Robust sparse Gaussian graphical modeling

Kei Hirose\textsuperscript{1}, Hironori Fujisawa\textsuperscript{2} and Jun Sese\textsuperscript{3}

\textsuperscript{1} Institute of Mathematics for Industry, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan.
\textsuperscript{2} The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan.
\textsuperscript{3} National Institute of Advanced Industrial Science and Technology, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan

E-mail: hirose@imi.kyushu-u.ac.jp, fujisawa@ism.ac.jp, sese.jun@aist.go.jp

Abstract

Gaussian graphical modeling has been widely used to explore various network structures, such as gene regulatory networks and social networks. We often use a penalized maximum likelihood approach with the $L_1$ penalty for learning a high-dimensional graphical model. However, the penalized maximum likelihood procedure is sensitive to outliers. To overcome this problem, we introduce a robust estimation procedure based on the $\gamma$-divergence. The proposed method has a descending property, which is known as a desirable property in robust statistics. The parameter estimation procedure is constructed using the Majorize-Minimization algorithm, which guarantees that the objective function monotonically decreases at each iteration. Extensive simulation studies showed that our procedure performed much better than the existing methods, in particular, when the contamination ratio was large. Two real data analyses were carried out to illustrate the usefulness of our proposed procedure.

Keywords: $\gamma$-divergence, graphical lasso, Majorize-Minimization algorithm, robust estimation.

1 Introduction

Gaussian graphical modeling has been widely used to investigate the conditional independence between two variables given other variables. Under a Gaussian assumption, the conditional independence between two variables corresponds to the zero entry of inverse covariance matrix (Edwards, 2000). A sparse estimation of the inverse covariance matrix, i.e., a method in which some of the elements of the inverse covariance matrix are shrunk to exactly zero, is often used to obtain the conditional independence graph.

In many applications, the number of variables is much larger than the number of observations. An example is the analysis of microarray gene expression data, in which
we discover the relation between pairs of genes. In such a case, the maximum likelihood estimate of the inverse covariance matrix does not exist. To overcome this problem, there has been a great deal of interest on the $L_1$ regularization, such as the lasso (Tibshirani, 1996), for estimating the sparse inverse covariance matrix. Meinshausen and Bühlmann (2006) proposed fitting the lasso regression to each variable, in which one variable is a response and the other variables are predictors. Their method does not guarantee that the non-zero pattern of the inverse covariance matrix is symmetric. Peng et al. (2009) introduced a joint regression, which is also based on the lasso regression and ensures the symmetry of the estimated inverse covariance matrix. Yuan and Lin (2007) considered the problem of maximizing the penalized log-likelihood function via the lasso (hereafter referred to as the graphical lasso). Among these methods, the graphical lasso has been becoming popular because of its computational efficiency (e.g., Friedman et al., 2008; Witten et al., 2011; Hsieh et al., 2011) and desirable statistical properties in high-dimensional settings (Rothman et al., 2008; Raskutti et al., 2008).

In practical situations, however, outliers are often observed or the distribution is heavy-tailed (Finegold and Drton, 2011; Fritsch et al., 2012). In such cases, the conventional estimation procedure may produce an inappropriate graph structure. To overcome this problem, a few researchers proposed the robust estimation procedures with the $L_1$ penalization. Liu et al. (2009) proposed the nonparanormal, in which a truncated marginal empirical distribution was adopted to remove outliers and a semiparametric Gaussian copula was used to treat a conditional independence structure. The model parameter was estimated by a standard algorithm of the graphical lasso, such as the blockwise coordinate descent algorithm (Friedman et al., 2008). Finegold and Drton (2011) introduced the tlasso, in which the underlying distribution was assumed to be the multivariate $t$-distribution with a heavy tail. The model parameters were estimated by the EM algorithm. Vinciotti and Hashem (2013) compared performances of various robust estimation procedures, including the nonparanormal and the tlasso, and they concluded the nonparanormal performed well in various situations. Sun and Li (2012) considered a modified likelihood approach based on the density power divergence (Basu et al., 1998) (hereafter referred to as the dp-lasso). The model parameter was estimated by the coordinate descent algorithm with a quadratic approximation (Tseng and Yun, 2009).

However, the above procedures have some drawbacks. The nonparanormal approach removes observations on both sides at the ratio $2\delta$, i.e., observations that have extremely large positive and negative values are removed at the same ratio $\delta$. The truncation parameter $\delta$ corresponds to the contamination ratio and must be selected beforehand. Liu et al. (2009) selected $\delta$ such that it achieved a desired rate of convergence of the estimator. Nevertheless, the selected truncation parameter tends to be too small when the contamination ratio is large, because $\delta \rightarrow 0$ as $n \rightarrow \infty$, where $n$ is the number of the
observations. In addition, the outliers may not exist on both sides at the same ratio \( \delta \).

In fact, the outliers of yeast gene expression data described in Section 6.1 have only large negative values. The \( t \)-lasso has the same drawback, because the heavy tail distribution implies that outliers are assumed on both sides. Furthermore, the \( t \)-lasso tends to yield a large variance of the estimator, because a heavier tail distribution often produces a smaller Fisher information. The \( dp \)-lasso approach has four tuning parameters to be determined, and is often unstable in our experience. In our simulation study, we observed that the above three estimation procedures performed poorly when the contamination ratio was large and the outliers were present on one side, and the estimators had large mean squared errors even when the number of observations was sufficiently large.

To handle the issues above, we propose the \( \gamma \)-lasso, which is a robust sparse estimation procedure of the inverse covariance matrix based on the \( \gamma \)-divergence (Fujisawa and Eguchi, 2008; Cichocki and Amari, 2010). The \( \gamma \)-lasso regards an observation whose likelihood value is small as an outlier, unlike the nonparanormal. As a result, the \( \gamma \)-lasso can appropriately treat the outliers even when they exist on only one side. In addition, we do not need to know the contamination ratio in advance. The \( \gamma \)-lasso tends to yield a much smaller variance of the estimator than the \( t \)-lasso, because the underlying distribution is assumed to be Gaussian. The parameter estimation algorithm is proposed using the Majorize-Minimization algorithm (MM algorithm, Hunter and Lange, 2004), which guarantees that the objective function monotonically decreases at each iteration. The proposed algorithm does not have any tuning parameters to be determined. As a result, the parameter estimation is more stable than the \( dp \)-lasso. In addition, the \( \gamma \)-lasso has a redescending property, which is known as a desirable property in robust statistics, so that the bias of the estimator is expected to be sufficiently small when an outlier takes a large value (Maronna et al., 2006). We conducted extensive Monte Carlo simulations to investigate the performance of the proposed procedure. The result showed that our procedure performed better than existing methods in most cases. The proposed procedure is available for use in the R package \texttt{rsggm}\footnote{Available at \url{http://cran.r-project.org/web/packages/rsggm}.}

The organization of this paper is given as follows. In Section 2, we introduce a robust estimation of the sparse inverse covariance matrix via the \( \gamma \)-divergence. Section 3 provides a parameter estimation procedure via the MM algorithm. In Section 4, we compare the proposed procedure with several existing methods. Section 5 investigates the effectiveness of our proposed procedure via Monte Carlo simulations. Section 6 describes two real data analyses of the gene expression data. Concluding remarks are given in Section 7. Some technical proofs are collected in Appendices.
2 Robust and sparse estimation of the inverse covariance matrix

2.1 Gaussian graphical model

Let \( X = (X_1, \ldots, X_p)^T \) be the \( p \)-dimensional multivariate-normally distributed random variable with mean vector \( \mu = (\mu_1, \ldots, \mu_p)^T \) and covariance matrix \( \Sigma = (\sigma_{ij}) \). Let the inverse covariance matrix of \( \Sigma \) be denoted by \( \Omega = (\omega_{ij}) \). It is well-known that each variable is written as \( X_i = \sum_{j \neq i} \beta_{ij} X_j + \delta_i \), where \( \beta_{ij} = -\omega_{ij} \omega_{ii} \) and \( \delta_i \sim N(0, 1/\omega_{ii}) \), and then the zero/non-zero element of the inverse covariance matrix corresponds to the conditional independence/dependence given other variables. The sparsity pattern of the inverse covariance matrix corresponds to the graph structure: there is an edge between \( i \)th and \( j \)th vertices if and only if \( \omega_{ij} \neq 0 \). We estimate the inverse covariance matrix by a sparse matrix to obtain a sparse graphical model.

2.2 Sparse estimation of the Gaussian graphical model

Suppose that we have a random sample of \( n \) observations \( x_1, \ldots, x_n \) from the \( p \)-dimensional normal population \( N(\mu, \Sigma) \). To estimate the sparse inverse covariance matrix, Yuan and Lin (2007) proposed minimizing the following penalized negative log-likelihood function:

\[
\ell_\lambda(\theta) = \ell(\theta) + \frac{\lambda}{2} \| \Omega - \text{diag}(\Omega) \|_1,
\]

where \( \ell(\theta) \) is a negative log-likelihood function given by \( \ell(\theta) = -\sum_{i=1}^{n} \log f(x_i; \theta) \), and \( \lambda \geq 0 \) is a tuning parameter which controls the balance between sparsity of parameters and goodness of fit to the data. Here \( f(x; \theta) \) is a density function of the multivariate normal distribution

\[
f(x; \theta) = (2\pi)^{-p/2} |\Omega|^{1/2} \exp \left( -\frac{1}{2} (x - \mu)^T \Omega (x - \mu) \right),
\]

and \( \theta \) is a model parameter expressed as \( \theta = (\mu^T, \text{vech}(\Omega)^T)^T \).

For any \( \Omega \), the penalized negative log-likelihood function \( \ell_\lambda(\theta) \) is minimized when the mean vector \( \mu \) is the sample mean \( \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \). The inverse covariance matrix \( \Omega \) is then estimated by the following minimization problem:

\[
\min_{\Omega} \left\{ -\log |\Omega| + \text{tr}(\Omega S) + \lambda \| \Omega - \text{diag}(\Omega) \|_1 \right\}, \tag{1}
\]

where \( S = (s_{ij}) \) is the sample covariance matrix. The problem of (1) is referred to as the graphical lasso (Witten et al., 2011). Several researchers have proposed efficient algorithms for solving the problem of (1), such as the blockwise coordinate descent approach (Friedman et al., 2008), the quadratic approximation (Hsieh et al., 2011), and the Alternating Direction Method of Multipliers (ADMM; Boyd et al., 2011).
2.3 Robust estimation via the $\gamma$-divergence

In practical situations, the estimate obtained by (1) is sensitive to outliers. To obtain a robust estimate, instead of the negative log-likelihood function $\ell(\theta)$, we use the negative $\gamma$-likelihood function [Fujisawa and Eguchi, 2008; Cichocki and Amari, 2010], given by

$$\ell_\gamma(\theta) = -\frac{1}{\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^{n} f(x_i; \theta)^\gamma \right\} + \frac{1}{1 + \gamma} \log \int f(x; \theta)^{1+\gamma} dx,$$

where $\gamma \geq 0$ is a tuning parameter which controls the balance between efficiency and robustness. Note that $\gamma \to +0$ corresponds to the negative log-likelihood function. The first term $\ell_1(\theta)$ can lead to a robust estimation, and the second term $\ell_2(\theta)$ makes the bias of the estimate sufficiently small.

We provide an intuitive explanation about how the negative $\gamma$-likelihood function can lead to the robust estimation. Suppose that $x_1$ is an outlier. The likelihood $f(x_1; \theta)$ is expected to be sufficiently small. Minimizing the negative log-likelihood function $\ell(\theta) = -\sum_{i=1}^{n} \log f(x_i; \theta)$ cannot make $f(x_1; \theta)$ extremely small, because $\ell(\theta) \to \infty$ as $f(x_1; \theta) \to 0$. On the other hand, with the negative $\gamma$-likelihood function $\ell_\gamma(\theta)$, the likelihood term $f(x_1; \theta)$ is naturally ignored, because we can easily obtain the following approximation:

$$\arg \min_{\theta} \ell_\gamma(\theta) \approx \arg \min_{\theta} \left[ -\frac{1}{\gamma} \log \left\{ \frac{1}{n-1} \sum_{i=2}^{n} f(x_i; \theta)^\gamma \right\} + \frac{1}{1 + \gamma} \log \int f(x; \theta)^{1+\gamma} dx \right].$$

The robust and sparse estimate is proposed by

$$\hat{\theta} = \arg \min_{\theta} \ell_\gamma(\theta),$$

where

$$\ell_{\gamma,\lambda}(\theta) = \ell_\gamma(\theta) + \frac{\lambda}{2} \| \Omega - \text{diag}(\Omega) \|_1.$$  (3)

We call $\ell_{\gamma,\lambda}(\theta)$ the penalized negative $\gamma$-likelihood function, and the minimization problem of (3) the $\gamma$-lasso.

2.4 Illustrative Example

We provide solution paths (estimates of $\omega_{ij}$ ($i, j = 1, \ldots, p, i < j$) as a function of $\sum_{i<j} |\omega_{ij}|$) of the $\gamma$-lasso and the ordinary graphical lasso when outliers exist. We gener-
ated \( n = 200 \) observations from a mixture distribution \( 0.9N(0, \Omega^{-1}) + 0.1N(5, I) \), where

\[
\Omega = \begin{pmatrix}
1.0 & 0.3 & 0.3 & 0.0 & 0.0 \\
0.3 & 1.0 & 0.0 & 0.0 & 0.3 \\
0.3 & 0.0 & 1.0 & 0.3 & 0.0 \\
0.0 & 0.0 & 0.0 & 1.0 & 0.0 \\
0.0 & 0.3 & 0.0 & 0.0 & 1.0 \\
\end{pmatrix},
\]

\( I \) is an identity matrix, and \( 5 \) is a 5-dimensional vector whose elements are 5. In this case, \( N(0, \Omega^{-1}) \) is our target model and \( N(5, I) \) corresponds to the contamination. Figure 1(a) shows the solution path of the graphical lasso applied to the uncontaminated data in which the outliers were removed from the original contaminated data. Figures 1(b) and (c) depict the solution paths of graphical lasso and \( \gamma \)-lasso, respectively, applied to the original contaminated data. For \( \gamma \)-lasso, we chose \( \gamma = 0.1 \).

![Graphical Lasso vs. \gamma-Lasso](image)

Figure 1: Solution paths (estimates of \( \omega_{ij} \) \( i, j = 1, \ldots, p, i < j \)) as a function of \( \sum_{i<j} |\hat{\omega}_{ij}| \) for the dataset generated from a mixture distribution. (a) The solution path of the graphical lasso applied to the uncontaminated data in which the outliers were removed from the original contaminated data. (b)-(c) The solution paths of graphical lasso and \( \gamma \)-lasso applied to the original contaminated data.

Clearly, the solution path of the graphical lasso in Figure 1(b) was completely different from Figure 1(a), which implies that the graphical lasso was highly sensitive to the outliers. However, the solution path of the \( \gamma \)-lasso in Figure 1(c) was almost the same as Figure 1(a), so that our method was robust against the outliers.

### 2.5 Redescending Property

Suppose that the estimating equation is given by \( \sum_{i=1}^{n} \psi(x_i; \theta) = 0 \). The estimating equation is said to have a redescending property if \( \lim_{\|x\| \to \infty} \psi(x; \theta^*) = 0 \), where \( \theta^* \) is
a true parameter of \( \theta \). The redescending property on M-estimation (see, e.g., Maronna et al. [2006]) is known as a desirable property in robust statistics, because the bias of the estimator is expected to be sufficiently small when an outlier takes a large value (Maronna et al. [2006]).

The \( \gamma \)-lasso minimizes the loss function in (3), so that the estimating equation is expressed as

\[
- \frac{1}{n} \sum_{i=1}^{n} f(x_i; \theta)^{\gamma} s(x_i; \theta) + \frac{\partial}{\partial \theta} \ell_2(\theta) + \frac{\lambda}{2} u = 0, \tag{4}
\]

where

\[
s(x; \theta) = \frac{\partial \log f(x; \theta)}{\partial \theta}, \quad u = (u_1, \ldots, u_{p(p+2)})^T, \quad u_j \in [-1, 1].
\]

Therefore, the kernel function \( \psi(x; \theta) \) is given by

\[
\psi(x; \theta) = f(x; \theta)^{\gamma} s(x; \theta) - f(x; \theta)^{\gamma} \frac{\partial}{\partial \theta} \ell_2(\theta) - f(x; \theta)^{\gamma} \frac{\lambda}{2} u. \tag{5}
\]

Because \( f(x; \theta) \) is the density function of the Gaussian distribution, we have

\[
f(x; \theta)^{\gamma} \frac{\partial \log f(x; \theta)}{\partial \mu} = (2\pi)^{-p/2} |\Omega|^{\gamma/2} \exp \left\{ -\frac{\gamma}{2} (x - \mu)^T \Omega (x - \mu) \right\} \times [\Omega (x - \mu)] \to 0 \quad \text{(as } \|x\| \to \infty),
\]

\[
f(x; \theta)^{\gamma} \frac{\partial \log f(x; \theta)}{\partial \omega_{jk}} = (2\pi)^{-p/2} |\Omega|^{\gamma/2} \exp \left\{ -\frac{\gamma}{2} (x - \mu)^T \Omega (x - \mu) \right\} \times \left[ \frac{1}{2} \frac{\partial}{\partial \omega_{jk}} \log |\Omega| - \frac{1}{2} (x - \mu)^T \frac{\partial \Omega}{\partial \omega_{jk}} (x - \mu) \right] \to 0 \quad \text{(as } \|x\| \to \infty).
\]

Therefore, the first term of the right side in (5) approaches 0 as \( \|x\| \to \infty \). The elements of \( u \) are bounded, so that the second and third terms of the right side in (5) also approach 0 as \( \|x\| \to \infty \). As a result, the estimating equation has a redescending property.

In numerical studies in Section 5, the proposed method is compared with three existing methods. It should be mentioned that two of them, the dp-lasso and \( t \)-lasso, do not have the redescending property, and one of them, the nonparanormal, is not an M-estimator. The redescending property for the dp-lasso and \( t \)-lasso is discussed in Sections 4.1 and 4.2.

### 3 Algorithm

Fujisawa and Eguchi (2008) proposed an iterative minimization algorithm that monotonically decreases the negative \( \gamma \)-likelihood function \( \ell_\gamma(\theta) \) given by (2) at each step. Their algorithm is based on the Pythagorean relation for the \( \gamma \)-divergence. Unfortunately, their idea cannot be directly applied to our minimization problem of the penalized negative
γ-likelihood $\ell_{\gamma,\lambda}(\theta)$ given by (3), due to the $L_1$ penalization. In this section, we propose an efficient minimization algorithm using the Majorize-Minimization algorithm (MM algorithm; Hunter and Lange, 2004; Lange, 2010).

### 3.1 Construction of the majorization function

Let $\theta^{(t)}$ be the estimate at the $t$th step. We construct a majorization function of $\ell_1(\theta)$, say $\tilde{\ell}_1(\theta|\theta^{(t)})$. The majorization function must satisfy the following properties:

\begin{align}
\tilde{\ell}_1(\theta|\theta^{(t)}) & \geq \ell_1(\theta), \quad (6) \\
\tilde{\ell}_1(\theta|\theta^{(t)}) & = \ell_1(\theta^{(t)}). \quad (7)
\end{align}

To construct a majorization function, first, we apply the Jensen’s inequality to the convex function $y = -\log x$ and we have

$$-\log \left( \sum_{i=1}^{n} w_i^{(t)} r_i^{(t)} \right) \leq - \sum_{i=1}^{n} w_i^{(t)} \log r_i^{(t)}, \quad (8)$$

where

\begin{align}
\hat{w}_i^{(t)} &= \frac{f(x_i; \theta^{(t)})^\gamma}{\sum_{j=1}^{n} f(x_j; \theta^{(t)})^\gamma} = \exp \left\{ -\frac{\gamma}{2} (x_i - \mu^{(t)})^T \Omega^{(t)} (x_i - \mu^{(t)}) \right\}, \quad (9) \\
\hat{r}_i^{(t)} &= \sum_{j=1}^{n} f(x_j; \theta^{(t)})^\gamma \frac{f(x_i; \theta)^\gamma}{f(x_i; \theta^{(t)})^\gamma}. \quad (10)
\end{align}

Here $\mu^{(t)}$ and $\Omega^{(t)}$ are the estimates of $\mu$ and $\Omega$ at the $t$th step, respectively. Note that $\sum_{i=1}^{n} w_i^{(t)} = 1$ and $w_i^{(t)} r_i^{(t)} = f(x_i; \theta)^\gamma$. Then, substituting (9) and (10) into (8) gives us

$$\ell_1(\theta) \leq - \sum_{i=1}^{n} w_i^{(t)} \log f(x_i; \theta) + C, \quad (11)$$

where $C = \frac{1}{\gamma} \sum_{i=1}^{n} w_i^{(t)} \log w_i^{(t)} + \frac{1}{\gamma} \log n$. Finally, we define $\tilde{\ell}_1(\theta|\theta^{(t)})$ as the right side of (11), i.e.,

$$\tilde{\ell}_1(\theta|\theta^{(t)}) = - \sum_{i=1}^{n} w_i^{(t)} \log f(x_i; \theta) + C. \quad (12)$$

It is shown that the function (12) satisfies the properties of (6) and (7).

The majorization function (12) is viewed as a weighted negative log-likelihood function with weights $w_i^{(t)}$ ($i = 1, \ldots, n$). When $x_i$ is an outlier, the corresponding likelihood $f(x_i; \theta^{(t)})$ is expected to be sufficiently small, so that the weight $w_i^{(t)}$, which is proportional to the $\gamma$th power of likelihood, is expected to be sufficiently small.
3.2 Update algorithm

We propose an update algorithm given by

$$\theta^{(t+1)} = \arg \min_{\theta} \tilde{\ell}_{\gamma,\lambda}(\theta|\theta^{(t)})$$

where

$$\tilde{\ell}_{\gamma,\lambda}(\theta|\theta^{(t)}) = \tilde{\ell}_1(\theta|\theta^{(t)}) + \ell_2(\theta) + \frac{\lambda}{2}\|\Omega - \text{diag}(\Omega)\|_1.$$ 

From the properties of the majorization function given by (6) and (7), the target function $$\ell_{\gamma,\lambda}(\theta)$$ in (3) monotonically decreases at each step: $$\ell_{\gamma,\lambda}(\theta^{(t)}) \geq \ell_{\gamma,\lambda}(\theta^{(t+1)})$$.

After a simple calculation of $$\ell_2(\theta)$$, which is detailed in Appendix A, we have

$$\tilde{\ell}_1(\theta|\theta^{(t)}) + \ell_2(\theta) = -\frac{1}{2(1 + \gamma)} \log |\Omega| + \frac{1}{2} \text{tr}(\Omega S_{w(\cdot)}(\mu)) + C'$$,

where

$$S_{w(\cdot)}(\mu) = \sum_{i=1}^{n} w_i^{(t)} (x_i - \mu)(x_i - \mu)^T,$$

and $$C'$$ is a constant. For any $$\Omega$$, the majorization function $$\tilde{\ell}_{\gamma,\lambda}(\theta|\theta^{(t)})$$ is minimized at

$$\mu^{(t+1)} = \sum_{i=1}^{n} w_i^{(t)} x_i.$$

For given $$\mu^{(t+1)}$$, the inverse covariance matrix $$\Omega^{(t+1)}$$ is obtained by minimizing the following function with respect to $$\Omega$$:

$$-\frac{1}{2(1 + \gamma)} \log |\Omega| + \frac{1}{2} \text{tr}(\Omega S_{w(\cdot)}(\mu^{(t+1)})) + \frac{\lambda}{2}\|\Omega - \text{diag}(\Omega)\|_1.$$ (13)

The above minimization problem corresponds to the graphical lasso in [1]. We can use a standard algorithm of the graphical lasso, such as the blockwise coordinate descent algorithm [Friedman et al., 2008], to obtain $$\Omega^{(t+1)}$$.

Remark 3.1 When $$\gamma = 0$$, the MM algorithm corresponds to the standard graphical lasso.

Remark 3.2 When $$\lambda = 0$$, our update algorithm is identical to that of Example 4.1 in [Fujisawa and Eguchi, 2008], in which the estimation algorithm is constructed by using the Pythagorean relation for the $$\gamma$$-divergence.

3.3 Computation of entire path of solutions

In practice, we set a sequence of decreasing regularization parameters $$\lambda_1, \ldots, \lambda_K$$ on the log scale, and an entire path of solutions is made by sequences of $$\lambda$$. The determination of the value of $$\lambda_1$$, which is a minimum value of $$\lambda$$ so that all of the non-diagonal elements of inverse covariance matrix are zeros, is provided in Appendix B. $$\lambda_K$$ is determined by $$\lambda_K = \delta \lambda_1$$, where $$\delta$$ is a positive value smaller than 1. In our R package rsggm, the default is $$\delta = 0.2$$ and $$K = 10$. 

9
4  Comparison with existing methods

4.1  dp-lasso

Sun and Li (2012) considered the problem of minimizing the following objective function based on the density power divergence (Basu et al., 1998):

\[ \ell_{\beta,\lambda}(\theta) = \ell_{\beta}(\theta) + \frac{\lambda}{2}\| \Omega - \text{diag}(\Omega) \|_1, \]  

(14)

where \( \ell_{\beta}(\theta) \) is the modified likelihood function based on the density power divergence given by

\[ \ell_{\beta}(\theta) = -\frac{1}{n\beta} \sum_{i=1}^{n} f(x_i; \theta)^{\beta} + \frac{1}{1+\beta} \int f(x; \theta)^{1+\beta} dx. \]  

(15)

Here \( \beta \geq 0 \) controls the balance between efficiency and robustness. Miyamura and Kano (2006) and Sun and Li (2012) used \( \ell_{\beta}(\theta) \) for robust estimation of the Gaussian graphical models. We call the minimization problem of (14) the dp-lasso.

The density power divergence is quite similar to the \( \gamma \)-divergence. The difference between \( \ell_{\gamma}(\theta) \) and \( \ell_{\beta}(\theta) \) is just the existence of the logarithm on \( \ell_{\gamma}(\theta) \). However, the estimators based on these two likelihood functions have much different robustness properties: the \( \gamma \)-lasso has the redescending property but the dp-lasso does not. In fact, the kernel function of estimating equation of the dp-lasso can be expressed as

\[ \psi(x; \theta) = -f(x; \theta)^{\gamma} s(x; \theta) + \frac{\partial}{\partial \theta} b_{\beta}(\theta) + \lambda \frac{1}{2} u, \]  

(16)

where

\[ b_{\beta}(\theta) = \frac{|\Omega|^{\beta/2}}{(1+\beta)^{1+\beta/2}(2\pi)^{p^{\beta/2}}}. \]

The first term of right side of (16) becomes zero as \( \|x\| \to \infty \). However, we have \( \frac{\partial}{\partial \omega_{jk}} b_{\beta}(\theta) \) does not always converges to 0 as \( \|x\| \to \infty \), and therefore \( \psi(x; \theta) \) does not go to zero as \( \|x\| \to \infty \).

For the \( \gamma \)-divergence, we proposed the MM algorithm by using the concavity of the function \( f(x) = \log x \), which guarantees that the objective function monotonically decreases at each step. On the other hand, it is difficult to derive the MM algorithm for the density power divergence. Sun and Li (2012) applied the coordinate descent algorithm using a quadratic approximation (Tseng and Yun, 2009), but it does not guarantee that the objective function monotonically decreases at each step. Furthermore, their algorithm has additional difficulties to be applied as follows:

- Their algorithm does not guarantee to have a positive definite inverse covariance matrix.
Their algorithm based on Tseng and Yun (2009) depends on four tuning parameters to be determined.

In the $\gamma$-lasso, the estimated inverse covariance matrix is positive definite, and the update algorithm does not have any tuning parameter to be determined in advance.

### 4.2 $t$-lasso

Finegold and Drton (2011) proposed a penalized maximum likelihood approach using a multivariate $t$-distribution instead of a Gaussian distribution, which was referred to as the $t$-lasso. The density function of the multivariate $t$-distribution with mean vector $\mu$, shape matrix $\Sigma$, and the degrees of freedom $\nu$ is

$$f(x; \mu, \Sigma, \nu) = \frac{\Gamma((\nu + p)/2)|\Sigma|^{-1/2}}{(\pi \nu)^{1/2} \Gamma(\nu/2) \{1 + (x - \mu)^T \Sigma^{-1}(x - \mu)/\nu\}^{(\nu + p)/2}}.$$

With a simple calculation of the EM algorithm (Finegold and Drton, 2011), we can obtain the following iterative algorithm:

$$\mu^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} u_i^{(t)} x_i$$

$$\Omega^{(t+1)} = \text{arg min}_\Omega \left\{ -\log |\Omega| + \text{tr} \left( \Omega S_{u_i^{(t)}}(\mu^{(t+1)}) \right) + \lambda \|\Omega - \text{diag} \Omega\|_1 \right\},$$

where

$$S_{u_i^{(t)}}(\mu) = \frac{1}{n} \sum_{i=1}^{n} u_i^{(t)} (x_i - \mu)(x_i - \mu)^T,$$

and

$$u_i^{(t)} = \frac{\hat{u}_i^{(t)}}{\sum_{j=1}^{n} \hat{u}_j^{(t)}}, \quad \hat{u}_i^{(t)} = \frac{\nu + p}{\nu + (x_i - \mu^{(t)})^T \Omega^{(t)}(x_i - \mu^{(t)})}.$$

The EM algorithm in the multivariate $t$-distribution turns out to be the problem of the weighted graphical lasso.

We recall that the $\gamma$-lasso is also based on the weighted graphical lasso. However, there are two significant differences between the EM algorithm and our algorithm as follows:

- In the EM algorithm, the first term of the complete-data log-likelihood function is $\frac{1}{2} \log \Omega$, whereas in the $\gamma$-lasso, the corresponding term is $\frac{1}{2(1+\gamma)} \log \Omega$.
- The formula of the weight implies that the $\gamma$-divergence is more robust than the multivariate $t$-distribution. For example, suppose that the $i$th observation is an outlier. As $\|x_i\| \to \infty$, we expect that the estimator is not affected by $x_i$. In the
EM algorithm, although \( u_i^{(t)} \rightarrow 0 \) as \( \|x_i\| \rightarrow \infty \), \( u_i^{(t)}(x_i - \hat{\mu}^{(t)})(x_i - \hat{\mu}^{(t)})^T \) does not become zero as \( \|x_i\| \rightarrow \infty \). This means the \( S_{u^{(t)}}(\hat{\mu}^{(t)}) \) is sensitive to the outlier when \( \|x_i\| \) is large. On the other hand, for the \( \gamma \)-divergence, \( w_i^{(t)}(x_i - \hat{\mu}^{(t)})(x_i - \hat{\mu}^{(t)})^T \) becomes zero as \( \|x_i\| \rightarrow \infty \). It is shown the tlasso does not have the redescending property. The kernel function of the estimating equation for tlasso is given by

\[
\psi(x; \theta) = -\frac{\partial}{\partial \theta} \log f(x; \mu, \Sigma, \nu) + \frac{\lambda}{2} u.
\]

In general, \( \frac{\partial}{\partial \omega_{jk}} \log f(x; \mu, \Sigma, \nu) \neq 0 \) as \( \|x\| \rightarrow \infty \), so that \( \psi(x; \theta) \) does not go to zero as \( \|x\| \rightarrow \infty \).

### 4.3 Nonparanormal

Liu et al. (2009) proposed the nonparanormal, which uses the semiparametric Gaussian copula for estimating the graph. The nonparanormal allows the transformation of non-normal data to normal data, which enables us to weaken the assumption of normality. Let \( h \) be a monotone and differentiable function such that \( h(X) = (h(X_1), \ldots, h(X_p))^T \) is multivariate-normally distributed with mean vector \( \mu \) and the covariance matrix \( \Omega^{-1} \). Liu et al. (2009) showed that when \( h_j(x) = \mu_j + \sqrt{\sigma_{jj}} F_j^{-1}(F_j(x)) \), \( X_i \) and \( X_j \) is conditionally independent if and only if \( \omega_{ij} = 0 \), where \( F_j(x) \) is the cumulative distribution function (CDF) of the marginal distribution of \( h(X) \), \( \Phi \) is the CDF of the standard normal distribution, and \( \sigma_{jj} \) is the \((j,j)\)th element of \( \Omega^{-1} \). Liu et al. (2009) estimated \( h_j(x) \) by \( \hat{h}_j(x) = \hat{\mu}_j + \sqrt{\hat{\sigma}_{jj}} \Phi^{-1}(\hat{F}_j(x)) \), where \( \hat{\mu}_j \) is the sample mean, \( \sqrt{\hat{\sigma}_{jj}} \) is the sample standard deviation, and \( \hat{F}_j \) is the truncated empirical CDF defined as

\[
\hat{F}_j(x) = \begin{cases} 
\delta_n & \text{if } \hat{F}_j(x) < \delta_n \\
\hat{F}_j(x) & \text{if } \delta_n \leq \hat{F}_j(x) \leq 1 - \delta_n \\
1 - \delta_n & \text{if } \hat{F}_j(x) > 1 - \delta_n
\end{cases}
\]

Here \( \hat{F}_j(x) \) is the empirical CDF and \( \delta_n \) is a truncation parameter. Liu et al. (2009) selected \( \delta_n = \frac{1}{4n^{1/4} \sqrt{\pi \log n}} \) to achieve the desired rate of convergence in high-dimensional settings.

However, the truncation parameter selected by Liu et al. (2009) may not appropriately treat the outliers:

- We have \( \lim_{n \to \infty} \delta_n = 0 \), which implies \( \delta_n \) is too small to detect the outliers for large samples when the contamination ratio is large.

- The nonparanormal cannot appropriately detect the outliers when the outliers are present only on one side, because the truncation is symmetric. In fact, our simulation
study presented in Section 5 also showed that the nonparanormal did not perform well when the outliers were present on one side.

Our proposed procedure, \( \gamma \)-lasso, does not have any truncation parameter, which implies the \( \gamma \)-lasso does not have an issue as above. Furthermore, the \( \gamma \)-lasso can appropriately treat the outliers even if the contamination ratio is large and/or the outliers are present only on one side, because the weight (9) of the outlier is expected to be sufficiently small.

5 Simulation study

5.1 Simulation model

In this simulation study, we used the following three simulation models:

\[
\begin{align*}
(i) & \quad (1-\varepsilon)N_p(0, \Omega^{-1}) + \varepsilon N_p(0, 30I), \\
(ii) & \quad (1-\varepsilon)N_p(0, \Omega^{-1}) + \varepsilon N_p(\eta_1, I), \\
(iii) & \quad (1-\varepsilon)N_p(0, \Omega^{-1}) + \varepsilon N_p(\eta_1^{(20)}, I),
\end{align*}
\]

where \( I \) is the identity matrix, \( 1 \) is the \( p \)-dimensional vector whose elements are one, \( 1^{(20)} \) is the \( p \)-dimensional vector whose first 20 elements are one and latter \( p - 20 \) elements are zeros, and \( \varepsilon \) (\( 0 \leq \varepsilon < 1 \)) is the contamination ratio. The number of variables was set to be \( p = 100 \). For the model (i), the distribution of outliers are symmetric but away from the central tendency. For the models (ii) and (iii), the outliers are present on one side of the mean direction of \( \eta_1 \) or \( \eta_1^{(20)} \).

We generated the inverse covariance matrix \( \Omega \) in a manner similar to Tan et al. (2014). First, we generated an adjacency matrix \( A = (A_{ij}) \) by the Barabási-Albert model (Barabási and Albert, 1999). Note that the degree distribution of the network generated by the Barabási-Albert model follows power-law. Many real-world networks are often considered as the scale-free networks, in which the degree distribution follows power-law (Barabási and Albert, 1999). Next, we created a matrix \( E = (E_{ij}) \) given by

\[
E_{ij} = \begin{cases} 
U[D] & \text{if } A_{ij} = 1 \\
0 & \text{otherwise}
\end{cases},
\]

where \( U[D] \) is a random sample from a uniform distribution with \( D = [-0.75, -0.25] \cup [0.25, 0.75] \). We calculated \( \hat{E} := (E + E^T)/2 \) and set \( \hat{\Omega} = \hat{E} + (0.1 - \lambda_{\min})I \), where \( \lambda_{\min} \) is the smallest eigenvalue of \( \hat{E} \). Finally, we set \( \Omega = L^{1/2}\hat{\Omega}L^{1/2} \), where \( L = \text{diag}(\Omega^{-1}) \). This procedure guarantees the positive definiteness of the inverse covariance matrix \( \Omega \).
5.2 Monte Carlo simulations

We conducted Monte Carlo simulations to investigate the performance of the proposed procedure. The values of $\varepsilon$ and $\eta$ were set to be $\varepsilon = 0, 0.05, 0.1, 0.3$ and $\eta = 5, 10$, respectively. We generated 100 datasets with $n = 200, 2000, 20000$. The tuning parameters were set to be $\gamma = 0.05, \beta = 0.05$, and $\nu = 1$. In the dp-lasso procedure, we were not able to obtain the solutions for a few datasets because of the non-convergence of the mean vector. We summarized the result without using such datasets for the dp-lasso. We compared (i) ROC curves and (ii) mean squared errors (MSEs) based on the non-diagonal parameters of $\Omega$.

5.2.1 ROC curve

The ROC curves were depicted in Figure 2. We obtain the following tendencies:

- Clearly, the $\gamma$-lasso significantly outperformed for most cases. An important point is that the ROC curves for the $\gamma$-lasso were essentially independent of contamination ratios, which implies that the performance of the $\gamma$-lasso was stable. On the other hand, the existing methods depended on the contamination ratios. For example, in the model (iii), the existing methods performed much worse than the $\gamma$-lasso for large contamination ratio.

- The $t$-lasso showed a similar performance to the $\gamma$-lasso in the model (i) in most cases. However, the $t$-lasso was clearly worse than the $\gamma$-lasso in the models (ii) as the contamination ratio became larger. In the model (iii), the $t$-lasso showed a poor performance. These poor performances may be caused by a symmetric multivariate $t$-distribution with a heavy tail. The symmetric distribution cannot appropriately treat the outliers when they are present on one side.

- The nonparanormal generally performed worse than the $\gamma$-lasso in most cases, especially when the contamination ratio was large. This will be because the truncation parameter of the truncated empirical distribution was selected by $\delta_n = 1 / 4n^{1/4} \sqrt{\pi \log n}$; as $n \to \infty$, $\delta_n \to 0$, which suggests that the outliers may not be detected for large sample sizes (e.g. $\delta_n = 0.0038$ when $n = 20000$).

- When $\varepsilon = 0$, the dp-lasso performed worse than the other methods for large sample sizes. When $n = 200$, the dp-lasso sometimes showed a similar performance to the $\gamma$-lasso.
Figure 2: ROC curves for simulated data. The $x$-axis indicates the number of estimated non-zero non-diagonal elements of $\Omega$ and the $y$-axis is the mean of the true positive rate for the non-zero non-diagonal elements of $\Omega$. 
5.2.2 MSE

The mean squared errors (MSEs) of the inverse covariance matrix $\Omega$ are given in Figure 3. We obtain the following tendencies:

- The $\gamma$-lasso clearly performed the best in most cases. In particular, when $n$ was large, the MSE approached zero as the estimated graph became dense (i.e., the value of $x$-axis in Figure 3 became large). On the other hand, for the other methods, the MSE did not approach zero except for the case of nonparanormal and the standard graphical lasso when $\varepsilon = 0$.

- The $t$-lasso performed well in terms of the ROC curve in the case (i), but did not perform well in terms of the MSE. This will be because the multivariate $t$-distribution with a heavy tail has a very large variance.

- The estimates of the dp-lasso and $t$-lasso were biased even when $\varepsilon = 0$.

- The nonparanormal often showed the second smallest MSE.

6 Gene expression data analyses

6.1 Galactose utilization

The yeast gene expression data were provided by Gasch et al. (2000). We restrict our attention to 8 genes involved in galactose utilization (Ideker et al., 2001). Finegold and Drton (2011) reported that 11 out of the 136 experiments showed unusually large negative values for 4 out of these 8 genes: GAL1, GAL2, GAL7, GAL10. Using the dataset, Vinciotti and Hashem (2013) compared several estimation procedures via $L_1$ penalization, including the standard graphical lasso, the $t$-lasso, and the nonparanormal. They also applied the graph estimation approach given by Meinshausen and Bühlmann (2006), in which the lasso regression (Tibshirani, 1996) is carried out to each variable. Vinciotti and Hashem (2013) applied not only the ordinary lasso regression, but also the adaptive lasso (Zou, 2006), and two typical robust estimation procedures, Least Absolute Deviation and Huber function, with the weighted lasso. The data were normalized to have sample mean 0 and sample deviation 1 before applying the above methods. The tuning parameter $\lambda$ was selected so that the number of edges was 9.

Vinciotti and Hashem (2013) estimated the edges using the original data and the uncontaminated data in which 11 outliers were removed from the original data. Let the original data and the uncontaminated data be denoted by $X$ and $X_{(-11)}$, respectively. A set of edges estimated using $X$ was not always the same as that estimated using $X_{(-11)}$. The difference was examined by the total agreement defined as follows: Let $A$ and $B$
Figure 3: MSEs for simulated data. The $x$-axis indicates the number of estimated non-zero non-diagonal elements of $\Omega$ and the $y$-axis is the mean of MSE on the non-diagonal elements of $\Omega$. 
be the set of edges estimated using $X$ and $X_{(-11)}$, respectively. Let $A^c = C - A$ and $B^c = C - B$, where $C$ is a set of edges of the complete graph. Let $\#D$ be the number of elements in the set $D$. The total agreement is given by $\{(\#(A \cap B) + \#(A^c \cap B^c))/\#C\}$. Vinciotti and Hashemi (2013) reported that the total agreement of the estimated edges was at most 0.86 among various robust estimation procedures described as above. We applied the $\gamma$-lasso with $\gamma = 0.05, 0.1, 0.5$ and dp-lasso with $\beta = 0.05, 0.1, 0.5$, which had not been applied yet. The results are given in Table 1. The $\gamma$-lasso showed that the total agreements were 1 for any $\gamma$, which means that the graph estimated using $X$ were completely same as that estimated using $X_{(-11)}$. Therefore, the $\gamma$-lasso was stable for any $\gamma$. The total agreement of the dp-lasso was also 1 for $\beta = 0.5$, whereas less than 0.8 for $\beta = 0.1$ and $\beta = 0.05$. Thus, the dp-lasso was sensitive to the tuning parameter $\beta$.

We also examined another normalization of the data, because the sample mean and sample deviation are not robust to outliers. The data were normalized with the robust estimates of mean and scale (median and adjusted median absolute deviation (MAD)) before the robust analyses. The results are also given in Table 1. In most cases, the performances based on the MAD were better than those based on the SD. Our procedure yielded the largest total agreement. When the dp-lasso was applied with $\beta = 0.5$, we were not able to find a tuning parameter $\lambda$ whose solution had 9 edges because of the instability of the solution path.

We also depicted the solution paths of the non-diagonal elements of inverse covariance matrix, which are given in Figure 4. Figure 6.1 is the solution path for the graphical lasso applied to $X_{(-11)}$, and the Figures 4(b)-(f) depict the solution paths of the various methods (standard graphical lasso, $\gamma$-lasso, dp-lasso, tlasso, nonparanormal) applied to $X$. Among Figures 4(b)-(f), only Figure 4(c) was similar to Figure 4(a). This means the $\gamma$-lasso was stable against the outliers, whereas the other methods were sensitive to the outliers.

Figure 6.1 showed the weight values of $\gamma$-lasso given by (9). We could find two additional data values whose weight values were enough small to be regarded as outliers. Figure 4(h) depicts the solution path of the graphical lasso applied to the data removing the 13 outliers (original 11 outliers and additional 2 outliers). Figure 4(c) was more similar to Figure 4(h) than Figure 4(a). The $\gamma$-lasso performed as if the 13 outliers were known in advance.

### 6.2 Gene function regulations

Yamada et al. (2014) selected $p = 11$ genes on E.coli with $n = 445$ gene expression levels (Faith et al., 2007), because gene function regulation relationships are well-known, as in Figure 5 (Alberts et al., 2014). The most characteristic point is that two network groups exist independently.
We conducted the principal component analysis to the dataset. The data values were normalized in advance with the robust estimates of mean and scale (median and adjusted median absolute deviation (MAD)). The scores of the first two principal components are plotted in Figure 6. There exist three clusters in the score plot. We see that 91.7% observations belong to a cluster located in the center of the score plot, and the remaining observations can be regarded as outliers, so that the dataset includes many outliers.

Figure 7 presents graphical models estimated by the graphical lasso, nonparanormal, tlasso with $\nu = 1$, the $\gamma$-lasso with $\gamma = 0.05$, and the dp-lasso with $\beta = 0.05$. The tuning parameter $\lambda$ was selected so that the number of edges was 10 and 15. When the number of edges was 10, the graphical lasso produced the edges between two independent groups, but all robust estimation procedures did not. When the number of edges was 15, all methods except for $\gamma$-lasso produced more than one edge connecting between the two groups. Therefore, the $\gamma$-lasso performed the best in terms of estimating fewer cross edges.

### 7 Concluding remarks

We have proposed a robust estimation procedure, $\gamma$-lasso, based on the $\gamma$-divergence for estimating the high-dimensional graphs. The parameter estimation procedure was constructed by using the MM algorithm with the standard algorithm used in the graphical lasso. Extensive simulation studies showed that $\gamma$-lasso performed much better than the
existing methods. Two data analyses illustrated that the $\gamma$-lasso worked well.

The proposed method is based on the $\gamma$-divergence. We can also adopt an extension of the $\gamma$-divergence, Hölder divergence (Kanamori and Fujisawa, 2014). The Hölder divergence allows us to estimate the contamination ratio as well as the model parameters. A distinguishing point is that we can use an iterative minimization algorithm based on the Pythagorean relation for the Hölder divergence, which is the same parameter estimation procedure as in the minimization of $\gamma$-divergence (Kanamori and Fujisawa, 2015). In a similar manner, the proposed parameter estimation procedure based on the MM algorithm is applicable even for the robust sparse Gaussian graphical modeling where the $\gamma$-divergence is replaced by the Hölder divergence.

The covariance estimation based on the likelihood function include a wide variety of statistical models such as the factor analysis, the probabilistic principal component analysis (Tipping and Bishop, 1999), and the canonical correlation analysis. As a future research topic, it is interesting to extend our method to various covariance estimation procedures.

Another important topic is the investigation of the asymptotic properties of the $\gamma$-lasso. In our simulation study in Section 5.2.2, the mean squared error (MSE) of the estimate approached 0 as the number of observations $n$ increased. In fact, Fujisawa and Eguchi (2008) showed that the MSE approaches 0 as $n \to \infty$ even if the contamination ratio is large. However, the authors showed this property only when the number of variables $p$ is fixed and there is no penalization. On the other hand, when no outliers are present, the MSE of the graphical lasso is close to zero when both $n$ and $p$ are sufficiently large (Rothman et al., 2008). Such an asymptotic property in the presence of the outliers is a future issue.

**Acknowledgement**

The authors would like to thank Dr. Hokeun Sun for providing an R code of the dp-lasso.

**Appendix A Derivation of $\ell_2(\theta)$**

We calculate $\ell_2(\theta)$. $\ell_2(\theta)$ can be expressed as

$$
\ell_2(\theta) = \frac{1}{1 + \gamma} \log \int f(x; \theta)^{1+\gamma} dx
= \frac{1}{1 + \gamma} \log \int (2\pi)^{-(1+\gamma)p/2}|\Omega|^{(1+\gamma)/2} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Omega (x - \mu) \right\} dx
$$

(17)
Because the probability of the multivariate normal distribution is one, we have

$$\int \exp \left\{ -\frac{1 + \gamma}{2} (x - \mu)^T \Omega (x - \mu) \right\} dx = (2\pi)^{p/2} (1 + \gamma)^{-p/2} |\Omega|^{-1/2}. \tag{18}$$

Substituting (18) into (17) gives us

$$\ell_2(\theta) = \frac{1}{1 + \gamma} \log \left\{ (2\pi)^{(1+\gamma)p/2} |\Omega|^{(1+\gamma)/2} (2\pi)^{p/2} (1 + \gamma)^{-p/2} |\Omega|^{-1/2} \right\}$$

$$= \frac{1}{1 + \gamma} \log \left\{ (2\pi)^{-p/2} |\Omega|^{1/2} (1 + \gamma)^{-p/2} \right\}$$

$$= \frac{1}{1 + \gamma} \left\{ -\frac{p}{2} \log(2\pi) + \frac{\gamma}{2} \log |\Omega| - \frac{p}{2} \log(1 + \gamma) \right\}.$$

**Appendix B  Determination of \( \lambda_1 \)**

The value of \( \lambda_1 \), which is the minimum value of \( \lambda \) so that all of the non-diagonal elements of inverse covariance matrix are zeros, is easily obtained. When \( \omega_{jk} = 0 \) for any \( j \neq k \), \( \sigma_{jk} = 0 \) as well. Therefore, the variance for each variable is estimated separately. For \( j \)th variable, the mean \( \mu_j \) and variance \( \sigma_{jj} \) are estimated by the iterative algorithm based on Example 4.1 of Fujisawa and Eguchi (2008):

$$\mu_j^{(t+1)} = \sum_{i=1}^{n} w_i^{(t)} x_{ij}, \quad \sigma_{jj}^{(t+1)} = \frac{1}{1 + \gamma} \sum_{i=1}^{n} w_i^{(t)} (x_{ij} - \mu_j^{(t+1)})^2,$$

where \( w_i^{(t)} \) is the weight given by (9). Let \( \hat{\mu} \) and \( \hat{\sigma}_{jj} \) be the estimate of \( \mu \) and \( \sigma_{jj} \) obtained by the above algorithm, respectively. Let the weight in (9) based on \( \hat{\mu} \) and \( \hat{\sigma}_{jj} \) be denoted by \( \hat{w}_i \). We obtain the following weighted sample covariance matrix:

$$S_w = \sum_{i=1}^{n} \hat{w}_i (x_i - \hat{\mu})(x_i - \hat{\mu})^T.$$

The value of \( \lambda_1 \) is then estimated by \( \lambda_1 = \|S_w - \text{Diag}(S_w)\|_\infty \). The basic idea is a necessary and sufficient condition for the solution to the graphical lasso problem given by Corollary 1 of Witten et al. (2011).

**References**

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., and Walter, P. (2014), *Molecular Biology of the Cell (6th edn)*, England: Garland Science.

Barabási, A.-L., and Albert, R. (1999), “Emergence of scaling in random networks,” *Science*, 286(5439), 509–512.
Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. (1998), “Robust and efficient estimation by minimising a density power divergence,” Biometrika, 85(3), 549–559.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011), “Distributed optimization and statistical learning via the alternating direction method of multipliers,” Foundations and Trends® in Machine Learning, 3(1), 1–122.

Cichocki, A., and Amari, S.-i. (2010), “Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities,” Entropy, 12(6), 1532–1568.

Edwards, D. (2000), Introduction to graphical modelling, New York: Springer.

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. (2007), “Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles,” PLoS Biol., 5(1), e8.

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

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

Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J.-B., and Thirion, B. (2012), “Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators,” Medical image analysis, 16(7), 1359–1370.

Fujiisawa, H., and Eguchi, S. (2008), “Robust parameter estimation with a small bias against heavy contamination,” Journal of Multivariate Analysis, 99(9), 2053–2081.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000), “Genomic expression programs in the response of yeast cells to environmental changes,” Molecular biology of the cell, 11(12), 4241–4257.

Hsieh, C.-J., Dhillon, I. S., Ravikumar, P. K., and Sustik, M. A. (2011), Sparse inverse covariance matrix estimation using quadratic approximation,, in Advances in Neural Information Processing Systems, pp. 2330–2338.

Hunter, D. R., and Lange, K. (2004), “A tutorial on MM algorithms,” The American Statistician, 58(1), 30–37.
Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001), “Integrated genomic and proteomic analyses of a systematically perturbed metabolic network,” *Science*, 292(5518), 929–934.

Kanamori, T., and Fujisawa, H. (2014), “Affine invariant divergences associated with proper composite scoring rules and their applications,” *Bernoulli*, 20(4), 2278–2304.

Kanamori, T., and Fujisawa, H. (2015), “Robust estimation under heavy contamination using unnormalized models,” *Biometrika*, 102, 559–572.

Lange, K. (2010), *Numerical analysis for statisticians*, Cambridge: Springer Science & Business Media.

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

Maronna, R., Martin, D., and Yohai, V. (2006), *Robust statistics*, England: John Wiley & Sons, Ltd.

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

Miyamura, M., and Kano, Y. (2006), “Robust Gaussian graphical modeling,” *Journal of Multivariate Analysis*, 97(7), 1525–1550.

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

Raskutti, G., Yu, B., Wainwright, M. J., and Ravikumar, P. K. (2008), Model Selection in Gaussian Graphical Models: High-Dimensional Consistency of $\ell_1$-regularized MLE,, in *Advances in Neural Information Processing Systems*, pp. 1329–1336.

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

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

Tan, K. M., London, P., Mohan, K., Lee, S.-I., Fazel, M., and Witten, D. (2014), “Learning Graphical Models With Hubs,” *Journal of Machine Learning Research*, 15, 3297–3331.

Tibshirani, R. (1996), “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288.
Tipping, M. E., and Bishop, C. M. (1999), “Probabilistic principal component analysis,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3), 611–622.

Tseng, P., and Yun, S. (2009), “A coordinate gradient descent method for nonsmooth separable minimization,” *Mathematical Programming*, 117(1-2), 387–423.

Vinciotti, V., and Hashem, H. (2013), “Robust methods for inferring sparse network structures,” *Computational Statistics & Data Analysis*, 67, 84–94.

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

Yamada, M., Sugiyama, M., and Sese, J. (2014), “Least-squares independence regression for non-linear causal inference under non-Gaussian noise,” *Machine learning*, 96(3), 249–267.

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

Zou, H. (2006), “The adaptive lasso and its oracle properties,” *Journal of the American Statistical Association*, 101(476), 1418–1429.
Figure 4: The solution paths (estimates of $\omega_{ij}$ ($i, j = 1, \ldots, p$, $i < j$) as a function of $\sum_{i<j} |\hat{\omega}_{ij}|$) and the weights of the $\gamma$-lasso in (9) for the yeast gene expression data. (a) The solution path for the graphical lasso applied to $X_{(-11)}$. (b)-(f) The solution paths made by the above methods applied to $X$. (g) The weight values of $\gamma$-lasso given by (9). We could find two additional data values whose weight values were enough small (denoted “+”) to be regarded as outliers in addition to the 11 outliers (denoted “△”). (h) The solution path of the graphical lasso applied to the data removing the 13 outliers (original 11 outliers and additional 2 outliers).
Figure 5: Gene function regulations on *E. coli* [Albers et al. 2014](#).

Figure 6: Score plot of PCA.
Figure 7: Graphical models when the number of edges was 10 and 15.