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: Covariance matrix estimation is a fundamental statistical task in many applications, but the sample covariance matrix is sub-optimal when the sample size is comparable to or less than the number of features. Such high-dimensional settings are common in modern genomics, where covariance matrix estimation is frequently employed as a method for inferring gene networks. To achieve estimation accuracy in these settings, existing methods typically either assume that the population covariance matrix has some particular structure, for example sparsity, or apply shrinkage to better estimate the population eigenvalues. In this paper, we propose a new approach to estimating high-dimensional covariance matrices that can have better performance than existing methods. We first frame covariance matrix estimation as a compound decision problem. This motivates defining a class of decision rules and using a nonparametric empirical Bayes g-modeling approach to estimate the optimal rule in the class. Simulation results and gene network inference in a single cell RNA-seq dataset show that our approach is comparable to or can outperform a number of state-of-the-art proposals in a wide range of settings.

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

1. Introduction

Covariance matrix estimation is a fundamental statistical problem that plays an essential role in various applications. However, in modern problems the number of features can be of the same order as or exceed the sample size. This high-dimensional setting is especially common in genomics, where covariance matrices are used to model gene networks but the number of genes can be much larger than the number of biological replicates (Schäfer and Strimmer, 2005; Markowetz and Spang, 2007). Accurate gene network estimation has become even more important with the advent of single-cell RNA-sequencing (scRNA-seq) technology, where interest now centers on inferring cell- or cell type-specific gene networks (Aibar et al., 2017; Dai et al., 2019; Stegle et al., 2015). For example, in Section 4.4 we study cell type-specific gene network inference using scRNA-seq data from honey bee brains. Need to do analysis

A common approach to estimating gene expression networks is to use the standard sample covariance or corrlation matrices (Langfelder and Horvath, 2008; Zhang and Horvath, 2005). However, these estimators behave poorly in the high-dimensional regime. To overcome these issues, various methods have been developed to estimate high-dimensional covariance matrices. These can roughly be divided into two groups, according to whether they impose assumptions of the structure of population covariance matrix.

Structured methods make assumptions about the form of the population covariance matrix. One popular class of methods assumes that the matrix is sparse, or has many zero entries. The most common strategy in this class is to threshold the entries of the sample covariance (Rothman et al., 2009; Cai and Liu, 2011), but penalized likelihood methods (Xue et al., 2012) have also been used. A second class of methods assume the data arise from a factor model (Fan et al., 2008), so that the covariance matrix has low intrinsic dimension. Other

common structured methods assume that the covariance matrix is banded (Li et al., 2017) or Toeplitz (Liu et al., 2017).

In contrast, unstructured methods do not make any assumptions about the population covariance matrix, yet still have lower estimation error than the sample covariance matrix. A first example was the linear shrinkage approach of Ledoit and Wolf (2004), which shrinks the sample covariance matrix toward a scaled identity matrix. More recently, nonlinear shrinkage methods were developed (Ledoit et al., 2012; Ledoit and Wolf, 2019; Lam et al., 2016). These shrink the eigenvalues of the sample covariance matrix in a data-adaptive fashion. Linear shrinkage can be viewed as a special case of nonlinear shrinkage, as it shrinks sample eigenvalues toward their global means.

Nonlinear shrinkage estimators have desirable optimality properties (Ledoit and Wolf, 2018) and show excellent performance. However, most existing nonlinear shrinkage estimators belong to the class of rotation-equivariant estimators (Bun et al., 2016; Stein, 1975, 1986), which modify the sample eigenvalue estimates but still use the sample eigenvectors. On the other hand, sample eigenvectors are not consistent when the dimension and the sample size increase at the same rate (Mestre, 2008). This suggests that certain classes of non-rotationally invariant estimators may outperform existing unstructured covariance matrix estimators.

Here we propose a new class of unstructured non-rotationally invariant estimators for high-dimensional covariance matrices. Our approach is based on interpreting the covariance matrix estimation problem as a compound decision problem (Robbins, 1951); see Section 2. This framing motivates us to vectorize the covariance matrix and solve the resulting vector estimation problem using a nonparametric empirical Bayes procedure, which has been shown in the compound decision literature to have excellent properties (Jiang and Zhang, 2009; Koenker and Mizera, 2014; Efron, 2019). 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. Our compound decision approach is conceptually similar to the empirical Bayes approach taken by Dey and Stephens (2018) to estimate a high-dimensional correlation matrix. We discuss in detail the connections with their work in Section 3.1. Need to add to numerical results

The article is organized as follows. In Section (3), we briefly review compound decision theory and then introduce our proposed approach. In Section (4) we illustrate the performance of our method in simulations and a gene expression dataset. Finally, Section (5) concludes with a discussion. Our procedure is implemented in the R package `cole`, available on GitHub. Is this implemented in the package yet?

2. Compound decision problem formulation

2.1 *Background*

Given n observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ independently generated from a p-dimensional $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, our goal is to find an estimator $\boldsymbol{\delta}(\boldsymbol{X})$ of $\boldsymbol{\Sigma}$. A common measure of estimation performance is 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. This paper constructs a new non-rotationally invariant $\boldsymbol{\delta}$ that in simulations and a data analysis has comparable or lower risk (1) compared to existing methods.

Our proposed approach is motivated by the observation that minimizing (1) is a type of compound decision problem, which we briefly review in this section. Introduced by Robbins (1951), compound decision 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 the squared error loss (Johnstone, 2017).

A key property of compound decision problems is that while a given Y_i or θ_i seems like it should offer no help for estimating θ_j when $j \neq i$, in fact borrowing information across all indices $i = 1, \ldots, n$ to estimate θ_j is superior to estimating each θ_j using the corresponding Y_j alone. A classical example of this phenomenon is the James-Stein estimator (James and Stein, 1961), 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 $\boldsymbol{Y}$. James and Stein (1961) showed 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 $\boldsymbol{\theta}$ (Brown and Greenshtein, 2009; Jiang and Zhang, 2009; Johnstone, 2017; Lindley, 1962; Fourdrinier et al., 2018).

The discussion above shows that covariance matrix estimation under the Frobenius risk (1) can be viewed as a compound decision problem. Furthermore, one special consequence of choosing (1) as the risk measure is that the matrix estimation problem becomes equivalent to a vector estimation problem, specifically the problem of estimating every component of the vector $(\sigma_{11}, \ldots, \sigma_{pp})^\top$ under average squared error loss. This interpretation is conceptually very similar to the classical Gaussian sequence problem.

We therefore propose to apply modern ideas for Gaussian sequence estimation to covariance matrix estimation. One important difference between estimating a covariance matrix versus a vector is that the former should have additional structure, and in particular should be at least positive semidefinite. Notably, however, this structure is not incentived by the Frobenius

risk (1). As a result, there can exist estimators of $\boldsymbol{\Sigma}$ that achieve low values of (1) but which are not positive semidefinite. This will in fact be true of one of the estimators we propose in Section 3.1. To resolve this issue, we project the result estimate into the space of positive-definite matrices; see Section 3.3.

2.2 *Connections to existing work*

Treating the covariance matrix estimation problem as a vector estimation compound decision problem may seem unintuitive, but many existing matrix estimation methods can also be interpreted as carrying out vector estimation. The simplest example is the sample covariance matrix $\boldsymbol{S} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$, which takes this form because the $\boldsymbol{X}$ are assumed to have mean zero. This can be thought of as estimating each component of $(\sigma_{11}, \ldots, \sigma_{pp})^\top$ using maximum likelihood. Less trivially, Cai and Liu (2011) studied sparse high-dimensional covariance matrices and explicitly appealed to the vector perspective. Their adaptive thresholding method applies a version of the soft thresholding method of Donoho and Johnstone (1995), originally developed to estimate a sparse mean vector in the Gaussian sequence problem, to each entry of the sample covariance matrix.

Interestingly, we can also show that the celebrated linear shrinkage covariance matrix estimator of Ledoit and Wolf (2004) can be essentially recovered as a solution to a vector estimation problem where the estimand is the $p^2 \times 1$ vector $(\sigma_{11}, \ldots, \sigma_{pp})^\top$. Their estimator is defined as

$$\hat{\boldsymbol{\Sigma}}_{\text{LW}} = \left(1 - \frac{b_n^2}{d_n^2}\right)\boldsymbol{S} + \frac{b_n^2}{d_n^2}\hat{\mu}\boldsymbol{I} \tag{3}$$

for $\hat{\mu} = \text{tr}(\boldsymbol{S})/p$, $d_n^2 = \|\boldsymbol{S} - \hat{\mu}\boldsymbol{I}\|_F^2$, and $b_n^2 = \min(d_n^2, \sum_{i=1}^n \|\boldsymbol{X}_i\boldsymbol{X}_i^\top - \boldsymbol{S}\|_F^2/n^2)$, where $\|\cdot\|_F$ denotes the Frobenius norm.

To see this, we first restrict attention to the class of linear decision rules for estimating each component of $(\sigma_{11}, \ldots, \sigma_{pp})^\top$, which we define to take the form

$$\delta_{jk}(\boldsymbol{X}) = \beta_S s_{jk} + \beta_I u_{jk} \tag{4}$$

for some parameters β_S and β_I, where s_{jk} is the jkth entry of $\boldsymbol{S}$ and u_{jk} is the jkth entry of $\boldsymbol{I}$. In other words, the estimate for σ_{jk} is linear in s_{jk}. Ideally we would like to estimate the optimal β_S and β_I by choosing the values that minimize the Frobenius risk (1), and the following proposition shows that it is possible to construct a good estimate of this risk for every fixed value of β_S and β_I.

PROPOSITION 1: Define $\hat{\Delta}^2_{jk} = \sum_{i=1}^n (X_{ij}X_{ik} - s_{jk})^2/n^2$ and

$$\hat{R}(\beta_S, \beta_I) = \frac{1}{p^2} \sum_{j,k=1}^{p} [(2\beta_S - 1)\hat{\Delta}^2_{jk} + \{(1-\beta_S)s_{jk} - \beta_I u_{jk}\}^2]. \tag{5}$$

If $p^{-2}\mathbb{E}\|\boldsymbol{S} - \boldsymbol{\Sigma}\|_F^2$ is bounded as $n \to \infty$, then

$$\lim_{n\to 0}\{\mathbb{E}\hat{R}(\beta_S, \beta_I) - R(\boldsymbol{\Sigma}, \boldsymbol{\delta})\} = 0$$

for the class of decision rules $\boldsymbol{\delta}$ defined in (4).

The condition on $\mathbb{E}\|\boldsymbol{S} - \boldsymbol{\Sigma}\|_F^2$ requires that the variances of the entries of $\boldsymbol{S}$ do not grow too quickly as n grows. It is also required by Ledoit and Wolf (2004) and is implied by their Lemma 3.1.

Proposition 1 shows that $\hat{R}(\beta_S, \beta_I)$ (5) is an asymptotically unbiased estimate of the true risk (1) for linear decision rules of the form (4). It is then reasonable to estimate the best estimator in this class by minimizing $\hat{R}(\beta_S, \beta_I)$ over β_S and β_I. Proposition 2 shows that this estimator is closely related to the Ledoit and Wolf (2004) estimator (3).

PROPOSITION 2: Let $\hat{\beta}_S$ and $\hat{\beta}_I$ denote the minimizers of $\hat{R}(\beta_S, \beta_I)$ and define the linear decision rule

$$\hat{\delta}_{jk}(\boldsymbol{X}) = \max(\hat{\beta}_S, 0)s_{jk} + \min\{\hat{\beta}_I, \hat{\mu}\}u_{jk}. \tag{6}$$

Then $\hat{\delta}_{jk}(\boldsymbol{X}) = \hat{\sigma}_{\mathrm{LW}jk}$, where $\hat{\sigma}_{\mathrm{LW}jk}$ is the jkth entry of $\boldsymbol{\Sigma}_{\mathrm{LW}}$ (3).

Proposition 2 therefore shows that the Ledoit and Wolf (2004) covariance matrix estimator can essentially be obtained as the solution to a pure vector estimation problem, without

considering the matrix structure of the target vector $(\sigma_{11}, \ldots, \sigma_{pp})^\top$. There are interesting connections between the estimator in Proposition 2 and classical estimators of a vector of normal means. First, instead of (6), consider the estimator $\tilde{\delta}_{jk}(\boldsymbol{X}) = \hat{\beta}_S s_{jk} + \hat{\beta}_I u_{jk}$, which directly plugs the estimated $\hat{\beta}_S$ and $\hat{\beta}_I$ into (4). It can be shown that $\tilde{\delta}_{jk}(\boldsymbol{X})$ is analogous to the Efron-Morris estimator for the Gaussian sequence problem (Efron and Morris, 1973). In other words, it can be shown that $\tilde{\delta}(\boldsymbol{X})$ 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 $(u_{11}, \ldots, u_{pp})^\top$ (Biscarri, 2019; Lindley, 1962; Stigler, 1990). A major difference between $\hat{\delta}_{jk}$ (6) and $\tilde{\delta}_{jk}$ is that $\hat{\beta}_S$ in the latter is replaced by $\max(\hat{\beta}_S, 0)$ in the former. This is equivalent to how the shrinkage factor in the classical James-Stein estimator is replaced by its positive part in the positive-part James-Stein estimator (Baranchik, 1964).

3. Method

3.1 *Proposed estimator*

Section 2 argues that treating covariance matrix estimation as a vector estimation problem can be a fruitful strategy, as evidenced by the linear estimator exhibited in Proposition 2. This suggests that estimators that are nonlinear in s_{jk} may have better performance.

We propose to consider an even larger class of estimators, 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 $\delta_i(\boldsymbol{Y}) = t(Y_i)$ for some function t that does not depend on the index i (Robbins, 1951). Here we generalize this idea 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 $(\sigma_{11}, \ldots, \sigma_{pp})^\top$, we define the class of separable rules to be

$$\mathcal{S} = \{\boldsymbol{\delta} : \delta_{kj} = \delta_{jk} = t(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k}),\ 1 \leqslant k < j \leqslant p, \quad \delta_{jj} = \widetilde{t}(\boldsymbol{X}_{\cdot j}),\ j = 1, \ldots, p)\}, \tag{7}$$

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 diagonal entries of $\boldsymbol{\Sigma}$ using a function $\tilde{t}$ of observations of the corresponding feature and off-diagonal entries using a function t of observations from the two corresponding features, where t and $\tilde{t}$ do not depend on the indices j and k. Furthermore, we enforce rule in $\mathcal{S}$ to give symmetric estimates of the off-diagonal entries.

This class of separable rules is reasonable to consider because it includes several common covariance estimators, including the sample covariance $(s_{11}, \ldots, s_{pp})^\top$, the class of adaptive thresholding estimators for sparse covariance matrices studied by Cai and Liu (2011), and the class of linear estimators (4) used by Ledoit and Wolf (2004), which can be expressed as

$$t(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k}) = \beta_S \boldsymbol{X}_{\cdot j}^\top \boldsymbol{X}_{\cdot k}/n,\ 1 \leqslant k < j \leqslant p,$$
$$\tilde{t}(\boldsymbol{X}_{\cdot j}) = \beta_S \boldsymbol{X}_{\cdot j}^\top \boldsymbol{X}_{\cdot j}/n + \beta_I,\ j = 1, \ldots, p.$$

Furthermore, whereas these three existing estimators are ultimately all functions of the s_{jk}, the class $\mathcal{S}$ allows for much more general rules that can be any function of the observations $\boldsymbol{X}_{\cdot j}$ and $\boldsymbol{X}_{\cdot k}$.

We propose to search for the optimal estimator within $\mathcal{S}$. The optimal separable estimator $\boldsymbol{\delta}^\star$ that minimizes the scaled squared Frobenius risk (1) over all rules in $\mathcal{S}$ will perform at least as well as the three estimators mentioned above, and may perform better. Targeting the optimal separable rule is standard in the compound decision literature (Zhang, 2003).

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 (Robbins, 1955; Zhang, 2003; Brown and Greenshtein, 2009; Jiang and Zhang, 2009; Efron, 2014, 2019). We take a similar approach here. $f(\cdot \mid \boldsymbol{\eta}_{jk})$, the density of $(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k})$, 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 $j \neq k$, $(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k})$ is comprised of n independent mean-zero

multivariate normals with covariance matrices

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

Now consider the Bayesian model for non-diagonal entries

$$(\boldsymbol{X}, \boldsymbol{X}') \mid \boldsymbol{\eta} \sim f(\cdot \mid \boldsymbol{\eta}), \quad \boldsymbol{\eta} \sim G_{nd}(a, b, \gamma) = \frac{2}{p(p-1)} \sum_{1 \leqslant k < j \leqslant p} I(\sigma_j \leqslant a, \sigma_k \leqslant b, r_{jk} \leqslant \gamma). \quad (8)$$

and diagonal entries

$$\boldsymbol{X} \mid a \sim \widetilde{f}(\cdot \mid a), \quad a \sim G_d(a) = \frac{1}{p} \sum_{j=1}^{p} I(\sigma_j \leqslant a). \quad (9)$$

the G_d is the marginal density for the first parameter of G_{nd}

$$\int dG_{nd}(a, b, \gamma) db d\gamma = dG_d(a) \quad (10)$$

By the fundamental theorem of compound decisions (Robbins, 1951; Jiang and Zhang, 2009), this is closely related to the vectorized covariance matrix estimation problem under Frobenius risk (1). Then we have the following proposition.

PROPOSITION 3: For any $\boldsymbol{\delta} \in \mathcal{S}$ (7), the Frobenius risk can be written as the convex combination of Bayesian risks with respect to priors G_{nd} and G_d.

$$\begin{aligned} R(\boldsymbol{\Sigma}, \boldsymbol{\delta}) =& \frac{1}{p^2}\{2 \sum_{1 \leqslant k < j \leqslant p}^{p} \int \{t(\boldsymbol{X}, \boldsymbol{X}') - \sigma_{jk}\}^2 f(t(\boldsymbol{X}, \boldsymbol{X}') \mid \boldsymbol{\eta}_{jk}) d\boldsymbol{X} d\boldsymbol{X}' \\ &+ \sum_{j=1}^{p} \int \{\widetilde{t}(\boldsymbol{X}) - \sigma_{jj}\}^2 \widetilde{f}(\boldsymbol{X} \mid s_{jj}) d\boldsymbol{X}\} \\ =& \frac{p-1}{p} \int \int \{t(\boldsymbol{X}, \boldsymbol{X}') - g(\boldsymbol{\eta})\}^2 f((\boldsymbol{X}, \boldsymbol{X}') \mid \boldsymbol{\eta}) dG_{nd}(\boldsymbol{\eta}) d\boldsymbol{X} d\boldsymbol{X}' \\ &+ \frac{1}{p} \int \int \{\widetilde{t}(\boldsymbol{X}) - a^2\}^2 \widetilde{f}(\boldsymbol{X} \mid a) dG_d(a) d\boldsymbol{X} \\ =& \frac{p-1}{p} \mathbb{E}[\{t(\boldsymbol{X}, \boldsymbol{X}') - g(\boldsymbol{\eta})\}^2] + \frac{1}{p} \mathbb{E}[\{\widetilde{t}(\boldsymbol{X}) - a^2\}^2], \end{aligned}$$

where $g(a, b, \gamma) = ab\gamma$ and the final expectation is the Bayes risk of estimating σ_{jk}.

The decision rule which minimizes the Frobenius risk depends on all the σ_j and r_{jk}. The optimal oracle separable rule $\boldsymbol{\delta}^\star$ therefore has jkth entry equal to $\delta^\star_{jk}(\boldsymbol{X}) = t^\star((\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k}))$

for $j > k$ and $\delta^\star_{jj}(\boldsymbol{X}) = \widetilde{t}^\star(\boldsymbol{X}_{\cdot j})$, where $t^\star = \mathbb{E}\{g(\boldsymbol{\eta}) \mid (\boldsymbol{X}, \boldsymbol{X}')\}$ and $\widetilde{t}^\star = \mathbb{E}\{a^2 \mid \boldsymbol{X}\}$ minimize the Bayes risk.

Based on this result, we propose the following empirical Bayes procedure. We first use nonparametric maximum likelihood (Kiefer and Wolfowitz, 1956) to estimate the priors G_{nd} and G_d. Under the Bayesian model (8) and (9), and the working assumption that the $(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k})$ are independent across jk, we estimate G_{nd} using

$$\hat{G}_{nd} = \underset{G_{nd} \in \mathcal{G}_{nd}}{\arg\max} \{ \prod_{1 \leqslant k \leqslant j < p} \int f(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k} \mid \boldsymbol{\eta}) dG_{nd}(\boldsymbol{\eta}) \} \{ \prod_{j=1}^{p} \int \widetilde{f}(\boldsymbol{X}_{\cdot j} \mid a) dG_d(a) \}, \tag{11}$$

where $\mathcal{G}_{nd}$ is the family of all distributions supported on $\mathbb{R}_+ \times \mathbb{R}_+ \times [-1, 1]$, G_d is determined by G_{nd} as indicated in (10). Of course, the $(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k})$ are not independent, so $\hat{G}_{nd}$ does not maximize a likelihood but rather a pairwise composite likelihood (Varin et al., 2011). Using $\hat{G}_{nd}$ and $\hat{G}_d$, we estimate the vectorized $\boldsymbol{\Sigma}$ using

$$\hat{\boldsymbol{\delta}}(\boldsymbol{X}) = (\hat{t}(\boldsymbol{X}_{\cdot 2}, \boldsymbol{X}_{\cdot 1}), \ldots, \hat{t}(\boldsymbol{X}_{\cdot p}, \boldsymbol{X}_{\cdot (p-1)}), \hat{t}(\boldsymbol{X}_{\cdot 1}), \ldots, \hat{t}(\boldsymbol{X}_{\cdot p}))$$

, where

$$\hat{t}(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k}) = \frac{\int g(\boldsymbol{\eta}) f(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k} \mid \boldsymbol{\eta}) d\hat{G}_{nd}(\boldsymbol{\eta})}{\int f(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k} \mid \boldsymbol{\eta}) d\hat{G}_{nd}(\boldsymbol{\eta})}, \quad \hat{t}(\boldsymbol{X}_{\cdot j}) = \frac{\int a^2 \widetilde{f}(\boldsymbol{X}_{\cdot j} \mid a) d\hat{G}_d(a)}{\int \widetilde{f}(\boldsymbol{X}_{\cdot j} \mid a) d\hat{G}_d(a)}. \tag{12}$$

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

Our proposed procedure is an example of what Efron (2014) calls g-modeling, an approach to empirical Bayes problems that proceeds by modeling the prior. A major advantage of nonparametric estimation of the prior is that it allows the data itself to determine how best to shrink the estimator. This kind of procedure is also adopted in Dey and Stephens (2018). In contrast, most existing methods shrink in a pre-determined direction, such as toward a diagonal matrix in the case of Ledoit and Wolf (2004). Theoretical justification of our proposed $\hat{\boldsymbol{\delta}}$ is difficult and is discussed in Section (5). Nevertheless, our numerical results in Section (4) show that in practice, our $\hat{\boldsymbol{\delta}}$ can outperform many existing covariance matrix estimators.

3.2 *Implementation*

Calculating the estimated prior $\hat{G}_{nd}$ (11) is difficult, as it is an infinite-dimensional optimization problem over the class of all probability distributions $\mathcal{G}_{nd}$ supported on $\mathbb{R}_+ \times \mathbb{R}_+ \times [-1, 1]$. Lindsay (1983) showed that the solution is atomic and is supported on at most $p(p+1)/2$ points. The EM algorithm has traditionally been used to estimate the locations of the support points and the masses at those points (Laird, 1978), 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 (Jiang and Zhang, 2009; Koenker and Mizera, 2014; Feng and Dicker, 2018). 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 $\boldsymbol{\xi}_\tau$, $\tau = 1, \ldots, D$. The support grid points are constructed symmetrically, satisfying $\boldsymbol{\xi}_{\tau+D/2} = (b_\tau, a_\tau, \gamma_\tau)$ for $\tau = 1, \ldots, D/2$. Since the empirical distribution of diagonal entries G_d is symmetric about the first two parameters, so we would to like enforce the weights w_τ also have symmetric property. We can then use the EM algorithm to estimate the masses $\hat{\boldsymbol{w}} = \{\hat{w}_1, \ldots, \hat{w}_D\}$ at those points via the iteration

$$\hat{w}_\tau^{(k)} = \hat{w}_{\tau+D/2}^{(k)} = \frac{2}{p(p-1)}\Big[\sum_{1\leqslant k<j\leqslant p} \frac{\hat{w}_\tau^{(k-1)}\{f(\boldsymbol{X}_j, \boldsymbol{X}_k \mid \boldsymbol{\xi}_\tau) + f(\boldsymbol{X}_j, \boldsymbol{X}_k \mid \boldsymbol{\xi}_{\tau+D/2})\}}{\sum_{l=1}^{D/2} \hat{w}_l^{(k-1)}\{f(\boldsymbol{X}_j, \boldsymbol{X}_k \mid \boldsymbol{\xi}_l) + f(\boldsymbol{X}_j, \boldsymbol{X}_k \mid \boldsymbol{\xi}_{l+D/2})\}} + \sum_{j=1}^{p} \frac{\hat{w}_\tau^{(k-1)}\{\widetilde{f}(\boldsymbol{X}_j \mid a_\tau) + \widetilde{f}(\boldsymbol{X}_j \mid a_{\tau+D/2})\}}{\sum_{l=1}^{D/2} \hat{w}_l^{(k-1)}\{\widetilde{f}(\boldsymbol{X}_j \mid a_l) + \widetilde{f}(\boldsymbol{X}_j \mid a_{l+D/2})\}}\Big]$$

over k and the iteration guarantees symmetry. Early stopping of the EM algorithm can be useful (Koenker et al., 2019), and more sophisticated convex optimization procedures can be used as well (Koenker and Mizera, 2014). Our proposed estimator (12) then becomes

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

Ideally, the grid points should be chosen to densely cover the parameter space. However, the fact that G_{nd} 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 (Saha et al., 2020), which sets the support points to equal the observed sample versions $\hat{\boldsymbol{\eta}}_{jk}$ 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 $p(p-1)/2$ lower triangular off-diagonal sample points (s_j, s_k, γ_{jk}) and find their symmetric points by exchanging their first two dimensions. We then use the $2K$ cluster centroids as our support points. 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 $(\boldsymbol{X}_j, \boldsymbol{X}_k)$ since it is the sufficient statistic of $(\boldsymbol{X}_j, \boldsymbol{X}_k)$ with respect to $\boldsymbol{C}_{jk}$. Based on the property of sufficient statistics,

$$\frac{f(\boldsymbol{X}_j, \boldsymbol{X}_k \mid \boldsymbol{\xi}_\tau)}{f(\boldsymbol{X}_j, \boldsymbol{X}_k \mid \boldsymbol{\xi}_l)} = \frac{f(\boldsymbol{S}_{jk} \mid \boldsymbol{\xi}_\tau)}{f(\boldsymbol{S}_{jk} \mid \boldsymbol{\xi}_l)} \tag{13}$$

it is equivalent to use $f(\boldsymbol{X}_j, \boldsymbol{X}_k \mid \boldsymbol{\xi}_\tau)$ instead to simplify the calculation. Although we assume $\boldsymbol{X}$ has mean of zero theoretically, in practice, we usually do not know the true means. In such case, we can centralize each $\boldsymbol{X}_{\cdot j}$ by their sample means.

3.3 *Positive definiteness correction*

Our proposed estimator (12) 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. Higham (1988) and Huang et al. (2017) showed that the projection of a $p \times p$

symmetric matrix $\boldsymbol{B}$ onto the space of positive semi-definite matrices is

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

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 Huang et al. (2017) 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

$$P_0(\boldsymbol{B}) = \boldsymbol{Q}\text{diag}\{\max(\lambda_1,c),\max(\lambda_2,c),\ldots,\max(\lambda_p,c)\}\boldsymbol{Q}^\top. \tag{14}$$

Huang et al. (2017) 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 $\alpha_K = 10$.

4. Numerical Results

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. In our EM algorithm, the maximum number of iterations (**??**) is 200 and it early stops when $\frac{L(\boldsymbol{w}^{(k+1)})-L(\boldsymbol{w}^{(k)})}{L(\boldsymbol{w}^{(k)})} \leqslant 10^{-4}$, where $L(\boldsymbol{w}^{(k)}) = \sum_{1\leqslant k\leqslant j\leqslant p}\log f(\boldsymbol{X}_{\cdot j},\boldsymbol{X}_{\cdot k}|\boldsymbol{w}^{(k)})$ is composite log likelihood of the observations.

4.1 *Models*

We considered six models for the population covariance matrix to explore the behavior of our method in different settings. 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 (Sparse). The standard deviations were independently generated from $\mathcal{U}(1,1.5)$

and the correlation matrix followed Model 2 of Cai and Liu (2011):

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

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 (Block). 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 (Dense). 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 (Dense2). 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 (11).
- Model 5 (Orth). With $\boldsymbol{U}$ a randomly generated orthonormal matrix, $\boldsymbol{\Sigma} = \boldsymbol{U}^T \text{diag}(\boldsymbol{l})\boldsymbol{U}$, where $\boldsymbol{l}$ was a vector of eigenvalues independently generated from $\mathcal{U}(1, 4)$. This referred to simulation settings from Lam et al. (2016) and Ledoit and Wolf (2019).
- Model 6 (Spiked). With $\boldsymbol{U}$ a randomly generated orthonormal matrix, $\boldsymbol{\Sigma} = \boldsymbol{U}^T \text{diag}(\boldsymbol{l})\boldsymbol{U}$, where $\boldsymbol{l}$ was a vector of eigenvalues where the first 3 entries were 4,3,2 and the remaining $p-3$ entries equaled 1.

In each scenario, we generated $n = 100$ samples from a p-variate $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $p =$

30, 100, or 200. Although $\mathbb{E}\boldsymbol{X} = 0$ in our setting, we assume it is unknown and use $\boldsymbol{S} = \frac{1}{n-1}(\boldsymbol{X} - \bar{\boldsymbol{X}})^\top(\boldsymbol{X} - \bar{\boldsymbol{X}})$. We generated 200 replicates and reported median errors under the following Frobenius norms, where $\hat{\boldsymbol{\Sigma}}$ is the estimated matrix with entries $\hat{\sigma}_{jk}$ and $\boldsymbol{\Sigma}$ is the true matrix with entries σ_{jk}:

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F = \{\sum_{j,k=1}^{p}(\hat{\sigma}_{jk} - \sigma_{jk})^2\}^{1/2}$$

which is a version of (1).

4.2 *Clustering-based exemplar algorithm*

We first studied the behavior of our K-means clustering-based exemplar algorithm for different K, described in Section (3.2). 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 6. 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.

4.3 *Methods compared*

In this subsection, 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 Ledoit and Wolf (2004) given in (3).
- QIS: the Quadratic-Inverse Shrinkage estimator of Ledoit and Wolf (2019), 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 Lam et al. (2016). 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 (Cai and Liu, 2011) 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.
- CorShrink: Empirical Bayes shrinkage estimation of correlation matrix (Dey and Stephens, 2018).

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 Ledoit and Wolf (2019), 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 Ledoit and Wolf (2004), and the nonlinear shrinkage estimators QIS and NERCOME are all rotation-invariant.
- OracMSG: It equals our proposed estimator (12) except sample grid points are cluster centers of true parameters $(\sigma_j, \sigma_k, r_{jk})$.

[Figure 2 about here.]

Figure 2 presents the Frobenius loss for different estimators. From Model 1 to Model 4 when the prior distributions of G_{nd} are simple, our MSG methods had the lowest errors across all settings. For Model 4 which has high correlations 0.9, the priority is less obvious. This is not surprising because our method assumed independence of $(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k})$. In some cases, 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_{nd} (8) and leverage it to provide much more accurate estimates. The nonlinear shrinkage estimators outperformed MSG in Model 5 and Model 6. From the plot, we can observe that CorShrink, of which the target is to estimate the correlation matrix, also had competitive performance in all settings. 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 (7) 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, 3 and 4, while the oracle rotiation-invariant estimator does not. Separable estimators seemed better for Models 1 to 4 while rotation-invariant estimators were superior in Models 6. They seem comparable in Model 5.

4.4 *Data analysis*

Covariance matrix estimation is often used to reconstruct gene networks (Markowetz and Spang, 2007). We applied our MSG and the other covariance matrix estimators described in Section (4.3) to gene network estimation using data from a small round blue-cell tumor microarray experiment (Khan et al., 2001), which was also studied by Cai and Liu (2011). Osareh and Shadgar (2009) 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 Cai and Liu (2011) and sorted the genes in decreasing order according to their F-statistic

$$F = \frac{1}{k-1}\sum_{m=1}^{k} n_m(\overline{x}_m - \overline{x})^2 / \frac{1}{n-k}\sum_{m=1}^{k}(n_m - 1)\hat{\sigma}_m^2 \tag{15}$$

where $k = 4$ is the number of patient categories, n_m, $\overline{x}_m$, and $\hat{\sigma}_k$ 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 Cai and Liu (2011) naturally produced a sparse estimated matrix, we thresholded the other matrix estimates to match the sparsity level of the Cai and Liu (2011) 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.]

5. Discussion

The class of separable covariance matrix estimators (7) 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 (12) 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 $Y_i \sim N(\theta_i, 1)$, Jiang and Zhang (2009) 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. Saha et al. (2020) 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 $(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k})$.

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 are 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 (11) and in our proposed estimator (12). 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.

References

Aibar, S., González-Blas, C. B., Moerman, T., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z. K., Wouters, J., and Aerts, S. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nature Methods* **14,** 1083–1086.

Baranchik, A. J. (1964). Multiple regression and estimation of the mean of a multivariate normal distribution. Technical Report 51, Stanford University.

Biscarri, W. D. (2019). *Statistical methods for binomial and Gaussian sequences*. PhD thesis, University of Illinois at Urbana-Champaign.

Brown, L. D. and Greenshtein, E. (2009). Nonparametric empirical bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics* **37,** 1685–1704.

Bun, J., Allez, R., Bouchaud, J.-P., and Potters, M. (2016). Rotational invariant estimator for general noisy matrices. *IEEE Transactions on Information Theory* **62,** 7475–7490.

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

Dai, H., Li, L., Zeng, T., and Chen, L. (2019). Cell-specific network constructed by single-cell rna sequencing data. *Nucleic Acids Research* **47,** e62–e62.

Dey, K. K. and Stephens, M. (2018). Corshrink: Empirical bayes shrinkage estimation of correlations, with applications. *bioRxiv* page 368316.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90,** 1200–1224.

Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statistical Science* **29,** 285–301.

Efron, B. (2019). Bayes, Oracle Bayes and Empirical Bayes. *Statistical Science* **34,** 177–201.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association* **68,** 117–130.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* **147,** 186–197.

Feng, L. and Dicker, L. H. (2018). Approximate nonparametric maximum likelihood for mixture models: A convex optimization approach to fitting arbitrary multivariate mixing distributions. *Computational Statistics & Data Analysis* **122,** 80–91.

Fourdrinier, D., Strawderman, W. E., and Wells, M. T. (2018). *Shrinkage estimation*. Springer.

Higham, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear algebra and its applications* **103,** 103–118.

Huang, C., Farewell, D., and Pan, J. (2017). A calibration method for non-positive definite covariance matrix in multivariate data analysis. *Journal of Multivariate Analysis* **157,** 45–52.

James, W. and Stein, C. M. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 367–379. Berkeley and Los Angeles, University of California Press.

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical bayes estimation of normal means. *The Annals of Statistics* **37,** 1647–1684.

Johnstone, I. M. (2017). Gaussian estimation: Sequence and wavelet models. Technical report, Department of Statistics, Stanford University, Stanford.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine* **7,** 673–679.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics* pages 887–906.

Koenker, R., Gu, J., et al. (2019). Comment: Minimalist g-modeling. *Statistical Science* **34,** 209–213.

Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association* **109,** 674–685.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* **73,** 805–811.

Lam, C. et al. (2016). Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *The Annals of Statistics* **44,** 928–953.

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 1–13.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88,** 365–411.

Ledoit, O. and Wolf, M. (2018). Analytical nonlinear shrinkage of large-dimensional covariance matrices. Technical Report 264, Department of Economics, University of Zurich.

Ledoit, O. and Wolf, M. (2019). Quadratic shrinkage for large covariance matrices. Technical Report 335, Department of Economics, University of Zurich.

Ledoit, O., Wolf, M., et al. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics* **40,** 1024–1060.

Li, J., Zhou, J., Zhang, B., and Li, X. R. (2017). Estimation of high dimensional covariance matrices by shrinkage algorithms. In *2017 20th International Conference on Information Fusion (Fusion)*, pages 1–8. IEEE.

Lindley, D. V. (1962). Discussion on professor stein's paper. *Journal of the Royal Statistical Society: Series B (Methodological)* **24,** 265–296.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: a general theory. *The Annals of Statistics* pages 86–94.

Liu, Y., Sun, X., and Zhao, S. (2017). A covariance matrix shrinkage method with toeplitz rectified target for doa estimation under the uniform linear array. *AEU-International Journal of Electronics and Communications* **81,** 50–55.

Markowetz, F. and Spang, R. (2007). Inferring cellular networks – a review. *BMC Bioinformatics* **8,** S5.

Mestre, X. (2008). On the asymptotic behavior of the sample estimates of eigenvalues and eigenvectors of covariance matrices. *IEEE Transactions on Signal Processing* **56,** 5353–5368.

Osareh, A. and Shadgar, B. (2009). Classification and diagnostic prediction of cancers using

gene microarray data analysis. *Journal of Applied Sciences* **9,** 459–468.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*. The Regents of the University of California.

Robbins, H. (1955). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 157–164. Berkeley and Los Angeles, University of California Press.

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

Saha, S., Guntuboyina, A., et al. (2020). On the nonparametric maximum likelihood estimator for gaussian location mixture densities with application to gaussian denoising. *Annals of Statistics* **48,** 738–762.

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,**.

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* **16,** 133–145.

Stein, C. (1975). Estimation of a covariance matrix. In *39th Annual Meeting IMS, Atlanta, GA, 1975*.

Stein, C. (1986). Lectures on the theory of estimation of many parameters. *Journal of Soviet Mathematics* **34,** 1373–1403.

Stigler, S. M. (1990). The 1988 Neyman memorial lecture: a Galtonian perspective on shrinkage estimators. *Statistical Science* pages 147–155.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* pages 5–42.

Xue, L., Ma, S., and Zou, H. (2012). Positive-definite l_1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association* **107,** 1480–1491.

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* **4,**.

Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods. *The Annals of Statistics* **31,** 379–390.

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. Using the fact that when $\mathbb{E}\boldsymbol{X} = \mathbf{0}$, $\mathbb{E}s_{jk} = \sigma_{jk}$ for all $j, k = 1, \ldots, p$. Therefore for the class of linear decision rules (4), the scaled Frobenius risk (1) equals

$$\begin{aligned}
R(\boldsymbol{\Sigma}, \boldsymbol{\delta}) &= \frac{1}{p^2}\sum_{j,k=1}^{p} \mathbb{E}\{(\beta_S s_{jk} + \beta_I u_{jk} - \sigma_{jk})^2\} \\
&= \frac{1}{p^2}\sum_{j,k=1}^{p} \mathbb{E}[s_{jk} - \sigma_{jk} - \{(1-\beta_S)s_{jk} - \beta_I u_{jk}\}]^2 \\
&= \frac{1}{p^2}\sum_{j,k=1}^{p} [\mathbb{E}\{(s_{jk} - \sigma_{jk})^2\} + \mathbb{E}\{(1-\beta_S)s_{jk} - \beta_I u_{jk}\}^2 - 2(1-\beta_S)\mathbb{E}\{s_{jk}(s_{jk} - \sigma_{jk})\}] \\
&= \frac{1}{p^2}\sum_{j,k=1}^{p} [(2\beta_S - 1)\mathrm{Var}(s_{jk}) + \mathbb{E}\{(1-\beta_S)s_{jk} - \beta_I u_{jk}\}^2] \\
&= \frac{1}{p^2}\sum_{j,k=1}^{p} \mathbb{E}[(2\beta_S - 1)\frac{n}{n-1}\hat{\Delta}_{jk}^2 + \{(1-\beta_S)s_{jk} - \beta_I u_{jk}\}^2],
\end{aligned}$$

with $\hat{\Delta}_{jk}$ defined in Proposition 2. Therefore

$$\begin{aligned}\mathbb{E}\hat{R}(\beta_S,\beta_I) - R(\boldsymbol{\Sigma},\boldsymbol{\delta}) &= \frac{-1}{n-1}(2\beta_S-1)\frac{1}{p^2}\sum_{j,k=1}^{p}\mathbb{E}\hat{\Delta}_{jk}^2 = \frac{-1}{n}(2\beta_S-1)\frac{1}{p^2}\sum_{j,k=1}^{p}\mathrm{Var}(s_{jk}) \\ &= \frac{-1}{n}(2\beta_S-1)\frac{1}{p^2}\mathbb{E}\|\boldsymbol{S}-\boldsymbol{\Sigma}\|_F^2 \to 0,\end{aligned}$$

where the last result follows because by assumption, $p^{-2}\mathbb{E}\|\boldsymbol{S}-\boldsymbol{\Sigma}\|_F^2$ is bounded as $n\to\infty$.

Proof of Proposition (2)

Proof. The estimator in Ledoit and Wolf (2004) is

$$\boldsymbol{\Sigma}_{LW} = (\frac{d_n^2-b_n^2}{d_n^2})_+\boldsymbol{S} + \min\{1,\frac{b_n^2}{d_n^2}\}\hat{\mu}\boldsymbol{I} \tag{A.1}$$

where $d_n^2 = \frac{1}{p}\sum_{j,k=1}^{p}s_{jk}^2 - \frac{1}{p^2}(\sum_{j=1}^{p}s_{jj})^2$, $b_n^2 = \frac{1}{pn^2}\sum_{j,k=1}^{p}\sum_{i=1}^{n}(X_{ij}X_{ik}-s_{jk})^2$, $\hat{\mu} = \frac{1}{p}\sum_{j=1}^{p}s_{jj}$.

We first rewrite the risk estimate $\hat{R}(\beta_S,\beta_I)$. Define $\boldsymbol{M} = (\sum_{j,k=1}^{p}\hat{\Delta}_{jk}^2, 0)^\top$, $\boldsymbol{\beta} = (\beta_S,\beta_I)^\top$, 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

$$p^2\hat{R}(\beta_S,\beta_I) = \boldsymbol{\beta}^\top(\boldsymbol{Z}^\top\boldsymbol{Z})\boldsymbol{\beta} - 2(\boldsymbol{Z}^\top\boldsymbol{v_S}-\boldsymbol{M})^\top\boldsymbol{\beta} - \mathbf{1}^\top\boldsymbol{M} + \boldsymbol{v}_S^\top\boldsymbol{v}_S,$$

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

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\hat{R}(\beta_S,\beta_I) = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}(\boldsymbol{Z}^\top\boldsymbol{v_S}-\boldsymbol{M}),$$

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

We will need to show $\hat{\mu}b_n^2/d_n^2 \to \hat{\beta}_I$ and $\hat{\beta}_I/\hat{\mu} + \hat{\beta}_S = 1$. 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 $\det(\boldsymbol{Z}^\top\boldsymbol{Z}) = p\sum_{j,k=1}^{p}s_{jk}^2 - (\sum_{j=1}^{p}s_{jj})^2 = p^2d_n^2$, it follows that

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

and in addition

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

Therefore

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

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

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

which is very closed to the coefficient of $\boldsymbol{I}$ in (A.1) Furthermore,

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

Proof of Proposition (3)

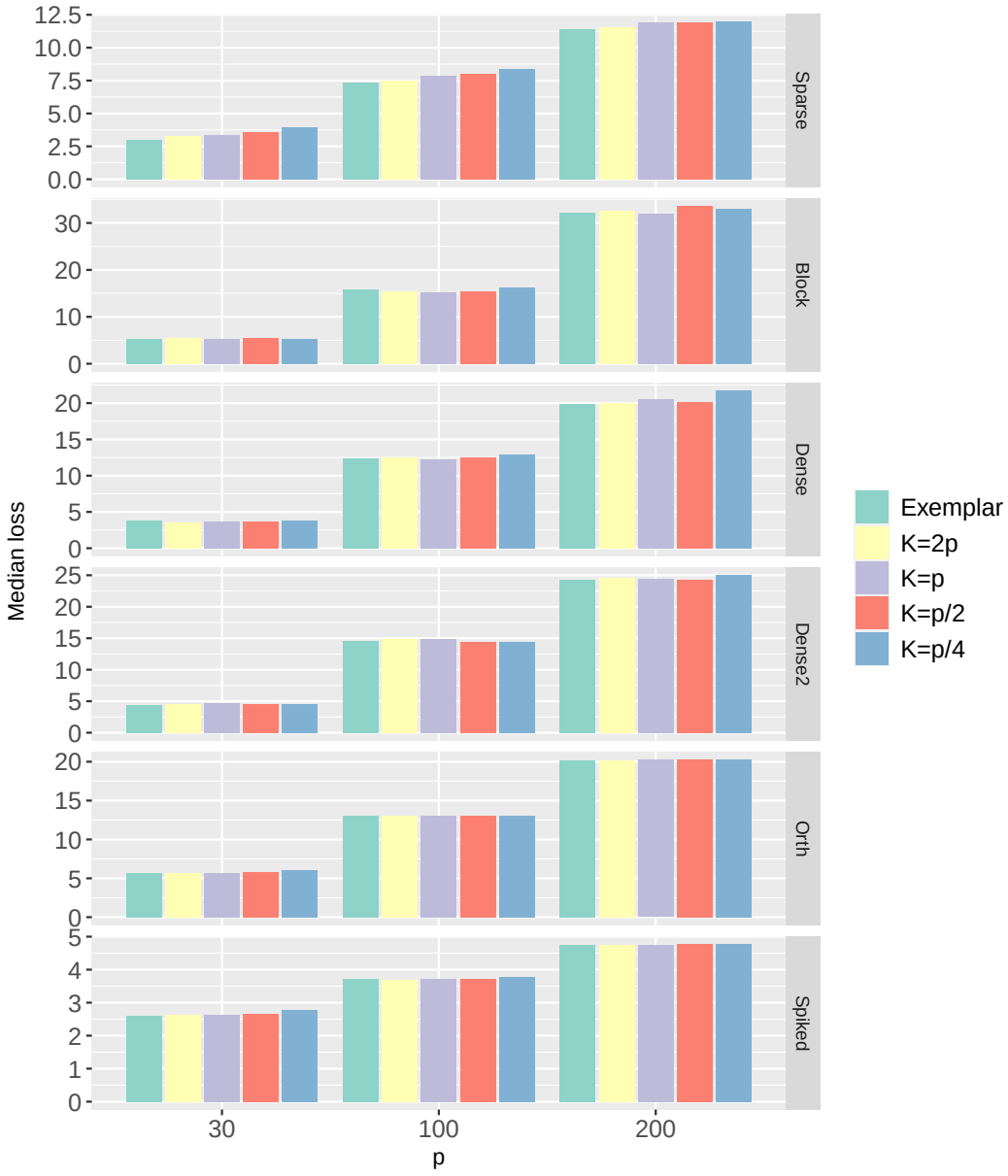

Figure 1. Median Frobenius norm errors over 200 replications. The Sparse, Block, Dense, Dense2, Orth and Spiked panels correspond to Models 1 through 6, respectively.

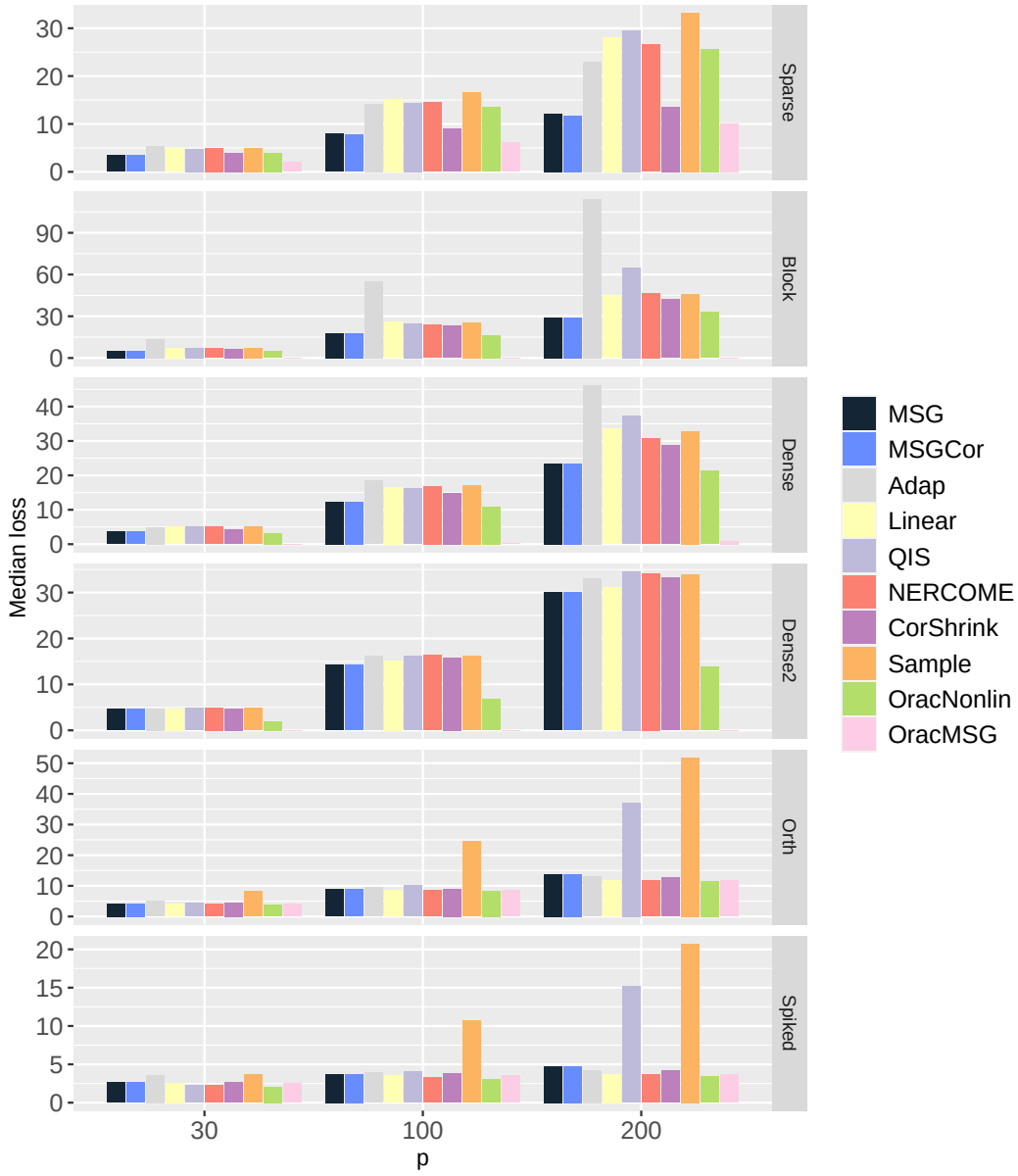

Figure 2. Median Frobenius norm errors over 200 replications. The Sparse, Block, Dense, Dense2, Orth and Spiked panels correspond to Models 1 through 6, respectively. In MSG, K-means clustering is applied with $K = p$

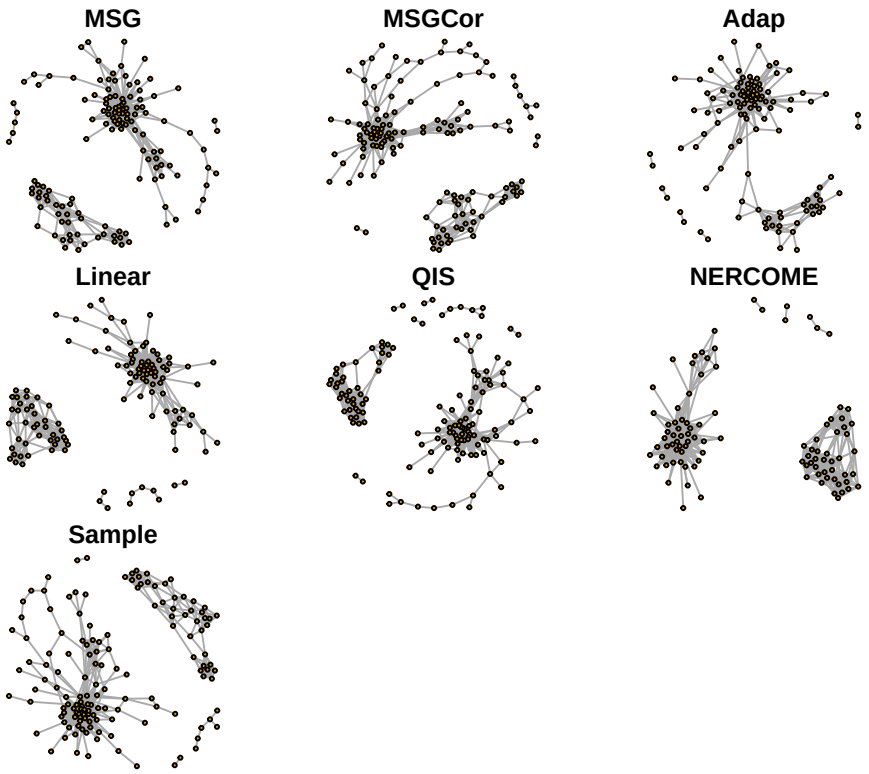

Figure 3. Gene networks recovered by the different covariance matrix estimation methods.

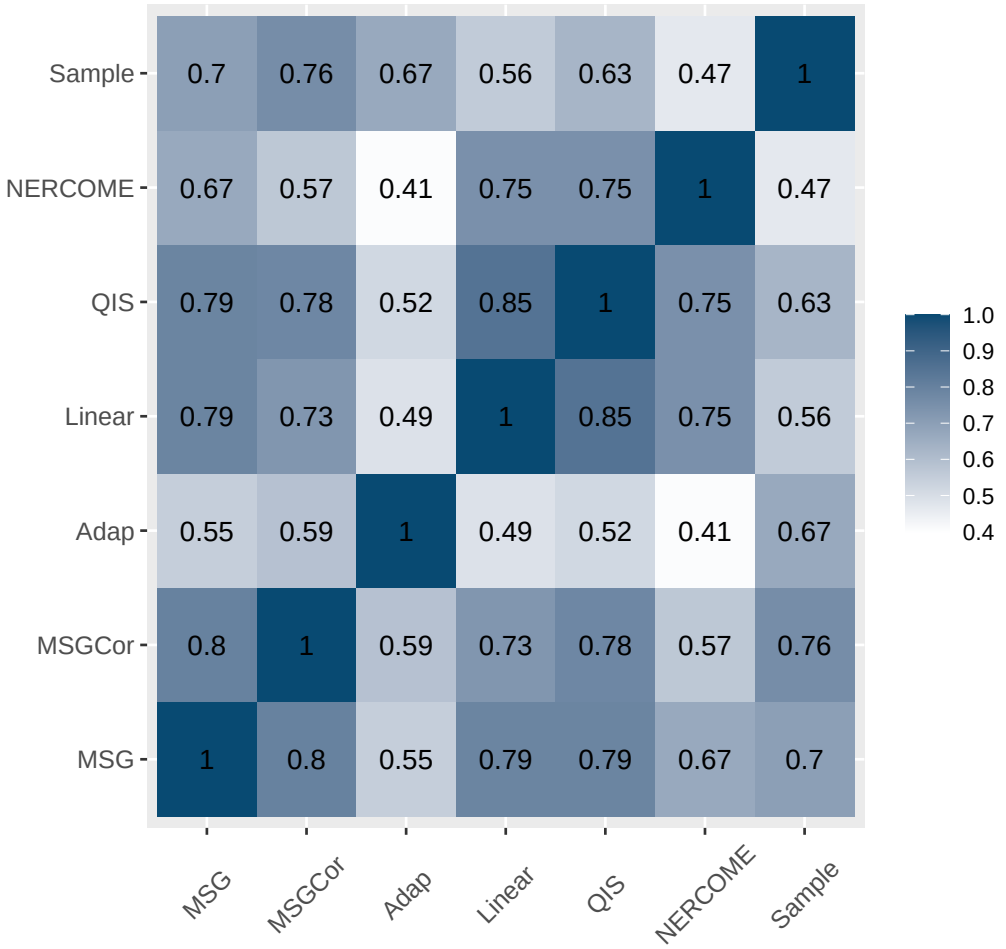

Figure 4. Similarities of gene degrees between the estimated networks. Each number reports the Jaccard index between the top 20% most connected genes of each pair of networks.

Table 1
Average running time for different ratios.

	p=30	p=100	p=200
Exemplar method	0.1357	6.0338	91.8942
$K = 2p$	0.0539	1.1883	10.7687
$K = p$	0.0357	0.7852	7.0807
$K = p/2$	0.0222	0.5231	4.6736
$K = p/4$	0.0158	0.3694	3.2749

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

	Frobenius	Spectral	Matrix ℓ_1
MSG	24.13 (2.54)	13.14(2.96)	46.53(8.49)
MSGCor	**24.04(2.56)**	**13.12(2.94)**	**46.00(8.67)**
Adap	28.93(2.03)	17.54(3.85)	49.66(8.47)
Linear	24.49(2.53)	13.65(3.14)	49.46(9.04)
QIS	28.76(2.79)	16.09(3.35)	58.73(9.74)
NERCOME	24.52(2.52)	13.19(3.07)	48.27(9.67)
Sample	28.43(2.54)	16.21(3.15)	56.08(9.89)