<table>
<thead>
<tr>
<th><strong>Statistica Sinica Preprint No:</strong> SS-12-328R3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Manuscript ID</strong></td>
</tr>
<tr>
<td><strong>URL</strong></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
</tr>
</tbody>
</table>
| **Complete List of Authors** | Cheng Wang  
Guangming Pan  
Tiejun Tong and  
Lixing Zhu |
| **Corresponding Author** | Cheng Wang |
| **E-mail** | cescwang@gmail.com |
| **Notice:** Accepted version subject to English editing. |
Shrinkage estimation of large dimensional precision matrix using random matrix theory

Cheng Wang\textsuperscript{1,3}, Guangming Pan\textsuperscript{2}, Tiejun Tong\textsuperscript{3} and Lixing Zhu\textsuperscript{3}

\textsuperscript{1}Shanghai Jiao Tong University
\textsuperscript{2}Nanyang Technological University
\textsuperscript{3}Hong Kong Baptist University

Abstract: This paper considers the ridge-type shrinkage estimation of the large dimensional precision matrix. The asymptotic optimal shrinkage coefficients and the theoretical loss are derived. Data-driven estimators for the shrinkage coefficients are also conducted based on the asymptotic results from random matrix theory. The new method is distribution-free and no assumption on the structure of the covariance matrix or the precision matrix is required. The proposed method also applies to situations where the dimension is larger than the sample size. Finally, numerical studies using simulated data and real data demonstrate that the proposed estimator performs better than the existing competitors in a wide range of settings.

Key words and phrases: Large dimensional data, random matrix theory, ridge-type estimator, shrinkage estimation, precision matrix.

1. Introduction

In multivariate statistical analysis, one often needs to estimate the precision matrix, i.e., the inverse of the covariance matrix. The estimation of the precision matrix has applications in many statistical problems, e.g., in linear discriminant analysis (Anderson, 2003), in Hotelling’s $T^2$ test (Hotelling, 1931), and in Markowitz mean-variance analysis (Markowitz, 1952). Let $n$ be the sample size, $p$ be the dimension of observation, and $\Sigma_p$ be the covariance matrix. When $p$ is fixed and $n$ is large, the inverse of the sample covariance matrix, $S_n^{-1}$, is commonly used to estimate the precision matrix $\Omega_p = \Sigma_p^{-1}$. For large dimensional data, however, $p$ can be as large as or even larger than $n$. As a consequence, the sample covariance matrix $S_n$ is close to or even a singular matrix. This brings in new challenges to the estimation of the precision matrix. One remedy to this problem is to apply the Moore-Penrose inverse of $S_n$ (Srivastava, 2007; Kubokawa and Srivastava, 2008). Such an estimator may perform poorly in practice since some of the
eigenvalues are zero or close to zero.

Let \( X_1, \ldots, X_n \) be a random sample independently from a multivariate distribution (See, Bai and Saranadasa 1996 or Chen, Zhang, and Zhong 2010),

\[
X_i = \Sigma_p^{1/2} Y_i + \mu_0, \quad i = 1, \ldots, n, \tag{1.2}
\]

where \( \mu_0 \) is a \( p \)-dimensional mean vector and \( \Sigma_p \) is a \( p \times p \) positive definite covariance matrix. Here \( Y = (Y_1, \ldots, Y_n) = (Y_{ij})_{p \times n} \) and \( \{Y_{ij}, i, j = 1, 2, \ldots\} \) are independent and identically distributed (i.i.d.) random variables with mean zero and variance one. Let also the sample covariance matrix be

\[
S_n = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X})(X_j - \bar{X})^T, \tag{1.3}
\]

where \( \bar{X} = \frac{1}{n} \sum_{j=1}^{n} X_j/n \) is the sample mean and the superscript \( T \) denotes the transpose of a matrix or vector. Note that, if the data are Gaussian distributed, then for \( p < n \), \( (n - 1)S_n \) follows a Wishart distribution and \( S_n^{-1}/(n-1) \) follows an inverse Wishart distribution. In addition, by the fact that

\[
E(S_n^{-1}) = \frac{n-1}{n-p-2} \Sigma_p^{-1}, \tag{1.4}
\]

an unbiased estimator of the precision matrix \( \Omega_p \) is \( (n-p-2)S_n^{-1}/(n-1) \). In this paper, we are interested in estimating \( \Omega_p \) without the Gaussian assumption. More specifically, our estimation will be distribution-free by using random matrix theory (RMT).

In the literature, under certain model structures such as sparsity or ordering, penalized methods have been widely proposed and applied (see, e.g. Friedman, Hastie, and Tibshirani 2008, Yuan 2010, Cai, Liu, and Luo 2011). Instead, when such prior information about the structure of covariance matrix is not available, one often considers the shrinkage methods to improve the standard estimators since the seminal paper by James and Stein (1961). For instance, Stein (1975) proposed to shrink each eigenvalue of the sample covariance matrix based on Stein’s loss function. See also Dey and Srinivasan (1985), Daniels and Kass (1999, 2001), Mestre and Lagunas (2006), Konno (2009), among many others. In a recent work by Ledoit and Wolf (2012), the authors derived the optimal shrinkage coefficients for each eigenvalue and proposed a nonlinear estimator for the precision matrix which significantly improves the standard estimator. Nevertheless, we note that all these methods require that \( p \) is less than \( n \) and none of the eigenvalues is zero.
To overcome the singularity problem when \( p \geq n \), Ledoit and Wolf (2004) proposed a shrinkage estimator for \( \Sigma_p \), which is a linear combination of \( S_n \) and the identity matrix \( I_p \) with respect to a quadratic loss function. Other works include Schäfer and Strimmer (2005), Warton (2008), and Fisher and Sun (2011). In contrast, very little work has been done for estimating the precision matrix directly. To the best of our knowledge, the only available work in this direction is proposed by Kubokawa and Srivastava (2008), in which the authors considered the following ridge-type estimator for the precision matrix,

\[
\hat{\Omega}_p = \alpha (S_n + \beta I_p)^{-1},
\]

(1.5)

where \( \alpha \) and \( \beta \) are two shrinkage coefficients. Note that the derivation of the shrinkage coefficients \( \alpha \) and \( \beta \) in (1.5) can be more challenging than those in estimating \( \Sigma_p \), especially when \( p \geq n \). As a solution, Kubokawa and Srivastava (2008) assumed that the data to be Gaussian distributed. They then employed an empirical Bayes method for specifying the shrinkage coefficients and demonstrated that the resulting estimator dominates the usual estimator.

In this paper, we propose to derive the optimal shrinkage coefficients \( \alpha \) and \( \beta \) for non-Gaussian data under the following loss function (Haff, 1979; Krishnamoorthy and Gupta, 1989; Yang and Berger, 1994),

\[
\frac{1}{p} tr (\hat{\Omega}_p \Sigma_p - I_p)^2.
\]

(1.6)

To proceed with the paper, we first study the asymptotic properties of the matrix \( \Sigma_p^{1/2} (S_n + \lambda I_p)^{-1} \Sigma_p^{1/2} \) and its relation with \( (S_n + \lambda I_p)^{-1} \) by using RMT in Section 2. We then study the theoretical loss of the ridge-type estimator, derive the optimal shrinkage coefficients \( \alpha \) and \( \beta \), and develop a data-driven shrinkage estimator for the precision matrix in Section 3. In Section 4, we conduct numerical studies with simulated data and real data to evaluate the performance of the proposed estimator and compare it with some existing methods. Finally, we conclude the paper in Section 5 and the technical proofs are provided in an online supplement.

2. Preliminary Results in RMT

We first introduce some notations in RMT. Suppose \( A_m \) is an \( m \times m \) Hermitian matrix with eigenvalues \( \lambda_j, j = 1, \ldots, m \). We define the empirical spectral distribution
(ESD) of the matrix $A_m$ as

$$F^A_m(x) = \frac{1}{m} \sum_{j=1}^{m} I(\lambda_j \leq x),$$

where $I(\cdot)$ denotes the indicator function. ESD plays an important role in multivariate analysis and many statistics can be expressed as the functionals of ESD, e.g., $\det(A_m) = \exp(m \int \log(x) dF^A_m(x))$ and $\text{tr}(A_m) = m \int x dF^A_m(x)$. For more details, see for example the recent monograph by Bai and Silverstein (2010) and the references therein. The limit distribution of $F^A_m$, if it exists and is non-random, is called the limiting spectral distribution (LSD) of the sequence $\{A_m\}$.

In RMT, the Stieltjes transform of $F$ is defined by

$$m_F(z) = \int \frac{1}{t-z} dF(t), \quad z \in \mathbb{C}^+ \equiv \{z \in \mathbb{C} : \text{Im}z > 0\},$$

and the inversion formula is

$$F\{[c,d]\} = \lim_{\eta \to 0^+} \frac{1}{\pi} \int_c^d \text{Im}F(\xi + i\eta) d\xi,$$

where $c < d$ are continuity points of $F$. By (2.4), $F$ is uniquely determined by its Stieltjes transform.

In this paper, we need the following two conditions:

(S1) Both $p$ and $n$ tend to infinity in such a way that $p/n \to y \in (0, \infty)$. Also, the fourth order moment of $Y_{ij}$ is bounded.

(S2) There exists constants $c_1$ and $c_2$ such that $c_1 \leq \lambda_{\min}(\Sigma_p) \leq \lambda_{\max}(\Sigma_p) \leq c_2$. That is, the extreme eigenvalues of $\Sigma_p$ are uniformly bounded. Further, $F^{\Sigma_p}$ tends to a non-random probability distribution $H$ as $p$ tends to infinity.

**Theorem 1** Assume that conditions (S1) and (S2) hold. As $n \to \infty$, we have $F^{\Sigma_p^{-1/2}(S_n + \lambda I_p)\Sigma_p^{-1/2}}$ converges almost surely to a non-random distribution $F$, whose Stieltjes transform $m(z)$ satisfies

$$m(z) = \int \frac{\lambda}{t-z + \frac{1}{1+ym(z)}} dH(t),$$

where $\lambda > 0$ and $z \in \mathbb{C}^+$. 
A ridge-type shrinkage estimator for large precision matrix

Note that Theorem 1 can also be derived from Theorem 1.2 in Ledoit and Pêché (2011) where the 12th moment is needed. By Silverstein (1995), the Stieltjes transform $m_0(z)$ of LSD of $S_n$ is the solution of the following equation,

$$m_0(z) = \int \frac{dH(t)}{t(1 - y - yzm_0(z)) - z}. \quad (2.6)$$

In addition, we have the following lemma.

**Lemma 1** For any $\lambda > 0$, $m_0(-\lambda)$ is the unique solution of the equation

$$m(-\lambda) = \int \frac{dH(t)}{t(1 - y + y\lambda m(-\lambda)) + \lambda}, \quad (2.7)$$

where $1 - y + y\lambda m(-\lambda) \geq 0$.

Note that the condition $1 - y + y\lambda m(-\lambda) \geq 0$ in Lemma 1 is necessary and can be regarded as a variant of the condition in Silverstein and Choi (1995). Here we use an example to illustrate this claim. Assuming $\Sigma_p = I_p$, (2.7) has the following two solutions:

$$m^{(1)}(-\lambda) = \frac{1}{2y\lambda}[-(1 - y + \lambda) + \sqrt{(1 - y + \lambda)^2 + 4y\lambda}],$$

$$m^{(2)}(-\lambda) = \frac{1}{2y\lambda}[-(1 - y + \lambda) - \sqrt{(1 - y + \lambda)^2 + 4y\lambda}],$$

whereas $1 - y + y\lambda m^{(1)}(-\lambda) > 0$ and $1 - y + y\lambda m^{(2)}(-\lambda) < 0$. This is the reason why Chen et al. (2011) claimed $m_0(-\lambda) = m^{(1)}(-\lambda)$, not $m_0(-\lambda) = m^{(2)}(-\lambda)$.

**Theorem 2** Assume that conditions (S1) and (S2) hold. For any $\lambda > 0$, as $n \to \infty$ we have

$$\frac{1}{p} tr(\Sigma_p^{1/2}(S_n + \lambda I_p)^{-1}\Sigma_p^{1/2}) \xrightarrow{a.s.} R_1(\lambda),$$

$$\frac{1}{p} tr(\Sigma_p^{1/2}(S_n + \lambda I_p)^{-1}\Sigma_p^{1/2})^2 \xrightarrow{a.s.} R_2(\lambda),$$

where $\xrightarrow{a.s.}$ represents convergence almost surely and

$$R_1(\lambda) = \frac{1 - \lambda m_0(-\lambda)}{1 - y(1 - \lambda m_0(-\lambda))},$$

$$R_2(\lambda) = \frac{1 - \lambda m_0(-\lambda)}{(1 - y(1 - \lambda m_0(-\lambda)))^2} - \frac{\lambda m_0(-\lambda) - \lambda^2 m_0'(-\lambda)}{(1 - y(1 - \lambda m_0(-\lambda)))^3}. $$
In addition, we have
\[
\frac{1}{p} \text{tr}((S_n + \lambda I_p)^{-1}) \xrightarrow{a.s.} m_0(-\lambda),
\]
\[
\frac{1}{p} \text{tr}((S_n + \lambda I_p)^{-2}) \xrightarrow{a.s.} m_0'(-\lambda) = \frac{dm_0(z)}{dz} \bigg|_{z = -\lambda}.
\]

In Section 3, we will use Theorem 2 to construct new estimators for the precision matrix \( \Omega_p \). Note also that Chen et al. (2011) proposed some similar results as those in Theorem 2, under the assumption that the data are Gaussian distributed. In this paper, we have relaxed their conditions by removing the Gaussian assumption.

3. Shrinkage Estimation of Precision Matrix

We consider the following ridge-type estimator for the precision matrix,
\[
\hat{\Omega}_p = \alpha(S_n + \beta I_p)^{-1},
\]
where \( \alpha > 0 \) and \( \beta > 0 \) are two shrinkage coefficients. By Theorem 2, we have
\[
\frac{1}{p} \text{tr}(\Sigma_p \hat{\Omega} - I_p)^2 \xrightarrow{a.s.} \alpha^2 R_2(\beta) - 2\alpha R_1(\beta) + 1
\]
\[
= R_2(\beta)(\alpha - \frac{R_1(\beta)}{R_2(\beta)})^2 + 1 - \frac{(R_1(\beta))^2}{R_2(\beta)}. \tag{3.5}
\]
By (3.5), the optimal \( \alpha \) is \( \alpha_{opt} = R_1(\beta)/R_2(\beta) \) for any fixed \( \beta \). This leads to the simplified loss function as
\[
L(\beta) = 1 - \frac{(R_1(\beta))^2}{R_2(\beta)}. \tag{3.6}
\]
Let \( L_0 = \min_{\beta > 0} L(\beta) \) be the minimum loss and \( \beta_{opt} = \arg \min_{\beta > 0} L(\beta) \) be the optimal parameter of \( \beta \).

Theorem 3 For any \( y < 1 \), we have \( L_0 = \min_{\gamma > 0} L_H(\gamma) \), where
\[
L_H(\gamma) = 1 - \left( \int \left( \frac{t}{t + \gamma} dH(t) \right)^2 \left( \frac{1}{\int \frac{t^2}{t + \gamma^2} dH(t)} - y \right) \right), \quad \gamma \geq 0.
\]
Furthermore, we have the following results:

1. When \( H(x) \) is a degenerate distribution at \( \sigma^2 \) (i.e. \( \Sigma_p = \sigma^2 I_p \)), the minimum loss is \( L_0 = 0 \).
II. For a general distribution \( H(x) \), \( L_H(\gamma) \) achieves its global minimum value \( L_0 \) at \( \gamma^* \) satisfying

\[
\frac{f_1(\gamma^*)f_3(\gamma^*) - f_2(\gamma^*)f_2(\gamma^*)}{f_2(\gamma^*)f_2(\gamma^*)f_1(\gamma^*) - f_2(\gamma^*)f_2(\gamma^*)} = y,
\]

(3.7)

where \( f_k(x) = \int \frac{1}{(1 + \lambda x)^k} dH(t) \). Correspondingly, \( L(\beta_{opt}) = L_0 \) where \( \beta_{opt} \) satisfies

\[
\gamma^* = \frac{\beta_{opt}}{1 - y(1 - \beta_{opt}n0(-\beta_{opt})).}
\]

Note that for simplicity, we have assumed that \( y < 1 \) in Theorem 3. A similar result, however, can be obtained for \( y \geq 1 \) as well. When \( H(x) \) is not a degenerate distribution, from the proof it is known that the optimal parameter \( \beta_{opt} \) will be located in a bounded interval \([C_1, C_2]\) where \( 0 < C_1 < C_2 < \infty \).

Note that \( \alpha_{opt} \) and \( \beta_{opt} \) are unknown and need to be estimated. In this paper, we consider a data-driven method for estimating \( \alpha_{opt} \) and \( \beta_{opt} \). Let

\[
\hat{R}_1(\lambda) = \frac{a_1(\lambda)}{1 - \tilde{y}a_1(\lambda)},
\]

\[
\hat{R}_2(\lambda) = \frac{a_1(\lambda)}{(1 - \tilde{y}a_1(\lambda))^3} - \frac{a_2(\lambda)}{(1 - \tilde{y}a_1(\lambda))^4},
\]

where \( \tilde{y} = p/n, \ a_1(\lambda) = 1 - \frac{1}{p}tr(\frac{1}{\lambda}S_n + I_p)^{-1}, \ a_2(\lambda) = \frac{1}{p}tr(\frac{1}{\lambda}S_n + I_p)^{-1} - \frac{1}{p}tr(\frac{1}{\lambda}S_n + I_p)^{-2}. \) Let also the empirical loss function of \( L(\lambda) \) be

\[
L_n(\lambda) = 1 - (\hat{R}_1(\lambda))^2/\hat{R}_2(\lambda).
\]

In addition, we define \( \beta_n^* = \arg \min_{\beta \in [C_1, C_2]} L_n(\beta) \) and \( \alpha_n^* = \hat{R}_1(\beta_n^*)/\hat{R}_2(\beta_n^*) \). In case that \( \beta_n^* \) may not be unique, we take the smallest solution throughout the paper.

**Theorem 4** Assume that conditions (S1) and (S2) hold. Then for any \( \lambda > 0 \),

\[
L_n(\lambda) \xrightarrow{a.s.} L(\lambda) \quad \text{as} \quad n \to \infty.
\]

In addition, we have

\[
\frac{1}{p}tr(\alpha_n^*(S_n + \beta_n^*I_p)^{-1} - I_p)^2 \xrightarrow{a.s.} L_0 \quad \text{as} \quad n \to \infty.
\]

(3.8)

Theorem 4 shows that the proposed estimator can achieve the minimum loss \( L_0 \) asymptotically. In view of this, we propose the new estimator of \( \Omega_p \) as

\[
\hat{\Omega}_p^* = \alpha_n^*(S_n + \beta_n^*I_p)^{-1}.
\]

(3.9)
For $\hat{\Omega}_p^*$, when $y < 1$, from the proof of Theorem 3 we can show that $L_0 < L_H(0) = y$. Then by noting that

$$\frac{1}{p} tr\left(\frac{n-p-2}{n-1} \Sigma_p S_n^{-1} - I_p\right)^2 \overset{a.s.}{\to} \frac{y}{1-y}.$$  \hfill (3.10)

the new estimator $\hat{\Omega}_p^*$ performs asymptotically better than the classical estimator $S_n^{-1}$ and also the unbiased estimator $(n-p-2)/(n-1)S_n^{-1}$. In addition, we note that the new estimator also applies to $y \geq 1$, in such situations where the estimators based on $S_n^{-1}$ or the non-zero eigenvalues of $S_n$ (Srivastava, 2007; Ledoit and Wolf, 2012; Bodnar et al., 2013) are no longer applicable.

4. Numerical Studies

4.1. Monte Carlo Simulation Study

Note that the components of data such as gene expression data may have different scales. As in Warton (2008), we propose a two-stage procedure to implement the new estimator. Specifically, we first normalize the data to eliminate the effect of different scales. By doing so, we are actually handling the sample correlation matrix $R_n$ and the proposed inverse correlation matrix estimator is $\hat{R}_n^{-1} = \alpha_n^* (R_n + \beta_n^* I_p)^{-1}$. We then use $\text{diag}(S_n)$ to rescale the inverse correlation matrix and estimate the precision matrix $\Omega_p$ by

$$\hat{\Omega}_{\text{New}} = (\text{diag}(S_n))^{-1/2} \hat{R}_n^{-1} (\text{diag}(S_n))^{-1/2}.$$  

In this section, we compare the new estimator with the following estimators in the literature:

- The scaled standard estimator (referred to as the SSE estimator)

$$\hat{\Omega}_{\text{SSE}} = \frac{n-p-2}{n-1} S_n^{-1} I(p < n) + \frac{p}{n-1} S_n^+ I(p \geq n),$$  \hfill (4.5)

where $S_n^+$ is the Moore-Penrose inverse of $S_n$. This estimator covers several methods including Stein (1975), Mestre and Lagunas (2006), Srivastava (2007) and Kubokawa and Srivastava (2008).

- The estimator in Efron and Morris (1976) (referred to as the EM estimator)

$$\hat{\Omega}_{\text{EM}} = \frac{n-p-2}{n-1} S_n^{-1} + \frac{p^2 + p-2}{(n-1) tr(S_n)} I_p.$$  \hfill (4.6)
A ridge-type shrinkage estimator for large precision matrix

- The empirical Bayes ridge-type estimator from Kubokawa and Srivastava (2008) (referred to as the KS estimator)

\[
\hat{\Omega}_{\text{KS}} = p((n-1)S_n + \text{tr}(S_n)I_p)^{-1}.
\]  

(4.7)

In addition, for illustration purposes, we have also included a recent shrinkage estimator that is designed for estimating the covariance matrix. Specifically, Fisher and Sun (2011) consider a combination between \(S_n\) and \(\text{diag}(S_n)\) to estimate \(\Sigma_p\), i.e., \(\hat{\Sigma}_{\text{FS}} = \hat{\lambda}S_n + (1 - \hat{\lambda})\text{diag}(S_n)\), where \(\hat{\lambda}\) is derived in Fisher and Sun (2011). We then estimate \(\Omega_p\) by

\[
\hat{\Omega}_{\text{FS}} = (\hat{\lambda}S_n + (1 - \hat{\lambda})\text{diag}(S_n))^{-1}.
\]  

(4.8)

We refer to it as the FS estimator.

To conduct simulation studies in a wide range of settings, we consider the following four models for generating the covariance matrix:

- (M1) \(\Sigma_1\) is diagonal with 20% of population eigenvalues being equal to 1, 40% equal to 3 and the other 40% equal to 10.
- (M2) \(\Sigma_2 = \Sigma_1^{1/2}\Sigma_0\Sigma_1^{1/2}\) where \(\Sigma_0 = (\sigma_{ij})_{p \times p}\) and \(\sigma_{ij} = 0.5|j-i|\) for \(1 \leq i, j \leq p\).
- (M3) \(\Sigma_3 = \Sigma_1^{1/2}\Sigma_{00}\Sigma_1^{1/2}\) where \(\Sigma_{00} = (\sigma_{ij})_{p \times p}\) and \(\sigma_{ij} = I(i = j) + 0.2I(i \neq j)\).
- (M4) \(\Sigma_4 = U_2\text{diag}(\lambda_1, \ldots, \lambda_p)U_2^T\) where \(\lambda_j = 2 + 0.125j, j = 1, \ldots, p\) and the rows of \(U_2\) are eigenvectors of \(\Sigma_0\).

Model 1 is a diagonal spiked example which was well studied in RMT (See, Bai and Silverstein 1998, Ledoit and Wolf 2012). Model 2 is an example of sparse matrix where the values of the entries decay as they move away from the diagonal. Model 3 serves as a dense matrix example and Model 4 is an example with many distinct eigenvalues. In addition, with respect to the random part \(Y = (Y_{ij})_{p \times n}\), we consider also four different distributions:

- Normal distribution: \(Y_{ij} \sim N(0, 1)\).
- Mixture normal distribution: \(Y_{ij} \sim 0.5N(0, 1) + 0.5N(1, 1)\).
- Student’s \(t\)-distribution: \(Y_{ij} \sim t(5)\).
• Log-normal distribution: \( \log(Y_{ij}) \sim N(0.5, 0.5^2) \).

In each scenario, \( Y_{ij} \) are further standardized to have unit variance.

With 100 simulations for each simulation setting, we report in Table 4.1 the average losses of (1.6) for the new estimator and the competitors, where the data are simulated from the normal distribution and the mixture normal distribution, respectively. Seven different combinations of \((p, n)\) are considered, with three combinations of \( p < n \), one combination of \( p = n \), and three combinations of \( p > n \). In particular, the combination of \((p, n) = (1000, 100)\) represents the popular setting of high-dimensional low-sample-size data. Note that the EM estimator is excluded in the last four combinations since \( S_n \) is singular when \( p \geq n \). From the results in Table 4.1, we observe that the new estimator \( \hat{\Omega}_{\text{New}} \) always outperforms the existing competitors, no matter whether \( p \) is less than \( n \) or not. There is no much difference in terms of which covariance matrix is used. In addition, when \( p \geq n \), the shrinkage estimators are always better than the SSE estimator. This demonstrates that the Moore-Penrose inverse \( S_n^+ \) does not perform well in large dimensional data.

Recall that the new estimator is distribution free and it does not require a Gaussian distribution. To investigate the practical performance under other distributions, we have conducted another simulation study where the data are simulated from Student’s \( t \)-distribution and the log-normal distribution, respectively. All other settings are kept the same as before. With 100 simulations for each setting, we report in Table 4.2 the average losses for \((p, n) = (100, 200)\) and \((200, 100)\). Together with the results in Table 4.1, we observe that the performance of the new estimator is only slightly affected by the violation of normality assumption. Meanwhile, the comparison results among the estimators remain similar for all the distribution considered. In addition, we have also conducted simulations for other combinations of \((p, n)\) and the comparative conclusions remain the same.

Our third simulation study is to investigate how \( p \) and \( n \) affect the performance of the new estimator. Since the performance of the estimators is quite consistent for different distributions and different covariance matrices, we consider here only the Gaussian data with covariance matrix from Model 2. Then for the ratio \( p/n \) being either 1/2 or 2, we plot the average losses in Figure 4.1, where the results for the KS and FS estimators are also included for comparison. We observe that the loss of the new estimator reduces quickly to the minimum loss \( L_0 \) when \( p \) or \( n \) is large. For instance, when \( n > 100 \), the
relative error between the average loss and the minimum loss is always less than 10%. Whereas for the other two estimators, the relative errors can be several times as large as the minimum loss. This shows that the proposed new estimator can be reliable for large dimensional data, no matter whether $p$ is less than $n$ or not.

![Graph](image)

Figure 4.1: Average losses of the new estimator and the competitors along with different combinations of $p$ and $n$, where the data are Gaussian distributed and the covariance matrix is generated using Model 2. Note that the minimum losses $L_0$ are also reported for comparison.

### 4.2. Real Data Analysis

Note that the shrinkage estimators of the covariance matrix are commonly applied to the linear discriminant analysis. See, for example, Friedman (1989), Srivastava and Kubokawa (2007), Kubokawa and Srivastava (2008), Fan et al. (2009), Cai et al. (2011), Fisher and Sun (2011), and among others. In this section, we illustrate the usefulness of the proposed shrinkage estimator using the Leukemia data in Golub et al. (1999)
Table 4.1: Empirical risks of the proposed estimator and existing estimators for normal and mixture normal distributions.

<table>
<thead>
<tr>
<th></th>
<th>Method</th>
<th>$(p, n)$</th>
<th>(N(0, 1))</th>
<th>(0.5N(0, 1) + 0.5N(1, 1))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>M1</td>
<td>M2</td>
</tr>
<tr>
<td>((50, 100))</td>
<td>SSE</td>
<td>1.0849</td>
<td>1.0905</td>
<td>1.0701</td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>1.4617</td>
<td>1.7308</td>
<td>1.9892</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.4641</td>
<td>0.5250</td>
<td>0.4996</td>
</tr>
<tr>
<td></td>
<td>FS</td>
<td>0.0240</td>
<td>0.3721</td>
<td>0.6362</td>
</tr>
<tr>
<td></td>
<td>New</td>
<td>0.0238</td>
<td>0.2423</td>
<td>0.1268</td>
</tr>
<tr>
<td>((100, 200))</td>
<td>SSE</td>
<td>1.0478</td>
<td>1.0516</td>
<td>1.0363</td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>1.4127</td>
<td>1.6660</td>
<td>2.4586</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.4654</td>
<td>0.5265</td>
<td>0.5019</td>
</tr>
<tr>
<td></td>
<td>FS</td>
<td>0.0111</td>
<td>0.3497</td>
<td>1.2435</td>
</tr>
<tr>
<td></td>
<td>New</td>
<td>0.0109</td>
<td>0.2343</td>
<td>0.0975</td>
</tr>
<tr>
<td>((100, 1000))</td>
<td>SSE</td>
<td>0.1135</td>
<td>0.1132</td>
<td>0.1126</td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>0.1290</td>
<td>0.1383</td>
<td>0.1682</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.8259</td>
<td>0.8378</td>
<td>0.8318</td>
</tr>
<tr>
<td></td>
<td>FS</td>
<td>0.0021</td>
<td>0.0942</td>
<td>0.1354</td>
</tr>
<tr>
<td></td>
<td>New</td>
<td>0.0021</td>
<td>0.0811</td>
<td>0.0343</td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>0.3516</td>
<td>0.5088</td>
<td>0.3969</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.0527</td>
<td>0.7336</td>
<td>1.5249</td>
</tr>
<tr>
<td></td>
<td>FS</td>
<td>0.0483</td>
<td>0.3780</td>
<td>0.2251</td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>0.3503</td>
<td>0.5028</td>
<td>0.3729</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.0225</td>
<td>0.6534</td>
<td>5.0668</td>
</tr>
<tr>
<td></td>
<td>FS</td>
<td>0.0189</td>
<td>0.3597</td>
<td>0.1597</td>
</tr>
<tr>
<td>((1000, 100))</td>
<td>SSE</td>
<td>0.9800</td>
<td>1.2232</td>
<td>1.0954</td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>0.4531</td>
<td>1.0757</td>
<td>0.6274</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.0235</td>
<td>0.7302</td>
<td>8.9487</td>
</tr>
<tr>
<td></td>
<td>FS</td>
<td>0.0251</td>
<td>0.4039</td>
<td>0.3228</td>
</tr>
<tr>
<td>((100, 100))</td>
<td>SSE</td>
<td>1.09e7</td>
<td>9.96e7</td>
<td>1.63e8</td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>0.3529</td>
<td>0.4348</td>
<td>0.3924</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.0243</td>
<td>0.5349</td>
<td>2.0889</td>
</tr>
<tr>
<td></td>
<td>FS</td>
<td>0.0237</td>
<td>0.3130</td>
<td>0.1409</td>
</tr>
</tbody>
</table>
Table 4.2: Empirical risks of the proposed estimator and existing estimators for Student’s $t$ and Log-normal distributions.

<table>
<thead>
<tr>
<th>$(p, n)$</th>
<th>Method</th>
<th>$t(5)$</th>
<th>$\ln N(0.5,0.5^2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M1</td>
<td>M2</td>
</tr>
<tr>
<td>(100, 200)</td>
<td>SSE</td>
<td>1.1349</td>
<td>1.1395</td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>1.5246</td>
<td>1.7657</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.4627</td>
<td>0.5265</td>
</tr>
<tr>
<td></td>
<td>FS</td>
<td>0.0307</td>
<td>0.3845</td>
</tr>
<tr>
<td></td>
<td>New</td>
<td>0.0306</td>
<td>0.2435</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.3530</td>
<td>0.5064</td>
</tr>
<tr>
<td></td>
<td>FS</td>
<td>0.0619</td>
<td>0.7394</td>
</tr>
<tr>
<td></td>
<td>New</td>
<td>0.0585</td>
<td>0.3727</td>
</tr>
</tbody>
</table>

and the breast cancer data in Hess et al. (2006). The Leukemia data contains a total of 7129 genes for 47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML). The breast cancer data has 22283 genes for 133 patients who may achieve pathologic Complete Response (pCR). Among the 133 patients, 34 of them achieved pCR, whereas the other 99 did not achieve pCR.

We apply the proposed method to the linear discriminant analysis (LDA) and consider five discriminant methods with their discriminant scores as follows:

- **LDA$_{SSE}$**: $d_i = (x_0 - \bar{x}_i)^T \hat{\Omega}_{SSE}(x_0 - \bar{x}_i)$ for $i = 1, 2$.
- **LDA$_{EM}$**: $d_i = (x_0 - \bar{x}_i)^T \hat{\Omega}_{EM}(x_0 - \bar{x}_i)$ for $i = 1, 2$.
- **LDA$_{KS}$**: $d_i = (x_0 - \bar{x}_i)^T \hat{\Omega}_{KS}(x_0 - \bar{x}_i)$ for $i = 1, 2$.
- **LDA$_{FS}$**: $d_i = (x_0 - \bar{x}_i)^T (\hat{\Sigma}_{FS})^{-1}(x_0 - \bar{x}_i)$ for $i = 1, 2$.
- **LDA$_{New}$**: $d_i = (x_0 - \bar{x}_i)^T \hat{\Omega}_{New}(x_0 - \bar{x}_i)$ for $i = 1, 2$.

Here, $x_0$ is the new observation for classification and $\bar{x}_i$ are the sample means of group $i$, respectively. The discriminant rules for the above methods are to classify $x_0$ to group 1 if $d_1 < d_2$, and to group 2 otherwise. For a more comprehensive comparison, we have also included three widely used classifiers in the literature: the diagonal linear discriminant analysis (DLDA) in Dudoit et al. (2002) or Bickel and Levina (2004), the
Table 4.3: The average misclassification rates (%) of the new discriminant method and other methods for the Leukemia data and the breast cancer data, respectively. The standard deviations are also given in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>Leukemia data</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p = 25</td>
<td>p = 50</td>
<td>p = 100</td>
<td>p = 500</td>
<td>p = 1000</td>
</tr>
<tr>
<td></td>
<td>DLDA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NSC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HCT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LDA_{SSE}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LDA_{SSE}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>29.824(10.301)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>LDA_{EM}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LDA_{FS}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LDA_{New}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Breast cancer data</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p = 25</td>
<td>p = 50</td>
<td>p = 100</td>
<td>p = 500</td>
<td>p = 1000</td>
</tr>
<tr>
<td></td>
<td>DLDA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>36.704(7.542)</td>
<td>34.867(7.371)</td>
<td>33.207(7.546)</td>
<td>31.657(7.003)</td>
<td>31.252(7.576)</td>
</tr>
<tr>
<td></td>
<td>NSC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>37.518(7.935)</td>
<td>35.637(7.669)</td>
<td>33.554(7.448)</td>
<td>31.391(6.680)</td>
<td>30.485(6.679)</td>
</tr>
<tr>
<td></td>
<td>HCT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>36.813(7.741)</td>
<td>34.127(7.249)</td>
<td>31.655(6.990)</td>
<td>28.851(6.172)</td>
<td>27.924(6.751)</td>
</tr>
<tr>
<td></td>
<td>LDA_{SSE}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LDA_{EM}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>36.034(7.299)</td>
<td>37.161(8.234)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>LDA_{KS}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>35.669(7.166)</td>
<td>33.715(6.726)</td>
<td>31.418(5.999)</td>
<td>28.809(5.241)</td>
<td>27.660(5.026)</td>
</tr>
<tr>
<td></td>
<td>LDA_{FS}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LDA_{New}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

nearest shrunken centroids (NSC) method in Tibshirani et al. (2003), and the higher criticism thresholding (HCT) method in Donoho and Jin (2008).

To assess the misclassification rates, for each data set we first randomly select $p = 25$, 50, 100, 500 or 1000 genes, and randomly divide the total samples into two distinct sets, one for the training set and the other one for the test set. Specifically, we fix the training set sizes being 23 ALL and 12 AML for the Leukemia data, and 17 pCP and 49 N-pCR for breast cancer data. We then repeat the above procedure 1000 times and report their average misclassification rates in Table 4.3. From the results, it is evident that the new discriminant method $\text{LDA}_{\text{New}}$ gives a comparable performance in both data sets along with different $p$ values. This demonstrates that our proposed estimator for the precision matrix can be useful in practice.

5. Discussions
In this paper, we consider a class of ridge-type estimators \( \hat{\Omega}_p = \alpha (S_n + \beta I_p)^{-1} \) of the precision matrix \( \Omega_p \). Under the loss function \( tr(\hat{\Omega}_p \Sigma_p - I_p)^2 / p \), the optimal shrinkage coefficients \( \alpha \) and \( \beta \) are determined and estimated consistently. The resulting estimator \( \hat{\Omega}_p^* = \alpha^*_n (S_n + \beta^*_n I_p)^{-1} \) has a simple and closed form. A different, but similar idea in spirit can also be found in Bodnar et al. (2013) in which \( \alpha S_n^{-1} + \beta I_p \) is constructed for the precision matrix. This idea can be traced back to, for example, Efron and Morris (1976), Haff (1977, 1979) and Yang and Berger (1994). Nevertheless, these estimators suffer the singularity problem when \( p \) is larger than or equal to \( n \). From this point of view, the proposed estimator \( \hat{\Omega}_p^* \) has extended the existing literature from small dimension to large dimension.

In the present paper, the shrinkage estimator is constructed between the sample covariance matrix and the identity matrix. Inspired by Schäfer and Strimmer (2005) and Fisher and Sun (2011), one may also consider other target matrices so that the resulting estimator is \( \hat{\Omega}_p = \alpha (S_n + \beta T)^{-1} \). Compared with the squared error loss function (Ledoit and Wolf, 2004; Warth, 2008; Fisher and Sun, 2011; Ledoit and Wolf, 2012), the loss function considered in this paper can accommodate the situations with extreme eigenvalues (Daniels and Kass, 2001). Note also that Stein’s loss function (Stein, 1975) can be another alternative of interest and it may deserve further study for the corresponding behavior of the proposed estimator.

Finally, we note that the proposed ridge-type shrinkage estimator may be suboptimal for high dimensional data with \( p \gg n \). When \( p \) is much larger than \( n \), to have a good estimate of \( \Omega_p \), one may need to rely on some prior information on the structure of the precision matrix. For instance, under the sparsity assumption that most of the off-diagonal elements in \( \Omega_p \) are zero or near zero, Cai et al. (2011) proposed a constrained \( \ell_1 \) minimization method for \( \Omega_p \). For more details, see also the recent review paper of Tong et al. (2014) and the references therein. Nevertheless, as argued in Ledoit and Wolf (2012), such prior information on the structure of the precision matrix may not always be available or even trustworthy. In these scenarios, the shrinkage methods can be considered and they will provide more or less improvement on the estimation. Note also that the estimators for sparse matrices may not be guaranteed to be well-conditioned (Xue et al., 2012; Rothman, 2012). The ridge-type shrinkage estimator is, however, always invertible and positive definite. Moreover, our proposed estimator has a very simple structure and the shrinkage coefficients can be easily calculated. Instead, most existing
methods for sparse precision matrices involve one or more tuning parameters and cross-validation procedures are often required for choosing the parameter values. Based on these facts, for large precision matrix with little structure information, we recommend the use of our proposed ridge-type shrinkage estimator.

Acknowledgments

The research of Guangming Pan was partially supported by the Ministry of Education, Singapore, under grant ARC 14/11. Tiejun Tong’s research was supported by Hong Kong Research grant HKBU202711 and Hong Kong Baptist University FRG grants FRG2/11-12/110 and FRG1/13-14/018. Lixing Zhu’s research was supported by a grant from the Research Grants Council of Hong Kong and a Faculty Research Grant (FRG) from Hong Kong Baptist University. The authors thank the editor, the associate editor, and two reviewers for their helpful comments and suggestions that have substantially improved an early version of this manuscript.

References


A ridge-type shrinkage estimator for large precision matrix


A ridge-type shrinkage estimator for large precision matrix


Cheng Wang
Department of Mathematics and Center for Statistics Science, Shanghai Jiao Tong University, Shanghai, China
Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong
E-mail: cescwang@gmail.com

Guangming Pan
School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore
E-mail: gmpan@ntu.edu.sg

Tiejun Tong
Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong
E-mail: tongt@hkbu.edu.hk

Lixing Zhu
Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong
E-mail: lzhu@hkbu.edu.hk