmann, Ignacio Ramirez, Guillermo Sapiro, and Yonina C. Eldar. C-hilasso: A collaborative hierarchical sparse modeling framework. *Signal Processing, IEEE Transactions on*, 59(9):4183–4198, May 2011.

Nicolas Städler, Peter Bühlmann, and Sara A. van de Geer. ℓ_1-penalization for mixture regression models. *TEST*, 19(2):209–256, 2010a. MR2677722.

Nicolas Städler, Peter Bühlmann, and Sara A. van de Geer. Rejoinder: ℓ_1-penalization for mixture regression models. *TEST*, 19(2):280–285, 2010b. MR2677728.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, September 2012. MR2999166.

Tingni Sun and Cun-Hui Zhang. Sparse matrix inversion with scaled lasso. *Journal of Machine Learning Research*, 14(1):3385–3418, November 2013. MR3144466.

Garth Tarr, Samuel Müller, and Neville C. Weber. Robust estimation of precision matrices under cellwise contamination. *Computational Statistics & Data Analysis*, 93:404–420, January 2016. MR3406222.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. MR1379242.

Andreas M. Tillmann and Marc E. Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *Information Theory, IEEE Transactions on*, 60(2):1248–1259, February 2014. MR3164973.

Sara A. van de Geer. The deterministic Lasso. In *Proceedings of Joint Statistical Meeting*, 2007.

Sara A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009. MR2576316.

Lieven Vandenberghe, Stephen Boyd, and Shao-Po Wu. Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications*, 19(2):499–533, 1998. MR1614078.

Roman Vershynin. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012a. MR2956207.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012b. MR2963170.

Ferdinand T. Wang and David W. Scott. The ℓ_1 method for robust nonparametric regression. *Journal of the American Statistical Association*, 89(425):65–76, 1994. MR1266287.

Jun-Kun Wang and Shou-de Lin. Robust inverse covariance estimation under noisy measurements. In *Journal of Machine Learning Research - W & CP 32(2) (ICML 2014)*, pages 928–936, 2014.

Joe Whittaker. *Graphical models in applied multivariate statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Chichester, 1990. MR1112133.

Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11:2261–2286, January 2010. MR2719856.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B. (Statistical Methodology)*, 68(1):49–67, February 2006. MR2212574.

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007. MR2367824.

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the LASSO selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008. MR2435448.

Yu Zhang, Jianxin Wu, and Jianfei Cai. Compact representation for image classification: To choose or to compress? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 907–914, 2014.

Peng Zhao and Bin Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006. MR2274449.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. MR2279469.

Index of terms