Statistica Sinica **25** (2015), 993-1008
doi:http://dx.doi.org/10.5705/ss.2012.328

SHRINKAGE ESTIMATION OF LARGE DIMENSIONAL PRECISION MATRIX USING RANDOM MATRIX THEORY

Cheng Wang[1,3], Guangming Pan[2], Tiejun Tong[3] and Lixing Zhu[3]

[1]*Shanghai Jiao Tong University,* [2]*Nanyang Technological University*
and [3]*Hong Kong Baptist University*

Abstract: This paper considers ridge-type shrinkage estimation of a 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. Numerical studies of simulated and real data demonstrate that the proposed estimator performs better than existing competitors in a wide range of settings.

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

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 a precision matrix has applications in such statistical problems as linear discriminant analysis (Anderson (2003)), Hotelling's T^2 test (Hotelling (1931)), and Markowitz mean-variance analysis (Markowitz (1952)). Let n be the sample size, p be the dimension of observation, and Σ_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 an independent random sample from a multivariate distribution (See, Bai and Saranadasa (1996) or Chen, Zhang, and Zhong (2010)),

$$X_i = \Sigma_p^{1/2} Y_i + \mu_0, \ i = 1, \ldots, n, \tag{1.1}$$

where μ_0 is a p-dimensional mean vector and Σ_p is a $p \times p$ positive definite covariance matrix. Here $\mathbb{Y} = (Y_1, \ldots, Y_n) = (Y_{ij})_{p\times n}$ 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 the sample covariance matrix be

$$S_n = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})(X_j - \bar{X})^T, \tag{1.2}$$

where $\bar{X} = \sum_{j=1}^{n} X_j/n$ is the sample mean and the superscript T denotes the transpose of a matrix or vector. 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. Since

$$E(S_n^{-1}) = \frac{n-1}{n-p-2}\Sigma_p^{-1}, \tag{1.3}$$

an unbiased estimator of the precision matrix Ω_p is $(n-p-2)S_n^{-1}/(n-1)$. In this paper, we are interested in estimating Ω_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)). When such prior information about the structure of covariance matrix is not available, one often considers shrinkage methods to improve the standard estimators after James and Stein (1961). Thus, 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 that significantly improves the standard estimator. Nevertheless, these methods require that p be less than n and none of the eigenvalues be zero.

To overcome the singularity problem when $p \geq n$, Ledoit and Wolf (2004) proposed a shrinkage estimator for Σ_p, 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). Very little work has been done for estimating the precision matrix directly. To the best of our knowledge, the only work in this direction is by Kubokawa and Srivastava (2008), in which the authors considered a ridge-type estimator for the precision matrix,

$$\hat{\Omega}_p = \alpha(S_n + \beta I_p)^{-1}, \tag{1.4}$$

where α and β are two shrinkage coefficients. Here the derivation of the shrinkage coefficients α and β in (1.4) can be more challenging than those in estimating Σ_p, especially when $p \geq n$. Kubokawa and Srivastava (2008) assumed the data to be Gaussian distributed and employed an empirical Bayes method for specifying the shrinkage coefficients, demonstrating that the resulting estimator dominates the usual estimator.

We propose to derive the optimal shrinkage coefficients α and β for non-Gaussian data under the loss function (Haff (1979); Krishnamoorthy and Gupta (1989); Yang and Berger (1994)),

$$\frac{1}{p} tr(\hat{\Omega}_p \Sigma_p - I_p)^2. \tag{1.5}$$

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 study the theoretical loss of the ridge-type estimator, derive the optimal shrinkage coefficients α and β, and develop a data-driven shrinkage estimator for the precision matrix in Section 3. In Section 4, we conduct numerical studies with simulated and real data to evaluate the performance of the proposed estimator and to compare it with some existing methods. We conclude the paper in Section 5 and proofs are provided in an online supplement.

2. Preliminary Results 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 $I(\cdot)$ is the indicator function. ESD plays an important role in multivariate analysis and many statistics can be expressed as functionals of it, e.g., $\det(A_m) = \exp(m \int \log(x) dF^{A_m(x)})$ and $\mathrm{tr}(A_m) = m \int x dF^{A_m(x)}$. For more details, see a 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

$$m_F(z) = \int \frac{1}{t - z} dF(t), \quad z \in \mathbb{C}^+ \equiv \{z \in \mathbb{C} : \mathrm{Im} z > 0\}, \tag{2.1}$$

and the inversion formula is

$$F\{[c, d]\} = \lim_{\eta \to 0^+} \frac{1}{\pi} \int_c^d \mathrm{Im} m_F(\xi + i\eta) d\xi, \tag{2.2}$$

where $c < d$ are continuity points of F. By (2.2), F is uniquely determined by its Stieltjes transform.

We need two conditions.

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

(S2) There exists constants c_1 and c_2 such that $c_1 \le \lambda_{min}(\Sigma_p) \le \lambda_{max}(\Sigma_p) \le c_2$; F^{Σ_p} tends to a non-random probability distribution H as p tends to infinity.

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

$$m(z) = \int \frac{1}{\lambda/t - z + 1/(1 + ym(z))} dH(t), \tag{2.3}$$

where $\lambda > 0$ *and* $z \in \mathbb{C}^+$.

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

$$m_0(z) = \int \frac{dH(t)}{t(1 - y - yzm_0(z)) - z}. \tag{2.4}$$

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

$$m(-\lambda) = \int \frac{dH(t)}{t(1 - y + y\lambda m(-\lambda)) + \lambda}, \tag{2.5}$$

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

The condition $1 - y + y\lambda m(-\lambda) \ge 0$ 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 point. Assuming $\Sigma_p = I_p$, (2.5) has the 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 why Chen et al. (2011) claimed $m_0(-\lambda) = m^{(1)}(-\lambda)$, not $m_0(-\lambda) = m^{(2)}(-\lambda)$.

Theorem 2. *Assume that* $(S1)$ *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 $\xrightarrow{a.s.}$ *is almost sure convergence and*

$$R_1(\lambda) = \frac{1 - \lambda m_0(-\lambda)}{1 - y(1 - \lambda m_0(-\lambda))},$$
$$R_2(\lambda) = \frac{1 - \lambda m_0(-\lambda)}{(1 - y(1 - \lambda m_0(-\lambda)))^3} - \frac{\lambda m_0(-\lambda) - \lambda^2 m_0'(-\lambda)}{(1 - y(1 - \lambda m_0(-\lambda)))^4}.$$

In addition, we have

$$\frac{1}{p} tr((S_n + \lambda I_p)^{-1}) \xrightarrow{a.s.} m_0(-\lambda),$$
$$\frac{1}{p} tr((S_n + \lambda I_p)^{-2}) \xrightarrow{a.s.} m_0'(-\lambda) = \frac{dm_0(z)}{dz}|_{z=-\lambda}.$$

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

3. Shrinkage Estimation of Precision Matrix

We consider a ridge-type estimator for the precision matrix,

$$\hat{\Omega}_p = \alpha(S_n + \beta I_p)^{-1}, \tag{3.1}$$

where $\alpha > 0$ and $\beta > 0$ are shrinkage coefficients. By Theorem 2, we have

$$\begin{aligned} \frac{1}{p} tr(\Sigma_p \hat{\Omega} - I_p)^2 &\xrightarrow{a.s.} \alpha^2 R_2(\beta) - 2\alpha R_1(\beta) + 1 \\ &= R_2(\beta)(\alpha - \frac{R_1(\beta)}{R_2(\beta)})^2 + 1 - \frac{(R_1(\beta))^2}{R_2(\beta)}. \end{aligned} \tag{3.2}$$

By (3.2), the optimal α is $\alpha_{\text{opt}} = R_1(\beta)/R_2(\beta)$ for any fixed β. This leads to the simplified loss function

$$L(\beta) = 1 - \frac{(R_1(\beta))^2}{R_2(\beta)}. \tag{3.3}$$

Let $L_0 = \min_{\beta>0} L(\beta)$ be the minimum loss and $\beta_{\text{opt}} = \arg\min_{\beta>0} L(\beta)$ be the optimal parameter of β.

Theorem 3. *For any $y<1$, we have $L_0=\min_{\gamma>0}L_H(\gamma)$, where*

$$L_H(\gamma)=1-\Big(\int\frac{t}{t+\gamma}dH(t)\Big)^2\Big(\Big(\int\frac{t^2}{(t+\gamma)^2}dH(t)\Big)^{-1}-y\Big),\ \gamma\geq 0.$$

Specially, when $H(x)$ is a degenerate distribution at σ^2, the minimum loss is $L_0=0$. For a general distribution $H(x)$, $L_H(\gamma)$ achieves its global minimum value L_0 at γ^* satisfying

$$\frac{f_1(\gamma^*)f_3(\gamma^*)-f_2(\gamma^*)f_2(\gamma^*)}{f_2(\gamma^*)f_2(\gamma^*)(f_1(\gamma^*)-f_2(\gamma^*))}=y, \tag{3.4}$$

where $f_k(x)=\int(t/(t+x))^k dH(t)$. Correspondingly, $L(\beta_{\rm opt})=L_0$ where $\beta_{\rm opt}$ satisfies

$$\gamma^*=\frac{\beta_{\rm opt}}{1-y(1-\beta_{\rm opt}m_0(-\beta_{\rm opt}))}.$$

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

Here $\alpha_{\rm opt}$ and $\beta_{\rm opt}$ are unknown and need to be estimated. We consider a data-driven method for estimating $\alpha_{\rm opt}$ and $\beta_{\rm opt}$. Let

$$\hat{R}_1(\lambda)=\frac{a_1(\lambda)}{1-\hat{y}a_1(\lambda)},$$
$$\hat{R}_2(\lambda)=\frac{a_1(\lambda)}{(1-\hat{y}a_1(\lambda))^3}-\frac{a_2(\lambda)}{(1-\hat{y}a_1(\lambda))^4},$$

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

$$L_n(\lambda)=1-\frac{(\hat{R}_1(\lambda))^2}{\hat{R}_2(\lambda)}.$$

We take $\beta_n^*=\arg\min_{\beta\in[C_1,C_2]}L_n(\beta)$ and $\alpha_n^*=\hat{R}_1(\beta_n^*)/\hat{R}_2(\beta_n^*)$. In case β_n^* is not unique, we specify the smallest solution.

Theorem 4. *Assume that $(S1)$ and $(S2)$ hold. For any $\lambda>0$, $L_n(\lambda)\xrightarrow{a.s.}L(\lambda)$ as $n\to\infty$. In addition,*

$$\frac{1}{p}tr(\alpha_n^*(S_n+\beta_n^*I_p)^{-1}\Sigma_p-I_p)^2\xrightarrow{a.s.}L_0\qquad\text{as}\quad n\to\infty. \tag{3.5}$$

Theorem 4 shows that the proposed estimator can achieve the minimum loss L_0 asymptotically. In view of this, we propose the estimator of Ω_p as

$$\hat{\Omega}_p^* = \alpha_n^*(S_n + \beta_n^* I_p)^{-1}. \tag{3.6}$$

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

$$\frac{1}{p} tr\Big(\frac{n-p-2}{n-1}\Sigma_p S_n^{-1} - I_p\Big)^2 \xrightarrow{a.s.} \frac{y}{1-y}, \tag{3.7}$$

the new estimator $\hat{\Omega}_p^*$ performs asymptotically better than the classical estimator S_n^{-1} and also the unbiased estimator $(n-p-2)/(n-1)S_n^{-1}$. 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, Gupta, and Parolya (2013)) are no longer applicable.

4. Numerical Studies

4.1. Monte Carlo simulation study

The components of data such as gene expression data can have different scales. As in Warton (2008), we propose a two-stage procedure to implement the new estimator. 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 $diag(S_n)$ to rescale the inverse correlation matrix and estimate the precision matrix Ω_p by

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

We compared the new estimator with the following three estimators.

I. The scaled standard estimator (referred to as the SSE estimator)

$$\hat{\Omega}_{\text{SSE}} = \frac{n-p-2}{n-1} S_n^{-1} I(p < n) + \frac{p}{n-1} S_n^{+} I(p \geq n), \tag{4.1}$$

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).

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

$$\hat{\Omega}_{\text{EM}} = \frac{n-p-2}{n-1} S_n^{-1} + \frac{p^2+p-2}{(n-1)tr(S_n)} I_p. \tag{4.2}$$

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

$$\hat{\Omega}_{\mathrm{KS}} = p((n-1)S_n + tr(S_n)I_p)^{-1}. \tag{4.3}$$

For illustration purposes, we also included a recent shrinkage estimator designed for estimating the covariance matrix. Fisher and Sun (2011) considered a combination between S_n and $diag(S_n)$ to estimate Σ_p, $\hat{\Sigma}_{\mathrm{FS}} = \hat{\lambda} S_n + (1 - \hat{\lambda})diag(S_n)$, where $\hat{\lambda}$ was estimated by Fisher and Sun (2011). We then estimate Ω_p by

$$\hat{\Omega}_{\mathrm{FS}} = (\hat{\lambda} S_n + (1-\hat{\lambda})diag(S_n))^{-1}. \tag{4.4}$$

We refer to this as the FS estimator.

To conduct simulation studies in a wide range of settings, we considered four models for generating the covariance matrix.

M1. Σ_1 is diagonal with 20% of the population eigenvalues equal to 1, 40% equal to 3, and the rest 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 $\sigma_{ij} = 0.5^{|i-j|}$ for $1 \le i, j \le p$.

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. $\Sigma_4 = U_2 diag(\lambda_1, \ldots, \lambda_p)U_2^T$, where $\lambda_j = 2 + 0.125j, j = 1, \ldots, p$ and the rows of U_2 are eigenvectors of Σ_0.

Model 1 is a diagonal spiked example which is well studied in RMT (Bai and Silverstein (1998), Ledoit and Wolf (2012)). Model 2 is an example of a sparse matrix whose 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. With respect to the random part $\mathbb{Y} = (Y_{ij})_{p\times n}$, we considered the standard normal $Y_{ij} \sim N(0,1)$, the mixture normal $Y_{ij} \sim 0.5N(0,1)+0.5N(1,1)$, Student's t-distribution $Y_{ij} \sim t(5)$, and the log-normal $\log(Y_{ij}) \sim N(0.5, 0.5^2)$. In each case, Y_{ij} was standardized to have unit variance.

With 100 simulations for each simulation setting, we report in Table 1 the average losses (1.5) for the new estimator and the competitors, where the data were simulated from the normal and the mixture normal distribution, respectively. Seven combinations (p, n) were considered, among which three were $p < n$, one were $p = n$, and the rest three were $p > n$. Specially, the combination $(p, n) = (1,000,\ 100)$ represents the popular setting of high-dimensional low-sample-size data. The EM estimator is excluded in the last four combinations since S_n is singular when $p \ge n$. From the results in Table 1, we observe that the new estimator $\hat{\Omega}_{\mathrm{New}}$ always outperforms the existing competitors, no matter whether p is less than n or not. There is not 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 and the Moore-Penrose inverse S_n^+ does not perform well in large dimensional data.

Since the new estimator is distribution free, to investigate its performance under other distributions, we conducted another simulation study where the data were simulated from Student's t-distribution and the log-normal distribution. All other settings were kept as before. With 100 simulations for each setting, we report in Table 2 the average losses for $(p, n) = (100, 200)$ and $(200, 100)$. Together with the results in Table 1, we see that the performance of the new estimator was only slightly affected by the violation of normality assumption; the comparison results among the estimators remained similar for all the distribution considered. We also conducted simulations for other combinations of (p, n), with the conclusions remaining the same.

Our third study was to investigate how p and n affect the performance of the new estimator. Since the performance of the estimators was quite consistent for different distributions and different covariance matrices, we considered only Gaussian data with covariance matrix from Model 2. For the ratio p/n either 1/2 or 2, we plot the average losses in Figure 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.

4.2. Real data analysis

Shrinkage estimators of the covariance matrix are commonly applied to linear discriminant analysis. See, for example, Friedman (1989), Srivastava and Kubokawa (2007), Kubokawa and Srivastava (2008), Fan, Feng, and Wu (2009), Cai, Liu, and Luo (2011), Fisher and Sun (2011), among others. Here, we illustrate the usefulness of the proposed shrinkage estimator with the Leukemia data in Golub et al. (1999) and the breast cancer data in Hess et al. (2006). The Leukemia data contains a total of 7,129 genes for 47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML). The breast cancer data has 22,283 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.

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

Table 1. Empirical risks of the proposed estimator and existing estimators for the normal and mixture normal distributions.

	$N(0,1)$				$0.5N(0,1)+0.5N(1,1)$			
Method	M1	M2	M3	M4	M1	M2	M3	M4
$(p,\ n)=(50,\ 100)$								
SSE	1.0849	1.0905	1.0701	1.0866	1.1071	1.0905	1.1039	1.1089
EM	1.4617	1.7308	1.9892	1.3955	1.4873	1.7379	1.9790	1.3676
KS	0.4641	0.5250	0.4996	0.3943	0.4639	0.5257	0.4990	0.3954
FS	0.0240	0.3721	0.6362	0.1348	0.0239	0.3680	0.6379	0.1319
New	0.0238	0.2423	0.1268	0.1164	0.0233	0.2409	0.1259	0.1149
$(p,\ n)=(100,\ 200)$								
SSE	1.0478	1.0516	1.0363	1.0398	1.0523	1.0423	1.0528	1.0584
EM	1.4127	1.6660	2.4586	1.3501	1.4407	1.6860	2.4189	1.3403
KS	0.4654	0.5265	0.5019	0.4100	0.4650	0.5252	0.5041	0.4112
FS	0.0111	0.3497	1.2435	0.1728	0.0105	0.3505	1.2575	0.1711
New	0.0109	0.2343	0.0975	0.1451	0.0103	0.2337	0.0938	0.1445
$(p,\ n)=(100,\ 1000)$								
SSE	0.1135	0.1132	0.1126	0.1131	0.1131	0.1123	0.1123	0.1122
EM	0.1290	0.1383	0.1682	0.1251	0.1285	0.1375	0.1679	0.1243
KS	0.8259	0.8378	0.8318	0.8169	0.8259	0.8378	0.8318	0.8170
FS	0.0021	0.0942	0.1354	0.0814	0.0020	0.0938	0.1351	0.0812
New	0.0021	0.0811	0.0343	0.0748	0.0020	0.0809	0.0340	0.0745
$(p,\ n)=(100,\ 50)$								
SSE	6.0464	8.5135	6.2353	3.7927	6.0684	8.5984	6.3137	3.8912
KS	0.3516	0.5088	0.3969	0.2071	0.3520	0.5085	0.4001	0.2069
FS	0.0527	0.7336	1.5249	0.2562	0.0515	0.7362	1.5335	0.2515
New	0.0483	0.3780	0.2251	0.1936	0.0470	0.3777	0.2291	0.1918
$(p,\ n)=(200,\ 100)$								
SSE	6.0486	8.4596	6.1888	4.3786	6.0110	8.5202	6.1581	4.3508
KS	0.3503	0.5028	0.3729	0.2376	0.3509	0.5029	0.3747	0.2375
FS	0.0225	0.6534	5.0668	0.2743	0.0221	0.6529	4.9641	0.2734
New	0.0218	0.3597	0.1597	0.2069	0.0212	0.3594	0.1629	0.2069
$(p,\ n)=(1000,\ 100)$								
SSE	0.9800	1.2232	1.0954	0.9475	0.9805	1.2275	1.0903	0.9475
KS	0.4531	1.0757	0.6274	0.2957	0.4539	1.0735	0.6158	0.2959
FS	0.0235	0.7302	8.9487	0.3534	0.0226	0.7285	9.6948	0.3520
New	0.0251	0.4039	0.3228	0.2506	0.0246	0.4038	0.3188	0.2501
$(p,\ n)=(100,\ 100)$								
SSE	1.09e7	9.96e7	1.63e8	3.05e8	2.61e7	4.61e7	3.42e8	1.19e8
KS	0.3529	0.4348	0.3924	0.2660	0.3524	0.4349	0.3926	0.2668
FS	0.0243	0.5349	2.0889	0.2069	0.0224	0.5292	2.0496	0.2038
New	0.0237	0.3130	0.1409	0.1658	0.0217	0.3112	0.1460	0.1643

Table 2. Empirical risks of the proposed estimator and existing estimators for Student's $t-$distribution and the log-normal distribution.

	$t(5)$				$\ln N(0.5, 0.5^2)$			
Method	M1	M2	M3	M4	M1	M2	M3	M4
				$(p,\ n) = (100,\ 200)$				
SSE	1.1349	1.1395	1.1443	1.1238	1.1447	1.1536	1.1522	1.1541
EM	1.5246	1.7657	2.5691	1.4270	1.5371	1.7902	2.5634	1.4607
KS	0.4627	0.5265	0.4996	0.4088	0.4631	0.5256	0.5014	0.4091
FS	0.0307	0.3845	1.3683	0.1956	0.0418	0.3964	1.3840	0.2054
New	0.0306	0.2435	0.1091	0.1581	0.0411	0.2476	0.1137	0.1643
				$(p,\ n) = (200,\ 100)$				
SSE	6.4237	9.0814	6.5239	4.6874	6.5340	9.1796	6.5612	4.6670
KS	0.3530	0.5064	0.3767	0.2378	0.3534	0.5064	0.3760	0.2380
FS	0.0619	0.7394	5.1303	0.3259	0.0837	0.7601	5.6854	0.3488
New	0.0585	0.3727	0.1813	0.2277	0.0803	0.3792	0.1899	0.2402

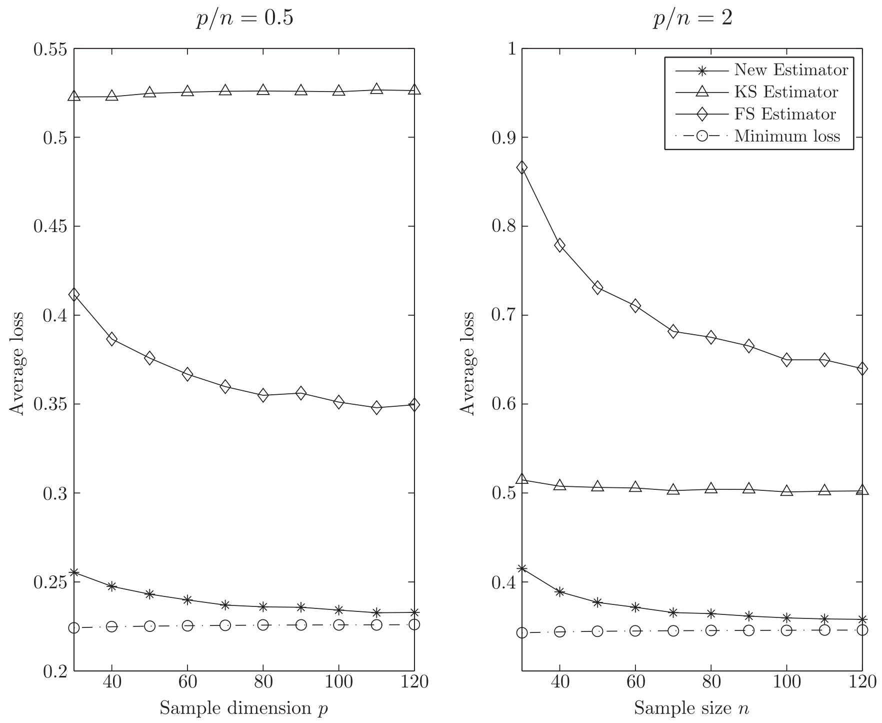

Figure 1. Average losses of the new estimator and the competitors for different combinations of p and n, where the data are Gaussian distributed and the covariance matrix is generated using Model 2. The minimum losses L_0 are reported for comparison.

LDA$_{\rm SSE}$: $d_i = (x_0 - \bar{x}_i)^T \hat{\Omega}_{\rm SSE}(x_0 - \bar{x}_i)$ for $i = 1, 2$.

LDA$_{\rm EM}$: $d_i = (x_0 - \bar{x}_i)^T \hat{\Omega}_{\rm EM}(x_0 - \bar{x}_i)$ for $i = 1, 2$.

LDA$_{\rm KS}$: $d_i = (x_0 - \bar{x}_i)^T \hat{\Omega}_{\rm KS}(x_0 - \bar{x}_i)$ for $i = 1, 2$.

LDA$_{\rm FS}$: $d_i = (x_0 - \bar{x}_i)^T (\hat{\Sigma}_{\rm FS})^{-1}(x_0 - \bar{x}_i)$ for $i = 1, 2$.

LDA$_{\rm New}$: $d_i = (x_0 - \bar{x}_i)^T \hat{\Omega}_{\rm New}(x_0 - \bar{x}_i)$ 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 methods were to classify x_0 to group 1 if $d_1 < d_2$, and to group 2 otherwise. For a more comprehensive comparison, we also included three widely used classifiers in the literature: the diagonal linear discriminant analysis (DLDA) in Dudoit, Fridlyand, and Speed (2002) or Bickel and Levina (2004), the 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 selected p = 25, 50, 100, 500 or 1,000 genes, and randomly divided the total samples into two distinct sets, one for the training set and the other for the test set. We fixed the training set sizes as 23 ALL and 12 AML for the Leukemia data, and 17 pCP and 49 N-pCR for the breast cancer data. We then repeated the procedure 1,000 times and now report their average misclassification rates in Table 3. From the results, it is evident that the new discriminant method LDA$_{\rm New}$ gives a comparable performance in both data sets for different p values. Our proposed estimator for the precision matrix can be useful in practice.

5. Discussion

We consider a class of ridge-type estimators $\hat{\Omega}_p = \alpha(S_n + \beta I_p)^{-1}$ of the precision matrix Ω_p. Under the loss function $tr(\hat{\Omega}_p \Sigma_p - I_p)^2/p$, the optimal shrinkage coefficients α and β 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 be found in Bodnar, Gupta, and Parolya (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, those estimators suffer from 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 existing methods from small to large dimensions.

The shrinkage estimator is constructed between the sample covariance matrix and the identity matrix. Inspired by Schäfer and Strimmer (2005) and Fisher

Table 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.

	$p = 25$	$p = 50$	$p = 100$	$p = 500$	$p = 1,000$
Leukemia data					
DLDA	24.723(8.587)	20.284(7.769)	16.719(7.484)	14.454(7.417)	13.413(7.277)
NSC	25.500(9.781)	20.786(8.684)	16.570(7.914)	10.859(7.119)	8.546(6.124)
HCT	25.870(9.647)	20.692(8.596)	16.124(7.606)	11.481(7.353)	9.786(7.127)
LDA_{SSE}	33.811(9.578)	27.611(9.394)	18.751(7.952)	12.949(6.152)	10.589(5.414)
LDA_{EM}	29.824(10.301)	NA	NA	NA	NA
LDA_{KS}	25.143(10.046)	21.016(9.237)	17.157(8.091)	10.273(5.338)	8.005(4.604)
LDA_{FS}	26.186(11.125)	20.811(10.406)	15.838(8.753)	7.949(4.741)	7.097(4.460)
LDA_{New}	23.595(9.026)	18.257(7.557)	14.414(6.464)	7.981(4.698)	7.154(4.457)
Breast cancer data					
DLDA	36.704(7.542)	34.867(7.371)	33.207(7.546)	31.657(7.003)	31.252(7.576)
NSC	37.518(7.935)	35.637(7.669)	33.554(7.448)	31.391(6.680)	30.485(6.679)
HCT	36.813(7.741)	34.127(7.249)	31.655(6.990)	28.851(6.172)	27.924(6.751)
LDA_{SSE}	36.124(6.229)	39.102(6.844)	36.921(6.367)	31.182(5.592)	30.364(5.544)
LDA_{EM}	36.034(7.299)	37.161(8.234)	NA	NA	NA
LDA_{KS}	35.669(7.166)	33.715(6.726)	31.418(5.999)	28.809(5.241)	27.660(5.026)
LDA_{FS}	34.751(7.968)	32.008(8.542)	28.812(8.657)	24.192(4.058)	26.584(4.520)
LDA_{New}	33.652(6.947)	30.782(7.423)	28.142(7.811)	24.155(4.077)	26.640(4.621)

and Sun (2011), one can 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); Warton (2008); Fisher and Sun (2011); Ledoit and Wolf (2012)), the loss function considered in this paper can accommodate situations with extreme eigenvalues (Daniels and Kass (2001)). Stein's loss function (Stein (1975)) is another alternative of interest and deserve further study for the corresponding behavior of the proposed estimator.

The proposed ridge-type shrinkage estimator may be suboptimal for high-dimensional data with p much larger than n. To have a good estimate of Ω_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 Ω_p are zero or near zero, Cai, Liu, and Luo (2011) proposed a constrained ℓ_1 minimization method for Ω_p. See also the recent review paper of Tong, Wang, and Wang (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 be available or even trustworthy. In such scenarios, shrinkage methods can be considered and they will provide more or less improvement on the estimation. Estimators for sparse matrices may not be guaranteed to be well-conditioned (Xue, Ma, and Zou (2012); Rothman (2012)) whereas the ridge-type shrinkage estimator is always invertible and positive definite. Our proposed estimator has a simple structure and the shrinkage coefficients can be easily calculated. 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. For large precision matrix with little structures information, we can recommend the use of the proposed ridge-type shrinkage estimator.

Acknowledgements

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, an associate editor, and two reviewers for their helpful comments and suggestions that have substantially improved an early version of this manuscript.

References

Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statist. Sinica* **6**, 311-330.

Bai, Z. and Silverstein, J. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Ann. Probab.* **26**, 316-345.

Bai, Z. and Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices.* Springer, New York.

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

Bodnar, T., Gupta, A. K. and Parolya, N. (2013). Optimal linear shrinkage estimator for large dimensional precision matrix. arXiv:1308.0931.

Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106**, 594-607.

Chen, L., Paul, D., Prentice, R. and Wang, P. (2011). A regularized Hotelling's T^2 test for pathway analysis in proteomic studies. *J. Amer. Statist. Assoc.* **106**, 1345-1360.

Chen, S., Zhang, L. and Zhong, P. (2010). Tests for high-dimensional covariance matrices. *J. Amer. Statist. Assoc.* **105**, 810-819.

Daniels, M. J. and Kass, R. E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *J. Amer. Statist. Assoc.* **94**, 1254-1263.

Daniels, M. J. and Kass, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics* **57**, 1173-1184.

Dey, D. K. and Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *Ann. Statist.* **13**, 1581-1591.

Donoho, D. and Jin, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Nat. Acad. Sci.* **105**, 14790-14795.

Dudoit, S., Fridlyand, J. and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.* **97**, 77-87.

Efron, B. and Morris, C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. *Ann. Statist.* **4**, 22-32.

Fan, J., Feng, Y. and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *Ann. Appl. Statist.* **3**, 521.

Fisher, T. and Sun, X. (2011). Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Comput. Statist. Data Anal.* **55**, 1909-1918.

Friedman, J. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.* **84**, 165-175.

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

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537.

Haff, L. (1977). Minimax estimators for a multinormal precision matrix. *J. Multivariate Anal.* **7**, 374-385.

Haff, L. (1979). An identity for the Wishart distribution with applications. *J. Multivariate Anal.* **9**, 531-544.

Hess, K., Anderson, K., Symmans, W., Valero, V., Ibrahim, N., Mejia, J., Booser, D., Theriault, R., Buzdar, A., Dempsey, P., et al. (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J. Clinical Oncology* **24**, 4236-4244.

Hotelling, H. (1931). The generalization of Student's ratio. *Ann. Math. Statist.* **2**, 360-378.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 361-379.

Konno, Y. (2009). Shrinkage estimators for large covariance matrices in multivariate real and complex normal distributions under an invariant quadratic loss. *J. Multivariate Anal.* **100**, 2237-2253.

Krishnamoorthy, K. and Gupta, A. (1989). Improved minimax estimation of a normal precision matrix. *Canad. J. Statist.* **17**, 91-102.

Kubokawa, T. and Srivastava, M. (2008). Estimation of the precision matrix of a singular Wishart distribution and its application in high-dimensional data. *J. Multivariate Anal.* **99**, 1906-1928.

Ledoit, O. and Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probab. Theory Related Fields* **151**, 233-264.

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

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.* **40**, 1024-1060.

Markowitz, H. (1952). Portfolio selection. *J. Finance* **7**, 77-91.

Mestre, X. and Lagunas, M. (2006). Finite sample size effect on minimum variance beamformers: Optimum diagonal loading factor for large arrays. *IEEE Trans. Signal Process.* **54**, 69-82.

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

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Mol. Biol.* **4**, 1175-1189.

Silverstein, J. (1995). Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *J. Multivariate Anal.* **55**, 331-339.

Silverstein, J. and Choi, S. (1995). Analysis of the limiting spectral distribution of large dimensional random matrices. *J. Multivariate Anal.* **54**, 295-309.

Srivastava, M. (2007). Multivariate theory for analyzing high dimensional data. *J. Japan Statist. Soc.* **37**, 53-86.

Srivastava, M. and Kubokawa, T. (2007). Comparison of discrimination methods for high dimensional data. *J. Japan Statist. Soc.* **37**, 123-134.

Stein, C. (1975). Estimation of a covariance matrix. 39*th Annual Meeting IMS Rietz Lecture*.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statist. Sci.* **18**, 104-117.

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

Warton, D. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *J. Amer. Statist. Assoc.* **103**, 340-349.

Xue, L., Ma, S. and Zou, H. (2012). Positive-definite ℓ_1-penalized estimation of large covariance matrices. *J. Amer. Statist. Assoc.* **107**, 1480-1491.

Yang, R. and Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. *Ann. Statist.* **22**, 1195-1211.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **99**, 2261-2286.

Department of Mathematics, Shanghai Jiao Tong University, Shanghai, China.

Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong.

E-mail: cescwang@gmail.com

School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore.

E-mail: gmpan@ntu.edu.sg

Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong.

E-mail: tongt@hkbu.edu.hk

Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong.

E-mail: lzhu@hkbu.edu.hk

(Received November 2012; accepted May 2014)