Author's personal copy

Journal of Multivariate Analysis 125 (2014) 222–232

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Non-parametric shrinkage mean estimation for quadratic loss functions with unknown covariance matrices

Cheng Wang [a,b], Tiejun Tong [b,*], Longbing Cao [c], Baiqi Miao [a]

[a] Department of Statistics and Finance, University of Science and Technology of China, Hefei, Anhui 230026, China

[b] Department of Mathematics, Hong Kong Baptist University, Hong Kong

[c] Advanced Analytics Institute, University of Technology, Sydney, NSW 2007, Australia

ARTICLE INFO

Article history:
Received 14 January 2013
Available online 15 January 2014

AMS subject classification:
62H12

Keywords:
High-dimensional data
Shrinkage estimator
Large p small n
U-statistic

ABSTRACT

In this paper, a shrinkage estimator for the population mean is proposed under known quadratic loss functions with unknown covariance matrices. The new estimator is non-parametric in the sense that it does not assume a specific parametric distribution for the data and it does not require the prior information on the population covariance matrix. Analytical results on the improvement of the proposed shrinkage estimator are provided and some corresponding asymptotic properties are also derived. Finally, we demonstrate the practical improvement of the proposed method over existing methods through extensive simulation studies and real data analysis.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

High-throughput molecular technologies that enable researchers to collect and monitor information at the genome level have revolutionized the field of biology in the past fifteen years. These data offer an unprecedented amount and diverse types of data that reveal different aspects of the biological processes. One such example is microarray data, where the expression levels of thousands of genes are measured simultaneously from each sample. These data have motivated the development of reliable biomarkers for disease subtypes classification and diagnosis, and for the identification of novel targets for drug treatment. Due to the cost and other experimental difficulties such as the availabilities of biological materials, it is common that high-throughput data are collected only in a limited number of samples. They are often referred to as high-dimension, low-sample-size data, or "large p small n" data where p is the number of genes or dimensions and n is the sample size. High-dimensional data pose many challenges to the traditional statistical and computational methods. Specifically, due to the small size n, there are more uncertainties associated with standard estimators of parameters such as the mean and variance estimations. As a consequence, statistical analyses based on such parameter estimation are usually unreliable.

In this work, our interest is conducing a more accurate estimator for the population mean μ under the "large p small n" setting [16,25]. An accurate estimate of μ is desired in many areas of statistical analysis, e.g., in linear discriminant analysis [1], diagonal linear discriminant analysis [10], Markowitz mean–variance analysis [22,12] and so on. Under the assumption that μ is sparse, Shao et al. [23] proposed a consistent estimator for μ under some regular conditions. However, in many real problems, there is often little prior information on μ and it may not necessarily have a sparse structure. In such situations, the shrinkage estimation of μ can be applied. Shrinkage estimation starts with the amazing result of James and Stein [17] that the commonly used sample mean of a normal distribution is inadmissible and can be improved by shrinkage

* Corresponding author.
E-mail address: tongt@hkbu.edu.hk (T. Tong).

0047-259X/$ – see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.12.012

Author's personal copy

estimators. We refer to them as James–Stein type estimators. Since then, there is a large body of literature in shrinkage estimation including [3,11,20,6,14,13,9] and etc. In the literature, most existing methods either assumed that the covariance matrix Σ_p is known or assumed that there exists an estimator of Σ_p that is invertible. As a common practice, if the sample covariance matrix S_n is used to estimate Σ_p, the sample size is required to be larger than the dimension, i.e. $n > p$, to avoid the singularity. Note that, however, for high-dimensional data it is common that p is much larger than n. To overcome the singularity problem for high-dimensional data, Tong et al. [25] proposed a new shrinkage estimator for μ by assuming that Σ_p has a diagonal structure which may not be realistic. Therefore, the traditional shrinkage methods cannot be applied to analyze high-dimensional data directly.

Inspired by Ledoit and Wolf [18], in this paper we consider the shrinkage estimation for μ under quadratic loss functions with unknown non-diagonal covariance matrices. The new estimator is non-parametric in the sense that it does not assume a specific parametric distribution for the data and it does not require the prior information on covariance matrix Σ_p. We will demonstrate by both theoretical and empirical studies that the proposed estimator has good properties for a wide range of settings. We will also show that the proposed method is better than the sample mean and the existing shrinkage methods even under a diagonal covariance matrix assumption.

The rest of the paper is organized as follows. Section 2 introduces the theoretical optimal shrinkage estimation under quadratic risks. Section 3 develops a data-driven shrinkage estimator and derives the asymptotic properties of the proposed estimator. We then conduct simulation studies on simulated data in Section 4 and using real data in Section 5 to evaluate the proposed optimal shrinkage estimator and compare it with existing shrinkage methods. Finally, we conclude the paper in Section 6 and provide the technical results in the Appendix.

2. Methodology

Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) observations satisfying the multivariate model

$$X_i = \Sigma_p^{1/2}\epsilon_i + \mu, \quad i = 1, \ldots, n, \tag{2.1}$$

where μ is a p-dimensional vector, Σ_p is a positive definite matrix and the random errors in $(\epsilon_{ij})_{p\times n} = (\epsilon_1, \ldots, \epsilon_n)$ are i.i.d. with zero mean, unit variance and finite fourth moment. Note that model (2.1) has been widely used in the literature such as [2,8]. In this paper, we do not assume that the data follow a multivariate normal distribution with mean μ and covariance matrix Σ_p. Given model (2.1), we consider to estimate μ under the following quadratic loss function [4,6,14],

$$L_Q(\delta) = n(\delta - \mu)'Q(\delta - \mu)/\mathrm{tr}(Q\,\Sigma_p), \tag{2.2}$$

where $\delta = \delta(X_1, \ldots, X_n)$ is the estimator of μ, Q is a known positive definite matrix, and $\mathrm{tr}(A)$ stands for the trace of matrix A. Note that for the standard sample mean $\bar{X} = (1/n)\sum_{k=1}^n X_k$, the risk function is $E[L_Q(\bar{X})] = 1$.

In the special case when $X_1, \ldots, X_n$ are multivariate normal distributed, James and Stein [17] showed that

$$\delta_{JS} = \left(1 - \frac{p-2}{n\bar{X}'\bar{X}}\right)\bar{X} \tag{2.3}$$

dominates $\bar{X}$ for any $p > 2$ under the assumption that $\Sigma_p = Q = I_p$. This result was then extended by Baranchik [3] to $\Sigma_p = \sigma^2 I_p$ with σ^2 unknown, and by Efron and Morris [11] to a Bayesian estimator. For a general unknown Σ_p, the James–Stein estimator has the form [20,4,6,14,13]

$$\delta_{JS} = \left(I - \frac{r(Q, S_n^{-1}, \bar{X})}{\bar{X}'S_n^{-1}\bar{X}}\right)\bar{X}, \tag{2.4}$$

where $r(Q, S_n^{-1}, \bar{X})$ is a measurable function of Q, S_n^{-1} and $\bar{X}$ with $0 \le r(Q, S_n^{-1}, \bar{X}) \le 2(G-2)/(n-G+2)$ and $S_n = \sum_{k=1}^n (X_k - \bar{X})(X_k - \bar{X})'/(n-1)$ being the sample covariance matrix. To guarantee S_n is invertible, $n > p$ is necessary which means the method is not applicable for the "large p small n" data.

To overcome the singularity problem, Tong et al. [25] considered a special situation where Σ_p is diagonal. Specifically, under the loss function with $Q = \Sigma_p^{-1}$ they constructed a hierarchical Bayesian model and then proposed the following shrinkage estimator,

$$\delta_T = \left(1 - \frac{(p-2)(n-1)}{n(n-3)\bar{X}'D_n^{-1}\bar{X}}\right)\bar{X} \tag{2.5}$$

where $D_n = \mathrm{diag}(S_n)$. Other related works for a diagonal Σ_p and a diagonal Q assumptions include [5,24]. Whereas for an arbitrary Q with non-diagonal Σ_p, it remains a challenging yet unanswered question under the "large p small n" setting. To address this question, we consider to estimate μ by a linear combination of $\bar{X}$ and $e = (1, \ldots, 1)'$,

$$\delta = \alpha\bar{X} + \beta e.$$

The following theorem derives the optimal shrinkage coefficients for model (2.1) under the quadratic loss (2.2) with an arbitrary known Q.

Author's personal copy

Theorem 2.1. *Consider the optimization problem,*

$$\min_{\alpha,\beta} E(\delta-\mu)'Q(\delta-\mu) \quad s.t.\ \delta=\alpha\bar{X}+\beta e, \tag{2.6}$$

where the coefficients α and β are non-random. The optimal shrinkage estimator is given as $\mu^*=\alpha^*\bar{X}+\beta^* e$ where

$$\alpha^*=\frac{\mu'Q\mu-\frac{(e'Q\mu)^2}{e'Qe}}{\mu'Q\mu+\frac{1}{n}\mathrm{tr}(Q\Sigma_p)-\frac{(e'Q\mu)^2}{e'Qe}},$$

$$\beta^*=\frac{\frac{1}{n}\mathrm{tr}(Q\Sigma_p)}{\mu'Q\mu+\frac{1}{n}\mathrm{tr}(Q\Sigma_p)-\frac{(e'Q\mu)^2}{e'Qe}}\frac{e'Q\mu}{e'Qe},$$

and the corresponding risk of μ^* is

$$E(L_Q(\mu^*))=\frac{\left(\mu-\frac{e'Q\mu}{e'Qe}e\right)'Q\left(\mu-\frac{e'Q\mu}{e'Qe}e\right)}{\left(\mu-\frac{e'Q\mu}{e'Qe}e\right)'Q\left(\mu-\frac{e'Q\mu}{e'Qe}e\right)+\frac{1}{n}\mathrm{tr}(Q\Sigma_p)}. \tag{2.7}$$

Note that the proposed shrinkage estimator can accommodate any shift of the grand mean, including the shift from μ to $\mu+ce$ where c is a constant. This is a similar idea as that in [21] where the author shrunk the observations to grand mean rather than to the origin. Also in [25], the authors applied their shrinkage method to the grand mean and so the final estimator was a linear combination of two different components. By Theorem 2.1, however, we point out that the method in [21] is not applicable for arbitrary Q. For this point, we will explain in the simulation study an example where the grand mean is zero but $e'Q\mu\neq 0$.

3. Data-driven shrinkage estimators for population means

Note that the shrinkage coefficients α^* and β^* are unknown and need to be estimated in practice. In this section, we propose to estimate them by U-statistics, motivated from [8,7,19]. Specifically, we estimate α^* and β^* by

$$\hat{\alpha}^*=\frac{Y_{1,n}-Y_{3,n}}{Y_{1,n}+Y_{2,n}-Y_{3,n}} \quad \text{and} \quad \hat{\beta}^*=\frac{Y_{2,n}}{Y_{1,n}+Y_{2,n}-Y_{3,n}}Y_{4,n}$$

where

$$Y_{1,n}=\frac{1}{p(n-1)}\sum_{i\neq j}X_i'QX_j,$$

$$Y_{2,n}=\frac{1}{np}\left(\sum_{k=1}^{n}X_k'QX_k-\frac{1}{n-1}\sum_{i\neq j}X_i'QX_j\right),$$

$$Y_{3,n}=\frac{1}{p(n-1)e'Qe}\sum_{i\neq j}e'QX_iX_j'Qe,$$

$$Y_{4,n}=\frac{1}{ne'Qe}\sum_{k=1}^{n}e'QX_k.$$

The resulting estimator of μ is then $\hat{\mu}^*=\hat{\alpha}^*\bar{X}+\hat{\beta}^*e$. To derive the asymptotic properties of the proposed estimator, we need the following regularity condition.

Assumption 3.1. There is a constant c_0 (not depending on p or n) such that

$$c_0^{-1}\leq \text{all eigenvalues of } \Sigma_p \quad \text{and} \quad Q\leq c_0.$$

Under Assumption 3.1, we have $\mathrm{tr}(\Sigma_pQ)/p=O(1)$. In this work, $o(1)$ denotes a sequence of random variables that converges to zero and $O(1)$ is short for a sequence that is bounded. Similarly, $o_p(1)$ and $O_p(1)$ are notations in probability. For more details, one may refer to [26]. Let

$$\pi_1=E(\bar{X}-\mu)'Q(\bar{X}-\mu)=\frac{1}{n}\mathrm{tr}(Q\Sigma_p),$$

$$\pi_2=\left(\mu-\frac{e'Q\mu}{e'Qe}e\right)'Q\left(\mu-\frac{e'Q\mu}{e'Qe}e\right).$$

The following theorems establish the rates of convergence for the proposed estimators and for the loss function.

Author's personal copy

Theorem 3.1. *Under Assumption 3.1,*

$$Y_{1,n} = \frac{n}{p}\mu' Q\mu + O_p\left(\frac{1}{\sqrt{p}}\right), \qquad Y_{2,n} = \frac{1}{p}\mathrm{tr}(\Sigma_p Q) + O_p\left(\frac{1}{\sqrt{np}}\right),$$
$$Y_{3,n} = \frac{n(\mu' Qe)^2}{pe' Qe} + O_p\left(\frac{1}{p}\right), \qquad Y_{4,n} = \frac{e' Qu}{e' Qe} + O_p\left(\frac{1}{\sqrt{np}}\right).$$

Further, we have

$$\hat{\alpha}^* = \alpha^* + O_p\left(\frac{1}{\sqrt{p}}\right) \quad \text{and} \quad \hat{\beta}^* = \beta^* + O_p\left(\frac{1}{\sqrt{np}}\right) + O_p\left(\frac{e' Qu}{p^{3/2}\left(\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2\right)}\right).$$

Therefore, $\hat{\alpha}^* - \alpha^* \xrightarrow{p} 0$ *and* $\hat{\beta}^* - \beta^* \xrightarrow{p} 0$ *as* $p \to \infty$ *and* $p^{-3/2} e' Qu \to 0$, *where* $\xrightarrow{p}$ *denotes convergence in probability.*

Theorem 3.2. *Under Assumption 3.1, under the "large p small n" setting the loss function of the shrinkage estimator* $\hat{\mu}^*$ *is*

$$L_Q(\hat{\mu}^*) = \frac{\pi_2}{\pi_1 + \pi_2} + O_p\left(\frac{1}{\sqrt{p}}\right). \tag{3.8}$$

By Theorem 3.2, we note that $\hat{\mu}^*$ behaves at least as well as $\bar{X}$ when p is large. The explicit improvement of $\hat{\mu}^*$ over $\bar{X}$ depends on π_1 and π_2. As in [18], we define the percentage relative improvement in average loss (PRIAL) over the sample mean as

$$\mathrm{PRIAL} = \frac{E(\bar{X} - \mu)' Q(\bar{X} - \mu) - (\hat{\mu}^* - \mu)' Q(\hat{\mu}^* - \mu)}{E(\bar{X} - \mu)' Q(\bar{X} - \mu)}. \tag{3.9}$$

We then have the following corollary.

Corollary 3.1. *Let* $s_n = \frac{n}{p}(\mu - \frac{e' Q\mu}{e' Qe} e)' Q(\mu - \frac{e' Q\mu}{e' Qe} e)$. *As* $p \to \infty$ *we have*

(I) *If* $s_n \to 0$, $\mathrm{PRIAL} \xrightarrow{p} 1$;
(II) *If* $s_n \to C_0$, $\mathrm{PRIAL} \xrightarrow{p} C_1 \in (0, 1)$;
(III) *If* $s_n \to \infty$, $\mathrm{PRIAL} \xrightarrow{p} 0$.

Therefore, the shrinkage estimator $\hat{\mu}^*$ always performs better than $\bar{X}$ under the loss function (2.2) when s_n is finite. In the extreme case when $s_n \to \infty$, $\hat{\mu}^*$ behaves similarly as $\bar{X}$.

4. Simulation studies

In this section, we conduct simulation studies to evaluate the performance of the proposed shrinkage estimator $\hat{\mu}^*$ and compare it with the following four estimators: the James–Stein estimator δ_{JS} in [3], the Berger–Bock estimator δ_{BB} in [5], the Tong et al. estimator δ_T in [25] and the CW estimator δ_{CW} proposed by Chételat and Wells [9].

Our first simulation will be conducted under the classical situation [17] where $Q = \Sigma_p = I_p$. In particular, we considered two scenarios with respect to the innovation random vector ϵ_{ij} in model (2.1):

(I) ϵ_{ij} are i.i.d from standard normal distribution that the data are distributed from Gaussian data with mean μ and an identity covariance matrix;
(II) $\sqrt{v/(v-2)} \times \epsilon_{ij}$ are i.i.d from t-distribution with v degrees of freedom where the constant $\sqrt{v/(v-2)}$ guarantees that ϵ_{ij} has zero mean and unit variance.

In both scenarios, we set $\mu \sim N_p(0, I_p)$ and all the simulation results are based on 10,000 simulations.

Table 1 reports the empirical risks of the proposed estimator and its competitors for both scenarios of distributions. We observe from Table 1 that overall the James–Stein estimator has the best performances since it utilized the fact that the population covariance matrix is an identity matrix and the proposed estimator has similar average losses as ones of the James–Stein estimator. All the estimators are applicable for variables with normal distributions and t-distributions. Moreover, the empirical risks for variables with t-distributions follow very similar patterns to those of Gaussian data. This is understandable as for data structure (2.1), there is no too much differences between Gaussian data and random data with other distributions such as t-distributions. For this reason, in the following part, we only report the simulation results for Gaussian data.

Note that the existing competitors for comparison, δ_{JS}, δ_B and δ_T, only work on a diagonal covariance matrix under the "large p small n" setting. Whereas for the proposed estimator $\hat{\mu}^*$, it works for both diagonal and non-diagonal covariance

Author's personal copy

Table 1
Empirical risks of the estimators under classical situation where $Q = \Sigma_p = I_p$ and $p = 100$.

n	Sample mean	James–Stein	Berger–Bock	Tong et al.	CW	Proposed
Normal random vectors						
10	1.0012	0.9088	0.9348	0.9094	0.9945	0.9090
20	0.9997	0.9508	0.9569	0.9509	0.9913	0.9508
50	0.9988	0.9790	0.9799	0.9790	0.9899	0.9790
t-distribution with $\upsilon = 5$						
10	1.0011	0.9079	0.9604	0.9097	0.9902	0.9080
20	1.0000	0.9511	0.9686	0.9515	0.9893	0.9511
50	1.0014	0.9814	0.9854	0.9815	0.9917	0.9815
t-distribution with $\upsilon = 10$						
10	0.9998	0.9075	0.9458	0.9080	0.9924	0.9076
20	1.0006	0.9519	0.9626	0.9520	0.9919	0.9520
50	0.9979	0.9776	0.9796	0.9776	0.9880	0.9776

Table 2
Empirical risks of the estimators under variance settings with covariance matrix Σ_1.

μ	τ	n	Sample mean	James–Stein	Berger–Bock	Tong et al.	Proposed
μ_1	0.5	10	1.0064	0.5160	0.6175	**0.4763**	0.4867
		25	1.0020	0.8092	0.7428	0.7195	**0.7184**
		50	0.9971	0.8853	0.7865	0.7787	**0.7766**
		100	0.9954	1.0426	0.9191	0.9182	**0.9179**
	1	10	0.9996	0.9524	0.8710	**0.8266**	0.8323
		25	1.0029	0.9231	0.8937	0.8827	**0.8811**
		50	0.9963	0.9868	0.9453	0.9430	**0.9429**
		100	1.0019	1.0024	0.9802	0.9795	**0.9793**
μ_2	0.5	10	1.0036	0.5693	0.6446	0.5083	**0.4229**
		25	1.0010	0.8119	0.7434	0.7165	**0.6160**
		50	0.9945	0.9188	0.8338	0.8268	**0.7535**
		100	0.9940	0.9752	0.9062	0.9044	**0.8605**
	1	10	0.9983	0.8989	0.8578	0.8006	**0.7300**
		25	0.9959	0.9759	0.9155	0.9062	**0.8612**
		50	0.9919	0.9927	0.9489	0.9470	**0.9223**
		100	0.9969	0.9949	0.9720	0.9716	**0.9586**

matrices. Thus in the following part, for a meaningful comparison, we will consider the quadratic loss function in [25]. Specifically, by letting Q be diagonal and let $Q^{-1} = \text{diag}(\Sigma_p)$, we have the following loss function,

$$L(\delta) = \frac{n}{p}(\delta - \mu)'[\text{diag}(\Sigma_p)]^{-1}(\delta - \mu), \tag{4.10}$$

where the constant n/p is applied to guarantee that $E[L(\bar{X})] = 1$. In applications, Q will be estimated from the diagonal elements of the sample covariance matrix.

We will simulate $X_1, \ldots, X_n$ independently from a p-dimensional multivariate normal distribution with mean μ and covariance matrix Σ_p. For μ, we consider two options:

(a) $\mu_1 = (\mu_{11}, \ldots, \mu_{1p})'$ where $\mu_{11}, \ldots, \mu_{1p}$ are i.i.d from $N(0, \tau^2)$;
(b) $\mu_2 = (\mu_{21}, \ldots, \mu_{2p})'$ where $\mu_{2k} = \tau$ for $k \le p/2$ and $\mu_{2k} = -\tau$ for $k > p/2$.

In both options, we consider $\tau = 0.5$ and 1 to represent different levels of mean heterogeneity. For Σ_p, we consider three covariance matrices:

(1) Σ_1 is diagonal with 20% of population eigenvalues being equal to 1, 40% begin equal to 3 and 40% being equal to 10;
(2) $\Sigma_2 = \Sigma_1^{1/2}\Sigma_0\Sigma_1^{1/2}$ where $\Sigma_0 = (\sigma_{ij})_{p\times p}$ and $\sigma_{ij} = \rho^{|i-j|}$ for $1 \le i, j \le p$;
(3) $\Sigma_3 = \Sigma_1^{1/2}\Sigma_{00}\Sigma_1^{1/2}$ where $\Sigma_{00} = (\sigma_{ij})_{p\times p}$ and $\sigma_{ij} = 1$ for $i = j$, $\sigma_{ij} = \rho$ for $i \ne j$.

We let ρ range from 0.1 to 0.9 for Σ_2 and from 0.1 to 0.5 for Σ_3 to represent four different levels of dependent structure.

The second simulation study is to evaluate the performance of $\hat{\mu}^*$ with existing methods when $\Sigma_p = \Sigma_1$, i.e., when the covariance matrix is diagonal. Let $p = 100$ throughout the simulations. We consider $n = 10, 25, 50$ and 100, to represent different levels of sample sizes. Table 2 reports the empirical risks of the estimators under various settings. First of all, we observe that all shrinkage methods have a smaller risk than the sample mean $\bar{X}$. This shows that for high-dimensional data, the shrinkage estimators do improve the standard estimation. Among the shrinkage estimators, δ_T and $\hat{\mu}^*$ are among the best in most settings. The James–Stein estimator δ_{JS} is not very compatible because it is restricted to a common variance

Author's personal copy

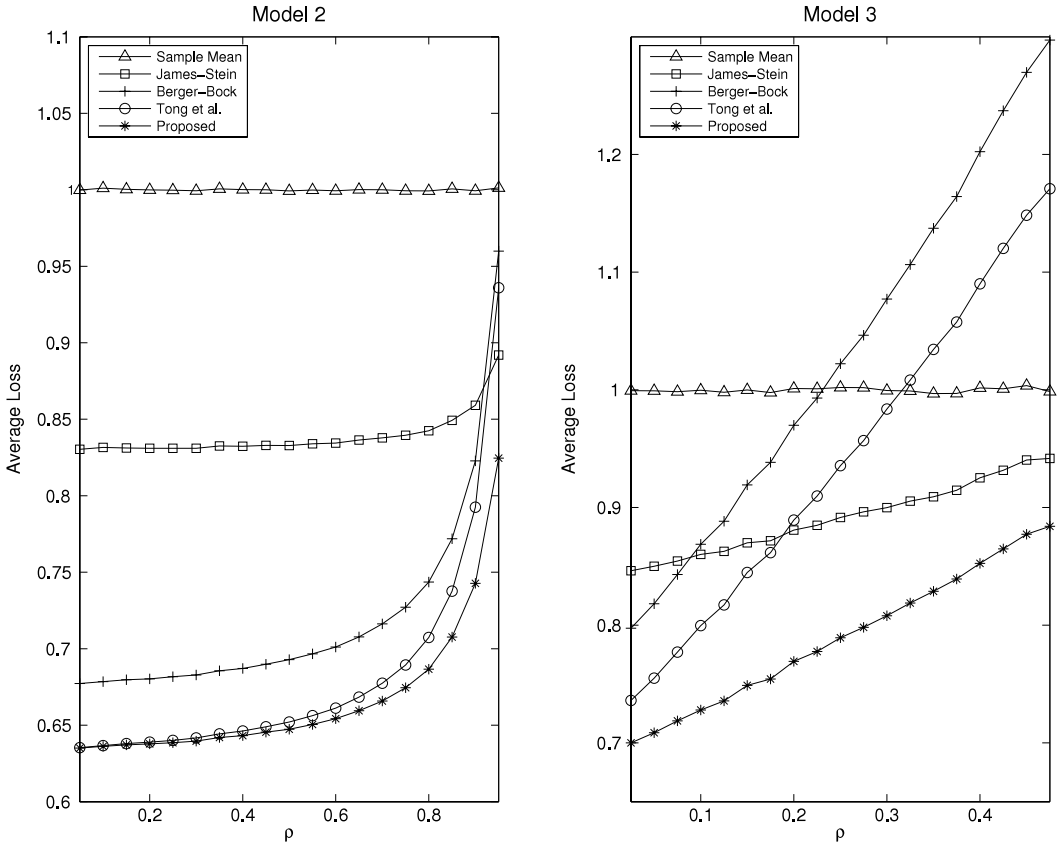

Fig. 1. Plots of the average losses of the proposed method and existing methods when the observations are correlated. Here $p = 100$, $n = 20$ and $\mu = \mu_1$ with $\tau = 0.5$.

assumption, and δ_B is only applicable for large sample sizes. Finally, for δ_T and $\hat{\mu}^*$, we note that they perform similarly when $\mu = \mu_1$, and $\hat{\mu}^*$ is better by a large margin than δ_T when $\mu = \mu_2$. In addition, when the mean heterogeneity increases from $\tau = 0.5$ to $\tau = 1$, the improvement of $\hat{\mu}^*$ over $\bar{X}$ decreases which is consistent with Corollary 3.1. We also observe that the improvements of the shrinkage estimators over the sample mean become smaller when n becomes larger. This is meaningful since for the large sample size scenario, the mean estimation itself is good enough and it is no longer necessary to borrow information from others to improve the estimation.

Finally, we will design simulation studies to evaluate the performance of $\hat{\mu}^*$ with existing methods when the covariance matrix is non-diagonal. This is to investigate the impact of the correlation coefficient ρ on the performance of the estimators. To achieve this, we plot in Fig. 1 the average losses of the estimators for covariance matrices Σ_2 and Σ_3 respectively. To save space, we only present the results for $p = 100$, $n = 20$, $\mu = \mu_1$ and $\tau = 0.5$; whereas the comparison patterns for other combination settings remain the similar. From the plots, it is evident that the proposed $\hat{\mu}^*$ provides a smaller average loss than the other estimators in most settings, no matter if ρ is small or not. We also note that (i) all the shrinkage estimators perform worse when ρ increases; and (ii) the risks of δ_B and δ_T may be even larger than 1 when the dependence structure is strong, say for Σ_3 with $\rho \geq 0.35$.

5. An application

In this section, we illustrate the proposed shrinkage estimator using the Leukemia data in [15]. The data set contains $p = 7129$ genes for 47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML), and is available online at the website http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi.

To evaluate the performance of the proposed estimator and compare it with existing methods, we randomly split the 47 ALL samples into the training set X_1 and the test set X_2. Specifically, we let the size of the training set range from 5 to 30 and the remaining samples assigned as the test set. Let $\bar{X}_1$ and $\bar{X}_2$ be the sample means of the training and test sets, respectively. We further standardize the ALL and AML sets so that each array has variance one across genes. For simplicity, we let $Q = I_p$ and let the loss function (4.10) be $L(\delta|\mu) = (n/p)(\delta - \mu)'(\delta - \mu)$. Then to compare the performance of the shrinkage estimator δ and the sample mean $\bar{X}_1$ based on the training set, we define the empirical partial risk (EPR) as

$$\text{EPR} = \frac{L(\bar{X}_2|\bar{X}_1) - L(\bar{X}_2|\delta)}{L(\bar{X}_2|\bar{X}_1)} = 1 - \frac{|\bar{X}_2 - \delta|^2}{|\bar{X}_2 - \bar{X}_1|^2}. \tag{5.11}$$

Author's personal copy

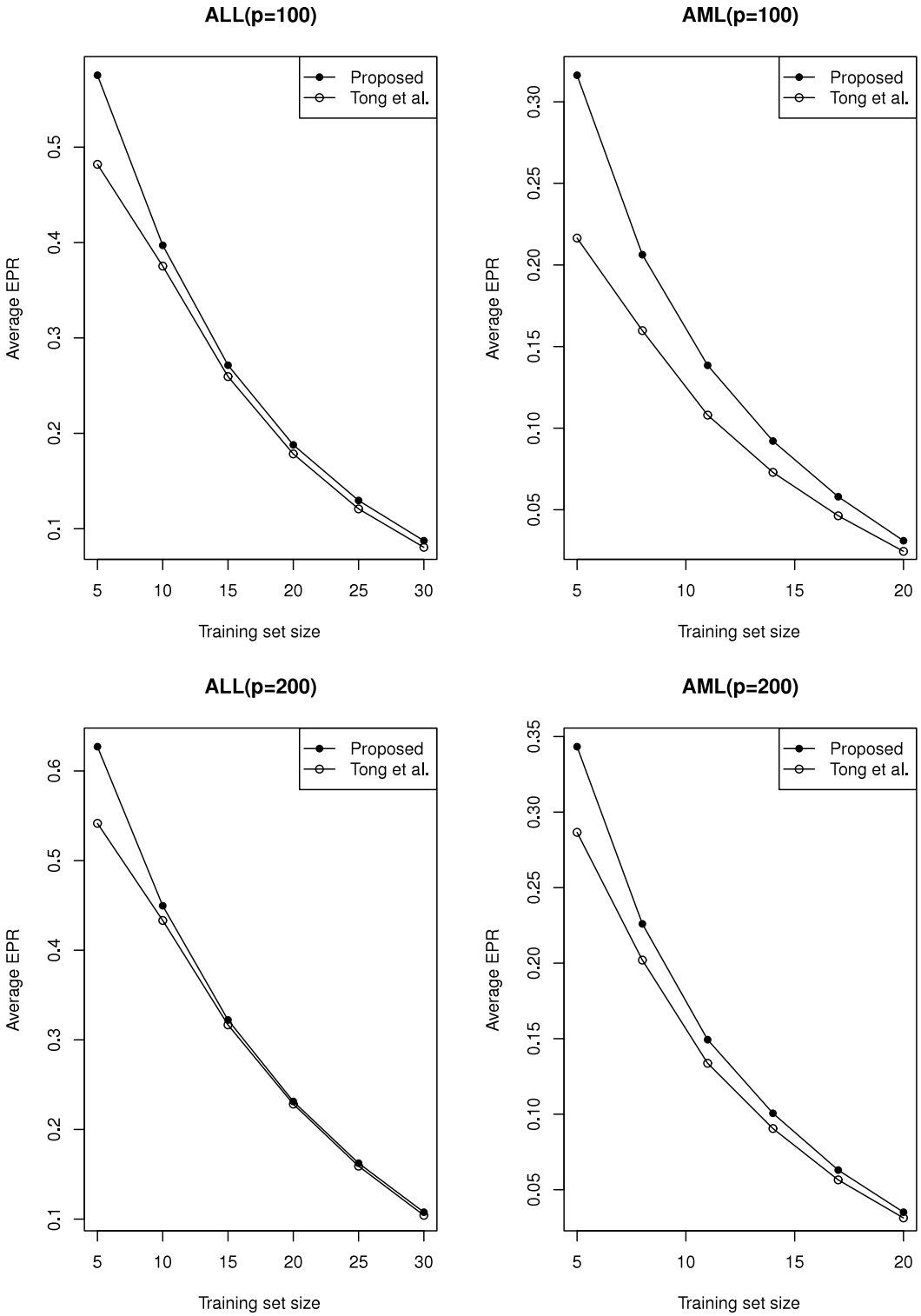

Fig. 2. The average EPRs of different shrinkage estimators on Leukemia data.

Intuitively, if δ estimates the true mean $\underline{\mu}$ more accurately than the sample mean $\bar{X}_1$, it will serve as a better proctor of μ and so $L(\bar{X}_2|\underline{\delta})$ will be smaller than $L(\bar{X}_2|\bar{X}_1)$. As a consequence, if the estimated EPR is larger than 0, we may clarify that δ is better than $\bar{X}_1$. Or equivalently, the EPR may represent the improvement of δ over $\bar{X}_1$.

With 10,000 simulations, we plot in Fig. 2 the average EPR using the first 100 and 200 genes of the AML and ALL sets with different sizes of the training set. Similarly as in Section 4, it is evident that the proposed estimator $\hat{\mu}^*$ outperforms the shrinkage estimator δ_T in most settings. We also note that the improvement of $\hat{\mu}^*$ over δ_T becomes smaller when the size of the training set increases. This shows that when the sample size is large, the performance of $\hat{\mu}^*$ over δ_T will be very similar. Meanwhile, the decreasing pattern of EPR on the training size indicates that both $\hat{\mu}^*$ and δ_T reduce to the sample mean $\bar{X}_1$ when the sample size is large.

Author's personal copy

6. Conclusion

The paper focuses on the shrinkage mean estimation under the "large p small n" setting. Specifically, we proposed a shrinkage estimator for the population mean under quadratic loss functions with unknown covariance matrix. Unlike existing methods in the literature, the proposed method does not assume a specific parametric distribution for the data and does not require any prior information on the covariance matrix. In this sense, the proposed estimator is a non-parametric shrinkage estimator and it works for both diagonal and non-diagonal covariance matrices. Except for the loss function (2.2), we note that another commonly used quadratic loss function is

$$L(\delta) = (\delta - \mu)^T \Sigma_p^{-1} (\delta - \mu). \tag{6.12}$$

When Σ_p is known, it results in a special case of the loss function (2.2). For the more realistic setting when Σ_p is unknown, if we do not assume a specific parametric distribution for the data, a shrinkage estimation for μ under the loss function (6.12) would be very difficult to achieve. Further research is warranted in this direction.

To verify the proposed estimator, we derived some analytical results on the estimator and on the optimal shrinkage coefficients. The estimators of the optimal shrinkage coefficients were also derived along with some asymptotic properties. We have also demonstrated through simulation studies using simulated data and real data that the proposed shrinkage estimator performs better than the sample mean estimation and the existing shrinkage methods under the "large p small n" setting. Finally, we note that the proposed method (i) extends the methods in [6,14] from the "small p large n" setting to the "large p small n" setting; and (ii) extends the methods in [5,25] from a diagonal covariance matrix assumption to a non-diagonal covariance matrix assumption. The proposed method has extensive applications in different areas including statistical genetics, epidemiology, ecology, and engineering sciences.

Acknowledgments

Cheng Wang's research was supported by NSF of China Grants (No. 11101397, 71001095 and 11271347). Tiejun Tong's research was supported by Hong Kong Research Grant HKBU202711 and Hong Kong Baptist University FRG Grands FRG2/10-11/020 and FRG2/11-12/110. Longbing Cao's research was supported by Australian Research Council Discovery Grant DP1096218 and Australian Research Council Linkage Grant LP100200774. The authors thank the editor, the associate editor, and two reviewers for their helpful comments and suggestions that have substantially improved the paper.

Conflict of Interest: None declared.

Appendix. Proofs of the theorems

A.1. Proof of Theorem 2.1

By direct calculation, we have

$$\begin{aligned} E(\delta-\mu)'Q(\delta-\mu) &= \alpha^2\left(\mu'Q\mu + \frac{1}{n}\mathrm{tr}(Q\Sigma_p)\right) + 2\alpha\mu'Q(\beta e-\mu) + (\beta e-\mu)'Q(\beta e-\mu) \\ &= \alpha^2\left(\mu'Q\mu + \frac{1}{n}\mathrm{tr}(Q\Sigma_p)\right) - (2\alpha-1)\mu'Q\mu + \beta^2 e'Qe - 2\beta(1-\alpha)e'Q\mu. \end{aligned}$$

This leads to the optimal weights as

$$\begin{aligned} \alpha^* &= \frac{\mu'Q\mu - \frac{(e'Q\mu)^2}{e'Qe}}{\mu'Q\mu + \frac{1}{n}\mathrm{tr}(Q\Sigma_p) - \frac{(e'Q\mu)^2}{e'Qe}} \\ \beta^* &= \frac{e'Q\mu}{e'Qe}(1-\alpha^*). \end{aligned}$$

Further, we have

$$\begin{aligned} E[L_Q(\mu^*)] &= \frac{\left(\mu'Q\mu - \frac{(e'Q\mu)^2}{e'Qe}\right)}{\mu'Q\mu + \frac{1}{n}\mathrm{tr}(Q\Sigma_p) - \frac{(e'Q\mu)^2}{e'Qe}} \\ &= \frac{\left(\mu - \frac{e'Q\mu}{e'Qe}e\right)' Q \left(\mu - \frac{e'Q\mu}{e'Qe}e\right)}{\left(\mu - \frac{e'Q\mu}{e'Qe}e\right)' Q \left(\mu - \frac{e'Q\mu}{e'Qe}e\right) + \frac{1}{n}\mathrm{tr}(Q\Sigma_p)}. \end{aligned}$$

Author's personal copy

A.2. Proof of Theorem 3.1

Without loss of generality, we assume that $E(\epsilon_{11}^4) = 3 + \Delta$. Then $E(Y_{1,n}) = (n/p)\mu'Q\mu$, $E(Y_{2,n}) = (1/p)\mathrm{tr}(\Sigma_p Q)$, $E(Y_{3,n}) = n(\mu'Qe)^2/(pe'Qe)$ and $E(Y_{4,n}) = e'Qu/(e'Qe)$. In addition, we have

$$\begin{aligned}
\mathrm{Var}(Y_{1,n}) &= \frac{2n}{p^2(n-1)}\mathrm{tr}(\Sigma_p Q \Sigma_p Q) = O\left(\frac{1}{p}\right),\\
\mathrm{Var}(Y_{2,n}) &= \frac{2}{p^2(n-1)}\mathrm{tr}(\Sigma_p Q \Sigma_p Q) + \frac{\Delta}{p^2 n}\mathrm{tr}((\Sigma_p^{1/2} Q \Sigma_p^{1/2}) \circ (\Sigma_p^{1/2} Q \Sigma_p^{1/2})) = O\left(\frac{1}{np}\right),\\
\mathrm{Var}(Y_{3,n}) &= \frac{2n}{p^2(n-1)}\left(\frac{e'Q\Sigma_p Qe}{e'Qe}\right)^2 = O\left(\frac{1}{p^2}\right),\\
\mathrm{Var}(Y_{4,n}) &= \frac{e'Q\Sigma_p Qe}{n(e'Qe)^2} = O\left(\frac{1}{np}\right),
\end{aligned}$$

where $A \circ B = (a_{ij}b_{ij})$ for matrices $A = (a_{ij})$ and $B = (b_{ij})$. This leads to

$$\begin{aligned}
Y_{1,n} &= \frac{n}{p}\mu'Q\mu + O_p\left(\frac{1}{\sqrt{p}}\right), \qquad Y_{2,n} = \frac{1}{p}\mathrm{tr}(\Sigma_p Q) + O_p\left(\frac{1}{\sqrt{np}}\right),\\
Y_{3,n} &= \frac{n(\mu'Qe)^2}{pe'Qe} + O_p\left(\frac{1}{p}\right), \qquad Y_{4,n} = \frac{e'Qu}{e'Qe} + O_p\left(\frac{1}{\sqrt{np}}\right).
\end{aligned}$$

For $\hat{\alpha}^*$ and $\hat{\beta}^*$, we have

$$\begin{aligned}
\hat{\alpha}^* - \alpha^* &= \frac{\frac{n}{p}\pi_2 + O_p\left(\frac{1}{\sqrt{p}}\right)}{\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2 + O_p\left(\frac{1}{\sqrt{p}}\right)} - \frac{\frac{n}{p}\pi_2}{\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2}\\
&= \frac{\left(\frac{n}{p}\pi_2 + \frac{n}{p}\pi_1\right) O_p\left(\frac{1}{\sqrt{p}}\right)}{\left(\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2 + O_p\left(\frac{1}{\sqrt{p}}\right)\right)\left(\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2\right)}\\
&= O_p\left(\frac{1}{\sqrt{p}\left(\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2\right)}\right)\\
&= O_p\left(\frac{1}{\sqrt{p}}\right),
\end{aligned}$$

and

$$\begin{aligned}
\hat{\beta}^* - \beta^* &= (1-\hat{\alpha}^*)Y_{4,n} - \beta^*\\
&= (1-\alpha^*)\left(Y_{4,n} - \frac{e'Qu}{e'Qe}\right) - (\hat{\alpha}^* - \alpha^*)Y_{4,n}\\
&= O_p\left(\frac{1}{\sqrt{np}}\right) + \left(\frac{e'Qu}{e'Qe} + O_p\left(\frac{1}{\sqrt{np}}\right)\right) O_p\left(\frac{1}{\sqrt{p}\left(\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2\right)}\right)\\
&= O_p\left(\frac{1}{\sqrt{np}}\right) + O_p\left(\frac{e'Qu}{p^{3/2}\left(\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2\right)}\right).
\end{aligned}$$

A.3. Proof of Theorem 3.2

First consider $(\mu^* - \mu)'Q(\mu^* - \mu)$. Note that

$$\begin{aligned}
(\mu^* - \mu)'Q(\mu^* - \mu) &= (\alpha^*(\bar{X} - \mu) + \beta^* e + (\alpha^* - 1)\mu)'Q(\alpha^*(\bar{X} - \mu) + \beta^* e + (\alpha^* - 1)\mu)\\
&= (\alpha^*)^2\left(\frac{1}{n^2}\sum_{k=1}^n \epsilon_k'\Sigma_p^{1/2}Q\Sigma_p^{1/2}\epsilon_k + \frac{1}{n^2}\sum_{i\neq j}\epsilon_i'\Sigma_p^{1/2}Q\Sigma_p^{1/2}\epsilon_j\right)\\
&\quad + \frac{2\alpha^*}{n}(\beta^* e + (\alpha^* - 1)\mu)'Q\Sigma_p^{1/2}\sum_{k=1}^n \epsilon_k + (\beta^* e + (\alpha^* - 1)\mu)'Q(\beta^* e + (\alpha^* - 1)\mu).
\end{aligned}$$

Author's personal copy

Then,

$$Var((\mu^*-\mu)'Q(\mu^*-\mu)) \le 2(\alpha^*)^4\left[\frac{2}{n^2}\mathrm{tr}(\Sigma_p Q\Sigma_p Q)+\frac{\Delta}{n^3}\sum_{k=1}^{p}\phi_{kk}^2\right] + \frac{8(\alpha^*)^2(1-\alpha^*)^2}{n}\left(\mu-\frac{e'Q\mu}{e'Qe}e\right)'Q\Sigma_p Q\left(\mu-\frac{e'Q\mu}{e'Qe}e\right),$$

where $\Sigma_p^{1/2}Q\Sigma_p^{1/2}=(\phi_{ij})_{p\times p}$.

By the definitions of π_1 and π_2, it is easy to verify that

$$\alpha^*=\frac{\pi_2}{\pi_1+\pi_2},$$
$$E(\mu^*-\mu)'Q(\mu^*-\mu)=\frac{\pi_1\pi_2}{\pi_1+\pi_2},$$
$$Var((\mu^*-\mu)'Q(\mu^*-\mu)) \le \frac{C_1}{n}\frac{\pi_1\pi_2^4+\pi_1^2\pi_2^3}{(\pi_1+\pi_2)^4}=\left(\frac{\pi_1\pi_2}{\pi_1+\pi_2}\right)^2 O\left(\frac{1}{p}\right).$$

Therefore,

$$(\mu^*-\mu)'Q(\mu^*-\mu)=\frac{\pi_1\pi_2}{\pi_1+\pi_2}\left(1+O_p\left(\frac{1}{\sqrt{p}}\right)\right). \tag{A.13}$$

By Theorem 3.1, we have

$$\hat\mu^*=\mu^*+O_p\left(\frac{1}{\sqrt{p}}\right)\bar X+\left(O_p\left(\frac{1}{\sqrt{np}}\right)+O_p\left(\frac{e'Qu}{p^{3/2}}\right)\right)e.$$

Note that

$$\begin{aligned}\hat\beta^*-\beta^* &= (1-\hat\alpha^*)Y_4-(1-\alpha^*)\frac{e'Qu}{e'Qe}\\ &= (\alpha^*-\hat\alpha^*)\frac{e'Qu}{e'Qe}+(1-\hat\alpha^*)\left(Y_4-\frac{e'Qu}{e'Qe}\right),\\ \hat\beta^*+\beta^* &= (1-\hat\alpha^*)Y_4+(1-\alpha^*)\frac{e'Qu}{e'Qe}\\ &= (2-\alpha^*-\hat\alpha^*)\frac{e'Qu}{e'Qe}+(1-\hat\alpha^*)\left(Y_4-\frac{e'Qu}{e'Qe}\right).\end{aligned}$$

We have

$$\begin{aligned}&(\hat\mu^*-\mu)'Q(\hat\mu^*-\mu)-(\mu^*-\mu)'Q(\mu^*-\mu)=(\hat\mu^*-\mu^*)'Q(\hat\mu^*+\mu^*-2u)\\ &=((\hat\alpha^*-\alpha^*)\bar X+(\hat\beta^*-\beta^*)e)'Q((\hat\alpha^*+\alpha^*)\bar X+(\hat\beta^*+\beta^*)e-2u)\\ &=((\hat\alpha^*-\alpha^*)(\bar X-\mu)+(\alpha^*-\hat\alpha^*)\left(\frac{e'Qu}{e'Qe}e-\mu\right)+(1-\hat\alpha^*)\left(Y_4-\frac{e'Qu}{e'Qe}\right)e)'Q\\ &\quad*\left((\hat\alpha^*+\alpha^*)(\bar X-\mu)+(2-\hat\alpha^*-\alpha^*)\left(\frac{e'Q\mu}{e'Qe}e-\mu\right)+(1-\hat\alpha^*)\left(Y_4-\frac{e'Qu}{e'Qe}\right)e\right)\\ &=(\hat\alpha^*-\alpha^*)\left[(\hat\alpha^*+\alpha^*)(\bar X-\mu)'Q(\bar X-\mu)+2(1-\hat\alpha^*-\alpha^*)(\bar X-\mu)'Q\left(\frac{e'Q\mu}{e'Qe}e-\mu\right)\right]\\ &\quad+2\hat\alpha^*(1-\hat\alpha^*)\left(Y_4-\frac{e'Qu}{e'Qe}\right)(\bar X-\mu)'Qe+\left(Y_4-\frac{e'Qu}{e'Qe}\right)^2(1-\hat\alpha^*)^2e'Qe\\ &\quad+(\hat\alpha^*-\alpha^*)(2-\hat\alpha^*-\alpha^*)\left(\mu-\frac{e'Qu}{e'Qe}e\right)'Q\left(\mu-\frac{e'Qu}{e'Qe}e\right)\\ &=\pi_1O_p\left(\frac{1}{\sqrt{p}}\right)+O_p\left(\frac{1}{\sqrt{p}\left(\frac{n}{p}\pi_1+\frac{n}{p}\pi_2\right)}\right)O_p\left(\sqrt{\frac{\pi_2}{n}}\right)+O_p\left(\frac{1}{n}\right)+O_p\left(\frac{\pi_2}{\sqrt{p}\left(\frac{n}{p}\pi_1+\frac{n}{p}\pi_2\right)}\right)\\ &=\pi_1\left(O_p\left(\frac{1}{\sqrt{p}}\right)+O_p\left(\frac{1}{p}\right)\right)=\pi_1O_p\left(\frac{1}{\sqrt{p}}\right),\end{aligned} \tag{A.14}$$

Author's personal copy

where we used the facts that

$$(\bar{X}-\mu)'Q(\bar{X}-\mu)=\pi_1\left(1+O_p\left(\frac{1}{\sqrt{p}}\right)\right),$$

$$(\bar{X}-\mu)'Q\left(\frac{e'Q\mu}{e'Qe}e-\mu\right)=O_p\left(\sqrt{\frac{\pi_2}{n}}\right),$$

$$(\bar{X}-\mu)'Qe=O_p\left(\sqrt{\frac{p}{n}}\right),$$

and

$$\hat{\alpha}^*-\alpha^*=O_p\left(\frac{1}{\sqrt{p}\left(\frac{n}{p}\pi_1+\frac{n}{p}\pi_2\right)}\right).$$

Finally, by (A.13) and (A.14) we have

$$(\hat{\mu}^*-\mu)'Q(\hat{\mu}^*-\mu)=\pi_1\left(\frac{\pi_2}{\pi_1+\pi_2}+O_p\left(\frac{1}{\sqrt{p}}\right)\right). \tag{A.15}$$

This completes the proof of Theorem 3.2.

References

[1] T. Anderson, An Introduction to Multivariate Statistical Analysis, in: Wiley Series in Probability and Statistics, 2003.
[2] Z. Bai, H. Saranadasa, Effect of high dimension: by an example of a two sample problem, Statist. Sinica 6 (1996) 311–330.
[3] A. Baranchik, A family of minimax estimators of the mean of a multivariate normal distribution, Ann. Math. Statist. 41 (1970) 642–645.
[4] J. Berger, Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss, Ann. Statist. 4 (1976) 223–226.
[5] J. Berger, M. Bock, Combining independent normal mean estimation problems with unknown variances, Ann. Statist. 4 (1976) 642–648.
[6] J. Berger, M. Bock, L. Brown, G. Casella, L. Gleser, Minimax estimation of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix, Ann. Statist. 5 (1977) 763–771.
[7] T. Cai, Z. Ma, Optimal hypothesis testing for high dimensional covariance matrices, 2012. ArXiv preprint arXiv:1205.4219.
[8] S. Chen, L. Zhang, P. Zhong, Tests for high-dimensional covariance matrices, J. Amer. Statist. Assoc. 105 (2010) 810–819.
[9] D. Chételat, M.T. Wells, Improved multivariate normal mean estimation with unknown covariance when p is greater than n, Ann. Statist. 40 (6) (2012) 3137–3160.
[10] S. Dudoit, J. Fridlyand, T. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, J. Amer. Statist. Assoc. 97 (2002) 77–87.
[11] B. Efron, C. Morris, Stein's estimation rule and its competitors-an empirical Bayes approach, J. Amer. Statist. Assoc. 68 (1973) 117–130.
[12] N. El Karoui, High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: risk underestimation, Ann. Statist. 38 (2010) 3487–3566.
[13] D. Fourdrinier, W. Strawderman, M. Wells, Robust shrinkage estimation for elliptically symmetric distributions with unknown covariance matrix, J. Multivariate Anal. 85 (2003) 24–39.
[14] L. Gleser, Minimax estimators of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix, Ann. Statist. 14 (1986) 1625–1633.
[15] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.
[16] J.T.G. Hwang, P. Liu, Optimal tests shrinking both means and variances applicable to microarray data analysis, Stat. Appl. Genet. Mol. Biol. 9 (2010) 36.
[17] W. James, C. Stein, Estimation with quadratic loss, in: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1, 1961, pp. 361–379.
[18] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivariate Anal. 88 (2004) 365–411.
[19] J. Li, S. Chen, Two sample tests for high-dimensional covariance matrices, Ann. Statist. 40 (2012) 908–940.
[20] P. Lin, H. Tsai, Generalized Bayes minimax estimators of the multivariate normal mean with unknown covariance matrix, Ann. Statist. 1 (1973) 142–145.
[21] D. Lindley, Discussion of professor Stein's paper: confidence sets for the mean of a multivariate normal distribution, J. Roy. Statist. Soc. Ser. B 24 (1962) 285–287.
[22] H. Markowitz, Portfolio selection, J. Finance 7 (1952) 77–91.
[23] J. Shao, Y. Wang, X. Deng, S. Wang, Sparse linear discriminant analysis by thresholding for high dimensional data, Ann. Statist. 39 (2011) 1241–1265.
[24] N. Shinozaki, Estimation of a multivariate normal mean with a class of quadratic loss functions, J. Amer. Statist. Assoc. 75 (1980) 973–976.
[25] T. Tong, L. Chen, H. Zhao, Improved mean estimation and its application to diagonal discriminant analysis, Bioinformatics 28 (2012) 531–537.
[26] A. Van der Vaart, Asymptotic Statistics, Cambridge University Press, 2000.