Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Multivariate Analysis 97 (2006) 1525–1550
Journal of Multivariate Analysis
www.elsevier.com/locate/jmva

Robust Gaussian graphical modeling

Masashi Miyamura*, Yutaka Kano

Division of Mathematical Science, Graduate School of Engineering Science, Osaka University, Japan

Received 13 September 2004
Available online 5 May 2006

Abstract

A new Gaussian graphical modeling that is robustified against possible outliers is proposed. The likelihood function is weighted according to how the observation is deviated, where the deviation of the observation is measured based on its likelihood. Test statistics associated with the robustified estimators are developed. These include statistics for goodness of fit of a model. An outlying score, similar to but more robust than the Mahalanobis distance, is also proposed. The new scores make it easier to identify outlying observations. A Monte Carlo simulation and an analysis of a real data set show that the proposed method works better than ordinary Gaussian graphical modeling and some other robustified multivariate estimators.

AMS 2000 subject classification: 62F35; 62J10

Keywords: Covariance selection; Graphical modeling; Robustness; Weighted maximum likelihood; Hypothesis testing

1. Introduction

Independence is an important concept in statistics. Conditional independence is also important, especially in multivariate analysis. To explore these relationships, graphical modeling [18,9,6] has been developed. Though there exist some types of graphical notations, in this paper we focus on conditional independence described with an undirected graph $G = (V, E)$, where V is a vertex set and E is an edge set. A vertex is associated with a random variable. For further discussion, we introduce some graphical notations. If vertices α and β are connected by an edge, they are said to be *adjacent*. If there is no edge between vertices, they are *non-adjacent*. A *path* is a sequence of distinct vertices that is included in E. For three distinct subsets $A, B, C \subset V$ of the vertex set, C is said to *separate* A from B if every path between A and B includes vertices of C. Assume that

* Corresponding author.
E-mail address: miyamura@sigmath.es.osaka-u.ac.jp (M. Miyamura).

0047-259X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2006.02.006

random variables Y_i $(i \in V)$ have a certain joint probability distribution P. If for any non-adjacent pair α, β,

$$Y_\alpha \perp\!\!\!\perp Y_\beta \mid \boldsymbol{Y}_{V\setminus\{\alpha,\beta\}},$$

then P is said to obey the *pairwise Markov property*. If for any triple A, B, C of disjoint subsets of V, C separates A and B in G, and

$$\boldsymbol{Y}_A \perp\!\!\!\perp \boldsymbol{Y}_B \mid \boldsymbol{Y}_C,$$

then P is said to obey the *global Markov property*. It has been proved that these two Markov properties are equivalent. (See [9] for a more detailed discussion.)

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^T$ be a random p-vector with a covariance matrix $\Sigma = (\sigma_{ij})$. Under Gaussian assumption, conditional independence is equivalent to zero partial correlation coefficient, that is,

$$Y_i \perp\!\!\!\perp Y_j \mid \boldsymbol{Y}_{V\setminus\{i,j\}} \quad \Leftrightarrow \quad \rho_{ij\cdot V\setminus\{i,j\}} = 0,$$

where V is an index set of all variables and $\rho_{ij\cdot V\setminus\{i,j\}}$ is a partial correlation between Y_i and Y_j given the variables $\boldsymbol{Y}_{V\setminus\{i,j\}}$, defined as

$$\rho_{ij\cdot V\setminus\{i,j\}} = -\frac{\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}},$$

where σ^{ij} is the (i, j)th element of the inverse matrix of Σ. Based on this property, Dempster [5] introduced the *covariance selection model* in which certain elements of Σ^{-1} are set to zero. The equivalence of the two Markov properties holds true under Gaussian assumption, which means that any covariance selection model can be described with an undirected graph. In other words, we can explore conditional independence based on partial correlation coefficient from observational data. Then a practical procedure for statistical estimation and evaluation of covariance selection models has been studied by many researchers [6,17,15].

However, collected data often involve several outliers. An outlier is defined as an observation that comes from a population other than a target population. The smaller the sample size, the more serious is the effect of outliers. The existence of outlying observations could lead to wrong analysis. It is well known that a traditional covariance estimator is seriously biased by outliers and thus correlations and partial correlations are also biased. Gaussian graphical modeling is often used for exploring multivariate structures only from observational data without any specific knowledge. In such a situation, biased estimates could lead to wrong models. Some robust covariance estimators, e.g., minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) have been proposed by Rousseeuw [13]. However, no one has derived these types of estimators for a structured covariance matrix and a test statistic for overall goodness of fit, hence, it is impossible to execute Gaussian graphical modeling procedure using these robust methods.

The main objective of the present paper is to improve a Gaussian graphical modeling procedure via robustified maximum likelihood estimation (MLE). Recently, some researchers presented an idea, downweighting observations with their own likelihoods [19,1,7]. We adopted this type of robustified MLE to derive an estimating equation for partial correlations and to construct an algorithm for obtaining estimates and test statistics.

Section 2 proposes a new procedure in detail, together with results concerning asymptotic variance of a robustified estimator. In addition, an appropriate value of the tuning parameter that

controls the degree of robustness is analyzed. In Section 3, we compare the proposed method with other robust methods, MVE and MCD. In Section 4, the proposed procedure is applied to a real data set. This application shows that our robustified procedure can construct an appropriate model even from contaminated data. Our results are summarized in Section 5.

2. Robustifying Gaussian graphical modeling

2.1. Robust maximum likelihood method

Consider a parametric statistical model $\{f(\boldsymbol{y}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ for observations $\{\boldsymbol{y}_i : i = 1, \ldots, n\}$, where $f(\boldsymbol{y}, \boldsymbol{\theta})$ is a probability distribution function and Θ is a parameter space of $\boldsymbol{\theta}$. Let $\Psi(\cdot)$ be a convex, increasing, and differentiable function on R^1.

For given data $\boldsymbol{y}_i$, we define a robustified log likelihood with $\Psi(\cdot)$ as

$$L(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \Psi\left\{\ell(\boldsymbol{y}_i, \boldsymbol{\theta})\right\} - b(\boldsymbol{\theta}) \tag{1}$$

with $\ell(\boldsymbol{y}_i, \theta) = \log f(\boldsymbol{y}_i, \theta)$, where

$$b(\boldsymbol{\theta}) = \int \Psi^* \{\ell(\boldsymbol{y}, \boldsymbol{\theta})\}\, \mathrm{d}\boldsymbol{y}$$

with

$$\Psi^*(z) = \int_0^z \exp(s)\frac{\partial \Psi(s)}{\partial s}\, \mathrm{d}s.$$

An estimator is defined as

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} \{L(\boldsymbol{\theta})\}.$$

Next, we derive an estimating equation. The first differentiation of the objective function (1) to be solved for $\boldsymbol{\theta}$ is

$$\frac{1}{n}\sum_{i=1}^{n} \psi\{\ell(\boldsymbol{y}_i, \boldsymbol{\theta})\} S(\boldsymbol{y}_i, \boldsymbol{\theta}) - \frac{\partial}{\partial \boldsymbol{\theta}} b(\boldsymbol{\theta}) = 0, \tag{2}$$

where

$$\psi(z) = \frac{\partial \Psi(z)}{\partial z} \quad \text{and} \quad S(\boldsymbol{y}, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{y}, \boldsymbol{\theta}).$$

Note that the second component of (2) is the expectation of the first component since

$$\frac{\partial b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \int \psi\{\ell(\boldsymbol{y}, \boldsymbol{\theta})\} S(\boldsymbol{y}, \boldsymbol{\theta}) f(\boldsymbol{y}, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{y} = \mathrm{E}\left[\psi\{\ell(\boldsymbol{y}, \boldsymbol{\theta})\} S(\boldsymbol{y}, \boldsymbol{\theta})\right].$$

Here, exchangeability between integration and differentiation is assumed. This means that the estimating equation (2) is unbiased and thus, the resulting estimator will be consistent for $\boldsymbol{\theta}$.

This method was proposed as Ψ-likelihood by Eguchi and Kano [7]. They defined the Ψ-likelihood as in (1) and showed its asymptotic characteristics.

To derive a specific estimator we choose

$$\Psi_\beta(z) = \begin{cases} \dfrac{\exp(\beta z) - 1}{\beta} & \text{for } \beta > 0, \\ z & \text{for } \beta = 0. \end{cases} \tag{3}$$

Obviously, $\Psi_\beta(\cdot)$ is convex, increasing and differentiable. Modified likelihood (1) with function (3) is termed β-likelihood.

The estimating equation (2) is then described as

$$\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{y}_i, \boldsymbol{\theta})^\beta S(\mathbf{y}_i, \boldsymbol{\theta}) - \mathrm{E}\Big[f(\mathbf{y}, \boldsymbol{\theta})^\beta S(\mathbf{y}, \boldsymbol{\theta})\Big] = 0. \tag{4}$$

We refer to a maximum β-likelihood estimator derived from this equation as a β-estimator and to β as a *robustness tuning parameter*. A larger value of β will result in more robust estimates. When $\beta = 0$, the weight is 1 for every observation, and no robust estimation is made. Indeed, the β-estimator with $\beta = 0$ is nothing but a MLE.

A similar idea of downweighting with respect to the model has also been proposed by Windham [19] and Basu et al. [1]. Windham [19] presented a procedure choosing parameter t such that

$$\frac{\sum_{i=1}^n S_t(y_i) f_t^{(1+\beta)}(y_i)}{\sum_{i=1}^n f_t^\beta(y_i)} = \frac{\int S_t(z) f_t^{(1+\beta)}(z)\,\mathrm{d}z}{\int f_t^{\beta+1}(z)\,\mathrm{d}z}$$

with a tuning parameter β, where $f_t(y)$ is a probability distribution function of y and $S_t(y)$ is its score function. Basu et al. [1] introduced a minimum divergence estimation method. They used *density power divergence* with a parameter β:

$$d_\beta(g, f) = \int \left\{ f^{(\beta+1)}(z) - \left(1 + \frac{1}{\beta}\right) g(z) f^\beta(z) + \frac{1}{\beta} g^{(\beta+1)}(z) \right\} \mathrm{d}z.$$

Jones et al. [8] compared the method of density power divergence with Windham's procedure and concluded that both the procedures show almost the same performance and the former gives a slightly better estimator in a special case.

Minami and Eguchi [12] have already pointed out that the β-divergence is related to the density power divergence as $D_\beta(g, f) = (\beta + 1) d_\beta(g, f)$, and thus, the same estimator is derived from these divergences. We use the β-likelihood procedure to robustify Gaussian graphical modeling.

2.2. Robustifying estimating equation via β-likelihood

Suppose that observations are generated as

$$\boldsymbol{y}_k \sim N(\boldsymbol{\mu}, \Sigma) \quad (k = 1, \ldots, n)$$

independently over k, where $\boldsymbol{y}_k$ is a $p \times 1$ vector, $\boldsymbol{\mu}$ is a $p \times 1$ mean vector and Σ is a $p \times p$ covariance matrix.

A covariance selection model is a model in which some non-diagonal elements of Σ^{-1} are restricted to zero. Let I be an index set of (u, v) that specifies zero elements of Σ^{-1}. I^{C} denotes the complement set of I. Let a_{ij} be an (i, j)th element of $A = \Sigma^{-1}$ and E_{ij} be a matrix having

the (i, j)th element as 1 and the rest 0. A parametrization of Σ^{-1} is then given as

$$\Sigma^{-1} = \sum_{(i,j)\in I^{\mathrm{C}}} a_{ij} E_{ij}. \tag{5}$$

For a saturated model, Eguchi and Kano [7] showed the following proposition on robustified estimators.

Proposition 1. *For a saturated model, estimators of a mean vector* $\boldsymbol{\mu}$ *and a covariance matrix* Σ *satisfy the following estimating equations*:

$$\frac{1}{n}\sum_{k=1}^{n} \exp(\beta z_k)(\boldsymbol{y}_k - \boldsymbol{\mu}) = \mathbf{0} \tag{6}$$

and

$$\frac{1}{n}\sum_{k=1}^{n} \exp(\beta z_k)\left\{\Sigma - (\boldsymbol{y}_k - \boldsymbol{\mu})(\boldsymbol{y}_k - \boldsymbol{\mu})^T\right\} = \frac{\beta}{(\beta+1)^{(p+2)/2}}\Sigma, \tag{7}$$

where

$$z_k = -\tfrac{1}{2}(\boldsymbol{y}_k - \boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{y}_k - \boldsymbol{\mu}).$$

Note that the robustified covariance matrix determined by (7) is positive definite since the β-likelihood $L_\beta(\boldsymbol{\theta})$ goes to negative infinity as an eigenvalue of Σ tends to zero.

In addition, for any covariance selection model, they derived a corresponding result, using differentiations of (5) in terms of a_{ij}.

Proposition 2. *In a covariance selection model, the estimating equation for a mean vector* $\boldsymbol{\mu}$ *is the same as in the saturated model, and the* β*-estimator* $\sigma_{ij}(i, j) \in I^{\mathrm{C}}$ *satisfies the following equation*:

$$\mathrm{tr}[E_{ij}(S - \Sigma)] = 0, \tag{8}$$

where

$$S = \frac{\frac{1}{n}\sum_{k=1}^{n} \exp(\beta z_k)(\boldsymbol{y}_k - \boldsymbol{\mu})(\boldsymbol{y}_k - \boldsymbol{\mu})^T}{\frac{1}{n}\sum_{k=1}^{n} \exp(\beta z_k) - \frac{\beta}{(\beta+1)^{(p+2)/2}}}.$$

Then robustified parameters $\sigma_{uv}(u, v) \in I$ are to be estimated with an algorithm given in the next section.

2.3. Iterative algorithm

The computation of covariance selection model is carried out by the iterative proportional fitting (IPF) algorithm of Speed and Kiiveri [15]. This algorithm is designed to solve the following problem. Given positive definite matrices G and H, find a positive definite matrix F such that

$$[F]_{ij} = [G]_{ij} \quad (i, j) \in I^{\mathrm{C}},$$
$$[F^{-1}]_{uv} = [H]_{uv} \quad (u, v) \in I.$$

Table 1
Algorithm for the robustified Gaussian graphical modeling

Given $\widehat{\boldsymbol{\mu}}^{(m)}, \widehat{\Sigma}^{(m)}, \boldsymbol{z}^{(m)}$. Assume $(u, v) \in I$, $(i, j) \in I^{\mathrm{C}}$. For all element of I, repeat the procedure below.

(1) Calculate $\widehat{\boldsymbol{\mu}}^{(m+1)}$ from (6).
(2) Calculate $\widehat{\Sigma}^{(m+1)}$.
  (a) Calculate $\widehat{\sigma}_{ij}^{(m+1)}$ from (8).
  (b) Compute complete matrix $\widehat{\Sigma}^{(m+1)}$ by IPF.
(3) Calculate $\boldsymbol{z}^{(m+1)}$.
(4) Repeat this procedure from 1 until Σ converges.

As described in the previous section, our robustified covariance matrix is positive definite. Hence, substituting G and H with $\widehat{\Sigma}$ and Σ^{-1}, respectively, we can obtain a complete robustified covariance matrix estimator for any covariance selection model. Table 1 presents an iterative algorithm for robustified Gaussian graphical modeling. Although we have not yet shown the convergence of this algorithm analytically, a numerical convergence is observed.

2.4. Test statistics

In conventional Gaussian graphical modeling, we can use the naive test or the z-transformed test for the hypothesis that a partial correlation coefficient equals zero, and the deviance test procedure for the overall fit. In our robustified method, however, the β-estimator for a partial correlation has a complicated distribution and thus it is difficult to derive exact test statistics. We count on asymptotic theory to obtain approximate distribution of test statistics.

We shall here introduce the vec-operators and the duplication matrices. Let $\mathrm{vec}(A)$ of a $p \times p$ matrix A denote a $p^2 \times 1$ vector formed from stacking column vectors of A and $\mathrm{v}(A)$ denote a $p^* \times 1$ vector formed from all elements of lower triangular part of A including diagonals (thus $p^* = p(p+1)/2$). The duplication matrix D_p is defined by the relation:

$$D_p \mathrm{v}(A) = \mathrm{vec}(A)$$

for any symmetric matrix A of order p. The Kronecker product $A \otimes B$ of matrices A and B is defined as a partitioned matrix with the (i, j)th block equal to $a_{ij}B$ (see, e.g. [10]).

Rewrite the estimating equation (4) as

$$\frac{1}{n}\sum_{i=1}^{n} h_{\boldsymbol{\theta}}(\boldsymbol{y}_i) = \boldsymbol{0}, \tag{9}$$

where $\boldsymbol{\theta}$ is a $p^* \times 1$ parameter vector that consists of non-duplicated elements of Σ^{-1}, and

$$h_{\boldsymbol{\theta}}(\boldsymbol{y}_i) = D_p^{+}\,\mathrm{vec}\left[\exp(\beta z_i)\left\{\Sigma^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^T\Sigma^{-1} - \Sigma^{-1}\right\} + \frac{\beta}{(\beta+1)^{(p+2)/2}}\Sigma^{-1}\right]$$

with

$$z_i = -\tfrac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}).$$

Here $D_p^{+}(= (D_p^T D_p)^{-1} D_p^T)$ denotes the Moore–Penrose generalized inverse of D_p.

Standard asymptotic theory shows that the asymptotic variance of $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is

$$\mathrm{E}\left[h_{\boldsymbol{\theta}\boldsymbol{\theta}^T}\right]^{-1} \mathrm{V}\left[h_{\boldsymbol{\theta}}\right] \mathrm{E}\left[h_{\boldsymbol{\theta}\boldsymbol{\theta}^T}\right]^{-1}. \tag{10}$$

In a covariance selection model, some elements of $\boldsymbol{\theta}$ are restricted to zero and they can be generally represented as $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{u})$ with a $q \times 1$ vector $\boldsymbol{u}$. Let Δ be a $p \times q$ matrix defined as

$$\Delta = \frac{\partial \boldsymbol{\theta}(\boldsymbol{u})}{\partial \boldsymbol{u}^T}.$$

The asymptotic covariance matrix of $\sqrt{n}(\widehat{\boldsymbol{u}} - \boldsymbol{u})$ is then expressible in the form

$$\left(\Delta^T \mathrm{E}\left[h_{\boldsymbol{\theta}\boldsymbol{\theta}^T}\right] \Delta\right)^{-1} \left(\Delta^T \mathrm{V}\left[h_{\boldsymbol{\theta}}\right] \Delta\right) \left(\Delta^T \mathrm{E}\left[h_{\boldsymbol{\theta}\boldsymbol{\theta}^T}\right] \Delta\right)^{-1}. \tag{11}$$

First, we obtain

$$\mathrm{V}\left[h_{\boldsymbol{\theta}}\right] = D_p^+ (\Sigma^{-1/2} \otimes \Sigma^{-1/2}) J (\Sigma^{-1/2} \otimes \Sigma^{-1/2}) (D_p^+)^T,$$

where J is a $p^2 \times p^2$ matrix. See Appendix A for a concrete representation of J. Second, we have

$$\mathrm{E}\left[h_{\boldsymbol{\theta}\boldsymbol{\theta}^T}\right] = D_p^+ \left[(\Sigma^{-1/2} \otimes \Sigma^{-1/2}) \left\{ K + \frac{\beta}{2(\beta+1)^{(p+2)/2}} \mathrm{vec}(I_p)\mathrm{vec}(I_p)^T \right\} (\Sigma^{1/2} \otimes \Sigma^{1/2}) + \frac{1}{(\beta+1)^{(p+2)/2}} (I_p \otimes I_p) \right] D_p,$$

where K is a $p^2 \times p^2$ matrix. See Appendix A for a concrete representation of K.

2.4.1. Statistical test concerning parameter restrictions

Based on the robustified estimate $\widehat{\sigma}^{ij} (i \neq j)$, we test the hypothesis $\sigma^{ij} = 0$ against the alternative hypothesis $\sigma^{ij} \neq 0$. For this purpose, we derive

$$z = \frac{\widehat{\sigma}^{ij}}{\sqrt{\frac{\widehat{\text{Asy-V}}(\widehat{\sigma}^{ij})}{n}}}, \tag{12}$$

where $\text{Asy-V}(\hat{\sigma}_{ij})$ denotes the asymptotic variance of $\widehat{\sigma}_{ij}$ obtained from (10). Then z is distributed asymptotically according to the standard normal distribution. If $|z| > 1.96$, we conclude that the null hypothesis should be rejected. When some elements of Σ^{-1} are restricted to zero, one can use the asymptotic variance (11) to form a z-test as in (12).

2.4.2. Testing a model fit

We construct a measure for overall goodness of fit of a model. Denote

$$\Gamma = \mathrm{E}[h_{\boldsymbol{\theta}\boldsymbol{\theta}^T}]^{-1} \mathrm{V}[h_{\boldsymbol{\theta}}] \mathrm{E}[h_{\boldsymbol{\theta}\boldsymbol{\theta}^T}]^{-1}.$$

We test the following hypothesis:

$$\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{u}) \quad \text{versus} \quad \mathrm{H}_1 : \boldsymbol{\theta} \text{ has no structure.}$$

For this, a test statistic T is defined as

$$T = n \left\{\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\widehat{\boldsymbol{u}})\right\}^T \left\{\widehat{\Gamma}^{-1} - \widehat{\Gamma}^{-1}\widehat{\Delta}(\widehat{\Delta}^T\widehat{\Gamma}^{-1}\widehat{\Delta})^{-1}\widehat{\Delta}^T\widehat{\Gamma}^{-1}\right\} \left\{\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\widehat{\boldsymbol{u}})\right\}$$
$$\left(= n \left\{\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\widehat{\boldsymbol{u}})\right\}^T \widehat{V} \left\{\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\widehat{\boldsymbol{u}})\right\}, \text{ say}\right).$$

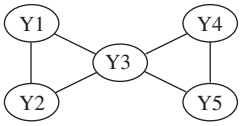

Fig. 1. Butterfly graph.

Since

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\widehat{\boldsymbol{u}})) = \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \sqrt{n}\Delta(\widehat{\boldsymbol{u}} - \boldsymbol{u}) + o_p(1),$$
$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{L} N(\mathbf{0}, \Gamma),$$

we obtain that the null distribution of T can be approximated by a χ^2 variate with degree of freedom $\mathrm{tr}(V\Gamma)$ when n is large enough (see [2]). Thus, one can evaluate a model fit with a χ^2 test.

2.5. Outlying score

With robustified estimates for a mean vector and a covariance matrix, we shall define an *outlying score*

$$(\boldsymbol{y}_k - \widehat{\boldsymbol{\mu}})^T \widehat{\Sigma}^{-1} (\boldsymbol{y}_k - \widehat{\boldsymbol{\mu}}) \tag{13}$$

for each observation. Although this score is similar to the Mahalanobis distance, estimators $\widehat{\boldsymbol{\mu}}$ and $\widehat{\Sigma}$ are robustified. When outliers strongly influence ordinary mean vector and covariance matrix, our outlying score measures the degree of outlying more precisely than the Mahalanobis distance. Note that the weight given to an observation $\boldsymbol{y}_k$ in the estimating equation (4) is a decreasing function of the outlying score.

2.6. Selection of the robustness tuning parameter

The robustness tuning parameter β plays an important role in our method. When $\beta = 0.0$, our method is identical to the traditional MLE method. A larger value of β leads to a more robust estimator, but to the inflation of the variance of a resultant estimator. Concerning this trade-off problem, Basu et al. [1] mentioned "There can be no universal way of selecting an appropriate α parameter when applying our estimation methods." Here, α corresponds to the parameter β in our procedure. Jones et al. [8] also made a similar statement.

To explore an appropriate value of β, we carry out a Monte Carlo simulation. We employ the sample size to be $n = 100$ and generate 1000 data sets. Each data set is generated based on a graph with five vertices (Fig. 1). This is well known as a butterfly graph [18,6]. Variables are generated from a multivariate normal distribution with zero mean vector and a partial correlation matrix in which every value of non-zero partial correlation coefficient is set to 0.25. Outliers are generated from the same multivariate normal distribution except the mean vector and they are mixed with probability 0.05. We make four types of means of outliers: $(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)^T = (0, 0, 0, 0, 0)^T, (1, 1, 1, 1, 1)^T, (2, 2, 2, 2, 2)^T, (3, 3, 3, 3, 3)^T$. This outlying pattern leads to decreasing each partial correlation coefficient, which means that a structure of the graph could be obscured by outliers. Values of β are increased from 0.0 to 1.0 by 0.1. In this setup, we fit a covariance selection model with a butterfly graph and compute mean, variance, and mean squared error (MSE) of $\widehat{\rho}_{12\cdot345}$, the partial correlation coefficient between Y_1 and Y_2 conditioned on the other variables.

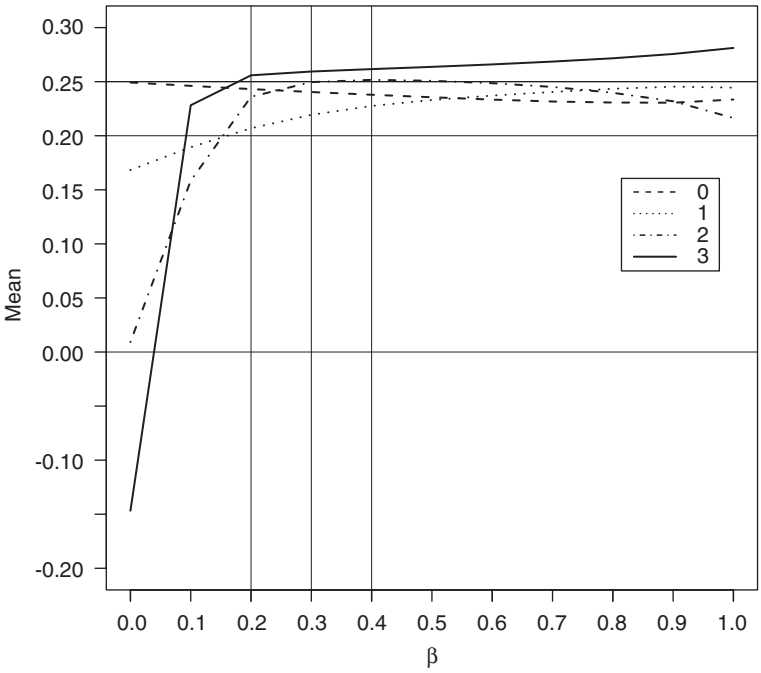

Fig. 2. Mean of $\widehat{\rho}_{12\cdot345}$.

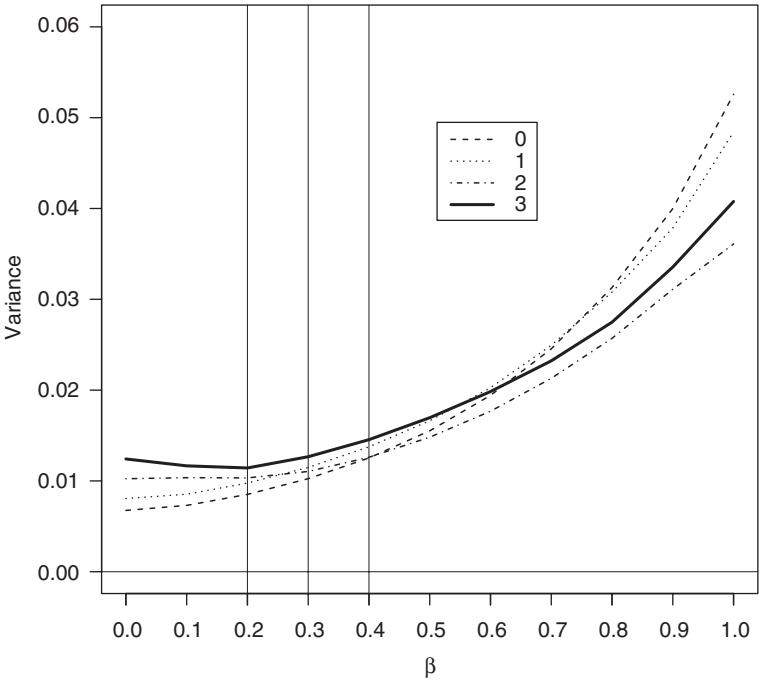

Fig. 3. Variance of $\widehat{\rho}_{12\cdot345}$.

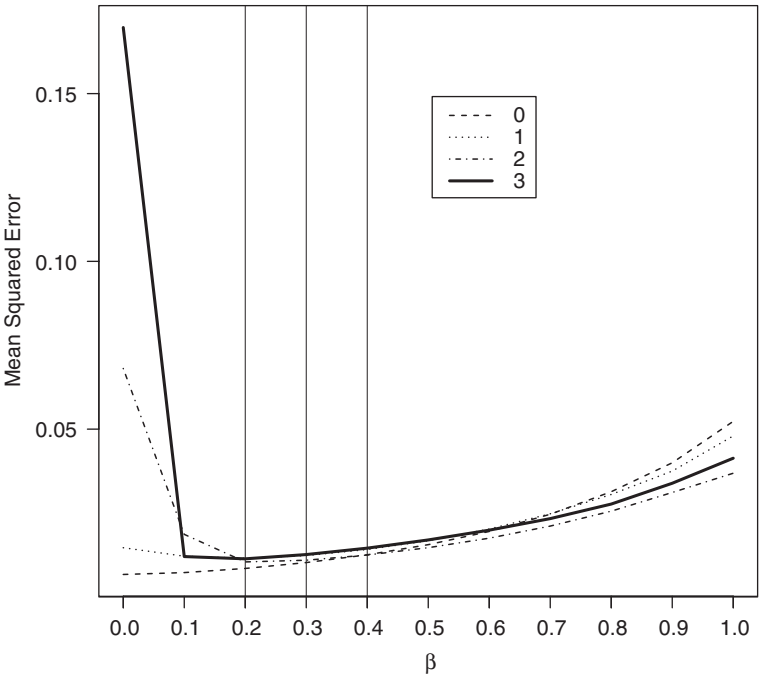

Fig. 4. Mean squared error of $\widehat{\rho}_{12\cdot345}$.

The result is shown in Figs. 2–4. In Fig. 2, the means of estimates are plotted for $\beta = 0.0, \ldots, 1.0$ for each outlying pattern. When $\beta = 0.0$ (not robustified), the bias of estimates becomes more serious as the outlier departs from the origin. It is seen that use of larger values of β reduces the biases substantially. At $\beta = 0.3$ or thereabouts, the biases of estimates are 0.05 or less for all outlying patterns. The variance of estimates $\widehat{\rho}_{12\cdot345}$ in Fig. 3 increases as β becomes larger. MSEs (Fig. 4) achieve smaller values at $\beta = 0.2$ or 0.3 in most patterns. The correction of biases decreases MSE for $\beta = 0.0$ to 0.3, whereas the lack of stability increases MSE for $\beta > 0.3$.

In this setting, i.e., a normal population contaminated with a few outliers, a better choice of a robustness tuning parameter β appears to be 0.3 or thereabouts.

3. Comparison with other robust methods

There are many other robust alternatives for the estimation of a mean vector and a covariance matrix in multivariate analysis [11]. One of the most commonly used estimators is the MVE proposed by Rousseeuw [13]. However, MVE estimator is not $\sqrt{n}$-consistent [4]. In contrast, the MCD estimator [13] has $\sqrt{n}$-consistency. Croux and Haesbroeck [3] gave an influence function of MCD and derived its asymptotic variance. MCD estimator requires heavy computational duty until recently. However, Rousseeuw and Van Driessen [14] proposed a new algorithm to compute MCD, which turns out to be extremely fast, even in high dimensions.

To compare these two alternatives with the proposed method, we conducted a Monte Carlo experiment. The setup is very similar to that in finding an appropriate value of β in the previous section, but the value of β varies from 0.0 to 0.5 by 0.05. For the MCD method, there is the

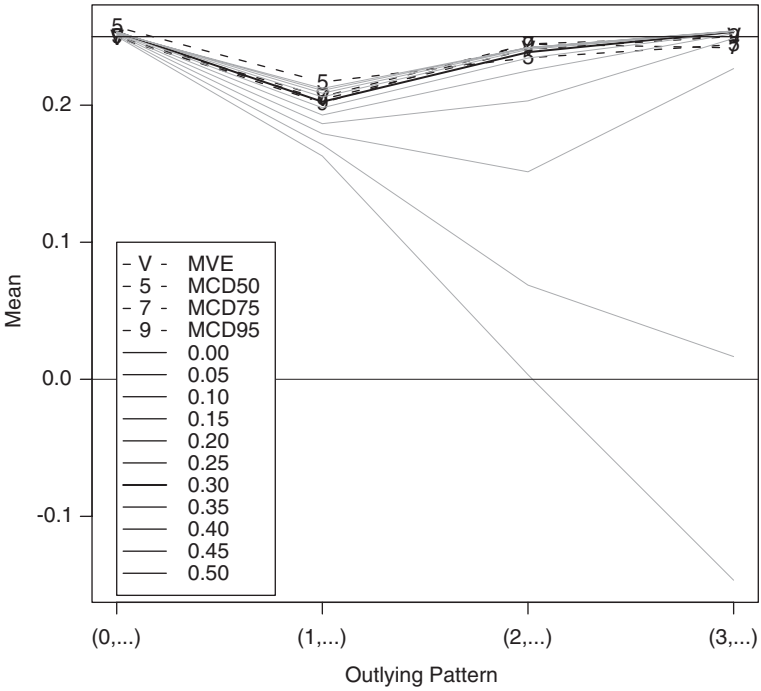

Fig. 5. Mean of $\widehat{\rho}_{12\cdot345}$. Solid lines are the proposed method; the thick solid line is the case $\beta = 0.3$; dashed with one character are alternatives, -V-: MVE, -5-: MCD50; -7-: MCD75; -9-: MCD95.

parameter $0.5 \leqslant \alpha \leqslant 1.0$ which specifies the mass of data determining the MCD. This results in an estimator with breakdown point $(1 - \alpha)$. Theoretically, even for $\alpha = 0.5$, that is, if half of the observations are outliers, MCD estimator holds robustness. However, at the same time, its efficiency is decreasing. Another default value is $\alpha = 0.75$. This value yields a better agreement with efficiency and high breakdown. In this simulation, we adapt three values $\alpha = 0.50$ (MCD50), 0.75 (MCD75), and 0.95 (MCD95). The third value corresponds to the known mass of outliers in this setup.

As in the previous section, we estimate a covariance selection model represented with a butterfly graph. Since there is no procedure to estimate a restricted covariance matrix for alternatives, we start the IPF algorithm with each robustified estimator. We calculate means, variances and MSEs of $\widehat{\rho}_{12\cdot345}$.

The results are shown in Figs. 5–8 First, Fig. 5 shows that alternatives works as efficiently as our method with a larger value of β. Our method has $\beta = 0.3$ or above and the alternatives attain a partial correlation coefficient which is larger than 0.20. Variances in Fig. 6 indicate that our method has relatively small variance when compared with the other methods. The result of MSE (Fig. 7) shows that our method with inappropriate tuning parameters does not work. Fig. 8 is an enlarged graph of MSE. Our method with $\beta = 0.30$ (the thick solid line) outperforms MVE, MCD50, and MCD75. For heavy outlying patterns, MCD95 shows the best performance. It should be noted, however, that the percentage of outlying cannot be known in practice.

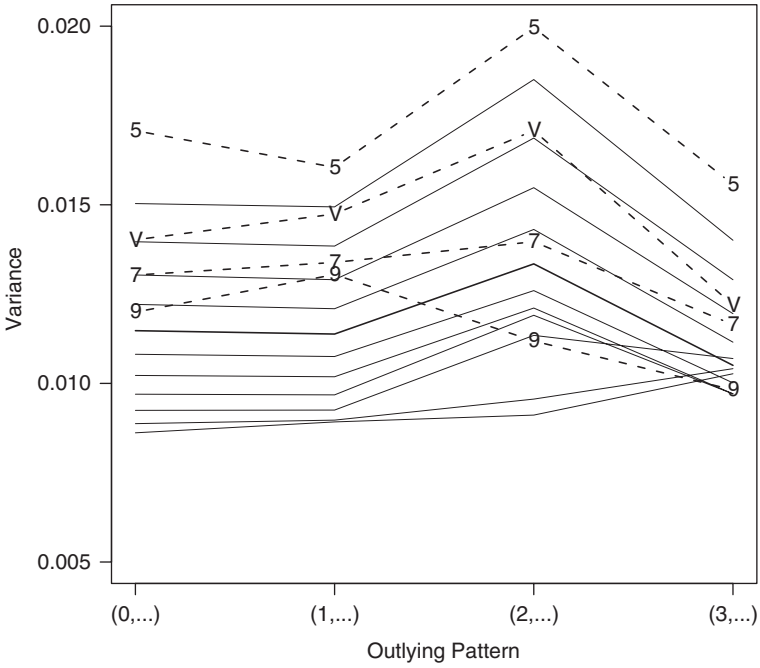

Fig. 6. Variance of $\widehat{\rho}_{12\cdot345}$. Solid lines are the proposed method; the thick solid line is the case $\beta = 0.3$; the dashed lines with one character are alternatives, -V-: MVE, -5-: MCD50; -7-: MCD75; -9-: MCD95.

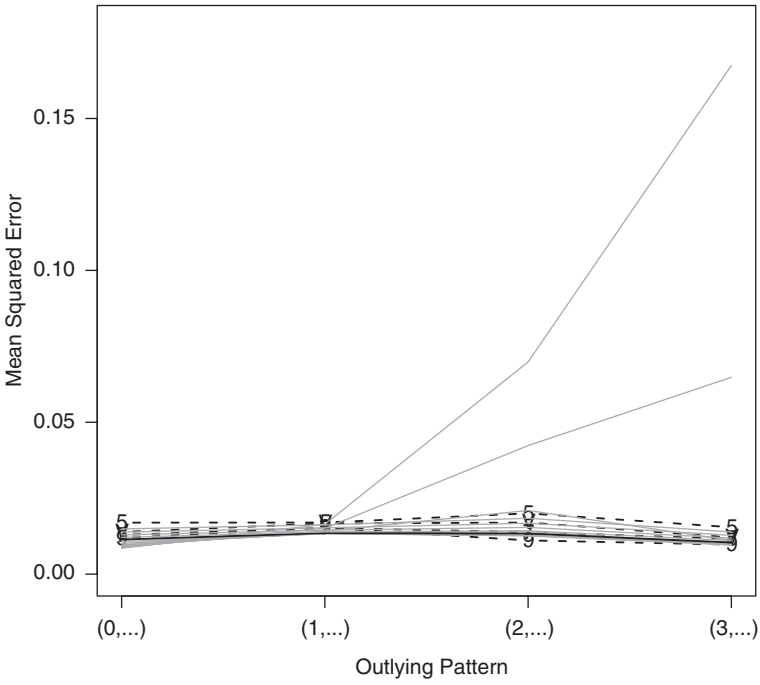

Fig. 7. Mean squared error of $\widehat{\rho}_{12\cdot345}$. $\beta = 0.00$ and 0.05 are inappropriate for seriously biased outliers.

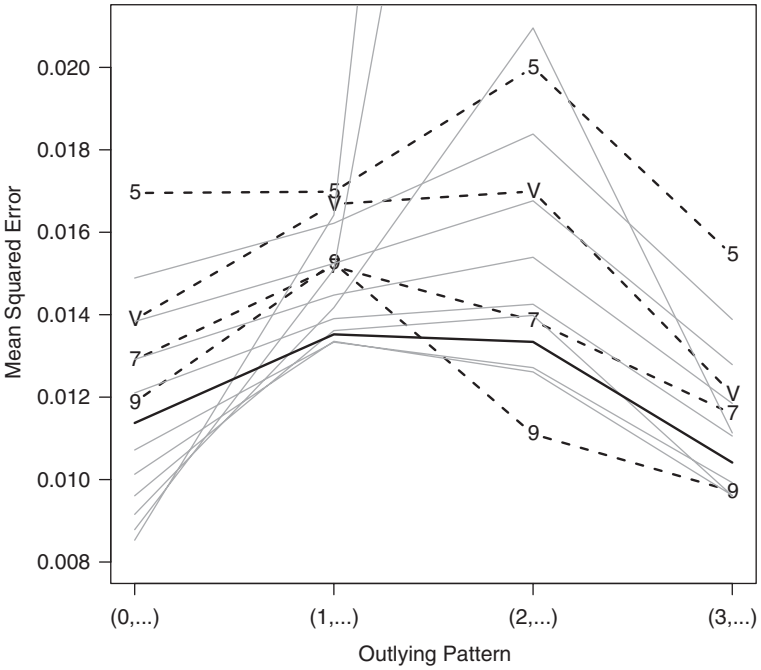

Fig. 8. Mean squared error of $\widehat{\rho}_{12\cdot345}$. Solid lines are values of the proposed method; the thick solid line is the case $\beta = 0.3$; the dotted lines with one character are alternatives, -V-: MVE, -5-: MCD50; -7-: MCD75; -9-: MCD95.

In summary, this experiment shows that (i) since MSE of MVE and MCD50 are larger than the others, these two methods are not attractive; (ii) MCD75 and MCD95 will attain a proper balance between robustness and efficiency; and (iii) the proposed method with an appropriate value of the tuning parameter β (e.g., $\beta = 0.3$) is better than these alternatives in terms of MSE.

4. Family height data

In this section, we consider a real example having three variables that consist of heights of a Son (S), his Father (F), and his Mother (M). The sample size of the data is 126.

Earlier studies [16] have confirmed that there is hereditary effect of height; it is especially stronger between a child and the same sex parent. Our aim of this analysis is to model relationships between sons and parents.

First, we start from the standard (not robustified) Gaussian graphical modeling procedure, namely backward stepwise selection described in [6]. Computed sample partial correlations are shown in Table 2. The minimum estimate is 0.154, which is the partial correlation between Son and Mother. The value does not indicate significance at $\alpha = 0.05$, applying the t-distributed statistics and z-transformed one. Next, we adopt $\beta = 0.30$ following the previous simulation result and analyze the data set with the proposed robustified method. Table 3 shows the result. The robustified estimates show that the minimum value is given at partial correlation between Mother and Father and the estimate between Mother and Son is significant. Model selection procedure based on the deviance test using the standard estimates and on the proposed test statistics using the robustified

Table 2
Sample partial coefficients

	S	F	M
S	1.000	0.356	0.228
F	0.317	1.000	0.249
	3.707	t-value	
	3.397	z-value	
M	0.154	0.184	1.000
	1.729	2.076	t-value
	1.711	2.034	z-value

Correlations (upper) and partial correlations (lower) with t- and z-values ($n = 126$).

Table 3
Robustified estimates ($\beta = 0.30$)

	S	F	M
S	1.000	0.492	0.332
F	0.471	1.000	0.160
	4.334	z-value	
M	0.295	−0.004	1.000
	2.885	−0.045	z-value

Correlations (upper) and partial correlations (lower) with z-values ($n = 126$).

Table 4
Overall goodness of fit

Model	Test Stat.	df	p-value
Deviance test			
S–F–M	3.035	1	0.081
F–S–M	4.352	1	0.037
S–F	11.100	2	0.004
$\beta = 0.3$			
S–F–M	8.321	1	0.004
F–S–M	0.002	1	0.964
S–F	10.319	2	0.006

Results for "deviance test" are based on the ordinary deviance value. Results for "$\beta = 0.3$" are based on the proposed test statistics.

estimates are carried out. The results are summarized in Table 4. Using the conventional procedure, Son–Father–Mother model (Fig. 9) is accepted. On the other hand, the robustified procedure leads to Father–Son–Mother model (Fig. 10). Table 5 shows robustified estimates of this model. The result derived from the robustified procedure shows two things: (i) the hereditary effect between a son and parents and (ii) heights of parents are independent. We would say that the robustified method derives a reasonable result.

Table 5
Robustified estimates ($\beta = 0.30$) of the selected model

	S	F	M
S	1.000	0.492	0.332
F	0.470 4.318	1.000 z-value	0.000
M	0.293 2.862	0.000 –	1.000 z-value

Correlations (upper) and Partial correlations (lower) with z-values ($n = 126$).

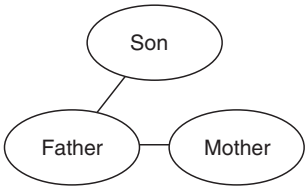

Fig. 9. Son–Father–Mother model.

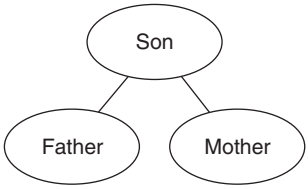

Fig. 10. Father–Son–Mother model.

To examine these different results, we compute the proposed outlying score (Table 6). Then ID29 is identified as the most outlying observation. Here, the son's height is 159 cm whereas his father and mother are 191 and 176 cm, respectively, that is, the tallest parents have a short son. Thus, the regressions on the son of parents are not fitted, which is shown in the scatter plot of residuals (Fig. 11). ID29 is a strong outlier and he influences the ordinary estimation and deviance for the overall goodness of fit.

We analyze this data set again, but without ID29. Sample partial coefficients (Table 7) are similar to the robustified estimates. The result of model selection (Table 8) shows that Father–Son–Mother model is better than Son–Father–Mother model.

To summarize this analysis, the ordinary procedure is hardly affected by the outlying ID29 so that it models the wrong relationships. On the other hand, the robustified method succeeds to model a reasonable relationship, Father–Son–Mother model.

5. Discussion

In this paper, we proposed robustified Gaussian graphical modeling procedure and showed that it has two advantages when compared with some alternative robust estimators. The first is

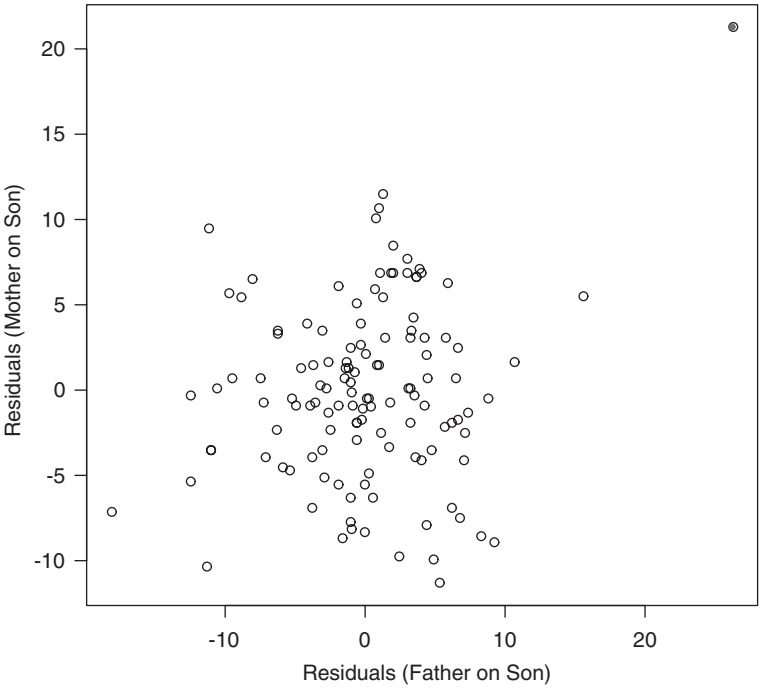

Fig. 11. Scatter plot of residuals. Dotted observation (●) is ID29.

Table 6
Outlying scores

ID	Score	Son	Father	Mother
29	10.32	159	191	176
102	3.56	158	180	160
77	3.03	163	155	165
⋮	⋮	⋮	⋮	⋮
114	0.03	173	170	157
46	0.03	170	168	158
4	0.01	171	169	156

Observations are sorted by the outlying score in decreasing order. The last three columns represent raw data. Three observations with the highest and lowest values are listed.

capability of estimating parameters with $\sqrt{n}$-consistency based on both a saturated model and a restricted model, and doing statistical test of those models. As we did in Section 3, it is possible to carry out the IPF algorithm with robustified covariance matrix by MVE or MCD. However, MVE does not have $\sqrt{n}$-consistency and thus deriving its asymptotic variance may be difficult. Though the MCD estimator has $\sqrt{n}$-consistency under a saturated covariance, its asymptotic behavior is not clear under restricted covariance structures. We have derived an asymptotic variance under both saturated and restricted models for statistical inference. The new method proposed in this paper is thus the only one that can test goodness of fit of restricted models. The second advantage

Table 7
Sample partial coefficients (without ID29)

	S	*F*	*M*
S	1.000	0.439	0.303
F	0.413	1.000	0.168
	5.009	*t*-value	
	4.272	*z*-value	
M	0.259	0.041	1.000
	2.962	0.453	*t*-value
	2.804	0.455	*z*-value

Correlations (upper) and partial correlations (lower) with *t*- and *z*-values ($n = 125$).

Table 8
Deviance test (without ID29)

Model	Deviance	df	*p*-value
S–F–M	8.682	1	0.003
F–S–M	0.207	1	0.649
S–F	12.262	2	0.002

of our estimator is a proper balance between robustness and efficiency. The comparative study in Section 3 shows that MCD with highest breakdown point $\alpha = 0.50$ loses efficiency. An alternative value $\alpha = 0.75$ achieves a better balance between robustness and efficiency. However, in almost all the contamination patterns, its MSE is larger than our procedure. In conclusion, the proposed method will give a better-robustified Gaussian graphical modeling procedure.

Choosing an appropriate value of the robustness tuning parameter β is very important in our procedure. In this paper, we found an appropriate value with a simulation study. However, this value would not be appropriate for some other situations. Other patterns of contamination could lead to different appropriate values of β. We leave this as an open question. Some alternative solutions for choosing β are (i) estimating β as well as other parameters, (ii) estimating a covariance matrix with previously estimated β, and (iii) choosing an appropriate value by evaluating the posterior probabilities of β with any Bayesian approach. However, these methods would require much complicated calculations and a larger sample size to obtain stable results.

Acknowledgments

The authors are grateful to anonymous reviewers for their helpful comments on earlier drafts. This research was supported by the Japan Society for the Promotion of Science (17-9452).

Appendix A. Asymptotic variance of robust estimator

A.1. Variance of $h_{\boldsymbol{\theta}}$

A.1.1. First differential

For $i = 1, \ldots, n$, $\mathbf{y}_i$ is a $p \times 1$ vector distributed according to a multivariate normal with a mean vector $\boldsymbol{\mu}$ and a covariance matrix Σ. The probability density function and its logarithm are

given as

$$f(\boldsymbol{y}_i) = \frac{|\Sigma^{-1}|^{1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})\right\},$$

$$\log f(\boldsymbol{y}_i) = -\frac{p}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma^{-1}| - \frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}).$$

Let $A = \Sigma^{-1}$. We consider differentiation with respect to A. The score function is then expressible as

$$\frac{\partial \log f(\boldsymbol{y}_i)}{\partial A} = \frac{1}{2}A^{-1} - \frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^T. \tag{A.1}$$

We have

$$\mathrm{E}\left[\psi_A(\boldsymbol{y}_i)\frac{1}{2}\left\{A^{-1} - (\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^T\right\}\right] = \frac{1}{2}\mathrm{E}\left[\psi_A(z_i^T z_i)\left\{A^{-1} - A^{-1/2} z_i z_i^T A^{-1/2}\right\}\right]$$
$$= \frac{1}{2}A^{-1/2}\mathrm{E}\left[\psi_A(z_i^T z_i)\left\{I_p - z_i z_i^T\right\}\right]A^{-1/2},$$

where $z_i = A^{1/2}(\boldsymbol{y}_i - \boldsymbol{\mu})$. Note that the first-order moment of z_i is a null vector and z_i is independently distributed, and then we obtain

$$\mathrm{E}\left[z_i z_i^T\right] = \frac{\mathrm{E}\left[z_i^T z_i\right]}{p} \times I_p. \tag{A.2}$$

It follows that

$$\mathrm{E}\left[\psi_A(\boldsymbol{y}_i)\left\{A - A(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^T A\right\}\right] = \frac{1}{p}\mathrm{E}\left[\psi_A(z_i)\left\{p - z_i^T z_i\right\}\right]A.$$

Assuming $\psi(x) = \exp(\beta x)$, we obtain

$$\frac{1}{p}\mathrm{E}\left[\psi_A(z_i)\left\{p - z_i^T z_i\right\}\right]$$
$$= \frac{1}{p}\int \psi_\beta\{\log f(z_i)\}(p - z_i^T z_i)\frac{1}{(2\pi)^{p/2}}\exp\left(-\frac{1}{2}z_i^T z_i\right)\mathrm{d}z_i$$
$$\propto \frac{1}{p}\int \exp\left\{\beta\left(-\frac{1}{2}z_i^T z_i\right)\right\}(p - z_i^T z_i)\frac{1}{(2\pi)^{p/2}}\exp\left(-\frac{1}{2}z_i^T z_i\right)\mathrm{d}z_i$$
$$= \frac{1}{p}\int (p - z_i^T z_i)\frac{1}{(2\pi)^{p/2}}\exp\left(-\frac{\beta+1}{2}z_i^T z_i\right)\mathrm{d}z_i$$
$$= \frac{1}{p(\beta+1)^{p/2}}\left\{p - \frac{p}{(\beta+1)}\right\}$$
$$= \frac{\beta}{p(\beta+1)^{(p+2)/2}}.$$

Hence the estimating equation of A is

$$\frac{1}{n}\sum_{j=1}^{n}\exp\left\{-\frac{\beta}{2}(\boldsymbol{y}_j - \boldsymbol{\mu})^T A(\boldsymbol{y}_j - \boldsymbol{\mu})\right\}\left\{A - A(\boldsymbol{y}_j - \boldsymbol{\mu})(\boldsymbol{y}_j - \boldsymbol{\mu})^T A\right\}$$
$$= \frac{\beta}{(\beta+1)^{(p+2)/2}}A. \tag{A.3}$$

It is also expressive as

$$\frac{1}{n}\sum_{i=1}^{n} h_{\boldsymbol{\theta}}(\boldsymbol{y}_i) = \mathbf{0}, \tag{A.4}$$

where

$$h_{\boldsymbol{\theta}}(\boldsymbol{y}_i) = D_p^{+}\mathrm{vec}\left[\exp\left\{-\frac{\beta}{2}(\boldsymbol{y}_i-\boldsymbol{\mu})^T A(\boldsymbol{y}_i-\boldsymbol{\mu})\right\}\left\{A(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^T A - A\right\} + \frac{\beta}{(\beta+1)^{(p+2)/2}}A\right].$$

A.1.2. Variance of $h_{\boldsymbol{\theta}}$

Note that

$$h_{\boldsymbol{\theta}}(\boldsymbol{y}_i) = D_p^{+}\mathrm{vec}\left[A^{1/2}\left\{a(\beta,\boldsymbol{y}_i)\times A^{1/2}(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^T A^{1/2} - b(\beta,\boldsymbol{y}_i)\times I_p\right\}A^{1/2}\right], \tag{A.5}$$

where

$$a(\beta,\boldsymbol{y}_i) = \exp\left\{-\frac{\beta}{2}(\boldsymbol{y}_i-\boldsymbol{\mu})^T A(\boldsymbol{y}_i-\boldsymbol{\mu})\right\}, \tag{A.6}$$

$$b(\beta,\boldsymbol{y}_i) = \exp\left\{-\frac{\beta}{2}(\boldsymbol{y}_i-\boldsymbol{\mu})^T A(\boldsymbol{y}_i-\boldsymbol{\mu})\right\} - \frac{\beta}{(\beta+1)^{(p+2)/2}}. \tag{A.7}$$

First, we derive $\mathrm{V}[h_{\boldsymbol{\theta}}]$. Since $\mathrm{E}[h_{\boldsymbol{\theta}}] = \mathbf{0}$, $\mathrm{V}[h_{\boldsymbol{\theta}}] = \mathrm{E}[h_{\boldsymbol{\theta}} h_{\boldsymbol{\theta}}^T]$.

Consider the following translation $\boldsymbol{y}_i$ to $\boldsymbol{z}_i$,

$$\boldsymbol{z_i} = A^{1/2}(\boldsymbol{y}_i-\boldsymbol{\mu}) \Rightarrow \boldsymbol{z}_i \overset{i.i.d.}{\sim} N(\mathbf{0}, I_p) \quad (i=1,\cdots,n), \tag{A.8}$$

then we write $\boldsymbol{z}$ for $\boldsymbol{z}_i$ for convenience.

Now we have

$$\begin{aligned}&\mathrm{V}[h_{\boldsymbol{\theta}}]\\ &= D_p^{+}(A^{1/2}\otimes A^{1/2})\\ &\quad\times\mathrm{E}\left[\mathrm{vec}\left\{a(\beta,z^Tz)zz^T - b(\beta,z^Tz)I_p\right\}\mathrm{vec}\left\{a(\beta,z^Tz)zz^T - b(\beta,z^Tz)I_p\right\}^T\right]\\ &\quad\times(A^{1/2}\otimes A^{1/2})\left(D_p^{+}\right)^T.\end{aligned} \tag{A.9}$$

Next, we calculate the expectation component in (A.9), that is,

$$\begin{aligned}&\mathrm{E}\left[\mathrm{vec}\left\{a(\beta,z^Tz)zz^T - b(\beta,z^Tz)I_p\right\}\mathrm{vec}\left\{a(\beta,z^Tz)zz^T - b(\beta,z^Tz)I_p\right\}^T\right]\\ &= \int \mathrm{vec}\left\{a(\beta,z^Tz)zz^T - b(\beta,z^Tz)I_p\right\}\mathrm{vec}\left\{a(\beta,z^Tz)zz^T - b(\beta,z^Tz)I_p\right\}^T\\ &\quad\times N(\boldsymbol{z}\mid\mathbf{0},I_p)\,\mathrm{d}\boldsymbol{z}.\end{aligned}$$

Since the first-order moment of z is a null vector, we only mention the second- and fourth-order moments:

$$\begin{cases} \textit{Case (i)}: & (az_j^2 - b)^2 & \text{for } j = 1, \ldots, p, \\ \textit{Case (ii)}: & (az_j^2 - b)(az_k^2 - b) & \text{for } j, k = 1, \ldots, p; \quad j \neq k, \\ \textit{Case (iii)}: & (az_j z_k)^2 & \text{for } j, k = 1, \ldots, p; \quad j \neq k. \end{cases}$$

For Case (i), we have

$$\mathrm{E}\left[(az_j^2 - b)^2\right] = \mathrm{E}\left[a^2 z_j^4\right] - 2\mathrm{E}\left[abz_j^2\right] + \mathrm{E}\left[b^2\right]. \tag{A.10}$$

Then the first component in (A.10) is

$$\begin{aligned} &\mathrm{E}\left[a^2 z_j^4\right] \\ &= \int z_j^4 \exp\left(-\frac{2\beta}{2} z_j^2\right) N(z_j \mid 0, 1)\, \mathrm{d}z_j \times \int \exp\left(-\frac{2\beta}{2} \boldsymbol{z}_{-i}^T \boldsymbol{z}_{-j}\right) N(\boldsymbol{z}_{-j} \mid \boldsymbol{0}, I_{p-1})\, \mathrm{d}\boldsymbol{z}_{-j} \\ &= \frac{1}{(2\beta+1)^{1/2}} \int z_j^4 \frac{(2\beta+1)^{1/2}}{(2\pi)^{1/2}} \exp\left(-\frac{2\beta+1}{2} z_j^2\right) \mathrm{d}z_j \\ &\quad \times \frac{1}{(2\beta+1)^{(p-1)/2}} \int \frac{(2\beta+1)^{(p-1)/2}}{(2\pi)^{(p-1)/2}} \exp\left(-\frac{2\beta+1}{2} \boldsymbol{z}_{-j}^T \boldsymbol{z}_{-j}\right) \mathrm{d}\boldsymbol{z}_{-j} \\ &= \frac{1}{(2\beta+1)^{1/2}} \frac{3}{(2\beta+1)^2} \times \frac{1}{(2\beta+1)^{(p-1)/2}} = \frac{3}{(2\beta+1)^{(p+4)/2}}, \end{aligned} \tag{A.11}$$

where $\boldsymbol{z}_{-j} = [z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_p]^T$.

And consider

$$\begin{aligned} ab &= \exp\left(-\frac{\beta}{2} \boldsymbol{z}^T \boldsymbol{z}\right) \left\{ \exp\left(-\frac{\beta}{2} \boldsymbol{z}^T \boldsymbol{z}\right) - \frac{\beta}{(\beta+1)^{(p+2)/2}} \right\} \\ &= \exp\left(-\frac{2\beta}{2} \boldsymbol{z}^T \boldsymbol{z}\right) - \frac{\beta}{(\beta+1)^{(p+2)/2}} \exp\left(\frac{\beta}{2} \boldsymbol{z}^T \boldsymbol{z}\right), \end{aligned}$$

and then the second component in (A.10) is

$$\begin{aligned} &\mathrm{E}\left[abz_j^2\right] \\ &= \int z_j^2 \exp\left(-\frac{2\beta}{2} z_j^2\right) N(z_j \mid 0, 1)\, \mathrm{d}z_j \times \int \exp\left(-\frac{2\beta}{2} \boldsymbol{z}_{-j}^T \boldsymbol{z}_{-j}\right) N(\boldsymbol{z}_{-j} \mid \boldsymbol{0}, I_{p-1})\, \mathrm{d}\boldsymbol{z}_{-j} \\ &\quad - \frac{\beta}{(\beta+1)^{(p+2)/2}} \int z_j^2 \exp\left(-\frac{\beta}{2} z_j^2\right) N(z_j \mid 0, 1)\, \mathrm{d}z_j \\ &\quad \times \int \exp\left(-\frac{\beta}{2} \boldsymbol{z}_{-j}^T \boldsymbol{z}_{-j}\right) N(\boldsymbol{z}_{-j} \mid \boldsymbol{0}, I_{p-1})\, \mathrm{d}\boldsymbol{z}_{-j} \\ &= \frac{1}{(2\beta+1)^{3/2}} \frac{1}{(2\beta+1)^{(p-1)/2}} - \frac{\beta}{(\beta+1)^{(p+2)/2}} \frac{1}{(\beta+1)^{3/2}} \frac{1}{(\beta+1)^{(p-1)/2}} \\ &= \frac{1}{(2\beta+1)^{(p+2)/2}} - \frac{\beta}{(\beta+1)^{p+2}}. \end{aligned} \tag{A.12}$$

Since

$$b^2 = \left\{\exp\left(-\frac{\beta}{2}z^T z\right) - \frac{\beta}{(\beta+1)^{(p+2)/2}}\right\}^2$$
$$= \exp\left(-\frac{2\beta}{2}z^T z\right) - \frac{2\beta}{(\beta+1)^{(p+2)/2}}\exp\left(-\frac{\beta}{2}z^T z\right) + \frac{\beta^2}{(\beta+1)^{p+2}},$$

the last component in (A.10) is

$$\mathrm{E}\left[b^2\right] = \mathrm{E}\left[\exp\left(-\frac{2\beta}{2}z^T z\right)\right] - \frac{2\beta}{(\beta+1)^{(p+2)/2}}\mathrm{E}\left[\exp\left(-\frac{\beta}{2}z^T z\right)\right] + \frac{\beta^2}{(\beta+1)^{p+2}}$$
$$= \frac{1}{(2\beta+1)^{p/2}} - \frac{2\beta}{(\beta+1)^{(p+2)/2}}\cdot\frac{1}{(\beta+1)^{p/2}} + \frac{\beta^2}{(\beta+1)^{p+2}}$$
$$= \frac{1}{(2\beta+1)^{p/2}} - \frac{2\beta(\beta+1)-\beta^2}{(\beta+1)^{p+2}}$$
$$= \frac{1}{(2\beta+1)^{p/2}} - \frac{\beta(\beta+2)}{(\beta+1)^{p+2}}. \tag{A.13}$$

With substitution of (A.11), (A.12) and (A.13) into (A.10), the expectation is obtained as

$$\mathrm{E}\left[\left(a - bz_j^2\right)^2\right]$$
$$= \frac{3}{(2\beta+1)^{(p+4)/2}} - \frac{2}{(2\beta+1)^{(p+2)/2}} + \frac{2\beta}{(\beta+1)^{p+2}} + \frac{1}{(2\beta+1)^{p/2}} - \frac{\beta(\beta+2)}{(\beta+1)^{p+2}}$$
$$= \frac{3-2(2\beta+1)+(2\beta+1)^2}{(2\beta+1)^{(p+4)/2}} - \frac{\beta^2}{(\beta+1)^{p+2}}$$
$$= \frac{4\beta^2+2}{(2\beta+1)^{(p+4)/2}} - \frac{\beta^2}{(\beta+1)^{p+2}}. \tag{A.14}$$

For Case (ii), we have

$$\mathrm{E}\left[(az_j^2 - b)(az_k^2 - b)\right] = \mathrm{E}\left[(az_j z_k)^2\right] - \mathrm{E}\left[abz_j^2\right] - \mathrm{E}\left[abz_k^2\right] + \mathrm{E}\left[b^2\right]. \tag{A.15}$$

The first component is

$$\mathrm{E}\left[(az_j z_k)^2\right]$$
$$= \int \exp\left(-\frac{2\beta}{2}z^T z\right) z_j^2 z_k^2 N(z \mid \mathbf{0}, I_p)\,\mathrm{d}z$$
$$= \int \exp\left(-\frac{2\beta}{2}z_j^2\right) N(z_j \mid 0,1)\,\mathrm{d}z_j \times \int \exp\left(-\frac{2\beta}{2}z_k^2\right) N(z_k \mid 0,1)\,\mathrm{d}z_k$$
$$\times \int \exp\left(-\frac{2\beta}{2}z_{-(j,k)}^T z_{-(j,k)}\right) N(z_{-(j,k)} \mid \mathbf{0}, I_{p-2})\,\mathrm{d}z_{-(j,k)}$$
$$= \frac{1}{(2\beta+1)^{3/2+3/2+(p-2)/2}} = \frac{1}{(2\beta+1)^{(p+4)/2}}. \tag{A.16}$$

The expectation is evaluated by substituting (A.12), (A.13) and (A.16) into (A.15). We have

$$\begin{aligned}\mathrm{E}\left[(az_j^2-b)(az_k^2-b)\right] &= \frac{1}{(2\beta+1)^{(p+4)/2}} - 2\times\left(\frac{1}{(2\beta+1)^{(p+2)/2}} - \frac{\beta}{(\beta+1)^{p+2}}\right)\\ &\quad + \frac{1}{(2\beta+1)^{p/2}} - \frac{\beta(\beta+2)}{(\beta+1)^{p+2}}\\ &= \frac{1-2(2\beta+1)+(2\beta+1)^2}{(2\beta+1)^{(p+4)/2}} - \frac{\beta^2+2\beta-2\beta}{(\beta+1)^{p+2}}\\ &= \frac{4\beta^2}{(2\beta+1)^{(p+4)/2}} - \frac{\beta^2}{(\beta+1)^{p+2}}. \end{aligned} \tag{A.17}$$

Note that Case (iii) is equivalent to (A.16).

As a consequence, the variance of $h_{\boldsymbol{\theta}}$ is a $p^*\times p^*$ matrix with typical elements:

$$\mathrm{E}\left[(az_j^2-b)^2\right] = \frac{4\beta^2+2}{(2\beta+1)^{(p+4)/2}} - \frac{\beta^2}{(\beta+1)^{p+2}}, \tag{A.18}$$

$$\mathrm{E}\left[(az_j^2-b)(az_k^2-b)\right] = \frac{4\beta^2}{(2\beta+1)^{(p+4)/2}} - \frac{\beta^2}{(\beta+1)^{p+2}}, \tag{A.19}$$

$$\mathrm{E}\left[(az_jz_k)^2\right] = \frac{1}{(2\beta+1)^{(p+4)/2}}. \tag{A.20}$$

When $\beta=0$, these three values result in 2, 0 and 1, respectively. This is the well-known result in the conventional maximum likelihood case.

A.2. Expectation of $h_{\boldsymbol{\theta}\boldsymbol{\theta}^T}$

A.2.1. Second differential

The first differentiation is

$$\begin{aligned} h_{\boldsymbol{\theta}}(\mathbf{y}_i) &= D_p^+\,\mathrm{vec}\left[\exp\left\{-\frac{\beta}{2}(\mathbf{y}_i-\boldsymbol{\mu})^TA(\mathbf{y}_i-\boldsymbol{\mu})\right\}\left\{A(\mathbf{y}_i-\boldsymbol{\mu})(\mathbf{y}_i-\boldsymbol{\mu})^TA-A\right\}\right.\\ &\quad \left.+\frac{\beta}{(\beta+1)^{(p+2)/2}}A\right]\\ &= D_p^+\,\mathrm{vec}\left[a(A)\left\{A(\mathbf{y}_i-\boldsymbol{\mu})(\mathbf{y}_i-\boldsymbol{\mu})^TA-A\right\}+bA\right], \end{aligned} \tag{A.21}$$

where

$$a(A) = \exp\left\{-\frac{\beta}{2}(\mathbf{y}_i-\boldsymbol{\mu})^TA(\mathbf{y}_i-\boldsymbol{\mu})\right\},$$

$$b = \frac{\beta}{(\beta+1)^{(p+2)/2}}.$$

Note that $a(A)$ is a scalar function.

Differentiating $h_{\boldsymbol{\theta}}$ with respect to A, we obtain

$$\begin{aligned} \mathrm{d}h_{\boldsymbol{\theta}} &= D_p^+\,\mathrm{d}\,\mathrm{vec}\left[a(A)\left\{A(\mathbf{y}_i-\boldsymbol{\mu})(\mathbf{y}_i-\boldsymbol{\mu})^TA-A\right\}+bA\right]\\ &= D_p^+\left[\mathrm{vec}\left\{A(\mathbf{y}_i-\boldsymbol{\mu})(\mathbf{y}_i-\boldsymbol{\mu})^TA\right\}\{\mathrm{d}a(A)\}+\mathrm{vec}\left[a(A)\,\mathrm{d}\left\{A(\mathbf{y}_i-\boldsymbol{\mu})(\mathbf{y}_i-\boldsymbol{\mu})^TA\right\}\right]\right.\\ &\quad \left.-\mathrm{vec}(A)\{\mathrm{d}a(A)\}-\mathrm{vec}\{a(A)\,\mathrm{d}A\}+b\,\mathrm{vec}(\mathrm{d}A)\right]. \end{aligned} \tag{A.22}$$

First,

$$\begin{aligned} \mathrm{d}a(A) &= \mathrm{d}\exp\left(-\frac{\beta}{2}(\boldsymbol{y}_i-\boldsymbol{\mu})^T A(\boldsymbol{y}_i-\boldsymbol{\mu})\right) \\ &= -\frac{\beta}{2}a(A)(\boldsymbol{y}_i-\boldsymbol{\mu})^T(\mathrm{d}A)(\boldsymbol{y}_i-\boldsymbol{\mu}), \end{aligned} \tag{A.23}$$

then the first component in (A.22) is

$$\begin{aligned} &-\frac{\beta}{2}a(A)\,\mathrm{vec}\left\{A(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^T A\right\}\,\mathrm{vec}\left\{(\boldsymbol{y}_i-\boldsymbol{\mu})(\mathrm{d}A)(\boldsymbol{y}_i-\boldsymbol{\mu})^T\right\} \\ &= -\frac{\beta}{2}a(A)\,\mathrm{vec}\left\{A(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^T A\right\}\left\{(\boldsymbol{y}_i-\boldsymbol{\mu})^T\otimes(\boldsymbol{y}_i-\boldsymbol{\mu})^T\right\}\mathrm{vec}(\mathrm{d}A) \\ &= -\frac{\beta}{2}a(A)\,\mathrm{vec}\left\{A(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^T A\right\}\,\mathrm{vec}\left\{(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^T\right\}^T\,\mathrm{vec}(\mathrm{d}A). \end{aligned}$$

The second component is

$$\begin{aligned} &a(A)\,\mathrm{vec}\left[(\mathrm{d}A)(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^T A\right] + a(A)\,\mathrm{vec}\left[A(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})(\mathrm{d}A)\right] \\ &= a(A)\left\{(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^T A\otimes I_p\right\}\mathrm{vec}(\mathrm{d}A) \\ &\quad + a(A)\left\{I_p\otimes A(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^T\right\}\mathrm{vec}(\mathrm{d}A). \end{aligned}$$

The third component is

$$\begin{aligned} &-\frac{\beta}{2}a(A)\,\mathrm{vec}(A)\,\mathrm{vec}\left\{(\boldsymbol{y}_i-\boldsymbol{\mu})^T(\mathrm{d}A)(\boldsymbol{y}_i-\boldsymbol{\mu})\right\} \\ &= -\frac{\beta}{2}a(A)\,\mathrm{vec}(A)\left\{(\boldsymbol{y}_i-\boldsymbol{\mu})^T\otimes(\boldsymbol{y}_i-\boldsymbol{\mu})^T\right\}\mathrm{vec}(\mathrm{d}A) \\ &= -\frac{\beta}{2}a(A)\,\mathrm{vec}(A)\,\mathrm{vec}\left\{(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^T\right\}^T\,\mathrm{vec}(\mathrm{d}A). \end{aligned}$$

Finally, the fourth component is

$$a(A)(I_p\otimes I_p)\,\mathrm{vec}(\mathrm{d}A).$$

Since $\mathrm{vec}(\mathrm{d}A)=\mathrm{d}\,\mathrm{vec}(A)=D_p\,\mathrm{dv}(A)$, we obtain the second differential with respect to A as

$$h_{\boldsymbol{\theta}\boldsymbol{\theta}^T}=\frac{\partial h_{\boldsymbol{\theta}}}{\partial \mathrm{v}(A)^T}=D_p^+C(\beta,A)D_p, \tag{A.24}$$

where

$$\begin{aligned} C(\beta,A) = &-\frac{\beta}{2}a(A)\,\mathrm{vec}\left\{A(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^T A\right\}\,\mathrm{vec}\left\{(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^T\right\}^T \\ &+\frac{\beta}{2}a(A)\,\mathrm{vec}(A)\,\mathrm{vec}\left\{(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^T\right\}^T + a(A)\left\{(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^T A\otimes I_p\right\} \\ &+a(A)\left\{I_p\otimes A(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^T\right\} - a(A)(I_p\otimes I_p) + b(I_p\otimes I_p). \end{aligned} \tag{A.25}$$

A.2.2. Expectation of $\mathrm{v}(h_{\theta\theta^T})$

Expectation of (A.24) is derived. Since $\mathrm{E}[h_{\theta\theta^T}] = D_p^+ \mathrm{E}[C(\beta, A)]D_p$, it is sufficient to derive the expectation of (A.25).

If $\boldsymbol{y}_i$ is translated to $\boldsymbol{z}_i = A^{1/2}(\boldsymbol{y}_i - \boldsymbol{\mu})$, then $\boldsymbol{z}_i$ is identically independent distributed with standard normal. We thus use $\boldsymbol{z}$ for $\boldsymbol{z}_i$ for convenience.

For the first line in (A.25),

$$-\frac{\beta}{2}\mathrm{E}\left[a(A)\,\mathrm{vec}\left\{A^{1/2}\boldsymbol{z}\boldsymbol{z}^T A^{1/2}\right\}\,\mathrm{vec}\left\{A^{-1/2}\boldsymbol{z}\boldsymbol{z}^T A^{-1/2}\right\}^T\right]$$
$$= (A^{1/2}\otimes A^{1/2})\left\{-\frac{\beta}{2}\mathrm{E}\left[a(\boldsymbol{z}^T\boldsymbol{z})\,\mathrm{vec}\left(\boldsymbol{z}\boldsymbol{z}^T\right)\mathrm{vec}\left(\boldsymbol{z}\boldsymbol{z}^T\right)^T\right]\right\}(A^{-1/2}\otimes A^{-1/2}). \quad \text{(A.26)}$$

Since the first-order moment of $\boldsymbol{z}$ is a null vector, we only mention the fourth- and second-order moments:

$$\begin{cases} \textit{Case (i)}: & z_j^4 \quad \text{for } j = 1, \ldots, p, \\ \textit{Case (ii)}: & z_j^2 z_k^2 \quad \text{for } j, k = 1, \ldots, p; \quad j \neq k. \end{cases}$$

For Case (i), it follows that

$$-\frac{\beta}{2}\mathrm{E}\left[a(\boldsymbol{z}^T\boldsymbol{z})z_j^4\right] = -\frac{\beta}{2}\times\frac{1}{(\beta+1)^{1/2}}\int z_j^4\frac{(\beta+1)^{1/2}}{(2\pi)^{1/2}}\exp\left(-\frac{\beta+1}{2}z_j^2\right)\mathrm{d}z_j$$
$$\times\frac{1}{(\beta+1)^{(p-1)/2}}\int\frac{(\beta+1)^{(p-1)/2}}{(2\pi)^{(p-1)/2}}\exp\left(-\frac{\beta+1}{2}\boldsymbol{z}_{-j}^T\boldsymbol{z}_{-j}\right)\mathrm{d}\boldsymbol{z}_{-j}$$
$$= -\frac{\beta}{2}\times\frac{3}{(\beta+1)^{5/2}}\times\frac{1}{(\beta+1)^{(p-1)/2}} = -\frac{3\beta}{2(\beta+1)^{(p+4)/2}}. \quad \text{(A.27)}$$

And for Case (ii),

$$-\frac{\beta}{2}\mathrm{E}\left[a(\boldsymbol{z}^T\boldsymbol{z})z_j^2 z_k^2\right]$$
$$= -\frac{\beta}{2}\times\frac{1}{(\beta+1)^{1/2}}\int z_j^2\frac{(\beta+1)^{1/2}}{(2\pi)^{1/2}}\exp\left(-\frac{\beta+1}{2}z_j^2\right)\mathrm{d}z_j$$
$$\times\frac{1}{(\beta+1)^{1/2}}\int z_k^2\frac{(\beta+1)^{1/2}}{(2\pi)^{1/2}}\exp\left(-\frac{\beta+1}{2}z_k^2\right)\mathrm{d}z_k$$
$$\times\frac{1}{(\beta+1)^{(p-2)/2}}\int\frac{(\beta+1)^{(p-2)/2}}{(2\pi)^{(p-2)/2}}\exp\left(-\frac{\beta+1}{2}\boldsymbol{z}_{-(i,j)}^T\boldsymbol{z}_{-(i,j)}\right)\mathrm{d}\boldsymbol{z}_{-(i,j)}$$
$$= -\frac{\beta}{2}\times\frac{1}{(\beta+1)^{3/2}}\times\frac{1}{(\beta+1)^{3/2}}\times\frac{1}{(\beta+1)^{(p-2)/2}}$$
$$= -\frac{\beta}{2(\beta+1)^{(p+4)/2}}. \quad \text{(A.28)}$$

The expectation of the second line in (A.25) is

$$\frac{\beta}{2}\mathrm{E}\left[a(\boldsymbol{z}^T\boldsymbol{z})\,\mathrm{vec}(A^{1/2}A^{1/2})\,\mathrm{vec}\left\{A^{-1/2}\boldsymbol{z}\boldsymbol{z}^T A^{-1/2}\right\}^T\right]$$
$$= (A^{1/2}\otimes A^{1/2})\left\{\frac{\beta}{2}\mathrm{E}\left[a(\boldsymbol{z}^T\boldsymbol{z})\,\mathrm{vec}(I_p)\,\mathrm{vec}\left(\boldsymbol{z}\boldsymbol{z}^T\right)^T\right]\right\}(A^{-1/2}\otimes A^{-1/2}). \quad \text{(A.29)}$$

The first moment of z_j $(j = 1, \ldots, p)$ is zero. Hence all we have to do is considering z_j^2:

$$\mathrm{E}[a(z^T z)z_j^2] = \frac{1}{(\beta+1)^{3/2}} \times \frac{1}{(\beta+1)^{(p-1)/2}} = \frac{1}{(\beta+1)^{(p+2)/2}}.$$

Then,

$$\begin{aligned}&\frac{\beta}{2}\mathrm{E}\left[a(z^T z)\,\mathrm{vec}(A^{1/2}A^{1/2})\,\mathrm{vec}\left\{A^{-1/2}zz^T A^{-1/2}\right\}^T\right]\\ &\quad = (A^{1/2}\otimes A^{1/2})\left[\frac{\beta}{2(\beta+1)^{(p+2)/2}}\left\{\mathrm{vec}(I_p)\,\mathrm{vec}(I_p)^T\right\}\right](A^{-1/2}\otimes A^{-1/2}).\end{aligned} \tag{A.30}$$

For the third and fourth lines in (A.25) can be decomposed as like

$$\begin{aligned}&\mathrm{E}\left[a(z^T z)\left\{A^{-1/2}zz^T A^{1/2}\otimes A^{-1/2}A^{1/2}\right\}\right]\\ &\quad = (A^{-1/2}\otimes A^{-1/2})\mathrm{E}\left[a(z^T z)\left(zz^T\otimes I_p\right)\right](A^{1/2}\otimes A^{1/2})\end{aligned}$$

and

$$\begin{aligned}&\mathrm{E}\left[a(z^T z)\left\{A^{1/2}A^{-1/2}\otimes A^{1/2}zz^T A^{-1/2}\right\}\right]\\ &\quad = (A^{1/2}\otimes A^{1/2})\mathrm{E}\left[a(z^T z)\left(I_p\otimes zz^T\right)\right](A^{-1/2}\otimes A^{-1/2}),\end{aligned}$$

respectively. Since both the expectation components are equivalent to that of (A.29), the expectations are

$$\frac{1}{(\beta+1)^{(p+2)/2}}(I_p\otimes I_p). \tag{A.31}$$

It is easy to derive the fifth expectation:

$$\mathrm{E}\left[a(z^T z)\right](I_p\otimes I_p) = \frac{1}{(\beta+1)^{p/2}}(I_p\otimes I_p). \tag{A.32}$$

The sum of $2\times$ (A.29), (A.32), and the expectation of the sixth line in (A.25) is

$$\frac{2-(\beta+1)+\beta}{(\beta+1)^{(p+2)/2}}(I_p\otimes I_p) = \frac{1}{(\beta+1)^{(p+2)/2}}(I_p\otimes I_p). \tag{A.33}$$

We thus obtain $\mathrm{E}[h_{\theta\theta^T}]$ from (A.27), (A.28), (A.30), and (A.33) as

$$\begin{aligned}&\mathrm{E}[h_{\theta\theta^T}]\\ &\quad = D_p^+\left[(A^{1/2}\otimes A^{1/2})\left\{K+\frac{\beta}{2(\beta+1)^{(p+2)/2}}\,\mathrm{vec}(I_p)\,\mathrm{vec}(I_p)^T\right\}(A^{-1/2}\otimes A^{-1/2})\right.\\ &\qquad \left.+\frac{1}{(\beta+1)^{(p+2)/2}}(I_p\otimes I_p)\right]D_p,\end{aligned} \tag{A.34}$$

where K is a $p^2\times p^2$ matrix with typical elements as (A.27) and (A.28). When $\beta = 0$, this expectation results in I_{p^*} $(p^* = p(p+1)/2)$.

References

[1] A. Basu, I.R. Harris, N.L. Hjort, M.C. Jones, Robust and efficient estimation by minimising a density power divergence, Biometrika 85 (3) (1998) 549–559.
[2] M.W. Browne, Asymptotically distribution-free methods for the analysis of covariance structures, Br. J. Math. Statist. Psychol. 37 (1984) 62–83.
[3] C. Croux, G. Haesbroeck, Influence function and efficiency of the minimum covariance determinant scatter matrix estimator, J. Multivar. Anal. 71 (1999) 161–190.
[4] L. Davies, The asymptotics of Rousseeuw's minimum volume ellipsoid estimator, Ann. Statist. 20 (4) (1992) 1828–1843.
[5] A.P. Dempster, Covariance selection, Biometrics 28 (1972) 157–175.
[6] D. Edwards, Introduction to Graphical Modeling, second ed., Springer, New York, 2000.
[7] S. Eguchi, Y. Kano, Robustifying maximum likelihood estimation, Technical Report, The Institute of Statistical Mathematics, Tokyo, June 2001.
[8] M.C. Jones, N.L. Hjort, I.R. Harris, A. Basu, A comparison of restated density-based minimum divergence estimators, Biometrika 88 (3) (2001) 865–873.
[9] S.L. Lauritzen, Graphical Models, Oxford Science Publications, Oxford, 1996.
[10] J.R. Magnus, H. Neudecker, Matrix Differential Calculus with Applications in Statistics and Econometrics, Revised Edition, Wiley, New York, 1999.
[11] R.A. Maronna, V.J. Yohai, Robust estimation of multivariate location and scatter, in: S. Kotz, C. Read, D. Banks (Eds.), Encyclopedia of Statistical Sciences Update, vol. 2, Wiley, New York, 1998, pp. 589–596.
[12] M. Minami, S. Eguchi, Robust blind source separation by beta divergence, Neural Comput. 14 (2002) 1859–1886.
[13] P.J. Rousseeuw, Multivariate estimation with high breakdown point, in: W. Grossmann, G. Pflug, I. Vincze, W. Wertz (Eds.), Mathematical Statistics and Applications, vol. B, Reidel, Dordrecht, 1985, pp. 283–297.
[14] P.J. Rousseeuw, K. Van Driessen, A fast algorithm for the minimum covariance determinant estimator, Technometrics 41 (3) (1999) 212–223.
[15] T.P. Speed, H.T. Kiiveri, Gaussian Markov distributions over finite graphs, Ann. Statist. 14 (1) (1986) 138–150.
[16] J.M. Tanner, W.J. Israelsohn, Parent–child correlation for body measurements of children between the ages one month and seven years, Ann. Hum. Genet. 26 (1963) 245–259.
[17] N. Wermuth, E. Scheidt, Fitting a covariance selection model to a matrix algorithm 105, J. Roy. Statist. Soc. Ser. C, Appl. Statist. 26 (1977) 88–92.
[18] J. Whittaker, Graphical Models in Applied Multivariate Statistics, Wiley Series in Probability and Mathematical Statistics, Wiley, New York, 1990.
[19] M.P. Windham, Robustifying model fitting, J. Roy. Statist. Soc. Ser. B, Methodol. 57 (3) (1995) 599–609.