Biometrics 000, 000–000 DOI: 000
000 0000

A Compound Decision Approach to Covariance Matrix Estimation

Huiqin Xin* and Sihai Dave Zhao**

Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, Illinois

**email:* huiqinx2@illinois.edu

***email:* sdzhao@illinois.edu

Summary: We propose an empirical Bayes method to estimate high-dimensional covariance matrices. Our procedure centers on vectorizing the covariance matrix and treating matrix estimation as a vector estimation problem. Drawing from the compound decision theory literature, we introduce a new class of decision rules that generalizes several existing procedures. We then use a nonparametric empirical Bayes g-modeling approach to estimate the oracle optimal rule in that class. This allows us to let the data itself determine how best to shrink the estimator, rather than shrinking in a pre-determined direction such as toward a diagonal matrix. Simulation results and a gene expression network analysis shows that our approach can outperform a number of state-of-the-art proposals in a wide range of settings, sometimes substantially.

Key words: Compound decision theory; g-modeling; nonparametric maximum likelihood; separable decision rule.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Covariance matrix estimation is a fundamental statistical problem that plays an essential role in various applications, such as portfolio management (**?**), genomics (**?**), and array signal processing (**?**). However, in modern problems the number of features can be of the same order as or exceed the sample size, and the standard sample covariance matrix estimator behaves poorly in this regime. To overcome these issues, various methods have been developed to estimate high-dimensional covariance matrix. These can roughly be divided into two groups, according to whether they impose assumptions about the structure of population covariance matrix.

Structured methods make structural assumptions about the population covariance matrix. One class models the population covariance matrix as sparse. The most common method to address this problem is thresholding (**??**). Penalized likelihood methods (**?**) can also estimate large-scale sparse covariance matrix by penalizing a log-likelihood function. Another class of methods assume the data arise from a factor model (**?**), so that the covariance matrix has low intrinsic dimension. Other common structured methods assume that the covariance matrix is banding (**?**) or Toeplitz matrix (**?**).

In contrast, unstructured methods do not make any assumptions about the population covariance matrix, yet can still outperform the sample covariance matrix. A first example was the linear shrinkage approach of **?**, which shrinks the sample covariance matrix toward a scaled identity matrix. More recently, nonlinear shrinkage methods were developed (**???**). These shrink the eigenvalues of the sample covariance matrix toward clusters. Linear shrinkage can be viewed as a special case of nonlinear shrinkage, as it shrinks sample eigenvalues toward their global mean.

Nonlinear shrinkage estimators have desirable optimality properties (**?**) and show excellent performance. However, they modify only the sample eigenvalues and not the sample

eigenvectors. It is known that sample eigenvectors are not consistent estimators of population eigenvectors when the dimension and the sample size increase at the same rate (**?**). This suggests that there may exist a class of unstructured estimators that can outperform nonlinear shrinkage.

Here we propose a new unstructured estimator for high-dimensional covariance matrices. Our approach centers on vectorizing the covariance matrix and treating matrix estimation as a vector estimation problem. We do this because it allows us to use a nonparametric empirical Bayes shrinkage procedure, which has been shown in the compound decision literature to have excellent properties (**??**). We then reassemble the estimated vector into matrix form and project onto the space of positive-definite matrices to give our final estimator. Surprisingly, though our vectorized approach essentially ignores the matrix structure, it can still substantially outperform a number of state-of-the-art proposals in simulations and a real data analysis.

The article is organized as follows. In Section 2, we briefly review compound decision theory and then introduce our proposed approach. In Section 3 we illustrate the performance of our method in simulations and a gene expression dataset. Finally, Section 4 concludes with a discussion. Our procedure is implemented in the R package `cole`, available on GitHub.

2. Method

2.1 *Compound decision problem formulation*

Suppose we have n observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ independently generated from a p-dimensional $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. The purpose of this paper is to find an estimator $\boldsymbol{\delta}(\boldsymbol{X})$ of $\boldsymbol{\Sigma}$ that minimizes the scaled squared Frobenius risk

$$R(\boldsymbol{\Sigma}, \boldsymbol{\delta}) = \frac{1}{p^2} \sum_{j,k=1}^{p} \mathbb{E}[\{\delta_{jk}(\boldsymbol{X}) - \sigma_{jk}\}^2], \tag{1}$$

where σ_{jk} is the jkth entry of $\boldsymbol{\Sigma}$ and $\delta_{jk}(\boldsymbol{X})$ is its corresponding estimate.

Our proposed approach is motivated by two observations. First, 1 shows that estimating $\boldsymbol{\Sigma}$ under Frobenius risk is equivalent to simultaneously estimating every component of the vector $(\sigma_{11}, \ldots, \sigma_{pp})^\top$ under a loss function that aggregates errors across components. Second, this type of vector estimation problem has been well-studied in the compound decision literature. Thus, recent advances in vector estimation may be profitably applied to covariance matrices.

We first briefly review compound decision problems. Introduced by **?**, these problems study the simultaneous estimation of multiple parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ given data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, with $Y_i \sim P_{\theta_i}$. Specifically, the goal is to develop a decision rule $\boldsymbol{\delta}(\boldsymbol{Y}) = (\delta_1(\boldsymbol{Y}), \ldots, \delta_n(\boldsymbol{Y}))$ that minimizes the compound risk

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} L(\theta_i, \delta_i(\boldsymbol{Y})) \tag{2}$$

where L is a loss function measuring the accuracy of $\delta_i(\boldsymbol{Y})$ as an estimate of θ_i. A classical example is the homoscedastic Gaussian sequence problem, where $Y_i \sim N(\theta_i, 1)$ independently and $L(t, d) = (t - d)^2$ is squared error loss (**?**).

A key property of compound decision problems is that while a given Y_i seems to offer no information about any specific θ_j when $j \neq i$, borrowing information across all components of $\boldsymbol{Y}$ to estimate $\boldsymbol{\theta}$ is superior to estimating each θ_i using the corresponding Y_i alone. A classical example of this phenomenon is the James-Stein estimator (**?**), which estimates $\boldsymbol{\theta}$ in the Gaussian sequence problem by shrinking each Y_i toward 0 by a factor that depends on all components of $\boldsymbol{Y}$. It is well-known that when $n \geqslant 3$, the James-Stein estimator dominates the maximum likelihood estimator, which simply estimates $\boldsymbol{\theta}$ using $\boldsymbol{Y}$. A long line of subsequent work has led to much more sophisticated and accurate procedures for estimating $\boldsymbol{\theta}$ (**?????**).

We propose to apply some of these recent vector estimation ideas to covariance matrix estimation. Clearly, covariance matrix estimation under the Frobenius risk 1 can be viewed as a compound decision problem. Furthermore, some existing covariance matrix estimation procedures can already be interpreted as taking a vector approach. The sample covariance

matrix $\boldsymbol{S}$, for instance, can be thought of as estimating each component of $(\sigma_{11}, \ldots, \sigma_{pp})^\top$ using maximum likelihood. Less trivially, **?** studied sparse high-dimensional covariance matrices and explicitly appealed to the vector perspective. Their adaptive thresholding method is a version of the soft thresholding method of **?**, which was originally developed to estimate a sparse mean vector in the Gaussian sequence problem.

Interestingly, we can also show that the celebrated linear shrinkage covariance matrix estimator of **?** can be interpreted as a vector estimator. The estimator is defined as

$$\widehat{\boldsymbol{\Sigma}}_{\text{LW}} = \widehat{\rho}_{\text{LW}} \boldsymbol{S} + (1 - \widehat{\rho}_{\text{LW}}) \widehat{\mu} \boldsymbol{I} \tag{3}$$

where $\widehat{\mu} = \text{tr}(\boldsymbol{S})/p$, and $\widehat{\rho}_{\text{LW}} = 1 - b_n^2/d_n^2$ for $d_n^2 = \|\boldsymbol{S} - \widehat{\mu}\boldsymbol{I}\|_F^2$ and $b_n^2 = \sum_{i=1}^n \|\boldsymbol{X}_i \boldsymbol{X}_i^\top - \boldsymbol{S}\|_F^2/n^2$. Now consider the problem of estimating the vectorized $\boldsymbol{\Sigma}$ under risk 1. We restrict attention to decision rules that estimate each component σ_{jk} using

$$\beta_S s_{jk} + \beta_I u_{jk}, \tag{4}$$

where s_{jk} is the jkth entry of $\boldsymbol{S}$, u_{jk} is the jkth entry of $\boldsymbol{I}$, and the class is indexed by the parameters (β_S, β_I). It is straightforward to show that

$$\widehat{R}(\beta_S, \beta_I) = \frac{1}{p^2} \sum_{j,k=1}^{p} [(2\beta_S - 1)\widehat{\Delta}_{jk}^2 + \{(1 - \beta_S)s_{jk} - \beta_I u_{jk}\}^2]$$

is an unbiased estimate of the risk 1, where $\widehat{\Delta}_{jk}^2 = \sum_{i=1}^n (X_{ij}X_{ik} - s_{jk})^2/n$ is the sample variance of s_{jk}. The optimal estimator in this class can now be chosen by minimizing $\widehat{R}(\beta_S, \beta_I)$ over β_S and β_I. It can be shown that this is equivalent to estimating the vector $(\sigma_{11}, \ldots, \sigma_{pp})^\top$ by shrinking $(s_{11}, \ldots, s_{pp})^\top$ toward the one-dimensional subspace spanned by $(u_{11}, \ldots, u_{pp})^\top$ (**??**). Proposition 1 shows that this subspace shrinkage estimator is identical to the **?** estimator 3.

PROPOSITION 1: Define the estimator $\widehat{\boldsymbol{\Sigma}}_V$ such that its jkth entry obeys $\widehat{\sigma}_{jk} = \widehat{\beta}_S s_{jk} + \widehat{\beta}_I u_{jk}$, where $(\widehat{\beta}_S, \widehat{\beta}_I) = \arg\min_{\beta_S, \beta_I} \widehat{R}(\beta_S, \beta_I)$. Then $\widehat{\boldsymbol{\Sigma}}_V = \widehat{\boldsymbol{\Sigma}}_{LW}$.

2.2 *Proposed estimator*

The previous section argues that treating covariance matrix estimation as a vector estimation problem can be a fruitful strategy, but discusses only estimators linear in s_{jk}. We propose to consider a larger class, the class of so-called separable rules. In the standard compound decision problem of estimating $\boldsymbol{\theta}$ using $\boldsymbol{Y}$, a separable decision rule $\boldsymbol{\delta}(\boldsymbol{Y})$ is one where $\delta_i(\boldsymbol{Y}) = t(Y_i)$ (**?**). Here we generalize this to the problem of estimating a vectorized matrix. For decision rules $\boldsymbol{\delta}(\boldsymbol{X}) = (\delta_{11}(\boldsymbol{X}), \ldots, \delta_{pp}(\boldsymbol{X}))$ that estimate $(\sigma_{11}, \ldots, \sigma_{pp})^\top$, define the class of separable rules

$$\mathcal{S} = \{\boldsymbol{\delta} : \delta_{jk} = t(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k}), j \neq k, \quad \delta_{jj} = \widetilde{t}(\boldsymbol{X}_{\cdot j}), j = 1, \ldots, p)\}, \tag{5}$$

where $\boldsymbol{X}_{\cdot j} = (X_{1j}, \ldots, X_{nj})^\top$ is the vector of observed values of the jth feature. In other words, rules in $\mathcal{S}$ estimate the jkth entry of the covariance matrix using a fixed function t of only observations from features j and k. The sufficient statistics of $(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k})$ is 2×2 matrix $\begin{bmatrix} s_{jj} & s_{jk} \\ s_{kj} & s_{kk} \end{bmatrix}$.

We propose to search for the optimal estimator within $\mathcal{S}$. This is sensible because $\mathcal{S}$ includes the sample covariance $(s_{11}, \ldots, s_{pp})^\top$, the class of linear estimators 4 used by **?**, which can be expressed as

$$\delta_{jk}(\boldsymbol{X}) = \beta_S \boldsymbol{X}_{\cdot j}^\top \boldsymbol{X}_{\cdot k}/n + \beta_I I(\boldsymbol{X}_{\cdot j} = \boldsymbol{X}_{\cdot k}),$$

and the class of adaptive thresholding estimators for sparse covariance matrices studied by **?**. Therefore the optimal separable estimator $\boldsymbol{\delta}^\star$ that minimizes the scaled squared Frobenius risk 1 will perform at least as well as these three estimators, and may perform better. Targeting the optimal separable rule is standard in the compound decision literature (**?**).

The optimal $\boldsymbol{\delta}^\star$ is an oracle estimator and cannot be calculated in practice, as the true risk is unknown. In the classical compound decision framework, empirical Bayes methods are used to estimate the oracle optimal separable rule (**??????**). We take a similar approach

here. To simplify notation, denote $(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k})^\top$ as $\boldsymbol{A}_{jk}$. $f(\cdot \mid \boldsymbol{\eta}_{jk})$, the density of $\boldsymbol{A}_{jk}$, depends on $\boldsymbol{\eta}_{jk} = (\sigma_j, \sigma_k, r_{jk})^\top$, where σ_j and σ_k are the true standard deviations of the jth and kth covariates and $r_{jk} = \sigma_{jk}/(\sigma_j\sigma_k)$ is their true correlation. When $r_{jk} \neq 1$, $\boldsymbol{A}_{jk}$ is comprised of n independent mean-zero multivariate normals with covariance matrices

$$\boldsymbol{C}_{jk} = \begin{bmatrix} \sigma_j^2 & \sigma_j\sigma_k r_{jk} \\ \sigma_j\sigma_k r_{jk} & \sigma_k^2 \end{bmatrix}.$$

When $r_{jk} = 1$, $\boldsymbol{A}_{jk}$ consists of mean-zero univariate normals with variances σ_j^2.

Now consider the Bayesian model for non-diagonal entries

$$\boldsymbol{A} \mid \boldsymbol{\eta} \sim f(\cdot \mid \boldsymbol{\eta}), \quad \boldsymbol{\eta} \sim G_{nd}(a, b, \gamma) = \frac{2}{p(p-1)} \sum_{1 \leqslant k < j \leqslant p} I(\sigma_j \leqslant a, \sigma_k \leqslant b, r_{jk} \leqslant \gamma). \tag{6}$$

and diagonal entries

$$\boldsymbol{A} \mid a \sim \widetilde{f}(\cdot \mid a), \quad a \sim G_d(a) = \frac{1}{p} \sum_{j=1}^{p} I(\sigma_j \leqslant a). \tag{7}$$

By the fundamental theorem of compound decisions (**??**), this is closely related to the vectorized covariance matrix estimation problem under Frobenius risk 1: for any separable $\boldsymbol{\delta} \in \mathcal{S}$ 5, the Frobenius risk can be written as

$$\begin{aligned} R(\boldsymbol{\Sigma}, \boldsymbol{\delta}) &= \frac{1}{p^2}\{ \sum_{1 \leqslant k < j \leqslant p}^{p} \int \{t(\boldsymbol{A}) - \sigma_{jk}\}^2 f(\boldsymbol{A} \mid \boldsymbol{\eta}_{jk}) d\boldsymbol{A} + \sum_{j=1}^{p} \int \{\widetilde{t}(\boldsymbol{X}) - \sigma_{jj}\}^2 \widetilde{f}(\boldsymbol{X} \mid s_{jj}) d\boldsymbol{X}\} \\ &= \frac{p-1}{p} \int \int \{t(\boldsymbol{A}) - g(\boldsymbol{\eta})\}^2 f(\boldsymbol{A} \mid \boldsymbol{\eta}) dG_{nd}(\boldsymbol{\eta}) d\boldsymbol{A} + \frac{1}{p} \int \int \{\widetilde{t}(\boldsymbol{X}) - a^2\}^2 \widetilde{f}(\boldsymbol{X} \mid a) dG_d(a) d\boldsymbol{X} \\ &= \frac{p-1}{p} \mathbb{E}[\{t(\boldsymbol{A}) - g(\boldsymbol{\eta})\}^2] + \frac{1}{p} \mathbb{E}[\{\widetilde{t}(\boldsymbol{X}) - a^2\}^2], \end{aligned}$$

where $g(a, b, \gamma) = ab\gamma$ and the final expectation is the Bayes risk of estimating σ_{jk}. The optimal oracle separable rule $\boldsymbol{\delta}^\star$ therefore has jkth entry equal to $\delta^\star_{jk}(\boldsymbol{X}) = t^\star(\boldsymbol{A}_{jk})$ for $j \neq k$ and $\delta^\star_{jj}(\boldsymbol{X}) = \widetilde{t}^\star(\boldsymbol{X}_{\cdot j})$, where $t^\star = \mathbb{E}\{g(\boldsymbol{\eta}) \mid \boldsymbol{A}\}$ and $\widetilde{t}^\star = \mathbb{E}\{g(a) \mid \boldsymbol{X}\}$ minimizes the Bayes risk.

Based on this result, we propose the following empirical Bayes procedure. We first use nonparametric maximum likelihood (**?**) to estimate the prior G_p. Under the Bayesian model **??**, and the working assumption that the $\boldsymbol{A}_{jk}$ are independent across jk, we estimate G_{nd}

using

$$\widehat{G}_{nd} = \underset{G \in \mathcal{G}_{nd}}{\arg\max} \prod_{1 \leqslant k < j < p} \int f(\boldsymbol{A}_{jk} \mid \boldsymbol{\eta}) dG(\boldsymbol{\eta}), \tag{8}$$

estimate G_d using

$$\widehat{G}_{d} = \underset{G \in \mathcal{G}_{d}}{\arg\max} \prod_{j=1}^{p} \int \widetilde{f}(\boldsymbol{X} \mid a) dG(a), \tag{9}$$

where $\mathcal{G}_{nd}$ is the family of all distributions supported on $\mathbb{R}_+ \times \mathbb{R}_+ \times [-1, 1]$, $\mathcal{G}_d$ is the family of all distributions supported on $\mathbb{R}$. Of course, the $\boldsymbol{A}_{jk}$ are not independent, so $\widehat{G}_p$ does not maximize a likelihood but rather a pairwise composite likelihood (**?**). Using $\widehat{G}_p$, we estimate the vectorized $\boldsymbol{\Sigma}$ using $\widehat{\boldsymbol{\delta}}(\boldsymbol{X}) = (\widehat{t}(\boldsymbol{A}_{21}), \ldots, \widehat{t}(\boldsymbol{A}_{p,p-1}), \widehat{t}(\boldsymbol{X}_{\cdot 1}), \ldots, \widehat{t}(\boldsymbol{X}_{\cdot p}))$, where

$$\widehat{t}(\boldsymbol{A}_{jk}) = \frac{\int g(\boldsymbol{\eta}) f(\boldsymbol{A}_{jk} \mid \boldsymbol{\eta}) d\widehat{G}_{nd}(\boldsymbol{\eta})}{\int f(\boldsymbol{A}_{jk} \mid \boldsymbol{\eta}) d\widehat{G}_{nd}(\boldsymbol{\eta})}, \quad \widehat{t}(\boldsymbol{X}_{\cdot j}) = \frac{\int a^2 \widetilde{f}(\boldsymbol{X}_{\cdot j} \mid a) d\widehat{G}_{d}(a)}{\int \widetilde{f}(\boldsymbol{X}_{\cdot j} \mid a) d\widehat{G}_{d}(a)}. \tag{10}$$

The $\widehat{t}$ estimates the Bayes rule $t^\star$ and $\widehat{\boldsymbol{\delta}}$ estimate the optimal oracle separable rule $\boldsymbol{\delta}^\star$.

Our proposed procedure is an example of what **?** calls g-modeling, an approach to empirical Bayes problems that proceeds by modeling the prior. A major advantage of nonparametric estiation of the prior is that it allows the data itself to determine how best to shrink the estimator. In contrast, most existing methods shrink in a pre-determined direction, such as toward a diagonal matrix in the case of **?**. Theoretical justification of our proposed $\widehat{\boldsymbol{\delta}}$ is difficult and is discussed in Section 4. Nevertheless, our numerical results in Section 3 show that in practice, our $\widehat{\boldsymbol{\delta}}$ can outperform many existing covariance matrix estimators.

2.3 *Implementation*

Calculating the estimated prior $\widehat{G}_p$ **??** is difficult, as it is an infinite-dimensional optimization problem over the class of all probability distributions $\mathcal{G}$. **?** showed that the solution is atomic and is supported on at most p^2 points. The EM algorithm has traditionally been used to estimate the locations of the support points and the masses at those points (**?**), but this is a difficult nonconvex optimization problem.

Instead, we maximize the pairwise composite likelihood over a fixed grid of support points,

similar to recent g-modeling procedures for standard compound decision problems; this restores convexity (**???**). Specifically, we assume that the prior for the $\boldsymbol{\eta}_{jk} = (\sigma_j, \sigma_k, r_{jk})^\top$ is supported on D fixed support points $\boldsymbol{\xi}_\tau$, $\tau = 1, \ldots, D$. We can then use the EM algorithm to estimate the masses $\widehat{\boldsymbol{w}} = \{\widehat{w}_1, \ldots, \widehat{w}_D\}$ at those points via the iteration

$$\widehat{w}_\tau^{(k)} = \frac{1}{p^2} \sum_{j,k=1}^{p} \frac{\widehat{w}_\tau^{(k-1)} f(\boldsymbol{A}_{jk} \mid \boldsymbol{\xi}_\tau)}{\sum_{l=1}^{D} \widehat{w}_l^{(k-1)} f(\boldsymbol{A}_{jk} \mid \boldsymbol{\xi}_l)}$$

over k. Early stopping of the EM algorithm can be useful (**?**), and more sophisticated convex optimization procedures can be used as well (**?**). Our proposed estimator 10 then becomes

$$\widehat{t}(\boldsymbol{A}_{jk}) = \frac{\sum_{\tau=1}^{D} g(\boldsymbol{\xi}_\tau) f(\boldsymbol{A} \mid \boldsymbol{\xi}_\tau) \widehat{w}_\tau}{\sum_{\tau=1}^{D} f(\boldsymbol{A} \mid \boldsymbol{\xi}_\tau) \widehat{w}_\tau}, \quad \widehat{t}(\boldsymbol{X}_{\cdot j}) = \frac{\sum_{\tau=1}^{D} a_\tau^2 \widetilde{f}(\boldsymbol{X}_{\cdot j} \mid a_\tau) \widehat{w}_\tau}{\sum_{\tau=1}^{D} \widetilde{f}(\boldsymbol{X}_{\cdot j} \mid a_\tau) \widehat{w}_\tau}.$$

Ideally, the grid points should be chosen to densely cover the parameter space. However, the fact that G_p is multivariate poses difficulties, as for example using a grid of d points in each dimension requires a total of $D = d^3$ grid points, which requires huge computational cost for even moderate d. Alternatively, we can use a so-called exemplar algorithm (**?**), which sets the support points to equal the observed sample versions $\widehat{\boldsymbol{\eta}}_{jk}$ of the $\boldsymbol{\eta}_{jk}$. This reduces the size of the support set, but even in this case the computation complexity grows like $O(p^2)$.

Here we propose a clustering-based exemplar algorithm to further improve computational efficiency. Let s_j, s_k, and γ_{jk} be the sample variances and correlation between the jth and kth covariates. We first apply K-means clustering to identify K clusters among the $p(p-1)/2$ off-diagonal sample points (s_j, s_k, γ_{jk}) and $\lceil K^{1/2} \rceil$ clusters among the p diagonal sample points $(s_j, s_j, 1)$. We then use the $K + \lceil K^{1/2} \rceil$ cluster centroids as our support points. We cluster the off- and on-diagonal observations separately to ensure that the support points $\boldsymbol{\xi}_\tau$ are such that $f(\boldsymbol{A}_{jk} \mid \boldsymbol{\xi}_\tau) \neq 0$ when both $j = k$ and $j \neq k$. Figure 1 shows that different K have similar estimation accuracy compared to the exemplar algorithm, while Table 1 shows that they can be significantly faster.

In our implementation, we use the density function of $\boldsymbol{S}_{jk} = \begin{bmatrix} s_{jj} & s_{jk} \\ s_{kj} & s_{kk} \end{bmatrix}$ instead of $\boldsymbol{A}_{jk}$

since it is the sufficient statistics of $\boldsymbol{A}_{jk}$ and easier to calculate. Based on the property of sufficient statistics,

$$\frac{f(\boldsymbol{A}_{jk} \mid \boldsymbol{\xi}_\tau)}{f(\boldsymbol{A}_{jk} \mid \boldsymbol{\xi}_l)} = \frac{f(\boldsymbol{S}_{jk} \mid \boldsymbol{\xi}_\tau)}{f(\boldsymbol{S}_{jk} \mid \boldsymbol{\xi}_l)} \tag{11}$$

So it is equivalent to use $f(\boldsymbol{A}_{jk} \mid \boldsymbol{\xi}_\tau)$ in calculation.

2.4 *Positive definiteness correction*

Our proposed estimator 10 is not guaranteed to be positive-definite. To correct this, we reshape our vector estimator back into a matrix and then identify the closest positive-definite matrix. **?** and **?** showed that the projection of a $p \times p$ symmetric matrix $\boldsymbol{B}$ onto the space of positive semi-definite matrices is

$$P_0(\boldsymbol{B}) = \arg\min_{\boldsymbol{A} \geqslant 0} \|\boldsymbol{A} - \boldsymbol{B}\| = \boldsymbol{Q}\text{diag}\{\max(\lambda_1, 0), \max(\lambda_2, 0), \ldots, \max(\lambda_p, 0)\}\boldsymbol{Q}^\top,$$

where $\|\cdot\|$ denotes the Frobenius norm, $\boldsymbol{Q}$ is the matrix of eigenvectors of $\boldsymbol{B}$, and $\lambda_1, \ldots, \lambda_p$ are its eigenvalues.

To guarantee positive-definiteness, we follow **?** and replace non-positive eigenvalues with a chosen positive value c smaller than the least positive eigenvalue $\lambda^+_{\min}$, so that the corrected estimate is

$$P_0(\boldsymbol{B}) = \boldsymbol{Q}\text{diag}\{\max(\lambda_1, c), \max(\lambda_2, c), \ldots, \max(\lambda_p, c)\}\boldsymbol{Q}^\top. \tag{12}$$

? suggest $c_\alpha = 10^{-\alpha}\lambda^+_{\min}$, where the parameter α is chosen to minimize $\|B - P_{c_\alpha}(B)\| + \alpha$ over a uniform partition of $\{\alpha_1, \ldots, \alpha_K\}$ of $[0, \alpha_K]$. In this paper we chose $K = 20$ and $\alpha_K = 10$.

3. Numerical Results

3.1 *Models*

We considered five models for the population covariance matrix. For the first four settings, $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{s})\boldsymbol{C}\text{diag}(\boldsymbol{s})$, where $\boldsymbol{C}$ is correlation matrix and $\boldsymbol{s}$ is a vector of standard deviations.

- Model 1. The standard deviations were independently generated from $\mathcal{U}(1, 1.5)$ and the

correlation matrix followed Model 2 of **?**:

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{A}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{p/2\times p/2} \end{pmatrix},$$

where the jkth entry of $\boldsymbol{A}_1$ is $a_{jk} = \max(1 - |j-k|/10, 0)$. This setting modeled a sparse covariance matrix.

- Model 2. The first $p/2$ standard deviations equaled 1, the last $p/2$ equaled 2, and the correlation matrix was

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{pmatrix},$$

where $\boldsymbol{C}_{11}$ and $\boldsymbol{C}_{22}$ were $p/2 \times p/2$ compound symmetric matrices with correlation parameters 0.8 and 0.2, respectively, and $\boldsymbol{C}_{12}$ and $\boldsymbol{C}_{21}$ were $p/2 \times p/2$ matrices with entries equal to 0.4. This model was designed such that larger σ_j and σ_k tended to correspond to larger r_{jk}.

- Model 3. The standard deviations were generated independently from $\mathcal{U}(1, 1.5)$ and $\boldsymbol{C}$ was a compound symmetric matrix with correlation parameter 0.7. This modeled a dense covariance matrix.
- Model 4. This setting was the same as Model 3 except with correlation parameter 0.9. This high level of dependence tested the robustness of the pairwise composite likelihood estimator **??**.
- Model 5. With $\boldsymbol{U}$ a randomly generated orthogonal matrix, $\boldsymbol{\Sigma} = \boldsymbol{U}^T \text{diag}(\boldsymbol{l})\boldsymbol{U}$, where $\boldsymbol{l}$ was a vector of eigenvalues where the first $p/2$ equaled 1 and the last $p/2$ equaled 4. This followed simulation settings from **?** and **?**.

In each scenario, we generated $n = 100$ samples from a p-variate $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $p = 30, 100$, or 200. We generated 200 replicates and reported average errors under the following three norms, where $\widehat{\boldsymbol{\Sigma}}$ is the estimated matrix with entries $\widehat{\sigma}_{jk}$ and $\boldsymbol{\Sigma}$ is the true matrix with entries σ_{jk}:

- Frobenius: $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F = \{\sum_{j,k=1}^{p}(\widehat{\sigma}_{jk} - \sigma_{jk})^2\}^{1/2}$, a version of 1,
- Spectral: $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 = \lambda_{\max}(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})$, the largest eigenvalue of $\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$, and
- Matrix ℓ_1: $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{L_1} = \max_{k=1,\ldots,p}\sum_{j=1}^{p}|\widehat{\sigma}_{jk} - \sigma_{jk}|$.

3.2 *Clustering-based exemplar algorithm*

We first studied the behavior of our K-means clustering-based exemplar algorithm for different K, described in Section 2.3. For a given p, we let $K = rp$ for different ratios ratios $r = 2, 1, 0.5, 0.25$. We compared these choices for K to the full exemplar method. For all these estimators, we show the result after applying positive-definiteness correction.

[Figure 1 about here.]

[Table 1 about here.]

Figure 1 presents the Frobenius norm error estimates from Model 1 to Model 5. Table 1 shows the running time only for Model 1, because the running time does not vary much across different models. The results show that different K exhibit similar performance and are comparable to the full exemplar method. Letting $K = p$ seemed to provide a good balance between accuracy and speed, so we implement our proposed method with $K = p$ in the rest of this paper.

3.3 *Methods compared*

In this section we refer to our approach using the abbreviation MSG: Matrix Shrinkage via G-modeling; we use MSGCor to refer to the version corrected for postive-definiteness. K-means clustering is applied on sample grids with $K = p$. We compared MSG and MSGCor to several existing high-dimensional covariance matrix estimation methods:

- Sample: the sample covariance matrix.
- Linear: the linear shrinkage estimator of **?** given in 3.

- QIS: the Quadratic-Inverse Shrinkage estimator of **?**, a recently developed nonlinear shrinkage method. QIS performs linear shrinkage on the sample eigenvalues of the covariance matrix in inverse eigenvalue space. A bandwidth parameter is required, which we choose following the paper's recommendation.
- NERCOME: the Nonparametric Eigenvalue-Regularized COvariance Matrix Estimator of **?**. This nonlinear shrinkage method randomly splits the samples into two groups, one for estimating eigenvectors and the other for estimating eigenvalues. Combining the estimates gives a matrix. Following the article, we repeated this procedure 50 times and took the final covariance matrix estimator to be the average of the individual matrices.
- Adap: the adaptive thresholding method of (**?**) for sparse covariance matrices, which applies soft thresholding to entries of the sample covariance matrix. The threshold method is adaptive to the entry's variance and involves a tuning parameter. We fixed the parameter at 2, as recommended.

In addition to the above estimators, we also implemented the two following oracle estimators, which cannot be implemented in practice as they require the unknown $\mathbf{\Sigma}$.

- OracNonlin: the optimal rotation-invariant covariance estimator, defined in **?**, with $\mathbf{\Sigma} = \boldsymbol{U}^T \text{diag}(\boldsymbol{l})\boldsymbol{U}$, where $\boldsymbol{U} = (\boldsymbol{u}_1 \dots \boldsymbol{u}_p)$ is the sample eigenvector matrix and $\boldsymbol{l} = (d_1, \dots, d_p)$ is composed of oracle eigenvalues $d_i = \boldsymbol{u}_i^T \mathbf{\Sigma} \boldsymbol{u}_i$. The sample covariance, the linear shrinkage estimator of **?**, and the nonlinear shrinkage estimators QIS and NERCOME are all rotation-invariant.
- OracMSG: the optimal covariance estimator in the class of separable estimators $\mathcal{S}$ 5. It equals our proposed estimator 10 except with the true G_p **??** instead of $\widehat{G}_p$ **??**. The adaptive thresholding method of **?** also targets a separable estimator.

[Figure 2 about here.]

Figure 2 presents the Frobenius loss for different estimators. Our MSG methods had the

lowest or near-lowest errors across all settings except for Model 4 which has high correlations 0.9. This is not surprising because our method assumed independence of $\boldsymbol{A}_{jk}$. In some cases, for example in Models 1 and 2, the improvement was substantial. Model 2 was especially interesting because the standard deviation and correlations were related. Our proposed empirical Bayes estimator was able to capture this dependence in its estimate of the prior G_p **??** and leverage it to provide much more accurate estimates. The nonlinear shrinkage estimators very slightly outperformed MSG in Model 5. In every setting, correcting MSG for positive-definiteness never increased the risk and decreased the risk in some cases. We also did experiments for Spectral norm and Matrix ℓ_1. The results for this two norms are very similar to Frobenius norm. One exception is Adap has the lowest error in terms of Matrix ℓ_1 norm in Model 1 because of its sparsity. Though our estimator was motivated in terms of the Frobenius norm error, it performed extremely well in terms of the other two norms as well.

Finally, the simulations show that the class of separable estimators 5 proposed in this paper is fundamentally different from the class of rotation-invariant estimators, as the oracle optimal estimators in these two classes behave very differently. For example, the oracle separable estimator had vanishing risk in Model 2, while the oracle rotiation-invariant estimator does not. Separable estimators seemed better for Models 1 and 2 while rotation-invariant estimators were superior in Models 3 and 4. They seem comparable in Model 5.

3.4 *Data analysis*

Covariance matrix estimation is often used to reconstruct gene networks (**?**). We applied our MSG and the other covariance matrix estimators described in Section 3.3 to gene network estimation using data from a small round blue-cell tumor microarray experiment (**?**), which was also studied by **?**. **?** report the expression of 2308 genes from 63 samples from four groups: 12 neuroblastoma, 20 rhabdomyosarcoma, 8 Burkitt lymphoma, and 23 Ewing's

sarcoma patients. In MSG, the clustering parameter K is still set as $p = 200$. We followed the same data preprocessing as **?** and sorted the genes in decreasing order according to their F-statistic

$$F = \frac{1}{k-1}\sum_{m=1}^{k} n_m(\overline{x}_m - \overline{x})^2 / \frac{1}{n-k}\sum_{m=1}^{k}(n_m - 1)\widehat{\sigma}_m^2 \tag{13}$$

where $k = 4$ is the number of patient categories, n_m, $\overline{x}_m$, and $\widehat{\sigma}_k$ represent the sample size, sample mean, and sample variance of the gene's expression in the mth category, respectively, and $\overline{x}$ is the global mean. We proceeded with the top 40 genes and bottom 160 genes.

We applied various methods to estimate the covariance matrix of these 200 genes. To measure the accuracy of the estimators, we split the 63 samples into two subsets $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, ensuring that each subset consisted of the same number of subjects from each of the four disease groups. After centering the variables to have zero mean, we used $\boldsymbol{X}_1$ to calculate covariance matrix estimates and compared these to the sample covariance matrix $\boldsymbol{S}_2$ of $\boldsymbol{X}_2$, which served as a proxy for the unknown true covariance matrix. We measured the errors using the Frobenius, spectral, and matrix ℓ_1 norms. We repeated this process 200 times.

Table 2 reports the average errors across the replications. Our MSG methods had the lowest average error. The positive-definiteness correction slightly reduced the risk as well. The linear shrinkage estimate was almost as accurate, but the other methods were much less accurate. These results suggest that our estimator can perform well in realistic settings, where the mean-zero multivariate normal distributional assumption on the data may not be met.

[Table 2 about here.]

In addition to comparing the numerical accuracies, we also investigated whether our estimator gave qualitatively different gene networks compared to the other approaches. First, Figure 3 illustrates the covariance matrices in network form, where each node represents a gene and each edge represents a non-zero covariance between the genes it connects. To

avoid completely connected graphs, we sparsified the matrix estimates by thresholding the smaller entries of each matrix to zero. Since the adaptive thresholding method of **?** naturally produced a sparse estimated matrix, we thresholded the other matrix estimates to match the sparsity level of the **?** estimate.

[Figure 3 about here.]

The results show several interesting features. First, there appear to be two major clusters, which are disconnected in every estimated network except for the one produced by the adaptive thresholding approach. Second, the larger cluster appears to contain two sub-clusters, and this finer structure was only recovered by MSG and QIS, and to a lesser extent the linear shrinkage estimator and NERCOME. Finally, the nodes in the networks estimated by QIS and NERCOME appear to be clustered more tightly together compared to in the other networks. These observations suggest that MSG produces qualitatively different networks, in addition to lower estimation errors.

Finally, we also compared the estimated degrees of the genes in the different networks. For each estimated network, we ordered the 200 genes by degree and then selected the top 20%, denoting this set J_k for the kth network. For each pair of networks k and k', we calculated the similarity between their most connected genes using Jaccard index $|J_k \cap J_{k'}|/|J_k \cup J_{k'}|$. Figure 4 visualizes these similarities. Interestingly, however, among all estimators, they were also the most similar to the unbiased sample covariance matrix. Together with the above results, this indicates that MSG may simultaneously give the lowest error and, at least in terms of degree estimation, the most unbiased results.

[Figure 4 about here.]

4. Discussion

The class of separable covariance matrix estimators 5 that we proposed in this paper appears to be very promising. Many existing procedures already explicitly or implicitly target this class, and our proposed estimate 10 of the optimal separable estimator outperforms a number of existing covariance matrix estimators. This is surprising because our approach vectorizes the matrix and therefore cannot take matrix structure, such as positive-definiteness, into account. This suggests that a vectorized approach combined with a positive-definiteness constraint may have improved performance. The resulting estimator would necessarily not be separable, because the estimate of the jkth entry would depend on more than just the jth and kth observed features, so the g-modeling estimation strategy is insufficient. More work is needed.

Though our estimator performs well in simulations and in real data, providing theoretical guarantees is difficult. In the standard mean vector estimation problem with $Y_i \sim N(\theta_i, 1)$, **?** showed that an empirical Bayes estimator based on a nonparametric maximum likelihood estimate of the prior on the θ_i can indeed asymptotically achieve the same risk as the oracle optimal separable estimator. However, this was in a simple model with a univariate prior distribution. **?** extended these results to multivariate $\boldsymbol{Y}_i \sim N(\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_i)$ with a multivariate prior on the $\boldsymbol{\theta}_i$, but assumed that the $\boldsymbol{Y}_i$ were independent. In contrast, our covariance matrix estimator is built from arbitrarily dependent $\boldsymbol{A}_{jk}$. These imposes significant theoretical difficulties that will require substantial work to address; we leave this for future research.

Finally, we have so far assumed that our data multivariate normal. To extend our procedure to non-normal data belonging to a parametric family, we can simply modify the density function $f(\cdot \mid \boldsymbol{\eta})$ in the nonparametric maximum compositive likelihood problem **??** and in our proposed estimator 10. If f is unknown or difficult to specify, alternative procedures may be necessary to approximate the optimal separable rule.

Acknowledgments

We thank Dr. Roger Koenker for his valuable comments.

Supporting Information

Web Appendices, Tables, and Figures referenced in Section 3 are available with this paper at the Biometrics website on Wiley Online Library.

Appendix

Proof of Proposition 1

Proof. We first rewrite the risk estimate $\widehat{R}(\beta_S, \beta_I)$. Define $\boldsymbol{M} = (\sum_{j,k=1}^{p} \widehat{\Delta}_{jk}^2, 0)^\top$, $\boldsymbol{\beta} = (\beta_S, \beta_I)^\top$, and the vectorized covariance matrices $\boldsymbol{v_S} = (s_{11}, \ldots, s_{pp})^\top$, $\boldsymbol{v_I} = (u_{11}, \ldots, u_{pp})^\top$, and $\boldsymbol{v_\Sigma} = (\sigma_{11}, \ldots, \sigma_{pp})^\top$. Then the unbiased risk estimator can be re-written as

$$\widehat{R}(\beta_S, \beta_I) = \boldsymbol{\beta}^\top (\boldsymbol{Z}^\top \boldsymbol{Z}) \boldsymbol{\beta} - 2(\boldsymbol{Z}\boldsymbol{v_S} - \boldsymbol{M})^\top \boldsymbol{\beta} - \mathbf{1}^\top \boldsymbol{M},$$

where $\boldsymbol{Z} = (\boldsymbol{v_S}, \boldsymbol{v_I})$. Therefore

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} (\boldsymbol{Z}\boldsymbol{v_S} - \boldsymbol{M}),$$

$$\widehat{\boldsymbol{v}}_{\boldsymbol{\Sigma}} = \boldsymbol{Z}\widehat{\boldsymbol{\beta}} = \boldsymbol{v_S} - \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{M}.$$

We will need to show $\widehat{\mu} b_n^2 / d_n^2 = \widehat{\beta}_I$ and $\widehat{\beta}_I / \widehat{\mu} + \widehat{\beta}_S = 1$. Since

$$\boldsymbol{Z}^\top \boldsymbol{Z} = \begin{pmatrix} s_{11} & \ldots & s_{pp} \\ u_{11} & \ldots & u_{pp} \end{pmatrix} \begin{pmatrix} s_{11} & u_{11} \\ \ldots & \ldots \\ s_{pp} & u_{pp} \end{pmatrix} = \begin{pmatrix} \sum_{j,k=1}^{p} s_{jk}^2 & \sum_{j=1}^{p} s_{jj} \\ \sum_{j=1}^{p} s_{jj} & p \end{pmatrix}$$

and $\det(\boldsymbol{Z}^\top \boldsymbol{Z}) = p \sum_{j,k=1}^{p} s_{jk}^2 - (\sum_{j=1}^{p} s_{jj})^2 = p^3 d_n^2$, it follows that

$$(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} = \frac{1}{p^3 d_n^2} \begin{pmatrix} p & -\sum_{j=1}^{p} s_{jj} \\ -\sum_{j=1}^{p} s_{jj} & \sum_{j,k=1}^{p} s_{jk}^2 \end{pmatrix},$$

and in addition

$$\boldsymbol{Z}^\top \boldsymbol{S} = \begin{pmatrix} \sum_{j,k=1}^p s_{jk}^2 \\ \sum_{j=1}^p s_{jj} \end{pmatrix}, \quad \boldsymbol{Z}^\top \boldsymbol{S} - \boldsymbol{M} = \begin{pmatrix} \sum_{j,k=1}^p s_{jk}^2 - \widehat{\Delta}_{jk}^2 \\ \sum_{j=1}^p s_{jj} \end{pmatrix}.$$

Therefore

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{S} \\ &= \frac{1}{p^3 d_n^2} \begin{pmatrix} p & -\sum_{j=1}^p s_{jj} \\ -\sum_{j=1}^p s_{jj} & \sum_{j,k=1}^p s_{jk}^2 \end{pmatrix} \begin{pmatrix} \sum_{j,k=1}^p s_{jk}^2 - \widehat{\Delta}_{jk}^2 \\ \sum_{j=1}^p s_{jj} \end{pmatrix} \\ &= \frac{1}{p^3 d_n^2} \begin{pmatrix} p \sum_{j,k=1}^p s_{jk}^2 - p \sum_{j,k=1}^p \widehat{\Delta}_{jk}^2 - (\sum_{j=1}^p s_{jj})^2 \\ (\sum_{j=1}^p s_{jj})(\sum_{j,k=1}^p \widehat{\Delta}_{jk}^2) \end{pmatrix}. \end{aligned}$$

The second component of $\widehat{\boldsymbol{\beta}}$ equals $\widehat{\beta}_I$, so

$$\begin{aligned} \widehat{\beta}_I &= \frac{1}{p^3 d_n^2} (\sum_{j=1}^p s_{jj})(\sum_{j,k=1}^p \widehat{\Delta}_{jk}^2) \\ &= \{(\sum_{j=1}^p s_{jj})/p\}\{(\sum_{j,k=1}^p \widehat{\Delta}_{jk}^2)/p^2\}/d_n^2 = \widehat{\mu} \frac{b_n^2}{d_n^2}. \end{aligned}$$

Furthermore,

$$\begin{aligned} \widehat{\beta}_I/\widehat{\mu} + \widehat{\beta}_S &= \frac{1}{p^3 d_n^2} p \sum_{j,k=1}^p \{s_{jk}^2 - p \sum_{j,k=1}^p \widehat{\Delta}_{jk}^2 - (\sum_{j=1}^p s_{jj})^2 + p \sum_{j,k=1}^p \widehat{\Delta}_{jk}^2\} \\ &= \frac{1}{p^3 d_n^2} \{p \sum_{j,k=1}^p s_{jk}^2 - (\sum_{j=1}^p s_{jj})^2\} = 1. \end{aligned}$$

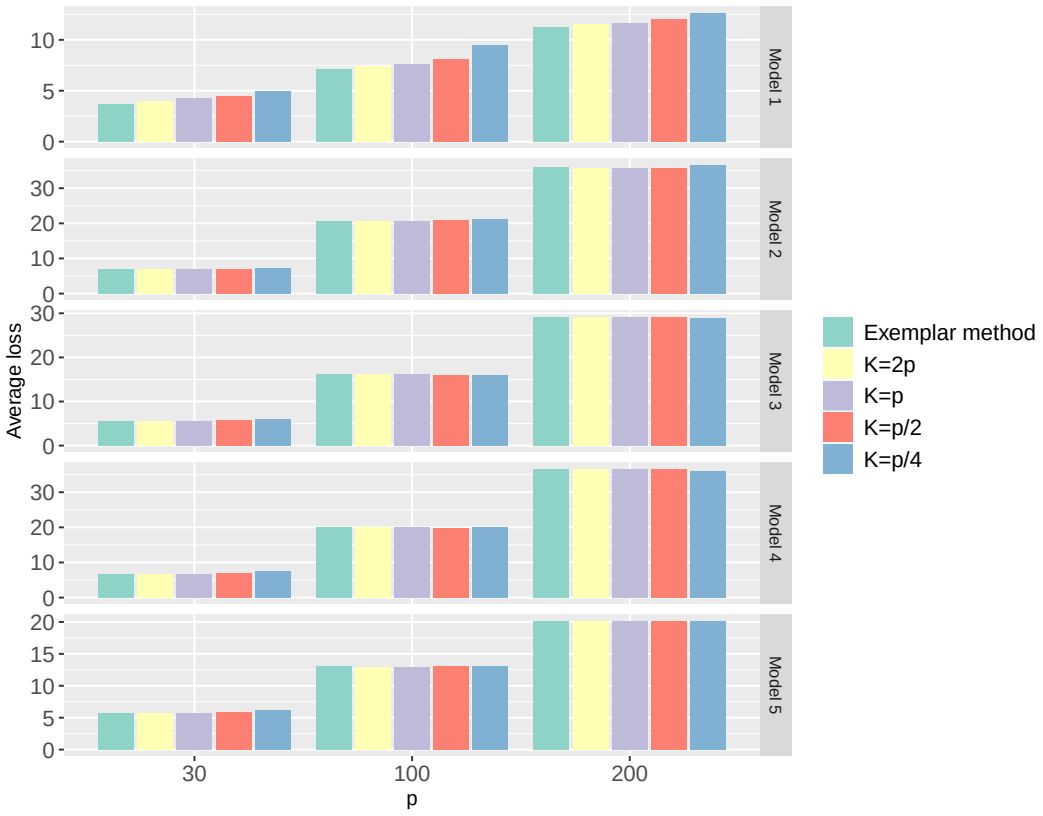

Figure 1. Average Frobenius norm errors over 200 replications. The Sparse, Block, Dense, Dense2, and Orth panels correspond to Models 1 through 5, respectively.

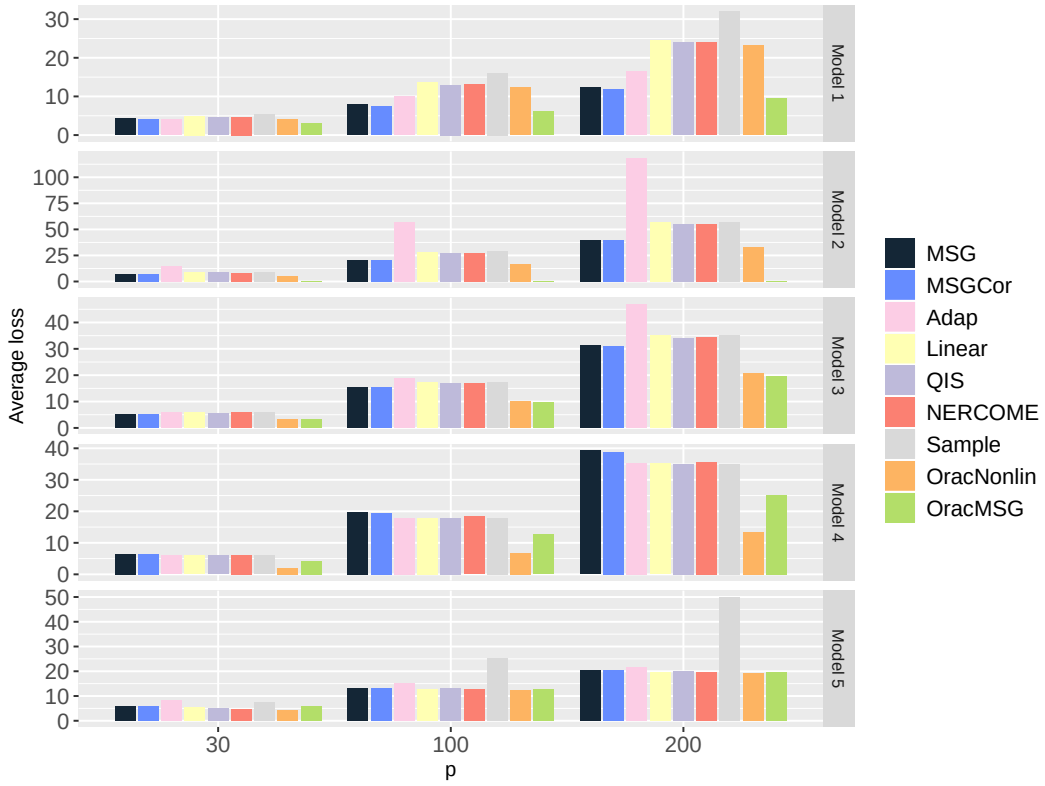

Figure 2. Average Frobenius norm errors over 200 replications. The Sparse, Block, Dense, Dense2, and Orth panels correspond to Models 1 through 5, respectively. In MSG, K-means clustering is applied with $K = p$

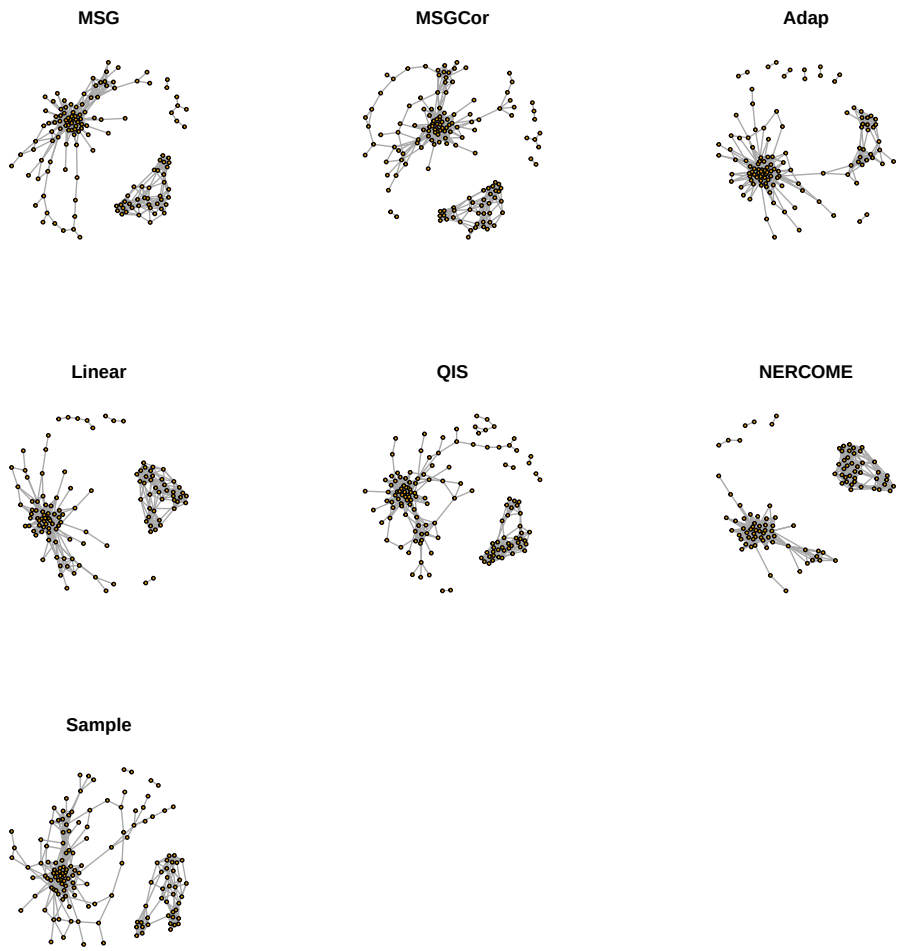

Figure 3. Gene networks recovered by the different covariance matrix estimation methods.

Figure 4. Similarities of gene degrees between the estimated networks. Each number reports the Jaccard index between the top 20% most connected genes of each pair of networks.

Table 1
Average running time for different ratios.

	p=30	p=100	p=200
Exemplar method	0.0748	4.2575	47.8616
$K = 2p$	0.0445	1.3166	12.2612
$K = p$	0.0291	0.8744	8.0862
$K = p/2$	0.0207	0.5861	5.3757
$K = p/4$	0.0153	0.4132	3.7415

Table 2
Average gene expression covariance matrix estimation errors. Bold entries highlight the smallest errors in each column.

	Frobenius	Spectral	Matrix ℓ_1
MSG	24.07	12.89	47.12
MSGCor	**23.90**	**12.87**	**46.63**
Adap	28.81	17.20	50.71
Linear	24.59	13.57	51.43
QIS	28.66	15.94	59.30
NERCOME	29.76	20.13	57.30
Sample	28.39	16.10	57.90