A comparison of methods for estimating the determinant of high-dimensional covariance matrix

Zongliang Hu[1], Kai Dong[1], Wenlin Dai[2] and Tiejun Tong[1,*]

[1]Department of Mathematics, Hong Kong Baptist University, Hong Kong

[2]CEMSE Division, King Abdullah University of Science and Technology, Jeddah, Saudi Arabia

*Email: tongt@hkbu.edu.hk

Abstract

The determinant of the covariance matrix for high-dimensional data plays an important role in statistical inference and decision. It has many real applications including statistical tests and information theory. Due to the statistical and computational challenges with high dimensionality, little work has been proposed in the literature for estimating the determinant of high-dimensional covariance matrix. In this paper, we estimate the determinant of the covariance matrix using some recent proposals for estimating high-dimensional covariance matrix. Specifically, we consider a total of eight covariance matrix estimation methods for comparison. Through extensive simulation studies, we explore and summarize some interesting comparison results among all compared methods. We also provide practical guidelines based on the sample size, the dimension, and the correlation of the data set for estimating the determinant of high-dimensional covariance matrix. Finally, from a perspective of the loss function, the comparison study in this paper may also serve as a proxy to assess the performance of the covariance matrix estimation.

Keywords: Covariance matrix; High-dimensional data; Log-determinant; Sparse matrix; Shrinkage estimation; Thresholding estimation.

1 Introduction

High-dimensional data are becoming more common in scientific research including gene expression study, financial engineering and signal processing. One significant feature of such data is that the dimension p is larger than the sample size n, the so-called "large p small n" data. For example, gene microarray often measures thousands of gene expression values simultaneously for each individual. However, due to the cost or the limited availability of patients, the number of samples in microarray experiments is usually much smaller than the number of genes. It is common to see microarray data with less than 10 samples (Kuster et al. 2011, Mokry et al. 2012, Kaur et al. 2012, Richard et al. 2014, Schurch et al. 2016). As seen in the literature, there are many statistical and computational challenges in analyzing the "large p small n" data.

Let $X_i = (x_{i1}, \ldots, x_{ip})^T$, $i = 1, \ldots, n$, be independent and identically distributed (i.i.d.) random vectors from the multivariate normal distribution $N_p(\mu, \Sigma)$, where μ is a p-dimensional mean vector and Σ is a covariance matrix of size $p \times p$. When p is larger than n, the sample covariance matrix S_n is a singular matrix. To overcome the singularity problem, various methods for estimating Σ have been proposed in the recent literature, e.g., the ridge-type estimators in Ledoit and Wolf (2003) and Fisher and Sun (2011), the sparse estimators in Bickel and Levina (2008), Cai and Yuan (2012), Rothman (2012) and Cai et al. (2013). Recently, Chen et al. (2013) and Basu and Michailidis (2015) considered sparse covariance matrix estimation for time series data based on certain dependence measures, which relaxes the independence assumption among samples. For more references, see also Tong et al. (2014), Cai et al. (2016) and Fan et al. (2016).

Apart from the covariance matrix estimation, there are situations where one needs an estimate of the determinant (or the log-determinant) of the covariance matrix for high-dimensional data. To illustrate it, we write the log-likelihood function of the data

as

$$\log(L) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^T\Sigma^{-1}(X_i - \mu),$$

where $|\Sigma|$ denotes the determinant of the covariance matrix Σ. In classic multivariate analysis, the determinant $|\Sigma|$, referred to as the generalized variance (GV), was introduced by Wilks (1932) and Wilks (1967) as a scalar measure of overall multidimensional scatter. It has many applications such as outlier detection, hypothesis testing, and classification. To cater for this demand, we present several examples as follows.

- Quadratic discriminant analysis (QDA) is an important method of classification. Assuming that the data in class k follows $N_p(\mu_k, \Sigma_k)$, the quadratic discriminant scores are given by

  $$d_k(Y) = (Y - \mu_k)^T\Sigma_k^{-1}(Y - \mu_k) + \log|\Sigma_k| - 2\log\pi_k, \qquad k = 1, \ldots, K,$$

  where Y is the new sample, K is the total number of classes, and π_k is the prior probability of observing a sample from class k. The classification rule is to assign Y to class k that minimizes $d_k(Y)$ among all classes. To implement QDA, it is obvious that we need an estimate of $|\Sigma_k|$ or $\log|\Sigma_k|$.

- To estimate the high-dimensional precision matrix $\Omega = \Sigma^{-1}$, Yuan and Lin (2007) and Friedman et al. (2008) proposed to solve the following optimization problem:

  $$\hat{\Omega} = \arg\min_{\Omega>0}\{\mathrm{tr}(S_n\Omega) - \log|\Omega| + \lambda\|\Omega\|_1\},$$

  where $\mathrm{tr}(\cdot)$ is the trace, $\|\cdot\|_1$ is the ℓ_1 norm, and λ is a tuning parameter. The purpose of the term, $\log|\Omega| = -\log|\Sigma|$, is to ensure that the optimization problem has a unique global positive definite minimizer (Rothman 2012). Other proposals in this direction include Banerjee et al. (2008), Witten and Tibshirani (2009), Ravikumar et al. (2011), Yin and Li (2013) and among others.

- In probability theory and information theory, the differential entropy is commonly used by extending the concept of entropy to the continuous probability distribution (Hastie et al. 2002, Bishop 2006). For a random vector from $N_p(\mu, \Sigma)$, the differential entropy is

$$h(\Sigma) = \frac{p}{2} + \frac{p \log(2\pi)}{2} + \frac{\log |\Sigma|}{2}.$$

- The minimum covariance determinant (MCD) method developed by Rousseeuw (1985) and Rousseeuw and Driessen (1999) is a robust estimator of multivariate scatter. MCD aims to find a subset with h samples (observations) having the smallest determinant of the covariance matrix. Specifically, let $\mathcal{S} = \{I \subset \{1, \ldots, n\} : \text{card}(I) = h\}$ be the collections of all subsets with h samples, where $\text{card}(I)$ is the cardinality of I. For any $I \in \mathcal{S}$, let S_I be the corresponding sample covariance. The subset with the minimum determinant is defined as

$$I_m = \arg\min_{I \in \mathcal{S}} \{|S_I|\}.$$

  When p is larger than n, MCD is ill-defined as S_I is singular. To generalize the MCD method to high-dimensional data, we need an estimate for the determinant of the high-dimensional covariance matrix. For instance, Ro et al. (2015) replaced $|S_I|$ with $|\text{diag}(S_I)|$, and Boudt et al. (2017) modified $|S_I|$ by shrinking the subset-based sample covariance matrix toward a target matrix.

- Multivariate analysis of variance (MANOVA) is a procedure for testing the equality of mean vectors across multiple groups. Wilks' Λ statistic for the hypothesis test (Anderson 1984) is given as

$$\Lambda = \frac{|E|}{|H + E|},$$

  where E is the within-group sum of squares and cross-product matrix, and H is the between-group sum of squares and cross-product matrix. However, E

is singular under the "large p small n" setting. To apply MANOVA for high-dimensional data, Tsai and Chen (2009) proposed replacing E with a shrinkage estimator, in which the shrinkage intensity is computed based on the method by Schäfer and Strimmer (2005). Ullah and Jones (2015) compared the powers of three types of regularized Wilks' Λ statistics, in which E was replaced by the lasso, ridge and shrinkage estimator, respectively.

From the above examples, it is evident that an estimator of GV, or $\log|\Sigma|$, plays an important role in high-dimensional data analysis. For ease of notation, we let

$$\theta = \log|\Sigma|$$

throughout the paper. In contrast to the covariance matrix estimation, the investigation of estimating θ is relatively overlooked in the literature. In practice, one often estimates the covariance matrix first and then uses it to compute the log-determinant. Chiu et al. (1996) considered a regression model and allowed the covariance matrix of response vector $X_i = (x_{i1}, \ldots, x_{ip})^T$ to vary with explanatory variables. In specific, they proposed modeling each element of $\log\Sigma$ as a linear function of the explanatory variables. One property of the transformation is that the log determinant $\log|\Sigma|$ is equal to $\mathrm{tr}(\log\Sigma)$, a summation of log eigenvalues of Σ. Recently, Cai et al. (2015) investigated the estimation of θ under various settings. Under some "moderate" setting with $p \leq n$, they proposed to estimate θ by the determinant of the sample covariance matrix, i.e., $\log|S_n|$. A central limit theorem was also established for $\log|S_n|$ in the setting where p can grow with n. For the "large p small n" data, however, they showed that it is impossible to estimate θ consistently, unless some structural assumption such as sparsity on the parameter can be imposed.

In this paper, we conduct a comprehensive simulation study that evaluates the performance of the existing methods for estimating θ. We follow a two-step procedure:

we first estimate Σ with the existing methods, and then estimate θ by the plug-in estimator, $\hat{\theta} = \log(|\hat{\Sigma}|)$. In Section 2, we consider a total of eight methods for estimating θ. A brief review on each of the methods is also given. In Section 3, we conduct simulation studies to evaluate and compare their performance under various settings. In particular, we will consider different types of correlation structures including a non-positive definite covariance matrix that is often ignored in the existing literature. We then explore and summarize some useful findings, and provide some practical guidelines for scientists in Section 4. Finally, we conclude the paper in Section 5 with some discussion. Technical details are provided in the Appendix.

2 Methods for Estimating θ

In this section, we review eight representative methods for estimating the covariance matrix, and then estimate the log-determinant θ using the eight estimates of Σ, respectively. We also propose a new method for estimating θ under the assumption of a diagonal covariance matrix. For ease of presentation, we divide the eight methods into four categories: diagonal estimation, shrinkage estimation, sparse estimation, and factor model estimation.

2.1 Diagonal Estimation

Method 1: Diagonal Estimator (DE)

Under the "large p small n" setting, one naive approach is to estimate Σ by the diagonal sample covariance matrix, i.e., $\text{diag}(S_n)$. This estimator was first considered in Dudoit et al. (2002) to propose a diagonal linear discriminant analysis. It was further considered in Bickel and Levina (2004) where the authors demonstrated that a diagonal covariance matrix estimation can be sometimes reasonable when p is much larger than n. Let $\text{diag}(\Sigma) = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ where σ_j^2 are the covariate-specific variances for

$j = 1, \ldots, p$, and $\text{diag}(S_n) = \text{diag}(s_1^2, \ldots, s_p^2)$ where s_j^2 are the sample variances of σ_j^2, respectively. By letting $\hat{\Sigma} = \text{diag}(S_n)$, we define the first estimator of θ as

$$\hat{\theta}_{(1)} = \log \left|\text{diag}(S_n)\right| = \sum_{j=1}^{p} \log s_j^2. \tag{1}$$

We refer to $\hat{\theta}_{(1)}$ as the diagonal estimator (DE). To be specific, DE is proposed to estimate $\log |\text{diag}(\Sigma)|$ rather than $\log |\Sigma|$.

Method 2: Improved Diagonal Estimator (IDE)

It is noteworthy that DE may not perform well as an estimate of $\log |\text{diag}(\Sigma)|$ when the sample size is small, mainly due to the unreliable estimates of the sample variances. Various approaches have been proposed to improving the variance estimation in the literature. See, for example, Baldi and Long (2001), Wright and Simon (2003), Cui et al. (2005), Tong and Wang (2007), and Tong et al. (2012).

To improve DE, we consider the optimal shrinkage estimator in Tong and Wang (2007),

$$\hat{\sigma}_j^2 = \{h_p(1) s_{pool}^2\}^{\alpha} \{h_1(1) s_j^2\}^{1-\alpha},$$

where $s_{pool}^2 = \prod_{i=1}^{p} (s_j^2)^{1/p}$, $h_p(1) = (\nu/2) \left\{\Gamma(\nu/2)/\Gamma(\nu/2 + 1/p)\right\}^p$ with $\nu = n - 1$, $\Gamma(\cdot)$ is the Gamma function, and $\alpha \in [0, 1]$ is the shrinkage parameter. Replacing s_j^2 in DE by $\hat{\sigma}_j^2$, we have

$$\hat{\theta} = \sum_{j=1}^{p} \log \hat{\sigma}_j^2 = \hat{\theta}_{(1)} + C, \tag{2}$$

where $C = \log \{h_p^{\alpha p}(1) h_1^{(1-\alpha)p}(1)\}$ is a constant.

The estimation structure in (2) shows that the DE estimator, $\hat{\theta}_{(1)}$, can be further improved. Specifically, if we have C_0 such that $E(\hat{\theta}_{(1)} + C_0) = \log |\text{diag}(\Sigma)|$, then C_0 defines as the optimal C value so that the estimator $\hat{\theta}_{(1)} + C_0$ minimizes the mean squared error in the family of estimators $\{\hat{\theta}_{(1)} + C : \ C \in (-\infty, \infty)\}$.

Theorem 1. *Let $s_j^2 = \sigma_j^2 \chi_{\nu,j}^2/\nu$, where $\chi_{\nu,j}^2$ are i.i.d random variables with a common chi-squared distribution with ν degrees of freedom, and $C_0 = -p\left\{\log(2/\nu) + \psi(\nu/2)\right\}$, where $\psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ is the digamma function. Then for any fixed $\nu > 0$, we have*

(1) $\hat{\theta}_{(1)} + C_0$ is an unbiased estimator of $\log|\mathrm{diag}(\Sigma)|$.

(2) Assume also that σ_j^2 are i.i.d random variables from a common distribution F and $E(\log\sigma_1^2) < \infty$. Then

$$\frac{1}{p}\left(\hat{\theta}_{(1)} + C_0 - \log|\mathrm{diag}(\Sigma)|\right) \xrightarrow{a.s.} 0 \qquad \text{as} \quad p \to \infty,$$

where $\xrightarrow{a.s.}$ denotes almost sure convergence.

The proof of Theorem 1 is given in the Appendix. By (2) and Theorem 1, we define the second estimator of θ as

$$\hat{\theta}_{(2)} = \sum_{j=1}^{p} \log s_j^2 - p\left\{\log(2/\nu) + \psi(\nu/2)\right\}.$$

We refer to $\hat{\theta}_{(2)}$ as the improved diagonal estimator (IDE).

2.2 Shrinkage Estimation

Recall that the sample covariance matrix S_n is singular when the dimension is larger than the sample size. To overcome the singularity problem, other than the diagonal methods in Section 2.1, one may also estimate the covariance matrix by the following convex combination:

$$S^* = \delta T + (1-\delta)S_n,$$

where T is the target matrix, and $\delta \in [0,1]$ is the shrinkage parameter. Both the target matrix and the shrinkage parameter play an important role in the shrinkage estimation. For instance, if we let $T = \mathrm{diag}(S_n)$ and $\delta = 1$, then S^* reduces to the DE estimator.

The appropriate choice of the target matrix has been extensively studied in the literature. See, for example, Ledoit and Wolf (2003), Schäfer and Strimmer (2005), Warton (2008), Warton (2011), and Fisher and Sun (2011) and the references therein. Note that T is often chosen to be positive definite and well-conditioned, and consequently, the final estimate S^* is also guaranteed positive definite and well-conditioned for any dimensionality. As suggested in Schäfer and Strimmer (2005) and Fisher and Sun (2011), we consider a popular target matrix for nonhomogeneous variances: the "diagonal, unequal variance" matrix, i.e., the diagonal sample covariance matrix $\text{diag}(S_n)$.

We also note that, given the target matrix, the estimation of the shrinkage parameter δ is also crucial to the final estimation. The available estimation methods for the shrinkage parameter are mainly: (1) the unbiased estimation, and (2) the consistent estimation. The unbiased estimation is replacing unknown terms in the optimal value by their unbiased estimators (Schäfer and Strimmer 2005). Whereas, the consistent estimation is replacing the unknown terms in the optimal shrinkage parameter with (n, p)-consistent estimators (Fisher and Sun 2011). Taken together, we present below the four methods for estimating the covariance matrix and consequently for estimating θ, respectively.

Method 3: Unbiased Shrinkage Estimator with $T = I$ (USIE)

Letting the target matrix be $T = I$, Schäfer and Strimmer (2005) proposed an unbiased estimator for the shrinkage parameter, denoted by $\hat{\delta}_1^*$. This leads to $S^* = \hat{\delta}_1^* I + (1 - \hat{\delta}_1^*) S_n$. We then define the third estimator of θ as

$$\hat{\theta}_{(3)} = \log |\hat{\delta}_1^* I + (1 - \hat{\delta}_1^*) S_n|. \tag{3}$$

Method 4: Consistent Shrinkage Estimator with $T = I$ (CSIE)

Letting the target matrix be $T = I$, Fisher and Sun (2011) proposed a consistent estimator for the shrinkage parameter, denoted by $\hat{\delta}_2^*$. This leads to $S^* = \hat{\delta}_2^* I + (1 -$

$\hat{\delta}_2^*)S_n$. We then define the fourth estimator of θ as

$$\hat{\theta}_{(4)} = \log|\hat{\delta}_2^* I + (1-\hat{\delta}_2^*)S_n|. \tag{4}$$

Method 5: Unbiased Shrinkage Estimator with $T = \text{diag}(S_n)$ (USDE)

Letting $T = \text{diag}(S_n)$, Schäfer and Strimmer (2005) also proposed an unbiased estimator for the shrinkage parameter, denoted by $\hat{\delta}_3^*$. This leads to $S^* = \hat{\delta}_3^*\text{diag}(S_n) + (1-\hat{\delta}_3^*)S_n$. We then define the fifth estimator of θ as

$$\hat{\theta}_{(5)} = \log|\hat{\delta}_3^*\text{diag}(S_n) + (1-\hat{\delta}_3^*)S_n|. \tag{5}$$

Method 6: Consistent Shrinkage Estimator with $T = \text{diag}(S_n)$ (CSDE)

Letting $T = \text{diag}(S_n)$, Fisher and Sun (2011) also proposed a consistent estimator for the shrinkage parameter, denoted by $\hat{\delta}_4^*$. This leads to $S^* = \hat{\delta}_4^*\text{diag}(S_n) + (1-\hat{\delta}_2^*)S_n$. We then define the sixth estimator of θ as

$$\hat{\theta}_{(6)} = \log|\hat{\delta}_4^*\text{diag}(S_n) + (1-\hat{\delta}_4^*)S_n|. \tag{6}$$

2.3 Sparse Estimation

When p is much larger than n, the shrinkage methods in Section 2.2 may not achieve a significant improvement over S_n. In such settings, to have a good estimate of Σ, one may have to impose some structural assumptions such as sparsity in the parameters. Recently, Cai et al. (2016) reviewed some methods on estimating structured high-dimensional covariance and precision matrix. A typical sparsity is to assume that most of the off-diagonal elements in the covariance matrix are zero. To estimate the covariance matrix under a sparsity condition, various thresholding-based methods have been proposed in the literature that aim to locate some "large" off-diagonal elements. See, for example, Bickel and Levina (2008), Karoui (2008), Rothman et al. (2009), Lam

and Fan (2009), Cai and Liu (2011), Cai and Yuan (2012), Cai and Zhou (2012), Mitra and Zhang (2014), and Wang et al. (2016). Particularly, the adaptive thresholding estimator proposed by Cai and Liu (2011) achieves the optimal rate of convergence over a large class of sparse covariance matrix under a wide spectral norms. Besides, it can be shown that the adaptive thresholding estimator also attains the optimal convergence rate under Bregman divergence losses over a large parameter class (Cai and Zhou 2012, Cai et al. 2016). Therefore, we also consider the sparsity methods as a representative and use them to estimate θ, i.e., the log-determinant of the covariance matrix.

Method 7: Adaptive Thresholding Estimator (ATE)

Bickel and Levina (2008) proposed a universal thresholding method where all entries in the sample covariance matrix are thresholded by a common value γ. They required that the variances σ_j^2 are uniformly bounded by a constant K, and consequently, the variances of the entries of the sample covariance matrix are also uniformly bounded. However, it was shown that a universal thresholding method is suboptimal over a certain class of sparse covariance matrices.

To improve the method above, Cai and Liu (2011) proposed an adaptive thresholding estimator for the covariance matrix:

$$\hat{\Sigma}^* = (\tilde{\sigma}_{ij}^*)_{p\times p} \quad \text{with} \quad \tilde{\sigma}_{ij}^* = s_{\gamma_{ij}}(s_{ij}),$$

where γ_{ij} is the corresponding threshold of $\tilde{\sigma}_{ij}^*$, and $s_{\gamma_{ij}}(\cdot)$ is a generalized thresholding operator (Rothman et al. 2009), which is specified as the soft thresholding throughout simulations. With the proper γ_{ij}, the estimator $\hat{\Sigma}^*$ adaptively achieves the optimal rate of convergence over a large class of sparse covariance matrix under the spectral norm. Now by $\hat{\Sigma}^*$, the seventh estimator of θ is

$$\hat{\theta}_{(7)} = \log|\hat{\Sigma}^*|. \tag{7}$$

We refer to $\hat{\theta}_{(7)}$ as the adaptive thresholding estimator (ATE).

2.4 Factor Model Estimation

The sparsity condition on the covariance matrix assumes that most of covariates are uncorrelated to each other. Note that, however, this assumption may not be realistic in practice. Recently, under the assumption of conditional sparsity, Fan et al. (2013) introduced a principle orthogonal complement thresholding method using the factor model. In this section, we briefly review their method and then apply it to estimate the log-determinant of the covariance matrix.

Method 8: Principal Orthogonal Complement Thresholding Estimator (POET)

Fan et al. (2013) considered the approximate factor model:

$$y_g = Bf_g + u_g, \quad g = 1, \ldots, G,$$

where $y_g = (y_{1g}, \ldots, y_{pg})^T$ is the observed response, $B = (b_1, \ldots, b_p)^T$ is the loading matrix, f_g is a $Q \times 1$ vector of common factors, and $u_g = (u_{1g}, \ldots, u_{pg})^T$ is the error vector. In this model, we can only observe y_g. Let

$$\Sigma = B\text{cov}(f_g)B^T + \Sigma_u, \quad g = 1, \ldots, G,$$

where Σ_u is the covariance matrix of u_g. To estimate Σ, Fan et al. (2013) applied the spectral decomposition on the sample covariance matrix:

$$S_n = \sum_{j=1}^{Q} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T + \hat{R}_Q,$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ are eigenvalues of S_n, $\hat{\xi}_j$, $j = 1, \ldots, p$ are the corresponding eigenvectors, and $\hat{R}_Q = \sum_{j=Q+1}^{p} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T$ is the principal orthogonal complement. For this decomposition, the first Q principal components were kept and the thresholding was applied on $\hat{R}_Q$. Here, the generalized thresholding operator can be used. In

addition, Fan et al. (2013) also introduced a method to obtain an estimation of Q, denoted by $\hat{Q}$. Their final estimator of Σ is

$$\hat{\Sigma}_{\hat{Q}} = \sum_{j=1}^{\hat{Q}} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T + \hat{R}_{\hat{Q}}^{\mathcal{T}}, \tag{8}$$

where $\hat{R}_{\hat{Q}}^{\mathcal{T}}$ is the thresholding result of $\hat{R}_Q$. Now by (8), we define the last estimator of θ as

$$\hat{\theta}_{(8)} = \log |\hat{\Sigma}_{\hat{Q}}|. \tag{9}$$

We refer to $\hat{\theta}_{(8)}$ as the principal orthogonal complement thresholding estimator (POET).

3 Simulation Studies

In this section, we compare the numerical performance of the aforementioned eight estimators. We consider five different setups. In the first setup, we generate data from the multivariate normal distribution, $N_p(0, \Sigma)$. In the second setup, we generate data from a mixture distribution where the covariance matrix is highly sparse. In the third setup, we simulate data from the log-normal distribution to assess the robustness of the eight methods under heavy-tailed data. In the forth setup, we consider a special case where the covariance matrix is degenerate and the data are generated from a degenerate multivariate normal distribution. And in the final setup, we use a realistic covariance matrix structure that is obtained from a real data. To compare these methods, we compute the mean squared error (MSE) as below:

$$\text{MSE}(\theta, \hat{\theta}) = \frac{1}{Mp} \sum_{m=1}^{M} (\hat{\theta}_m - \theta)^2,$$

where M is the repeated times. Throughout the simulations, we take $M = 500$.

3.1 Normal Data

In this setup, we consider a block diagonal structure for the covariance matrix. This structure is widely adopted in the literature, e.g., Guo et al. (2007) and Pang et al. (2009). Specifically, we let

$$\Sigma_2 = D^{1/2} R(\rho) D^{1/2},$$

where $D = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ with σ_j^2 being i.i.d. from the distribution $\chi_5^2/5$, and R follows a block diagonal structure:

$$R(\rho) = \begin{pmatrix} \Sigma_\rho & 0 & \cdots & \cdots & 0 \\ 0 & \Sigma_{-\rho} & 0 & \ddots & \vdots \\ \vdots & 0 & \Sigma_\rho & 0 & \vdots \\ \vdots & \ddots & 0 & \Sigma_{-\rho} & \ddots \\ 0 & \cdots & \cdots & \ddots & \ddots \end{pmatrix}_{p \times p}.$$

In our simulations, we consider $\Sigma_\rho = (\sigma_{ij}(\rho))_{q \times q}$ with $\sigma_{ij}(\rho) = \rho^{|i-j|}$ for $1 \le i, j \le q$. In addition, we set $\rho = 0$, 0.3, 0.6 or 0.9, to represent different levels of dependence, and $(p, q) = (50, 5)$ or $(300, 10)$, respectively.

Figures 1 and 2 display the log(MSE) of the eight methods for different levels of dependence, dimension and sample size. From these figures, we have the following findings. When the covariates are uncorrelated, IDE gives the best performance under a high dimension (e.g., $p = 300$). However, if the dimension is not large (e.g., $p = 50$), and the covariates are uncorrelated or weak correlated, shrinking the covariance matrix toward an identity matrix leads to a better performance under a small sample size. This is because when the sample size is small, the variances of the entries of the sample covariance matrix is large. Hence, CSIE and USIE stabilize both diagonal and off-diagonal entries and, at the same time, an identity target possesses an explicit structure which in turn requires little data to fit. Consequently, the resulting estimators have a good bias–variance tradeoff. In addition, when the correlation and dimension are both

large, imposing additional structure assumptions is necessary. Under this situation, ATE and POET turn out to be the best two methods among the eight methods unless the sample size is relatively small. When the sample size is small, the pattern of ATE is very similar to that of DE. When the sample size and dimension are both large, ATE outperforms all other methods except for POET.

Figure 3 displays the performance of the eight methods for different levels of dependence with $p = 300$. The pattern is consistent with Figure 2. In particular, as the correlation and sample size are large, the performance of POET is satisfactory. From Figures 1 and 2, however, we note that the log(MSE) of POET tends to be oscillating as the sample size increases. This may due to that POET depends on the estimated number of factors K. In Fan et al. (2013), the authors used a consistent estimator for K and showed that POET is robust to over-estimated number of factors under the spectral norm. Our simulations in Table 1, however, show that the robustness for estimating the covariance matrix may not hold any more when the purpose is for estimating the determinant. In particular for small sample sizes, either over-estimated or under-estimated K leads to a large bias for the determinant estimator.

3.2 Mixture Normal Data

In this setup, we consider a mixture model where the random vectors are generated from

$$X \sim \alpha_1 f_1(X) + \alpha_2 f_2(X),$$

where $f_1(X)$ and $f_2(X)$ are the density functions of $N_p(\mu_3, \Sigma_3)$ and $N_p(\mu_4, \Sigma_4)$, respectively. For the covariance matrices, we consider a sparse block diagonal structure as follows:

$$\Sigma_3 = D^{1/2} R(\rho) D^{1/2} \quad \text{and} \quad \Sigma_4 = D^{1/2} R(-\rho) D^{1/2},$$

where $D = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ with σ_j^2 being i.i.d. from the distribution $(1/5)\chi_5^2$, and $R(\rho)$ being the same as in Setup II. For simplicity, we set $\alpha_1 = \alpha_2 = 1/2$ and $\mu_1 = \mu_2 = 0$. Under this setting, the covariance matrix of X is simplified as $(\Sigma_3 + \Sigma_4)/2$, which results in a highly sparse matrix where the odd off-diagonal parts in diagonal blocks are zeros. We set $(p, q) = (50, 5)$ or $(300, 10)$, and $\rho = 0$, 0.3, 0.6 or 0.9.

Figures 4 and 5 display the log(MSE) of the eight methods under different levels of dependence and sample size. When the sample size is large and the covariates are uncorrelated, IDE gives the best performance. When the sample size is small and the dimension is not large (e.g., $n = 5, p = 50$), shrinking the covariance matrix toward an identity matrix (e.g., USIE and CSIE) outperforms the other methods except that the correlation is very large (e.g., $\rho = 0.9$). However, as the sample size and dimension are both large, the shrinkage methods will become suboptimal. Instead, if the correlation is also large (e.g., $\rho = 0.6$), ATE and POET outperform the other methods in most settings. As aforementioned, the performance of POET is not stable and may not be satisfactory when the sample size is not large.

3.3 Heavy-tailed Data

In this setup, we consider to simulate heavy-tailed data from a log-normal distribution, $\ln N(\mu, \sigma^2)$, where the mean and variance are $e^{\mu+\sigma^2/2}$ and $(e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$, respectively. First of all, we generate n independent random vectors $Z_i = (z_{i1}, \ldots, z_{ip})^T$, where all the components of Z_i are sampled independently from $\ln N(0, 1)$. Let $X_i = \Sigma^{1/2} Z_i^*$ with $Z_i^* = (z_{i1} - e^{1/2}, \ldots, z_{ip} - e^{1/2})^T / \{e(e-1)\}^{1/2}$, and Σ is a $p \times p$ positive definite matrix. Consequently, the mean vector and covariance matrix of X_i are $0_{p\times 1}$ and $\Sigma_{p\times p}$, respectively. For the covariance matrix, we consider the block diagonal structure as described in Section 3.1. We set $(p, q) = (50, 5)$ or $(300, 10)$, and $\rho = 0$, 0.3, 0.6 or 0.9.

Figures 6 and 7 display the log(MSE) of the eight methods under different levels

of dependence and sample size. When the dimension and correlation are both small, USIE and CSIE outperform the other methods. The reason is similar as the discussion in Section 3.1, the heavy-tailed data may lead to unstable estimates of the entries of Σ, hence shrinking towards a simple identity target, which requires little data to fit, stabilizes the sample covariance matrix. In addition, as shown in Figure 7, when the dimension is large and the correlation is not small, ATE and POET are the only two methods that have a better performance than the other methods except that the sample size is small. Finally, we also note that IDE cannot provide a satisfactory performance even if the covariates are uncorrelated. As demonstrated in Theorem 1, IDE estimator is derived under the normal distribution and may not be robust to heavy-tailed data.

3.4 Degenerate Normal Data

To further investigate the performance of the eight methods, we consider a non-positive definite covariance matrix in which the positive definite assumption of the covariance matrix is violated. Note that this new setting is often overlooked in the literature. To construct a non-positive definite covariance matrix, we define the affine transformation C as

$$C = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & 0 & 1 & 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 & 1 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 & 1 & 0 \\ 0 & \cdots & 0 & 1/\sqrt{p-4} & 1/\sqrt{p-4} & \cdots & 1/\sqrt{p-4} & 0 \end{pmatrix}_{p\times p}.$$

We then apply the affine transformation to the covariance matrix in Setup II and form

$$\Sigma_5 = C\Sigma_2 C^T.$$

It is obvious that $|\Sigma_5| = 0$ since $|C| = 0$. We set $(p, q) = (50, 5)$, and $\rho = 0$, 0.3, 0.6 or 0.9. Note that the log-determinant of Σ_5 is negative infinity. Hence, for this degenerate setting, the MSE is defined on the determinant rather than on the log-determinant. Specifically, it is

$$\text{MSE}(e^{\theta}, e^{\hat{\theta}}) = \frac{1}{Mp} \sum_{m=1}^{M} \left(e^{\hat{\theta}_m} - e^{\theta} \right)^2.$$

Figure 8 shows the log(MSE) of all eight methods for different levels of dependence and sample size. We can see that the simulation results are different from those in the previous three setups. POET gives the best performance among the eight methods. In addition, we note that, under the non-positive definite setting, POET performs extremely well when the sample size is very small. For this phenomenon, we explore the possible reasons in the next paragraph.

To estimate Σ, Fan et al. (2013) applied the spectral decomposition on the sample covariance matrix:

$$S_n = \sum_{j=1}^{Q} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T + \hat{R}_Q.$$

If the sample size is much smaller than the dimension p, most eigenvalues of S_n are zeros. This leads to $\hat{R}_Q$, the principal orthogonal complement of the largest Q eigenvalues, is nearly a zero matrix. And consequently, the final estimator of POET, $\hat{\Sigma}_{\hat{Q}} = \sum_{j=1}^{\hat{Q}} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T + \hat{R}_{\hat{Q}}^{\mathcal{T}}$, tends to be highly degenerate for small sample sizes rather than for large sample sizes.

Finally, it is noteworthy that when the correlation is strong, the log(MSE) of POET is also fluctuant as the sample size increases. This again verifies that both the correlation and sample size have a large impact on the performance of POET.

3.5 Real Data

In this setup, we consider to generate a realistic covariance matrix from the Myeloma data (Zhan et al. 2007), which is a real microarray data set including a total of 54,675 genes, with 351 samples in the first group and 208 samples in the second group. To generate the covariance matrix, we first select 100 genes randomly from the first group and then compute the sample covariance matrix using the selected genes, denoted by Σ_r. Next, to evaluate the performance of the estimators under different levels of dependence, we follow Tong et al. (2013) and define the true covariance matrix as

$$\Sigma_1 = (1-\rho)\text{diag}(\Sigma_r) + \rho\Sigma_r,$$

where ρ controls the level of dependence. We set $\rho = 0$, $1/3$, $2/3$ or 1. Note that $\rho = 0$ corresponds to a diagonal covariance matrix, and $\rho = 1$ treats the generated sample covariance matrix as the true covariance matrix.

Figure 9 shows the log(MSE) of the eight methods for different levels of dependence and sample size. The comparison results are summarized as follows. When the sample size and correlation are both small, the methods that shrinking the covariance matrix toward the identity matrix (e.g., USIE and CSIE) lead to a good performance. When the covariates are uncorrelated and the sample size is large, IDE has the best performance. In addition, when the sample size is large and the correlation is moderate (e.g., $n = 80$ and $\rho = 2/3$), shrinking the sample covariance matrix toward a diagonal target matrix (e.g., USDE and CSDE) has a good performance. When the correlation and sample size are both large, ATE outperforms or is at least comparable to USDE and CSDE. Finally, POET is not stable and very sensitive to both the correlation and the sample size. When the correlation and sample size is not large, POET may fail to provide a satisfactory performance owing to the largely increased bias compared with the other methods.

4 Conclusion

In this section, we summarize some useful findings of the comparison results and also provide some practical guidelines for researchers.

1. Diagonal estimation

   The diagonal estimator, DE, is the simplest method for estimating the determinant of high-dimensional covariance matrix. It assumes that all covariates are uncorrelated. For independent normal data, IDE is an unbiased estimator of $\log|\text{diag}(\Sigma)|$ and also provides the best performance, especially when the dimension is large. For such settings, IDE can be recommended for estimating the determinant of high-dimensional covariance matrix. In addition, we note that IDE is not robust and may lead to an unsatisfactory performance when the independent normal assumption is violated.

2. Shrinkage estimation

   For the shrinkage estimation, different choices of the target matrix and shrinkage parameter result in different performance for the determinant estimation. In general, when the dimension is not large (e.g., $p = 50$), the shrinkage towards an identity target matrix (e.g., CSIE and USIE) performs well under the small sample size and weak correlation. This pattern is more evident for the heavy-tailed data. With a diagonal target matrix, CSDE, the consistent estimator of Fisher and Sun (2011), has a similar performance with USDE. However, CSDE and USDE are seldom become the best method especially when the sample size is not large.

   For the shrinkage estimators, the optimal shrinkage intensity can be specified without any further turning parameters. Consequently, the time consuming procedures such as cross-validation or bootstrap can be avoided. Table 2 shows the

computational time of the eight methods. As we can see, the shrinkage methods are much faster than ATE and POET. More importantly, if the sample size is very small as $n = 5$, 10, selecting the turning parameters in ATE and POET by cross-validation may result in a large bias. Under this situation, the shrinkage estimators (e.g., shrinkage towards an explicit target matrix) can be very attractive. Nevertheless, as the sample size increases or the correlation is strong, the performance of the shrinkage methods may not be as competitive as the sparse method and the factor model method.

3. Sparse estimation

   ATE presents its robust property in our settings. Specifically, when the sample size is not very small, ATE performs better or comparably to the other seven methods under various data structures and different levels of dependence. In practice, if the sample size is not very small and we have no prior information about the dependence level of the covariates, the sparse estimator can be recommended for estimating the determinant of high-dimensional covariance matrix.

   As shown in the simulations, when the sample size is very small, the performance of ATE is not attractive as the shrinkage estimators or even the diagonal estimators. For possible reasons, we note that an adaptive thresholding parameter in ATE is needed in practice. When the sample size is very small, however, their proposed cross-validation method may not provide a reliable estimate for the optimal threshold value.

4. Factor model estimation

   The factor model estimation, POET, is very attractive for strong correlated data sets when the sample size is not small. Fan et al. (2013) assumed that the data are weakly correlated after extracting the common factors which can result in

high levels of dependence among the covariates. This implies that POET may provide a good performance if the data are strong correlated. Note also that POET can select $K = 0$ automatically if the true covariance matrix is sparse. Then consequently, their method will degenerate to the sparse estimation such as the hard thresholding estimation in Bickel and Levina (2008) or ATE in Cai and Liu (2011).

POET, however, depends on the number of factors K, which is unknown in practice. To investigate the impact of the factors under different sample sizes and different levels of dependence, we simulated the MSE of POET for the log-determinant of the covariance under Setup II. Results from Table 1 show that K has a large impact on the determinant estimation. When the correlation is strong, $\hat{K}$, a consistent estimator of K, usually leads to a large MSE. Fan et al. (2013) demonstrated that POET is robust to over-estimated and sensitive to under-estimated factors. For the finite sample size, they suggested to chose a relatively large K (e.g., not less than 8). However, our simulation studies showed that the robustness for estimating the covariance matrix may not hold any more for estimating the determinant. In particular, for small sample size, both under-estimated and over-estimated factors will give a bad performance of POET. In view of this, we believe that future research is needed for selecting the optimal K when the factor model method is applied to estimate the determinant of the covariance matrix.

To conclude, the sample size, the dependence level and the dimension of the data sets take a great impact on the accuracy of estimation. In practice, we may need to select an appropriate estimation method according to the sample size and the prior information on the correlation structure of the covariates. When such prior information is not available, we recommend to use ATE (Cai and Liu 2011) to estimate the deter-

minant of high-dimensional covariance matrix, which is robust to various correlations and data structures.

5 Discussion

In this paper, we have compared a total of eight methods for estimating the log-determinant of the high-dimensional covariance matrix. The performance of the eight methods depends on the sample size, the dependence structure and the dimension of the data. When the sample size is not small, we note that ATE (Cai and Liu 2011) is always able to provide an average or above average performance among the eight methods. Hence, if there is little prior information about the structure of the covariance matrix, we recommend to use ATE to estimate the log-determinant θ, or GV, in practice. In terms of computational time, the shrinkage methods are more convenient than ATE and POET because the latter two methods need to select the penalty parameters via cross-validation.

Note that the log-determinant of a covariance matrix is a scalar, the two-step procedure may not provide the best estimation for θ. One possible future direction is to consider circumventing the full covariance matrix estimation, and estimating the log-determinant directly. Note that $\log|\Sigma| = \mathrm{tr}(\log\Sigma)$, which is essentially a summation of the log-eigenvalues of Σ. This suggests that the random matrix theory or the spectrum analysis may provide feasible solutions to estimate the log-determinant more accurately. The comparison study in this paper may also serve as a proxy to assess the performance of the covariance matrix estimation. Specifically, from a perspective of the loss function, if we define the loss function as

$$\mathrm{Loss}(\hat{\Sigma},\Sigma) = (\log|\hat{\Sigma}| - \log|\Sigma|)^2 \quad \text{or} \quad \mathrm{Loss}(\hat{\Sigma},\Sigma) = (|\hat{\Sigma}| - |\Sigma|)^2,$$

then the conducted simulations in Section 3 provide essentially a comparison for the eight methods for estimating Σ rather than θ. Of course, we do not intend to claim

that the above loss functions should be consistently recommended. In contrast, for evaluating the covariance matrix estimation, other popular methods are also available in the literature. For instance, by letting L as the likelihood function and $\hat{L}$ as the corresponding estimator, we may consider any of the distance between the log-likelihood and the estimated log-likelihood as the criterion to evaluate the performance:

$$D(L, \hat{L}) = \{\log(L) - \log(\hat{L})\}^2.$$

In addition, we can also consider any of the following loss functions:

- $\text{Loss}(\hat{\Sigma}, \Sigma) = \|\hat{\Sigma} - \Sigma\|_2 = \sqrt{\lambda_{\max}\{(\hat{\Sigma} - \Sigma)^T(\hat{\Sigma} - \Sigma)\}}$, where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue (Karoui 2008, Rothman et al. 2009, Fan et al. 2011).
- $\text{Loss}(\hat{\Sigma}, \Sigma) = \|\hat{\Sigma} - \Sigma\|_F = \sqrt{\sum_{i,j}(\hat{\sigma}_{ij} - \sigma_{ij})^2}$, where $\Sigma = (\sigma_{ij})_{p\times p}$ and $\hat{\Sigma} = (\hat{\sigma}_{ij})_{p\times p}$ (Cai and Liu 2011, Fan et al. 2013).
- $\text{Loss}(\hat{\Sigma}, \Sigma) = \|\hat{\Sigma} - \Sigma\|_{\max} = \max_{i,j} |\hat{\sigma}_{ij} - \sigma_{ij}|$ (Fan et al. 2013).

Further research is needed to investigate which loss function provides the best criterion for evaluating the estimation methods of the covariance matrix.

Finally, it is noteworthy that there are another category of publications in the literature on calculating the log-determinant of the covariance matrix (Barry and Pace 1999, Zhang and Leithead 2007, Boutsidis et al. 2017, Peng and Wang 2015, Han et al. 2015, Fitzsimons et al. 2017a, Fitzsimons et al. 2017b). We now point out that they are very different from the proposed study in our paper. Specifically, these papers assume that the covariance matrix Σ is known, yet as the dimension is very large, the canonical methods (e.g., the Choleskey decomposition) for computing $\log|\Sigma|$ require a total of $O(p^3)$ operations and may not be feasible in practice. The above papers have proposed more efficient algorithms including the random matrix theory and the spectrum analysis for fast computation of $\log|\Sigma|$.

Acknowledgments

Tiejun Tong's research was supported by the National Natural Science Foundation of China grant (No. 11671338), and the Hong Kong Baptist University grants FRG2/15-16/019, FRG2/15-16/038 and FRG1/16-17/018. The authors thank the editor, the associate editor and two reviewers for their constructive comments that have led to a substantial improvement of the paper.

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, New York: Wiley.

Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes, *Bioinformatics* **17**: 509–519.

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

Barry, R. P. and Pace, R. K. (1999). Monte carlo estimates of the log determinant of large sparse matrices, *Linear Algebra and its applications* **289**: 41–54.

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models, *The Annals of Statistics* **43**: 1535–1567.

Bickel, P. J. and Levina, E. (2004). Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations, *Bernoulli* **10**: 989–1010.

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

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, New York: Springer.

Boudt, K., Rousseeuw, P., Vanduffel, S. and Verdonck, T. (2017). The minimum regularized covariance determinant estimator, *arXiv preprint arXiv:1701.07086*.

Boutsidis, C., Drineas, P., Kambadur, P., Kontopoulou, E.-M. and Zouzias, A. (2017). A randomized algorithm for approximating the log determinant of a symmetric positive definite matrix, *Linear Algebra and its Applications*, in press.

Cai, T., Liang, T. and Zhou, H. (2015). Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions, *Journal of Multivariate Analysis* **137**: 161–172.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation, *Journal of the American Statistical Association* **106**: 672–684.

Cai, T., Ren, Z. and Zhou, H. (2013). Optimal rates of convergence for estimating Toeplitz covariance matrices, *Probability Theory and Related Fields* **156**: 101–143.

Cai, T., Ren, Z. and Zhou, H. (2016). Estimating structured high-dimensional covariance and precision matrices: optimal rates and adaptive estimation, *Electronic Journal of Statistics* **10**: 1–59.

Cai, T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding, *The Annals of Statistics* **40**: 2014–2042.

Cai, T. and Zhou, H. (2012). Optimal rates of convergence for sparse covariance matrix estimation, *The Annals of Statistics* **40**: 2389–2420.

Chen, X., Xu, M. and Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series, *The Annals of Statistics* **41**: 2994–3021.

Chiu, T. Y. M., Leonard, T. and Tsui, K. W. (1996). The matrix-logarithmic covariance model, *Journal of the American Statistical Association* **91**: 198–210.

Cui, X., Hwang, J. T. G., Qiu, J., Blades, N. J. and Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates, *Biostatistics* **6**: 59–75.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association* **97**: 77–87.

Fan, J., Liao, Y. and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices, *The Econometrics Journal* **19**: C1–C32.

Fan, J., Liao, Y. and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models, *The Annals of Statistics* **39**: 3320–3356.

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion), *Journal of the Royal Statistical Society: Series B* **75**: 603–680.

Fisher, T. J. and Sun, X. (2011). Improved stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix, *Computational Statistics and Data Analysis* **55**: 1909–1918.

Fitzsimons, J., Cutajar, K., Osborne, M., Roberts, S. and Filippone, M. (2017a). Bayesian inference of log determinants, *arXiv preprint arXiv:1704.01445*.

Fitzsimons, J., Granziol, D., Cutajar, K., Osborne, M., Filippone, M. and Roberts, S. (2017b). Entropic trace estimates for log determinants, *arXiv preprint arXiv:1704.07223*.

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

Guo, Y., Hastie, T. and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays, *Biostatistics* **8**: 86–100.

Han, I., Malioutov, D. and Shin, J. (2015). Large-scale log-determinant computation through stochastic Chebyshev expansions, *Proceedings of the 32nd International Conference on Machine Learning*, pp. 908–917.

Hastie, T., Tibshirani, R. and Friedman, J. (2002). *The Elements of Statistical Learning*, New York: Springer.

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

Kaur, S., Archer, K. J., Devi, M. G., Kriplani, A., Strauss, J. F. and Singh, R. (2012). Differential gene expression in granulosa cells from polycystic ovary syndrome patients with and without insulin resistance: identification of susceptibility gene sets through network analysis, *Journal of Clinical Endocrinology and Metabolism* **97**: E2016–E2021.

Kuster, D. W., Merkus, D., Kremer, A., van IJcken, W. F., de Beer, V. J., Verhoeven, A. J. and Duncker, D. J. (2011). Left ventricular remodeling in swine after myocardial infarction: a transcriptional genomics approach, *Basic Research in Cardiology* **106**: 1269–1281.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation, *The Annals of Statistics* **37**: 42–54.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection, *Journal of Empirical Finance* **10**: 603–621.

Mitra, R. and Zhang, C. (2014). Multivariate analysis of nonparametric estimates of large correlation matrices, *arXiv preprint arXiv:1403.6195*.

Mokry, M., Hatzis, P., Schuijers, J., Lansu, N., Ruzius, F. P., Clevers, H. and Cuppen, E. (2012). Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady-state RNA levels identify differentially regulated functional gene classes, *Nucleic Acids Research* **40**: 148–158.

Pang, H., Tong, T. and Zhao, H. (2009). Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data, *Biometrics* **65**: 1021–1029.

Peng, W. and Wang, H. (2015). Large-scale log-determinant computation via weighted l_2 polynomial approximation with prior distribution of eigenvalues, *International Conference on High Performance Computing and Applications*, Springer, pp. 120–125.

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

Richard, A. C., Lyons, P. A., Peters, J. E., Biasci, D., Flint, S. M., Lee, J. C., McKinney, E. F., Siegel, R. M. and Smith, K. G. (2014). Comparison of gene expression microarray data with count-based RNA measurements informs microarray interpretation, *BMC Genomics* **15**: 649–659.

Ro, K., Zou, C., Wang, Z. and Yin, G. (2015). Outlier detection for high-dimensional data, *Biometrika* **102**: 589–599.

Rothman, A. J. (2012). Positive definite estimators of large covariance matrices, *Biometrika* **99**: 733–740.

Rothman, A. J., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices, *Journal of the American Statistical Association* **104**: 177–186.

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

Rousseeuw, P. J. and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator, *Technometrics* **41**: 212–223.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Statistical Applications in Genetics and Molecular Biology* **4**: 32.

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G. and Owen-Hughes, T. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use, *RNA* **22**: 839–851.

Tong, T., Feng, Z., Hilton, J. S. and Zhao, H. (2013). Estimating the proportion of true null hypotheses using the pattern of observed p-values, *Journal of Applied Statistics* **40**: 1949–1964.

Tong, T., Jang, H. and Wang, Y. (2012). James-Stein type estimators of variances, *Journal of Multivariate Analysis* **107**: 232–243.

Tong, T., Wang, C. and Wang, Y. (2014). Estimation of variances and covariances for high-dimensional data: a selective review, *WIREs Computational Statistics* **6**: 255–264.

Tong, T. and Wang, Y. (2007). Optimal shrinkage estimation of variances with applications to microarray data analysis, *Journal of the American Statistical Association* **102**: 113–122.

Tsai, C. A. and Chen, J. J. (2009). Multivariate analysis of variance test for gene set analysis, *Bioinformatics* **25**: 897–903.

Ullah, I. and Jones, B. (2015). Regularised MANONA for high-dimensional data, *Australian and New Zealand Journal of Statistics* **57**: 377–389.

Wang, T., Berthet, Q. and Samworth, R. J. (2016). Statistical and computational trade-offs in estimation of sparse principal components, *The Annals of Statistics* **44**: 1896–1930.

Warton, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices, *Journal of the American Statistical Association* **103**: 340–349.

Warton, D. I. (2011). Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations, *Biometrics* **67**: 116–123.

Wilks, S. (1967). "Multidimensional Statistical Scatter", *in* T. W. Andreson (ed.), *Collected Papers: Contributions to Mathematical Statistics*, John Wiley and Sons, New York, pp. 597–614.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance, *Biometrika* **24**: 471–494.

Witten, D. M. and Tibshirani, R. (2009). Covariance-regularized regression and classification for high dimensional problems, *Journal of the Royal Statistical Society: Series B* **71**: 615–636.

Wright, G. W. and Simon, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments, *Bioinformatics* **19**: 2448–2455.

Yin, J. and Li, H. (2013). Adjusting for high-dimensional covariates in sparse precision matrix estimation by ℓ_1-penalization, *Journal of Multivariate Analysis* **116**: 365–381.

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

Zhan, F., Barlogie, B., Arzoumanian, V., Huang, Y., Williams, D. R., Hollmig, K., Pineda-Roman, M., Tricot, G., van Rhee, F., Zangari, M., Dhodapkar, M. and Shaughnessy Jr, J. D. (2007). Gene-expression signature of benign monoclonal gammopathy evident in multiple myeloma is linked to good prognosis, *Blood* **109**: 1692–1700.

Zhang, Y. and Leithead, W. E. (2007). Approximate implementation of the logarithm of the matrix determinant in Gaussian process regression, *Journal of Statistical Computation and Simulation* **77**: 329–348.

Appendix: Proof of Theorem 1

(1) From $s_j^2 = \sigma_j^2 \chi_{\nu,j}^2/\nu$, we have $\log s_j^2 = \log \sigma_j^2 + \log(\chi_{\nu,j}^2/\nu)$. Then, $\sum_{j=1}^p \log s_j^2 = \sum_{j=1}^p \log \sigma_j^2 + p \log \chi_{\nu,j}^2 - p \log \nu$. Further,

$$E\left(\sum_{j=1}^p \log s_j^2\right) = \sum_{j=1}^p \log \sigma_j^2 + p\{\log 2 + \psi(\nu/2)\} - p \log \nu.$$

This leads to

$$\begin{aligned} E\left\{\hat{\theta}_{(1)} + C_0\right\} &= E\left(\sum_{j=1}^p \log s_j^2\right) - p\{\log 2 + \psi(\nu/2)\} + p \log \nu \\ &= \sum_{j=1}^p \log \sigma_j^2 = \log|\mathrm{diag}(\Sigma)|. \end{aligned}$$

Hence, $\hat{\theta}_{(1)} + C_0$ is an unbiased estimator of $\log|\mathrm{diag}(\Sigma)|$.

(2) For $E(\log \sigma_1^2) < \infty$, we have

$$\frac{1}{p}\sum_{j=1}^p \log \sigma_j^2 \xrightarrow{a.s.} E(\log \sigma_1^2) \quad as\ p \to \infty.$$

Since $E(\log s_1^2) = E\{E(\log s_1^2|\sigma_1^2)\} = E(\log \sigma_1^2) + \log(2/\nu) + \psi(\nu/2)$, we have

$$\frac{1}{p}\sum_{j=1}^p \log s_j^2 - \log(2/\nu) - \psi(\nu/2) \xrightarrow{a.s.} E(\log \sigma_1^2) \quad as\ p \to \infty.$$

By the above two results, it yields that

$$\frac{1}{p}\sum_{j=1}^p \log s_j^2 - \log(2/\nu) - \psi(\nu/2) - \frac{1}{p}\sum_{j=1}^p \log \sigma_j^2 \xrightarrow{a.s.} 0 \quad as\ p \to \infty.$$

Finally, we have

$$\begin{aligned} \frac{1}{p}\left\{\hat{\theta}_{(1)} + C_0 - \log|\mathrm{diag}(\Sigma)|\right\} &= \frac{1}{p}\sum_{j=1}^p \log s_j^2 - \log(2/\nu) - \psi(\nu/2) - \frac{1}{p}\sum_{j=1}^p \log \sigma_j^2 \\ &\xrightarrow{a.s.} 0 \quad as\ p \to \infty. \quad \square \end{aligned}$$

Figure 1: Log MSEs for data from normal distribution with p=50. The sample size ranges from 5 to 50. In all figures, "1" to "8" represent the eight methods: DE (Bickel and Levina, 2004), IDE, USIE (Schäfer and Strimmer, 2005), CSIE (Fisher and Sun, 2011), USDE (Schäfer and Strimmer, 2005), CSDE (Fisher and Sun, 2011), ATE (Cai and Liu, 2011), and POET (Fan et al., 2013), respectively.

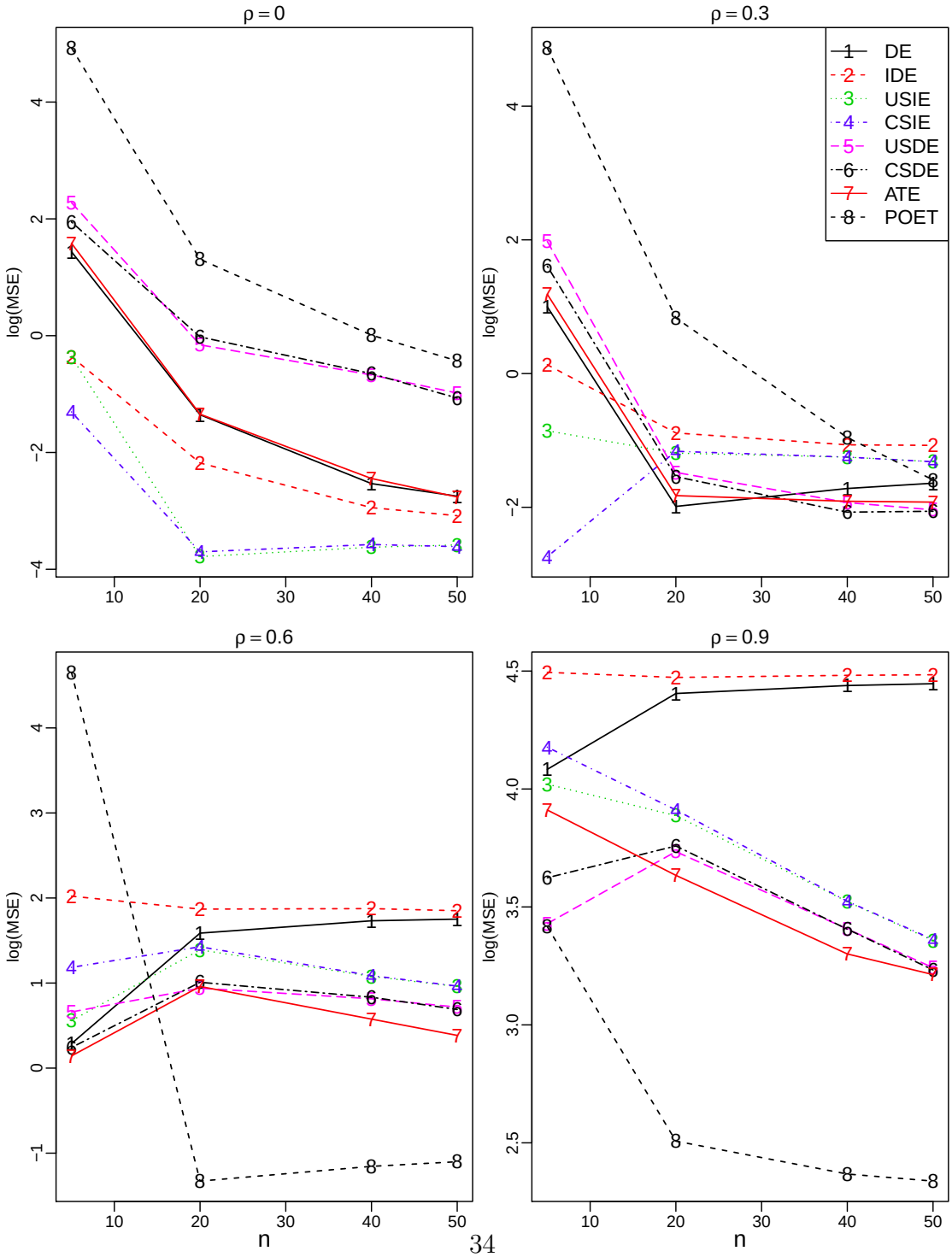

Figure 2: Log MSEs for data from normal distribution with p=300. The sample size ranges from 10 to 200. In all figures, "1" to "8" represent the eight methods: DE (Bickel and Levina, 2004), IDE, USIE (Schäfer and Strimmer, 2005), CSIE (Fisher and Sun, 2011), USDE (Schäfer and Strimmer, 2005), CSDE (Fisher and Sun, 2011), ATE (Cai and Liu, 2011), and POET (Fan et al., 2013), respectively.

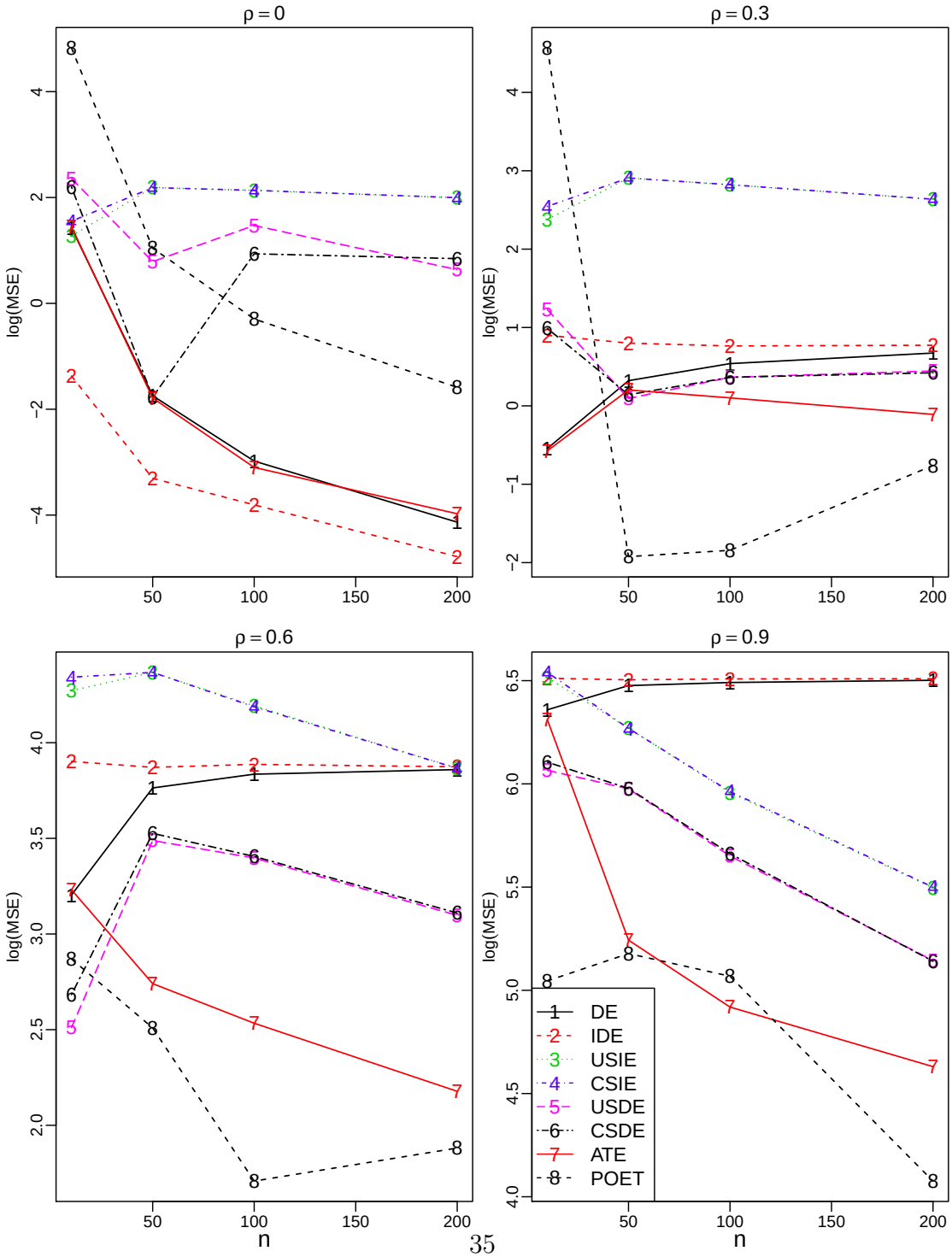

Figure 3: Log MSEs for data from normal distribution with p=300, and ρ ranging from 0 to 0.9. In all figures, "1" to "8" represent the eight methods: DE (Bickel and Levina, 2004), IDE, USIE (Schäfer and Strimmer, 2005), CSIE (Fisher and Sun, 2011), USDE (Schäfer and Strimmer, 2005), CSDE (Fisher and Sun, 2011), ATE (Cai and Liu, 2011), and POET (Fan et al., 2013), respectively.

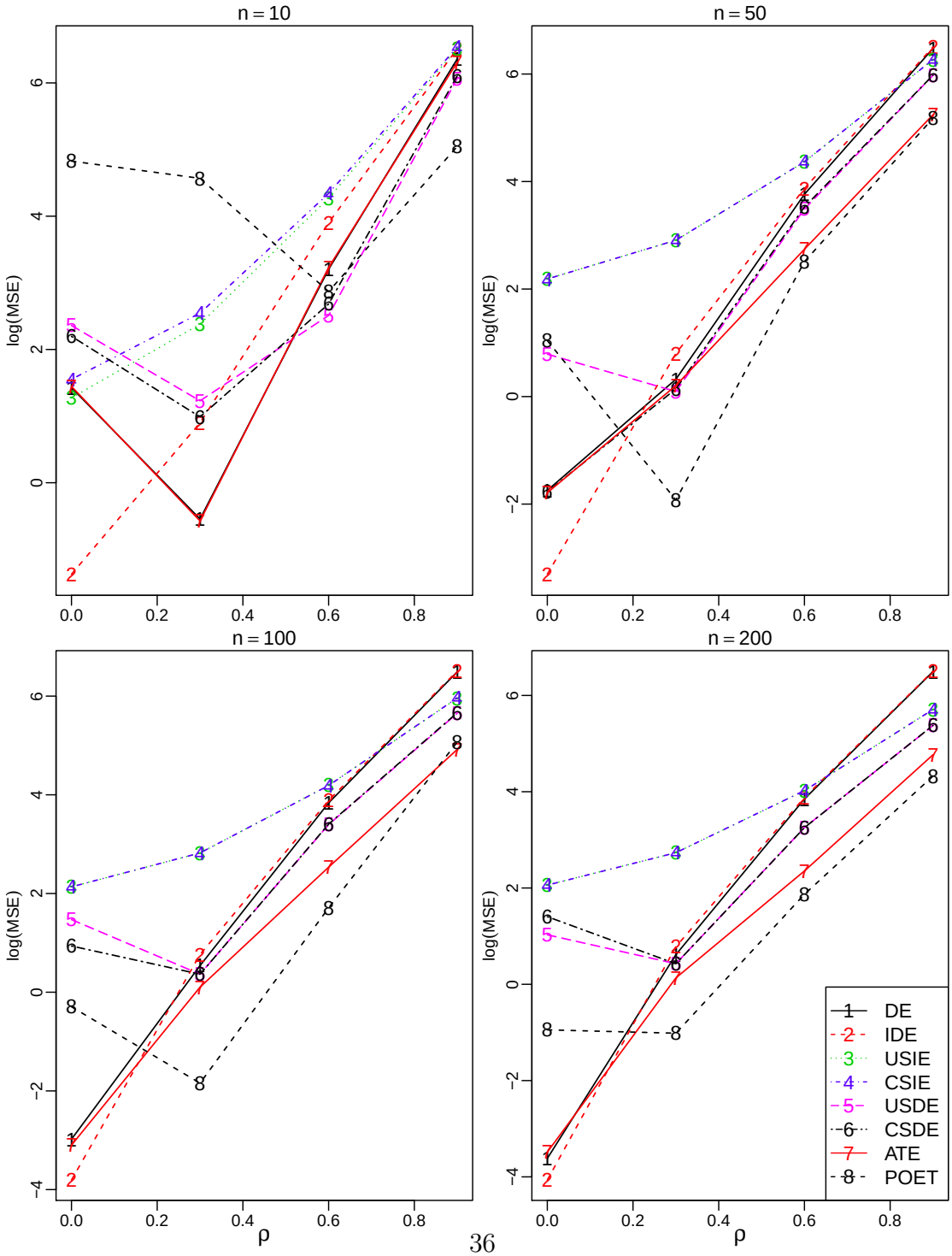

Table 1: MSEs of $\hat{\theta}$ for data from normal distribution with $\rho = 0.3, 0.6, 0.9$, $n = 10, 40$ and $p = 50, 100$, respectively. The number of factors K is either fixed or estimated by the method in Fan et al. (2013), denoted by $\hat{K}$. All MSEs are rounded to integer numbers. The minimum MSE of each line is highlighted.

		ρ	$K=0$	$K=1$	$K=2$	$K=4$	$K=6$	$K=\hat{K}$
$n=10$	$p=50$	0.3	**17**	272	660	2752	8820	667
		0.6	156	**38**	246	2543	7771	288
		0.9	2675	1179	**338**	710	4667	375
	$p=100$	0.3	**33**	944	2487	14261	37946	2481
		0.6	862	**58**	665	9643	30972	673
		0.9	14031	7386	2733	**695**	14648	2767
		ρ	$K=0$	$K=1$	$K=4$	$K=8$	$K=12$	$K=\hat{K}$
$n=40$	$p=50$	0.3	7	**5**	82	339	752	18
		0.6	91	44	**17**	241	691	15
		0.9	1359	909	**109**	272	558	531
	$p=100$	0.3	38	**8**	203	1066	3200	37
		0.6	526	272	**18**	529	2159	140
		0.9	6457	3816	712	**81**	1170	2303

Table 2: The time consumption of computing $\hat{\theta}$ with DE (Bickel and Levina, 2004), IDE, USIE (Schäfer and Strimmer, 2005), CSIE (Fisher and Sun, 2011), USDE (Schäfer and Strimmer, 2005), CSDE (Fisher and Sun, 2011), ATE (Cai and Liu, 2011), and POET (Fan et al., 2013), respectively. In ATE and POET, the turning parameter was selected based on 5-fold cross validation. The data is generated as described in Section 3.1. Timings (seconds) of 10 runs with Intel Core(TM) 3.20GH processor.

$n=100, p=300$	DE	IDE	USIE	CSIE	USDE	CSDE	ATE	POET
$\rho=0.0$	0.52	0.59	16.3	0.70	15.8	0.71	258	359
$\rho=0.9$	0.50	0.55	16.1	0.71	16.0	0.67	259	361

Figure 4: Log MSEs for data from mixture normal distribution with p=50, and ρ ranging from 0 to 0.9. In all figures, "1" to "8" represent the eight methods: DE (Bickel and Levina, 2004), IDE, USIE (Schäfer and Strimmer, 2005), CSIE (Fisher and Sun, 2011), USDE (Schäfer and Strimmer, 2005), CSDE (Fisher and Sun, 2011), ATE (Cai and Liu, 2011), and POET (Fan et al., 2013), respectively.

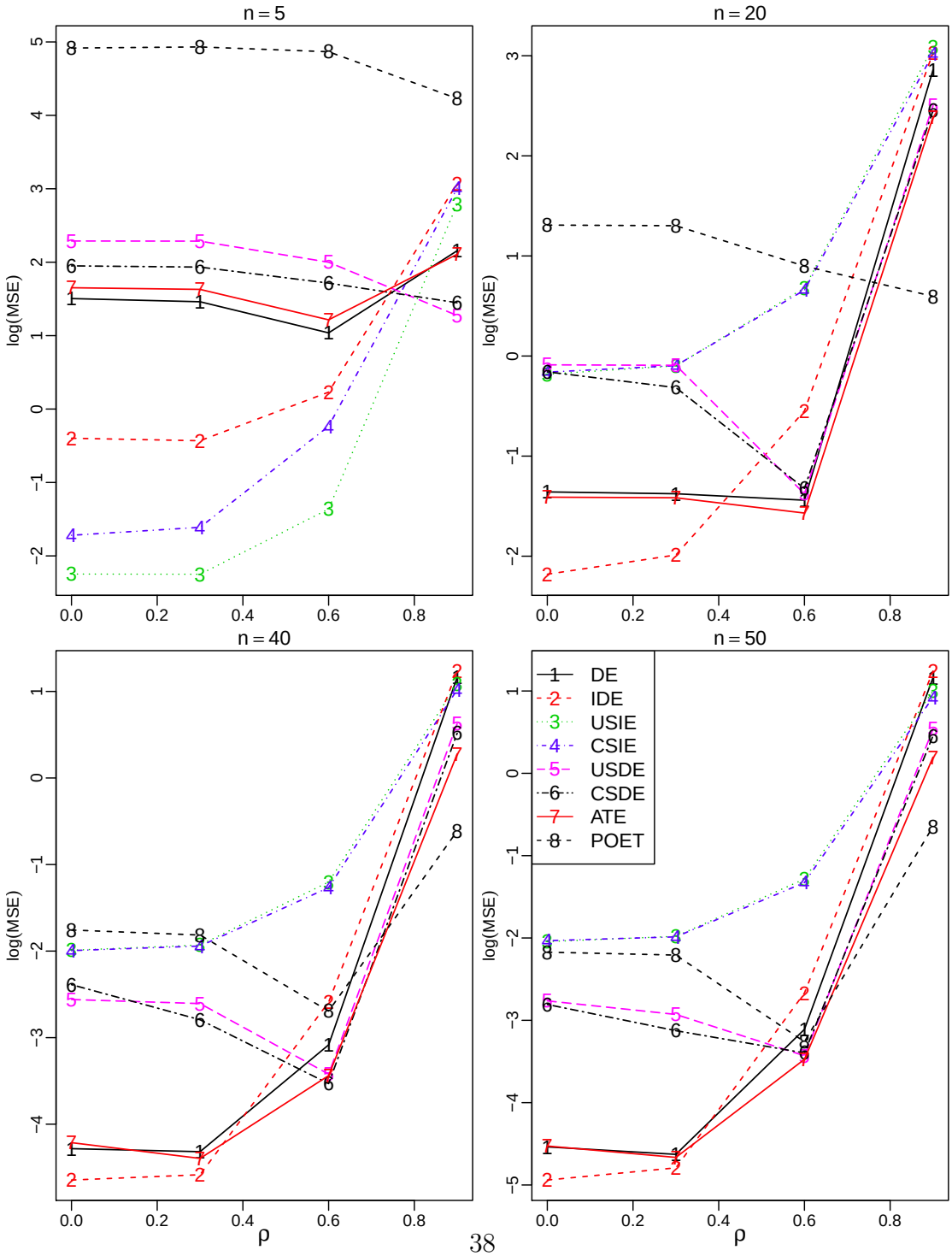

Figure 5: Log MSEs for data from mixture normal distribution with p=300, and ρ ranging from 0 to 0.9. In all figures, "1" to "8" represent the eight methods: DE (Bickel and Levina, 2004), IDE, USIE (Schäfer and Strimmer, 2005), CSIE (Fisher and Sun, 2011), USDE (Schäfer and Strimmer, 2005), CSDE (Fisher and Sun, 2011), ATE (Cai and Liu, 2011), and POET (Fan et al., 2013), respectively.

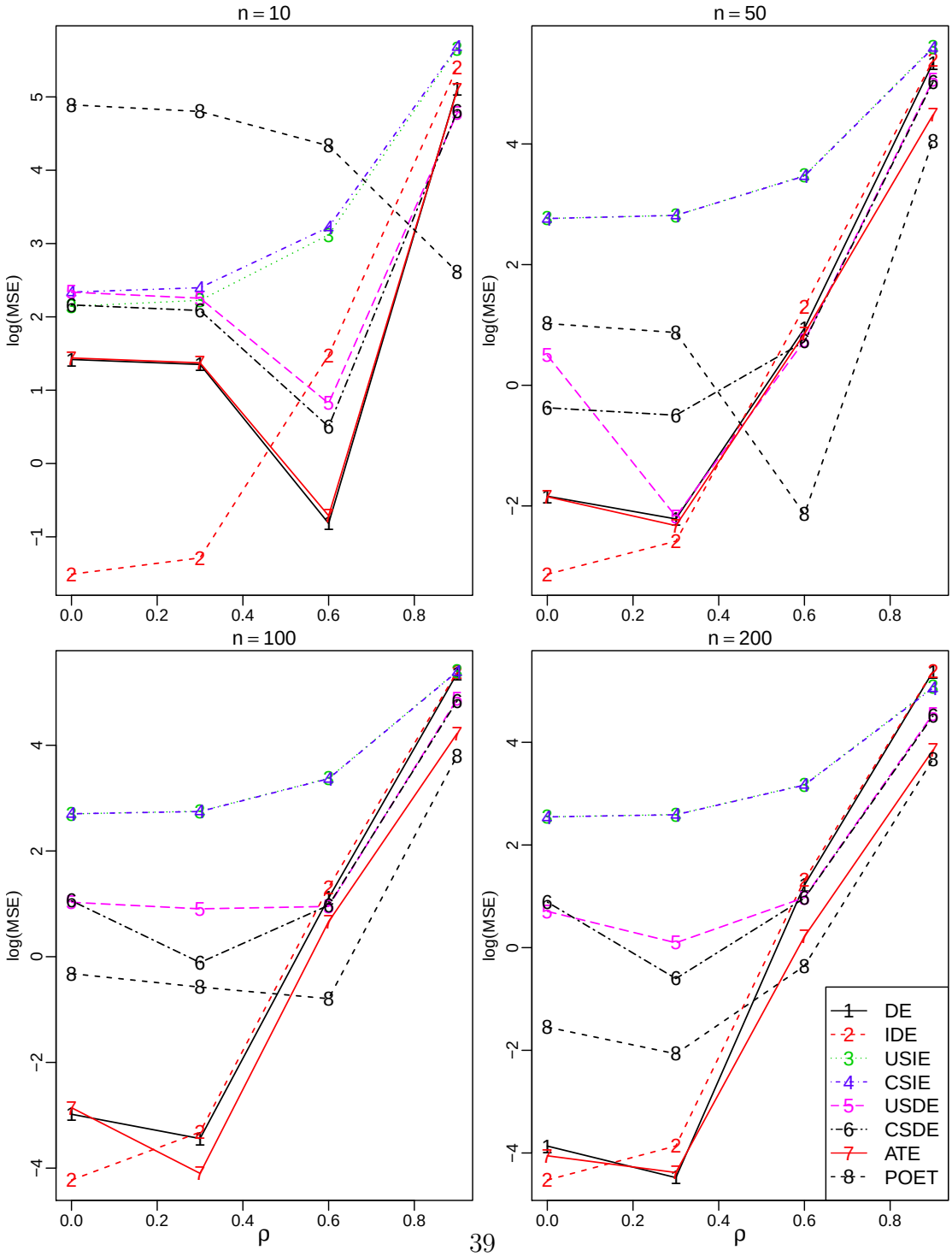

Figure 6: Log MSEs for data from heavy-tailed distribution with p=50, and ρ ranging from 0 to 0.9. In all figures, "1" to "8" represent the eight methods: DE (Bickel and Levina, 2004), IDE, USIE (Schäfer and Strimmer, 2005), CSIE (Fisher and Sun, 2011), USDE (Schäfer and Strimmer, 2005), CSDE (Fisher and Sun, 2011), ATE (Cai and Liu, 2011), and POET (Fan et al., 2013), respectively.

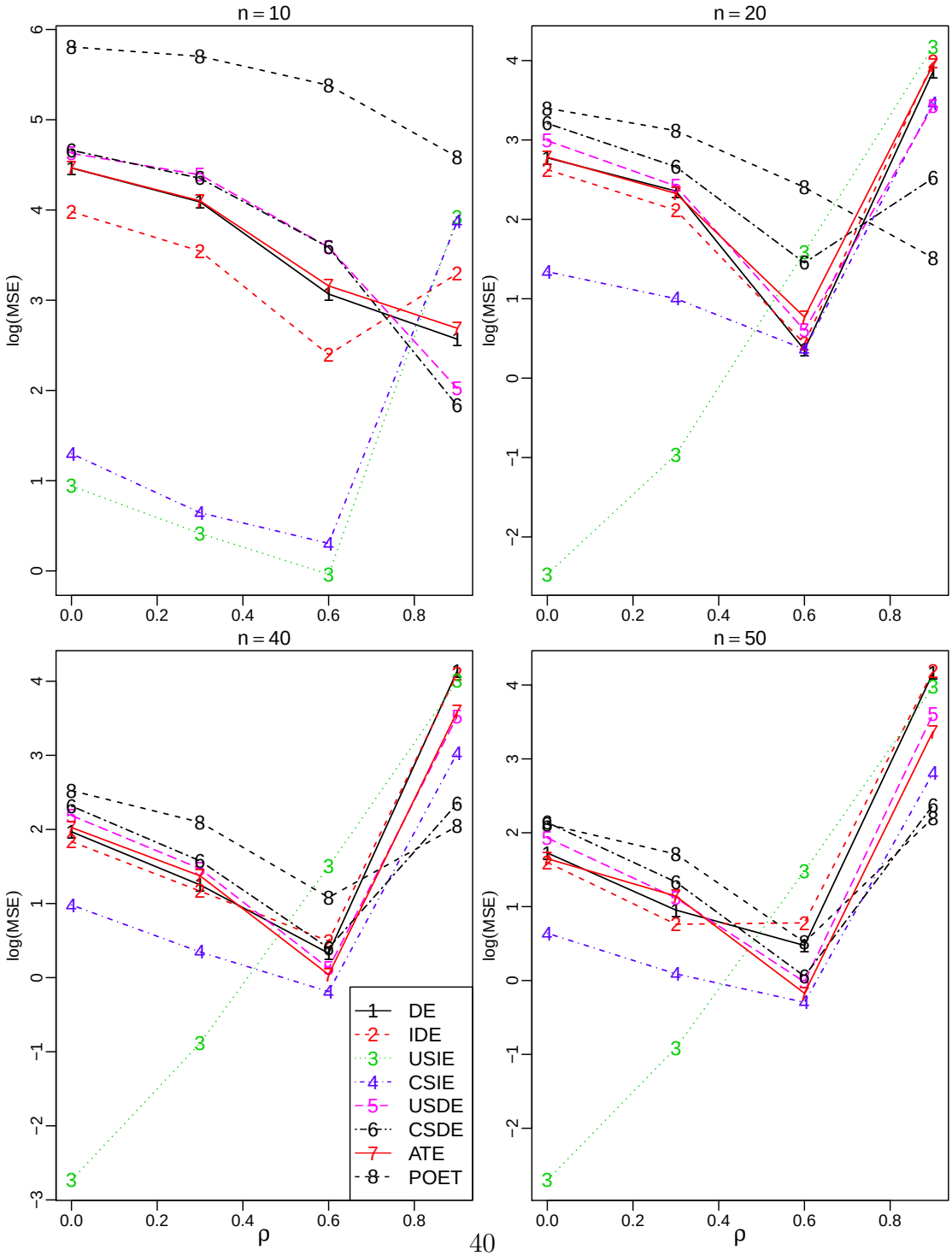

Figure 7: Log MSEs for data from heavy-tailed distribution with p=300, and ρ ranging from 0 to 0.9. In all figures, "1" to "8" represent the eight methods: DE (Bickel and Levina, 2004), IDE, USIE (Schäfer and Strimmer, 2005), CSIE (Fisher and Sun, 2011), USDE (Schäfer and Strimmer, 2005), CSDE (Fisher and Sun, 2011), ATE (Cai and Liu, 2011), and POET (Fan et al., 2013), respectively.

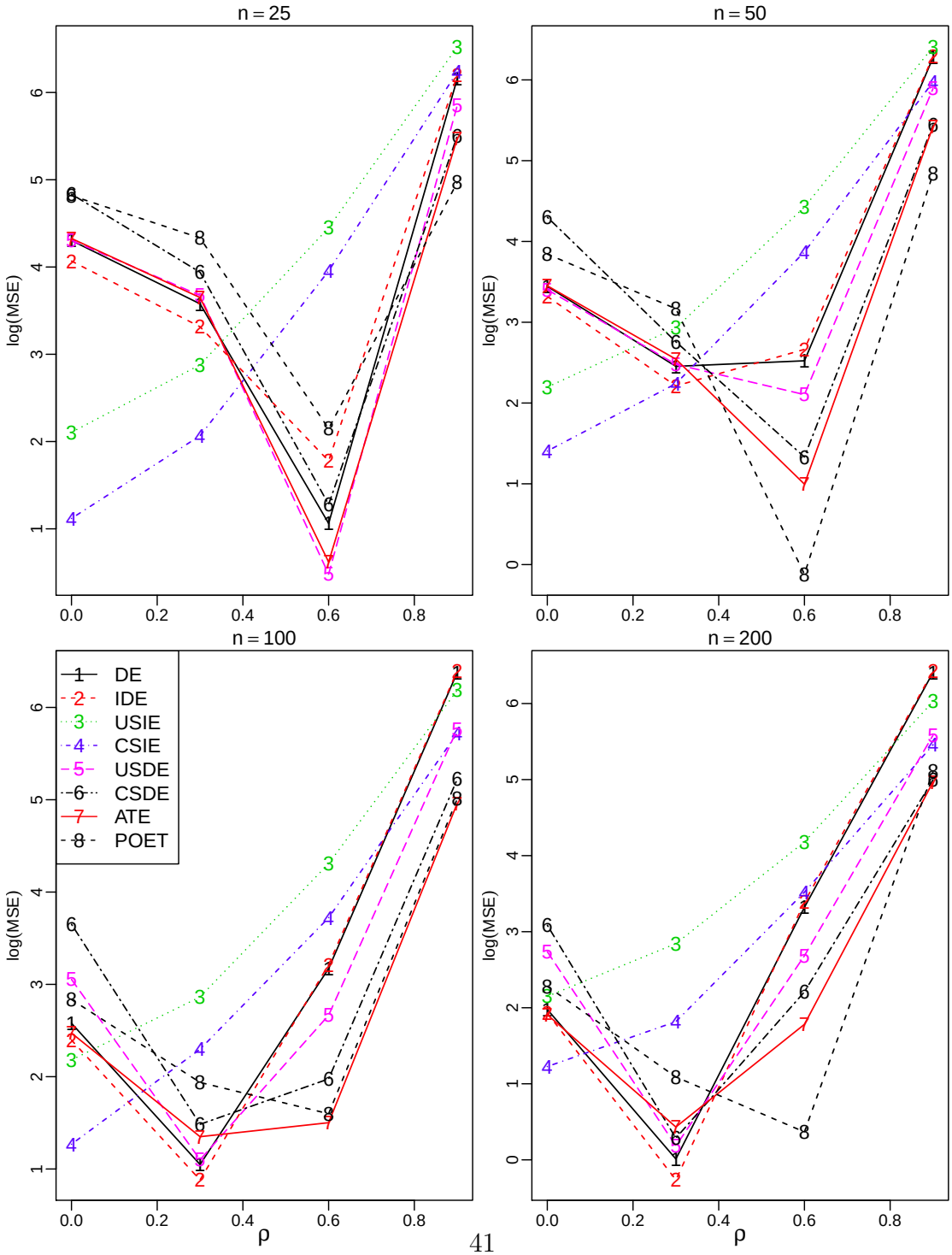

Figure 8: Log MSEs for data from degenerate normal distribution with p=50. The sample size ranges from 5 to 50. In all figures, "1" to "8" represent the eight methods: DE (Bickel and Levina, 2004), IDE, USIE (Schäfer and Strimmer, 2005), CSIE (Fisher and Sun, 2011), USDE (Schäfer and Strimmer, 2005), CSDE (Fisher and Sun, 2011), ATE (Cai and Liu, 2011), and POET (Fan et al., 2013), respectively.

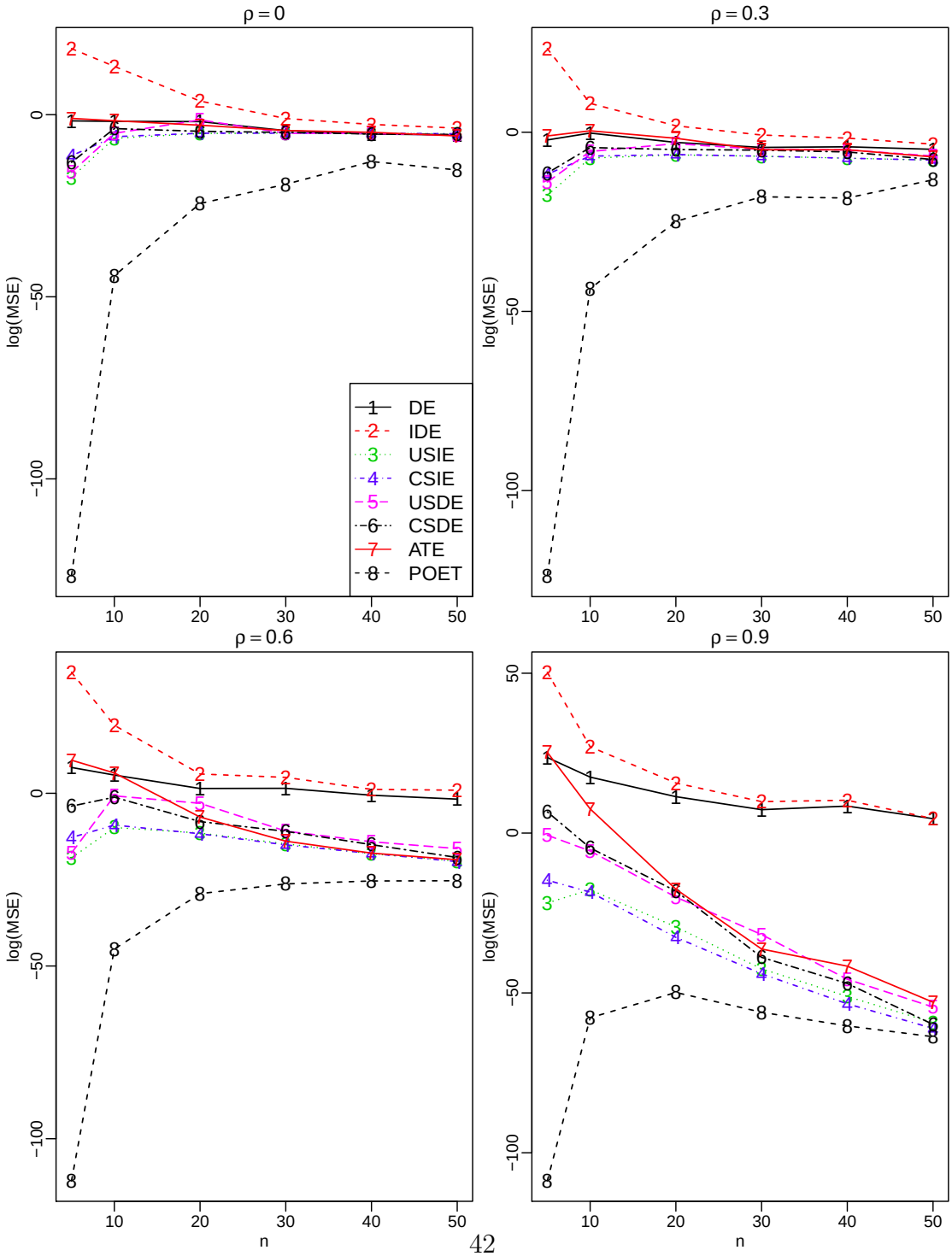

Figure 9: Log MSEs for real data with p=100. The sample size ranges from 10 to 80. In all figures, "1" to "8" represent the eight methods: DE (Bickel and Levina, 2004), IDE, USIE (Schäfer and Strimmer, 2005), CSIE (Fisher and Sun, 2011), USDE (Schäfer and Strimmer, 2005), CSDE (Fisher and Sun, 2011), ATE (Cai and Liu, 2011), and POET (Fan et al., 2013), respectively.