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 study a new approach to estimating high-dimensional covariance matrices. 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 an RNA-seq experiment in mouse show that our approach is comparable to or can outperform a number of state-of-the-art proposals, particularly when the sample eigenvectors are poor estimates of the population eigenvectors.

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. 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). For example, in Section 5 we study brain region-specific gene networks in mouse using bulk RNA-sequencing data.

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. Many of these methods can be interpreted as rotationally invariant estimators, which shrink

the sample eigenvalues in a data-adaptive fashion (Bun et al., 2016; Stein, 1975, 1986). One example is the linear shrinkage approach of Ledoit and Wolf (2004), which shrinks the eigenvalues toward their global mean. Some more recently developed nonlinear shrinkage methods have been shown to be asymptotically optimal in this class (Ledoit et al., 2012; Ledoit and Wolf, 2019; Lam et al., 2016).

Though these optimal estimators perform extremely well, Bun et al. (2016) showed that the eigenvectors of any rotationally invariant estimate must be the same as those of the sample covariance matrix. This can be problematic because sample eigenvectors are not consistent when the dimension and the sample size increase at the same rate (Mestre, 2008). This suggests that unstructured but non-rotationally invariant covariance matrix estimators may be worth exploring, as these can modify both the sample eigenvectors as well as the eigenvalues. The difficulty is to identify a reasonable class of estimators to pursue.

In this paper we define and explore one such class of estimators. This approach is based on interpreting the covariance matrix estimation problem as a compound decision problem (Robbins, 1951). Section 2 shows how this framing motivates vectorizing the covariance matrix and solving the resulting vector estimation problem using nonparametric empirical Bayes procedures studied in the compound decision literature (Jiang and Zhang, 2009; Koenker and Mizera, 2014; Efron, 2019). The vector estimate can then be reassembled into matrix form and projected into the space of positive-definite matrices to give a final covariance matrix estimator. Surprisingly, though this vectorized approach essentially ignores the matrix structure, it can sometimes outperform a number of state-of-the-art proposals in simulations and a real data analysis. Our results suggest that vectorized approaches may be preferrable when the sample eigenvectors are poor estimates of the population eigenvectors. Our compound decision approach is closely related to the correlation matrix estimator of

Dey and Stephens (2018). We discuss in detail the connections with their work in Section 3.2.

The article is organized as follows. In Section 3, we briefly review compound decision theory and then introduce our proposed estimator. In Section 4 we illustrate its performance in simulations and a gene coexpression network inference problem. Finally, Section 6 concludes with a discussion. Our procedure is implemented in the R package `cole`, available on GitHub at `http://github.com/sdzhao/cole`.

2. Covariance matrix estimation as a compound decision problem

2.1 *Problem formulation*

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 $\boldsymbol{\Sigma}$. This section argues that covariance matrix estimation can be viewed as a compound decision problem when estimation performance is measured using 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 compound decision formulation will motivate a new class of unstructured non-rotationally invariant estimators.

Compound decision problems involve 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 $Y_i \sim P_{\theta_i}$ (Robbins, 1951). 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 $n^{-1} \sum_{i=1}^{n} \mathbb{E} L(\theta_i, \delta_i(\boldsymbol{Y}))$, 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 classical 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).

Covariance matrix estimation under the Frobenius risk (1) can therefore be viewed as a compound decision problem where the goal is to simultaneously estimate the components of the vector

$$\boldsymbol{\sigma} = (\sigma_{11}, \ldots, \sigma_{1p}, \ldots, \sigma_{p1}, \ldots, \sigma_{pp})^\top \tag{2}$$

under average squared error loss. 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-definite. This will in fact be true for one of the estimators we propose in Section 3.2. To resolve this issue, we can project any estimate into the space of positive-definite matrices; see Section 3.4.

The value of this formulation is that it allows us to apply ideas from the compound decision literature to covariance matrix estimation. A key property of compound decision problems is that while information about index i seems like it should offer no help for estimating a parameter with a different index j, in fact borrowing information across all indices is superior to considering each index alone. An example of this phenomenon is the James-Stein estimator (James and Stein, 1961), which estimates $\boldsymbol{\theta} = E\boldsymbol{Y}$ 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). A natural next step is therefore to apply modern ideas for Gaussian sequence estimation to covariance matrix estimation, which we do in Section 3.

2.2 *Reinterpretation of existing methods*

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 reinterpreted as carrying out vector estimation. The simplest example is the sample covariance matrix $\boldsymbol{S} = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$, which takes this form because the $\boldsymbol{X}$ are assumed to have mean zero. As another example, 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.

As a further illustration of the potential of our new formulation, we can show that the linear shrinkage covariance matrix estimator of Ledoit and Wolf (2004) can be essentially recovered as the solution to a compound decision problem. As described in Section 2.1, we first view the problem of minimizing the scaled Frobenius norm (1) as simultaneously estimating the components of the vector $\boldsymbol{\sigma}$ (2) under squared error. Then a standard approach in the compound decision literature is to first define a class of decision rules for estimating each component and then identify the member of that class that minimizes the compound risk. For example, in the Gaussian sequence problem the Efron-Morris estimator (Efron and Morris, 1973) can be viewed as the optimal linear decision rule (Fourdrinier et al., 2018; Stigler, 1990).

Following this recipe, we also restrict attention to the class of linear decision rules where σ_{jk} will be estimated by

$$\delta_{jk}(\boldsymbol{X}) = \beta_S s_{jk} + \beta_I u_{jk} \tag{3}$$

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 $\boldsymbol{I}$. Ideally we would like to estimate the optimal β_S and β_I by choosing the values that

minimize the Frobenius risk (1). Though we cannot directly calculate (1) because it involves the unknown σ_{jk}, the following proposition shows that we can construct a good estimate of this risk for every fixed value of β_S and β_I.

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

$$\hat{R}_L(\beta_S, \beta_I) = \frac{1}{p^2} \sum_{j,k=1}^{p} [(2\beta_S - 1)\hat{\Delta}_{jk}^2 + \{(1-\beta_S)s_{jk} - \beta_I u_{jk}\}^2]. \tag{4}$$

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

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

for the class of decision rules $\boldsymbol{\delta}$ defined in (3). $\|\cdot\|_F$ denotes the Frobenius norm $\|\boldsymbol{A}\|_F^2 = \sum_{j,k=1}^p A_{jk}^2$.

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 $\hat{R}_L(\beta_S, \beta_I)$ (4) is an asymptotically unbiased estimate of the true risk (1) for linear decision rules of the form (3). It is then reasonable to estimate the best estimator in this class by minimizing $\hat{R}_L(\beta_S, \beta_I)$ over β_S and β_I. A slight modification of this procedure turns out to recover the Ledoit and Wolf (2004) estimator

$$\hat{\boldsymbol{\Sigma}}_{\text{LW}} = \left(1 - \frac{b_n^2}{d_n^2}\right)\boldsymbol{S} + \frac{b_n^2}{d_n^2}\hat{\mu}\boldsymbol{I}, \tag{5}$$

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

PROPOSITION 2: Let $\hat{\beta}_S$ and $\hat{\beta}_I$ denote the minimizers of $\hat{R}_L(\beta_S, \beta_I)$ 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 $\hat{\delta}_{jk}(\boldsymbol{X}) = \hat{\sigma}_{\text{LW}jk}$, where $\hat{\sigma}_{\text{LW}jk}$ is the jkth entry of $\boldsymbol{\Sigma}_{\text{LW}}$ (5).

There are also interesting connections between (6) 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 (3). It can be shown that $\tilde{\delta}_{jk}(\boldsymbol{X})$ is equivalent to estimating $\boldsymbol{\sigma}$ by shrinking the vector $(s_{11}, \ldots, s_{1p}, \ldots, s_{p1}, \ldots, s_{pp})^\top$ toward the one-dimensional subspace spanned by $(u_{11}, \ldots, u_{1p}, \ldots, u_{p1}, \ldots, u_{pp})^\top$ (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 similar to how replacing the shrinkage factor in the classical James-Stein estimator by its positive part results can improve its accuracy (Baranchik, 1964).

3. Methods

3.1 *New class of estimators*

Proposition 2 of Section 2.2 suggests that applying ideas from the compound decision literature to estimating the vector $\boldsymbol{\sigma}$ (2) can be a fruitful strategy for covariance matrix estimation. However, Proposition 2 considers only a linear decision rule, whereas more recent work in compound decision theory has focused on the much larger class of so-called separable rules (Brown and Greenshtein, 2009; Jiang and Zhang, 2009; Zhang, 2003). In the standard compound decision problem of estimating $\boldsymbol{\theta} = E\boldsymbol{Y}$ 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). Linear rules such as (6) are included in this class, so that the best separable rule should perform at least as well as the best linear rule.

Motivated by these ideas, here we propose a new class of covariance matrix estimators by generalizing the notion of separable rules 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

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

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 t_d and off-diagonal entries using t_{od}, where t_{od} and t_d 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, the class of adaptive thresholding estimators for sparse covariance matrices studied by Cai and Liu (2011), and the class of linear estimators (3) used by Ledoit and Wolf (2004), which can be expressed as

$$t_{od}(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k}) = \beta_S \boldsymbol{X}_{\cdot j}^\top \boldsymbol{X}_{\cdot k}/n,\ 1 \leqslant k < j \leqslant p,$$

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

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 $\boldsymbol{X}_{\cdot k}$.

The class $\mathcal{S}$ constitutes a fundamentally different approach to covariance matrix estimation compared to the existing methods described in Section 1. Estimators in $\mathcal{S}$ do not assume that $\boldsymbol{\Sigma}$ has any particular structure and also are not necessarily rotationally invariant. Intuitively, they should perform better than rotationally invariant estimators when the sample eigenvectors of $\boldsymbol{S}$ are very different from the population eigenvectors. Results in Section 4 suggest that this is indeed the case.

3.2 *Proposed estimator*

We propose to search for the optimal estimator $\boldsymbol{\delta}^\star$ within $\mathcal{S}$ (7). We can first provide a closed-form expression for $\boldsymbol{\delta}^\star$. Let $f_2(\cdot \mid a, b, \gamma)$ be the density of a bivariate normal centered at zero with standard deviations a and b and correlation γ, $f_1(\cdot \mid a)$ be the density of a

univariate normal with mean zero and standard deviation a, and

$$\begin{aligned} G_{od}(a,b,\gamma) &= \frac{1}{p(p-1)} \sum_{1\leqslant j,k\leqslant p, j\neq k} I(\sigma_j \leqslant a, \sigma_k \leqslant b, r_{jk} \leqslant \gamma), \\ G_d(a) &= \int dG_{od}(a,b,\gamma) dbd\gamma = \frac{1}{p}\sum_{j=1}^{p} I(\sigma_j \leqslant a), \end{aligned} \tag{8}$$

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.

Proposition 3: The optimal separable estimator

$$\boldsymbol{\delta}^\star = \arg\min_{\boldsymbol{\delta}\in\mathcal{S}} R(\boldsymbol{\Sigma}, \boldsymbol{\delta}) = \begin{cases} t^\star_{od}(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k}), & 1 \leqslant k < j \leqslant p, \\ t^\star_d(\boldsymbol{X}_{\cdot j}), & j = 1, \ldots, p, \end{cases} \tag{9}$$

obeys

$$\begin{aligned} t^\star_{od}(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k}) &= \frac{\int ab\gamma f_{2n}(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k} \mid a,b,\gamma) dG_{od}(a,b,\gamma)}{\int f_{2n}(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k} \mid a,b,\gamma) dG_{od}(a,b,\gamma)}, \\ t^\star_d(\boldsymbol{X}_{\cdot j}) &= \frac{\int a^2 f_{1n}(\boldsymbol{X}_{\cdot j} \mid a) dG_d(a)}{\int f_{1n}(\boldsymbol{X}_{\cdot j} \mid a) dG_d(a)} \end{aligned} \tag{10}$$

where

$$f_{2n}(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k} \mid a,b,\gamma) = \prod_{i=1}^{n} f_2(X_{ij}, X_{ik} \mid a,b,\gamma), \quad f_{1n}(\boldsymbol{X}_{\cdot j} \mid a) = \prod_{i=1}^{n} f_1(X_{ij} \mid a). \tag{11}$$

Proposition 3 illustrates the key property of compound decision problems: that borrowing information across indices can help estimate a single parameter. In this case, the optimal separable rule for estimating σ_{jk} borrows from all of the σ_j and r_{jk}. This rule also has an interesting Bayesian interpretation, as $t^\star_{od}$ and $t^\star_d$ are formally equivalent to posterior expectations when the $(\sigma_j, \sigma_k, r_{jk})$ are independently and identically distributed according to the prior G_{od}. However, the result in Proposition 3 holds even when the parameters are not random. Proposition 3 is a direct extension of the fundamental theorem of compound decisions (Robbins, 1951; Zhang, 2003).

The optimal $\boldsymbol{\delta}^\star$ (9) is an oracle estimator and cannot be calculated in practice. A common approach in the compound decision literature is to use empirical Bayes ideas to estimate the oracle optimal separable decision rule (Robbins, 1955; Zhang, 2003; Brown and Greenshtein,

2009; Jiang and Zhang, 2009; Efron, 2014, 2019). We propose to estimate $\boldsymbol{\delta}^\star$ using what Efron (2014) calls g-modeling, where we proceed as if the $(\sigma_j, \sigma_k, r_{jk})$ were truly random and then estimate the prior G_{od} from the data. Specifically, under the working assumption that the $(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k})$ are independent for different (j, k), we propose to calculate estimate G_{od} and G_d (8) using nonparametric maximum likelihood (Kiefer and Wolfowitz, 1956):

$$\hat{G}_{od} = \underset{G_{od} \in \mathcal{G}_{od}}{\arg\max} \prod_{1 \leqslant k < j \leqslant p} \int f_{2n}(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k} \mid a, b, \gamma) dG_{od}(a, b, \gamma) \prod_{j=1}^{p} \int f_{1n}(\boldsymbol{X}_{\cdot j} \mid a) dG_d(a), \quad (12)$$

where $\mathcal{G}_{od}$ is the family of all distributions supported on $\mathbb{R}_+ \times \mathbb{R}_+ \times [-1, 1]$ and G_d is determined by G_{od} as indicated in (8). Of course, the $(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k})$ are not independent across different (j, k) unless the true covariance $\boldsymbol{\Sigma}$ is diagonal, so $\hat{G}_{od}$ does not maximize a likelihood but rather a pairwise composite likelihood (Varin et al., 2011). Using $\hat{G}_{od}$ and $\hat{G}_d$, we propose to estimate the vectorized $\boldsymbol{\Sigma}$ using

$$\hat{\boldsymbol{\delta}}(\boldsymbol{X}) = (\hat{t}_d(\boldsymbol{X}_{\cdot 1}), \ldots, \hat{t}_{od}(\boldsymbol{X}_{\cdot 1}, \boldsymbol{X}_{\cdot p}), \ldots, \hat{t}_{od}(\boldsymbol{X}_{\cdot p}, \boldsymbol{X}_{\cdot 1}), \ldots, \hat{t}_d(\boldsymbol{X}_{\cdot p})) \quad (13)$$

where $\hat{t}_{od}$ and $\hat{t}_d$ are obtained by plugging $\hat{G}_{od}$ and $\hat{G}_d$ (12) into the expressions for $t^\star_{od}$ and $t^\star_d$ from (10).

Our proposed procedure is similar to the method of Dey and Stephens (2018), who introduce an empirical Bayes approach to estimating a correlation matrix. They first apply Fisher's Z-transform to each off-diagonal entry of the sample correlation matrix and model the means of the transformed entries as independent random variables arising independently from some prior distribution. They then estimate the prior from the data, calculate the posterior expectations of the means, transform them back into a correlation matrix, and select the closest positive-definite matrix.

The main difference between our approaches is that Dey and Stephens (2018) do not shrink the variances while we do, as our $\hat{t}_d$ estimates σ_j using its supposed posterior expectation assuming a prior distribution of $\hat{G}_d$. Furthermore, our Proposition 3 shows that our shrinkage approach can be interpreted as viewing the triples $(\sigma_j, \sigma_k, r_{jk})$ as if they were drawn from

the trivariate prior G_{od}, which allows for possible dependencies between r_{jk} and the σ_j. For example, model 2 in our simulations in Section 4 considers a covariance matrix where entries with larger standard deviations also tend to have larger correlations. In contrast to the method of Dey and Stephens (2018), our approach can learn this structure if it exists and take advantage of it for more accurate estimation.

3.3 *Implementation*

Calculating the estimated prior $\hat{G}_{od}$ (12) is difficult, as it is an infinite-dimensional optimization problem over the class of all probability distributions $\mathcal{G}_{od}$ 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 G_{od} is supported on D fixed support points $(a_\tau, b_\tau, \gamma_\tau)$, $\tau = 1, \ldots, D$. Since G_{od} is symmetric in its first two coordinates, we construct the support points so that D is even and for $\tau > D/2$, $(a_\tau, b_\tau, \gamma_\tau) = (b_{\tau - D/2}, a_{\tau - D/2}, \gamma_{\tau - D/2})$. We also constrain $\hat{G}_{od}$ such that the estimated mass at the point $(a_\tau, b_\tau, \gamma_\tau)$ is equal to the mass at $(b_\tau, a_\tau, \gamma_\tau)$. Finally, we use the EM algorithm to estimate the masses $\hat{w}_\tau$. Letting $\hat{w}_\tau^{(k)}$ denote the estimate of the τth mass point as the kth iteration, it can be shown that for

$\tau = 1, \ldots, D/2$, the update formula is

$$\begin{aligned}
&\hat{w}_\tau^{(k)} = \hat{w}_{\tau+D/2}^{(k)} \\
&= \frac{2}{p(p+1)} \left[\sum_{j=1}^{p} \frac{\hat{w}_\tau^{(k-1)} \{ f_{1n}(\boldsymbol{X}_j \mid a_\tau) + f_{1n}(\boldsymbol{X}_j \mid a_{\tau+D/2}) \}}{\sum_{l=1}^{D} \hat{w}_l^{(k-1)} f_{1n}(\boldsymbol{X}_j \mid a_l)} + \right. \\
&\quad \left. \sum_{1 \leqslant k < j \leqslant p} \frac{\hat{w}_\tau^{(k-1)} \{ f_{2n}(\boldsymbol{X}_j, \boldsymbol{X}_k \mid a_\tau, b_\tau, \gamma_\tau) + f_{2n}(\boldsymbol{X}_j, \boldsymbol{X}_k \mid a_{\tau+D/2}, b_{\tau+D/2}, \gamma_{\tau+D/2}) \}}{\sum_{l=1}^{D} \hat{w}_l^{(k-1)} f_{2n}(\boldsymbol{X}_j, \boldsymbol{X}_k \mid a_l, b_l, \gamma_l)} \right].
\end{aligned}$$

In practice, we center each $\boldsymbol{X}_{\cdot j}$ and instead of using the density functions f_{1n} and f_{2n} (11), we use the densities of their sufficient statistics: the sample variance s_j^2, in the case of f_{1n}, and the 2×2 matrices with diagonal entries s_j^2 and s_k^2 and off-diagonal entries equal to the sample covariances s_{jk}, in the case of f_{2n}.

The support points need to be carefully chosen. Ideally, the points $(a_\tau, b_\tau, \gamma_\tau)$ would densely cover the parameter space. However, using a grid of d points in each dimension would require a total of $D = d^3$ points, which would incur a huge computational cost for even moderate d. Alternatively, we could use the so-called exemplar method (Saha et al., 2020), which sets the support points to equal the observed sample versions of the $(\sigma_j, \sigma_k, r_{jk})$. This would reduce the size of the support set, but the computation complexity would still grow like $O(p^2)$.

Instead, we use a clustering-based exemplar algorithm to further improve computational efficiency. Let s_j, s_k, and γ_{jk} be the sample standard deviations and correlation between the jth and kth covariates. We first apply K-means clustering to identify K cluster centroids using the $p(p-1)/2$ lower triangular off-diagonal sample points (s_j, s_k, γ_{jk}). We then swap the first two coordinates of each point to construct another set of K centroids. Finally, we use these $2K$ centroids as the support points for $\hat{G}_{od}$. We evaluate the impact of changing K in Section 4.

Our procedure is implemented in the R package `cole`, available on GitHub. In our implementations of the EM algorithm in Section 4 we stopped iterating when the relative difference between two successive log-likelihoods was less than 10^{-4}, up to a maximum of

200 iterations. Instead of EM, we attempted to maximize the composite likelihood (12) using the interior point solver MOSEK, as advocated by Koenker and Mizera (2014), but we ran into computational difficulties because for even moderate sample sizes n, the values of the densities f_{1n} and f_{2n} at certain support points could be extremely small. These small values caused problems for the R wrapper for MOSEK. Our implementation of the EM algorithm was more robust to these small density values.

3.4 *Positive definiteness correction*

Our proposed estimator (13) 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

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

where $\boldsymbol{Q}$ 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

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

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 our implementations we used $K = 20$ and $\alpha_K = 10$.

4. Numerical Results

4.1 *Covariance matrix models*

We considered six models for the population covariance matrix to explore the behavior of our proposed method in simulations. In each model, we generated certain parameters randomly

but then fixed them across all replications of our simulations. For readability, below we give each model a short name that we will refer to in the following figures and captions.

- Model 1 (sparse): we generated a sparse $\boldsymbol{\Sigma}$ where most features were not correlated. We let half of the standard deviations equal 1 and the other half equal 1.5. We set the correlation matrix following Model 2 of Cai and Liu (2011):
$$\begin{pmatrix} \boldsymbol{A}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{p/2\times p/2} \end{pmatrix},$$
where the jkth entry of $\boldsymbol{A}_1$ is $a_{jk} = \max(1 - |j-k|/10, 0)$.
- Model 2 (hypercorrelated): we generated a $\boldsymbol{\Sigma}$ that we call "hypercorrelated" because its parameters are in a sense dependent on each other, as we let larger standard deviations σ_j and σ_k correspond to larger r_{jk}. Specifically, we let the first $p/2$ standard deviations equal 1, the last $p/2$ equal 2, and the correlation matrix equal
$$\begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{pmatrix},$$
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. As discussed in Section 3.2, we expect that our proposed approach will be able to take advantage of the dependencies between the σ_j and the r_{jk}, unlike the empirical Bayes approach of Dey and Stephens (2018) that does not shrink standard deviations.
- Model 3 (dense-0.7): we generated a $\boldsymbol{\Sigma}$ where the correlation between all features was 0.7, in contrast to our sparse Model 1. Otherwise, we generated the standard deviations as in Model 1.
- Model 4 (dense-0.9): this setting was the same as Model 3 except with correlation parameter 0.9. This high level of dependence tests the robustness of the pairwise composite likelihood estimator (12).
- Model 5 (orthogonal): instead of generating $\boldsymbol{\Sigma}$ by specifying its correlation matrix, we

randomly generated a $\boldsymbol{U}$ from the Haar measure on the space of all orthonormal matrices and a vector of eigenvalues $\boldsymbol{l}$ independently generated from $\mathcal{U}(1,4)$. We then let $\boldsymbol{\Sigma} = \boldsymbol{U}^T \text{diag}(\boldsymbol{l})\boldsymbol{U}$. This simulation setting was used in Lam et al. (2016) and Ledoit and Wolf (2019).

- Model 6 (spiked): this setting was the same as model 5 except that $\boldsymbol{l}$ was a vector of eigenvalues where the first 3 entries were 4,3,2 and the remaining $p-3$ entries equaled 1, so that $\boldsymbol{\Sigma}$ was a spiked covariance matrix.

4.2 *Methods compared*

We compared several high-dimensional covariance matrix estimation procedures.

- Our proposed estimator (13) and its positive-definiteness-corrected version described in Section 3.4. In the following figures and captions we refer to these as MSG (Matrix Shrinkage via G-modeling) and MSGCor, respectively.
- 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.
- Linear: the linear shrinkage estimator of Ledoit and Wolf (2004) given in (5).
- 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 chose 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 Lam et al. (2016), we repeated this procedure 50

times and took the final covariance matrix estimator to be the average of the individual matrices.

- CorShrink: the empirical Bayes correlation matrix estimation method of Dey and Stephens (2018). To estimate $\mathbf{\Sigma}$, we used CorShrink to estimate the correlation matrix and then scaled its entries using the sample standard deviations.
- Sample: the sample covariance matrix.

In addition to the above data-driven 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, derived in Ledoit and Wolf (2019), of $\boldsymbol{U}^T \text{diag}(\boldsymbol{l})\boldsymbol{U}$, 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 $d_i = \boldsymbol{u}_i^T \mathbf{\Sigma} \boldsymbol{u}_i$.
- OracMSG: we implemented our proposed estimator (13) except that the support points are generated by clustering the true parameters $(\sigma_j, \sigma_k, r_{jk})$ instead of their sample versions.

4.3 *Clustering-based exemplar algorithm*

We first explored the consequences of choosing different numbers of cluster centroids in our clustering-based exemplar algorithm described in Section 3.3. For each model, we generated $n = 100$ samples from a p-variate $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, 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 loss function

$$\frac{1}{p}\left\{\sum_{j,k=1}^{p}(\hat{\sigma}_{jk} - \sigma_{jk})^2\right\}^{1/2}, \tag{15}$$

where $\hat{\mathbf{\Sigma}}$ is the estimated matrix with entries $\hat{\sigma}_{jk}$ and $\mathbf{\Sigma}$ is the true matrix with entries σ_{jk}. We then fit our proposed method, after positive-definitness correction, for different K, where we let $K = rp$ for $r = 2, 1, 0.5, 0.25$.

[Figure 1 about here.]

[Table 1 about here.]

Figure 1 shows that across all models, different K give similar estimation accuracies compared to using all sample points in the exemplar algorithm. Table 1 shows that they can be significantly faster; we only provide results from Model 1 because the times did not vary much across different models. 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.

4.4 *Estimation accuracies*

We next compared the estimation accuracies of the methods in Section 4.2 under loss (15), following the same simulation scheme as in Section 4.3. The results are visualized in Figure 2. Our proposed methods could outperform existing methods, especially in Models 1 through 4. This pattern also held when we ran simulations with $p = 1000$ in the Supporting Information. The corrected version consistently outperformed the uncorrected version. Our methods could outperform the CorShrink method of Dey and Stephens (2018) in many cases, likely because it was able to accurately shrink the standard deviations.

[Figure 2 about here.]

MSG and MSGCor were especially good in Model 2, where the standard deviation and correlations were related, because it was able to capture this dependence in its estimate of the prior G_{od} (8) and leverage it to provide much more accurate estimates. It was substantially better than CorShrink in this setting, further illustrating the advantage of shrinking the standard deviations. In the Supporting Information we also compare MSG and MSGCor to CorShrink for estimating correlation instead of covariance matrices. We find that in that case, CorShrink performs slightly better than MSG and MSGCor except in Model 2, likely because the additional flexibility that our method trades off lower bias for higher variance.

However, in Models 5 and 6 MSG and MSGCor were worse than the best of the other methods, though they were still comparable. In the Supporting Information we found this pattern was exacerbated when $p = 1000$. This observation is consistent with the behavior of the oracle estimator OracMSG, which was dramatically better than the oracle rotationally-invariant estimator OracNonlin in Models 1 through 4 but was comparable to it in Models 5 and 6. As discussed in Section 3.1, this suggests that the population eigenvectors may be closer to the sample eigenvectors in Models 5 and 6. To explore this possibility, we recorded the Frobenius norms of the differences between the matrices of sample and population eigenvectors in each of our simulation settings. The results in Figure 3 seem to support our hypothesis that our class of separable estimators is best-suited for estimating covariance matrices whose sample and population eigenvalues differ greatly.

[Figure 3 about here.]

5. Data analysis

Covariance matrix estimation is often used to reconstruct gene networks (Markowetz and Spang, 2007). We used our proposed methods and the other covariance matrix estimators described in Section 4.2 for gene network estimation in a study of the transcriptomic response to social aggression in various brain regions in mouse (Saul et al., 2017). As part of the control condition, mice were exposed to a paper cup and bulk RNA-sequencing data was collected from the amygdala, frontal cortex, and hypothalamus at 30, 60, and 90 minutes after exposure. Data were collected from five mice at each time point, but two of the 30-minute hypothalamus samples were sequenced using a different library preparation than the other samples and so were left out of this analysis. The data are available from the NCBI Gene Expression Omnibus (GEO; `http://www.ncbi.nlm.nih.gov/geo/`) under accession number GSE80345.

Our goal was to estimate brain region-specific gene networks in these mice, assuming that the networks were the same across time points. Understanding how these networks differ between brain regions may provide insight into the different functions of the different regions. We built coexpression networks of the top 200 genes that were most differentially expressed between the regions. To identify these genes, we first considered only genes that were expressed in each of our samples and had at least one count per million mapped reads in at least 3 samples. We then performed a Kruskal-Wallis test and retained the 200 genes with the lowest p-values.

We then converted the gene expression values to log-counts per million mapped reads and applied the methods from Section 4.2 to estimate the covariance matrix of these 200 genes in each brain region. When implementing our K-means clustering-based exemplar algorithm, we let $K = 400$. We also only implemented MSGCor, since it was the best performer in our simulations.

[Table 2 about here.]

To measure the accuracy of the estimators, we randomly split the samples into a training and testing set. We estimated the covariance matrix on the training set and measured accuracy using the Frobenius norm of the difference between the estimated matrix and the sample covariance matrix calculated from the test set. We set the training sample size equal to 10, which left five samples in the test sets for the amygdala and frontal cortex and only three samples in the test set for hypothalamus. We therefore dropped hypothalamus from the analysis because we felt that three samples was unsuitable for calculating a surrogate of the true covariance matrix. We repeated this process 100 times. Table (2) reports the median errors and interquartile ranges across the replications. Many of the different methods had comparable performance, and the different brain regions favored different methods. Nevertheless, our proposed MSGCor had among the best performances in both regions.

[Figure 4 about here.]

[Figure 5 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. Figures 4 and 5 illustrate the estimated amygdala and frontal cortex 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. The network representations suggest that the resulting networks look fairly similar except for the ones produced by linear shrinkage and NERCOME. That our procedure produces networks that look similar to the unbiased sample covariance matrix networks is comforting.

6. Discussion

The class of separable covariance matrix estimators (7) that we explored in this paper appears to be promising. This is somewhat 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.

Providing theoretical guarantees for our estimator 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, though simulations in the Supporting Information suggest that our procedures can still provide relatively accurate estimates when the normality assumption is violated. To extend our procedure to non-normal data belonging to a parametric family, we can simply modify the density function f_1 and f_2 in the nonparametric maximum compositive likelihood problem (12) and in our proposed estimator (13). If f_1 and f_2 are unknown or difficult to specify, alternative procedures may be necessary to approximate the optimal separable rule.

Acknowledgments

We thank Dr. Roger Koenker and three anonymous referees for their very valuable comments.

References

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.

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.

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. 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. (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.

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.

Saul, M. C., Seward, C. H., Troy, J. M., Zhang, H., Sloofman, L. G., Lu, X., Weisner, P. A., Caetano-Anolles, D., Sun, H., Zhao, S. D., Chandrasekaran, S., Sinha, S., and Stubbs, L. (2017). Transcriptional regulatory dynamics drive coordinated metabolic and neural response to social challenge in mice. *Genome Research* **27,** 959–972.

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,**.

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 are available with this paper at the Biometrics website on Wiley Online Library.

Appendix

Proof of Proposition 1

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 (3), 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

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

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

We first rewrite the risk estimate $\hat{R}_L(\beta_S, \beta_I)$. 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 risk estimate can be re-written as

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

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

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

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

Since

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

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

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

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 $\hat{\beta}_I$, and

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

so $\min(\hat{\mu}, \hat{\beta}_I) = \hat{\mu} b_n^2/d_n^2$, which is the coefficient of the second term in the Ledoit and Wolf

estimator (5). Furthermore,

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

so $\min(\hat{\beta}_S, 0) = 1 - b_n^2/d_n^2$, which is the coefficient of the first term in (5).

Proof of Proposition 3

For any $\boldsymbol{\delta} \in \mathcal{S}$ (7), the Frobenius risk (1) can be written as

$$\begin{aligned}R(\boldsymbol{\Sigma}, \boldsymbol{\delta}) =& \frac{2}{p^2}\sum_{1\leqslant k<j\leqslant p}^{p}\int\{t_{od}(\boldsymbol{X}, \boldsymbol{X}') - \sigma_{jk}\}^2 f_{2n}(\boldsymbol{X}, \boldsymbol{X}' \mid \sigma_j, \sigma_k, r_{jk})d\boldsymbol{X}d\boldsymbol{X}' +\\ & \frac{1}{p^2}\sum_{j=1}^{p}\int\{t_d(\boldsymbol{X}) - \sigma_{jj}\}^2 f_{1n}(\boldsymbol{X} \mid s_{jj})d\boldsymbol{X}\\ =& \frac{p-1}{p}\int\int\{t_{od}(\boldsymbol{X}, \boldsymbol{X}') - ab\gamma\}^2 f_{2n}(\boldsymbol{X}, \boldsymbol{X}' \mid a, b, \gamma)dG_{od}(a, b, \gamma)d\boldsymbol{X}d\boldsymbol{X}' +\\ & \frac{1}{p}\int\int\{t_d(\boldsymbol{X}) - a^2\}^2 f_{1n}(\boldsymbol{X} \mid a)dG_d(a)d\boldsymbol{X}.\end{aligned}$$

Therefore for any t_{od},

$$\begin{aligned}&\int\{t_{od}(\boldsymbol{X}, \boldsymbol{X}') - ab\gamma\}^2 f_{2n}(\boldsymbol{X}, \boldsymbol{X}' \mid a, b, \gamma)dG_{od}(a, b, \gamma)\\ =& \int\{t_{od}(\boldsymbol{X}, \boldsymbol{X}') - t_{od}^\star(\boldsymbol{X}, \boldsymbol{X}')\}^2 f_{2n}(\boldsymbol{X}, \boldsymbol{X}' \mid a, b, \gamma)dG_{od}(a, b, \gamma) +\\ & 2\{t_{od}(\boldsymbol{X}, \boldsymbol{X}') - t_{od}^\star(\boldsymbol{X}, \boldsymbol{X}')\}\int\{t_{od}^\star(\boldsymbol{X}, \boldsymbol{X}') - ab\gamma\}f_{2n}(\boldsymbol{X}, \boldsymbol{X}' \mid a, b, \gamma)dG_{od}(a, b, \gamma) +\\ & \int\{t_{od}^\star(\boldsymbol{X}, \boldsymbol{X}') - ab\gamma\}^2 f_{2n}(\boldsymbol{X}, \boldsymbol{X}' \mid a, b, \gamma)dG_{od}(a, b, \gamma)\\ \geqslant& \int\{t_{od}^\star(\boldsymbol{X}, \boldsymbol{X}') - ab\gamma\}^2 f_{2n}(\boldsymbol{X}, \boldsymbol{X}' \mid a, b, \gamma)dG_{od}(a, b, \gamma),\end{aligned}$$

because the middle cross term is equal to zero by the definition of $t_{od}^\star$. Similarly, it can be shown that for any t_d,

$$\int\{t_d(\boldsymbol{X}) - a^2\}^2 f_{1n}(\boldsymbol{X} \mid a)dG_d(a) \geqslant \int\{t_d^\star(\boldsymbol{X}) - a^2\}^2 f_{1n}(\boldsymbol{X} \mid a)dG_d(a).$$

This implies that $R(\boldsymbol{\Sigma}, \boldsymbol{\delta}) \geqslant R(\boldsymbol{\Sigma}, \boldsymbol{\delta}^\star)$ for any $\boldsymbol{\delta} \in \mathcal{S}$.

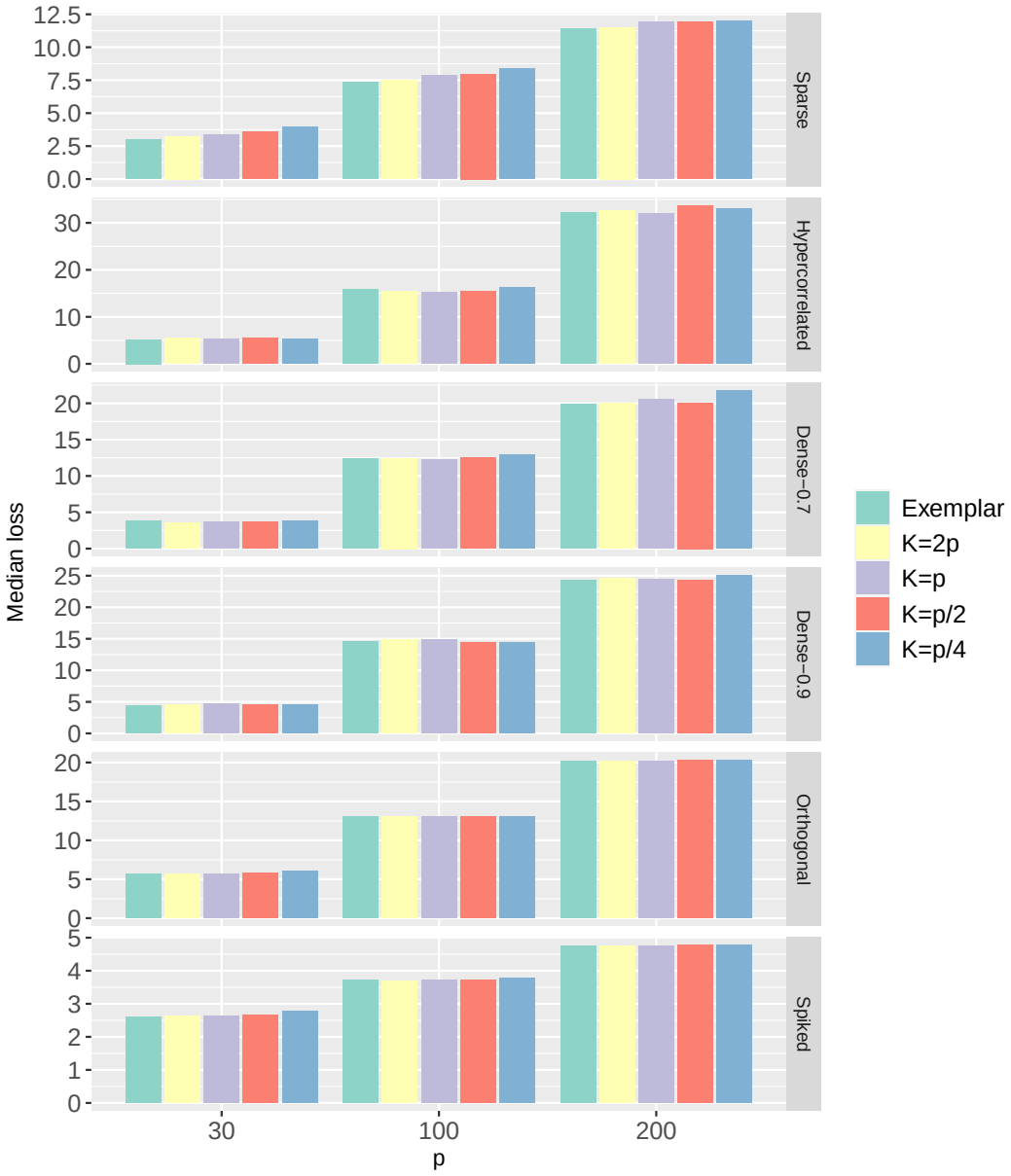

Figure 1. Median Frobenius norm errors over 200 replications for our proposed MSGCor. Exact numerical results and interquartile ranges are provided in the Supporting Information. Sparse: Model 1; Hypercorrelated: Model 2; Dense-0.7: Model 3; Dense-0.9: Model 4; Orthogonal: Model 5; Spiked: Model 6.

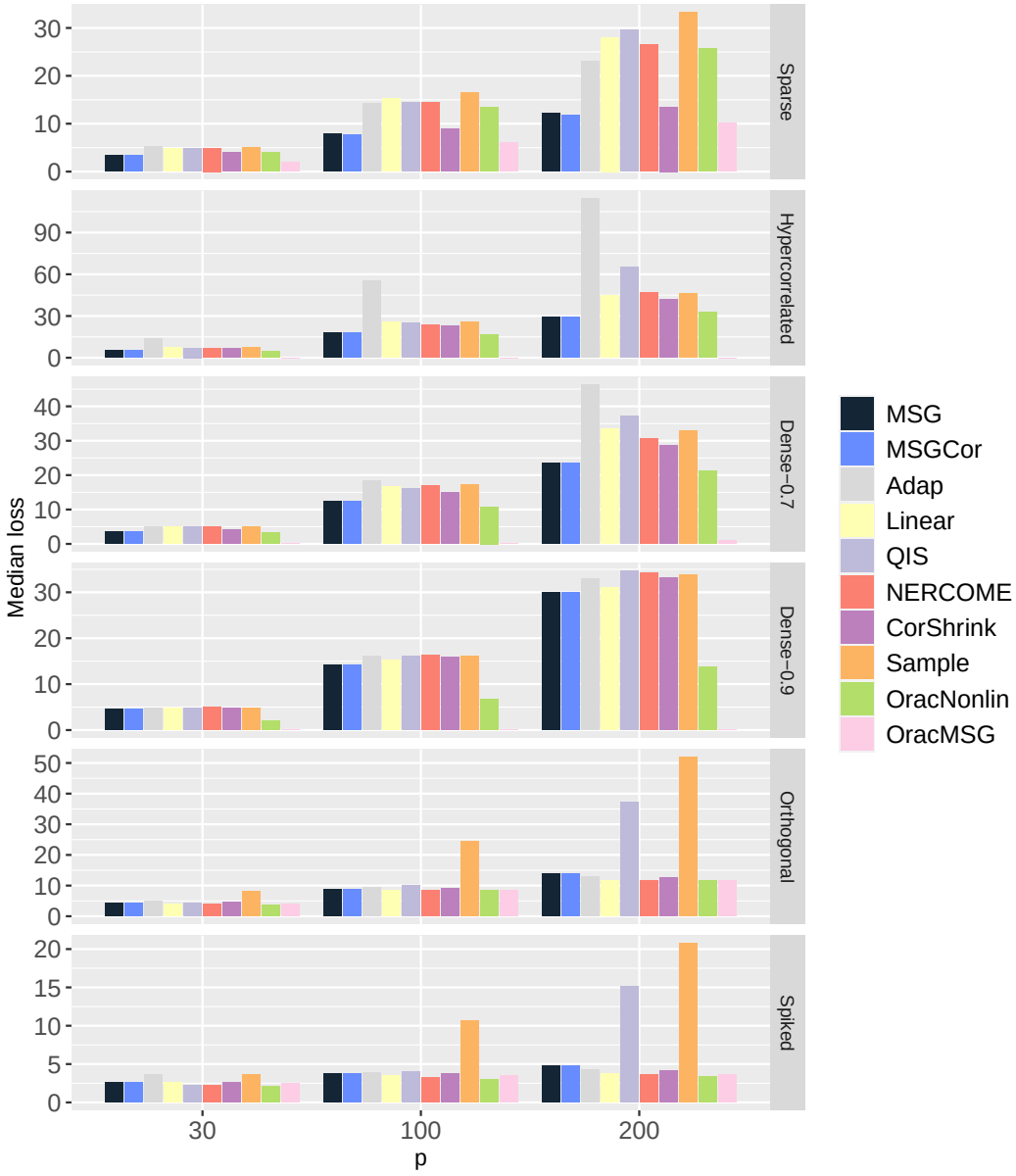

Figure 2. Median Frobenius norm errors over 200 replications. Exact numerical results and interquartile ranges are provided in the Supporting Information. MSG and MSGCor were implemented with $K = p$. Sparse: Model 1; Hypercorrelated: Model 2; Dense-0.7: Model 3; Dense-0.9: Model 4; Orthogonal: Model 5; Spiked: Model 6.

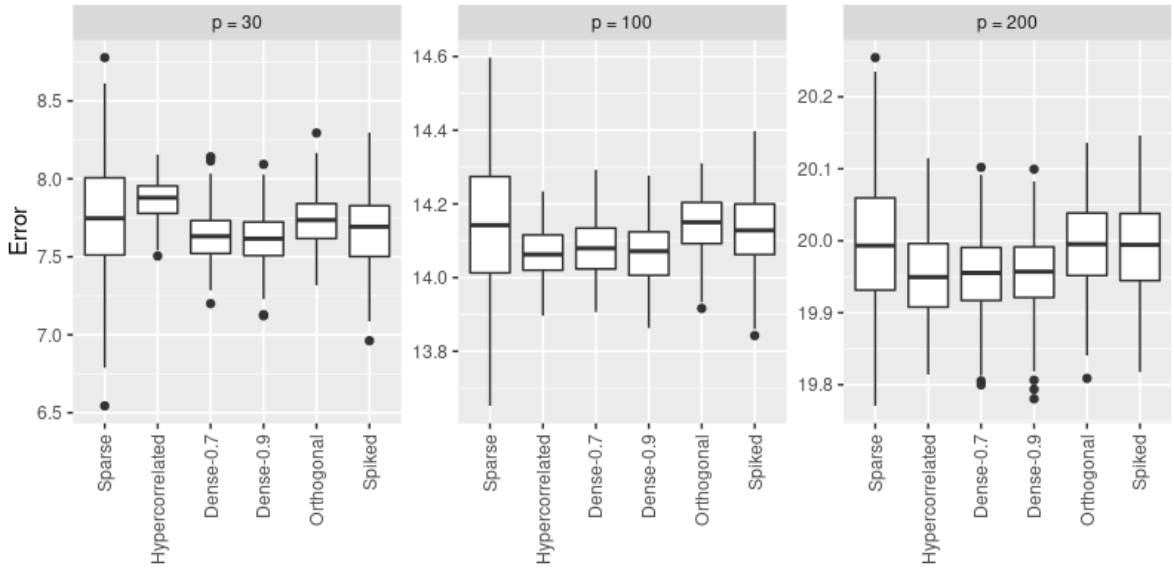

Figure 3. Estimation errors of sample eigenvectors over 200 replications. Sparse: Model 1; Hypercorrelated: Model 2; Dense-0.7: Model 3; Dense-0.9: Model 4; Orthogonal: Model 5; Spiked: Model 6.

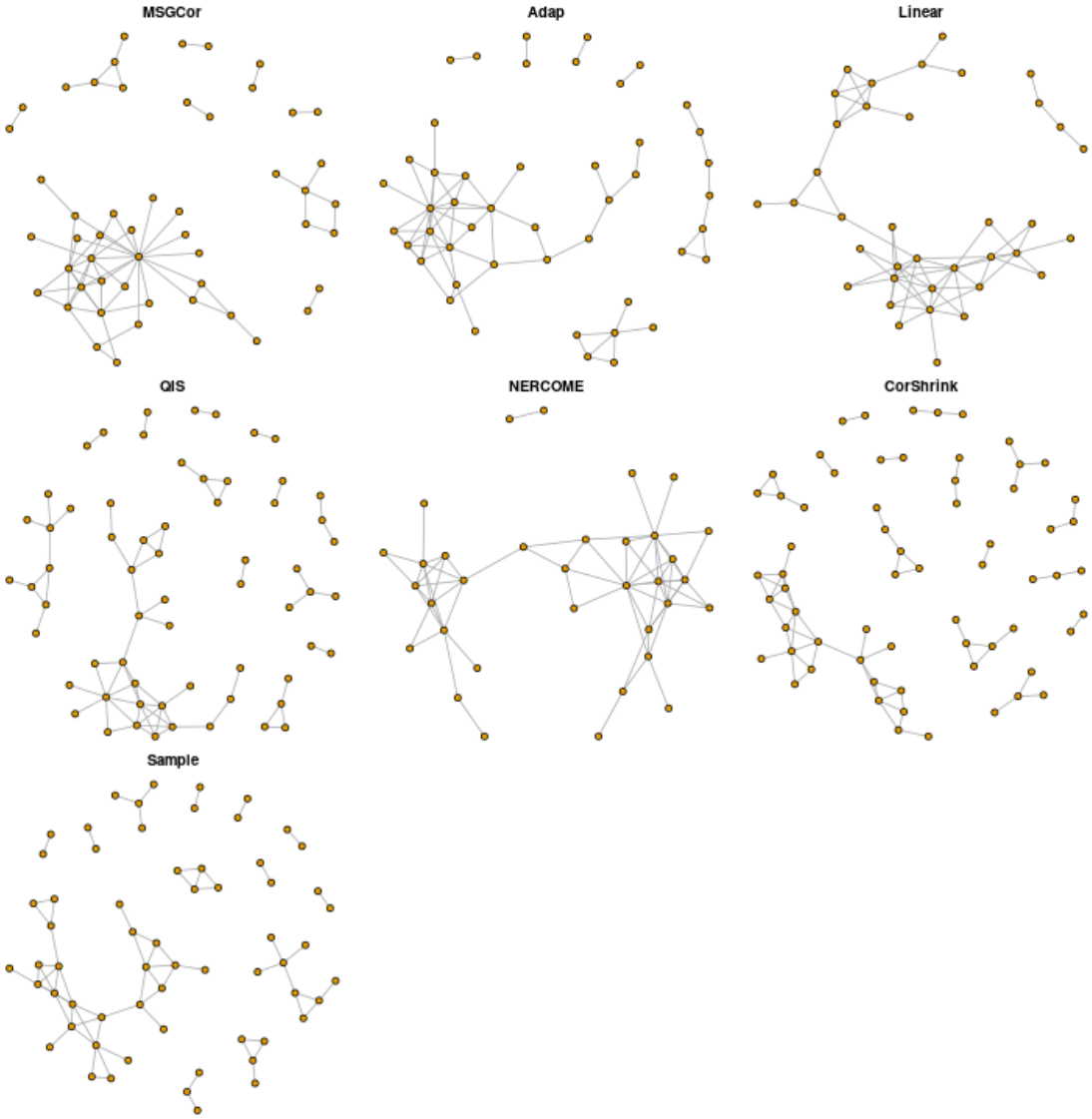

Figure 4. Amygdala gene networks recovered by the different covariance matrix estimation methods.

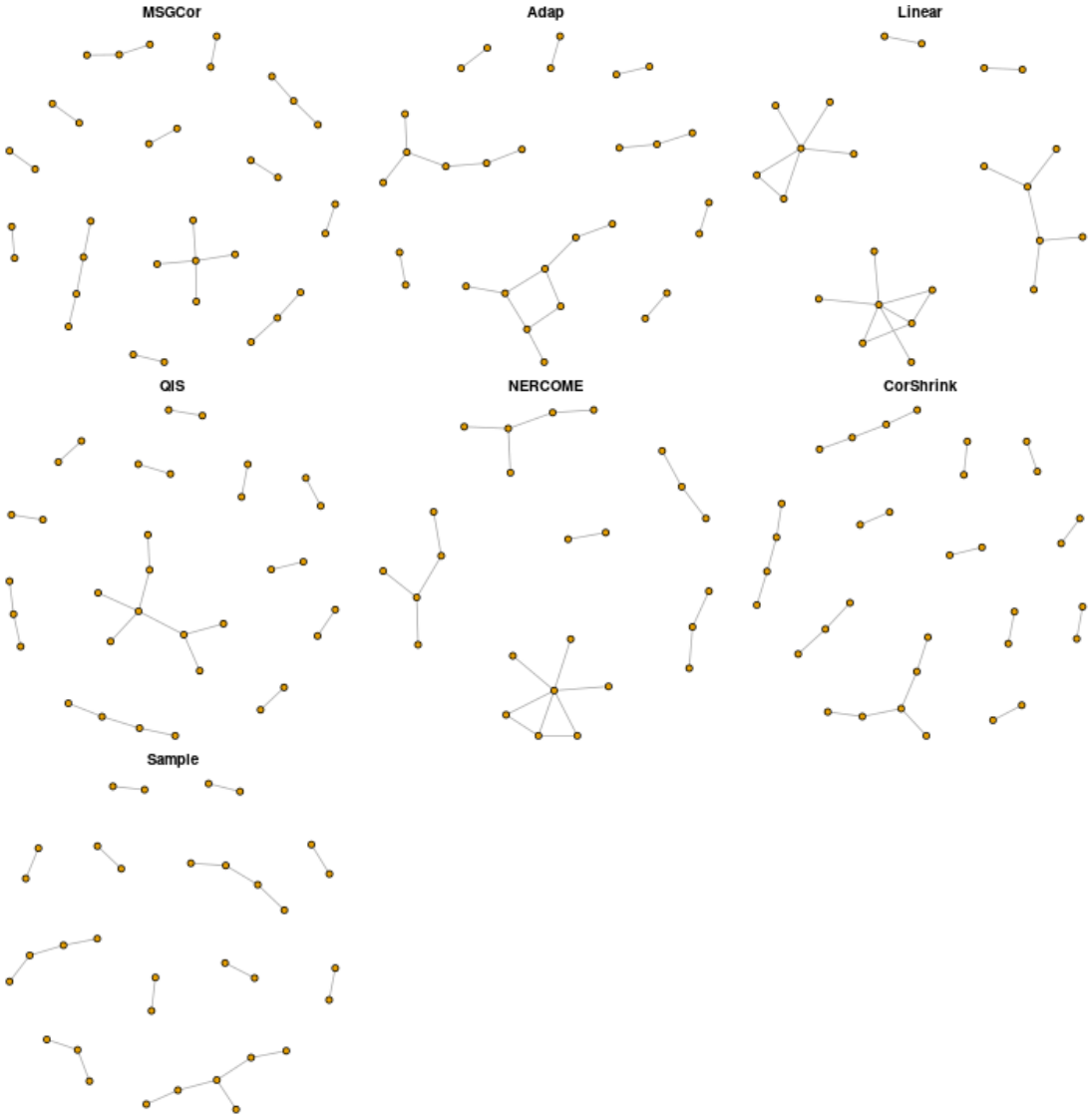

Figure 5. Frontal cortex gene networks recovered by the different covariance matrix estimation methods.

Table 1
Average running time of MSGCor under Model 1 for different K.

	p=30	p=100	p=200
Full example 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
Median gene expression covariance matrix estimation errors (25% and 75% quantiles in parentheses). Bold text highlights the smallest median errors in each column.

Brain region	Amygdala	Frontal cortex
MSGCor	**2.26 (1.98, 2.57)**	**2.3 (2.15, 2.55)**
Adap	2.6 (2.18, 2.97)	2.44 (2.18, 2.71)
Linear	2.3 (2.05, 2.55)	2.35 (2.19, 2.53)
QIS	2.61 (2.33, 2.86)	2.79 (2.64, 2.93)
NERCOME	2.37 (2.14, 2.62)	**2.3 (2.14, 2.53)**
CorShrink	**2.26 (2, 2.56)**	2.36 (2.2, 2.56)
Sample	2.61 (2.33, 2.85)	2.79 (2.64, 2.95)