



UC3M Working Papers  
Statistics and Econometrics  
17-06  
ISSN 2387-0303  
May 2017

Departamento de Estadística  
Universidad Carlos III de Madrid  
Calle Madrid, 126  
28903 Getafe (Spain)  
Fax (34) 91 624-98-48

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

Ginette Lafit<sup>a</sup>, Francisco J. Nogales<sup>b</sup>

### Abstract

---

Robust estimation of Gaussian Graphical models in the high-dimensional setting is becoming increasingly important since large and real 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.

---

**Keywords:** Gaussian Graphical Models; Cellwise Contamination; Robust Correlation Estimation; Winsorization

---

<sup>a</sup> Department of Statistics, Universidad Carlos III de Madrid.

<sup>b</sup> Department of Statistics and UC3M-BS Institute of Financial Big Data, Universidad Carlos III de Madrid.

Acknowledgements: the authors acknowledge financial support from the Spanish Ministry of Education and Science, research project MTM2013-44902-P.

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

Ginette Lafit\*

Department of Statistics, Universidad Carlos III de Madrid

and

Francisco J. Nogales

Department of Statistics and

UC3M-BS Institute of Financial Big Data, Universidad Carlos III de Madrid

## Abstract

Robust estimation of Gaussian Graphical models in the high-dimensional setting is becoming increasingly important since large and real 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.

*Keywords:* Gaussian Graphical Models; Cellwise Contamination; Robust Correlation Estimation; Winsorization.

---

\*Ginette Lafit is Ph.D. student, Department of Statistics, Universidad Carlos III de Madrid (E-mail: glafit@est-econ.uc3m.es). Francisco J. Nogales is Associate Professor, Department of Statistics and UC3M-BS Institute of Financial Big Data, Universidad Carlos III de Madrid (E-mail: fcojavier.nogales@uc3m.es). The research of Ginette Lafit and Francisco J. Nogales was funded by the Spanish government project MTM2013-44902-P.

# 1 Introduction

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  $\mathbf{X} = (X_1, \dots, X_p)$ , 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  $V = \{1, \dots, p\}$  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  $\Omega = (\omega_{ij})$ , 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  $\Omega$  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  $\Omega$  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. 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  $\ell_1$  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 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  $\ell_1$  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  $\ell_1$  penalized maximum likelihood problem. Consistency of penalized likelihood procedures were also explored. Rothman et al. (2008) estimate convergence rates under the Frobenius 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 Se-

lector) 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  $\ell_1$ -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 article 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](#) we present the winsorized correlation matrix estimator, which is able to identify structural bivariate outliers under the cellwise contamination mechanism. In [Section 4](#) we present a theoretical analysis of the method. In [Section 5](#) we present numerical results on simulated data under different contamination mechanisms. [Section 6](#) 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.

## 2 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  $\Gamma = (\rho_{ij})$ . 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:

$$\mathbf{Y} = (I - B)\mathbf{X} + B\mathbf{Z} \quad (2.1)$$

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

$$B = \begin{bmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_p \end{bmatrix} \quad (2.2)$$

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  $P(B_1 = B_2 = \dots = B_p) = 1$ . 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  $\varepsilon$  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  $B_1, \dots, B_p$  are independent:

$$P(B_1 = 1) = \dots = P(B_p = 1) = \varepsilon \quad (2.3)$$

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  $\Omega$  can be obtained by plugging a robust correlation matrix estimator, denote by  $\hat{\Gamma}$ , into the following  $\ell_1$ -regularized log-determinant program (see [Öllerer and Croux, 2015](#); [Loh and Tan, 2015](#)):

$$\hat{\Omega} = \underset{\Omega \succ 0}{\operatorname{argmin}} \{ \operatorname{tr}(\Omega \hat{\Gamma}) - \log \det(\Omega) + \lambda \|\Omega\|_{1, \text{off}} \} \quad (2.4)$$

where  $\lambda > 0$  is the regularizing constant of the off-diagonal  $\ell_1$  regularizer

$$\|\Omega\|_{1, \text{off}} := \sum_{i \neq j} |\omega_{ij}| \quad \text{for } i, j = 1, \dots, p \quad (2.5)$$

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

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  $\varepsilon$  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  $\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(n)}$  be a sample of size  $n$  where  $\mathbf{Y}^{(k)} = (Y_1^{(k)}, \dots, Y_p^{(k)})^T \in \mathbb{R}^p$  for  $k = 1, \dots, n$ . 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  $(Y_i^{(1)}, \dots, Y_i^{(n)})$ . 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 non-paranormal 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  $\Gamma^{n \times n}$ .

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

$$T_{\delta_n}(y) = \begin{cases} \delta_n & \text{if } \hat{F}_i(y) < \delta_n \\ y & \text{if } \delta_n \leq \hat{F}_i(y) \leq (1 - \delta_n) \\ (1 - \delta_n) & \text{if } \hat{F}_i(y) > (1 - \delta_n), \end{cases} \quad (2.6)$$

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

$$\hat{\rho}_{ij}^{n_{pn}} = \frac{\frac{1}{n} \sum_{k=1}^n \hat{f}_i(Y_i^{(k)}) \hat{f}_j(Y_j^{(k)})}{\sqrt{\frac{1}{n} \sum_{k=1}^n \hat{f}_i^2(Y_i^{(k)})} \cdot \sqrt{\frac{1}{n} \sum_{k=1}^n \hat{f}_j^2(Y_j^{(k)})}}. \quad (2.7)$$

Then, the precision matrix nonparanormal estimator is computed by plugging  $\Gamma^{n_{pn}}$  into the  $\ell_1$  log-determinant program (2.4). 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 SKEP-TIC 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  $Y_i^{(1)}, \dots, Y_i^{(n)}$  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

$$\hat{\rho}_{ij}^\rho = \frac{\sum_{k=1}^n (r_i^{(k)} - \bar{r}_i)(r_j^{(k)} - \bar{r}_j)}{\sqrt{\sum_{k=1}^n (r_i^{(k)} - \bar{r}_i)^2 \sum_{k=1}^n (r_j^{(k)} - \bar{r}_j)^2}}. \quad (2.8)$$

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

$$\hat{\rho}_{ij}^S = \begin{cases} 2\sin(\frac{\pi}{6}\hat{\rho}_{ij}^\rho) & \text{for } i \neq j \\ 1 & \text{for } i = j \end{cases} \quad (2.9)$$

Liu et al. (2012) show that when the data contamination is low, the nonparanormal estimator is slightly more efficient than the nonparanormal SKEPTIC. But when the contamination increases the later significantly 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 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  $\ell_1$  log-determinant divergence optimization problem defined in (2.4).

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 com-

puted 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  $i, j = 1, \dots, p$ , follow a bivariate Gaussian distribution with mean  $\boldsymbol{\mu}_J = (\mu_i, \mu_j)$ , covariance  $\boldsymbol{\sigma}_J^2 = (\sigma_i^2, \sigma_j^2)$  and correlation matrix  $\Gamma_J$ . Let's compute the squared population Mahalanobis distance as follows

$$d_k^2 = \left( \frac{Y_i^{(k)} - \mu_i}{\sigma_i}, \frac{Y_j^{(k)} - \mu_j}{\sigma_j} \right) (\Gamma_J)^{-1} \left( \frac{Y_i^{(k)} - \mu_i}{\sigma_i}, \frac{Y_j^{(k)} - \mu_j}{\sigma_j} \right)^T. \quad (3.1)$$

We define the following weights

$$w_k(d_k^2) = \begin{cases} \sqrt{c^2/d_k^2} & \text{if } d_k^2 > c^2 \\ 1 & \text{if } d_k^2 \leq c^2 \end{cases} \quad (3.2)$$

where  $c^2$  is given by  $Pr(\chi_2^2 > c^2) = \varepsilon$  and  $\varepsilon$  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  $\mathbf{Y}_J^{(k)} = (Y_i^{(k)}, Y_j^{(k)})^T$  for  $i, j = 1, \dots, p$  and  $k = 1, \dots, n$ , the following Proposition, due to [Cerioli \(2010\)](#), refers to the distribution of the Mahalanobis distance of the observations for which  $w_k = 1$ .

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

$$E(\mathbf{Y}_J^{(k)} | w_k = 1) = \boldsymbol{\mu}_J \quad \text{and} \quad Cor(\mathbf{Y}_J^{(k)} | w_k = 1) = \kappa_\varepsilon^{-1} \Gamma_J \quad (3.3)$$

where

$$\kappa_\varepsilon = \frac{1 - \varepsilon}{P(\chi_2^2 > \chi_{2,1-\varepsilon}^2)}. \quad (3.4)$$

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

$$\begin{aligned}
(\hat{\mu}_i^\varepsilon, \hat{\mu}_j^\varepsilon) &= \left( \frac{1}{w_\varepsilon} \sum_{k=1}^n w_k Y_i^{(k)}, \frac{1}{w_\varepsilon} \sum_{k=1}^n w_k Y_j^{(k)} \right) \\
(\hat{\sigma}_i^\varepsilon, \hat{\sigma}_j^\varepsilon) &= \left( \left( \frac{\kappa_\varepsilon}{w_\varepsilon - 1} \sum_{k=1}^n w_k (Y_i^{(k)} - \hat{\mu}_i^\varepsilon)^2 \right)^{1/2}, \left( \frac{\kappa_\varepsilon}{w_\varepsilon - 1} \sum_{k=1}^n w_k (Y_j^{(k)} - \hat{\mu}_j^\varepsilon)^2 \right)^{1/2} \right) \\
\hat{\rho}_{ij}^\varepsilon &= \frac{\kappa_\varepsilon}{w_\varepsilon - 1} \sum_{k=1}^n w_k \left( \frac{Y_i^{(k)} - \hat{\mu}_i^\varepsilon}{\hat{\sigma}_i^\varepsilon} \right) \left( \frac{Y_j^{(k)} - \hat{\mu}_j^\varepsilon}{\hat{\sigma}_j^\varepsilon} \right),
\end{aligned} \tag{3.5}$$

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

A direct result from Proposition 1 is that we can obtain consistent estimators of  $\boldsymbol{\mu}_J$  and  $\Sigma_J$  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  $\hat{\Gamma}_J^0 = (\rho_{ij}^0)$  be a robust and affine equivariant correlation estimator of the correlation matrix  $\Gamma_J$ . We will use  $\hat{\Gamma}_J^0$  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:

$$d_{k, \hat{\Gamma}_J^0}^2 = \left( \frac{Y_i^{(k)} - \hat{\mu}_i^0}{\hat{\sigma}_i^0}, \frac{Y_j^{(k)} - \hat{\mu}_j^0}{\hat{\sigma}_j^0} \right) (\hat{\Gamma}_J^0)^{-1} \left( \frac{Y_i^{(k)} - \hat{\mu}_i^0}{\hat{\sigma}_i^0}, \frac{Y_j^{(k)} - \hat{\mu}_j^0}{\hat{\sigma}_j^0} \right)^T. \tag{3.6}$$

We propose two estimators to compute the correlation matrix  $\hat{\Gamma}_J^0$  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 opposite quadrants that contains most of the standardized 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](#))

$$\Psi(Y_i, Y_j) = \begin{cases} \left( \psi_{c_1} \left( \frac{Y_i - \hat{\mu}_i^0}{\hat{\sigma}_i^0} \right), \psi_{c_1} \left( \frac{Y_j - \hat{\mu}_j^0}{\hat{\sigma}_j^0} \right) \right) & \text{if } \left( \frac{Y_i - \hat{\mu}_i^0}{\hat{\sigma}_i^0} \right) \left( \frac{Y_j - \hat{\mu}_j^0}{\hat{\sigma}_j^0} \right) \geq 0 \\ \left( \psi_{c_2} \left( \frac{Y_i - \hat{\mu}_i^0}{\hat{\sigma}_i^0} \right), \psi_{c_2} \left( \frac{Y_j - \hat{\mu}_j^0}{\hat{\sigma}_j^0} \right) \right) & \text{if } \left( \frac{Y_i - \hat{\mu}_i^0}{\hat{\sigma}_i^0} \right) \left( \frac{Y_j - \hat{\mu}_j^0}{\hat{\sigma}_j^0} \right) < 0, \end{cases} \quad (3.7)$$

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  $\hat{\Gamma}_j^0$  using the Spearman's rho as in equation (2.9). This approach is denoted by *Spearman's Winsorization*.

Therefore, given an affine equivariant robust correlation estimator  $\hat{\Gamma}_j^0$  (i.e. Adjusted Winsorized correlation coefficient or Spearman's rho), we estimate the robust Mahalanobis distance as in equation (3.6), then the outlier-free bivariate transformed data is computed as follows

$$\Psi_W(Y_i^{(k)}) = \begin{cases} \sqrt{c^2/d_{k, \hat{\Gamma}_j^0}^2} \left( \frac{Y_i^{(k)} - \hat{\mu}_i^0}{\hat{\sigma}_i^0} \right) & \text{if } d_{k, \hat{\Gamma}_j^0}^2 > c^2 \\ \frac{Y_i^{(k)} - \hat{\mu}_i^0}{\hat{\sigma}_i^0} & \text{if } d_{k, \hat{\Gamma}_j^0}^2 \leq c^2, \end{cases} \quad (3.8)$$

where  $c^2$  is given by  $P(\chi_2^2 > c^2) = \varepsilon$  and  $\varepsilon$  is the proportion of outliers we want to control assuming that the majority of the data follows a bivariate Gaussian distribution.

Given the observations  $(Y_1^{(k)}, \dots, Y_p^{(k)})^T$ , the winsorized correlation matrix  $\hat{\Gamma}^W = (\hat{\rho}_{ij}^W)$  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  $\Gamma^W$  into the  $\ell_1$  log-determinant divergence (2.4).

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 1, Panel (a), shows the scatter plot of contaminated data. Figure 1, Panel (b), 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 1, Panel (c), we observe that 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.

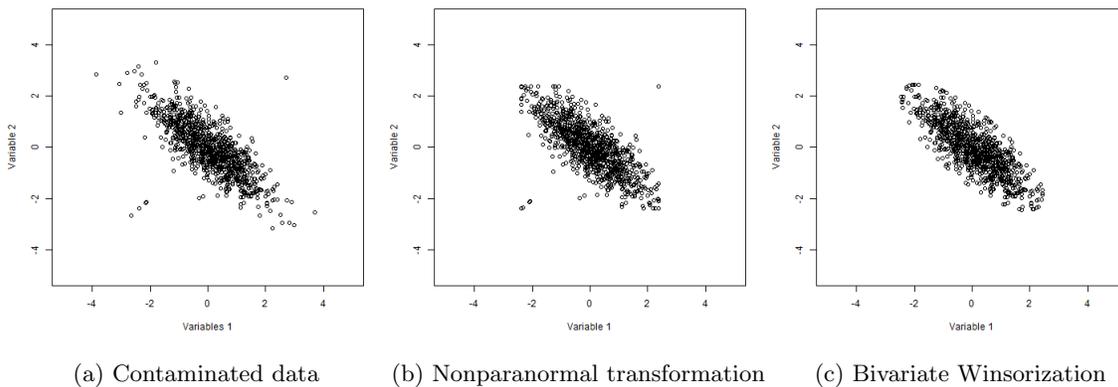


Figure 1: Illustration of nonparanormal transformation and bivariate winsorization under bivariate contamination.

In the next section we state some analytical properties of the bivariate winsorized pairwise scatter estimate is the same as the affine equivariant robust correlation estimates used to compute the Mahalanobis distance. 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.

## 4 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  $\mathbf{Y}_J^{(1)}, \dots, \mathbf{Y}_J^{(n)}$  be independent bivariate random vectors that follow a distribution in an elliptical family with density

$$f(\mathbf{y}_J) = \det(\Gamma_J)^{-1/2} h \left( \left( \frac{Y_i - \mu_i}{\sigma_i}, \frac{Y_j - \mu_j}{\sigma_j} \right)^T (\Gamma_J)^{-1} \left( \frac{Y_i - \mu_i}{\sigma_i}, \frac{Y_j - \mu_j}{\sigma_j} \right) \right) \quad (4.1)$$

where  $h : [0, \infty) \rightarrow [0, \infty)$  is assumed to be known. Under the assumption that the vector  $\mathbf{Y}_J = (Y_i, Y_j)^T$  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 (3.6). We apply the bivariate transformation in (3.8) and we compute the bivariate winsorized correlation estimator  $\hat{\rho}_{ij}^W$ .

Let  $w : [0, \infty) \rightarrow [0, 1]$  be the function defined in (3.2), 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  $n \rightarrow \infty$ . Let  $\boldsymbol{\theta}^* = (\mu_i, \mu_j, \sigma_i, \sigma_j, \rho_{ij})$  denote the true vector of parameters. Assuming that the estimates  $\hat{\boldsymbol{\theta}}^0$  are affine equivariant and consistent in probability (i.e.  $\hat{\boldsymbol{\theta}}^0 \rightarrow \boldsymbol{\theta}^*$  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 1.** *Let  $\mathbf{Y}_J^{(1)}, \dots, \mathbf{Y}_J^{(n)}$  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 (4.1). Suppose that  $w : [0, \infty) \rightarrow [0, 1]$  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,*

$$\hat{\rho}_{ij}^W - c_1 \rho_{ij} = o_p(1/\sqrt{n}) + o_p(\hat{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^*) + \frac{1}{n} \sum_{k=1}^n \left\{ w(d_k^2) \left( \frac{Y_i^{(k)} - \mu_i}{\sigma_i} \right) \left( \frac{Y_j^{(k)} - \mu_j}{\sigma_j} \right) - c_1 \rho_{ij} \right\}, \quad (4.2)$$

where the constant  $c_1$  is given by

$$c_1 = \pi \int_0^\infty w(r^2) h(r^2) r^3 dr > 0. \quad (4.3)$$

*Proof.* Theorem 1 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

$$\begin{aligned} \Psi_1(\mathbf{y}_J, \boldsymbol{\theta}) &= w(d^2(\boldsymbol{\theta})) \mathbf{y}_J \\ \Psi_2(\mathbf{y}_J, \boldsymbol{\theta}, \mathbf{t}) &= w(d^2(\boldsymbol{\theta})) (\mathbf{y}_J - \mathbf{t})(\mathbf{y}_J - \mathbf{t})^T. \end{aligned} \quad (4.4)$$

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

$$\begin{aligned} \hat{\boldsymbol{\mu}}_J^W &= \frac{1}{n} \sum_{k=1}^n w(\hat{d}_k^2) \mathbf{Y}_J^{(k)} \\ \hat{\Sigma}_J^W &= \frac{1}{n} \sum_{k=1}^n w(\hat{d}_k^2) (\mathbf{Y}_J^{(k)} - \hat{\boldsymbol{\mu}}_J^W)(\mathbf{Y}_J^{(k)} - \hat{\boldsymbol{\mu}}_J^W)^T. \end{aligned} \quad (4.5)$$

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

$$\begin{aligned}\hat{\boldsymbol{\mu}}_J^W &= \int \Psi_1(\mathbf{y}_J, \boldsymbol{\theta}) dP_n(\mathbf{y}_J) \\ \hat{\Sigma}_J^W &= \int \Psi_2(\mathbf{y}_J, \boldsymbol{\theta}, \hat{\boldsymbol{\mu}}_J^W) dP_n(\mathbf{y}_J),\end{aligned}\tag{4.6}$$

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

Moreover, estimates in (4.6) can be written as:

$$\begin{aligned}\int \Psi_1(\mathbf{y}_J, \hat{\boldsymbol{\theta}}^0) &= \int \Psi_1(\mathbf{y}_J, \hat{\boldsymbol{\theta}}^0) dP(\mathbf{y}_J) + \int \Psi_1(\mathbf{y}_J, \boldsymbol{\theta}^*) d(P_n - P)(\mathbf{y}_J) \\ &+ \int \left( \Psi_1(\mathbf{y}_J, \hat{\boldsymbol{\theta}}^0) - \Psi_1(\mathbf{y}_J, \boldsymbol{\theta}^*) \right) d(P_n - P)(\mathbf{y}_J),\end{aligned}\tag{4.7}$$

Suppose that  $\Sigma_J = B^2$  where  $B$  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  $\hat{\Sigma}_J^0 = B_n^2$  be affine equivariant location and scatter estimates such that  $(\hat{\boldsymbol{\mu}}_J^0 - \boldsymbol{\mu}_J, B_n - B)$  are consistent in probability. Then, using the result in [Lopuhaä \(1999\)](#) the first term in the right-hand side of (4.7) 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:

$$\int \Psi_1(\mathbf{y}_J, \boldsymbol{\theta}^*) d(P_n - P)(\mathbf{y}_J) = \frac{1}{n} \sum_{k=1}^n w(d_k^2) (\mathbf{Y}_J^{(k)} - \boldsymbol{\mu}_J).\tag{4.8}$$

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

$$\hat{\boldsymbol{\mu}}_J^W - \boldsymbol{\mu}_J = \frac{1}{n} \sum_{k=1}^n w(d_k^2) (\mathbf{Y}_J^{(k)} - \boldsymbol{\mu}_J) + c_0(\hat{\boldsymbol{\mu}}_J^0 - \boldsymbol{\mu}_J) + o_p(1/\sqrt{n}) + o_p(\hat{\boldsymbol{\mu}}_J^0 - \boldsymbol{\mu}_J, \hat{\Sigma}_J^0 - \Sigma_J)\tag{4.9}$$

the constants are given by

$$c_0 = 2\pi \int_o^\infty w(r^2)[h(r^2) + h'(r^2)r^2]rdr \quad (4.10)$$

$$c_1 = \pi \int_o^\infty w(r^2)h(r^2)r^3dr > 0. \quad (4.11)$$

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

$$\begin{aligned} \int \Psi_2(\mathbf{y}_J, \hat{\boldsymbol{\theta}}^0, \hat{\boldsymbol{\mu}}_J^W) &= c_1 \Sigma_J + c_2 \{ \text{tr}(B^{-1}(B_n - B))\Sigma_J + 2B^{-1}(B_n - B)\Sigma_J \} \\ &+ \frac{1}{n} \sum_{k=1}^n \{ w(d_k^2)(\mathbf{Y}_J^{(k)} - \boldsymbol{\mu}_J)(\mathbf{Y}_J^{(k)} - \boldsymbol{\mu}_J)^T - c_1 \Sigma_J \} \\ &+ o_p(1/\sqrt{n}) + o_p(\hat{\boldsymbol{\mu}}_J^0 - \boldsymbol{\mu}_J, B_n - B, \hat{\boldsymbol{\mu}}_J^W - \boldsymbol{\mu}_J), \end{aligned} \quad (4.12)$$

where  $B^{-1}(B_n - B) = (B_n - B)B^{-1} = A_n$ ,  $A_n$  is  $o_p(1)$  and the constant  $c_2$  is given by

$$c_2 = \pi \int_o^\infty w(r^2) \left[ r^2 h(r^2) + \frac{r^4}{2} h'(r^2) \right] r dr. \quad (4.13)$$

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:

$$\hat{\Gamma}_J^W = \int \Psi_2(\hat{\mathbf{y}}_J, \boldsymbol{\theta}) dP_n(\hat{\mathbf{y}}_J). \quad (4.14)$$

Using the result in (4.12) we obtain (4.2). □

Theorem 1 shows that the bivariate winsorized correlation estimate of  $\rho_{ij}$  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

Spearman's 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  $\rho_{ij}$  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  $\hat{\Gamma}$  converges to the true correlation matrix  $\Gamma$  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 1.** *Assume there exists a constant  $C$  such that the robust bivariate winsorized correlation coefficient estimator satisfies the following concentration bound*

$$Pr(|\hat{\rho}_{ij}^W - \rho_{ij}| > \epsilon) \leq 4exp(-Cn\epsilon^2) \quad (4.15)$$

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  $\Omega$  and by  $\mathcal{A}$  the support set of the off-diagonal elements in  $\Omega$ . Moreover, we define by  $K_\Gamma = \|\Gamma\|_\infty = \max_i \sum_j |\rho_{ij}|$  to be the matrix  $\ell_\infty$  norm of the true correlation matrix  $\Gamma$ , the matrix  $H_{\mathcal{A}\mathcal{A}} = [\Omega^{-1} \otimes \Omega^{-1}]_{\mathcal{A}\mathcal{A}}$  and the parameter  $K_H = \|H_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty$ . The following Theorem shows that is we plug a robust estimate of the correlation matrix, that achieves the optimal parametric bound in (4.15), 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 2.** *If there exists a constant  $\kappa \in (0, 1)$  such that  $\|H_{\mathcal{A}^c\mathcal{A}}(H_{\mathcal{A}\mathcal{A}})^{-1}\|_{\ell_\infty} < 1 - \kappa$ . Let  $\hat{\Omega}^W$  be the unique solution of the log-determinant program (2.4) with regularization parameter  $\lambda_n = \frac{8}{\kappa} \sqrt{\frac{\log 4n}{Cp^\tau}}$  for some  $\tau > 2$ . Then, if the sample size is lower bounded as*

$$n > \frac{\log(4/\max\{C^{-1/2}, 6(1+8\kappa^{-1})d\max\{K_\Gamma K_H, K_\Gamma^3 K_H^3\}\})}{Cp^{2\tau}}, \quad (4.16)$$

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

$$\|\hat{\Omega}^W - \Omega\|_\infty \leq \{2(1+8\kappa^{-1})K_H\} \sqrt{\frac{\log 4n}{Cp^\tau}}. \quad (4.17)$$

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  $|\omega_{ij}| > \{2(1+8\kappa^{-1})K_H\} \sqrt{\frac{\log 4n}{Cp^\tau}}$ .

If we consider that  $K_\Gamma$ ,  $K_H$  and  $\kappa$  remain constant as a function of  $(n, p, d)$ , we can obtain an asymptotic bound for the elementwise- $\ell_\infty$  norm  $\|\hat{\Omega}^W - \Omega\|_\infty \leq O\left(\sqrt{\frac{\log 4n}{Cp^\tau}}\right)$ . Assuming the concentration bound in Lemma 1, Theorem 2 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  $\hat{\rho}_{ij}^0$  converges to  $\rho_{ij}$  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  $\hat{\Gamma}^W$  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.

## 5 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  $\ell_1$  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 (2.4). In particular we consider the following correlation matrix estimates: “Adjusted Winsorization”, for the pairwise correlation matrix estimator using bivariate winsorization 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 winsorization 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. \(2012\)](#).

### 5.1 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  $\Omega$ :

1. AR(1) Model:  $\omega_{ii} = 1$ ,  $\omega_{i,i+1} = \omega_{i-1,i} = 0.4$  and 0 otherwise.
2. Erdős-Rényi random graph:  $\Omega = D(A + (|\lambda_{\min}(A)| + 0.2)I_p)D$  where  $A$  is a zero diagonal matrix where  $a_{ij} = 0.3a$ , such that  $a$  is independently generated and Bernoulli distributed with probability 0.01 and  $\lambda_{\min}(A)$  is the minimum eigenvalue of matrix  $A$ .  $D$  is a diagonal matrix with  $d_{ii} = 1$  for  $i = 1, \dots, p/2$  and  $d_{ii} = 3$  for  $i = p/2+1, \dots, p$ . The matrix is standardized to have unit diagonals.

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

$$\mathbf{Y} = (I - B)\mathbf{X} + B\mathbf{Z} \quad (5.1)$$

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

$$P(B_1 = 1) = \dots = P(B_p = 1) = \varepsilon \quad (5.2)$$

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  $i = 1, \dots, p$  and covariance matrix  $\Sigma^z = \Omega^{-1}$ . In the second contamination mechanism we assume that  $\mathbf{Z}$  is multivariate Gaussian distributed with mean  $\mu_i^z = 10$  for  $i = 1, \dots, p$  and covariance matrix  $\Sigma^z = 0.2I_p$ . 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 (5.1):  $\varepsilon = \{0, 0.05, 0.1\}$ . 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  $\Omega$  and  $\hat{\Omega}$ , given by the expectation of the squared of the Frobenius norm:

$$\text{MSE}(\hat{\Omega}) = E(\|\hat{\Omega} - \Omega\|_F^2). \quad (5.3)$$

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

ratio distance computed as

$$\text{LRT}(\hat{\Omega}) = \text{tr}(\hat{\Omega}(\Omega)^{-1}) - \log(\det(\hat{\Omega}(\Omega)^{-1})) - p. \quad (5.4)$$

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  $\hat{\Omega}$ . Let FP be the false non-zero elements and FN be the false zero elements estimated by  $\hat{\Omega}$ . The classification performance measures are then defined as follows:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad \text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5.5)$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (5.6)$$

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

$$\lambda^* = \underset{\lambda > 0}{\text{argmin}} \left\{ -\log(\det(\hat{\Omega})) + \text{tr}(\hat{\Omega}\hat{\Gamma}) + h \frac{\log(n)}{2n} \right\} \quad (5.7)$$

where  $h$  is the number of non-zero off-diagonal elements in  $\hat{\Omega}$ , and  $\hat{\Gamma}$  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](#)).

## 5.2 Simulation Results

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

Regarding the support recovery under the first contamination mechanism, Panel (a) of

Figures 2 and 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 2 and 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 bivariate winsorization significantly outperform the non-paranormal SKEPTIC proposed by Liu et al. (2012) and the winsorized normal-score nonparanormal from Liu et al. (2009).

Tables 1 and 2 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 Erdős-Rényi random graph, respectively. For clean data, the “Sample Correlation” slightly 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.

Regarding the second contamination specification, simulation results can be sent upon request. Under this contamination mechanism the performance of the bivariate winsorized

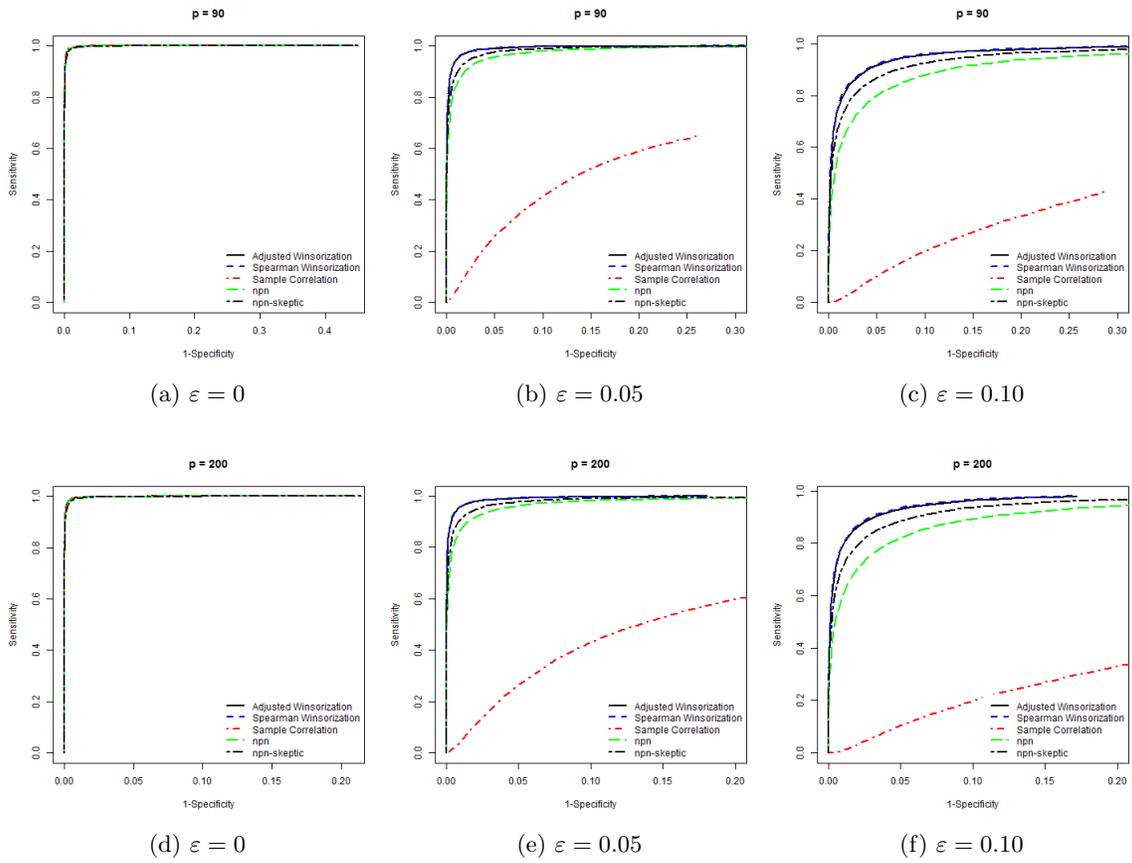


Figure 2: AR(1)-Model Specification. ROC curves under the first contamination mechanism over 100 replications.

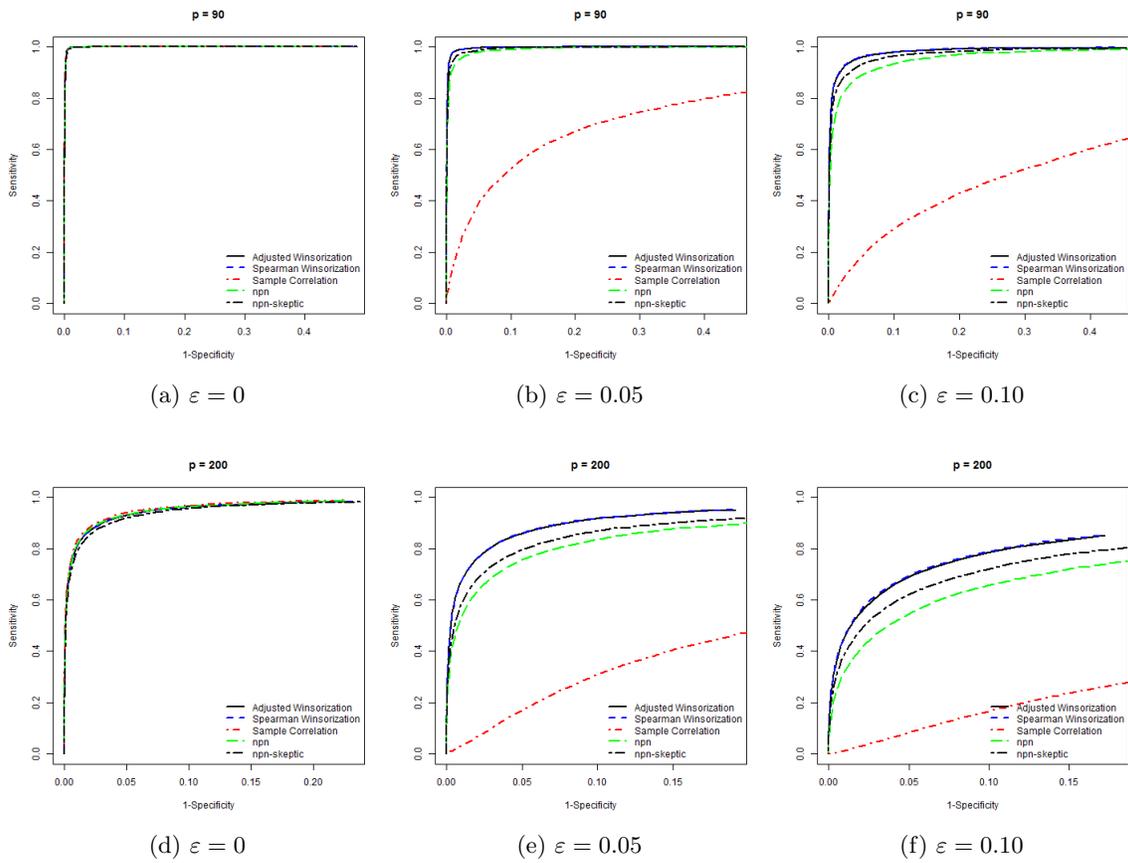


Figure 3: Erdős-Rényi Specification. ROC curves under the first contamination mechanism over 100 replications.

estimates to recover the true GGM, for the AR(1) Model and Erdős-Rényi random graph, confirms the insights of the first contamination mechanism.

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

Table 2: Erdős-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 Winsorization	90	10.118 (0.410)	10.168 (0.464)	13.537 (0.535)	13.274 (0.526)	17.953 (0.893)	17.082 (0.588)
	200	32.129 (0.752)	43.434 (0.482)	36.125 (0.825)	45.066 (0.365)	37.976 (0.746)	43.658 (0.254)
Adjusted Winsorization	90	10.083 (0.407)	10.126 (0.463)	13.381 (0.537)	13.131 (0.532)	17.767 (0.904)	16.956 (0.599)
	200	33.693 (0.712)	45.995 (0.425)	35.99 (0.834)	44.988 (0.375)	37.846 (0.764)	43.603 (0.261)
Sample Correlation	90	10.049 (0.400)	10.093 (0.455)	22.758 (0.311)	22.771 (0.135)	23.213 (0.105)	22.234 (0.038)
	200	32.073 (0.746)	43.405 (0.492)	39.995 (0.132)	49.502 (0.035)	39.996 (0.030)	46.808 (0.010)
npn	90	10.273 (0.412)	10.360 (0.456)	16.016 (0.633)	16.690 (0.509)	20.239 (0.757)	20.525 (0.436)
	200	35.589 (0.667)	48.757 (0.353)	37.265 (0.702)	46.883 (0.299)	38.834 (0.533)	46.321 (0.184)
npn-SKEPTIC	90	10.863 (0.455)	11.661 (0.482)	15.281 (0.585)	16.770 (0.493)	19.267 (0.800)	20.637 (0.499)
	200	35.283 (0.691)	48.508 (0.370)	36.977 (0.697)	48.104 (0.314)	38.317 (0.648)	47.387 (0.229)

## 6 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](#)):

$$\mathbf{Y}^{(k)} = (I - B)\mathbf{X}^{(k)} + B\mathbf{Z}^{(k)} \quad \text{for } k = 1, \dots, n \quad (6.1)$$

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  $B$  is the contamination indicator matrix where  $P(B_1 = 1) = \dots = P(B_p = 1) = \varepsilon$  (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

$$\delta_t(Y^{(k)}) = -\frac{1}{2}\log(\det(\hat{\Omega})) - \frac{1}{2}d^2(Y^{(k)}, \hat{\mu}_t, \hat{\Omega}) + \log\hat{\pi}_t, \quad (6.2)$$

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

squared Mahalanobis distance. The classification rule is

$$\hat{t}(Y^{(k)}) = \operatorname{argmax} \delta_t(Y^{(k)}) \quad \text{for } t = 1, \dots, T. \quad (6.3)$$

To perform model selection for  $\lambda$  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.

## 6.1 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) chemotherapy of patients with pathological complete response (pCR) and with residual disease (RD). The importance of analyzing the subject response to neoadjuvant (preoperative) chemotherapy, 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 signed-rank test is performed to select the 113 most significant genes.

Table 3 displays the average classification performance and the number of missclassified 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: Comparison of average pCR classification errors over 100 replications with standard deviation in brackets.

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

## 6.2 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 signed-rank test is performed

Table 4: Comparison of average leukemia classification errors over 100 replications with standard deviation in brackets.

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

to select the 50 most significant genes.

Table 4 displays the average classification performance and the number of missclassified 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” outperforms “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](#)).

## 7 Conclusions

In this article 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. 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  $\ell_1$  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 equivariant 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 high dimensions.

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  $\Omega$ . 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  $\ell_1$  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.

## SUPPLEMENTARY MATERIAL

**R script for Adjusted Winsorization** R script cor.hub containing code to estimate the bivariate winsorized correlation matrix using adjusted winsorization describe in the article. (.R file)

**R script for Spearman Winsorization** R script cor.spearman containing code to estimate the bivariate winsorized correlation matrix using Spearman’s rho describe in the article. (.R file)

## References

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.
- Cai, T., W. Liu, and X. Luo (2011). A constrained  $\ell_1$  minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* 106(494), 594–607.
- Ceroli, 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association* 53(284), 814–861.
- 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.

- Liu, H., F. Han, M. Yuan, J. Lafferty, and L. Wasserman (2012). High-dimensional semi-parametric 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.
- Maronna, R. A. (1976, 01). Robust  $m$ -estimators of multivariate location and scatter. *The Annals of Statistics* 4(1), 51–67.
- 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.
- 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.
- 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, G. Raskutti, B. Yu, et al. (2011). High-dimensional covariance estimation by minimizing 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.
- 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.
- 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.
- 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.
- Wang, H., R. Li, and C.-L. Tsai (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 94(3), 553–568.
- 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.
- 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.
- 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.