Robust and Sparse Estimation of Large Precision Matrices

Ginette Lafit

Doctoral Thesis

Universidad Carlos III de Madrid

Department of Statistics

Advisors: Francisco J. Nogales

Ruben H. Zamar

July 2017

Abstract

The thesis considers the estimation of sparse precision matrices in the high-dimensional setting. First, we introduce an integrated approach to estimate undirected graphs and to perform model selection in high-dimensional Gaussian Graphical Models (GGMs). The approach is based on a parametrization of the inverse covariance matrix in terms of the prediction errors of the best linear predictor of each node in the graph. We exploit the relationship between partial correlation coefficients and the distribution of the prediction errors to propose a novel forward-backward algorithm for detecting pairs of variables having nonzero partial correlations among a large number of random variables based on i.i.d. samples. Then, we are able to establish asymptotic properties under mild conditions. Finally, numerical studies through simulation and real data examples provide evidence of the practical advantage of the procedure, where the proposed approach outperforms state-of-the-art methods such as the Graphical lasso and CLIME under different settings.

Furthermore, we study the problem of robust estimation of GGMs in the high-dimensional setting when the data may contain outlying observations. We propose a robust precision matrix estimator under the cellwise contamination mechanism that is robust against structural bivariate outliers. This framework exploits robust pairwise weighted correlation coefficient estimates, where the weights are computed by the Mahalanobis distance with respect to an affine equivariant robust correlation coefficient estimator. We show that the convergence rate of the proposed estimator is the same as the correlation coefficient used to compute the Mahalanobis distance. We conduct numerical simulation under different contamination settings to compare the graph recovery performance of different robust estimators. The proposed method is then applied to the classification of tumors using gene expression data. We show that our procedure can effectively recover the true graph under cellwise data contamination.

Acknowledgements

> "Kiss me while I calculate and calibrate and
> heaven's sake don't make me explain."
> — Fiona Apple

First, I would like to thank Javier Nogales for supervising my thesis. He gave me excellent help and support and was always open for questions. I would like to express my sincere gratitude to Ruben Zamar for his patience, motivation, and immense knowledge. I could not have imagined having better advisors and mentors for my Ph.D study. Their guidance helped me in all the time of research and writing of this thesis. They have the talent of creating a very motivating and reassuring atmosphere that wide my research from various perspectives. Besides, I would like to thank Marcelo Ruiz for his insightful comments, contributions and valuable ideas.

I am grateful to the Department of Statistics of Universidad Carlos III de Madrid, for providing the financial support to carry out this thesis. In particular, I want to show my gratitude to Susana Linares and Francisco García Saavedra for their help and administrative assistance.

Furthermore, I want to thank all my colleagues of the Department of Statistics for creating a familial working environment which made my time at the department truly enjoyable. In particular, I thank Andres Benchimol, Diego Ayma, María Guadarrama and Jorge Herrera for many stimulating discussions.

My time in Madrid was made enjoyable in large part due to the many friends that became a part of my life. I am grateful to my lovely friends: Gisela, Mar, Florinda, Elena and Gabri.

Last but not the least, I would like to thank my family: my parents and to my brother and sister and the rest of the family for their unconditional love and support throughout writing this thesis they have cherished with me every great moment and encourage me to be disobedient and to follow my intuitions.

To the memory of Lito

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Many statistical and practical problems require the estimation of different measures of linear dependence to infer whether there exists an association between a pair of variables, when we conditioned on the rest of them. Statistically, this is measured by the partial correlation coefficient. Thus, when two variables are linearly and conditionally associated, the partial correlation coefficient is different from zero (Edwards, 2000).

Let $\mathbf{x} = (X_1, ..., X_p)^T \in \mathbb{R}^p$ be a multivariate and centered vector with covariance matrix $\mathbf{\Sigma}$ and precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$. Then, the linear relation between two variables excluding the effect of the others can be measured by the *partial correlation coefficient* defined as follows:

Definition 1. *The partial correlation coefficient between two variables (X_i, X_j), condition on the rest of them and denoted by ρ_{ij}, is defined as the correlation coefficient between X_i and X_j when the effect of the other variables is eliminated.*

Assuming that the vector $\mathbf{x} = (X_1, ..., X_p)^T$ is multivariate Gaussian distributed with mean zero and covariance matrix $\mathbf{\Sigma}$. A notable result is that the partial correlation coefficients are closely related with the elements of the precision matrix $\mathbf{\Omega}$ (see Lauritzen, 1996). Hence, we can relate the nonzero entries in the precision matrix with the nonzero partial correlation coefficients. This procedure was first proposed by Dempster (1972) and is denoted by *Covariance Selection*. Let ω_{ij} be the ij-element of the precision matrix $\mathbf{\Omega}$. Then, the partial correlation coefficient between variables X_i and X_j is given by the following expression:

$$\rho_{ij} = -\frac{\omega_{ij}}{[\omega_{ii}\omega_{jj}]^{1/2}} \quad for\ i, j = 1, ..., p. \tag{1.1.1}$$

A problem that is closely related with covariance selection is the recovery of the support of the precision matrix (i.e. non-zero elements of $\mathbf{\Omega}$). This problem is connected with model selection in undirected graphical models. An undirected graph, denoted by $\mathcal{G} = (V, E)$ is defined as the set of p vertices, denoted by V, that represents the p variables and the set of edges, given by $E \subseteq V \times V$. The undirected graph establishes that if the variables X_i and X_j are connected, then the pair $(i, j) \in E$, and the variables X_i and X_j are adjacent. Furthermore, the edge (i, j) is excluded from E if and only if X_i and X_j are independent given $(X_k, k \in V \setminus \{i, j\})$. In particular, if $\mathbf{x} = (X_1, ..., X_p)^T$ is multivariate Gaussian distributed then, the conditionally independence between X_i and X_j is equivalent to $\omega_{ij} = 0$. Thus, recovering the structure of the graph $\mathcal{G} = (V, E)$ is equivalent to estimate the non-zero elements in $\mathbf{\Omega}$. This is known in the literature as *Gaussian Graphical Models* (GGMs) (Lauritzen, 1996).

Given an independent and identically distributed random sample $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ from the distribution of $\mathbf{x}$, the most natural estimator of $\mathbf{\Sigma}$ is the empirical covariance matrix defined as:

$$\mathbf{S} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^T, \tag{1.1.2}$$

where $\bar{\mathbf{x}} = n^{-1}\sum_{k=1}^{n}\mathbf{x}_k$. For data sets in which the ratio between the dimension p and the number of available observation n is less than one and negligible, the inverse of the empirical covariance matrix can be used as an estimate for $\mathbf{\Omega}$. However, when $p > n$, $\mathbf{S}$ is not invertible. Moreover, when the ratio p/n is approximately one, $\mathbf{S}$ is still invertible but ill-conditioned, meaning that its inverse will amplify the estimation error, which can be observed by the presence of small eigenvalues (Ledoit and Wolf, 2004). Moreover, from the asymptotic point of view, when both n and p are large (i.e. $p = O(n)$), the empirical covariance matrix is not a consistent estimator (see El Karoui, 2008). Therefore, in the high-dimensional setting, traditional methods that relies on the optimization of a discrete function (Speed and Kiiveri, 1986; Lauritzen, 1996; Edwards, 2000) do not work well due to the lack of a pivotal estimator like the empirical covariance matrix. Hence, different methods focus on obtaining an estimator of $\mathbf{\Sigma}$ that can be inverted and is well-conditioned.

Moreover, in the high-dimensional setting several covariance selection procedures focus on the assumption that the precision matrix is sparse, that is, $\mathbf{\Omega}$ is mostly composed by zero elements. This suggests that even when we are dealing with $p = O(n)$, the dimension of the problem may still be tractable since the number of edges will grow slowly than the number of observations (Meinshausen and Bühlmann, 2006).

Sparse Gaussian Graphical Models have been apply to the construction of gene regulatory networks (see Wille et al., 2004; Li and Gui, 2006; Kiiveri, 2011). In the genetic literature, a well known result is that the process in which the cell controls the interaction between RNAs and proteins can be theoretically represented as a network. This approach explores dependence relations between genes through the estimation of the corresponding precision matrix. Moreover, this framework assume that the patterns of variation in gene expressions will be predicted by a small subset of other genes. Hence, genetic networks are not fully connected, suggesting that genetic networks are intrinsically sparse and hence the associated precision matrix is mostly composed by zero elements outside its principal diagonal.

Finally, We have to note that there is a growing literature on *Direct Acyclic Graphical Models* or *Bayesian Networks* (see Spirtes et al., 2000; Kalisch and Bühlmann, 2007; Bühlmann et al., 2010). These are defined as graphical models in which the edges have directional arrows but not directed cycles (see Chapter 2 Lauritzen, 1996). However, in the present thesis we focus on the estimation of GGM where all edges are undirected. For the remainder of this chapter we review existing approaches to estimate the precision matrix and we present the organization and outline of the thesis.

1.2 Precision Matrix Estimation

Existing methods to estimate undirected GGM in the high-dimensional setting can be classified in two classes: the nodewise regression methods and maximum likelihood methods. The nodewise regression method was proposed by Meinshausen and Bühlmann (2006). This method estimate for each node in the graph the set of partial correlated variables. Penalized likelihood methods include Yuan and Lin (2007), Banerjee et al. (2008) and Friedman et al. (2008), among others. These methods propose to estimate the precision matrix by penalizing the log-likelihood function.

Notation. Given a vector $\mathbf{v} \in \mathbb{R}^p$ and parameter $a \in [1, \infty)$, we use $\| \mathbf{v} \|_a$ to denote the usual ℓ_a norm. We consider $\| \mathbf{v} \|_1 = \sum_{i=1}^p |v_i|$, $\| \mathbf{v} \|_2 = \sqrt{\sum_{i=1}^p v_i^2}$ and $\| \mathbf{v} \|_\infty = \max_i |v_i|$. Given a matrix $\mathbf{U} = (u_{ij}) \in \mathbb{R}^{p \times p}$ we define the elementwise ℓ_∞ norm $\| \mathbf{U} \|_\infty = \max_{1 \leq i \leq p, 1 \leq j \leq p} |u_{ij}|$, the spectral norm $\| \mathbf{U} \|_2 = \sup_{\|\mathbf{x}\|_2 \leq 1} \| \mathbf{Ux} \|_2$, the matrix ℓ_1 norm $\| \mathbf{U} \|_{L_1} = \max_{1 \leq j \leq p} \sum_{i=1}^p |u_{ij}|$, the Frobenius norm $\| \mathbf{U} \|_F = \sqrt{\sum_{i,j} u_{ij}^2}$ and the elementwise ℓ_1 norm $\| \mathbf{U} \|_1 = \sum_{i=1}^p \sum_{j=1}^p |u_{ij}|$. The notation $\mathbf{U} \succ 0$ indicates that $\mathbf{U}$ is positive definite.

1.2.1 Estimation by Multiple Regression

The methods that rely on estimation by multiple regression aim to detect pair of variables with non-zero partial correlation based on independent identically distributed samples. Meinshausen and Bühlmann (2006) develop a procedure, called *Neighborhood Selection*, that uses lasso penalty (Tibshirani, 1996) to estimate the conditional independence structure of a set of variables $\mathbf{x} = (X_1, ..., X_p)^T$ which are Gaussian distributed with mean zero and covariance $\mathbf{\Sigma}$. The neighborhood of a variable is defined in the following way:

Definition 2. *The neighborhood of a node $i \in V$, denotes by $\mathcal{A}_i$, is the smallest subset of the set of vertices that not contain the node i (i.e. $V \setminus \{i\}$) such that, given all the nodes in the neighborhood, X_i is conditionally independent of all the remaining variables that do not belong to the subset $\mathcal{A}_i$.*

For each node $i \in V$, the optimal predictor for X_i, $\beta^i \in \mathbb{R}^p$, is defined as

$$\boldsymbol{\beta}^i = \underset{\beta \in \mathbb{R}^p, \beta_i = 0}{\arg\min} \parallel X_i - \sum_{j=1}^{p} \beta_j X_j \parallel_2^2, \tag{1.2.1}$$

Meinshausen and Bühlmann (2006) show that

$$\beta_j^i = -\frac{\omega_{ij}}{\omega_{ii}}. \tag{1.2.2}$$

Given an identically and independent distributed random sample $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the Neighborhood Selection approach aims to estimate individually the neighborhood of a variable X_i assuming that the precision matrix $\mathbf{\Omega}$ is sparse. The Neighborhood Selection procedure estimates for each node $i \in V$ the vector of coefficients $\boldsymbol{\beta}^i \in \mathbb{R}^p$ by minimizing the following lasso regression for each variable:

$$\hat{\boldsymbol{\beta}}^i = \underset{\beta \in \mathbb{R}^p, \beta_i = 0}{\arg\min} \left\{ n^{-1} \sum_{k=1}^{n} \left(X_{ki} - \sum_{j=1}^{p} \beta_j X_{kj} \right)^2 + \lambda_1 \parallel \boldsymbol{\beta} \parallel_1 \right\}, \tag{1.2.3}$$

where $\lambda_1 > 0$ is the penalty parameter in the lasso regression and $\| \boldsymbol{\beta} \|_1 = \sum_{j=1}^{p} |\beta_j|$ is the ℓ_1 norm of the coefficient vector. Hence, Meinshausen and Bühlmann (2006) define the estimated set of neighborhoods for node $i \in V$ as:

$$\hat{\mathcal{A}}_i = \{j \in V : \hat{\beta}^i_j \neq 0\}, \tag{1.2.4}$$

Hence, neighborhood selection estimates p separate lasso regressions. The elements of $\boldsymbol{\Omega}$ are estimated to be non-zero if $\hat{\beta}^i_j \neq 0$ or $\hat{\beta}^j_i \neq 0$. The solution will depend on the penalty parameter λ_1, large values will tend to increase the set of variables included in the neighborhood, while low values will shrink the set.

Meinshausen and Bühlmann (2006) prove that under certain conditions it is possible to estimate consistently the neighborhood of each node in the graph when the dimension grows exponentially as a function of the sample size. Hence, the Neighborhood Selection approach has an exponentially fast convergence rate and the conditional independence structure of a multivariate Gaussian distribution can be estimated consistently by combining the neighborhood estimates for all variables.

The neighborhood selection procedure is a simple method to recover the neighborhood of a variable consistently under a set of suitable assumptions. Moreover is also computationally fast for large data sets. However, it does not take into account the intrinsic symmetry of $\boldsymbol{\Omega}$. This could produce a loss in efficiency and contradictory neighborhoods. Also, the same penalty is used for all the p lasso regressions, which is only efficient when the distribution of the network is nearly uniform (Peng et al., 2009).

Peng et al. (2009) develop a procedure, called *Sparse Partial Correlation Estimation* (SPACE), that is based on a joint sparse regression model that simultaneously perform neighborhood selection and preserves symmetry. SPACE can detect pairs of variables that have non-zero partial correlations among a large number of random variables based on independent identically distributed samples with mean zero and covariance matrix $\mathbf{\Sigma}$. Using the fact that X_i can be expressed as $X_i = \sum_{j\neq i} \beta_j^i X_j + \varepsilon_i$, such that ε_i is uncorrelated with X_{-i}, we can write β_j^i in terms of the coefficients of the precision matrix $\mathbf{\Omega}$:

$$\beta_j^i = -\frac{\omega_{ij}}{\omega_{ii}} = \rho_{ij}\sqrt{\frac{\omega_{jj}}{\omega_{ii}}}, \tag{1.2.5}$$

where ρ_{ij} is the partial correlation coefficient and ω_{ij} the coefficients of the precision matrix $\mathbf{\Omega}$.

Given this result, Peng et al. (2009) define the following loss function

$$L_n(\boldsymbol{\theta}, \mathbf{\Omega}, \mathbf{X}) = \frac{1}{2}\left(\sum_{i=1}^{p} w_i \sum_{k=1}^{n}\left(X_{ki} - \sum_{j\neq i}\rho_{ij}\sqrt{\frac{\omega_{jj}}{\omega_{ii}}}X_{kj}\right)^2\right), \tag{1.2.6}$$

where $\boldsymbol{\theta} = (\rho_{12}, ..., \rho_{(p-1)p})$, $\boldsymbol{\omega} = \{\omega_{ii}\}_{i=1}^{p}$, $\mathbf{X} = \{\mathbf{x}_k\}_{k=1}^{n}$ and $\mathbf{w} = \{w_i\}_{i=1}^{p}$ are nonnegative weights.

The partial correlation vector $\boldsymbol{\theta} = (\rho_{12}, ..., \rho_{(p-1)p})$ is estimated by minimizing the following penalized loss function

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\omega}, \mathbf{X}) = L_n(\boldsymbol{\theta}, \boldsymbol{\omega}, \mathbf{X}) + \lambda_2 \parallel \boldsymbol{\theta} \parallel_1, \tag{1.2.7}$$

where λ_2 is the regularization parameter and $\parallel \boldsymbol{\theta} \parallel_1 = \sum_{1\leq i<j\leq p} |\rho_{ij}|$ is an ℓ_1 penalty.

There are three different types of weights that can be considered: (1) uniform weights , $w_i = 1$, (2) residual variance based weights $w_i = \hat{\omega}_{ii}$ and (3) degree based weights, where w_i is proportional to the estimated degree of X_i (i.e. this is related with the number of variables such that $\hat{\rho}_{ij} \neq 0$ for $j \neq i$).

Peng et al. (2009) show that under certain conditions the SPACE method produces consistent estimates and can identify the correct neighborhood when both $n \to \infty$ and $p \to \infty$.

1.2.2 Penalized Maximum Likelihood

Under the assumption that $\mathbf{x} = (X_1, ..., X_p)^T$ are normally distributed with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The objective of the following procedures is to estimate the precision matrix $\boldsymbol{\Omega}$ imposing some type of penalization on the likelihood function to control the number of zeros in the precision matrix.

Yuan and Lin (2007) propose a method to estimate sparse graphical models. Assuming that the data is multivariate Gaussian, a ℓ_1 penalty is imposed on the likelihood function to estimate $\boldsymbol{\Omega}$. Therefore, the objective is to maximize the following ℓ_1 regularized log-likelihood function:

$$\max_{\boldsymbol{\Omega} \succ 0} \left\{ log|\boldsymbol{\Omega}| - tr(\mathbf{S}\boldsymbol{\Omega}) - \lambda_3 \sum_{i=1}^{p} \sum_{j=1}^{p} |\omega_{ij}| \right\}. \tag{1.2.8}$$

where $\lambda_3 > 0$ is the regularization parameter.

From the optimization problem in (1.2.8), we can define the sub-gradient equation for minimizing the log-likelihood

$$\mathbf{W} - \mathbf{S} - \lambda_3 \boldsymbol{\Lambda} = 0, \tag{1.2.9}$$

where $\mathbf{W} = \boldsymbol{\Omega}^{-1}$, $\Lambda_{ij} \in sign(\omega_{ij})$ (i.e. $\Lambda_{ij} = sign(\omega_{ij})$ if $\omega_{ij} \neq 0$ and $\boldsymbol{\Omega}_{ij} = 0$ if $\Lambda_{ij} \in [-1, 1]$).

Given that the problem in (1.2.8) is convex, Banerjee et al. (2008) consider the estimation of the covariance matrix $\boldsymbol{\Sigma}$. Assuming that W is an estimate of $\boldsymbol{\Sigma}$, they show that the solution can be obtained by optimizing over each row and corresponding column of W in a block coordinate descending fashion. To illustrate the idea, if the matrix W and S are partitioned in the following way

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{w}_{12} \\ \mathbf{w}_{12}^T & w_{22} \end{bmatrix} \quad \mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{s}_{12} \\ \mathbf{s}_{12}^T & s_{22} \end{bmatrix},$$

the solution for $\mathbf{w}_{12}$ is given by:

$$\mathbf{w}_{12} = \arg\min_{\mathbf{y}} \left\{ \mathbf{y}^T \mathbf{W}_{11}^{-1} \mathbf{y} : \| \mathbf{y} - \mathbf{s}_{12} \|_\infty \leq \lambda_3 \right\}. \tag{1.2.10}$$

Which is identical to solve the following dual problem:

$$\mathbf{w}_{12} = \underset{\boldsymbol{\beta}}{min} \left\{ \frac{1}{2} \| \mathbf{W}_{11}^{1/2} \boldsymbol{\beta} - \mathbf{b} \|^2 + \lambda_4 \| \boldsymbol{\beta} \|_1 \right\}, \tag{1.2.11}$$

where $\mathbf{b} = \mathbf{W}_{11}^{-1/2} s_{12}$. Consequently, if $\boldsymbol{\beta}$ solves the problem in (1.2.11) then $\mathbf{w}_{12} = \mathbf{W}_{11}\boldsymbol{\beta}$ solves equation (1.2.10).

Friedman et al. (2008) state that the problem in (1.2.11) resembles a lasso regression and is similar to the first order condition, given by equation (1.2.9), when optimizing the log-likelihood imposing and ℓ_1 penalty. To illustrate the similarities between the two approaches, we can partition $\mathbf{W}$ and $\boldsymbol{\Omega}$ in the following way:

$$\mathbf{W}\boldsymbol{\Omega} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{w}_{12} \\ \mathbf{w}_{12}^T & w_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{12}^T & \boldsymbol{\Omega}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{(p-1)} & \mathbf{0}_{(p-1)\times p} \\ \mathbf{0}_{p\times(p-1)} & 1 \end{bmatrix}. \tag{1.2.12}$$

The sub-gradient equation (1.2.9) can be written in terms of the elements of the partition matrices as

$$\mathbf{w}_{12} - \mathbf{s}_{12} - \lambda_4 \boldsymbol{\gamma}_{12} = 0. \tag{1.2.13}$$

Given the result in Banerjee et al. (2008) we can conveniently write the problem in (1.2.13) as a lasso regression problem:

$$\mathbf{W}_{11}\boldsymbol{\beta} - \mathbf{s}_{12} + \lambda_4 \mathbf{v} = 0, \tag{1.2.14}$$

where $\boldsymbol{\beta} = \mathbf{W}_{11}^{-1}\mathbf{w}_{12}$ and $\mathbf{v} = \boldsymbol{\gamma}_{12}$ such that $\mathbf{v} \in sign(\boldsymbol{\beta})$ elementwise.

If we set $\mathbf{W}_{11} = \mathbf{S}_{11}$, the estimates of the problem in (1.2.8) are the same to those proposed by Meinshausen and Bühlmann (2006), in which a lasso regression is fitted to each of the p variables taking as regressors the rest $(p-1)$ variables. However, Banerjee et al. (2008) show that this approach does not yield the maximum likelihood estimator, since in general $\mathbf{W}_{11} \neq \mathbf{S}_{11}$. Therefore, to solve this problem, Friedman et al. (2008) propose a new algorithm, the *Graphical Lasso* or *Glasso*, that uses an estimate of the upper block of $\mathbf{W}$ instead of using $\mathbf{S}_{11}$ directly. The Glasso algorithm works in the following manner:

1. Start with $\mathbf{W} = \mathbf{S} + \lambda_4 \mathbf{I}_p$. The diagonal of $\mathbf{W}$ remains unchanged in the following steps.

2. For each $j = 1, 2, ..., p, 1, 2, ..., p, ...$ solve the Lasso problem in (1.2.11), that takes as input the inner products $\mathbf{W}_{11}$ and $\mathbf{s}_{12}$. This gives a $(p-1)$ vector of solution $\widehat{\boldsymbol{\beta}}$. Fill in the corresponding row and column of $\mathbf{W}$ using $\mathbf{w}_{12} = \mathbf{W}_{11}\widehat{\boldsymbol{\beta}}$.

3. Continue until convergence.

Step 2 implies a permutation of the rows and columns to make the target column the last. Thus each lasso problem can be efficiently solved by a coordinate descent algorithm (see Friedman et al., 2008). Since, the estimated $\widehat{\boldsymbol{\beta}}$ will be sparse the computation of $\mathbf{w}_{12} = \mathbf{W}_{11}\widehat{\boldsymbol{\beta}}$ will be fast. The Glasso algorithm will estimate $\widehat{\boldsymbol{\Sigma}} = \mathbf{W}$, and the precision matrix can be recovered solving the system of equations in (1.2.12).

A disadvantage of the Graphical lasso algorithm is that the ℓ_1 penalty tends to produce biases even in the simple regression setting, due to the linear increase of penalty on regression coefficients. Fan et al. (2009) propose two different procedures to remedy this situation, one of them relies on non-concave penalties, the *Smoothly Clipped Absolute Deviation* (SCAD), and the other is the *Adaptive Lasso*.

Fan and Li (2001) consider that a good penalty function should be able to produce and estimator with three properties: (1) the estimator should be unbiased when the true parameter is large to avoid unnecessary modeling bias. (2) It should produce sparsity, this implies that the resulting estimator is a thresholding rule which automatically sets small estimated coefficients to zero to reduce model complexity. (3) The resulting estimator should be continuous in data to avoid instability in model prediction. Fan and Li (2001) and Fan et al. (2009) show that these properties are achieved by the SCAD penalty, which is symmetric on a quadratic spline $[0, \infty)$ where the first order derivative is given by

$$SCAD'_{\lambda_5,a} = V\Big\{I(|x| < \lambda_5) + \frac{(a\lambda_5 - |x|)+}{(a-1)\lambda_5}\Big\}, \tag{1.2.15}$$

for $x \geq 0$, where $\lambda_5 > 0$ and $a > 2$ are two tuning parameters. When $a \to \infty$ equation (1.2.15) corresponds to the ℓ_1 penalty. This penalty function leaves large values of $|x|$ not excessively penalized and makes the solution continuous. Using the SCAD penalty, the optimization problem is given by

$$\max_{\mathbf{\Omega}\succ 0}\left\{log|\mathbf{\Omega}| - tr(\mathbf{S\Omega}) - \sum_{i=1}^{p}\sum_{j=1}^{p} SCAD_{\lambda_5,a}(|\omega_{ij}|)\right\}. \tag{1.2.16}$$

Fan and Li (2001) demonstrate that the rates of convergence for the penalized likelihood estimators depend on the regularization parameter. While for the ℓ_1 penalized likelihood the oracle property does not hold, the SCAD penalty performs as well as the oracle procedure in terms of selecting the correct model.

Another procedure that achieves the three properties mention before is the Adaptive Lasso penalty (Zou, 2006), which set a different weight to each component, thus the problem is to optimize the following restricted log-likelihood function:

$$\max_{\mathbf{\Omega}\succ 0}\left\{log|\mathbf{\Omega}| - tr(\mathbf{S\Omega}) - \lambda_6\sum_{i=1}^{p}\sum_{j=1}^{p} w_{ij}|\omega_{ij}|\right\}, \tag{1.2.17}$$

where $w_{ij} = 1/|\widetilde{\omega}_{ij}|^{\gamma}$ for some $\gamma > 0$ and any consistent estimator $\widetilde{\mathbf{\Omega}} = [\widetilde{\omega}_{ij}]_{1 \leq i,j \leq p}$, the initial estimates of $\widetilde{\mathbf{\Omega}}$ can be obtained by estimating the precision matrix applying a lasso penalty.

The SCAD penalty is a more flexible approach than the Adaptive Lasso, since an element being estimated zero can escape from zero in the next iteration. While the Adaptive Lasso absorbs zeros in each iteration producing estimates that are always sparser than the initial values (see Fan et al., 2009).

Finally, another method that considers the dual problem of optimizing a ℓ_1 penalized likelihood function to estimate $\mathbf{\Omega}$ is developed by Cai et al. (2011) and denoted by *Constrained ℓ_1-minimization for Inverse Matrix Estimation* (CLIME). Let $\widehat{\mathbf{\Omega}}_1$ to be the solution of the following optimization problem:

$$min \parallel \mathbf{\Omega} \parallel_1 \quad subject\ to\ \mid \mathbf{\Sigma\Omega} - I \mid_{\infty} \leq \lambda_7 \quad \mathbf{\Omega} \in \mathbb{R}^{p \times p}, \tag{1.2.18}$$

where λ_7 is a tuning parameter. This problem does not impose the symmetry condition on $\mathbf{\Omega}$, thus the solution, $\widehat{\omega}^1_{ij}$, is not symmetric. Therefore, the final CLIME estimator of the precision matrix is obtained in the following way:

$$\widehat{\omega}_{ij} = \widehat{\omega}_{ji} = \widehat{\omega}^1_{ij} 1(\mid \widehat{\omega}^1_{ij} \mid \leq \widehat{\omega}^1_{ij}) + \widehat{\omega}^1_{ji} 1(\mid \widehat{\omega}^1_{ij} \mid > \widehat{\omega}^1_{ij}). \tag{1.2.19}$$

This implies that between $\widehat{\omega}^1_{ij}$ and $\widehat{\omega}^1_{ji}$, the one that is selected is the one with the smallest magnitude. The method will also guarantee that $\widehat{\mathbf{\Omega}}$ is positive definite.

Cai et al. (2011) prove that the rate of convergence of CLIME outperforms the rates of convergence for ℓ_1 Penalized Maximum Likelihood estimators and also satisfies that

$$\parallel \widehat{\mathbf{\Omega}} - \mathbf{\Omega} \parallel_2 = O_p \left(\sqrt{\frac{log\ p}{n}} \right). \tag{1.2.20}$$

1.3 Robust Precision Matrix Estimation

One of the main drawback of the popular procedures to estimate the precision matrix is that they are not well suited to handle noisy data (contaminated by outliers). The existing approaches to estimate the precision matrix and recover the support of the GGM use as input the empirical covariance matrix. The empirical covariance and correlation matrix estimates are very sensitive to the presence of multidimensional outliers (Alqallaf et al., 2002). The violation of the Gaussian assumption may result in poor recovery of the GGM and biased estimation of the precision matrix (see Finegold and Drton, 2011; Liu et al., 2012; Sun and Li, 2012). Moreover, in the high-dimensional setting, the fraction of perfectly observed rows may be very small. If all components of a row have an independent chance of being contaminated, then the probability that a case is perfectly observed is small. To deal with outliers in the high-dimensional setting, Alqallaf et al. (2009) propose a contamination model where the contamination in each variable is independent from other variables (i.e. componentwise outliers). It allows for cellwise contamination that can be applied to explain the contamination mechanism in Microarrays experiments (see Troyanskaya et al., 2001; Liu et al., 2003). The cellwise contamination model lacks the affine equivariant property. Henceforth, existing approaches for robust covariance estimation such as M-estimates (Hampel, 1973; Maronna, 1976), Minimum Volume Ellipsoid (MVE) and Minimum Covariance Determinant (MCD) estimators (Rousseeuw, 1985, 1984) and the Stahel-Donoho (SD) estimators (Stahel, 1981; Donoho, 1982), may not be reliable in high-dimensional data sets.

To deal with outliers in high-dimensional data sets, many procedures construct robust covariance and correlation matrices by using pairwise robust correlation coefficients. Liu et al. (2009) and Liu et al. (2012) propose to apply a univariate monotone transformation to make the data Gaussian distributed. Then, a robust precision estimator of the correlation matrix can be computed from the transformed data. The main drawback of these procedures is that they are not robust under the presence of structural bivariate outliers which could lead to a misleading graph support recovery.

1.4 Organization and Outline of the Thesis

The thesis is structure as follows. In Chapter 2 we present an approach to estimate undirected graphs and to perform model selection in high dimensional Gaussian Graphical Models. We consider a parametrization of the precision matrix in terms of the prediction errors of the best linear predictor of each node in the graph. We exploit the relationship between partial correlation coefficients and the distribution of the prediction errors. We propose a novel forward-backward algorithm for detecting pairs of variables having nonzero partial correlations among a large number of random variables based on i.i.d. samples. We establish asymptotic properties under mild conditions. The proposed algorithm outperforms the existing methods, such as the Graphical lasso and CLIME, when we compare the graph recovery and numerical performance under different settings. Numerical studies through simulation and real data examples provide evidence of the theoretical advantage of the procedure.

Chapter 3 is concerned with robust estimation of Gaussian Graphical Models in the high-dimensional setting when the data may contain outlying observations. These outliers can lead to drastically wrong inference on the intrinsic graph structure. Several procedures apply univariate transformations to make the data Gaussian distributed. However, these transformations do not work well under the presence of structural bivariate outliers. We propose a robust precision matrix estimator under the cellwise contamination mechanism that is robust against structural bivariate outliers. This estimator exploits robust pairwise weighted correlation coefficient estimates, where the weights are computed by the Mahalanobis distance with respect to an affine equivariant robust correlation coefficient estimator. We show that the convergence rate of the proposed estimator is the same as the correlation coefficient used to compute the Mahalanobis distance. We conduct numerical simulation under different contamination settings to compare the graph recovery performance of different robust estimators. Finally, the proposed method is then applied to the classification of tumors using gene expression data. We show that our procedure can effectively recover the true graph under cellwise data contamination. Chapter 4 concludes the thesis.

Chapter 2

A Stepwise Approach for High-Dimensional Gaussian Graphical Models

High-dimensional Gaussian Graphical models (GGMs) have been widely used to represent the linear dependency between variables. The idea underlying GGMs is to measure linear dependencies by estimating partial correlations to infer whether there is an association between a pair of variables, conditionally on the rest of them. Moreover, there is a close relation between the nonzero partial correlation coefficients and the nonzero entries in the inverse of the covariance matrix (Lauritzen, 1996; Edwards, 2000). This procedure is known as covariance selection and is widely used to identify the conditional independence in an undirected graph from a set of independently indentically distributed observations (Dempster, 1972).

In a high-dimensional framework, when the dimension p is larger than the number of available observations n, the sample covariance matrix is badly conditioned and its inverse tends to amplify the estimation error (Ledoit and Wolf, 2004). From the asymptotic point of view, when both n and p are large (i.e. $p = O(n)$), the sample covariance matrix is not a consistent estimator (El Karoui, 2008). To deal with this problem, several covariance selection procedures have been proposed based on the assumption that the inverse of the covariance matrix (i.e. precision matrix) is sparse. This implies that most of the variables are conditionally independent.

Existing methods to estimate the GGM can be classified in three classes: the nodewise regression methods, maximum likelihood methods and limited order partial correlations methods. The nodewise regression method was proposed by Meinshausen and Bühlmann (2006). This method estimate a lasso regression for each node in the graph. Peng et al. (2009) present a procedure that simultaneously performs neighborhood selection for all variables to estimate joint sparse regressions, applying an active-shooting to solve the lasso. Yuan (2010) replaces the lasso regression with a Dantzig selector. Liu and Wang (2012) propose an asymptotically tuning-free procedure that estimates the precision matrix in a column-by-column fashion. Zhou et al. (2011) propose an estimator for the precision matrix base on ℓ_1 regularization and thresholding to infer a sparse undirected graphical model. Ren et al. (2015) propose a nodewise regression approach to obtain asymptotically efficient estimation of each entry of the precision matrix under sparseness conditions.

Penalized likelihood methods include Yuan and Lin (2007), Banerjee et al. (2008) and Friedman et al. (2008), among others. These methods propose to estimate the precision matrix by penalizing the log-likelihood function. Friedman et al. (2008) propose the Graphical lasso (Glasso) procedure to estimate sparse precision matrices fitting a modified lasso regression to each variable and solving the problem by a coordinate descent algorithm. Rates of convergence under the Frobeniuos norm were study by Rothman et al. (2008). Ravikumar et al. (2008) and Ravikumar et al. (2011) obtain convergence rates under the elementwise ℓ_∞ norm and the spectral norm assuming subgaussian distributions. Lam and Fan (2009) and Fan et al. (2009) propose methods to diminish the bias imposed by the ℓ_1 penalty by introducing a non-convex

SCAD penalty. Cai et al. (2011) propose an estimator called CLIME that estimate precision matrices for both sparse and non-sparse graphs, without imposing a specific sparsity pattern, by solving the dual of an ℓ_1 penalized maximum likelihood problem. They establish convergence rates under the elementwise ℓ_∞ norm and Frobenius norm. A greedy forward-backward algorithm to optimize a Gaussian log-likelihood loss was proposed by Johnson et al. (2011). This algorithm starts with an empty set of active variables and adds (and removes) variables to the active set. In doing so they use a greedy coordinate ascent algorithm, this implies optimizing a quadratic function for every pair of nodes in each iteration. They show that the greedy algorithm requires a restricted eigenvalue condition on the true precision matrix. This condition is weaker than the irrepresentable condition impose by the ℓ_1 regularized log-likelihood methods (see Ravikumar et al., 2011).

Limited order partial correlation procedures use lower order partial correlations to test for conditional independence relations. Spirtes et al. (2000) propose the PC-algorithm, which works in an iterative procedure, it starts with a complete undirected graph and deletes edges based on conditional independence decisions. This algorithm works in the worst case in exponential time, but under sparsity assumptions the computational complexity is reduced to polynomial time. Kalisch and Bühlmann (2007) and Rütimann et al. (2009) modify the PC-algorithm to estimate direct acyclic graphs. However, their approach has a large degree of computational complexity. Liang et al. (2015) propose an equivalent measure of partial correlations based on the Markov property and adjacency faithfulness to estimate partial correlation coefficients. Huang et al. (2016) propose a partial correlation screening approach that uses a screen step and a clean step. In the screen step, the algorithm select a reduce neighborhood set for each node using a stagewise algorithm. In the clean step, the algorithm removes false positives and uses the resultant neighborhood set to reconstruct each row of the precision matrix.

In this article, we present an approach to estimate partial correlations and to perform model selection in high dimensional Gaussian Graphical Models (GGM) based on a forward-backward algorithm. Our method is motivated by the relation between the partial correlation coefficients and the elements of the precision matrix. We propose to parametrize the precision matrix in terms of the prediction errors of the best linear predictor of each variable. Hence, we convert the original problem of estimating the precision matrix into that of the covariance matrix estimation. We estimate the edge set applying a forward-backward algorithm. The algorithm begins with an empty edge set and gradually adds and removes edges from the edge set. In the forward step the algorithm finds the best next candidate given by the pair of variables with the largest absolute empirical partial correlation coefficient, otherwise the algorithm terminates. In the backward step the algorithm removes the unlikely edges. The maximization of this coefficient is related to the maximization of a empirical information divergence measure. We the call this procedure *Graphical Stepwise*. Compare with existing methods, our approach is able to provide a set of edges associated with the largest absolute partial correlation coefficients for a given threshold. Moreover, under mild conditions we show that the Graphical Stepwise procedure is able to consistently estimate the true set of edges.

The rest of the chapter is organized as follows. In the next section we present a parametrization of the precision matrix which exploits the relation between the elements of the precision matrix and a system of linear regressions. In Section 2.2 we present the Graphical Stepwise procedure to estimate a set of edges with a forward-backward algorithm. In Section 2.3 we establish the consistency of the proposed approach. Section 2.4 presents simulation results and real data analysis. We compare the numerical and classification performance of the method with that of Glasso and CLIME. In Section 2.5 we conclude the article with a brief discussion. Proof of the main results are presented in Section 2.6.

2.1 Undirected Graphical Models

To make the manuscript self-contained, we first introduce some definitions and analytical results for Graphical Models. Suppose that $\mathbf{x} = (X_1, \ldots, X_p)^T$ is a p-variate random vector with joint distribution F. The conditional independence structure of the distribution can be represented by a graphical model $\mathcal{G} = (V, E)$, where $V = \{1, \ldots, p\}$ is the set of nodes and $E \subseteq V \times V$ the set of edges. The graph is called undirected when all the edges are undirected. If two nodes i and $j \in V$ form and edge, then i and j are adjacent or neighbors (Lauritzen, 1996; Edwards, 2000).

Definition 3. *Let $\mathcal{G} = (V, E)$ be an undirected graph. The set of neighbors of a node $i \in V$ is denoted as $\mathcal{A}_i$ and is defined by the following set of nodes:*

$$\mathcal{A}_i = \{k \in V \setminus \{i\} : (i, k) \in E\}. \tag{2.1.1}$$

Associated with and undirected graph $\mathcal{G}$ and the probability distribution F we can assume a range of different Markov properties. If $I \subseteq V$ let $X_I = \{X_i : i \in I\}$.

Definition 4. *(Pairwise Markov Property) We say that F satisfies the Pairwise Markov property (P) with respect to an undirected graph $\mathcal{G}$, if for any pair of unconnected nodes $(i, j) \notin E$*

$$X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}.$$

Definition 5. *(Local Markov Property) We say that F satisfies the Local Markov property (L) with respect to an undirected graph $\mathcal{G}$, if for any node $i \in V$*

$$X_i \perp\!\!\!\perp X_{V \setminus (\mathcal{A}_i \cup \{i\})} \mid X_{\mathcal{A}_i}.$$

A stronger notion is the Global Markov property. Let I, J, S three disjoint and nonempty subsets of V. We denote $X_I \perp\!\!\!\perp X_J \mid X_S$ when X_I is independent of X_J

conditioned on X_S; the set S is said to separate I from J if for every node $i \in I$ and $j \in J$, all paths from i to j have at least one node in S. We introduce the following definition.

Definition 6. *(Global Markov Property) The probability distribution F satisfies the Global Markov property (G) with respect to $\mathcal{G}$ if for every triplet of disjoint and nonempty subsets of nodes I, J, S it holds that $X_I \perp\!\!\!\perp X_J \mid X_S$ whenever S separates I and J in $\mathcal{G}$.*

The three Markov properties are related as follows.

Proposition 1. *For any undirected graph $\mathcal{G}$ and any probability distribution F, it holds that*

$$\text{(G)} \Longrightarrow \text{(L)} \Longrightarrow \text{(P)}.$$

The following theorem states that if F has a positive and continuous density function with respect to a product measure then the Markov properties are all equivalent (see Lauritzen, 1996).

Theorem 1. *(Pearl and Paz, 1985) If F has a strictly positive and continuous density function with respect to a product measure, then*

$$\text{(G)} \Longleftrightarrow \text{(L)} \Longleftrightarrow \text{(P)}.$$

A special class of undirected graphical models is the conditionally independence graph. This is a graphical model with undirected graph $\mathcal{G}$ and probability distribution F where the pairwise Markov property holds. Hence, if $(i, j) \notin E$ then $X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}$.

2.1.1 Gaussian Graphical Models

Suppose that $\mathbf{x} = (X_1, \ldots, X_p)^T$ has a joint Gaussian distribution with mean $\boldsymbol{\mu} = \mathbf{0}$ and $p \times p$ covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j=1\ldots,p}$ and let $\mathcal{G} = (V, E)$ be the associated graphical model . If $\boldsymbol{\Sigma}$ is regular then $\boldsymbol{\Omega} = (\omega_{ij})_{i,j=1\ldots,p}$ denote the precision matrix $\boldsymbol{\Sigma}^{-1}$.

Consider X_i and X_j with $i < j$ and let $\mathbf{x}_{i,j} = (X_i, X_j)^T$ and $\mathbf{x}_{-\{i,j\}}$ the random vector containing the rest of the variables (the indexes in $V \setminus \{i, j\}$ in ascending order) and let $\mathbf{x}_{0-\{i,j\}} \in \mathbb{R}^{(p-2)\times 1}$ be a (fixed) vector. By Proposition C.5. of Lauritzen (1996), the conditional distribution of $\mathbf{x}_{i,j}$ given $\mathbf{x}_{-\{i,j\}} = \mathbf{x}_{0-\{i,j\}}$ is normally distributed with covariance matrix $\boldsymbol{\Sigma}_{i,j|-\{i,j\}}$ given by the inverse of the matrix

$$A_{i,j} = \begin{pmatrix} \omega_{ii} & \omega_{ij} \\ \omega_{ji} & \omega_{jj} \end{pmatrix}.$$

Hence

$$\boldsymbol{\Sigma}_{i,j|-\{i,j\}} = A_{i,j}^{-1} = \frac{1}{\omega_{ii}\omega_{jj} - \omega_{ij}\omega_{ji}} \begin{pmatrix} \omega_{jj} & -\omega_{ij} \\ -\omega_{ji} & \omega_{ii} \end{pmatrix}. \tag{2.1.2}$$

By the normality, conditionally to $\mathbf{x}_{-\{i,j\}} = \mathbf{x}_{0-\{i,j\}}$, X_i and X_j are independent if and only if $\omega_{ij} = \omega_{ji} = 0$ (hereafter, for the sake of simplicity, we will omit the vector $\mathbf{x}_{0-\{i,j\}}$).

The partial correlation between variables X_i and X_j, ρ_{ij}, is defined as the correlation coefficient of the conditional distribution of $\mathbf{x}_{\{i,j\}}$ given $\mathbf{x}_{-\{i,j\}}$ and so, by (2.1.2),

$$\rho_{ij} = \frac{-\omega_{ij}}{[\omega_{ii}\omega_{jj}]^{1/2}} \tag{2.1.3}$$

and, as a consequence, the following proposition is established (see Proposition of 5.2 of Lauritzen, 1996).

Proposition 2. *(Conditional Independence) Assume that $\boldsymbol{x} = (X_1, \dots, X_p)^T$ is multivariate Gaussian distributed with regular covariance matrix $\boldsymbol{\Sigma}$. Then, it holds that $\forall i, j \in V$, with $i \neq j$:*

$$X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}} \Longleftrightarrow \omega_{ij} = 0 \tag{2.1.4}$$

or, equivalently,

$$(i, j) \notin E \Longleftrightarrow \omega_{ij} = 0. \tag{2.1.5}$$

Note that, since the joint Gaussian density function is positive and continuous, as a result of applying Theorem 1 the GGMs also satisfies Local and Global Markov properties.

We are interested to express partial correlation in terms of regression coefficients. Let $\mathbf{x}_{-i} \in \mathbb{R}^{(p-1)\times 1}$ the random vector with the set of indexes $V \setminus \{i\}$ in ascending order,

$$\begin{aligned} \boldsymbol{\Sigma}_{-i,-i} &= \operatorname{Cov}(\mathbf{x}_{-i}, \mathbf{x}_{-i}) \qquad (2.1.6) \\ \boldsymbol{\Sigma}_{i,-i} &= \operatorname{Cov}(X_i, \mathbf{x}_{-i}) \\ &= \operatorname{Cov}(X_i, \mathbf{x}_{-i}) = (\operatorname{Cov}(X_i, X_1), \dots, \operatorname{Cov}(X_i, X_{i-1})) \end{aligned}$$

and $\boldsymbol{\Sigma}_{-i,i} = \operatorname{Cov}(\mathbf{x}_{-i}, X_i)$.

Note that, by the same proposition Proposition C5 of Lauritzen (1996), the conditional distribution of $X_i \mid \mathbf{x}_{-i}$ satisfies

$$X_i \mid \mathbf{x}_{-i} \sim N\left(\mu_{i|-i}, \Sigma_{i|-i}\right), \tag{2.1.7}$$

where $\mu_{i|-i} = \boldsymbol{\Sigma}_{i,-i}\boldsymbol{\Sigma}_{-i,-i}^{-1}\mathbf{x}_{-i}$ and $\Sigma_{i|-i} = \boldsymbol{\Sigma}_{ii} - \boldsymbol{\Sigma}_{i,-i}\boldsymbol{\Sigma}_{-i,-i}^{-1}\boldsymbol{\Sigma}_{-i,i}$.

Then, for each node $i \in V$ the optimal predictor for X_i, $\boldsymbol{\beta}^i$, given the remaining variables is defined as (see Meinshausen and Bühlmann, 2006):

$$\begin{aligned} \boldsymbol{\beta}^i &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p:\beta_i=0}{\operatorname{argmin}} \parallel X_i - \sum_{j\in V}^{p} \beta_j X_j \parallel_2^2 \\ &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p:\beta_i=0}{\operatorname{argmin}} \mathrm{E}_{X_i|\mathbf{x}_{-i}} \left(X_i - \sum_{j\in V} \beta_j X_j \right)^2 . \end{aligned} \tag{2.1.8}$$

The following Lemma summarizes very well known properties related to multivariate normal distribution and its proof is given in Section 2.6.

Lemma 1.

a) For every node $i \in V$ the optimal predictor for X_i, $\boldsymbol{\beta}^i \in \mathbb{R}^p$, satisfies

$$\beta_j^i = -\frac{\omega_{ij}}{\omega_{ii}}, \forall j \neq i \tag{2.1.9}$$

and $\beta_i^i = 0$ (by definition).

b) Given $i, j \in V$, the partial correlation ρ_{ij} is related to β_j^i by the following expression

$$\rho_{ij} = sign(\beta_j^i)\sqrt{\beta_i^j \beta_j^i}. \tag{2.1.10}$$

c) For every $i \in V$, X_i satisfies the following regression model

$$X_i = \sum_{j\neq i} \beta_j^i X_j + \varepsilon_i = \boldsymbol{\beta}^i \boldsymbol{x}_{-i}^T + \varepsilon_i$$

where β_j^i is given by (2.1.9) and the error term $\varepsilon_i \sim N\left(0, \Sigma_{i|-i}\right)$ is independent of $\boldsymbol{x}_{-i}$.

As a simple consequence of the previous Lemma the errors vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)^T$ has a $\mathrm{N}\left(\mathbf{0}, \boldsymbol{\Psi}\right)$ distribution with $\mathrm{Var}(\varepsilon_i)$ and $\mathrm{Cov}(\varepsilon_i, \varepsilon_j)$ given by

$$\psi_{i,i} = \omega_{ii}^{-1} \text{ and } \psi_{i,j} = \omega_{ij}/(\omega_{ii}\omega_{jj}), \tag{2.1.11}$$

respectively.

As a generalization of equation (2.1.8), for a every node i and $\mathcal{A}_i \subseteq V \setminus \{i\}$ the optimal prediction of X_i given the set of variables $\{X_k : k \in \mathcal{A}_i\}$ is characterized by the vector $\boldsymbol{\beta}^{i,\mathcal{A}_i}$ (see Meinshausen and Bühlmann, 2006) by

$$\begin{aligned} \boldsymbol{\beta}^{i,\mathcal{A}_i} &= \underset{\beta \in \mathbb{R}^p : \beta_j = 0, \forall j \notin \mathcal{A}_i}{\operatorname{argmin}} \parallel \mathbf{X}_i - \sum_{j \in V} \beta_j \mathbf{X}_j \parallel_2^2 \\ &= \underset{\beta \in \mathbb{R}^p : \beta_j = 0, \forall j \notin \mathcal{A}_i}{\operatorname{argmin}} \mathrm{E}\left(X_i - \sum_{j \in V} \beta_j X_j \right)^2 . \end{aligned} \tag{2.1.12}$$

2.1.2 Inverse Covariance Estimation

Let $\mathbf{x} = (X_1, \ldots, X_p)^T \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$, where $\boldsymbol{\Sigma}$ is the covariance matrix, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ the precision matrix and let $\mathcal{G} = (V, E)$ be the associated graphical model as before. Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a random sample of the random vector $\mathbf{x}$ where $\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})^T$, $k = 1, \ldots, n$. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n \times p}$ be the data matrix.

In gaussian graphical models (GGMs) we are interested in recovering, based on $\mathbf{X}$, the underlying graph structure and this problem corresponds to determining which off-diagonal entries of $\boldsymbol{\Omega}$ are non-zero. Considering (2.1.5) and (2.1.9)

$$E = \{(i,j) : \omega_{ij} \neq 0\} = \left\{(i,j) : \beta_j^i \neq 0\right\}.$$

The neighborhood of a node $i \in V$ is defined as

$$\mathcal{A}_i = \{j \in V \setminus \{i\} : (i,j) \in V\}, \tag{2.1.13}$$

and, in consequence, the estimation of the inverse covariance matrix is equivalent to the estimation of the set of neighborhoods of the nodes $i \in V$.

As we mentioned in the introduction when $p > n$ the sample covariance matrix is badly conditioned and its inverse tends to amplify the estimation error. For large precision matrix estimation Fan et al. (2016) proposed the innovated scalable efficient estimation (ISEE) breaking the large-scale precision matrix estimation into smaller-scale linear regression problems. In the following we give a brief explanation of ISEE.

Consider the linear transformation, termed as innovation (in time series literature),

$$\widetilde{\mathbf{x}} = \mathbf{\Omega}\mathbf{x} \tag{2.1.14}$$

and note that the unobservable $p-$variate random vector $\widetilde{\mathbf{x}}$ has a $\mathrm{N}(\mathbf{0}, \mathbf{\Omega})$ distribution.

If A, B are subsets of V define (the sub–matrix of $\mathbf{\Omega}$) $\mathbf{\Omega}_{A,B} = (\omega_{ij})_{i \in A, j \in B}$ and for $A = B$ let $\mathbf{\Omega}_A = \mathbf{\Omega}_{A,A}$. Consider a partition $(A_l)_{l=1}^{L}$ of the set of nodes and, for every A_l, let $\mathbf{x}_{A_l}$ be the sub–vector of $\mathbf{x}$ formed by its components with indexes in A_l. So, by the definition of $\widetilde{\mathbf{x}}$ we can write the sub–vector $\widetilde{\mathbf{x}}_{A_l}$ as

$$\widetilde{\mathbf{x}}_{A_l} = \mathbf{\Omega}_{A_l}\boldsymbol{\eta}_{A_l}$$

with $\boldsymbol{\eta}_{A_l} = \mathbf{x}_{A_l} + \mathbf{\Omega}_{A_l}^{-1}\mathbf{\Omega}_{A_l,A_l^c}$. Note now that the result in (3.1.6) can be generalized to $\mathbf{x}_{A_l}|\mathbf{x}_{A_l^c} \sim \mathrm{N}\left(-\mathbf{\Omega}_{A_l}^{-1}\mathbf{\Omega}_{A_l,A_l^c}, \mathbf{\Omega}_{A_l}^{-1}\right)$ suggesting the following multivariate linear regression model

$$\mathbf{x}_{A_l} = \mathbf{C}_{A_l}\mathbf{x}_{A_l^c} + \boldsymbol{\eta}_{A_l} \tag{2.1.15}$$

where $\mathbf{C}_{A_l}$ is a matrix of regression coefficients and $\boldsymbol{\eta}_{A_l}$ is the vector of model errors. Equation (2.1.15) suggest to fit this model and to propose, using some regression technique, an estimate $\widehat{\boldsymbol{\eta}}_{A_l}$ considering $\boldsymbol{\eta}_{A_l}$ as a residual vector.

Hence, $\mathbf{\Omega}_{A_l}$ can be estimated by the inverse of the sample covariance matrix, $\widehat{\mathbf{\Omega}}_{A_l}$, of the model residual vector $\widehat{\boldsymbol{\eta}}_{A_l}$ and, finally, it is possible to estimate $\widetilde{\mathbf{x}}_{A_l}$ by

$\widehat{\mathbf{x}}_{A_l} = \widehat{\mathbf{\Omega}}_{A_l}\widehat{\boldsymbol{\eta}}_{A_l}$. By stacking all these sub-vectors $(\widetilde{\mathbf{x}}_{A_l})_{l=1}^{L}$ together it is possible to estimate the oracle innovated vector $\widetilde{\mathbf{x}}$.

We will briefly discuss below the implementation of the ISSE that will use in our forward-backward method in the next section. Given $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n\times p}$ the data matrix, the oracle empirical matrix $\widetilde{\mathbf{X}} \in \mathbb{R}^{n\times p}$ is defined as, (see Fan et al., 2016),

$$\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{\Omega}. \tag{2.1.16}$$

If A is a subset of the set of nodes V, using matrix notation, the model (2.1.15) can be written

$$\mathbf{X}_A = \mathbf{C}_A\mathbf{X}_{A^c} + \mathbf{E}_A \tag{2.1.17}$$

where $\mathbf{X}_A, \mathbf{X}_{A^c}$ are the sub–matrices of $\mathbf{X}$ with columns in A and A^c respectively, $\mathbf{E}_A$ is an $n \times |A|$ model error matrix with rows as i.i.d. copies of $\widehat{\boldsymbol{\eta}}_A^T$. The corresponding matrix $\widetilde{\mathbf{X}}_A$ can be written as

$$\widetilde{\mathbf{X}}_A = \mathbf{\Omega}_A\mathbf{E}_A.$$

Note now that for every node $i \in A$, it is possible to define the univariate linear regression model for response $\mathbf{X}_i$ ($i$the column of $\mathbf{X}$) given by the linear regression

$$\mathbf{X}_i = \mathbf{X}_{A^c}\boldsymbol{\beta}^i + \mathbf{E}_i \tag{2.1.18}$$

where $\mathbf{E}_i$ is the corresponding column of the matrix $\mathbf{E}_A$ and $\boldsymbol{\beta}^i \in \mathbb{R}^{(p-|A|)\times 1}$ is the column of the regression coefficients matrix $\mathbf{C}_A$.

As we mentioned before, using some technique regression we obtain an estimation, $\widehat{\boldsymbol{\beta}}^i$, of the $\boldsymbol{\beta}^i$ coefficients and so

$$\widehat{\mathbf{E}}_i = \mathbf{X}_i - \mathbf{X}_{A^c}\widehat{\boldsymbol{\beta}}^i \text{ and } \widehat{\mathbf{E}}_A = \left(\widehat{\mathbf{E}}_i\right)_{i\in A} \tag{2.1.19}$$

This equations give a natural estimate, $\widehat{\mathbf{\Omega}}_A$, of the sub–matrix $\mathbf{\Omega}_A$ defined by

$$\widehat{\mathbf{\Omega}}_A = \left(n^{-1}\widehat{\mathbf{E}}_A^T\widehat{\mathbf{E}}_A\right)^{-1} \tag{2.1.20}$$

and so, a plug-in estimator for the unobservable sub–matrix $\widetilde{\mathbf{X}}_A$ is $\widehat{\mathbf{E}}_A\widehat{\mathbf{\Omega}}_A$.

When A ranges the partition $(A_l)_{l=1}^L$ we obtain the estimate $\widehat{\mathbf{X}}$ of the the oracle empirical matrix $\widetilde{\mathbf{X}}$ and an initial estimate $\widehat{\mathbf{\Omega}}$ of the precision matrix $\mathbf{\Omega}$:

$$\widehat{\mathbf{X}} = \left(\widehat{\mathbf{X}}_{A_l}\right)_{1\leq l\leq L} \text{ and } \widehat{\mathbf{\Omega}} = \frac{1}{n}\widehat{\mathbf{X}}^T\widehat{\mathbf{X}} \tag{2.1.21}$$

To estimate the regression coefficients, ISEE introduced by Fan et al. (2016) use penalized least square with the scale Lasso. Sun and Zhang (2012) show that under mild regularities conditions, the residual error and the regression coefficients, estimated with the scaled lasso, are consistent and asymptotically normal. However, this may result in a dense network, thus the final sparse precision matrix estimator is computed by thresholding the elements of $n^{-1}\hat{\mathbf{X}}^T\hat{\mathbf{X}}$ (Bickel and Levina, 2008a).

To show that the lasso can consistently estimate the neighborhood of each node, we need to assume the irrepresentable condition and the beta-min condition on the size of the minimal absolute value of non-zero regression coefficients (see Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; van de Geer and Bühlmann, 2009; Bühlmann and Van De Geer, 2011). This assumption may be restricted, hence we proposed an alternative approach to estimate the oracle empirical matrix $\tilde{\mathbf{X}}$ without resorting to ℓ_1 regularization procedures.

In the next section we propose a non-regularized greedy forward-backward algorithm to recover the graph structure and this technique is based, partially, in the estimation of the precision matrix based on innovation although our method use least square regression.

For each node $i \in V$, the optimal predictor $\boldsymbol{\beta}^{i,\mathcal{A}_i} = \boldsymbol{\beta}^i$ based on the variables in the neighborhood set $\mathcal{A}_i$ was defined in (2.1.12). Let $\widehat{\boldsymbol{\beta}}^i$ the estimated coefficient vector and let $\widehat{\mathbf{E}}_i$ be the estimated residual error of the best predictor of X_i given the variables in the neighborhood set

$$\widehat{\mathbf{E}}_i = \mathbf{X}_i - \sum_{j=1}^{p} \widehat{\beta}_j^i \mathbf{X}_j. \tag{2.1.22}$$

By (2.1.20) if $\widehat{\boldsymbol{\Omega}}_i = n^{-1}\widehat{\mathbf{E}}_i^T\widehat{\mathbf{E}}_i$ and if $\widehat{\mathbf{X}} \in \mathbb{R}^{n\times p}$ denotes the estimator of the oracle empirical matrix $\widetilde{\mathbf{X}}$, the estimate of the i-th element of the oracle empirical matrix $\widetilde{\mathbf{X}}$ is

$$\widehat{\mathbf{X}}_i = \widehat{\mathbf{E}}_i\widehat{\boldsymbol{\Omega}}_i \quad \text{for } i \in V. \tag{2.1.23}$$

The estimator of the precision matrix Ω is the sample covariance matrix of the estimate oracle empirical matrix

$$\widehat{\boldsymbol{\Omega}} = n^{-1}\widehat{\mathbf{X}}^T\widehat{\mathbf{X}}. \tag{2.1.24}$$

So, for an undirected graph $\mathcal{G} = (V, E)$, if we can efficiently recover the support of $\boldsymbol{\Omega}$

$$\text{supp}(\boldsymbol{\Omega}) = \{\{i, j\} \in V : (i, j) \in E\}, \tag{2.1.25}$$

then we can estimate the precision matrix as the sample covariance matrix of a random vector that depends on the residuals of the best linear predictor for each node and, so, $\text{supp}(\boldsymbol{\Omega})$ can be estimated by $\text{supp}(\widehat{\boldsymbol{\Omega}})$.

2.2 The Proposed Method

In this section, we present a new forward-backward greedy algorithm to recover the structure of a GGM. This procedure estimates, sequentially, the oracle empirical matrix and as a consequence obtains an estimation of the precision matrix. We call this procedure *Graphical Stepwise*.

The algorithm begins with an empty set of edges (i.e. $\hat{E}^{(0)} = \emptyset$ where the superscripts denotes the number of steps). Moreover, the initial neighborhoods sets are given by $\hat{\mathcal{A}}_i^{(0)} = \emptyset$ for each $i \in V$. In the forward step, the initial optimal edge is given by the pair of variables with the largest absolute empirical correlation:

$$(i^*, j^*) = \underset{(i,j)\in(\hat{E}^{(0)})^c}{\operatorname{argmax}} \mid \widehat{\text{Cor}}(\mathbf{X}_i, \mathbf{X}_j) \mid . \tag{2.2.1}$$

The edge set and the neighborhood sets are updated as: $\hat{E}^{(1)} \leftarrow \{(i^*, j^*)\}$, $\hat{\mathcal{A}}_{i^*}^{(1)} \leftarrow \{j^*\}$ and $\hat{\mathcal{A}}_{j^*}^{(1)} \leftarrow \{i^*\}$. Next, the algorithm sets $k = k+1$ and estimates the prediction errors of X_{i^*} and X_{j^*} using the result in (2.1.12). If the corresponding absolute empirical correlation is smaller or equal to a threshold γ (i.e. $|\widehat{\text{Cor}}(\mathbf{X}_{i^*}, \mathbf{X}_{j^*})| \leq \gamma$), the algorithm stops and outputs the current estimate set of edges.

At step k, the current edge set is given by $\hat{E}^{(k-1)}$. In the forward step, the procedure finds the edge that maximizes the absolute value of the empirical partial correlation coefficient which is define as the empirical correlation coefficient between the estimated prediction errors $\widehat{\mathbf{E}}_i^{(k-1)}$ and $\widehat{\mathbf{E}}_j^{(k-1)}$ of the best linear predictors of X_i and X_j based on the set of variables in the neighborhood sets $\hat{\mathcal{A}}_i^{(k-1)}$ and $\hat{\mathcal{A}}_j^{(k-1)}$, respectively for $(i, j) \in (\hat{E}^{(k-1)})^c$:

$$\widehat{\rho}_{ij}^{(k-1)} = \widehat{\text{Cor}}(\widehat{\mathbf{E}}_i^{(k-1)}, \widehat{\mathbf{E}}_j^{(k-1)}). \tag{2.2.2}$$

Then the best next candidate is given by the pair of nodes with the largest absolute empirical partial correlation coefficient:

$$(i^*, j^*) = \underset{(i,j)\in(\hat{E}^{(k-1)})^c}{\operatorname{argmax}} \mid \widehat{\text{Cor}}(\hat{\mathbf{E}}_i^{(k-1)}, \hat{\mathbf{E}}_j^{(k-1)}) \mid . \tag{2.2.3}$$

Next, the edge set and the neighborhood sets are updated as $\hat{E}^{(k)} \leftarrow \hat{E}^{(k-1)} \cup \{(i^*, j^*)\}$, $\hat{\mathcal{A}}_{i^*}^{(k)} \leftarrow \hat{\mathcal{A}}_{i^*}^{(k-1)} \cup \{j^*\}$ and $\hat{\mathcal{A}}_{j^*}^{(k)} \leftarrow \hat{\mathcal{A}}_{j^*}^{(k-1)} \cup \{i^*\}$. If the corresponding absolute empirical partial correlation is smaller than γ, the algorithm stops and outputs the current estimate set of edges. Otherwise, the algorithm sets $k = k + 1$ and estimates the prediction errors of X_{i^*} and X_{j^*} using the result in (2.1.12) assuming that the neighborhood sets are given by $\hat{\mathcal{A}}_{i^*}^{(k)}$ and $\hat{\mathcal{A}}_{j^*}^{(k)}$. In the backward step, the algorithm eliminates the unlikely edges in $E^{(k)}$. Thus, the procedure selects the pair of nodes from the edge set with the minimum absolute empirical partial correlation coefficient:

$$(i_*, j_*) = \underset{(i,j) \in (\hat{E}^{(k)})}{\operatorname{argmin}} \mid \widehat{\operatorname{Cor}}(\hat{\mathbf{E}}_i^{(k)}, \hat{\mathbf{E}}_j^{(k)}) \mid, \tag{2.2.4}$$

if the corresponding absolute empirical partial correlation corresponding to nodes (i_*, j_*) is smaller than γ, the algorithm removes the pair of variables from the edge set: $\hat{E}^{(k-1)} \leftarrow \hat{E}^{(k)} - \{(i_*, j_*)\}$, $\hat{\mathcal{A}}_{i_*}^{(k-1)} \leftarrow \hat{\mathcal{A}}_{i_*}^{(k)} - \{j_*\}$ and $\hat{\mathcal{A}}_{j_*}^{(k-1)} \leftarrow \hat{\mathcal{A}}_{j_*}^{(k)} - \{i_*\}$. Algorithm 1 summarizes the Graphical Stepwise procedure to learn the structure of a GGM.

The Graphical Stepwise procedure outputs the estimated set of edges, denoted by $\hat{E}$ and the corresponding prediction errors $\hat{\mathbf{E}}_i$ for $i \in V$. To estimate the precision matrix we convert the original problem of estimating $\boldsymbol{\Omega}$ into that of a covariance estimation problem. Let $\hat{\mathbf{X}} \in \mathbb{R}^{n \times p}$ be the estimator of the oracle empirical matrix $\tilde{\mathbf{X}}$ define in (2.1.16). The diagonal elements of $\hat{\boldsymbol{\Omega}}$ are computed as $\widehat{\omega}_{ii} = n^{-1} \widehat{\mathbf{E}}_i^T \widehat{\mathbf{E}}_i$, and the estimate of the i-th element of the oracle empirical matrix $\tilde{\mathbf{X}}$ is

$$\widehat{\mathbf{X}}_i = \widehat{\mathbf{E}}_i \widehat{\boldsymbol{\Omega}}_i \quad \text{for } i \in V. \tag{2.2.5}$$

The estimator of the off-diagonal elements of the precision matrix $\boldsymbol{\Omega}$ are given by the covariance of the oracle empirical matrix estimate for all pairs of nodes that belong to the estimated set of edges $\hat{E}$:

$$\begin{aligned} \hat{\omega}_{ij} &= n^{-1}\hat{\mathbf{X}}_i^T\hat{\mathbf{X}}_j \qquad (i,j) \in \hat{E} \\ \hat{\omega}_{ij} &= 0 \qquad\qquad\qquad (i,j) \notin \hat{E} \end{aligned} \tag{2.2.6}$$

The Graphical Stepwise procedure is able to estimate an undirected graph that contains the set of edges with the largest absolute empirical partial correlations coefficients for a given threshold. Also, it is possible to introduce previous knowledge when we construct the network. For instance, if we have information that some variables are conditionally independent we can exclude them from the candidate set of edges $\hat{E}$. Moreover, we can reduce the candidate set of edges by performing a correlation screening proposed by Fan and Lv (2008). We propose to obtain a reduce candidate set of edges by thresholding the empirical correlation matrix by a constant $\xi > 0$. Then, the set of candidate edges is defined as follows

$$\hat{\mathcal{E}}_\xi = \{(i,j) : |\widehat{\mathrm{Cor}}(\mathbf{X}_i, \mathbf{X}_j)| > \xi,\ i,j = 1,\ldots,p\}. \tag{2.2.7}$$

Luo et al. (2014) and Liang et al. (2015) show that for an appropriate choice of ξ, the true set of edges is contained in the candidate set of edges $\hat{\mathcal{E}}_\xi$ with high probability, when p grows exponentially with n.

Algorithm 1: Graphical Stepwise algorithm for Gaussian covariance estimation

```
input  : $\mathbf{X}$, Stopping Threshold $\gamma > 0$
output: Edge Set Estimation $\hat{E}$, Residual Error Estimation $\hat{\mathbf{E}}_i$ for $i \in V$

Initialize $\hat{E}^{(0)} \leftarrow \emptyset$, $\hat{\mathcal{A}}_i^{(0)} \leftarrow \emptyset$, $\hat{\boldsymbol{E}}_i^{(0)} \leftarrow \boldsymbol{X}_i$ for $i \in V$, $k \leftarrow 1$;
while true do
    Forward Step;
    $(i^*, j^*) \leftarrow \text{argmax}_{(i,j)\in(\hat{E}^{(k-1)})^C} \mid \widehat{\text{Cor}}(\hat{\mathbf{E}}_i^{(k-1)}, \hat{\mathbf{E}}_j^{(k-1)}) \mid$;
    $\hat{E}^{(k)} \leftarrow \hat{E}^{(k-1)} \cup \{(i^*, j^*)\}$;
    $\hat{\mathcal{A}}_{i^*}^{(k)} \leftarrow \hat{\mathcal{A}}_{i^*}^{(k-1)} \cup \{j^*\}$;
    $\hat{\mathcal{A}}_{j^*}^{(k)} \leftarrow \hat{\mathcal{A}}_{j^*}^{(k-1)} \cup \{i^*\}$;
    if $\mid \widehat{Cor}(\hat{\boldsymbol{E}}_{i^*}^{(k-1)}, \hat{\boldsymbol{E}}_{j^*}^{(k-1)}) \mid \leq \gamma$ then
        break
    end
    $\hat{\beta}^{i^*} = \text{argmin}_{\beta\in\mathbb{R}^p:\beta_l=0, l\in V\setminus\hat{\mathcal{A}}_{i^*}^{(k)}} \parallel \mathbf{X}_{i^*} - \sum_{l=1}^{p} \beta_l \mathbf{X}_l \parallel_2^2$ ;
    $\hat{\mathbf{E}}_{i^*}^{(k)} = \mathbf{X}_{i^*} - \sum_{l=1}^{p} \hat{\beta}_l^{i^*} \mathbf{X}_l$;
    $\hat{\beta}^{j^*} = \text{argmin}_{\beta\in\mathbb{R}^p:\beta_l=0, l\in V\setminus\hat{\mathcal{A}}_{j^*}^{(k)}} \parallel \mathbf{X}_{j^*} - \sum_{l=1}^{p} \beta_l \mathbf{X}_l \parallel_2^2$ ;
    $\hat{\mathbf{E}}_{j^*}^{(k)} = \mathbf{X}_{j^*} - \sum_{l=1}^{p} \hat{\beta}_l^{j^*} \mathbf{X}_l$;
    $k \leftarrow k + 1$;
    while true do
        Backward Step;
        $(i_*, j_*) \leftarrow \text{argmin}_{(i,j)\in\hat{E}^{(k-1)}} \mid \widehat{\text{Cor}}(\hat{\mathbf{E}}_i^{(k-1)}, \hat{\mathbf{E}}_j^{(k-1)}) \mid$ ;
        if $\mid \widehat{Cor}(\hat{\boldsymbol{E}}_{i_*}^{(k-1)}, \hat{\boldsymbol{E}}_{j_*}^{(k-1)}) \mid \leq \gamma$ then
            break
        end
        $\hat{E}^{(k-1)} \leftarrow \hat{E}^{(k)} - \{(i_*, j_*)\}$;
        $\hat{\mathcal{A}}_{i_*}^{(k-1)} \leftarrow \hat{\mathcal{A}}_{i_*}^{(k)} - \{j_*\}$;
        $\hat{\mathcal{A}}_{j_*}^{(k-1)} \leftarrow \hat{\mathcal{A}}_{j_*}^{(k)} - \{i_*\}$;
        $\hat{\beta}^{i_*} = \text{argmin}_{\beta\in\mathbb{R}^p:\beta_l=0, l\in V\setminus\hat{\mathcal{A}}_{i_*}^{(k-1)}} \parallel \mathbf{X}_{i_*} - \sum_{l=1}^{p} \beta_l \mathbf{X}_l \parallel_2^2$ ;
        $\hat{\mathbf{E}}_{i_*}^{(k-1)} = \mathbf{X}_{i_*} - \sum_{l=1}^{p} \hat{\beta}_l^{i_*} \mathbf{X}_l$;
        $\hat{\beta}^{j_*} = \text{argmin}_{\beta\in\mathbb{R}^p:\beta_l=0, l\in V\setminus\hat{\mathcal{A}}_{j_*}^{(k-1)}} \parallel \mathbf{X}_{j_*} - \sum_{l=1}^{p} \beta_l \mathbf{X}_l \parallel_2^2$ ;
        $\hat{\mathbf{E}}_{j_*}^{(k-1)} = \mathbf{X}_{j_*} - \sum_{l=1}^{p} \hat{\beta}_l^{j_*} \mathbf{X}_l$;
        $k \leftarrow k - 1$;
    end
end
```

2.2.1 The Proposed Method: Relation with the Information Divergence Measure

The Kullback-Leibler information divergence between two densities f and g for the random vector $\mathbf{x}$ is defined as

$$I(f;g) = \mathrm{E}_f\left[\log\left(f(\mathbf{x})/g(\mathbf{x})\right)\right]$$

where E_f denotes expectation with respect to the density f (see Whittaker, 2009, pp. 87–104).

Let $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ be random vectors with joint density function $f_{\mathbf{x},\mathbf{y},\mathbf{z}}$ and let $f_{\mathbf{y}|\mathbf{x}}$ and $f_{\mathbf{z}|\mathbf{x}}$ the densities of the conditional distribution of $\mathbf{y} \mid \mathbf{x}$ and $\mathbf{z} \mid \mathbf{x}$, respectively; let $f_{\mathbf{x}}$ be the density function of the vector $\mathbf{x}$. The information (Inf) in $\mathbf{y}$ about $\mathbf{z}$ conditional on $\mathbf{x}$ is defined by

$$\mathrm{Inf}\left(\mathbf{y} \perp\!\!\!\perp \mathbf{z} \mid \mathbf{x}\right) = I\left(f_{\mathbf{x},\mathbf{y},\mathbf{z}}; f_{\mathbf{z}|\mathbf{x}} f_{\mathbf{y}|\mathbf{x}} f_{\mathbf{x}}\right)$$

and it represents a measure of the average amount of information in $f_{\mathbf{x},\mathbf{y},\mathbf{z}}$ against the independence of $\mathbf{y}$ and $\mathbf{z}$ conditional on $\mathbf{x}$ because $\mathrm{Inf}\left(\mathbf{y} \perp\!\!\!\perp \mathbf{z} \mid \mathbf{x}\right) = 0$ if and only if $\mathbf{y} \perp\!\!\!\perp \mathbf{z} \mid \mathbf{x}$.

Let $\mathbf{x} = (X_1, \dots, X_p)^T$ be a random vector with multivariate Gaussian distribution and let $\mathcal{G} = (V, E)$ be its graphical model. For a fixed pair of nodes i, j consider $\mathbf{x}_{i,j} = (X_i, X_j)^T$ and $\mathbf{x}_{-\{i,j\}}$ (the random vector containing the rest of the variables) defined in Subsection 2.1.1.

The following proposition claims that information divergence for measuring the conditional independence of a pair of random variables X_i and X_j, given the remaining variables, has a simple expression as a function of the partial correlation (of the two variables). Although its proof is a consequence of Proposition 6.4.6 of Whittaker (2009) we give a straightforward proof for this particular case.

Proposition 3. *For every $i, j \in V$ it holds*

$$\mathit{Inf}\big(X_i \perp\!\!\!\perp X_j \mid \boldsymbol{x}_{-\{i,j\}}\big) = -\frac{1}{2}\mathit{log}(1 - \rho_{ij}^2), \tag{2.2.8}$$

where ρ_{ij} is the partial correlation coefficient between variables X_i and X_j conditioned on $\boldsymbol{x}_{-\{i,j\}}$.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a random sample $\mathbf{x}$ and let $\widehat{\mathrm{Cor}}(\hat{\mathbf{E}}_i^{(k-1)}, \hat{\mathbf{E}}_j^{(k-1)})$ be the empirical partial correlation coefficient defined as the empirical correlation coefficient between the estimated prediction errors $\widehat{\mathbf{E}}_i^{(k-1)}$ and $\widehat{\mathbf{E}}_j^{(k-1)}$ of the best linear predictors of X_i and X_j based on the set of variables in the neighborhood set $\hat{\mathcal{A}}_i^{(k-1)}$ and $\hat{\mathcal{A}}_j^{(k-1)}$. Lemma 1 relates the correlation between the regression errors and partial coefficient regression and Proposition 3 establishes the relation with the information divergence.

Hence, by the previous considerations, at step k the Graphical Stepwise algorithm select the pair of nodes (i^*, j^*) that satisfies

$$(i^*, j^*) = \underset{(i,j)\in(\hat{E}^{(k-1)})^C}{\mathrm{argmax}} -\frac{1}{2}\log(1 - \widehat{\mathrm{Cor}}(\hat{\mathbf{E}}_i^{(k-1)}, \hat{\mathbf{E}}_j^{(k-1)})^2). \tag{2.2.9}$$

So, at step k the algorithm select the pair (i^*, j^*) with largest absolute empirical partial correlation coefficient or largest empirical divergence information.

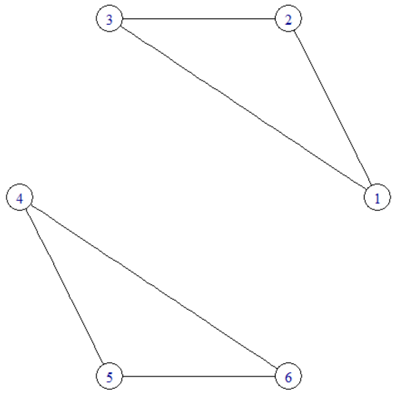


Figure 2.1: Undirected block graph (when $p = 6$). Nodes are represented by circles and undirected edges are represented by lines.

2.2.2 The Proposed Method: Example

The technique proposed in Section 2.2 will now be illustrated numerically on the 6×6 precision matrix $\mathbf{\Omega}$:

$$\mathbf{\Omega} = \begin{bmatrix} 1.0 & 0.5 & 0.5 & 0.0 & 0.0 & 0.0 \\ 0.5 & 1.0 & 0.5 & 0.0 & 0.0 & 0.0 \\ 0.5 & 0.5 & 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 & 0.5 & 0.5 \\ 0.0 & 0.0 & 0.0 & 0.5 & 1.0 & 0.5 \\ 0.0 & 0.0 & 0.0 & 0.5 & 0.5 & 1.0 \end{bmatrix} \tag{2.2.10}$$

This is a block graph $\mathcal{G} = (V, E)$ where the set of nodes are given by $V = \{1, 2, 3, 4, 5\}$ and the set of edges is given by $E = \{(1, 2), (1, 3), (2, 3), (4, 5), (4, 6), (5, 6)\}$. The graphical representation is shown in Figure 2.1.

We now show how the Graphical Stepwise procedure works. In doing so, we draw $n = 100$ independent samples from a multivariate Gaussian distribution with mean zero and covariance matrix $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$. To select the threshold γ we follow the scheme proposed by Cai et al. (2011). We generate a training sample of size $n = 100$ from a multivariate Gaussian distribution with mean zero and covariance matrix $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$, and an independent sample of the same distribution for validating the thresholding parameter. From the training data, we apply Algorithm 1 for 50 different values of γ. The optimal parameter is given by the one that minimize the log-likelihood loss defined by:

$$-\log(\det(\hat{\mathbf{\Omega}})) + \operatorname{tr}(\hat{\mathbf{\Omega}}\mathbf{X}^T\mathbf{X}) - p. \tag{2.2.11}$$

We replicate each simulation experiment 100 times.

First, we show how the algorithm works in each step for an specific replication where the thresholding parameter is set to $\gamma = 0.165$. The algorithm is initialized at step $k = 0$ assuming that all variables are conditionally independent. The initial set of edges is $\hat{E}^{(0)} = \emptyset$ and the the initial neighborhoods sets are given by $\hat{\mathcal{A}}_i^{(0)} = \emptyset$ for $i \in V$. The initial residual errors are given by $\hat{\mathbf{E}}_i^{(0)} = \mathbf{X}_i$ for $i \in V$. The empirical correlation matrix of the estimated residual errors is computed as the empirical correlation matrix of $\mathbf{X}$:

$$\widehat{\operatorname{Cor}}(\hat{\mathbf{E}}^{(0)}, \hat{\mathbf{E}}^{(0)}) = \begin{bmatrix} 1.00 & -0.28 & -0.37 & 0.03 & -0.01 & -0.06 \\ -0.28 & 1.00 & -0.38 & -0.11 & 0.08 & 0.10 \\ -0.37 & -0.38 & 1.00 & -0.06 & 0.05 & -0.00 \\ 0.03 & -0.11 & -0.06 & 1.00 & -0.33 & -0.30 \\ -0.01 & 0.08 & 0.05 & -0.33 & 1.00 & -0.31 \\ -0.06 & 0.10 & -0.00 & -0.30 & -0.31 & 1.00 \end{bmatrix} \tag{2.2.12}$$

In the forward step, the initial optimal edge is given by the pair of variables with the largest absolute empirical correlation. From the result in (2.2.12), we observe that the largest absolute partial correlation is given by the ordered pair $(i^*, j^*) = (2, 3)$ with $|\widehat{\operatorname{Cor}}(\hat{\mathbf{E}}_2^{(0)}, \hat{\mathbf{E}}_3^{(0)})| = 0.38$. Given that the corresponding absolute empirical correlation is not smaller or equal to $\gamma = 0.165$ the algorithm sets $k = 1$. Then the edge set and the neighborhood sets are updated as: $\hat{E}^{(1)} = \{(2,3)\}$, $\hat{\mathcal{A}}_2^{(1)} = \{3\}$ and $\hat{\mathcal{A}}_3^{(1)} = \{2\}$. The prediction errors of X_2 and X_3 are estimated using the result in (2.1.12). The empirical correlation matrix of the estimated residual errors at step $k = 1$ is given by:

$$\widehat{\text{Cor}}(\hat{\mathbf{E}}^{(1)}, \hat{\mathbf{E}}^{(1)}) = \begin{bmatrix} 1.00 & -0.45 & -0.51 & 0.03 & -0.01 & -0.06 \\ -0.45 & 1.00 & 0.38 & -0.15 & 0.11 & 0.11 \\ -0.51 & 0.38 & 1.00 & -0.11 & 0.09 & 0.04 \\ 0.03 & -0.15 & -0.11 & 1.00 & -0.33 & -0.30 \\ -0.01 & 0.11 & 0.09 & -0.33 & 1.00 & -0.31 \\ -0.06 & 0.11 & 0.04 & -0.30 & -0.31 & 1.00 \end{bmatrix} \tag{2.2.13}$$

In the backward step, the algorithm eliminates the unlikely edges in $E^{(1)}$. Thus, the procedure selects the pair of edges from the edge set with the minimum absolute partial correlation. Since the absolute empirical partial correlation of the edge $(2,3)$ is not smaller than γ, we do not eliminate the edge.

Next, we add an additional edge. In the forward step we select the ordered pair of nodes with the maximum absolute correlation between the empirical residuals in (2.2.13). Which is given by $(i^*, j^*) = (1,3)$. Given that $|\widehat{\text{Cor}}(\hat{\mathbf{E}}_1^{(1)}, \hat{\mathbf{E}}_3^{(1)})| = 0.51$ is larger than γ, The algorithm keeps running. Then, we set $k = 2$ and the edge set and the neighborhood sets are updated as: $\hat{E}^{(2)} = \{(1,3),(2,3)\}$, $\hat{\mathcal{A}}_1^{(2)} = \{3\}$ and $\hat{\mathcal{A}}_3^{(2)} = \{1,2\}$. The prediction errors of X_1 and X_3 are estimated using the result in (2.1.12). The empirical correlation matrix of the estimated residual errors at step

$k = 2$ is given by:

$$\widehat{\mathrm{Cor}}(\hat{\mathbf{E}}^{(2)}, \hat{\mathbf{E}}^{(2)}) = \begin{bmatrix} 1.00 & -0.48 & 0.31 & 0.01 & 0.00 & -0.06 \\ -0.48 & 1.00 & 0.32 & -0.15 & 0.11 & 0.11 \\ 0.31 & 0.32 & 1.00 & -0.13 & 0.11 & 0.03 \\ 0.01 & -0.15 & -0.13 & 1.00 & -0.33 & -0.30 \\ 0.00 & 0.11 & 0.11 & -0.33 & 1.00 & -0.31 \\ -0.06 & 0.11 & 0.03 & -0.30 & -0.31 & 1.00 \end{bmatrix} \tag{2.2.14}$$

In the backward step, we choose from the edges in $\hat{E}^{(2)}$ the edge with the minimum absolute empirical partial correlation (i.e. $\widehat{\mathrm{Cor}}(\hat{\mathbf{E}}_1^{(2)}, \hat{\mathbf{E}}_3^{(2)}) = 0.31$), since the coefficient is not smaller than $\gamma = 0.165$, we do not eliminate the ordered pair from the edge set.

In the following run of the algorithm the maximum correlation of the residuals errors in (2.2.14) is $|\widehat{\mathrm{Cor}}(\hat{\mathbf{E}}_1^{(2)}, \hat{\mathbf{E}}_2^{(2)})| = 0.48$. The algorithm does not stop and we set $k = 3$. The edge set and the neighborhood sets are updated as: $\hat{E}^{(3)} = \{(1,3),(1,2),(2,3)\}$, $\hat{\mathcal{A}}_1^{(3)} = \{2,3\}$ and $\hat{\mathcal{A}}_2^{(3)} = \{1,3\}$. We estimate the prediction errors $\hat{\mathbf{E}}_1^{(3)}$ and $\hat{\mathbf{E}}_2^{(3)}$ and we compute the correlation matrix of the empirical residual errors as:

$$\widehat{\mathrm{Cor}}(\hat{\mathbf{E}}^{(3)}, \hat{\mathbf{E}}^{(3)}) = \begin{bmatrix} 1.00 & 0.48 & 0.53 & -0.07 & 0.06 & -0.01 \\ 0.48 & 1.00 & 0.54 & -0.16 & 0.13 & 0.09 \\ 0.53 & 0.54 & 1.00 & -0.13 & 0.11 & 0.03 \\ -0.07 & -0.16 & -0.13 & 1.00 & -0.33 & -0.30 \\ 0.06 & 0.13 & 0.11 & -0.33 & 1.00 & -0.31 \\ -0.01 & 0.09 & 0.03 & -0.30 & -0.31 & 1.00 \end{bmatrix} \tag{2.2.15}$$

In the backward step we do not eliminate any ordered pair from the edge set since all estimated partial correlations in absolute value are larger than γ.

In the following step, the additional edge is given by $(i^*, j^*) = (4,5)$ and $|\widehat{\mathrm{Cor}}(\hat{\mathbf{E}}_4^{(3)}, \hat{\mathbf{E}}_5^{(3)})|$ is larger than γ, the algorithm sets $k = 4$: $\hat{E}^{(4)} = \{(1,3),(1,2),(2,3),(4,5)\}$ and $\hat{\mathcal{A}}_4^{(4)} = \{5\}$, $\hat{\mathcal{A}}_5^{(4)} = \{4\}$. The prediction errors of X_4 and X_4 are estimated using the result in (2.1.12) and the correlation matrix of the empirical residual errors is:

$$\widehat{\mathrm{Cor}}(\hat{\mathbf{E}}^{(4)}, \hat{\mathbf{E}}^{(4)}) = \begin{bmatrix} 1.00 & 0.48 & 0.53 & -0.05 & 0.05 & -0.01 \\ 0.48 & 1.00 & 0.54 & -0.13 & 0.08 & 0.09 \\ 0.53 & 0.54 & 1.00 & -0.10 & 0.07 & 0.03 \\ -0.05 & -0.13 & -0.10 & 1.00 & 0.33 & -0.43 \\ 0.05 & 0.08 & 0.07 & 0.33 & 1.00 & -0.43 \\ -0.01 & 0.09 & 0.03 & -0.43 & -0.43 & 1.00 \end{bmatrix} \qquad (2.2.16)$$

The backward step does not eliminate any edge.

At the beginning of step $k = 6$, the additional edge is $(i^*, j^*) = (5,6)$ and the estimated correlation between the residuals in absolute value is larger than γ (i.e. $|\widehat{\mathrm{Cor}}(\hat{\mathbf{E}}_5^{(4)}, \hat{\mathbf{E}}_6^{(4)})| = 0.43$). We set $k = 5$, $\hat{E}^{(5)} = \{(1,3),(1,2),(2,3),(4,5),(5,6)\}$ and $\hat{\mathcal{A}}_5^{(5)} = \{4,6\}$, $\hat{\mathcal{A}}_6^{(5)} = \{5\}$. We compute the residuals errors of X_5 and X_6 and we update the empirical correlation matrix of the residual errors:

$$\widehat{\mathrm{Cor}}(\hat{\mathbf{E}}^{(5)}, \hat{\mathbf{E}}^{(5)}) = \begin{bmatrix} 1.00 & 0.48 & 0.53 & -0.05 & 0.03 & 0.01 \\ 0.48 & 1.00 & 0.54 & -0.13 & 0.11 & 0.14 \\ 0.53 & 0.54 & 1.00 & -0.10 & 0.07 & 0.06 \\ -0.05 & -0.13 & -0.10 & 1.00 & 0.29 & -0.45 \\ 0.03 & 0.11 & 0.07 & 0.29 & 1.00 & 0.27 \\ 0.01 & 0.14 & 0.06 & -0.45 & 0.27 & 1.00 \end{bmatrix} \qquad (2.2.17)$$

In the following round, the additional edge is given by $(i^*, j^*) = (4,6)$ and the corresponding estimated correlation between the empirical errors is larger than the threshold value. We set $k = 6$, $\hat{E}^{(6)} = \{(1,3),(1,2),(2,3),(4,5),(4,6),(5,6)\}$ and

$\hat{\mathcal{A}}_4^{(6)} = \{5, 6\}$, $\hat{\mathcal{A}}_6^{(6)} = \{4, 5\}$. We compute the residuals errors of X_4 and X_6 and the empirical correlation matrix of the residual errors:

$$\widehat{\text{Cor}}(\hat{\mathbf{E}}^{(6)}, \hat{\mathbf{E}}^{(6)}) = \begin{bmatrix} 1.00 & 0.48 & 0.53 & -0.05 & 0.03 & -0.01 \\ 0.48 & 1.00 & 0.54 & -0.07 & 0.11 & 0.09 \\ 0.53 & 0.54 & 1.00 & -0.08 & 0.07 & 0.02 \\ -0.05 & -0.07 & -0.08 & 1.00 & 0.46 & 0.45 \\ 0.03 & 0.11 & 0.07 & 0.46 & 1.00 & 0.45 \\ -0.01 & 0.09 & 0.02 & 0.45 & 0.45 & 1.00 \end{bmatrix} \tag{2.2.18}$$

The backward step does not eliminate any of the edges since the correlation between the residual errors in absolute value are larger than γ.

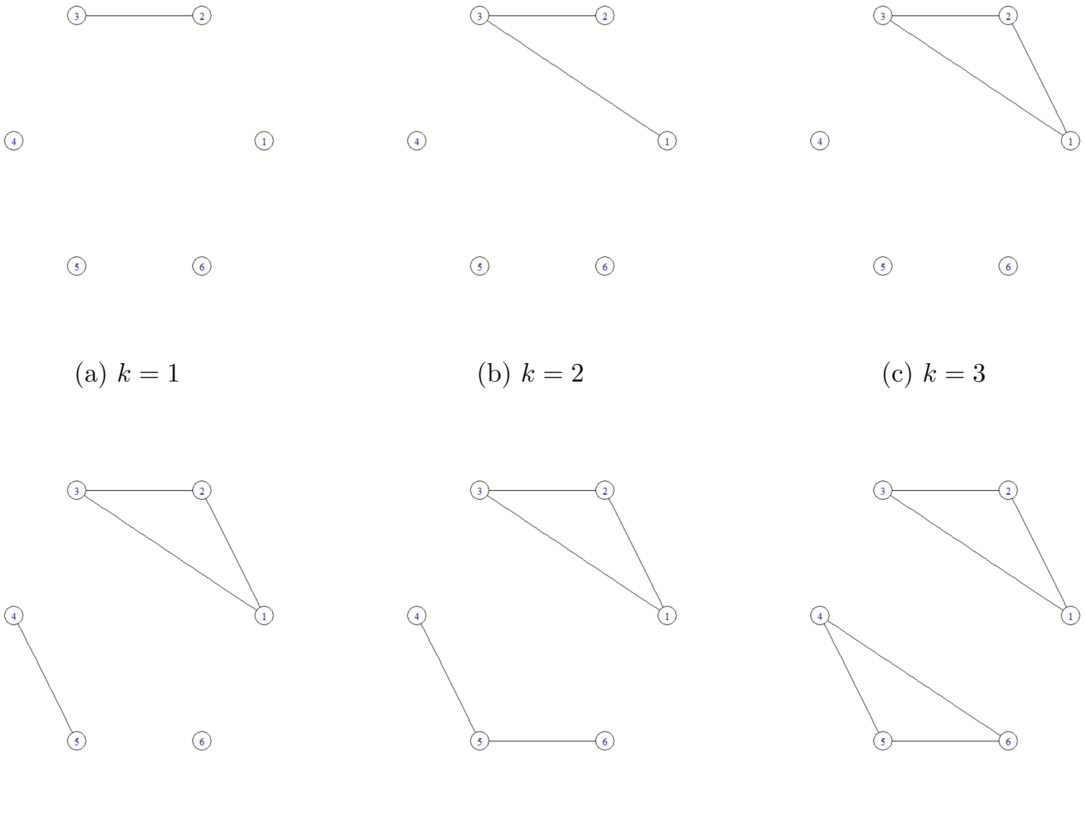
In the following run, the additional edge is $(i^*, j^*) = (2, 5)$. Since $|\widehat{\text{Cor}}(\hat{\mathbf{E}}_2^{(6)}, \hat{\mathbf{E}}_5^{(6)})| = 0.11$ is smaller than $\gamma = 0.165$, the algorithm stops and output the estimated set of edges $\hat{E} = \{(1, 3), (1, 2), (2, 3), (4, 5), (4, 6), (5, 6)\}$ and the corresponding prediction errors $\hat{\mathbf{E}}_i$ for $i \in V$. We estimate the oracle empirical matrix $\tilde{\mathbf{X}}$ and the precision matrix using the result in (3.2.5):

$$\hat{\boldsymbol{\Omega}} = \begin{bmatrix} 0.90 & 0.49 & 0.55 & 0.00 & 0.00 & 0.00 \\ 0.49 & 1.13 & 0.62 & 0.00 & 0.00 & 0.00 \\ 0.55 & 0.62 & 1.20 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.96 & 0.40 & 0.44 \\ 0.00 & 0.00 & 0.00 & 0.40 & 0.78 & 0.40 \\ 0.00 & 0.00 & 0.00 & 0.44 & 0.40 & 1.00 \end{bmatrix} \tag{2.2.19}$$

Table 2.1 and Figure 2.2 summarize the results for each step of the Graphical Stepwise procedure. Finally, Figure 2.3 shows the corresponding frequency of the zeros identified for each entry of $\boldsymbol{\Omega}$ out of 100 replications. We observe that the Graphical Stepwise procedure is able to recover the support of $\boldsymbol{\Omega}$ in an efficient way.

Table 2.1: Graphical Stepwise algorithm. Selected edges in forward and backward step.

	Forward Step		Backward Step		
Step	(i^*, j^*)	$\widehat{\text{Cor}}\hat{\mathbf{E}}_{i^*}^{(k-1)}, \hat{\mathbf{E}}_{j^*}^{(k-1)})$	(i_*, j_*)	$\widehat{\text{Cor}}(\hat{\mathbf{E}}_{i_*}^{(k)}, \hat{\mathbf{E}}_{j_*}^{(k)})$	$\hat{E}^{(k)}$
$k = 1$	$(2,3)$	-0.38	$(2,3)$	0.38	$\{(2,3)\}$
$k = 2$	$(1,3)$	-0.51	$(1,3)$	0.31	$\{(1,3);(2,3)\}$
$k = 3$	$(1,2)$	-0.48	$(1,2)$	0.48	$\{(1,2);(1,3);(2,3)\}$
$k = 4$	$(4,5)$	-0.33	$(4,5)$	0.33	$\{(1,2);(1,3);(2,3);(4,5)\}$
$k = 5$	$(5,6)$	-0.43	$(5,6)$	0.27	$\{(1,2);(1,3);(2,3);(4,5);(5,6)\}$
$k = 6$	$(4,6)$	-0.45	$(5,6)$	0.45	$\{(1,2);(1,3);(2,3);(4,5);(4,6);(5,6)\}$
$k = 7$	$(2,5)$	0.11	-	-	$\{(1,2);(1,3);(2,3);(4,5);(4,6);(5,6)\}$



(a) $k = 1$

(b) $k = 2$

(c) $k = 3$

(d) $k = 4$

(e) $k = 5$

(f) $k = 6$

Figure 2.2: Graph representation of the estimated undirected graphs in each step of the Graphical Stepwise Algorithm.

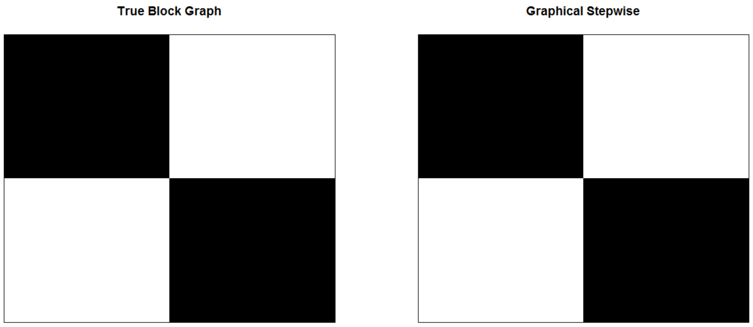


Figure 2.3: Heatmaps of the frequency of the zeros identified for each entry of Ω (when $p = 6$) out of 100 replications. White represents 100 zeros identified out of 100 runs, and black represents 0/100.

2.3 Analytical Properties

We conjecture some analytical properties for the Graphical Stepwise procedure to estimate the set of edges of a GGM. The analysis is related with the theoretical properties of the PC algorithm (Spirtes et al., 2000) proposed by Kalisch and Bühlmann (2007) and the ψ-learning algorithm proposed by Liang et al. (2015).

Assume that observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are identically independently distributed with $\mathbf{x}_1 \in \mathbb{R}^p$ and probability distribution F. We let the dimension grow as a function of the sample size. Thus, we rewrite the dimension p as p_n, the distribution F as $F^{(n)}$ and the undirected graph $\mathcal{G} = (V, E)$ as $\mathcal{G}^{(n)} = (V^{(n)}, E^{(n)})$. We define the set of edges:

$$E^{(n)} = \{(i, j) : \rho_{ij} \neq 0,\ i, j = 1, \ldots, p_n\}. \tag{2.3.1}$$

To establish consistency of the Graphical Stepwise procedure, we make the following assumptions (see Section 3.3 Liang et al., 2015):

(A1) The distribution $F^{(n)}$ is multivariate Gaussian.

(A2) The dimension $p_n = O(\exp(n^{\delta}))$ for some constant $0 \leq \delta < 1$.

(A3) The correlation coefficients satisfy:

$$\inf\{|\operatorname{Cor}(X_i, X_j)|;\ \operatorname{Cor}(X_i, X_j) \neq 0,\ i, j = 1, \ldots, p_n,\ i \neq j\} \geq c_0 n^{-\kappa}, \tag{2.3.2}$$

where $c_0 > 0$ and $0 < \kappa < (1-\delta)/2$. Moreover,

$$\sup\{|\operatorname{Cor}(X_i, X_j)|;\ i, j = 1, \ldots, p_n,\ i \neq j;\} \leq M_r < 1, \tag{2.3.3}$$

for some constant $0 < M_r < 1$.

(A4) There exist a constant $c_1 > 0$ and $0 \leq \tau \leq 1-2\kappa$ for some $0 < \kappa < (1-\delta)/2$ such that $\lambda_{\max}(\mathbf{\Sigma}) \leq c_1 n^{\tau}$, where $\lambda_{\max}(\mathbf{\Sigma})$ is the largest eigenvalue of the covariance matrix $\mathbf{\Sigma}$.

Assumption (A1) is often used in graphical modeling. When we compare our approach with that of Kalisch and Bühlmann (2007) and Liang et al. (2015), we do not require the distribution $F^{(n)}$ to be faithful to the graph $\mathcal{G}^{(n)}$. The adjacency faithfulness condition restricts the class of probability distributions. Assumption (A2) allows for exponential growth of the dimension as a function of the sample size. (A3) ensures we can detect nonzero correlations, and restrict the linear dependencies of the variables by requiring an upper bound $0 < M_r < 1$. Assumption (A4) restricts the rate of growth of the maximum eigenvalue of $\mathbf{\Sigma}$ as the sample size increases.

Let denote the empirical partial correlations coefficients between variables X_i and X_j conditioned on the variables $X_{\mathcal{S}_{ij}}$, where the conditioning set is defined as $\mathcal{S}_{ij} = \{k \in V \setminus \{i, j\}\}$, as $\widehat{\operatorname{Cor}}(X_i, X_j | X_{\mathcal{S}_{ij}})$. Let $q_n = \max_{i,j \in \{1,\ldots,p_n\}} |\mathcal{S}_{ij}|$ be the maximal cardinality of the subset set $\mathcal{S}_{ij}$ given in the Graphical Stepwise procedure. Also, we note that $\mathcal{S}_{ij}$ denotes every possible conditioning set. In order to ensure that the conditioning sets are bounded we need to reduce the problem of estimating a GGM form a high-dimensional setting $(n < p)$ to a low-dimensional setting $n > q_n$. Then, we can reduce the conditioning set $\mathcal{S}_{ij}$ by performing a correlation screening proposed by Fan and Lv (2008). For a given threshold $\xi_n > 0$, let $\hat{\mathcal{E}}_{\xi_n,i}$ be the set of candidate neighborhood set of node $i \in V_n$

$$\hat{\mathcal{E}}_{\xi_n,i} = \{j : j \neq i, |\widehat{\mathrm{Cor}}(\mathbf{X}_i, \mathbf{X}_j)| > \xi_n,\ i, j = 1, \ldots, p_n\}. \tag{2.3.4}$$

Lemma 2 gives a probabilistic upper bound for the candidate neighborhood set of node i (see Lemma 2 in Liang et al., 2015).

Lemma 2. *Assume (A1), (A2), (A3) and (A4) hold. Let and* $\xi_n = 2/3c_1 n^{-\kappa}$. *Then, for each node* i,

$$Pr\Big\{|\hat{\mathcal{E}}_{\xi_n,i}| \leq O(n^{2\kappa+\tau})\Big\} \geq 1 - c_2\, exp(-c_3 n^{1-2\kappa}), \tag{2.3.5}$$

for some constants c_2 *and* c_3.

Let γ_n denote the threshold value used to screen the partial correlations in the Graphical Stepwise algorithm. Let $\widehat{E}_{\gamma_n}$ denote the estimated set of edges. We define:

$$\widehat{E}_{\gamma_n} = \{(i, j) : |\widehat{\mathrm{Cor}}(X_i, X_j | X_{\mathcal{S}_{ij}})| > \gamma_n,\ i, j = 1, \ldots, p_n\}. \tag{2.3.6}$$

To establish the consistency of $\widehat{E}_{\gamma_n}$, we assume that the population partial correlations, denoted by $\mathrm{Cor}(X_i, X_j | X_{\mathcal{S}_{ij}})$, with non-zero coefficients in $E^{(n)}$ satisfy the following condition:

(A5) The partial correlation coefficients satisfy:

$$\inf\{|\ \mathrm{Cor}(X_i, X_j | X_{\mathcal{S}_{ij}})\ |;\ \rho_{ij} \neq 0,\ i, j = 1, \ldots, p_n,\ i \neq j; |\mathcal{S}_{ij}| \leq q_n\} \geq c_4 n^{-d}, \tag{2.3.7}$$

where $q_n = O(n^{2\kappa+\tau})$, $0 < c_4 < \infty$ and $0 < d < (1-\delta)/2$. Moreover,

$$\sup\{|\operatorname{Cor}(X_i, X_j|X_{\mathcal{S}_{ij}})|;\ i,j = 1,\dots,p_n,\ i \neq j; |\mathcal{S}_{ij}| \leq q_n\} \leq M < 1, \quad (2.3.8)$$

for some constant $0 < M < 1$.

Assumption (A5) ensures that we can detect non-zero partial correlation coefficients and restrict the linear dependencies between the variables by requiring an upper bound $0 < M < 1$.

The following Lemma is concerned with the uniform consistency of the estimated partial correlations coefficients, which is adopted from Corollary 1 in Kalisch and Bühlmann (2007) and assuming that the cardinality of the conditioning sets is bounded (see Lemma 2 in Liang et al., 2015).

Lemma 3. *Assume (A1) and (A5) hold and $q_n < n-4$. Then, for any $0 < \gamma < 2$,*

$$\begin{aligned}\sup_{i,j\in\{1,\dots,p_n\}} & Pr\left\{|\widehat{Cor}(X_i, X_j|X_{\mathcal{S}_{ij}}) - Cor(X_i, X_j|X_{\mathcal{S}_{ij}})| > \gamma\right\} \\ & \leq c_5(n-q_n-2)exp\left\{(n-q_n-4)log\left(\frac{4-\gamma^2}{4+\gamma^2}\right)\right\},\end{aligned} \quad (2.3.9)$$

for some constant $0 < c_5 < \infty$ depending on M in (A4) only.

In Algorithm 1, we select sequentially ordered pair of nodes by finding the variables with the largest absolute empirical partial correlation and then adding to the selected active set of edges in each step. Theorem 2 establishes the consistency of the Graphical Stepwise algorithm.

Theorem 2. *Consider a GGM with distribution $F^{(n)}$ and underlying undirected graph $\mathcal{G}^{(n)} = (V^{(n)}, E^{(n)})$. Assume (A1)-(A5) hold and let $\gamma_n = \frac{1}{2}c_4 n^{-d}$. Then,*

$$Pr\left\{\widehat{E}_{\gamma_n} = E^{(n)}\right\} \geq 1 - o(1), \quad as\ n \to \infty. \quad (2.3.10)$$

To prove Theorem 2 we follow the analysis propose by Kalisch and Bühlmann (2007), Bühlmann et al. (2010) and Bühlmann and Van De Geer (2011). Detail of the proof is given in Section 2.6.

2.4 Numerical Results

In this section, we conduct an exhaustive numerical study on the performance of the Graphical Stepwise procedure. We consider the empirical performance for simulated and real data. We compare our estimation method with the Graphical lasso (Glasso) proposed by Friedman et al. (2008) and CLIME proposed by Cai et al. (2011). These methods aim to estimate $\mathbf{\Omega}$ by solving the following ℓ_1 penalized-likelihood problem

$$\min_{\mathbf{\Omega}\succ 0} -\log(\det(\mathbf{\Omega})) + \operatorname{tr}(\mathbf{\Omega}\mathbf{X}^T\mathbf{X}) + \theta \parallel \mathbf{\Omega} \parallel_1, \tag{2.4.1}$$

where $\theta \geq 0$ is the regularization parameter and $\parallel \mathbf{\Omega} \parallel_1 = \sum_{j\neq i} \mid \omega_{ij} \mid$ the element-wise ℓ_1 norm.

2.4.1 Simulation Experiments

In this section we present simulation experiments to examine the performance of the proposed method to estimate high-dimensional GGMs. We first investigate the numerical and classification performance of our procedure and we compare with that of Graphical lasso (Glasso) and CLIME.

We draw n independent samples from a multivariate Gaussian distribution with mean zero and covariance matrix $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$. We fix $n = 100$ and consider different values of $p = \{90, 100, 120\}$. We consider four different specifications for the population precision-matrix $\mathbf{\Omega}$:

1. AR(1) Model: $\omega_{ii} = 1$, $\omega_{i,i+1} = \omega_{i-1,i} = 0.4$ and 0 otherwise.

2. AR(2) Model: $\omega_{ii} = 1$, $\omega_{i,i+1} = \omega_{i-1,i} = 0.4$, $\omega_{i,i+2} = \omega_{i-2,i} = 0.2$ and 0 otherwise.

3. 2-nearest-neighbor graph: p points are randomly selected from a unit square and all pairwise distances among the points are computed. Then, the 2 nearest neighbors of each node are selected. The entries of the precision matrix are randomly chosen from the interval $[-1, -0.5] \cup [0.5, 1]$. To ensure that the precision matrix is positive definite the matrix is normalized as: $\mathbf{\Omega} + (\lambda(\mathbf{\Omega})_{\min} + 0.2)I_p$ where $\lambda(\mathbf{\Omega})_{\min}$ refers to the smallest eigenvalue.

4. Block Graph: $\mathbf{\Omega}$ is a block diagonal matrix with block size $p/4$. Each block has off-diagonal elements equal to 0.5 and diagonal elements equal to 1. The matrix is guarantee to be positive definite. The resulting matrix is randomly permuted by rows and columns. The graph have approximately 90 edges when $p = 60$, 150 edges when $p = 100$ and 180 edges when $p = 120$.

To select the threshold γ we follow the scheme proposed by Cai et al. (2011). For each model, we generate a training sample of size $n = 100$ from a multivariate Gaussian distribution with mean zero and covariance matrix $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$, and an independent sample of the same distribution for validating the thresholding parameter. From the training data, we estimate the different precision matrices estimators for 50 different values of γ. The optimal parameter is given by the one that minimize the log-likelihood loss defined by:

$$-\log(\det(\widehat{\mathbf{\Omega}})) + \mathrm{tr}(\widehat{\mathbf{\Omega}}\mathbf{X}^T\mathbf{X}) - p.$$

The same scheme is used to choose the regularization parameter of Glasso and CLIME. We replicate each simulation experiment 100 times.

To evaluate the performance of the method, we study specific assessment measures to evaluate support recovery and numerical performance. To compare the numerical performance, we compute the Frobenius and spectral norm between $\boldsymbol{\Omega}$ and $\widehat{\boldsymbol{\Omega}}$. Moreover, we evaluate the performance of the estimator $\widehat{\boldsymbol{\Omega}}$ with the expected value of the Likelihood Ratio Test (LRT), measured by $\mathrm{E}(\mathrm{LRT}(\widehat{\boldsymbol{\Omega}}))$, where $\mathrm{LRT}(\widehat{\boldsymbol{\Omega}})$ is the likelihood ratio distance computed as

$$\mathrm{LRT}(\widehat{\boldsymbol{\Omega}}) = \mathrm{tr}(\widehat{\boldsymbol{\Omega}}\boldsymbol{\Omega}^{-1}) - \log(\det(\widehat{\boldsymbol{\Omega}}\boldsymbol{\Omega}^{-1})) - p, \tag{2.4.2}$$

small values imply a better performance of the method in estimating the true $\boldsymbol{\Omega}$ (see Danilov et al., 2012).

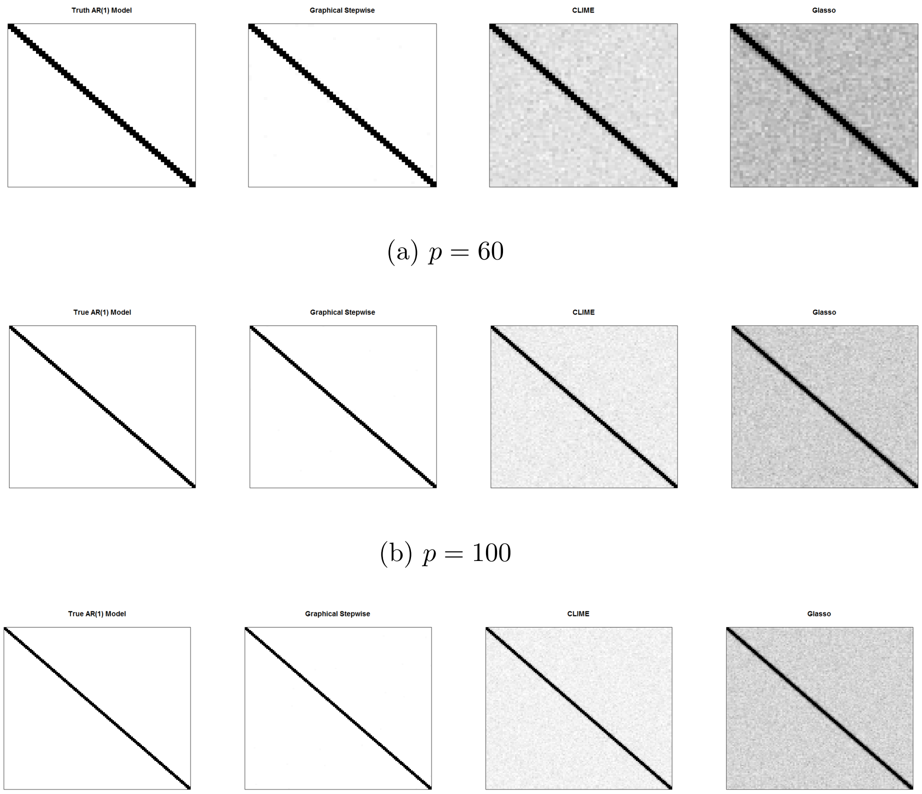
The graph structure recovery is evaluated by specificity, sensitivity, and Mathews correlation coefficient (MCC) criteria, defined as follows:

$$\text{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \quad \text{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{2.4.3}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TNFP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}, \tag{2.4.4}$$

where TP be the true non-zero elements and TN be the true zero elements estimated by $\widehat{\boldsymbol{\Omega}}$ and FP be the false non-zero elements and FN be the false zero elements estimated by $\widehat{\boldsymbol{\Omega}}$.

We first evaluate the estimation performance. Table 2.2 reports the average and standard errors of LRT and matrix losses. We see that Graphical Stepwise uniformly outperforms Glasso. The improvement is more significant when the graph is sparse and $n < p$. When we compare the Graphical Stepwise average LRT and Frobenius norm with that of CLIME, we observe that our procedure outperforms CLIME and is specially favorable when p is large and $\boldsymbol{\Omega}$ is sparse. We observe that for the AR(1) and 2-nearest-neighbor specifications, CLIME shows a slightly better behavior than Graphical Stepwise in terms of spectral norm.



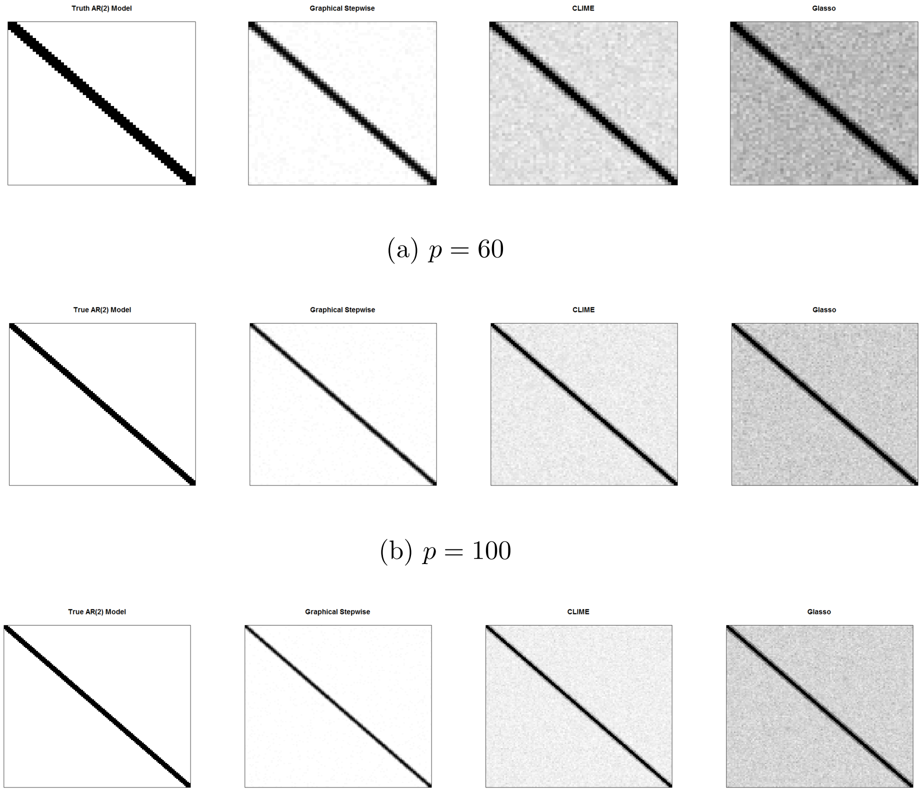
(a) $p = 60$

(b) $p = 100$

(c) $p = 120$

Figure 2.4: AR(1) Model. Heatmaps of the frequency of the zeros identified for each entry of $\mathbf{\Omega}$ out of 100 replications. White represents 100 zeros identified out of 100 runs, and black represents 0/100.

Regarding the support recovery, Table 2.3 shows the classification performance for the four specifications. We observe that Graphical Stepwise significantly outperforms CLIME and Glasso in terms of the overall classification performance measure by the MCC criteria. Our procedure estimate more sparse graphs than Glasso and CLIME. Glasso tends to introduce erroneous non-zero elements. To illustrate the recovery performance, Figures 2.4 to 2.7 show the heatmaps out of 100 replications. We observe that Graphical Stepwise estimates a sparsity pattern that is closely related with the true model. We also note that CLIME and Glasso tend to introduce false positive edges.

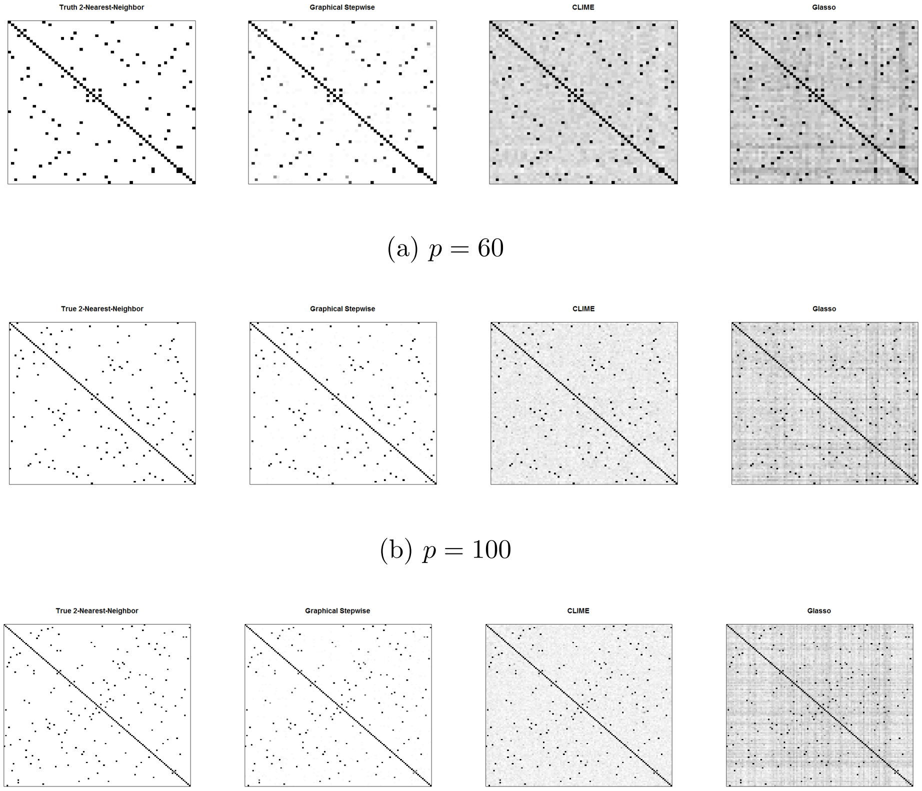


(a) $p = 60$

(b) $p = 100$

(c) $p = 120$

Figure 2.5: AR(2) Model. Heatmaps of the frequency of the zeros identified for each entry of $\boldsymbol{\Omega}$ out of 100 replications. White represents 100 zeros identified out of 100 runs, and black represents 0/100.

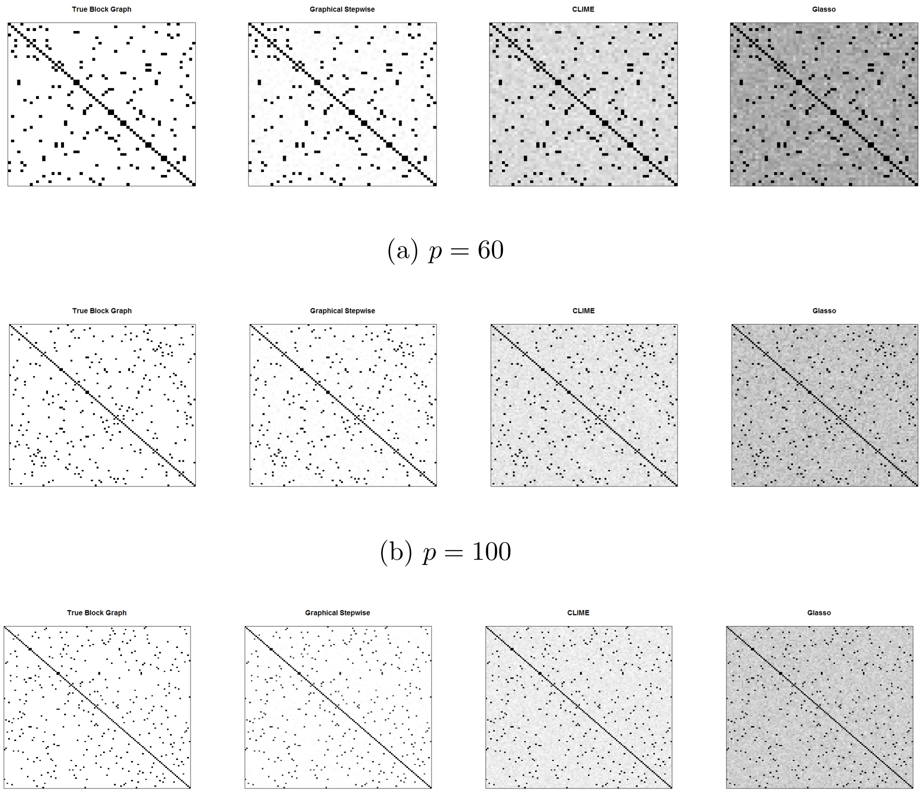


(a) $p = 60$

(b) $p = 100$

(c) $p = 120$

Figure 2.6: 2-nearest-neighbor graph. Heatmaps of the frequency of the zeros identified for each entry of $\mathbf{\Omega}$ out of 100 replications. White represents 100 zeros identified out of 100 runs, and black represents 0/100.



(a) $p = 60$

(b) $p = 100$

(c) $p = 120$

Figure 2.7: Block graph. Heatmaps of the frequency of the zeros identified for each entry of $\mathbf{\Omega}$ out of 100 replications. White represents 100 zeros identified out of 100 runs, and black represents 0/100.

Table 2.2: Comparison of average numerical performance for four models over 100 replications with standard deviation in brackets.

		Graphical Stepwise			CLIME			Glasso		
Model	p	LRT	Frobenius	Spectral	LRT	Frobenius	Spectral	LRT	Frobenius	Spectral
	60	2.203	1.914	0.749	4.378	2.563	0.730	4.571	2.876	0.783
		(0.448)	(0.217)	(0.178)	(0.392)	(0.139)	(0.049)	(0.318)	(0.113)	(0.033)
AR(1)	100	4.076	2.536	0.824	8.533	3.646	0.803	9.093	4.097	0.853
		(0.704)	(0.218)	(0.193)	(0.593)	(0.194)	(0.044)	(0.403)	(0.137)	(0.029)
	120	4.843	2.744	0.822	10.811	4.089	0.816	11.526	4.637	0.876
		(0.795)	(0.217)	(0.131)	(0.669)	(0.178)	(0.046)	(0.433)	(0.116)	(0.027)
	60	6.024	3.268	1.030	5.612	3.701	1.120	6.032	4.011	1.187
		(0.763)	(0.226)	(0.127)	(0.392)	(0.179)	(0.066)	(0.264)	(0.116)	(0.038)
AR(2)	100	11.870	4.565	1.136	10.516	5.280	1.242	11.474	5.661	1.296
		(1.054)	(0.234)	(0.150)	(0.489)	(0.161)	(0.044)	(0.315)	(0.121)	(0.029)
	120	14.956	5.152	1.133	13.287	5.944	1.274	14.355	6.351	1.327
		(1.147)	(0.236)	(0.119)	(0.522)	(0.197)	(0.045)	(0.356)	(0.123)	(0.027)
	60	2.707	2.305	0.864	3.471	2.704	0.861	3.862	2.878	0.921
		(0.496)	(0.220)	(0.162)	(0.246)	(0.132)	(0.081)	(0.211)	(0.104)	(0.057)
Neighbor	100	4.065	2.712	0.862	6.214	3.395	0.809	7.200	3.804	0.877
		(0.640)	(0.217)	(0.147)	(0.457)	(0.149)	(0.059)	(0.339)	(0.112)	(0.047)
	120	5.569	3.216	0.952	7.969	3.790	0.854	9.277	4.288	0.929
		(0.731)	(0.239)	(0.207)	(0.516)	(0.156)	(0.065)	(0.374)	(0.122)	(0.040)
	60	4.517	2.987	1.245	5.738	4.651	1.402	7.552	5.593	1.543
		(0.830)	(0.333)	(0.222)	(0.470)	(0.279)	(0.073)	(0.291)	(0.182)	(0.046)
Block	100	9.702	4.287	1.373	11.412	6.843	1.570	14.329	7.923	1.686
		(1.472)	(0.396)	(0.174)	(0.640)	(0.308)	(0.061)	(0.316)	(0.135)	(0.028)
	120	14.921	5.799	1.520	14.763	7.941	1.644	17.928	8.940	1.733
		(1.569)	(0.351)	(0.089)	(0.553)	(0.235)	(0.052)	(0.282)	(0.157)	(0.027)

Table 2.3: Comparison of average support recovery for four models over 100 replications with standard deviation in brackets.

		Graphical Stepwise			CLIME			Glasso		
Model	p	Sensitivity	Specificity	MCC	Sensitivity	Specificity	MCC	Sensitivity	Specificity	MCC
AR(1)	60	0.983	0.999	0.971	0.999	0.830	0.374	1.000	0.743	0.295
		(0.017)	(0.001)	(0.021)	(0.003)	(0.028)	(0.031)	(0.000)	(0.022)	(0.015)
	100	0.975	0.999	0.965	0.999	0.900	0.392	1.000	0.804	0.275
		(0.018)	(0.001)	(0.017)	(0.003)	(0.019)	(0.034)	(0.001)	(0.016)	(0.013)
	120	0.975	0.999	0.968	0.999	0.921	0.403	1.000	0.823	0.267
		(0.016)	(0.000)	(0.014)	(0.003)	(0.012)	(0.032)	(0.000)	(0.011)	(0.010)
AR(2)	60	0.651	0.990	0.714	0.773	0.841	0.382	0.834	0.723	0.297
		(0.091)	(0.005)	(0.052)	(0.045)	(0.030)	(0.028)	(0.037)	(0.025)	(0.018)
	100	0.566	0.994	0.657	0.705	0.898	0.356	0.775	0.803	0.272
		(0.079)	(0.002)	(0.041)	(0.030)	(0.017)	(0.025)	(0.030)	(0.017)	(0.014)
	120	0.523	0.995	0.633	0.669	0.917	0.348	0.751	0.824	0.259
		(0.072)	(0.002)	(0.040)	(0.033)	(0.017)	(0.028)	(0.028)	(0.016)	(0.011)
Neighbor	60	0.847	0.997	0.858	0.975	0.826	0.329	0.973	0.787	0.290
		(0.070)	(0.002)	(0.039)	(0.024)	(0.034)	(0.033)	(0.022)	(0.022)	(0.016)
	100	0.909	0.999	0.902	0.987	0.906	0.339	0.991	0.843	0.259
		(0.042)	(0.001)	(0.030)	(0.013)	(0.017)	(0.032)	(0.011)	(0.013)	(0.013)
	120	0.877	0.998	0.873	0.972	0.926	0.352	0.974	0.856	0.250
		(0.041)	(0.001)	(0.030)	(0.017)	(0.013)	(0.030)	(0.015)	(0.014)	(0.013)
Block	60	0.956	0.988	0.871	0.995	0.813	0.422	0.993	0.673	0.303
		(0.031)	(0.005)	(0.037)	(0.008)	(0.028)	(0.031)	(0.010)	(0.027)	(0.015)
	100	0.938	0.990	0.831	0.976	0.874	0.408	0.970	0.772	0.292
		(0.044)	(0.003)	(0.031)	(0.015)	(0.019)	(0.030)	(0.016)	(0.018)	(0.012)
	120	0.797	0.995	0.795	0.960	0.902	0.417	0.953	0.804	0.288
		(0.045)	(0.001)	(0.034)	(0.016)	(0.013)	(0.020)	(0.020)	(0.018)	(0.011)

2.4.2 Analysis of Breast Cancer Data

We apply the procedure to evaluate gene expression profiling to breast cancer patients data to predict who may achieve pathological complete response (pCR). Using normalized gene expression data of patients in stages I-III of breast cancer data analyzed by Hess et al. (2006), we aim to predict response state to neoadjuvant (preoperative) chemoterapy of patients with pathological complete response (pCR) and with residual disease (RD). The importance of study the subject response to neoadjuvant (preoperative) chemoterapy, resides in the fact that complete eradication of all invasive cancer (i.e. pCR) is associated with long-term cancer free survival.

The data set consist of 22,283 gene expression levels of 133 subjects, with 34 pCR and 99 RD, respectively. We follow the analysis scheme proposed by Fan et al. (2009) and Cai et al. (2011). The data is randomly split into the training and testing set, we repeat this procedure 100 times. The testing set is formed by randomly selecting 5 pCR subjects and 16 RD subjects (approximately 1/6 subjects in each group). The remaining subjects form the training set. From the training set a Wilcox singed-rank test is performed to select the 113 most significant genes.

Based on the estimate of the precision matrix, we apply a linear discriminant analysis (LDA) to predict whether a patient may achieve pathological complete response (pCR). From the training set, we compute the mean and precision matrix estimates. For the test data we compute the linear discriminant score as follows

$$\delta_r(\mathbf{X}_i) = \mathbf{X}_i^T \widehat{\boldsymbol{\Omega}} \widehat{\boldsymbol{\mu}}_r - \frac{1}{2} \boldsymbol{\mu}_r^T \widehat{\boldsymbol{\Omega}} \boldsymbol{\mu}_r + \log \widehat{\pi}_r \quad \text{for } i = 1, \ldots, n, \tag{2.4.5}$$

where $\widehat{\pi}_r$ is the proportion of group r subjects in the training set, $\boldsymbol{\mu}_r$ the sample mean of group r and $\widehat{\boldsymbol{\Omega}}$ the precision matrix estimate for the whole training set. The classification rule is taken to be

$$\widehat{r}(\mathbf{X}_i) = \text{argmax}\, \delta_t(\mathbf{X}_i) \quad \text{for } r = 1, 2. \tag{2.4.6}$$

To perform model selection we use 5-fold cross validation on the training data.

Table 3.7 displays the average classification performance and the number of miss-classified tumor samples for each precision matrix estimator. We can see that Graphical Stepwise and CLIME improve over Glasso in terms of Sensitivity, MCC and the classification error. While all three methods give similar Specificity performance. Graphical Stepwise is slightly better than CLIME in classifying the pCR subjects, which is measure by the MCC and Testing Set Error values. Moreover, our procedure estimates sparse precision matrices then, we can obtain simpler models with small number of edges which is usually favorable for interpreting real data sets.

Table 2.4: Comparison of average pCR classification errors over 100 replications with standard deviation in brackets.

	Sensitivity	Specificity	MCC	Test Error Set	Sparsity
Graphical Stepwise	0.718	0.804	0.484	0.216	0.004
	(0.187)	(0.083)	(0.179)	(0.076)	(0.002)
CLIME	0.652	0.812	0.437	0.226	0.788
	(0.210)	(0.082)	(0.199)	(0.078)	(0.045)
Glasso	0.522	0.828	0.345	0.245	0.230
	(0.216)	(0.086)	(0.214)	(0.077)	(0.022)

2.5 Conclusions

In this article, we have presented an integrated approach to estimate undirected graphs and to perform model selection in high dimensional Gaussian Graphical Models (GGMs). We consider a parametrization of the precision matrix in terms of the prediction errors of the best linear predictor of each node in the graph. We exploit the relationship between partial correlation coefficients and the distribution of the prediction errors. We propose a novel forward-backward algorithm for detecting pairs of variables having nonzero partial correlations among a large number of random variables based on i.i.d. samples. We obtain a set of the most probable edges in a GGM which are related with the largest absolute partial correlation coefficients. The position of the new non-zero element in the precision matrix corresponds to the pair of variables with the largest absolute partial correlation conditioned on the set of active nodes previously detected. We show that under mild conditions the Graphical Stepwise procedure is able to consistently estimated the set of true edges. The novelty of the approach is that we can obtain a set of more probable edges in a GGM for a given threshold value without resorting to penalized regression procedures. The Graphical Stepwise has good numerical and GGM classification performance when sparse precision matrices are estimated. Simulation studies show that the procedure is able to detect the true set of edges. The numerical examples indicate that our procedure outperforms existing algorithms, such as the Graphical lasso and CLIME. Applications to real data to perform a classification analysis show that our approach has a satisfactory predictive performance. Finally, we note that the computation time of the Graphical Stepwise algorithm is similar to that of CLIME and it depends strongly on the value of the threshold.

There are several possible extensions of our method. In the forward-backward procedure we use a constant threshold value to select edges. A possible extension is to consider that the threshold varies with the effective sample size, which is given by $n - |\mathcal{S}_{ij}| - 3$, where $\mathcal{S}_{ij}$ is the separator set od nodes i and j. This adjustment is able to improve the performance of the Graphical Stepwise procedure. Furthermore, we can apply multiple testing hypothesis procedure for selecting the non-zero partial correlation coefficients. Liang et al. (2015) introduce a generalized Bayesian method for conducting multiple hypothesis testing that can be apply to test for conditional independence in our procedure. Finally, our procedure can be extended to binary graphical models by replacing linear regressions with logistic regression (see Ravikumar et al., 2010). For non-Gaussian random variables we could apply the nonparanormal transformation proposed by Liu et al. (2009) or the rank-based partial correlation coefficient proposed by Harris and Drton (2013).

2.6 Proof of Main Results

2.6.1 Proof of Lemma 1

Proof. We will prove item a). The proof of item b) can be found in pp. 32–33 of Kurowicka and Cooke (2006) and the proof of c) is straightforward.

Let $\mathbf{v}^T = (v_1, \ldots, v_{p-1}) \in \mathbb{R}^{(p-1)}$ be a vector. By definition (see (2.1.8)) the optimal predictor $\boldsymbol{\beta}^i$ minimizes, over the set $\{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_i = 0\}$, the conditional expectation $h(\boldsymbol{\beta}) = \mathrm{E}_{X_i|\mathbf{x}_{-i}=\mathbf{v}} \left(X_i - \sum_{j \in V} \beta_j X_j \right)^2$. By (2.1.6) and (3.1.6), $\mathrm{E}_{X_i|\mathbf{x}_{-i}=\mathbf{v}} (X_i) = \mu_{i|-i}$ and $\Sigma_{i|-i} = \mathrm{Var}_{X_i|\mathbf{x}_{-i}} (X_i)$ and, in consequence,

$$\begin{aligned} h(\boldsymbol{\beta}) &= \mathrm{E}_{X_i|\mathbf{x}_{-i}=\mathbf{v}} \left(X_i^2 \right) - 2 \sum_{j \in V} \mathrm{E}_{X_i|\mathbf{x}_{-i}=\mathbf{v}} \left(X_i \beta_j X_j \right) + \\ &\quad + \mathrm{E}_{X_i|\mathbf{x}_{-i}=\mathbf{v}} \left(\sum_{j \in V} \beta_j X_j \right)^2 \\ &= \mu_{i|-i}^2 + \Sigma_{i|-i} - 2\mu_{i|-i} \sum_{j \in V} \beta_j v_j + \left(\sum_{j \in V} \beta_j v_j \right)^2 . \end{aligned}$$

Hence the minimum value of $h(\boldsymbol{\beta})$ solves the set of equations:

$$\begin{aligned} \forall k \in V \setminus i:\ 0 &= \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_k} \text{ or, equivalently} \\ \forall k \in V \setminus i:\ 0 &= -2\mu_{i|-i} v_k + v_k \left(\sum_{j \in V} \beta_j v_j \right). \end{aligned}$$

So, the solution $\boldsymbol{\beta}^i$ satisfies

$$\mu_{i|-i} = \mathbf{x}_{-i}^T \boldsymbol{\beta}^i \tag{2.6.1}$$

(given $\mathbf{x}_{-i} = \mathbf{v}$).

By equation (C.4) of Lauritzen (1996, p. 256) $\omega_{ii}^{-1} \boldsymbol{\Omega}_{i,-i} = -\boldsymbol{\Sigma}_{i,-i} \boldsymbol{\Sigma}_{-i,-i}^{-1}$ (see (2.1.6)). Hence, by this equality and considering that $\mu_{i|-i} = \boldsymbol{\Sigma}_{i,-i} \boldsymbol{\Sigma}_{-i,-i}^{-1} \mathbf{x}_{-i}$ (see (3.1.6)) we have

$$-\omega_{ii}^{-1} \boldsymbol{\Omega}_{i,-i} \mathbf{x}_{-i} = -\boldsymbol{\Sigma}_{i,-i} \boldsymbol{\Sigma}_{-i,-i}^{-1} \mathbf{x}_{-i} = \mu_{i|-i} \tag{2.6.2}$$

By (2.6.1) and (2.6.2) $\boldsymbol{\beta}^i = -\omega_{ii}^{-1} \boldsymbol{\Omega}_{i,-i}$ minimices $h(\boldsymbol{\beta})$ and then if β_j^i is the j-th component of $\boldsymbol{\beta}^i$ then $\beta_j^i = -\dfrac{\omega_{i,j}}{\omega_{ii}}$. □

2.6.2 Proof of Proposition 3

Proof. Let $I = V \setminus \{i, j\}$, $J = \{i, j\}$, $\mathbf{x} = (X_1, \ldots, X_p)^T$, $\mathbf{x}_J = (X_i, X_j)^T$ and $\mathbf{x}_I = \mathbf{x}_{-\{i,j\}}$, as before, the vector containing the remaining variables with indexes in the set $V \setminus \{i, j\}$ considered in ascending order.

Let $f_{X_i|\mathbf{x}_I}$ and $f_{X_j|\mathbf{x}_I}$ be the conditional distributions of X_i and X_j given $\mathbf{x}_I$. By definition and a very simple calculation

$$\begin{aligned} \mathrm{Inf}\left(X_i \perp\!\!\!\perp X_j \mid \mathbf{x}_I\right) &= \mathrm{E}\left(\log \frac{f_{\mathbf{x}}}{f_{X_i|\mathbf{x}_I} f_{X_j|\mathbf{x}_I} f_{\mathbf{x}_I}}\right) \\ &= \mathrm{E}\left(\log \frac{f_{\mathbf{x}_J|\mathbf{x}_I}}{f_{X_i|\mathbf{x}_I} f_{X_j|\mathbf{x}_I}}\right) \\ &= \mathrm{E}\left(\log f_{\mathbf{x}_J|\mathbf{x}_I}\right) - 2\mathrm{E}\left(\log f_{X_i|\mathbf{x}_I}\right) \end{aligned} \tag{2.6.3}$$

Using (C.2) of Lauritzen, 1996

$$X_i \mid \mathbf{x}_I \sim N(\mu_{i|I}, \Sigma_{i|I}). \tag{2.6.4}$$

where $\mu_{i|I} = \mathbf{\Sigma}_{i,I}\mathbf{\Sigma}_{II}^{-1}\mathbf{x}_I$ and $\Sigma_{i|I} = \mathbf{\Sigma}_{ii} - \mathbf{\Sigma}_{i,I}\mathbf{\Sigma}_{II}^{-1}\Sigma_{I,i}$.

Analogously, the conditional distribution of $\mathbf{x}_J$ given $\mathbf{x}_I$ is also Gaussian distributed

$$\mathbf{x}_J \mid \mathbf{x}_I \sim N(\boldsymbol{\mu}_{J|I}, \mathbf{\Sigma}_{J|I}). \tag{2.6.5}$$

where $\boldsymbol{\mu}_{J|I} = \mathbf{\Sigma}_{JI}\mathbf{\Sigma}_{II}^{-1}\mathbf{x}_I$ and $\mathbf{\Sigma}_{J|I} = \mathbf{\Sigma}_{JJ} - \mathbf{\Sigma}_{JI}\mathbf{\Sigma}_{II}^{-1}\mathbf{\Sigma}_{IJ}$

From the definition of Gaussian density functions we have

$$\begin{aligned} \mathrm{E}\left(\log f_{X_i|\mathbf{x_I}}\right) &= -\frac{1}{2}\left(\log(2\pi) + \log(\Sigma_{i|I})\right) - \frac{1}{2}\frac{\mathrm{E}(X_i - \mu_{i|I})^2}{\Sigma_{i|I}} \\ &= -\frac{1}{2}\left(\log(2\pi) + \log(\Sigma_{i|I})\right) - \frac{1}{2}. \end{aligned} \tag{2.6.6}$$

The partial correlation coefficient, introduced in Subsection 2.1.1, was defined as the correlation coefficient of the conditional distribution of $\mathbf{x}_J$ given $\mathbf{x}_I$. So, denoting for $k = i, j : Z_k = \left(X_k - \mu_{k|I}\right)$ the centered variables, we have that

$$\begin{aligned}\mathrm{E}\left(\log f_{\mathbf{x}_J|\mathbf{x}_I}\right) &= -\log\left(2\pi\Sigma_{i|I}\Sigma_{j|I}\sqrt{1-\rho_{ij}^2}\right) \\ &\quad -\frac{1}{2(1-\rho_{ij}^2)}\left(\frac{\mathrm{E}(Z_i)^2}{\Sigma_{i|I}}+\frac{\mathrm{E}(Z_j)^2}{\Sigma_{j|I}}-2\rho_{ij}\frac{\mathrm{E}(Z_iZ_j)}{(\Sigma_{i|I}\Sigma_{j|I})^{1/2}}\right) \\ &= -\log(2\pi)-\log\left(\Sigma_{i|I}\Sigma_{j|I}\sqrt{1-\rho_{ij}^2}\right)-1.\end{aligned} \tag{2.6.7}$$

Replacing (2.6.6) and (2.6.7) in (2.6.3) the statement in the proposition is proved.

□

2.6.3 Proof of Theorem 2

Proof. We follow the scheme of the proof of Lemma 4 in Kalisch and Bühlmann (2007). The Graphical Stepwise algorithm estimates partial correlations conditioning on the set of neighborhoods that corresponds to the active variables at each step. Let $A_{ij|\mathcal{S}_{ij}}$ denote the event an error occur in the Graphical Stepwise procedure when testing partial correlations for zero at nodes (i,j) conditioning on the set $\mathcal{S}_{ij}$. Hence,

$$\Pr\left\{\text{an error occur in } \widehat{E}_{\gamma_n}\right\} = \Pr\{\cup_{i\neq j}A_{ij|\mathcal{S}_{ij}}\} \leq O(p_n^{2+q_n})\sup_{i\neq j}\Pr\{A_{ij|\mathcal{S}_{ij}}\}. \tag{2.6.8}$$

Let $A^I_{ij|\mathcal{S}_{ij}}$ and $A^{II}_{ij|\mathcal{S}_{ij}}$ denote the false positive and false negative errors, respectively. Then,

$$A_{ij|\mathcal{S}_{ij}} = A^I_{ij|\mathcal{S}_{ij}} \cup A^{II}_{ij|\mathcal{S}_{ij}}, \tag{2.6.9}$$

where:

$$\begin{aligned}&\text{False positive error } A^I_{ij|\mathcal{S}_{ij}} : \ |\widehat{\mathrm{Cor}}(X_i,X_j|X_{\mathcal{S}_{ij}})| > \frac{c_4}{2}n^{-d} \quad \text{and } \rho_{ij}=0 \\ &\text{False negative error } A^{II}_{ij|\mathcal{S}_{ij}} : \ |\widehat{\mathrm{Cor}}(X_i,X_j|X_{\mathcal{S}_{ij}})| \leq \frac{c_4}{2}n^{-d} \quad \text{and } \rho_{ij}\neq 0.\end{aligned} \tag{2.6.10}$$

Then, using Lemma 3 and that $\log((4-a^2)/(4+a^2))$ converge to $-a^2/2$ as $a \to 0$, there exist a constant $0 < C < \infty$ such that:

$$\begin{aligned} &\sup_{ij} \Pr\{A^{I}_{ij|\mathcal{S}_{ij}}\} = \sup_{ij} \Pr\left\{|\widehat{\mathrm{Cor}}(X_i, X_j|X_{\mathcal{S}_{ij}}) - \mathrm{Cor}(X_i, X_j|X_{\mathcal{S}_{ij}})| > \frac{c_4}{2} n^{-d}\right\} \\ &\sup_{ij} \Pr\{A^{I}_{ij|\mathcal{S}_{ij}}\} \leq O(n-q_n)\mathrm{exp}\left\{-Cn^{-2d}(n-q_n)\right\}, \end{aligned} \tag{2.6.11}$$

The probability of the false negative error is given by:

$$\sup_{ij} \Pr\{A^{II}_{ij|\mathcal{S}_{ij}}\} = \sup_{ij} \Pr\left\{|\widehat{\mathrm{Cor}}(X_i, X_j|X_{\mathcal{S}_{ij}})| \leq \frac{c_4}{2} n^{-d}\right\}. \tag{2.6.12}$$

Moreover, by assumption (A4) $\min_{ij} |\widehat{\mathrm{Cor}}(X_i, X_j|X_{\mathcal{S}_{ij}})| \geq c_4 n^{-d}$,

$$\sup_{ij} \Pr\{A^{II}_{ij|\mathcal{S}_{ij}}\} \leq \sup_{ij} \Pr\left\{|\widehat{\mathrm{Cor}}(X_i, X_j|X_{\mathcal{S}_{ij}}) - \mathrm{Cor}(X_i, X_j|X_{\mathcal{S}_{ij}})| > \frac{c_4}{2} n^{-d}\right\}. \tag{2.6.13}$$

By Lemma 3, we have

$$\sup_{ij} \Pr\{A^{II}_{ij|\mathcal{S}_{ij}}\} \leq O(n-q_n)\mathrm{exp}\left\{-Cn^{-2d}(n-q_n)\right\}, \tag{2.6.14}$$

for some constant $0 < C < \infty$. From the results in (2.6.11) and (2.6.14), we have

$$\Pr\left\{\text{an error occur in } \widehat{E}_{\gamma_n}\right\} \leq O(p_n^{2+q_n}(n-q_n))\mathrm{exp}\left\{-Cn^{-2d}(n-q_n)\right\} = o(1), \tag{2.6.15}$$

since $0 < d < (1-\delta)/2$, also in (A2) we assume that $\log(p_n) = n^{\delta}$ for some $0 \leq \delta < 1$ and from Lemma 2 we have $q_n = O(n^{2\kappa+\tau})$. □

Chapter 3

Robust and Sparse Estimation of High-dimensional Precision Matrices via Bivariate Outlier Detection

We consider the problem of estimating high-dimensional undirected graphs when the data possibly contains anomalies that are difficult to visualize and clean. Given n independent samples of a p-dimensional random vector $\mathbf{x} = (X_1, \ldots, X_p)$, we can represent the linear dependency between variables by an undirected graph. The conditional dependence structure of the distribution can be represented by a graphical model, $\mathcal{G} = (V, E)$, where $V = \{1, \ldots, p\}$ is the set of nodes and E the set of edges in $V \times V$. The undirected graph establishes that if the variables X_i and X_j are connected, then they are adjacent (Lauritzen, 1996). Statistically, we can measure linear dependencies by estimating partial correlations to infer whether there is an association between a pair of variables, conditionally on the rest of them. Furthermore, we can relate the nonzero entries in the precision matrix, denoted by $\boldsymbol{\Omega} = (\omega_{ij})$, with the nonzero partial correlation coefficients (Edwards, 2000). This procedure is known as covariance selection and is widely used to identify the conditional independence restrictions in an undirected graph (Dempster, 1972). In particular, under a Gaussian distribution, the nonzero entries of the precision matrix imply that each pair of variables is conditionally dependent when controlling for the rest of them. These are known in the literature as Gaussian Graphical Models (GGMs) (Lauritzen, 1996).

In a high-dimensional framework, the estimation of $\mathbf{\Omega}$ is not straightforward because of the lack of a pivotal estimator such as the empirical covariance matrix. Moreover, when the dimension p is larger than the number of available observations, the sample covariance matrix is not invertible. And even when the ratio p/n is approximately (but less than) one, the sample covariance matrix is badly conditioned and its inverse tends to amplify the estimation error, which can be observed by the presence of small eigenvalues (Ledoit and Wolf, 2004). From the asymptotic point of view, when both n and p are large (i.e. $p = O(n)$), the sample covariance matrix is not a consistent estimator (El Karoui, 2008). To deal with this problem, several covariance selection procedures have been proposed based on the assumption that $\mathbf{\Omega}$ is mostly composed by zero elements. This suggests that even when $p = O(n)$ the dimension of the problem may still be tractable since the number of edges will grow more slowly than the number of observations (Meinshausen and Bühlmann, 2006).

Several precision matrix estimators have been proposed in the literature. Bickel and Levina (2008a) and Bickel and Levina (2008b) propose banding and thresholding estimators and obtain rates of convergence in the operator norm. El Karoui (2008) propose a hard thresholding estimator for the convariance matrix assuming a flexible notion of sparsity. Other procedures rely on the idea of regression to parameterize a covariance or precision matrix. Some of these include regression-based interpretation of a Cholesky decomposition of the covariance matrix (Pourahmadi, 2007; Rothman et al., 2010). Meinshausen and Bühlmann (2006) propose the neighborhood selection procedure that consistently estimates sparse high-dimensional graphs by estimating a lasso regression for each node in the graph. Peng et al. (2009) present a procedure that simultaneously performs neighborhood selection for all variables to estimate joint sparse regressions, applying an active-shooting to solve the lasso. Yuan (2010) replaces the lasso regression with a Dantzig selector. Liu and Wang (2012) propose an asymptotically tuning-free procedure that estimates the precision matrix in a column-by-column fashion. Zhou et al. (2011) propose an estimator for the precision matrix base on an ℓ_1 regularization and thresholding to infer a sparse undirected graphical model. Ren et al. (2015) propose a nodewise regression approach to obtain assymptotically efficient estimation of each entry of the precision matrix under sparseness conditions.

Penalized likelihood methods have also been introduced for estimating sparse precision matrices. Yuan and Lin (2007) propose to estimate the precision matrix by penalizing the log-likelihood function. Convex and fast algorithms were developed by Banerjee et al. (2008) and Friedman et al. (2008). Friedman et al. (2008) propose the Graphical lasso (Glasso) procedure to estimate sparse precision matrices fitting a modified lasso regression to each variable and solving the problem by a coordinate descent algorithm. Lam and Fan (2009) and Fan et al. (2009) propose methods to diminish the bias imposed by the ℓ_1 penalty by introducing a non-convex SCAD penalty. Cai et al. (2011) estimate precision matrices for both sparse and non-sparse matrices, without imposing a specific sparsity pattern solving the dual of an ℓ_1 penalized maximum likelihood problem. Consistency of penalized likelihood procedures were also explored. Rothman et al. (2008) estimate convergence rates under the Frobeniuos norm and Yuan and Lin (2007), Ravikumar et al. (2008) and Ravikumar et al. (2011) estimate convergence rates for subgaussian distributions.

One of the main drawback of the popular estimation procedures is that they are not well suited to handle noisy data (contaminated by outliers). The existing approaches to estimate the precision matrix and recover the support of the GGM use as input the empirical covariance matrix. The empirical covariance and correlation matrix estimates are very sensitive to the presence of multidimensional outliers (Alqallaf et al., 2002). The violation of the Gaussian assumption may result in poor recovery of the GGM and biased estimation of the precision matrix (see Finegold and Drton, 2011; Liu et al., 2012; Sun and Li, 2012). In the high-dimensional setting, the fraction of perfectly observed rows may be very small. If all components of a row have an independent chance of being contaminated, then the probability that a case is perfectly observed is small. Alqallaf et al. (2009) propose a contamination model where the contamination in each variable is independent from other variables (i.e. componentwise outliers). It generalizes the classical Tukey-Huber row-wise contamination model (see Tukey, 1962; Huber et al., 1964) and allows for cellwise contamination that can be applied to explain the contamination mechanism in Microarrays experiments (see Troyanskaya et al., 2001; Liu et al., 2003). The cellwise contamination model lacks the affine equivariant property. Henceforth, existing approaches for robust covariance estimation such as M-estimates (Maronna, 1976), Minimum Volume Ellipsoid (MVE) and Minimum Covariance Determinant (MCD) estimators (Rousseeuw, 1985, 1984) and the Stahel-Donoho (SD) estimators (Stahel, 1981; Donoho, 1982), may not be reliable in high-dimensional data sets since the operations to compute affine equivariant estimates tend to propagate the effect of multivariate outliers. Also, these estimators downweight contaminated observations to reduce their influence, which produces a significant loss of information when $n < p$.

To deal with outliers in high-dimensional data sets, many procedures construct robust covariance and correlation matrices by using pairwise robust correlation coefficients. Liu et al. (2009) propose to apply a univariate monotone transformation to make the data Gaussian distributed. Then, a robust precision estimator of the correlation matrix can be computed from the transformed data. The estimated correlation matrix is plugged into the existing parametric procedures (the Graphical Lasso, CLIME, or graphical Dantzig Selector) to obtain the final estimate of the inverse correlation matrix and the graph. Liu et al. (2012) and Xue et al. (2012) propose to estimate the unknown correlation matrix with robust nonparametric rank-based statistics Spearman's rho and Kendall's tau. Finegold and Drton (2011) propose to use multivariate t-distribution for more robust inference of graphs. However, there is not a direct relationship between the zero elements on the estimated precision matrix and the conditional independences when a t-distribution is assumed. Sun and Li (2012) propose a robust estimator of the GGM through ℓ_1-penalization of a robustified likelihood function. Öllerer and Croux (2015) and Loh and Tan (2015) propose robust precision matrix estimation under the cellwise contamination setting. These methods estimate robust pairwise scatter covariance using rank-based statistics and plug them into the existing parametric procedures. Öllerer and Croux (2015), and Loh and Tan (2015) analyze the breakdown property of the Graphical lasso and CLIME.

The robust correlation matrix based on univariate transformations to achieve normality are not robust under the presence of structural bivariate outliers which could lead to a misleading graph support recovery. We propose an approach to robustly estimate a Gaussian Graphical Model when there is cellwise contamination in the data. Following the idea of Khan et al. (2007), we estimate robust correlation coefficients applying a bivariate winsorization to the data given an affine equivariant robust correlation coefficient. This transformation allows us to identify bivariate outliers. The proposed correlation matrix is plugged into a parametric procedure to compute the precision matrix. We show that the bivariate winsorized pairwise correlation coefficient converges to the true parameter at the same rate as the affine equivariant correlation coefficient. This result suggests that if the robust correlation coefficient estimator, which is used to winzorize the data, converges to the true parameter at the optimal parametric rate, then the bivariate winsorized correlation matrix achieves the optimal parametric rate of convergence in terms of both precision matrix estimation and graph recovery.

Finally, we perform simulation studies and show that under different contamination settings our procedure outperforms the normal-score based nonpararnomal estimator proposed by Liu et al. (2009) and the nonparanormal SKEPTIC proposed by Liu et al. (2012). We also apply our procedure to the classification of tumors using gene expression data. We show that our procedure achieves good classification performance. The empirical results suggest that, by using bivariate winsorization on the data based on some affine equivariant robust correlation estimate, we can efficiently recover the GGM under cellwise contamination.

The rest of the Chapter is organized as follows. In the next section we briefly review the cellwise contamination model and the existing approaches to estimate robust precision matrices. In Section 3.2 we present the winsorized correlation matrix estimator, which is able to identify structural bivariate outliers under the cellwise contamination mechanism. In Section 3.3 we present a theoretical analysis of the method. In Section 3.4 we present numerical results on simulated data under different contamination mechanisms. Section 3.5 presents the results based on real data where the problem is the classification of tumors using gene expression data. Finally, we discuss the connections to existing methods and possible future directions.

3.1 Problem Setup

In this Section we consider the problem of estimating a high-dimensional undirected graph when the data possibly contains anomalies that are difficult to visualize and clean. A robust statistic must be able to efficiently model the bulk of data points, be resistant to model deviations, and to perform well under the correct model. The performance of a robust estimator can be analyzed with contamination or mixture models. We introduce a general contamination model able to capture properties of high-dimensional outliers, gross errors or missing values, among other perturbed observations. In high-dimension, the fraction of perfectly observed rows may be very small. To deal with this issue, Alqallaf et al. (2009) propose a contamination model where the contamination in each variable is independent from other variables (i.e. componentwise outliers).

Suppose the random vector $\mathbf{x} = (X_1, \dots, X_p)$ has a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and correlation matrix $\boldsymbol{\Gamma} = (\rho_{ij})$. The linear dependency between variables are represented by an undirected graph $\mathcal{G} = (V, E)$, where $V = \{1, \dots, p\}$ is the set of nodes and E the set of edges in $V \times V$. The contamination model can be written as follows:

$$\mathbf{y} = (\mathbf{I} - \mathbf{B})\mathbf{x} + \mathbf{B}\mathbf{z} \tag{3.1.1}$$

where $\mathbf{I}$ is a $p \times p$ identity matrix, $\mathbf{z} \in \mathbb{R}^p$ an arbitrary random vector and $\mathbf{B}$ is the contamination indicator matrix:

$$\mathbf{B} = \begin{bmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_p \end{bmatrix} \tag{3.1.2}$$

and each B_j is a Bernoulli random variable with $P(B_j = 1) = \varepsilon$.

The classical contamination setting or row-wise contamination model, proposed by Tukey (1962) and extended by Huber et al. (1964), assume that $B_1, \dots, B_p$ are fully dependent $P(B_1 = B_2 = \ldots = B_p) = 1$. Then, the observed variable $\mathbf{y}$ is a mixture of two independent distributions. Under this model a fraction $(1 - \varepsilon)$ of the rows are multivariate Gaussian distributed and a fraction ε are outliers. Furthermore, the percentage of contaminated cases is preserved under affine equivariant transformations.

But the Tukey-Huber model does not adequately represent the reality of many multivariate high-dimensional data sets. This model assumes that the majority of the cases are not contaminated. When $p > n$, downweighting an entire case may be inconvenient. The main drawback is that the probability of a perfectly observed row became very small when the number of variables increases (i.e. $p = O(n)$).

Alqallaf et al. (2009) propose an alternative model where the contamination in each variable is independent from other variables (i.e. componentwise outliers). In this model, the variables $B_1, \ldots, B_p$ are independent:

$$P(B_1 = 1) = \ldots = P(B_p = 1) = \varepsilon \tag{3.1.3}$$

Then, the probability of an outlier occurring in the each variable is the same. In this model the probability that a row is not contaminated is $(1-\varepsilon)^p$, which decreases with p. This model allows for cellwise contamination and is denoted by fully independent contamination model.

The fully independent contamination model lacks of affine equivariance. Under the cellwise contamination, each column has on average $(1-\varepsilon)$ of clean observations. Then, linear combinations of these columns produce an increment in the number of contaminated cases (i.e. outlier propagation). Henceforth, in the high-dimensional setting, robust affine equivariant methods are not robust against propagation of outliers.

Under the cellwise contamination model, a robust estimation of the precision matrix $\mathbf{\Omega}$ can be obtained by plugging a robust correlation matrix estimator, denote by $\hat{\mathbf{\Gamma}}$, into the following ℓ_1-regularized log-determinant program (see Öllerer and Croux, 2015; Loh and Tan, 2015):

$$\hat{\mathbf{\Omega}} = \underset{\mathbf{\Omega} \succ 0}{\operatorname{argmin}} \{ \operatorname{tr}(\mathbf{\Omega}\hat{\mathbf{\Gamma}}) - \operatorname{logdet}(\mathbf{\Omega}) + \lambda \parallel \mathbf{\Omega} \parallel_{1,\text{off}} \} \tag{3.1.4}$$

where $\lambda > 0$ is the regularizing constant of the off-diagonal ℓ_1 regularizer

$$\| \mathbf{\Omega} \|_{1,\text{off}} := \sum_{i \neq j} |\omega_{ij}| \quad \text{for } i, j = 1, \ldots, p \tag{3.1.5}$$

Ravikumar et al. (2011) show that, for any positive λ and $\hat{\mathbf{\Gamma}}$ with strictly positive diagonals elements, the problem has a unique solution and the resulting matrix is positive definite (i.e. $\hat{\mathbf{\Omega}} \succ 0$).

Classical approaches for robust scatter estimation such as M-estimates (Maronna, 1976), Minimum Volume Ellipsoid (MVE) and Minimum Covariance Determinant (MCD) estimators (Rousseeuw, 1985, 1984) and the Stahel-Donoho (SD) estimators (Stahel, 1981; Donoho, 1982), are not well suited when the contamination mechanism operates on individual variables (columns) rather than individual cases (rows). Under cellwise contamination each column in the data table contains on average a fraction of ε contaminated observations. Classical affine equivariant estimators apply linear combination of the columns on the original data. This spreads the contamination in one of the cells of an observation over all its components.

To deal with high-dimensional cellwise outliers, Alqallaf et al. (2002) propose to use coordinated wise outlier insensitive transformations to estimate pairwise scatter estimates. These procedures operate one variable at a time and guarantee the protection against outlier propagation.

Let $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}$ be a sample of size n where $\mathbf{y}^{(k)} = (Y_1^{(k)}, \ldots, Y_p^{(k)})^T \in \mathbb{R}^p$ for $k = 1, \ldots, n$. Let's assume that there exists an appropriate score function, denoted by $f_i(Y_i)$, that preserves monotone ordering and commute with permutations of the components of $(Y_i^{(1)}, \ldots, Y_i^{(n)})$. Huber (2011) defines the pairwise robust correlations coefficients through the Person correlation coefficient computed on the outlier free univariate transformed data.

To estimate the robust pairwise correlation matrix, Liu et al. (2009) propose the nonparanormal model where the random vector $\mathbf{y} = (Y_1, \dots, Y_p)^T$ is replaced by the transformed variable $f(\mathbf{y}) = (f_1(Y_1), \dots, f_p(Y_p))^T$ such that $f(\mathbf{y})$ is multivariate Gaussian with mean zero and correlation matrix denoted by $\mathbf{\Gamma}^{npn}$.

Let $\hat{F}_i(t) = \frac{1}{n+1} \sum_{k=1}^{n} I(Y_i^{(k)})$ be the scaled empirical cumulative function of Y_i. To estimate the nonparanormal transformation, Liu et al. (2009) define the coordinated wise transformation function $\hat{f}_i(t) = \Phi^{-1}\left(T_{\delta_n}[\hat{F}_j]\right)$ where $\Phi^{-1}(\cdot)$ is the standard Gaussian quantile function and T_{δ_n} is a winsorization operator defined as

$$T_{\delta_n}(y) = \begin{cases} \delta_n & \text{if} \quad \hat{F}_i(y) < \delta_n \\ y & \text{if} \quad \delta_n \leq \hat{F}_i(y) \leq (1-\delta_n) \\ (1-\delta_n) & \text{if} \quad \hat{F}_i(y) > (1-\delta_n), \end{cases} \tag{3.1.6}$$

where $\delta_n = \frac{1}{4n^{1/4}\sqrt{\pi \text{logn}}}$ is a truncation parameter. The nonparanormal estimate of the correlation matrix is computed as follows

$$\hat{\rho}_{ij}^{npn} = \frac{\frac{1}{n}\sum_{k=1}^{n} \hat{f}_i(Y_i^{(k)})\hat{f}_j(Y_j^{(k)})}{\sqrt{\frac{1}{n}\sum_{k=1}^{n} \hat{f}_i^2(Y_i^{(k)})} \cdot \sqrt{\frac{1}{n}\sum_{k=1}^{n} \hat{f}_j^2(Y_j^{(k)})}}. \tag{3.1.7}$$

Then, the precision matrix nonparanormal estimator is computed by plugging $\mathbf{\Gamma}^{npn}$ into the ℓ_1 log-determinant program (3.1.4). Liu et al. (2009) establish convergence rate for estimating the precision matrix in the Frobenious and spectral norm when p is restricted to a polynomial order of n.

Liu et al. (2012) show that rate of convergence of the nonparanormal estimator is not optimal. Liu et al. (2012) and Xue et al. (2012) present an alternative procedure that applies rank based methods to estimate the pairwise correlation matrix without computing explicitly the marginal transformations. This approach is called the nonparanormal SKEPTIC and achieves the optimal parametric rate of convergence in terms of both precision matrix estimation and graph recovery.

Let $r_i^{(k)}$ be the rank of $Y_i^{(k)}$ among $Y_i^{(1)}, \ldots, Y_i^{(n)}$ and $\bar{r}_i = \frac{1}{n}\sum_{k=1}^n r_i^{(k)} = \frac{n+1}{2}$. The Spearman's rho statistics can be computed as follows

$$\hat{\rho}_{ij}^{\rho} = \frac{\sum_{k=1}^n (r_i^{(k)} - \bar{r}_i)(r_j^{(k)} - \bar{r}_j)}{\sqrt{\sum_{k=1}^n (r_i^{(k)} - \bar{r}_i)^2 \sum_{k=1}^n (r_j^{(k)} - \bar{r}_j)^2}}. \tag{3.1.8}$$

The nonparanormal model implies that $(f_i(Y_i), f_j(Y_j))$ follows a bivariate normal distribution with correlation parameter ρ_{ij}^{npn}. A classical result due to Kendall and Gibbons (1990) and Kruskal (1958) shows that $\rho_{ij}^{npn} = 2\sin\left(\frac{\pi}{6}\rho_{ik}^{\rho}\right)$. Henceforth, the correlation matrix of the nonparanormal model can be alternatively computed as follows:

$$\hat{\rho}_{ij}^{S} = \begin{cases} 2\sin(\frac{\pi}{6}\hat{\rho}_{ij}^{\rho}) & \text{for} \quad i \neq j \\ 1 & \text{for} \quad i = j \end{cases} \tag{3.1.9}$$

Liu et al. (2012) show that when the data contamination is low, the nonparanormal estimator is slightly more efficient than the nonparonormal SKEPTIC. But when the contamination increases the later siginificantly outperforms the normal-score based estimator proposed by Liu et al. (2009).

The main drawback of the univariate outlier insensitive transformations is their lack of robustness against structural outliers (see Alqallaf et al., 2009). This type of outliers can only be handled via robust affine equivariant methods. In the next section we propose an alternative robust pairwise correlation coefficient estimator that apply robust affine equivariant methods to the bivariate data. This method applies a bivariate winsorization that shrinks observations to the border of a tolerance ellipse so that outlying observations are appropriately downweight to obtain a robust correlation coefficient estimate that allows for protection against structural bivariate outliers.

3.2 The Proposed Winsorized Correlation Matrix

In this section, we propose to estimate the precision matrix by computing an affine equivariant transformation to the bivariate data. This transformation takes into account the orientation of the bivariate data and allows for protection against structural bivariate outliers. Then, a pairwise correlation matrix is computed from the outlier free bivariate transformed data. The resulting correlation matrix is plugged into the ℓ_1 log-determinant divergence optimization problem defined in (3.1.4).

To obtain a correlation estimator that is robust against structural bivariate outliers we could apply affine equivariant bivariate M estimators (Maronna, 1976). However, in the high-dimensional setting we require fast robust correlation estimates. Following the idea of Khan et al. (2007), we estimate the robust correlation coefficients applying a bivariate winsorization to the bivariate data given an affine equivariant robust correlation coefficient. In order to compute a correlation matrix that is robust against bivariate outliers, we are going to use reweighted robust pairwise estimators of scatter, where the weights are computed by the Mahalanobis distance with respect to an affine equivariant robust correlation estimator.

Let the vector $\mathbf{x}_J = (X_i, X_j)^T$, for $i, j = 1, \ldots, p$, follow a bivariate Gaussian distribution with mean $\boldsymbol{\mu}_J = (\mu_i, \mu_j)$, covariance $\boldsymbol{\sigma}^2_J = (\sigma_i^2, \sigma_j^2)$ and correlation matrix $\boldsymbol{\Gamma}_J$. Let's compute the squared population Mahalanobis distance as follows

$$d_k^2 = \left(\frac{Y_i^{(k)} - \mu_i}{\sigma_i}, \frac{Y_j^{(k)} - \mu_j}{\sigma_j} \right) (\boldsymbol{\Gamma}_J)^{-1} \left(\frac{Y_i^{(k)} - \mu_i}{\sigma_i}, \frac{Y_j^{(k)} - \mu_j}{\sigma_j} \right)^T . \tag{3.2.1}$$

We define the following weights

$$w_k(d_k^2) = \begin{cases} \sqrt{c^2/d_k^2} & \text{if} \quad d_k^2 > c^2 \\ 1 & \text{if} \quad d_k^2 \leq c^2 \end{cases} \tag{3.2.2}$$

where c^2 is given by $Pr(\chi^2_2 > c^2) = \varepsilon$ and ε is the proportion of outliers we want to control assuming that the majority of the data follows a bivariate Gaussian distribution.

Assuming we observe the vector of bivariate observations $\mathbf{y}_J^{(k)} = \left(Y_i^{(k)}, Y_j^{(k)}\right)^T$ for $i, j = 1, \ldots, p$ and $k = 1, \ldots, n$, the following Proposition, due to Cerioli (2010), refers to the distribution of the Mahalanobis distance of the observations for which $w_k = 1$.

Proposition 4. *The distribution of $\boldsymbol{y}_J^{(k)}$ conditioned on $w_k = 1$ is a truncated bivariate Gaussian distribution with*

$$E(\boldsymbol{y}_J^{(k)}|w_k = 1) = \boldsymbol{\mu}_J \quad \textit{and} \quad Cor(\boldsymbol{y}_J^{(k)}|w_k = 1) = \kappa_\varepsilon^{-1}\boldsymbol{\Gamma}_J \tag{3.2.3}$$

where

$$\kappa_\varepsilon = \frac{1-\varepsilon}{P(\chi^2_2 > \chi^2_{2,1-\varepsilon})}. \tag{3.2.4}$$

If we denote $w_\varepsilon = \sum_{k=1}^n w_k$ and

$$\begin{aligned}
(\hat{\mu}_i^\varepsilon, \hat{\mu}_j^\varepsilon) &= \left(\frac{1}{w_\varepsilon}\sum_{k=1}^n w_k Y_i^{(k)}, \frac{1}{w_\varepsilon}\sum_{k=1}^n w_k Y_j^{(k)}\right) \\
(\hat{\sigma}_i^\varepsilon, \hat{\sigma}_j^\varepsilon) &= \left(\left(\frac{\kappa_\varepsilon}{w_\varepsilon - 1}\sum_{k=1}^n w_k (Y_i^{(k)} - \hat{\mu}_i^\varepsilon)^2\right)^{1/2}, \left(\frac{\kappa_\varepsilon}{w_\varepsilon - 1}\sum_{k=1}^n w_k (Y_j^{(k)} - \hat{\mu}_j^\varepsilon)^2\right)^{1/2}\right) \\
\hat{\rho}_{ij}^\varepsilon &= \frac{\kappa_\varepsilon}{w_\varepsilon - 1}\sum_{k=1}^n w_k \left(\frac{Y_i^{(k)} - \hat{\mu}_i^\varepsilon}{\hat{\sigma}_i^\varepsilon}\right)\left(\frac{Y_j^{(k)} - \hat{\mu}_j^\varepsilon}{\hat{\sigma}_j^\varepsilon}\right),
\end{aligned} \tag{3.2.5}$$

then $\hat{\boldsymbol{\Gamma}}_J^\varepsilon = (\hat{\rho}_{ij}^\varepsilon)$ and $w_\varepsilon/n = (1-\varepsilon) + O_p(1/\sqrt{n})$ and it follows that $E(\hat{\boldsymbol{\mu}}_J^\varepsilon) \to \boldsymbol{\mu}_J$ and $E(\hat{\boldsymbol{\Gamma}}_J^\varepsilon) \to \boldsymbol{\Gamma}_J$.

A direct result from Proposition 4 is that we can obtain consistent estimators of $\boldsymbol{\mu}_J$ and $\boldsymbol{\Gamma}_J$ applying a bivariate winsorization to the observations of $\mathbf{y}_J^{(k)}$. To obtain robust estimates against two-dimensional structural outliers we propose to estimate the Mahalanobis distance using some affine equivariant robust correlation coefficient. To do that, we can define n bivariate standardized observations $\left(\frac{Y_i^{(k)}-\hat{\mu}_i^0}{\hat{\sigma}_i^0}, \frac{Y_j^{(k)}-\hat{\mu}_j^0}{\hat{\sigma}_j^0}\right)$ where $\hat{\mu}_i^0$ is a robust scale estimate and $\hat{\sigma}_i^0$ is a robust location estimate. Now let $\hat{\mathbf{\Gamma}}_J^0 = (\rho_{ij}^0)$ be a robust and affine equivariant correlation estimator of the correlation matrix $\mathbf{\Gamma}_J$. We will use $\hat{\mathbf{\Gamma}}_J^0$ as a diagnostic tool to identify two-dimensional structural outlying observations. If the initial robust estimator reflects the bulk of data, then the outlying observation will have a large Mahalanobis distance and the outlying observations could be downweighted in order to minimize their influence. We define the Mahalanobis distance estimate as follows:

$$d^2_{k,\hat{\mathbf{\Gamma}}_J^0} = \left(\frac{Y_i^{(k)}-\hat{\mu}_i^0}{\hat{\sigma}_i^0}, \frac{Y_j^{(k)}-\hat{\mu}_j^0}{\hat{\sigma}_j^0}\right) (\hat{\mathbf{\Gamma}}_J^0)^{-1} \left(\frac{Y_i^{(k)}-\hat{\mu}_i^0}{\hat{\sigma}_i^0}, \frac{Y_j^{(k)}-\hat{\mu}_j^0}{\hat{\sigma}_j^0}\right)^T. \tag{3.2.6}$$

We propose two estimators to compute the correlation matrix $\hat{\mathbf{\Gamma}}_J^0$ and to perform the bivariate winsorization. First, we apply the *Adjusted Winsorization* proposed by Khan et al. (2007). This approach takes into account the quadrants relative to the coordinatewise medians and considers two tuning constants to perform univariate winsorization of the data. A larger tuning constant c_1 is used to winsorize the points lying in the two diagonally oppose quadrants that contains most of the standardize data. A smaller tuning constant c_2 is used to winsorize the remaining data. We set $c_1 = 2$ and $c_2 = \sqrt{h}c_1$ where $h = n_2/n_1$, n_1 is the number of observations in the major quadrants and $n_2 = n - n_1$. The adjusted winsorization is then defined as (see Khan, 2006)

$$\Psi(Y_i, Y_j) = \begin{cases} \left(\psi_{c_1}\left(\frac{Y_i-\hat{\mu}_i^0}{\hat{\sigma}_i^0}\right), \psi_{c_1}\left(\frac{Y_j-\hat{\mu}_i^0}{\hat{\sigma}_i^0}\right)\right) & \text{if} \quad \left(\frac{Y_i-\hat{\mu}_i^0}{\hat{\sigma}_i^0}\right)\left(\frac{Y_j-\hat{\mu}_j^0}{\hat{\sigma}_j^0}\right) \geq 0 \\ \left(\psi_{c_2}\left(\frac{Y_i-\hat{\mu}_i^0}{\hat{\sigma}_i^0}\right), \psi_{c_2}\left(\frac{Y_j-\hat{\mu}_i^0}{\hat{\sigma}_i^0}\right)\right) & \text{if} \quad \left(\frac{Y_i-\hat{\mu}_i^0}{\hat{\sigma}_i^0}\right)\left(\frac{Y_j-\hat{\mu}_j^0}{\hat{\sigma}_j^0}\right) < 0, \end{cases} \tag{3.2.7}$$

where $\psi_c(y) = \min\{\max\{-c, y\}, c\}$ is a non-decreasing symmetric function and c_1 and c_2 are previous constants. Then, the correlation coefficient estimator $\hat{\rho}_J^0$ is obtained by computing the Pearson correlation on the adjusted winsorized data. In the second alternative, we compute $\hat{\mathbf{\Gamma}}_J^0$ using the Spearman's rho as in equation (3.1.9). This approach is denoted by *Spearman's Winsorization*.

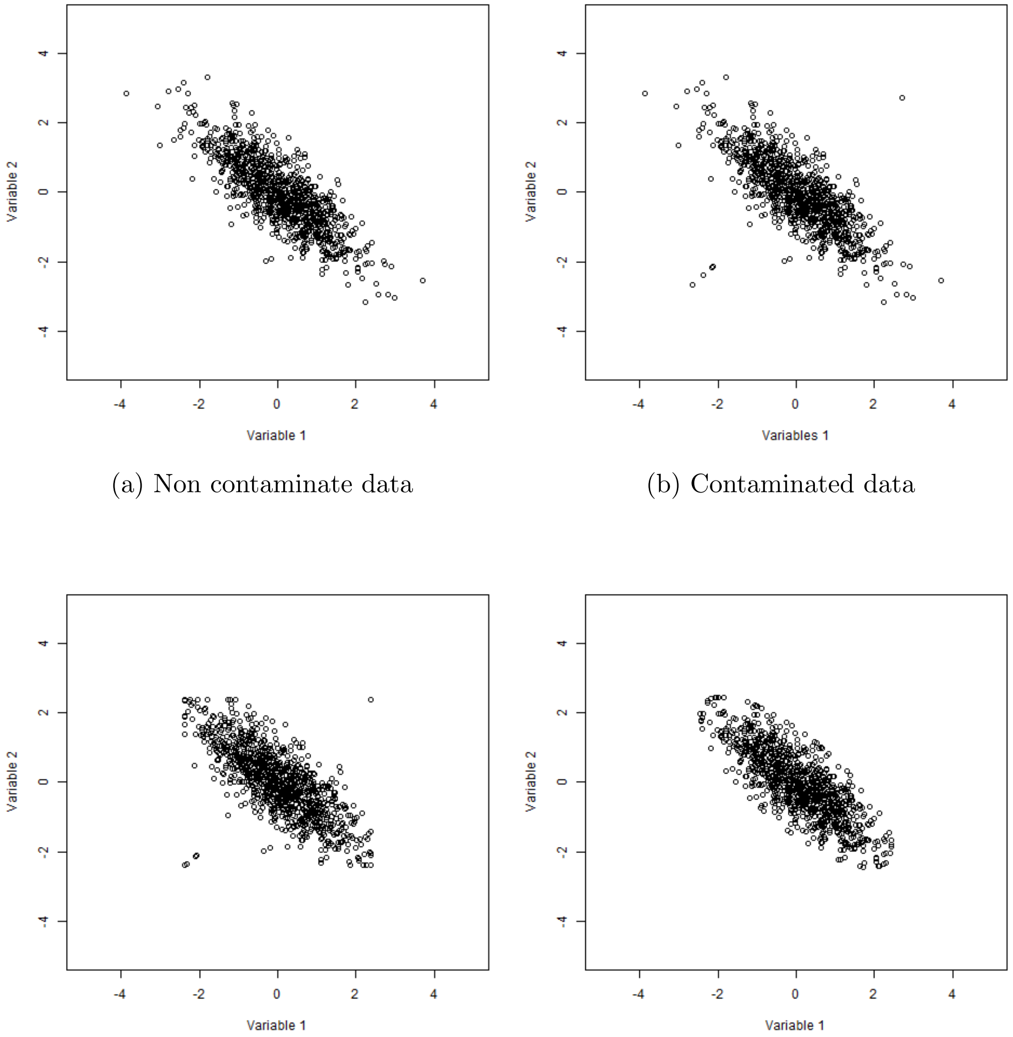
Therefore, given an affine equivariant robust correlation estimator $\hat{\mathbf{\Gamma}}_J^0$ (i.e. Adjusted Winsorized correlation coefficient or Spearman's rho), we estimate the robust Mahalanobis distance as in equation (3.2.6), then the outlier-free bivariate transformed data is computed as follows

$$\Psi_W(Y_i^{(k)}) = \begin{cases} \sqrt{c^2/d^2_{k,\hat{\mathbf{\Gamma}}_J^0}} \left(\frac{Y_i^{(k)} - \hat{\mu}_i^0}{\hat{\sigma}_i^0}\right) & \text{if} \quad d^2_{k,\hat{\mathbf{\Gamma}}_J^0} > c^2 \\ \frac{Y_i^{(k)} - \hat{\mu}_i^0}{\hat{\sigma}_i^0} & \text{if} \quad d^2_{k,\hat{\mathbf{\Gamma}}_J^0} \leq c^2, \end{cases} \tag{3.2.8}$$

where c^2 is given by $P(\chi^2_2 > c^2) = \varepsilon$ and ε is the proportion of outliers we want to control assuming that the majority of the data follows a bivariate Gaussian distribution.

Given the observations $(Y_1^{(k)}, \ldots, Y_p^{(k)})^T$, the winsorized correlation matrix $\hat{\mathbf{\Gamma}}^W = (\hat{\rho}_{ij}^W)$ is obtained by computing the Pearson correlation coefficient with respect to the bivariate winsorized data. The robust precision matrix is estimated by plugging the winsorized correlation matrix $\mathbf{\Gamma}^W$ into the ℓ_1 log-determinant divergence (3.1.4).

To show how the bivariate winsorization works under cellwise contamination, we simulate data from a bivariate Gaussian distribution where the correlation is set equal to -0.8. We select $n = 1000$ and generate 5 structural bivariate outliers. Figure 3.1, Panel (a) and (b), shows the scatter plot of the uncontaminated and contaminated simulated data, respectively. Figure 3.1, Panel (c), shows the scatter plot when we apply the non-paranormal transformation (see Liu et al., 2009). The non-paranormal transformation shrinks the correlation outliers to the boundary of a square. However, it does not take into account the orientation of the data and the effect of the structural outliers is not significantly downweighted. In Figure 3.1, Panel (d), we observe that



(a) Non contaminate data

(b) Contaminated data

(c) Nonparanormal transformation

(d) Bivariate Winsorization

Figure 3.1: Illustration of nonparanormal tranformation and bivariate winsorization under bivariate contamination. The nonparanormal transformation does not take into account the orientation of the bivariate data. The bivariate data winsorization shrinks points outside an ellipse that connects points of equal Mahalanobis distance (2.45) towards its boundary.

the bivariate transformation shrinks the outliers to the boundary of an ellipse of equal Mahalanobis distance. Henceforth, the influence of the bivariate outliers, when we compute the robust correlation coefficient, is appropriately downweighted.

Table 3.1: Estimation performance of the bivariate winsorized correlation coefficient under non contamination over 100 replications with standard deviations in brackets.

n	Sample Correlation	Adjusted Winsorization	Spearman Winsorization
20	-0.795(0.078)	-0.784(0.083)	-0.783(0.082)
30	-0.790(0.069)	-0.789(0.069)	-0.788(0.069)
50	-0.791(0.060)	-0.790(0.062)	-0.789(0.063)
100	-0.799(0.037)	-0.797(0.038)	-0.797(0.039)
1000	-0.800(0.012)	-0.800(0.011)	-0.799(0.012)
10,000	-0.801(0.004)	-0.801(0.004)	-0.801(0.004)

We also study the intrinsic bias of $\hat{\rho}_{ij}^{W}$ when there is no contamination (i.e. $\varepsilon = 0$) and assuming we observe the vector $\mathbf{x}$ that is Gaussian distributed with mean $\boldsymbol{\mu}$ and correlation matrix $\boldsymbol{\Gamma}$. The intrinsic bias of the bivariate winsorized estimator occurs due to the data transformation. To compare the bivariate winsorized correlation coefficient and the true correlation coefficient empirically, we generate random samples of sizes: 20, 30, 50, 1000, 10,000 from a bivariate normal distribution with mean zero and correlation coefficient equal to -0.8. We compare the empirical correlation coefficient denoted by "Sample Correlation", the bivariate winzorized correlation coefficient based on adjusted winsorization (i.e. "Adjusted Winsorization") and the bivariate winzorized correlation coefficient based on Spearman's correlation (i.e. "Spearman Winsorization"). We observe from Table 3.1 that the intrinsic bias diminishes as n increases. For small sample sizes the bivariate winsorized correlation estimator tends to underestimate the true correlation coefficient. When we compare the "Adjusted Winsorization" estimator with the "Spearman Winsorization", we observe that the later slightly underestimates the true correlation coefficient for small sample sizes. This is due to the fact that the Spearman's rho is computed using univariate rank transformation, while the adjusted winsorization operates directly on the data.

In terms of computational time, the correlation coefficient based on bivariate winsorization has a computational complexity of $O(n\log(n))$ (see Khan et al., 2007). The log-determinant divergence can be computed with the Graphical Lasso algo-

Table 3.2: Average computing times (in seconds) over 100 replications with standard deviations in brackets.

	$p = 90$	$p = 200$
Adjusted Winsorization	1.290(0.044)	6.462(0.098)
Spearman Winsorization	1.040(0.027)	4.604(0.079)
Sample Correlation	0.029(0.007)	0.191(0.012)
npn	0.210(0.009)	0.445(0.016)
npn-SKEPTIC	0.208(0.009)	0.453(0.014)

rithm proposed by Friedman et al. (2008). The fast implementation of the Graphical Lasso algorithm makes use of the block diagonal structure in the graphical lasso solution, the computational complexity is $O(p^2 + (p-q)^3)$, where q is the number of fully unconnected nodes (see Witten et al., 2011; Mazumder and Hastie, 2012). We conduct a numerical experiment to compare the computational time of the proposed bivariate winzorization plug-in estimators with the robust estimators based on univariate transformations. We simulate Gaussian data from an AR(1) model: $\mathbf{\Omega}_{ii} = 1$, $\mathbf{\Omega}_{i,i-1} = \mathbf{\Omega}_{i-1,i} = 0.4$, and zero otherwise. We chose a regularization parameter so that the solution contains roughly the actual number of non-zero elements in the true model. Table 3.2 shows the average computing times (in seconds) over 100 replications. We observe that estimating correlations coefficients via bivariate winsorization takes more time than computing the univariate winsorized normal-score nomparanormal transformation (i.e. "npn") from Liu et al. (2009) and the Spearman's rho non-paranormal SKEPTIC from Liu et al. (2012). However, the bivariate winsorization is much more accurate when there are structural bivariate outliers in the data. Therefore, the gain in robustness compensated the extra computing time.

In the next section we prove that the rate of convergence of the bivariate winsorized pairwise scatter estimate is the same as the affine equivariant robust correlation estimates used to compute the Mahalanobis distance (i.e. Adjsuted Winsorized correlation coefficient or Spearman's rho). This result suggests that if the initial robust correlation coefficient estimate converges to the true parameter at the optimal parametric rate, then the winsorized precision matrix achieves the optimal parametric rates of convergence in terms of both precision matrix estimation and graph recovery.

3.3 Analytical Properties

In this section we establish some analytical properties for the proposed bivariate winsorized correlation estimator. The main conclusion drawn from the theoretical results is that the location and scatter estimates computed from the bivariate winsorized data have the same rate of convergence as the affine equivariant robust location and pairwise scatter estimates used to compute the Mahalanobis distance.

Let $\mathbf{y}_J^{(1)}, \ldots, \mathbf{y}_J^{(n)}$ be independent bivariate random vectors that follow a distribution in an elliptical family with density

$$f(\mathbf{y}_J) = \det(\mathbf{\Gamma}_J)^{-1/2} h\left(\left(\frac{Y_i - \mu_i}{\sigma_i}, \frac{Y_j - \mu_j}{\sigma_j}\right)^T (\mathbf{\Gamma}_J)^{-1} \left(\frac{Y_i - \mu_i}{\sigma_i}, \frac{Y_j - \mu_j}{\sigma_j}\right)\right) \tag{3.3.1}$$

where $h : [0, \infty) \to [0, \infty)$ is assumed to be known. Under the assumption that the vector $\mathbf{y}_J = (Y_i, Y_i)^T$ is bivariate Gaussian distributed, the function h corresponds to $h(r) = (2\pi)e^{r/2}$. Moreover, we assume the following smoothness conditions on the function h:

(H1) h is continuous differentiable.

(H2) h has finite fourth moment: $\int (\mathbf{y}_J^T \mathbf{y}_J)^2 h(\mathbf{y}_J^T \mathbf{y}_J) d\mathbf{y}_J < \infty$.

Let $\hat{\boldsymbol{\theta}}^0 = (\hat{\mu}_i^0, \hat{\mu}_j^0, \hat{\sigma}_i^0, \hat{\sigma}_j^0, \hat{\rho}_{ij}^0)$ denote robust and affine equivariant estimators of location and scatter. We will use these estimates as diagnostic tool to identify structural bivariate outliers. Let $\hat{d}_k^2$ be the Mahalanobis distance computed as in (3.2.6). We apply the bivariate transformation in (3.2.8) and we compute the bivariate winsorized correlation estimator $\hat{\rho}_{ij}^W$.

Let $w : [0, \infty) \to [0, 1]$ be the function defined in (3.2.2), that satisfies the following condition

(W) w is bounded and of bounded variation and almost everywhere continuous on $[0, \infty)$.

We study the asymptotic behavior of $\hat{\rho}_{ij}^W$ as $n \to \infty$. Let $\boldsymbol{\theta}^* = (\mu_i, \mu_j, \sigma_i, \sigma_j, \rho_{ij})$ denoted the true vector of parameters. Assuming that the estimates $\hat{\boldsymbol{\theta}}^0$ are affine equivariant and consistent in probability (i.e. $\hat{\boldsymbol{\theta}}^0 \to \boldsymbol{\theta}^*$ in probability), the next Theorem analyzes the asymptotic properties of the bivariate winsorized correlation coefficient. The proof follows the analysis for reweighted estimators of multivariate location and scatter of Lopuhaä (1999).

Theorem 3. *Let $\boldsymbol{y}_J^{(1)}, \ldots, \boldsymbol{y}_J^{(n)}$ be independent bivariate random vectors with parameter vector $\boldsymbol{\theta}^* = (\mu_i, \mu_j, \sigma_i, \sigma_j, \rho_{ij})$ which are assumed to have density function defined in (3.3.1). Suppose that $w : [0, \infty) \to [0, 1]$ satisfies (W) and h satisfies (H1) and (H2). Let $\hat{\boldsymbol{\theta}}^0$ be affine equivariant and consistent estimate in probability of $\boldsymbol{\theta}^*$. Then,*

$$\hat{\rho}_{ij}^W - c_1\rho_{ij} = o_p(1/\sqrt{n}) + o_p(\hat{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^*) + \frac{1}{n}\sum_{k=1}^{n}\left\{ w(d_k^2)\left(\frac{Y_i^{(k)} - \mu_i}{\sigma_i}\right)\left(\frac{Y_j^{(k)} - \mu_j}{\sigma_j}\right) - c_1\rho_{ij}\right\}, \tag{3.3.2}$$

where the constant c_1 is given by

$$c_1 = \pi \int_o^\infty w(r^2)h(r^2)r^3 dr > 0. \tag{3.3.3}$$

Proof. Theorem 3 can be proved by adapting the proof in Lopuhaä (1999). The Mahalanobis distance can be written as a function of the vector $\boldsymbol{\theta}$. Thus, we define the following functions

$$\begin{aligned} \Psi_1(\mathbf{y}_J, \boldsymbol{\theta}) &= w\left(d^2(\boldsymbol{\theta})\right)\mathbf{y}_J \\ \Psi_2(\mathbf{y}_J, \boldsymbol{\theta}, \mathbf{t}) &= w\left(d^2(\boldsymbol{\theta})\right)(\mathbf{y}_J - \mathbf{t})(\mathbf{y}_J - \mathbf{t})^T. \end{aligned} \tag{3.3.4}$$

We define the bivariate adjusted winsorization estimates of location and covariance as follows

$$\begin{aligned}\hat{\boldsymbol{\mu}}_J^W &= \frac{1}{n}\sum_{k=1}^{n} w\left(\hat{d}_k^2\right)\mathbf{y}_J^{(k)} \\ \hat{\boldsymbol{\Sigma}}_J^W &= \frac{1}{n}\sum_{k=1}^{n} w\left(\hat{d}_k^2\right)(\mathbf{y}_J^{(k)}-\hat{\boldsymbol{\mu}}_J^W)(\mathbf{y}_J^{(k)}-\hat{\boldsymbol{\mu}}_J^W)^T.\end{aligned} \tag{3.3.5}$$

Then, $\hat{\boldsymbol{\mu}}_J^W$ and $\hat{\boldsymbol{\Sigma}}_J^W$ can be written as:

$$\begin{aligned}\hat{\boldsymbol{\mu}}_J^W &= \int \Psi_1(\mathbf{y}_J,\boldsymbol{\theta})dP_n(\mathbf{y}_J) \\ \hat{\sigma}_J^W &= \int \Psi_2(\mathbf{y}_J,\boldsymbol{\theta},\hat{\boldsymbol{\mu}}_J^W)dP_n(\mathbf{y}_J),\end{aligned} \tag{3.3.6}$$

where P_n denotes the empirical measure corresponding to $\mathbf{y}_J^{(1)},\ldots,\mathbf{y}_J^{(n)}$.

Moreover, estimates in (3.3.6) can be written as:

$$\begin{aligned}\int \Psi_1(\mathbf{y}_J,\hat{\boldsymbol{\theta}}^0) = \int \Psi_1(\mathbf{y}_J,\hat{\boldsymbol{\theta}}^0)dP(\mathbf{y}_J) + \int \Psi_1(\mathbf{y}_J,\boldsymbol{\theta}^*)d(P_n-P)(\mathbf{y}_J) \\ + \int \left(\Psi_1(\mathbf{y}_J,\hat{\boldsymbol{\theta}}^0)-\Psi_1(\mathbf{y}_J,\boldsymbol{\theta}^*)\right)d(P_n-P)(\mathbf{y}_J),\end{aligned} \tag{3.3.7}$$

Suppose that $\boldsymbol{\Sigma}_J = \mathbf{D}^2$ where $\mathbf{C}$ belongs to the class of positive definite symmetric matrices. Let $\hat{\boldsymbol{\mu}}_J^0 = (\hat{\mu}_i^0,\hat{\mu}_j^0)^T$ and $\hat{\boldsymbol{\Sigma}}_J^0 = \mathbf{C}_n^2$ be affine equivariant location and scatter estimates such that $(\hat{\boldsymbol{\mu}}_J^0-\boldsymbol{\mu}_J,\mathbf{C}_n-\mathbf{C})$ are consistent in probability. Then, using the result in Lopuhaä (1999) the first term in the right-hand side of (3.3.7) is $c_0(\hat{\boldsymbol{\mu}}_J^0-\boldsymbol{\mu}_J)+o_p(\hat{\boldsymbol{\mu}}_J^0-\boldsymbol{\mu}_J,B_n-B)$ and the third term is $o_p(1/\sqrt{n})$. The second term is equal to:

$$\int \Psi_1(\mathbf{y}_J,\boldsymbol{\theta}^*)d(P_n-P)(\mathbf{y}_J) = \frac{1}{n}\sum_{k=1}^{n} w(d_k^2)(\mathbf{y}_J^{(k)}-\boldsymbol{\mu}_J). \tag{3.3.8}$$

This proves the expansion for $\hat{\boldsymbol{\mu}}_J^W$:

$$\hat{\boldsymbol{\mu}}_J^W-\boldsymbol{\mu}_J = \frac{1}{n}\sum_{k=1}^{n} w(d_k^2)(\mathbf{y}_J^{(k)}-\boldsymbol{\mu}_J)+c_0(\hat{\boldsymbol{\mu}}_J^0-\boldsymbol{\mu}_J)+o_p(1/\sqrt{n})+o_p(\hat{\boldsymbol{\mu}}_J^0-\boldsymbol{\mu}_J,\hat{\boldsymbol{\Sigma}}_J^0-\sigma_J) \tag{3.3.9}$$

the constants are given by

$$c_0 = 2\pi \int_o^\infty w(r^2)[h(r^2) + h'(r^2)r^2]rdr \tag{3.3.10}$$

$$c_1 = \pi \int_o^\infty w(r^2)h(r^2)r^3dr > 0. \tag{3.3.11}$$

In a similar way, using that the expansion of $\hat{\boldsymbol{\mu}}_J^W$ implies that $\hat{\boldsymbol{\mu}}_J^W \to \boldsymbol{\mu}_J$, it can be shown that

$$\begin{aligned}\int \Psi_2(\mathbf{y}_J, \hat{\boldsymbol{\theta}}^0, \hat{\boldsymbol{\mu}}_J^W) = c_1\boldsymbol{\Sigma}_J + c_2\{\mathrm{tr}(B^{-1}(B_n - B))\boldsymbol{\Sigma}_J + 2\mathbf{C}^{-1}(\mathbf{C}_n - \mathbf{C})\boldsymbol{\Sigma}_J\}\\ + \frac{1}{n}\sum_{k=1}^{n}\{w(d_k^2)(\mathbf{y}_J^{(k)} - \boldsymbol{\mu}_J)(\mathbf{y}_J^{(k)} - \boldsymbol{\mu}_J)^T - c_1\sigma_J\}\\ + o_p(1/\sqrt{n}) + o_p(\hat{\boldsymbol{\mu}}_J^0 - \boldsymbol{\mu}_J, \mathbf{C}_n - \mathbf{C}, \hat{\boldsymbol{\mu}}_J^W - \boldsymbol{\mu}_J),\end{aligned} \tag{3.3.12}$$

where $\mathbf{C}^{-1}(\mathbf{C}_n - \mathbf{C}) = (\mathbf{C}_n - \mathbf{C})\mathbf{C}^{-1} = \mathbf{A}_n$, $\mathbf{A}_n$ is $o_p(1)$ and the constant c_2 is given by

$$c_2 = \pi \int_o^\infty w(r^2)\left[r^2h(r^2) + \frac{r^4}{2}h'(r^2)\right]rdr. \tag{3.3.13}$$

Finally, let define the vector of standardized observations $\hat{\mathbf{y}}_J = \left(\frac{Y_i^{(k)} - \hat{\mu}_i^W}{\hat{\sigma}_i^W}, \frac{Y_j^{(k)} - \hat{\mu}_j^W}{\hat{\sigma}_j^W}\right)^T$ The bivariate winsorized correlation matrix can be define as:

$$\hat{\boldsymbol{\Gamma}}_J^W = \int \Psi_2(\hat{\mathbf{y}}_J, \boldsymbol{\theta})dP_n(\hat{\mathbf{y}}_J). \tag{3.3.14}$$

Using the result in (3.3.12) we obtain (3.3.2). □

Theorem 3 shows that the bivariate winsorized correlation estimate of ρ_{ij} works as well as the affine equivariant robust estimator $\hat{\rho}_{ij}^0$ used to identify structural bivariate outliers. Hence, if $\hat{\rho}_{ij}^0$ converges at a rate slower than $\sqrt{n}$, then the bivariate winsorized estimator $\hat{\rho}_{ij}^W$ converges to $c_1\rho_{ij}$ at the same rate.

We propose to use the correlation coefficient based on adjusted winsorization and the Spearmans' rho as diagnostic tool to estimate the Mahalanobis distance and obtain robustness against two-dimensional outliers. Khan (2006) shows that under certain regularity condition, the correlation coefficient based on adjusted winsorized data is consistent and asymptotically normal. Liu et al. (2012) and Xue et al. (2012) show that the Spearman's rank correlation estimate is consistent and converge to ρ_{ij} with the optimal parametric rate.

Regarding the precision matrix estimator, Ravikumar et al. (2008) and Ravikumar et al. (2011) study the sufficient condition on the estimated correlation matrix in order to achieve the optimal parametric rate in high-dimension. A sufficient condition to ensure consistency and graph recovery of the precision matrix estimator, at the minimax optimal rate, is given by the condition that the robust correlation matrix estimate $\hat{\mathbf{\Gamma}}$ converges to the true correlation matrix $\mathbf{\Gamma}$ at the optimal parametric rate (see Liu et al., 2012; Xue et al., 2012).

The following Lemma, adopted from Ravikumar et al. (2011), shows that if the bivariate winsorized correlation coefficient works as well as the usual sample correlation estimator based on uncontaminated data, then the bivariate winsorized correlation estimate achieves the optimal parametric rate.

Lemma 4. *Assume there exists a constant C such that the robust bivarite winsorized correlation coefficient estimator satisfies the following concentration bound*

$$Pr(|\hat{\rho}_{ij}^{W} - \rho_{ij}| > \epsilon) \leq 4exp(-Cn\epsilon^2) \tag{3.3.15}$$

for any $\epsilon \in (0, C^{-1/2})$.

Let denote by $d = \max_j \sum_{i \neq j} I_{\omega_{ij} \neq 0}$ to be the maximal degree over the underlying graph corresponding to $\mathbf{\Omega}$ and by $\mathcal{A}$ the support set of the off-diagonal elements in $\mathbf{\Omega}$. Moreover, we define by $K_{\mathbf{\Gamma}} = \parallel \mathbf{\Gamma} \parallel_\infty = \max_i \sum_j |\rho_{ij}|$ to be the matrix ℓ_∞ norm of the true correlation matrix $\mathbf{\Gamma}$, the matrix $\mathbf{H}_{\mathcal{A}\mathcal{A}} = [\mathbf{\Omega}^{-1} \otimes \mathbf{\Omega}^{-1}]_{\mathcal{A}\mathcal{A}}$ and the parameter $K_{\mathbf{H}} =\| \mathbf{H}^{-1}_{\mathcal{A}\mathcal{A}} \|_{\infty}$. The following Theorem shows that if we plug a robust estimate of the correlation matrix, that achieves the optimal parametric bound in (3.3.15), into the Graphical Lasso algorithm (Friedman et al., 2008), then the precision matrix estimate achieves the optimal rate of convergence in term of both precision matrix estimation and graph recovery.

Theorem 4. *If there exists a constant $\kappa \in (0,1)$ such that $\| \boldsymbol{H}_{\mathcal{A}^c\mathcal{A}}(\boldsymbol{H}_{\mathcal{A}\mathcal{A}})^{-1} \|_{\ell_\infty} < 1-\kappa$. Let $\hat{\mathbf{\Omega}}^W$ be the unique solution of the log-determinant program (3.1.4) with regularization parameter $\lambda_n = \frac{8}{\kappa}\sqrt{\frac{log4n}{Cp^\tau}}$ for some $\tau > 2$. Then, if the sample size is lower bounded as*

$$n > \frac{log\left(4/max\{C^{-1/2}, 6(1+8\kappa^{-1})d\ max\{K_{\mathbf{\Gamma}}K_H, K^3_{\mathbf{\Gamma}}K^3_H\}\}\right)}{Cp^{2\tau}}, \tag{3.3.16}$$

then with probability greater than $1 - 1/p^{\tau-2}$ we have that the estimated $\hat{\mathbf{\Omega}}^W$ satisfies the elementwise-ℓ_∞ bound:

$$\| \hat{\mathbf{\Omega}}^W - \mathbf{\Omega} \|_\infty \leq \{2(1+8\kappa^{-1})K_{\boldsymbol{H}}\}\sqrt{\frac{log4n}{Cp^\tau}}. \tag{3.3.17}$$

Moreover, the corresponding estimated edge set $\hat{E}$ is a subset of the true set of edges E and includes all edges (i,j) with $|\mathbf{\Omega}_{ij}| > \{2(1+8\kappa^{-1})K_{\boldsymbol{H}}\}\sqrt{\frac{log4n}{Cp^\tau}}$.

If we consider that $K_{\mathbf{\Gamma}}$, $K_{\mathbf{H}}$ and κ remain constant as a function of (n,p,d), we can obtain an asymptotic bound for the elementwise-ℓ_∞ norm $\| \hat{\mathbf{\Omega}}^W - \mathbf{\Omega} \|_\infty \leq O\left(\sqrt{\frac{\text{log4n}}{Cp^\tau}}\right)$. Assuming the concentration bound in Lemma 4, Theorem 4 can be prove by adapting the proof presented in Ravikumar et al. (2011).

From the theoretical results, we observe that if the affine equivariant robust correlation coefficient estimate $\hat{\rho}^0_{ij}$ converges to ρ_{ij} in probability at the optimal parametric rate, then the bivariate winsorized correlation coefficient $\hat{\rho}^W$ converges at the same rate as $\hat{\rho}^0$. Thus, if we plug the estimated correlation matrix $\hat{\mathbf{\Gamma}}^W$ into the parametric Graphical lasso, the robust precision matrix estimate based on bivariate winsorized data achieves the optimal minimax rate under the same conditions that when the data is not contaminated.

3.4 Empirical Performance in Simulated Data

In this section we analyze the empirical performance of the proposed methods through simulated data using different contamination mechanisms. We focus on the performance of the precision matrix estimators when we plug-in a robust correlation matrix onto the ℓ_1 log-determinant divergence function. To do that, we use the Graphical lasso algorithm proposed by Friedman et al. (2008) to solve the convex optimization problem in (3.1.4). In particular we consider the following correlation matrix estimates: "Adjusted Winsorization", for the pairwise correlation matrix estimator using bivariate winzorization when the correlation coefficient used to compute the Mahalanobis distance is estimated with the adjusted winsorized data. "Spearman Winsorization", for the pairwise correlation matrix estimator using bivariate winzorization when the Mahalanobis distance is computed using the Spearman's rho. "Sample Correlation", for the empirical correlation matrix. "npn" is the winsorized normal-score nonparanormal estimator from Liu et al. (2009). Finally, "npn-SKEPTIC" represents the non-paranormal SKEPTIC using Spearman's rho from Liu et al. (2012).

3.4.1 Simulation Framework

We present simulation experiments to examine the performance of the proposed methods to estimate the precision matrix under different contamination mechanisms. We consider two different specifications for the population precision matrix $\mathbf{\Omega}$:

1. AR(1) Model: $\mathbf{\Omega}_{ii} = 1$, $\mathbf{\Omega}_{i,i+1} = \mathbf{\Omega}_{i-1,i} = 0.4$ and 0 otherwise.

2. Erdös-Rényi random graph: $\mathbf{\Omega} = \mathbf{D}(\mathbf{M} + (|\lambda_{min}(\mathbf{M})| + 0.2)\mathbf{I}_p)\mathbf{D}$ where $\mathbf{M}$ is a zero diagonal matrix where $m_{ij} = 0.3m$, such that m is independently generated and Bernoulli distributed with probability 0.01 and $\lambda_{min}(\mathbf{M})$ is the minimum eigenvalue of matrix $\mathbf{M}$. $\mathbf{D}$ is a diagonal matrix with $d_{ii} = 1$ for $i = 1, \ldots, p/2$ and $d_{ii} = 3$ for $i = p/2 + 1, \ldots, p$. The matrix is standardized to have unit diagonals.

We assume that the random vector $\mathbf{x} = (X_1, \ldots, X_p)^T$ is Gaussian distributed with mean zero and covariance matrix $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$. We study the performance of the precision matrix estimator under the fully independent contamination model:

$$\mathbf{y} = (\mathbf{I} - \mathbf{B})\mathbf{x} + \mathbf{B}\mathbf{z} \tag{3.4.1}$$

assuming that the variables $B_1, \ldots, B_p$ are independent:

$$P(B_1 = 1) = \ldots = P(B_p = 1) = \varepsilon \tag{3.4.2}$$

We follow Öllerer and Croux (2015) and we study two contamination mechanisms. In the first contamination mechanism we assume that $\mathbf{z}$ is multivariate Gaussian distributed with mean $\mu_i^z = 10$ for $i = 1, \ldots, p$ and covariance matrix $\mathbf{\Sigma}^z = \mathbf{\Omega}^{-1}$. In the second contamination mechanism we assume that $\mathbf{z}$ is multivariate Gaussian distributed with mean $\mu_i^z = 10$ for $i = 1, \ldots, p$ and covariance matrix $\mathbf{\Sigma}^z = 0.2I_p$. We robust standardized the data using the median as a robust location estimator and the median absolute deviation as a robust scale measure. We set the sample size $n = 100$ and the dimension $p = \{90, 200\}$. We select three values for the probability that a variable is contaminated in model (3.4.1): $\varepsilon = \{0, 0.05, 0.1\}$. We generate 100 replicates for each simulation experiment.

To evaluate the performance of the proposed methods we study specific assessment measures to evaluate numerical performance and support recovery. To compare the numerical performance, we compute the Mean Squared Error (MSE) between $\boldsymbol{\Omega}$ and $\hat{\boldsymbol{\Omega}}$, given by the expectation of the squared of the Frobenius norm:

$$\text{MSE}(\hat{\boldsymbol{\Omega}}) = E(\| \hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega} \|_F^2). \tag{3.4.3}$$

Moreover, we evaluate the performance of the estimator $\hat{\boldsymbol{\Omega}}$ with the expected value of the Likelihood Ratio Test (LRT), measured by E(LRT($\hat{\boldsymbol{\Omega}}$)), where LRT($\hat{\boldsymbol{\Omega}}$) is the likelihood ratio distance computed as

$$\text{LRT}(\hat{\boldsymbol{\Omega}}) = \text{tr}(\hat{\boldsymbol{\Omega}}(\boldsymbol{\Omega})^{-1}) - \log(\det(\hat{\boldsymbol{\Omega}}(\boldsymbol{\Omega})^{-1})) - p. \tag{3.4.4}$$

Small values of either the MSE and LRT imply a better performance of the method in estimating the true precision matrix (see Danilov et al., 2012).

To study the support recovery we use specificity, sensitivity, and Mathews correlation coefficient (MCC) criteria. Let TP be the true non-zero elements and TN be the true zero elements estimated by $\hat{\boldsymbol{\Omega}}$. Let FP be the false non-zero elements and FN be the false zero elements estimated by $\hat{\boldsymbol{\Omega}}$. The classification performance measures are then defined as follows:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad \text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.4.5}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{3.4.6}$$

To select the optimal tuning parameter λ^* in the log-determinant divergence problem, we choose the Bayesian Information Criteria (BIC):

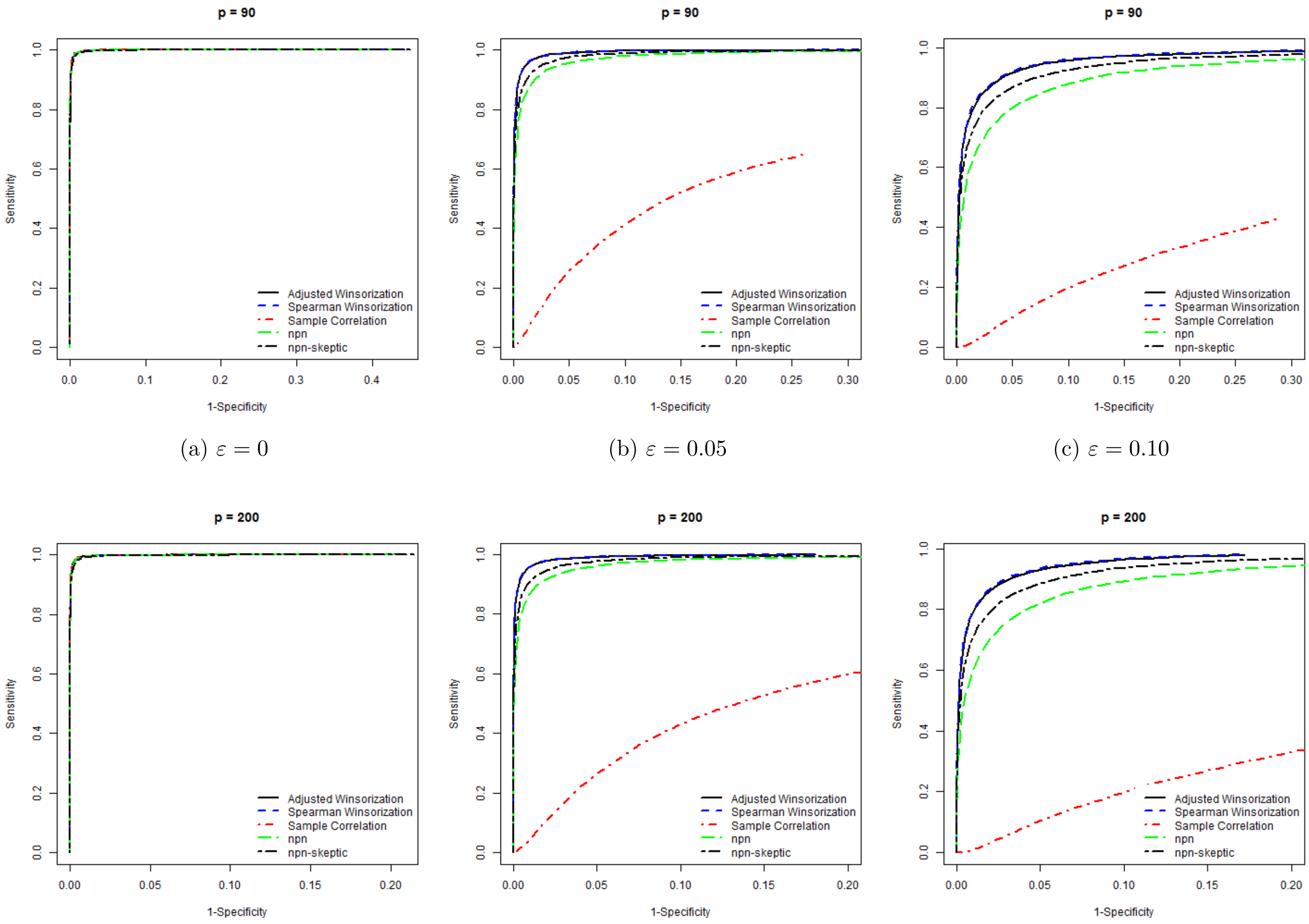
$$\lambda^* = \underset{\lambda > 0}{\text{argmin}} \left\{ -\log(\det(\hat{\boldsymbol{\Omega}})) + \text{tr}(\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\Gamma}}) + h\frac{\log(n)}{2n} \right\} \tag{3.4.7}$$

where h is the number of non-zero off-diagonal elements in $\hat{\mathbf{\Omega}}$, and $\hat{\mathbf{\Gamma}}$ the robust correlation estimate. The BIC has shown to have satisfactory performance for selecting the regularization parameter and for estimating the precision matrix (see Wang et al., 2007; Chen and Chen, 2008).

3.4.2 Simulation Results

We present detailed analysis based on numerical simulations under the first contamination mechanism for the two proposed specifications of $\mathbf{\Omega}$.

Regarding the support recovery under the first contamination mechanism, Panel (a) of Figures 3.2 and 3.3 illustrate the overall performance of different plug-in correlation estimates to robustly estimate the precision matrix under the first contamination mechanism for the full path of regularization parameters. For clean data, when the probability that a variable is contaminated is zero (i.e. $\varepsilon = 0$), the performance of the robust precision matrix estimates is similar to "Sample Correlation". Under contamination, the performance of the different estimates change. Panel (b) and Panel (c) of Figures 3.2 and 3.3 show that under cellwise contamination (i.e. $\varepsilon = 0.05$ and $\varepsilon = 0.10$), "Sample Correlation" becomes very sensitive to the presence of cellwise outliers. When $\varepsilon = 0.05$, we observe that the support recovery of "Adjusted Winsorization" and "Spearman Winsorization" performs slightly better than the robust estimates based on univariate outlier insensitive transformations. When $\varepsilon = 0.10$ the precision matrix estimates based on bivarite winsorization significantly outperform the non-paranormal SKEPTIC proposed by Liu et al. (2012) and the winsorized normal-score nonparanormal from Liu et al. (2009).



(a) $\varepsilon = 0$

(b) $\varepsilon = 0.05$

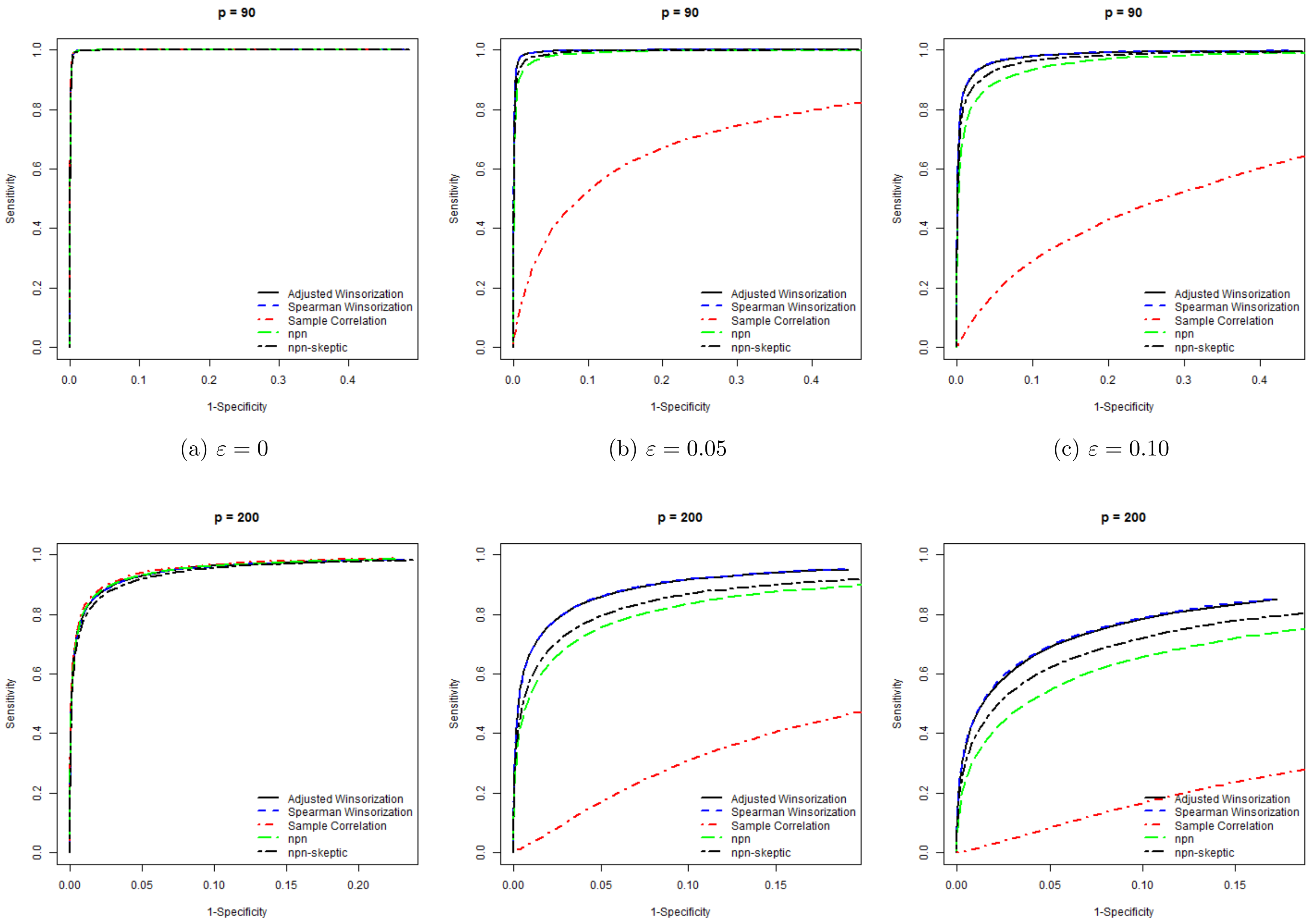
(c) $\varepsilon = 0.10$

(d) $\varepsilon = 0$

(e) $\varepsilon = 0.05$

(f) $\varepsilon = 0.10$

Figure 3.2: AR(1)-Model Specification. ROC curves under the first contamination mechanism over 100 replications.



(a) $\varepsilon = 0$

(b) $\varepsilon = 0.05$

(c) $\varepsilon = 0.10$

(d) $\varepsilon = 0$

(e) $\varepsilon = 0.05$

(f) $\varepsilon = 0.10$

Figure 3.3: Erdös-Rényi Specification. ROC curves under the first contamination mechanism over 100 replications.

Tables 3.3 and 3.4 show the results for the numerical performance for the optimal regularization parameter under the first contamination mechanism when the precision matrix is specified as in the AR(1) Model and Erdös-Rényi random graph, respectively. For clean data, the "Sample Correlation" sightly outperforms the robust plug-in estimators. The performance of the estimates based on bivariate winsorization is comparable with that of the empirical correlation matrix. Also, they slightly outperform the non-paranormal SKEPTIC and the winsorized normal-score nonparanormal estimator. When the probability that a variable contains outliers is positive, "Sample Correlation" performs very poorly in terms of efficiency on the precision matrix estimation. We observe that the robust estimators of the precision matrices have similar performance in terms of the expected likelihood ratio test and the mean squared error as the contamination increases. The similarity in their numerical performance is related with the fact that the BIC criteria selects models that contain a large number of false negatives.

Second contamination mechanism. Regarding the second contamination specification, Figure 3.4 and 3.5 illustrate the overall performance of the bivariate winsorized estimators to recover the true GGM for AR(1) Model and Erdös-Rényi random graph, respectively. We observe from the ROC curves that the robust precision matrix estimates behave in a similar way as the first contamination mechanism. When the probability that a variable contains outliers is low, the bivariate winzorized estimators perform slightly better than "npn" and "npn-SKEPTIC". When the probability is high, "Adjusted winsorization" and "Spearman Winsorization" significantly outperform the rank-based procedures. Tables 3.5 and 3.6 show the numerical performance under the second contamination mechanism. We observe that under cellwise contamination the precision matrix estimates based on bivariate winsorization exhibit satisfactory numerical performance.

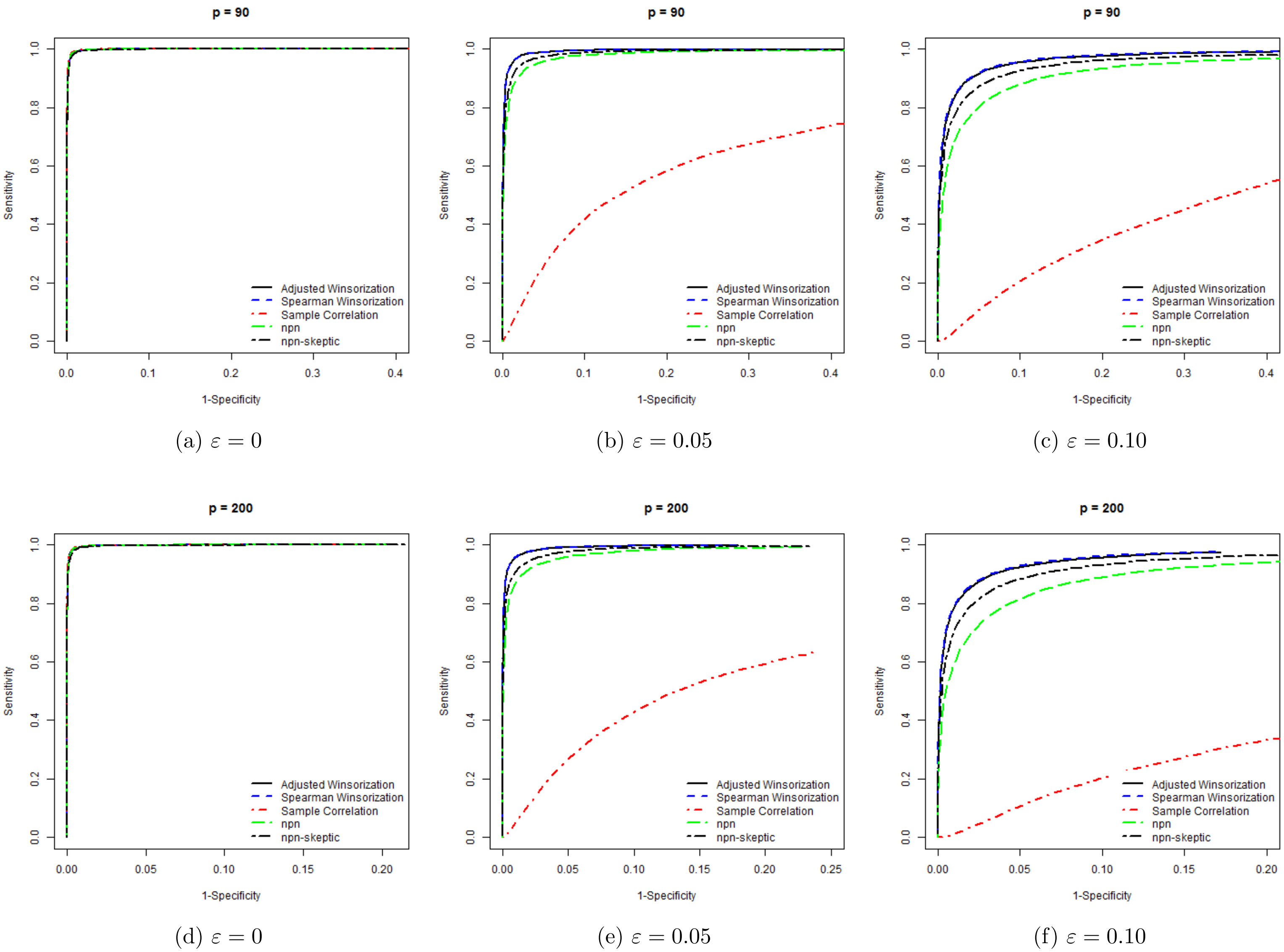
Table 3.3: AR(1)-Model Specification. Numerical performance under the first contamination mechanism over 100 replications with standard deviation in brackets.

		$\varepsilon = 0$		$\varepsilon = 0.05$		$\varepsilon = 0.10$	
	p	LRT	MSE	LRT	MSE	LRT	MSE
Spearman Winzorization	90	13.468	15.964	19.701	22.455	26.902	34.092
		(0.597)	(0.736)	(0.657)	(0.728)	(0.190)	(0.196)
	200	32.592	40.122	57.933	82.646	60.349	76.702
		(0.859)	(1.014)	(0.233)	(0.240)	(0.249)	(0.254)
Adjusted Winsorization	90	13.374	15.849	19.518	22.249	26.773	35.061
		(0.593)	(0.732)	(0.663)	(0.737)	(0.133)	(0.136)
	200	34.799	44.587	57.844	82.555	60.059	78.421
		(0.862)	(1.010)	(0.247)	(0.254)	(0.193)	(0.196)
Sample Correlation	90	12.446	13.980	27.646	34.239	27.668	34.269
		(0.558)	(0.689)	(0.057)	(0.047)	(0.003)	(0.018)
	200	32.348	39.813	60.731	79.059	61.431	77.764
		(0.855)	(1.014)	(0.047)	(0.030)	(0.024)	(0.009)
npn	90	13.784	16.363	25.734	36.320	26.587	34.873
		(0.586)	(0.707)	(0.174)	(0.179)	(0.178)	(0.185)
	200	33.369	41.086	57.883	82.594	59.479	79.909
		(0.875)	(1.028)	(0.241)	(0.248)	(0.166)	(0.171)
npn-SKEPTIC	90	13.566	16.093	25.467	37.259	26.041	35.457
		(0.621)	(0.757)	(0.160)	(0.165)	(0.210)	(0.218)
	200	35.219	45.080	57.251	84.174	58.483	81.026
		(0.853)	(0.997)	(0.261)	(0.268)	(0.212)	(0.219)

As a summary, simulation results show that bivariate winsorization have better support recovery performance in comparison with rank-based procedures. In general, both “Adjusted Winsorization” and “Spearman Winsorization” have satisfactory overall numerical performance properties. In terms of which method should be used, we observe that “Adjusted Winsorization” is slightly more efficient than “Spearman Winsorization” when the uncontaminated data is Gaussian distributed. This is due to the fact that the Spearman’s rho is computed using univariate rank transformations, while adjusted winsorization operates directly on the data.

Table 3.4: Erdös-Rényi Specification. Numerical performance under the first contamination mechanism over 100 replications with standard deviation in brackets.

		$\varepsilon = 0$		$\varepsilon = 0.05$		$\varepsilon = 0.10$	
	p	LRT	MSE	LRT	MSE	LRT	MSE
Spearman Winzorization	90	10.118	10.168	13.537	13.274	17.953	17.082
		(0.410)	(0.464)	(0.535)	(0.526)	(0.893)	(0.588)
	200	32.129	43.434	36.125	45.066	37.976	43.658
		(0.752)	(0.482)	(0.825)	(0.365)	(0.746)	(0.254)
Adjusted Winsorization	90	10.083	10.126	13.381	13.131	17.767	16.956
		(0.407)	(0.463)	(0.537)	(0.532)	(0.904)	(0.599)
	200	33.693	45.995	35.99	44.988	37.846	43.603
		(0.712)	(0.425)	(0.834)	(0.375)	(0.764)	(0.261)
Sample Correlation	90	10.049	10.093	22.758	22.771	23.213	22.234
		(0.400)	(0.455)	(0.311)	(0.135)	(0.105)	(0.038)
	200	32.073	43.405	39.995	49.502	39.996	46.808
		(0.746)	(0.492)	(0.132)	(0.035)	(0.030)	(0.010)
npn	90	10.273	10.360	16.016	16.690	20.239	20.525
		(0.412)	(0.456)	(0.633)	(0.509)	(0.757)	(0.436)
	200	35.589	48.757	37.265	46.883	38.834	46.321
		(0.667)	(0.353)	(0.702)	(0.299)	(0.533)	(0.184)
npn-SKEPTIC	90	10.863	11.661	15.281	16.770	19.267	20.637
		(0.455)	(0.482)	(0.585)	(0.493)	(0.800)	(0.499)
	200	35.283	48.508	36.977	48.104	38.317	47.387
		(0.691)	(0.370)	(0.697)	(0.314)	(0.648)	(0.229)



(a) $\varepsilon = 0$

(b) $\varepsilon = 0.05$

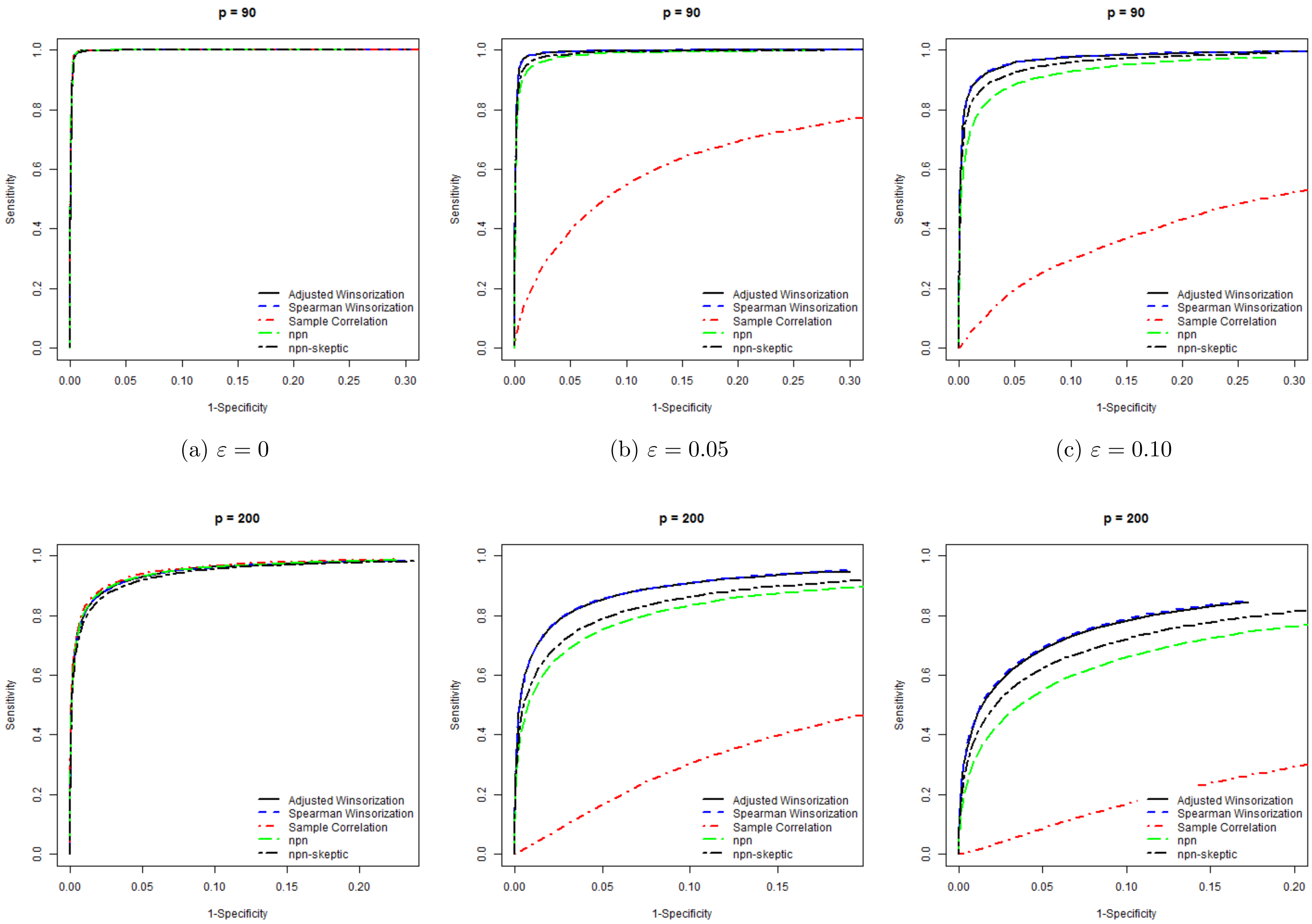
(c) $\varepsilon = 0.10$

(d) $\varepsilon = 0$

(e) $\varepsilon = 0.05$

(f) $\varepsilon = 0.10$

Figure 3.4: AR(1)-Model Specification. ROC curves under the second contamination mechanism over 100 replications.



(a) $\varepsilon = 0$

(b) $\varepsilon = 0.05$

(c) $\varepsilon = 0.10$

(d) $\varepsilon = 0$

(e) $\varepsilon = 0.05$

(f) $\varepsilon = 0.10$

Figure 3.5: Erdös-Rényi Specification. ROC curves under the second contamination mechanism over 100 replications.

Table 3.5: AR(1)-Model Specification. Numerical performance under the second contamination mechanism over 100 replications with standard deviation in brackets.

		$\varepsilon = 0$		$\varepsilon = 0.05$		$\varepsilon = 0.10$	
	p	LRT	MSE	LRT	MSE	LRT	MSE
Spearman Winzorization	90	13.468	15.964	19.702	22.484	26.889	34.078
		(0.597)	(0.736)	(0.711)	(0.783)	(0.198)	(0.203)
	200	32.592	40.122	57.987	82.701	60.470	76.823
		(0.859)	(1.014)	(0.192)	(0.196)	(0.220)	(0.222)
Adjusted Winsorization	90	13.374	15.849	19.523	22.283	26.759	35.047
		(0.593)	(0.732)	(0.715)	(0.791)	(0.140)	(0.143)
	200	34.799	44.587	57.904	82.616	60.153	78.516
		(0.862)	(1.010)	(0.200)	(0.205)	(0.161)	(0.162)
Sample Correlation	90	12.446	13.980	27.424	34.609	27.855	33.985
		(0.558)	(0.689)	(0.054)	(0.048)	(0.020)	(0.007)
	200	32.348	39.813	60.723	79.057	61.426	77.761
		(0.855)	(1.014)	(0.051)	(0.038)	(0.024)	(0.007)
npn	90	13.784	16.363	25.730	36.317	26.570	34.858
		(0.586)	(0.707)	(0.207)	(0.216)	(0.180)	(0.186)
	200	33.369	41.086	57.923	82.635	59.719	78.082
		(0.875)	(1.028)	(0.219)	(0.224)	(0.218)	(0.221)
npn-SKEPTIC	90	13.566	16.093	25.454	37.246	26.060	36.655
		(0.621)	(0.757)	(0.194)	(0.202)	(0.137)	(0.142)
	200	35.219	45.080	57.277	84.200	58.559	81.103
		(0.853)	(0.997)	(0.216)	(0.222)	(0.197)	(0.202)

Table 3.6: Erdös-Rényi Specification. Numerical performance under the second contamination mechanism over 100 replications with standard deviation in brackets.

		$\varepsilon = 0$		$\varepsilon = 0.05$		$\varepsilon = 0.10$	
	p	LRT	MSE	LRT	MSE	LRT	MSE
Spearman Winzorization	90	10.118	10.168	13.537	13.274	17.953	17.082
		(0.410)	(0.464)	(0.535)	(0.526)	(0.893)	(0.588)
	200	32.129	43.434	36.125	45.066	37.976	43.658
		(0.752)	(0.482)	(0.825)	(0.365)	(0.746)	(0.254)
Adjsuted Winsorization	90	10.083	10.126	13.381	13.131	17.767	16.956
		(0.407)	(0.463)	(0.537)	(0.532)	(0.904)	(0.599)
	200	33.693	45.995	35.99	44.988	37.846	43.603
		(0.712)	(0.425)	(0.834)	(0.375)	(0.764)	(0.261)
Sample Correlation	90	10.049	10.093	22.758	22.771	23.213	22.234
		(0.400)	(0.455)	(0.311)	(0.135)	(0.105)	(0.038)
	200	32.073	43.405	39.995	49.502	39.996	46.808
		(0.746)	(0.492)	(0.132)	(0.035)	(0.030)	(0.010)
npn	90	10.273	10.360	16.016	16.690	20.239	20.525
		(0.412)	(0.456)	(0.633)	(0.509)	(0.757)	(0.436)
	200	35.589	48.757	37.265	46.883	38.834	46.321
		(0.667)	(0.353)	(0.702)	(0.299)	(0.533)	(0.184)
npn-SKEPTIC	90	10.863	11.661	15.281	16.770	19.267	20.637
		(0.455)	(0.482)	(0.585)	(0.493)	(0.800)	(0.499)
	200	35.283	48.508	36.977	48.104	38.317	47.387
		(0.691)	(0.370)	(0.697)	(0.314)	(0.648)	(0.229)

3.5 Robust Cancer Classification based on Gene Expression Data

Microarrays experiments have being widely used to study the behavior of genes under various conditions. Microarrays raw data consist of image files and is subject to different preprocessing steps (Wu and Irizarry, 2007). First, probe intensities are adjusted for optical noise or nonspecific binding. Then, the data is adjusted to remove systematic bias due to different experimental designs. This task is often called *normalization*. As a result, gene expression data is often subject to numerous sources of experimental and preprocessing errors (Daye et al., 2012) and it may contain outliers. Moreover, the violation of the Gaussian assumption can lead to bias in the recovery of the true undirected graph and estimation of the precision matrix.

In this section we focus on the performance of robust precision matrices estimators for the classification of tumors using gene expression data. The different estimators are compared using two gene expression profile studies. For each study the data have being preprocessed, including image analysis of the microarray probe intensities, normalization and selection of differential expressed genes.

For an observed gene expression profile k we write the cellwise contamination model in the following form (see Alqallaf et al., 2002):

$$\mathbf{y}^{(k)} = (\mathbf{I} - \mathbf{B})\mathbf{X}^{(k)} + \mathbf{B}\mathbf{Z}^{(k)} \quad \text{for } k = 1, \ldots, n \tag{3.5.1}$$

where $\mathbf{y}^{(k)}$ denotes the observed gene expression vector of p genes in mRNA sample k. The unobservable random vector of gene expression levels $\mathbf{X}^{(k)}$ is assumed to be Gaussian distributed, $\mathbf{z}^{(k)} \in \mathbb{R}^p$ is an arbitrary random vector and $\mathbf{B}$ is the contamination indicator matrix where $P(B_1 = 1) = \ldots = P(B_1 = 1) = \varepsilon$ (i.e. the probability of an outlier occurring in the each gene is the same). The mRNA samples belong to

T known tumor classes, so a class label $t^{(k)} \in \{1, \dots, T\}$ can be predicted from the expression profiles $\mathbf{y}^{(k)} = (Y_i^{(k)}, \dots, Y_p^{(k)})^T$.

Based on the robust estimate of the precision matrix of the gene expression levels, we apply a linear discriminant analysis (LDA) to predict tumor classes. The different predictors are compared based on randomly splitting the data into training and testing sets. From the training set, we compute the robust center, scale and precision matrix estimates. For the test data we compute the linear discriminant score as follows

$$\delta_t(\mathbf{y}^{(k)}) = -\frac{1}{2}\log(\det(\hat{\mathbf{\Omega}})) - \frac{1}{2}d^2(\mathbf{y}^{(k)}, \hat{\boldsymbol{\mu}}_t, \hat{\mathbf{\Omega}}) + \log\hat{\pi}_t, \tag{3.5.2}$$

where $\hat{\pi}_t$ is the proportion of subjects in group t in the training set, $\hat{\boldsymbol{\mu}}_t$ the within class mean estimate, $\hat{\mathbf{\Omega}}$ the precision matrix estimate for the whole training set and $d^2(\cdot)$ is the squared Mahalanobis distance. The classification rule is

$$\hat{t}(\mathbf{y}^{(k)}) = \operatorname{argmax} \delta_t(\mathbf{y}^{(k)}) \quad \text{for } t = 1, \dots, T. \tag{3.5.3}$$

To perform model selection for λ we use 5-fold cross validation on the training data. Next, we analyze the performance of the bivariate winsorized precision matrix for the classification of tumors from gene expression datasets.

3.5.1 Analysis of Breast Cancer Data

We apply the procedure to evaluate gene expression profiling to breast cancer patients data to predict who may achieve pathological complete response (pCR). Using normalized gene expression data of patients in stages I-III of breast cancer data analyzed by Hess et al. (2006), we aim to predict response stated to neoadjuvant (preoperative) chemoterapy of patients with pathological complete response (pCR) and with residual disease (RD). The importance of analyzing the subject response to neoadjuvant (preoperative) chemoterapy, resides in the fact that complete eradication of all invasive cancer (i.e. pCR) is associated with long-term cancer free survival.

The data set consist of 22,283 gene expression levels of 133 subjects, with 34 pCR and 99 RD, respectively. We follow the analysis scheme proposed by Fan et al. (2009) and Cai et al. (2011). The data is randomly split into the training and testing set, and we repeat this procedure 100 times. The testing set is formed by randomly selecting 5 pCR subjects and 16 RD subjects (approximately 1/6 subjects in each group). The remaining subjects form the training set. From the training set, a Wilcox singed-rank test is performed to select the 113 most significant genes.

Table 3.7 displays the average classification performance and the number of miss-classified pCR subjects (Test Set Error) for each precision matrix estimator. We observe that "Sample Correlation" has the worst performance in predicting the pCR subjects in comparison with the robust precision matrix estimates. The overall classification performance measure by MCC criteria shows that "Adjusted Winsorization" outperforms the other procedures. From the results, we observe that the bivariate winsorized estimators improve over "npn" and "npn-SKEPTIC" in terms of the sensitivity and MCC, while all of them give similar specificity.

Table 3.7: Comparison of average pCR classification errors over 100 replications with standard deviation in brackets.

	Sensitivity	Specificity	MCC	Test Set Error	# of edges
Spearman Winzorization	0.558	0.816	0.366	0.246	2039.340
	(0.198)	(0.092)	(0.202)	(0.080)	(87.990)
Adjusted Winsorization	0.556	0.814	0.360	0.247	2006.820
	(0.196)	(0.085)	(0.189)	(0.073)	(90.722)
Sample Correlation	0.512	0.813	0.317	0.259	1891.240
	(0.215)	(0.089)	(0.222)	(0.080)	(90.703)
npn	0.540	0.816	0.345	0.250	2185.910
	(0.212)	(0.082)	(0.220)	(0.081)	(78.147)
npn-SKEPTIC	0.528	0.821	0.341	0.249	1978.700
	(0.214)	(0.086)	(0.225)	(0.081)	(76.069)

3.5.2 Analysis of Leukemia Data

The Leukemia dataset comes from a study of gene expression in two types of acute leukemia: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML), and was described by Golub et al. (1999). It has been shown that is critical for determining the chemotherapy regime to obtain discriminating tumor tissues between ALL and AML. Gene expression levels were measured using Affymetrix high-density oligonucleotide arrays. The raw data set consists of 6,817 gene expression levels of 38 bone marrow samples (27 ALL and 11 AML). The data was preprocessed and reduced to a subset of 3,051 with the most differential gene expression values.

The preprocessed data is randomly split into the training and testing set, and we repeat this procedure 100 times. The training set is formed by randomly selecting 25 cases and the testing set by randomly selecting 13 tissue samples. The training set is formed by 18 ALL samples and 7 AML samples. From the training set, a Wilcox singed-rank test is performed to select the 50 most significant genes.

Table 3.8 displays the average classification performance and the number of miss-classified tumor samples for each precision matrix estimator. The bivariate winsorized estimate based on adjusted winsorization has the better overall performance measure by MCC. We see that "Adjusted Winsorization" and "Spearman Winsorization" out-performs "npn" and "npn-SKEPTIC" in Sensitivity and MCC. In terms of Specificity all estimators have good performance in estimating false negatives. When we compare the rank-based procedures we observe that the winsorized normal-score nonparanormal estimator has better performance than the non-paranormal SKEPTIC estimator. This is due to the fact that when the contamination is low the "npn" is slightly more efficient than the nonparanormal SKEPTIC (see Liu et al., 2012).

Table 3.8: Comparison of average leukemia classification errors over 100 replications with standard deviation in brackets.

	Sensitivity	Specificity	MCC	Test Set Error	# of edges
Spearman Winzorization	0.870	0.959	0.841	0.063	380.410
	(0.195)	(0.070)	(0.191)	(0.074)	(29.026)
Adjusted Winsorization	0.903	0.956	0.860	0.057	382.290
	(0.179)	(0.071)	(0.174)	(0.069)	(31.672)
Sample Correlation	0.887	0.961	0.857	0.057	379.120
	(0.197)	(0.070)	(0.183)	(0.071)	(30.951)
npn	0.797	0.926	0.743	0.107	360.470
	(0.232)	(0.092)	(0.199)	(0.081)	(23.916)
npn-SKEPTIC	0.760	0.927	0.717	0.115	352.370
	(0.255)	(0.091)	(0.236)	(0.091)	(17.170)

3.6 Conclusions

In this article we have presented a method to robustly perform model selection in a Gaussian Graphical model when the data contain outliers. Several authors, including Liu et al. (2009) and Liu et al. (2012), have proposed robust estimators for the precision matrix in the high-dimensional setting. These methods are based on univariate outliers insensitive transformations to achieve normality. These transformations guarantee the protection against outlier propagation. However, they are not robust under the presence of structural bivariate outliers which may lead to misleading graph support recovery. Our approach is able to handle structural bivariate outliers while protecting against outlier propagation.

We estimate a high-dimensional and sparse robust precision matrix by plugging a robust correlation matrix estimate into a constraint ℓ_1 log-determinant divergence. We estimate the robust correlation matrix applying robust affine equivariant methods to the bivariate data and compute robust pairwise weighted correlation estimates, where the weights are computed by the Mahalanobis distance with respect to an affine equivariante robust correlation estimate. The proposed transformation applies a bivariate winsorization that shrinks observations to the border of a tolerance ellipse so that outlying observations are appropriately downweight to obtain a robust correlation estimate against two-dimensional structural outliers.

We analyze the analytic properties of the proposed bivariate winsorized pairwise scatter estimate and show that the rate of convergence is the same as the affine equivariant estimates used as a diagnostic tool to identify outlying observations. Furthermore, we show that if the initial robust affine equivariant correlation coefficient converges to the true correlation at the optimal parametric rate, then the bivariate winsorized precision matrix estimate achieves the optimal parametric rate in highdimensions.

Finally, we conducted extensive numerical simulations under different contamination settings to compare graph recovery performance of different robust estimators. We show that the proposed precision matrix estimate is robust against structural bivariate outliers and works well under the cellwise contamination model. The numerical simulations show that the bivariate winsorized transformation outperforms the existing rank-based methods when we aim to recover the support of $\mathbf{\Omega}$. Moreover, the proposed methods were then applied to the classification of tumors using gene expression data and we obtained satisfactory and promising prediction results.

There are several future directions of research. First, it would be interesting to derived specific concentration bounds for the Spearman's bivariate winsorization and the adjusted bivariate winsorization correlation coefficient. The performance of the bivariate winsorized estimate could also be studied under alternative precision matrix estimators such as CLIME (Cai et al., 2011), neighborhood selection with the lasso (Meinshausen and Bühlmann, 2006) and neighborhood Dantzig selector (Yuan, 2010). Also, we would like to establish the breakdown properties of the pairwise weighted correlation estimates under the cellwise contamination model. It would be important to determine the breakdown properties of the Graphical lasso when the bivariate winsorized correlation matrix is plugged into the ℓ_1 log-determinant divergence. Moreover, the proposed bivariate winsorized correlation coefficient could be used to perform robust correlation screening to deal with ultrahigh-dimensional data (see Li et al., 2012). Finally, it would be possible to study the bivariate outliers detection approach to estimate high-dimensional and sparse undirected graphs under more general elliptical distributions such as the multivariate $t-$distributions and nonparanormal models.

Chapter 4

Conclusion and Future Research Lines

The thesis considered the problem of estimating sparse Gaussian Graphical Models (GGMs) in the high-dimensional setting. Also, we study the problem of estimating robust precision matrices when the dataset may contain a fraction of outliers that are difficult to visualize and clean

In Chapter 2, we have introduced an approach to estimate undirected graphs and to perform model selection in high dimensional Gaussian Graphical Models. We consider a parametrization of the precision matrix in terms of the prediction errors of the best linear predictor of each node in the graph. We exploit the relationship between partial correlation coefficients and the distribution of the prediction errors. We propose a novel forward-backward algorithm for detecting pairs of variables having nonzero partial correlations among a large number of random variables based on i.i.d. samples. We obtain a set of the most probable edges in a GGM which are related with the largest absolute partial correlation coefficients. The position of the new non-zero element in the precision matrix corresponds to the pair of variables with the largest absolute partial correlation conditioned on the set of active nodes previously detected. We show that under mild conditions the Graphical Stepwise procedure is able to consistently estimated the set of true edges. The novelty of the approach is that we can obtain a set of more probable edges in a GGM for a given threshold value without resorting to penalized regression procedures. The Graphical Stepwise has good numerical and GGM classification performance when sparse precision matrices are estimated. Simulation studies show that the procedure is able to detect the true set of edges. The numerical examples indicate that our procedure outperforms existing algorithms, such as the Graphical lasso and CLIME. Applications to real data to perform a classification analysis show that our approach has a satisfactory predictive performance.

In Chapter 3, we have presented a method to robustly estimate a Gaussian Graphical model when the data contain outliers. Several authors, including Liu et al. (2009) and Liu et al. (2012), have proposed robust estimators for the precision matrix in the high-dimensional setting. These methods are based on univariate outliers insensitive transformations to achieve normality. These transformations guarantee the protection against outlier propagation. However, they are not robust under the presence of structural bivariate outliers which may lead to misleading graph support recovery. We estimated a high-dimensional and sparse robust precision matrix by plugging a robust correlation matrix estimate into a constraint ℓ_1 log-determinant divergence. We estimated the robust correlation matrix applying robust affine equivariant methods to the bivariate data and compute robust pairwise weighted correlation estimates, where the weights are computed by the Mahalanobis distance with respect to an affine equivariante robust correlation estimate. The proposed transformation applies a bivariate winsorization that shrinks observations to the border of a tolerance ellipse so that outlying observations are appropriately downweight to obtain a robust correlation estimate against two-dimensional structural outliers. We analyzed the analytic properties of the proposed bivariate winsorized pairwise scatter estimate and show that the rate of convergence is the same as the affine equivariant estimates used as a diagnostic tool to identify outlying observations. Furthermore, we show that if the initial robust affine equivariant correlation coefficient converges to the true correlation at the optimal parametric rate, then the bivariate winsorized precision matrix estimate achieves the optimal parametric rate in highdimensions. Finally, we conducted extensive numerical simulations under different contamination settings to compare graph recovery performance of different robust estimators. We showed that the proposed precision matrix estimate is robust against structural bivariate outliers and works well under the cellwise contamination model. The numerical simulations showed that the bivariate winsorized transformation outperforms the existing rank-based methods when we aim to recover the support of Ω. Moreover, the proposed methods were then applied to the classification of tumors using gene expression data and we obtained satisfactory and promising prediction results.

4.1 Future Research Lines

There are several possible extensions of the approaches presented in Chapter 2. In the forward-backward procedure we use a constant threshold value to select edges. A possible extension is to consider that the threshold varies with the effective sample size. This adjustment is able to improve the performance of the Graphical Stepwise procedure. Furthermore, we can apply multiple testing hypothesis procedure for selecting the optimal partial correlation coefficients. Liang et al. (2015) introduce a generalized Bayesian method for conducting multiple hypothesis testing that can be apply to test for conditional independence in our procedure. The procedure can be extended to binary graphical models by replacing linear regressions with logistic regression (see Ravikumar et al., 2010). For non-Gaussian random variables we could apply the non-paranormal transformation proposed by Liu et al. (2009) or the rank-based partial correlation coefficient proposed by Harris and Drton (2013). Finally, the procedure can be apply to covariate-adjusted precision matrix estimation (see Yin and Li (2011), Cai et al. (2012), Chen et al. (2016)). This model is apply when the graph structure among variables arises from both intrinsic connections and external effects, adjusting the effect of these external effects is of importance for understanding the underlying graph structure.

There are several future directions of research for robust estimation of Gaussian Graphical Models. First, it would be interesting to derived specific concentration bounds for the bivariate winsorization correlation coefficients. The performance of the bivariate winsorized estimate could also be studied under alternative precision matrix estimators such as CLIME (Cai et al., 2011), neighborhood selection with the lasso (Meinshausen and Bühlmann, 2006) and neighborhood Dantzig selector (Yuan, 2010). Also, we would like to establish the breakdown properties of the pairwise weighted correlation estimates under the cellwise contamination model. It would be important to determine the breakdown properties of the Graphical lasso when the bivariate winsorized correlation matrix is plugged into the ℓ_1 log-determinant divergence. Moreover, the proposed bivariate winsorized correlation coefficient could be used to perform robust correlation screening to deal with ultrahigh-dimensional data (see Li et al., 2012). Finally, it would be possible to study the bivariate outliers detection approach to estimate high-dimensional and sparse undirected graphs under more general elliptical distributions such as the multivariate $t-$distributions and nonparanormal models.

Bibliography

Alqallaf, F., S. V. Aelst, V. J. Yohai, and R. H. Zamar (2009). Propagation of outliers in multivariate data. *The Annals of Statistics 37*(1), 311–331.

Alqallaf, F. A., K. P. Konis, R. D. Martin, and R. H. Zamar (2002). Scalable robust covariance and correlation estimates for data mining. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 14–23. ACM.

Banerjee, O., L. El Ghaoui, and A. d'Aspremont (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research 9*, 485–516.

Bickel, P. J. and E. Levina (2008a). Covariance regularization by thresholding. *The Annals of Statistics 36*(6), 2577–2604.

Bickel, P. J. and E. Levina (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics 36*(1), 199–227.

Bühlmann, P., M. Kalisch, and M. H. Maathuis (2010). Variable selection in high-dimensional linear models: partially faithful distributions and the pc-simple algorithm. *Biometrika*, asq008.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media.

Cai, T., W. Liu, and X. Luo (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association 106*(494), 594–607.

Cai, T. T., H. Li, W. Liu, and J. Xie (2012). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika 100*(1), 139–156.

Cerioli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association 105*(489), 147–156.

Chen, J. and Z. Chen (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771.

Chen, M., Z. Ren, H. Zhao, and H. Zhou (2016). Asymptotically normal and efficient estimation of covariate-adjusted gaussian graphical model. *Journal of the American Statistical Association 111*(513), 394–406.

Danilov, M., V. J. Yohai, and R. H. Zamar (2012). Robust estimation of multivariate location and scatter in the presence of missing data. *Journal of the American Statistical Association 107*(499), 1178–1186.

Daye, Z. J., J. Chen, and H. Li (2012). High-dimensional heteroscedastic regression with an application to eQTL data analysis. *Biometrics 68*(1), 316–326.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, 157–175.

Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. Technical report, Technical report, Harvard University, Boston. URL http://www-stat. stanford. edu/˜ donoho/Reports/Oldies/BPMLE. pdf.

Edwards, D. (2000). *Introduction to Graphical Modelling*. Springer Science & Business Media.

El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics 36*(6), 2717–2756.

Fan, J., Y. Feng, and Y. Wu (2009). Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics 3*(2), 521–541.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911.

Fan, Y., J. Lv, et al. (2016). Innovated scalable efficient estimation in ultra-large gaussian graphical models. *The Annals of Statistics 44*(5), 2098–2126.

Finegold, M. and M. Drton (2011). Robust graphical modeling of gene networks using classical and alternative t-distributions. *The Annals of Applied Statistics*, 1057–1080.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*(3), 432–441.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science 286*(5439), 531–537.

Hampel, F. R. (1973). Robust estimation: A condensed partial survey. *Probability Theory and Related Fields 27*(2), 87–104.

Harris, N. and M. Drton (2013). Pc algorithm for nonparanormal graphical models. *Journal of Machine Learning Research 14*(1), 3365–3383.

Hess, K. R., K. Anderson, W. F. Symmans, V. Valero, N. Ibrahim, J. A. Mejia, D. Booser, R. L. Theriault, A. U. Buzdar, P. J. Dempsey, et al. (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of Clinical Oncology 24*(26), 4236–4244.

Huang, S., J. Jin, and Z. Yao (2016). Partial correlation screening for estimating large precision matrices, with applications to classification. *The Annals of Statistics 44*(5), 2018–2057.

Huber, P. J. (2011). *Robust Statistics.* Springer.

Huber, P. J. et al. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics 35*(1), 73–101.

Johnson, C. C., A. Jalali, and P. Ravikumar (2011). High-dimensional sparse inverse covariance estimation using greedy methods. *arXiv preprint arXiv:1112.6411*.

Kalisch, M. and P. Bühlmann (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *The Journal of Machine Learning Research 8*, 613–636.

Kendall, M. and J. Gibbons (1990). *Rank correlation methods.* A Charles Griffin Book. E. Arnold.

Khan, J. A., S. Van Aelst, and R. H. Zamar (2007). Robust linear model selection based on least angle regression. *Journal of the American Statistical Association 102*(480), 1289–1299.

Khan, M. J. A. (2006). *Robust Linear Model Selection for High-dimensional Datasets.* Ph. D. thesis, University of British Columbia.

Kiiveri, H. T. (2011). Multivariate analysis of microarray data: differential expression and differential connection. *BMC bioinformatics 12*(1), 42.

Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association 53*(284), 814–861.

Kurowicka, D. and R. Cooke (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling.* John Wiley.

Lam, C. and J. Fan (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics 37*(6B), 4254–4278.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford University Press.

Ledoit, O. and M. Wolf (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis 88*(2), 365–411.

Li, G., H. Peng, J. Zhang, and L. Zhu (2012). Robust rank correlation based screening. *The Annals of Statistics 40*(3), 1846–1877.

Li, H. and J. Gui (2006). Gradient directed regularization for sparse gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics 7*(2), 302–317.

Liang, F., Q. Song, and P. Qiu (2015). An equivalent measure of partial correlation coefficients for high-dimensional gaussian graphical models. *Journal of the American Statistical Association 110*(511), 1248–1265.

Liu, H., F. Han, M. Yuan, J. Lafferty, and L. Wasserman (2012). High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics 40*(4), 2293–2326.

Liu, H., J. Lafferty, and L. Wasserman (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research 10*(Oct), 2295–2328.

Liu, H. and L. Wang (2012). Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. *arXiv preprint arXiv:1209.2437*.

Liu, L., D. M. Hawkins, S. Ghosh, and S. S. Young (2003). Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences 100*(23), 13167–13172.

Loh, P.-L. and X. L. Tan (2015). High-dimensional robust precision matrix estimation: Cellwise corruption under epsilon-contamination. *arXiv preprint arXiv:1509.07229*.

Lopuhaä, H. P. (1999). Asymptotics of reweighted estimators of multivariate location and scatter. *Annals of Statistics 27*(5), 1638–1665.

Luo, S., R. Song, and D. Witten (2014). Sure screening for gaussian graphical models. *arXiv preprint arXiv:1407.7819*.

Maronna, R. A. (1976, 01). Robust m-estimators of multivariate location and scatter. *The Annals of Statistics 4*(1), 51–67.

Mazumder, R. and T. Hastie (2012). The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics 6*, 2125.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*(3), 1436–1462.

Öllerer, V. and C. Croux (2015). Robust high-dimensional precision matrix estimation. In *Modern Nonparametric, Robust and Multivariate Methods*, pp. 325–350. Springer.

Pearl, J. and A. Paz (1985). *Graphoids: A Graph-based Logic for Reasoning about Relevance Relations.* University of California (Los Angeles). Computer Science Department.

Peng, J., P. Wang, N. Zhou, and J. Zhu (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association 104*(486), 735–746.

Pourahmadi, M. (2007). Cholesky decompositions and estimation of a covariance matrix: orthogonality of variance–correlation parameters. *Biometrika 94*(4), 1006–1013.

Ravikumar, P., G. Raskutti, M. J. Wainwright, and B. Yu (2008). Model selection in gaussian graphical models: High-dimensional consistency of l1-regularized MLE. In *NIPS*, pp. 1329–1336.

Ravikumar, P., M. J. Wainwright, J. D. Lafferty, et al. (2010). High-dimensional ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics 38*(3), 1287–1319.

Ravikumar, P., M. J. Wainwright, G. Raskutti, B. Yu, et al. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics 5*, 935–980.

Ren, Z., T. Sun, C.-H. Zhang, H. H. Zhou, et al. (2015). Asymptotic normality and optimalities in estimation of large gaussian graphical models. *The Annals of Statistics 43*(3), 991–1026.

Rothman, A. J., P. J. Bickel, E. Levina, J. Zhu, et al. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics 2*, 494–515.

Rothman, A. J., E. Levina, and J. Zhu (2010). A new approach to cholesky-based covariance regularization in high dimensions. *Biometrika 97*(3), 539–550.

Rousseeuw, P. J. (1984). Least Median of Squares Regression. *Journal of the American statistical association 79*(388), 871–880.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications 8*, 283–297.

Rütimann, P., P. Bühlmann, et al. (2009). High dimensional sparse covariance estimation via directed acyclic graphs. *Electronic Journal of Statistics 3*, 1133–1160.

Speed, T. and H. Kiiveri (1986). Gaussian markov distributions over finite graphs. *The Annals of Statistics*, 138–150.

Spirtes, P., C. N. Glymour, and R. Scheines (2000). *Causation, Prediction, and Search.* MIT press.

Stahel, W. A. (1981). *Breakdown of Covariance Estimators.* Fachgruppe für Statistik, Eidgenössische Techn. Hochsch.

Sun, H. and H. Li (2012). Robust gaussian graphical modeling via l1 penalization. *Biometrics 68*(4), 1197–1206.

Sun, T. and C.-H. Zhang (2012). Scaled sparse linear regression. *Biometrika 99*(4), 879–898.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics 17*(6), 520–525.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics 33*(1), 1–67.

van de Geer, S. A. and P. Bühlmann (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics 3*, 1360–1392.

Wang, H., R. Li, and C.-L. Tsai (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika 94*(3), 553–568.

Whittaker, J. (2009). *Graphical Models in Applied Multivariate Statistics.* Wiley Publishing.

Wille, A., P. Zimmermann, E. Vranová, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelić, P. von Rohr, L. Thiele, et al. (2004). Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome biology 5*(11), R92.

Witten, D. M., J. H. Friedman, and N. Simon (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics 20*(4), 892–900.

Wu, Z. and R. A. Irizarry (2007). A statistical framework for the analysis of microarray probe-level data. *The Annals of Applied Statistics 1*(2), 333–357.

Xue, L., H. Zou, et al. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics 40*(5), 2541–2571.

Yin, J. and H. Li (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The annals of applied statistics 5*(4), 2630.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research 11*, 2261–2286.

Yuan, M. and Y. Lin (2007). Model selection and estimation in the gaussian graphical model. *Biometrika 94*(1), 19–35.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research 7*(Nov), 2541–2563.

Zhou, S., P. Rütimann, M. Xu, and P. Bühlmann (2011). High-dimensional covariance estimation based on gaussian graphical models. *The Journal of Machine Learning Research 12*, 2975–3026.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.