BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 4 2012, pages 531–537
doi:10.1093/bioinformatics/btr690

Gene expression

Advance Access publication December 14, 2011

Improved mean estimation and its application to diagonal discriminant analysis

Tiejun Tong [1,*], Liang Chen [2] and Hongyu Zhao [3,4]

[1]Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, [2]Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, [3]Department of Epidemiology and Public Health and [4]Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: High-dimensional data such as microarrays have created new challenges to traditional statistical methods. One such example is on class prediction with high-dimension, low-sample size data. Due to the small sample size, the sample mean estimates are usually unreliable. As a consequence, the performance of the class prediction methods using the sample mean may also be unsatisfactory. To obtain more accurate estimation of parameters some statistical methods, such as regularizations through shrinkage, are often desired.

Results: In this article, we investigate the family of shrinkage estimators for the mean value under the quadratic loss function. The optimal shrinkage parameter is proposed under the scenario when the sample size is fixed and the dimension is large. We then construct a shrinkage-based diagonal discriminant rule by replacing the sample mean by the proposed shrinkage mean. Finally, we demonstrate via simulation studies and real data analysis that the proposed shrinkage-based rule outperforms its original competitor in a wide range of settings.

Contact: tongt@hkbu.edu.hk

Received on August 24, 2011; revised on November 24, 2011; accepted on December 8, 2011

1 INTRODUCTION

With the advent of high-throughput technologies, learning high-dimensional complex models is critical in many disciplines such as biology, genetics, epidemiology, geology, ecology, neurology and engineering. One such example is microarray data, where the expression levels of thousands of genes are measured simultaneously from each sample. Due to the cost and/or 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 referred to as high-dimensional data with small sample size, or 'large G small n' data, where G is the number of dimensions and n is the sample size. High-dimensional data pose many challenges to traditional statistics methods. Specifically, due to the small n, there are more uncertainties associated with standard estimations of parameters such as the mean and variance estimations. As a consequence, statistical analyses based on such parameter estimation are usually unreliable. To obtain more accurate estimation of parameters some statistical methods, such as regularizations through shrinkage, may yield better results.

Shrinkage-based methods have been proposed in recent years to improve the variance estimation for 'large G small n' data. See for example, Baldi and Long (2001), Storey and Tibshirani (2003), Wright and Simon (2003), Smyth (2004), Cui *et al*. (2005), Tong and Wang (2007), Opgen-Rhein and Strimmer (2007) and Wang et al. (2009), among many others. In contrast to the advances on variance estimation, little attention has been paid to improving the mean estimation for high-dimensional data until recently (Hausser and Strimmer, 2009; Hwang and Liu, 2010). In this article, we investigate the family of shrinkage estimators for the mean value which is tailored to the high-dimensional data such as microarrays. Specifically, we will propose the optimal shrinkage parameter under the quadratic loss function for the data when G tends to be infinite.

Class prediction with high-dimensional data has been recognized as a very important problem and received much attention in different fields such as genomics, proteomics, brain images, medicine and machine learning. For high-dimensional data with small sample sizes, it is known that the traditional classification methods such as the linear discriminant analysis are not applicable as the sample covariance is going to singular when G is greater than n. To overcome the singularity problem, Dudoit *et al*. (2002) introduced two diagonalized discriminant rules, the diagonal linear discriminant analysis (DLDA) and the diagonal quadratic discriminant analysis (DQDA). When the sample size is small, DLDA performed remarkably well compared with more sophisticated classifiers in terms of both accuracy and stability (Dettling, 2004; Lee *et al*., 2005). In addition, DLDA is easy to implement and is not sensitive to the number of predictor variables.

Though DLDA performed well for high-dimensional small sample size data, there is still room to improve it (Huang *et al*., 2010; Pang *et al*., 2009; Pang *et al*., 2010). In particular, we notice that the mean estimation (the sample mean) in DLDA will be unreliable when the sample size is not sufficiently large. As a consequence, the performance of DLDA may also be unsatisfactory. With this insight, we propose in this article an improved version of DLDA, which replaces the sample mean by the optimal shrinkage estimator. We expect that the proposed shrinkage-based DLDA will improve the classification accuracy in practice.

The remainder of the article is organized as follows. In Section 2, we investigate the family of shrinkage estimators for the mean value under the quadratic loss function. With the nature of high-dimensional data, we assume that the variances are unequal and unknown. Under regularity conditions, we discuss the choices of the

*To whom correspondence should be addressed.

shrinkage parameter and then evaluate their practical performance via numerical studies. In Section 3, we construct a shrinkage-based diagonal discriminant rule by replacing the sample mean by the proposed shrinkage mean. Simulation studies will also be conducted to evaluate its performance over its original competitor. In Section 4, we use the leukemia data to demonstrate that the proposed method is widely applicable and performs well. Finally, we conclude the article in Section 5 with a brief discussion.

2 IMPROVED MEAN ESTIMATION

Let $\boldsymbol{X}_j=(X_{1j},\ldots,X_{Gj})^T$, $j=1,\ldots,n$, be independent G-dimensional vectors normally distributed with mean $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_G)^T$ and covariance matrix Σ. Let $\bar{\boldsymbol{X}}=\sum_{j=1}^n \boldsymbol{X}_j/n$ be the sample mean. Let $\|\bar{\boldsymbol{X}}\|^2=\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}$ and $\|\bar{\boldsymbol{X}}\|_\Sigma^2=\bar{\boldsymbol{X}}^T\Sigma^{-1}\bar{\boldsymbol{X}}$ for any invertible matrix Σ. In this section, we consider estimating $\boldsymbol{\mu}$ with respect to the quadratic loss function.

2.1 Motivation

Shrinkage estimation of means has a long history starting with the seminal paper of James and Stein (1961) that the sample mean is inadmissible for $G\geq 3$. When $\Sigma=I$, James and Stein (1961) showed that

$$\hat{\boldsymbol{\mu}}_{JS}=\left(1-\frac{(G-2)/n}{\|\bar{\boldsymbol{X}}\|^2}\right)\bar{\boldsymbol{X}}$$

dominates $\bar{\boldsymbol{X}}$ for any $G\geq 3$. When $\Sigma=\sigma^2 I$ with σ^2 unknown, the James-Stein type estimator is given as (Baranchik, 1970)

$$\hat{\boldsymbol{\mu}}_{JS}=\left(1-\frac{(G-2)/n}{\|\bar{\boldsymbol{X}}\|^2/\hat{\sigma}_0^2}\right)\bar{\boldsymbol{X}},$$

where $\hat{\sigma}_0^2$ is the pooled estimator of σ^2. More generally, when Σ is non-diagonal and unknown, the James-Stein type estimator has the form (Lin and Tsai, 1973; Gleser, 1986)

$$\hat{\boldsymbol{\mu}}_{JS}=\left(1-\frac{(G-2)/n}{\|\bar{\boldsymbol{X}}\|_{\hat{\Sigma}}^2}\right)\boldsymbol{X},$$

where $\hat{\Sigma}=\sum_{j=1}^n(\boldsymbol{X}_j-\boldsymbol{X})(\boldsymbol{X}_j-\boldsymbol{X})^T/(n-G+2)$ is the pooled sample covariance matrix. To guarantee $\hat{\Sigma}$ non-singular, we require that $n>G$. Note that this is not the case for high-dimensional data where G can be much larger than n. Therefore, the existing shrinkage methods for estimating $\boldsymbol{\mu}$ break down and cannot apply to the high-dimensional data directly.

2.2 Proposed mean estimator

To overcome the singularity problem, we assume Σ is diagonal and denote it by $D=\mathrm{diag}(\sigma_1^2,\ldots,\sigma_G^2)$. As mentioned in Hwang *et al.* (2009), most existing shrinkage estimators of $\boldsymbol{\mu}$ in the literature required the variances σ_i^2, $i=1,\ldots,G$, to be either equal or unequal but known. When $\sigma_i^2=\sigma^2$ for all i, the problem reduces to $D=\sigma^2 I$ with σ^2 unknown (James and Stein, 1961; Montazeri et al., 2010). When the σ_i^2 are unequal but known, the problem reduces to the case considered in Efron and Morris (1973). In this article, we consider the assumption of σ_i^2 being both unequal and unknown. Note that the same assumption has been commonly used in the recent literature, e.g. in Berger and Bock (1976), Dudoit *et al.* (2002), Bickel and Levina (2004), Cui *et al.* (2005), Tong and Wang (2007) and Hwang and Liu (2010) where the correlations among genes are ignored due to the small sample size.

Consider the following hierarchical Bayesian model with conjugate priors,

$$\begin{aligned} X_{ij}|\mu_i,\sigma_i^2 &\overset{\text{i.i.d.}}{\sim} N(\mu_i,\sigma_i^2),\\ \mu_i|\sigma_i^2 &\sim N(0,\sigma_i^2/\tau_0),\\ \sigma_i^2 &\sim \text{Inv-}\chi^2(\nu_0,\sigma_0^2), \end{aligned}$$

where $i=1,\ldots,G$, $j=1,\ldots,n$, $N(\cdot)$ is the standard normal distribution, and Inv-$\chi^2(\cdot)$ is the scaled inverse chi-squared distribution with unknown hyperparameters $(\tau_0,\nu_0,\sigma_0^2)$. The joint prior density of μ_i and σ_i^2 is given as

$$f(\mu_i,\sigma_i^2)\propto(\sigma_i^2)^{-(\nu_0+3)/2}\exp\left\{-\frac{1}{2\sigma_i^2}(\nu_0\sigma_0^2+\tau_0\mu_i^2)\right\}.$$

Let $X_i=\sum_{j=1}^n X_{ij}/n$ and $s_i^2=\sum_{j=1}^n(X_{ij}-X_i)^2/(n-1)$ be the sample mean and the sample variance for gene i, respectively. Let $\bar{\boldsymbol{X}}=(X_1,\ldots,X_G)^T$ and $S=\mathrm{diag}(s_1^2,\ldots,s_G^2)$. By Gelman *et al.* (2004), the posterior distribution of (μ_i,σ_i^2) can be represented as

$$f(\mu_i,\sigma_i^2|X_{i1},\ldots,X_{in})=N(\mu_{i,p},\frac{\sigma_i^2}{\tau_0+n})\cdot\text{Inv-}\chi^2(\nu_0+n,\sigma_{i,p}^2)$$

where

$$\begin{aligned} \mu_{i,p}&=\frac{n}{\tau_0+n}\bar{X}_i,\\ \sigma_{i,p}^2&=\frac{1}{\nu_0+n}\left(\nu_0\sigma_0^2+(n+1)s_i^2+\frac{n\tau_0}{\tau_0+n}\bar{X}_i^2\right). \end{aligned}$$

The posterior mean estimator of μ_i is then

$$\hat{\mu}_i=\mu_{i,p}=\left(1-\frac{1}{n/\tau_0+1}\right)\bar{X}_i,\qquad i=1,\ldots,G. \tag{1}$$

Note that τ_0 in (1) is an unknown parameter. We propose to estimate n/τ_0+1 by the form of $\|\bar{\boldsymbol{X}}\|_S^2/r$, where r is a shrinkage parameter to be tuned. We then have the following estimator for $\boldsymbol{\mu}$,

$$\hat{\boldsymbol{\mu}}(r)=\left(1-\frac{r}{\|\bar{\boldsymbol{X}}\|_S^2}\right)\bar{\boldsymbol{X}}. \tag{2}$$

Under the quadratic loss function

$$L(\hat{\boldsymbol{\mu}},\boldsymbol{\mu},D)=\frac{n}{G}\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_D^2, \tag{3}$$

Fourdrinier *et al.* (2003) showed that the estimator (2) dominates $\bar{\boldsymbol{X}}$ when $0<r<2(G-2)/(n+1)$.

2.3 Optimal shrinkage parameter

In this section, we propose the optimal shrinkage parameter of $r>0$ within the family of estimator (2). In Appendix A, we show that the optimal shrinkage parameter is given as

$$r_{\text{opt}}=\frac{(G-2)E(1/\|\boldsymbol{X}\|_S^2)}{nE\left[\|\boldsymbol{X}\|_D^2/(\|\boldsymbol{X}\|_S^2)^2\right]}. \tag{4}$$

Note that for high-dimensional data, n is usually small but G is large. We have the follow asymptotic result.

LEMMA 1. *Define* $N_m = \#\{k : (a_k / \sum_{i=1}^{k} a_i) \geq 1/m\}$, *where* $a_i = \bar{X}_i^2/\sigma_i^2$ *and* $\#\{A\}$ *denotes the total number of elements in set* A. *Assume that* $\sup_m(N_m/m) < \infty$ *almost surely. Then for any fixed* $n \geq 4$, *we have* $\|\bar{\boldsymbol{X}}\|_D^2/\|\bar{\boldsymbol{X}}\|_S^2 \xrightarrow{a.s.} (n-3)/(n-1)$ *as* $G \to \infty$.

The proof of Lemma 1 is given in Appendix B. By Lemma 1, for high-dimensional data, we have $\hat{r}_{\text{opt}} \approx [(n-1)(G-2)]/[n(n-3)]$ and the optimal shrinkage estimator is

$$\hat{\boldsymbol{\mu}}(\hat{r}_{\text{opt}}) = \left(1 - \frac{\hat{r}_{\text{opt}}}{\|\bar{\boldsymbol{X}}\|_S^2}\right)\bar{\boldsymbol{X}}. \tag{5}$$

Let $\tilde{r} = (G-2)/(n+1)$ be the middle point of the range $0 < r < 2(G-2)/(n+1)$ in Fourdrinier *et al*. (2003). It is clear that when n is large, $\hat{r}_{\text{opt}}$ and $\tilde{r}$ are asymptotically equivalent. While for high-dimensional data with small sample sizes, the difference between these two estimators can be large. For instance, when $n=5$, the ratio $\hat{r}_{\text{opt}}/\tilde{r}$ is given as 2.4; and when $n=6$, the ratio is 1.9. The practical performance of these two estimators will be studied in the next section.

2.4 Evaluation

The first simulation is to evaluate the performance of $\hat{\boldsymbol{\mu}}(\hat{r}_{\text{opt}})$ with the estimators $\bar{\boldsymbol{X}}$, $\hat{\boldsymbol{\mu}}_{JS}$ and $\hat{\boldsymbol{\mu}}(\tilde{r})$. Assume that $G=100$. We consider four different values of n, ranging from 5, 10, 20 to 50, to represent different levels of sample sizes. We draw σ_i^2 from the scaled chi-square distribution $\chi_{n-1}^2/(n-1)$, and μ_i from $N(0,\tau^2)$, where $\tau = 0.2$, 0.6 or 1 representing different levels of mean heterogeneity. For each gene i, we simulate n observations from the normal distribution $N(\mu_i, \sigma_i^2)$. We repeat the process 5000 times for each setting and report the simulated average risk $\text{AR} = \sum_{k=1}^{5000} L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}, D)/5000$ in Table 1 for the four estimators, respectively.

Under the quadratic loss function (3), the sample mean $\bar{\boldsymbol{X}}$ has a constant risk at 1, as reported in the simulations. The standard errors of these average risks are all around $0.14/\sqrt{5000} = 0.0020$. Therefore, the improvements of the three shrinkage estimators over $\bar{\boldsymbol{X}}$ are all statistically significant. We observe that $\hat{\boldsymbol{\mu}}(\hat{r}_{\text{opt}})$ has a smaller average risk than both $\hat{\boldsymbol{\mu}}_{JS}$ and $\hat{\boldsymbol{\mu}}(\tilde{r})$ in most settings, especially when the sample size is small. In addition, the improvement of $\hat{\boldsymbol{\mu}}(\hat{r}_{\text{opt}})$ over $\bar{\boldsymbol{X}}$ increases when the mean heterogeneity decreases, especially when τ is near zero. We also observe that the improvements of the shrinkage estimators over the sample mean become smaller when n becomes larger. This indicates that for the large sample size scenario, it is no longer necessary to borrow information from other genes to improving the mean estimation.

Note that we have assumed the grand mean to be zero in the above simulation. Note that when the grand mean has a shift from zero, the term $\sum_{i=1}^{G} \mu_i^2/\sigma_i^2$ will tend to be larger so that the improvement of $\hat{\boldsymbol{\mu}}(\hat{r}_{\text{opt}})$ over $\bar{\boldsymbol{X}}$ will be diminished correspondingly. In such situations, Lindley (1962) suggested to apply the shrinkage method to the deviations $X_{ij} - \bar{X}_{\cdot\cdot}$ rather than to the original observations X_{ij}, which leads to the following variation of the estimator (5),

$$\tilde{\boldsymbol{\mu}}_{\hat{r}_{\text{opt}}} = \bar{\boldsymbol{X}}_{\cdot\cdot} + \left(1 - \frac{\hat{r}_{\text{opt}}}{\|\boldsymbol{X} - \bar{\boldsymbol{X}}_{\cdot\cdot}\|_S^2}\right)(\bar{\boldsymbol{X}} - \bar{\boldsymbol{X}}_{\cdot\cdot}), \tag{6}$$

Table 1. Average risks of the estimators under various settings

τ		$n=5$	$n=10$	$n=20$	$n=50$
0.2	$\bar{\boldsymbol{X}}$	0.997	1.001	0.996	1.000
	$\hat{\boldsymbol{\mu}}_{JS}$	0.337	0.362	0.485	0.682
	$\hat{\boldsymbol{\mu}}(\tilde{r})$	0.514	0.408	0.493	0.683
	$\hat{\boldsymbol{\mu}}(\hat{r}_{\text{opt}})$	0.339	0.359	0.483	0.682
0.6	$\bar{\boldsymbol{X}}$	1.004	1.001	1.003	0.999
	$\hat{\boldsymbol{\mu}}_{JS}$	0.889	0.836	0.897	0.951
	$\hat{\boldsymbol{\mu}}(\tilde{r})$	0.851	0.840	0.898	0.951
	$\hat{\boldsymbol{\mu}}(\hat{r}_{\text{opt}})$	0.801	0.827	0.895	0.951
1	$\bar{\boldsymbol{X}}$	0.998	0.997	0.998	1.001
	$\hat{\boldsymbol{\mu}}_{JS}$	0.973	0.932	0.957	0.984
	$\hat{\boldsymbol{\mu}}(\tilde{r})$	0.933	0.932	0.957	0.984
	$\hat{\boldsymbol{\mu}}(\hat{r}_{\text{opt}})$	0.912	0.927	0.956	0.983

Table 2. Average risks of the estimators (5) and (6) under various settings

μ_0		$n=5$	$n=10$	$n=20$	$n=50$
0	$\hat{\boldsymbol{\mu}}(\hat{r}_{\text{opt}})$	0.732	0.770	0.849	0.926
	$\tilde{\boldsymbol{\mu}}(\hat{r}_{\text{opt}})$	0.736	0.772	0.850	0.926
1	$\hat{\boldsymbol{\mu}}(\hat{r}_{\text{opt}})$	0.928	0.940	0.966	0.982
	$\tilde{\boldsymbol{\mu}}(\hat{r}_{\text{opt}})$	0.737	0.771	0.852	0.928
2	$\hat{\boldsymbol{\mu}}(\hat{r}_{\text{opt}})$	0.977	0.982	0.993	0.993
	$\tilde{\boldsymbol{\mu}}(\hat{r}_{\text{opt}})$	0.738	0.773	0.853	0.928

where $\bar{X}_{\cdot\cdot} = \sum_{i=1}^{G}\sum_{j=1}^{n} X_{ij}/(nG)$ is the grand sample mean across all the genes, and $\bar{\boldsymbol{X}}_{\cdot\cdot} = (\bar{X}_{\cdot\cdot}, \ldots, \bar{X}_{\cdot\cdot})^T$ is a vector of size G with common values.

To evaluate the performance of the Lindley-type estimator (6), we simulate μ_i from $N(\mu_0, 0.5^2)$ for three different values of μ_0: 0, 1 and 2. All other settings remain the same as before. We repeat 5000 simulations for each setting and report the average risks of both $\hat{\boldsymbol{\mu}}(\hat{r}_{\text{opt}})$ and $\tilde{\boldsymbol{\mu}}(\hat{r}_{\text{opt}})$ in Table 2. We observe that the two shrinkage estimators perform similarly when $\mu_0 = 0$. When μ_0 is away from zero, $\hat{\boldsymbol{\mu}}(\hat{r}_{\text{opt}})$ may perform unsatisfactory while the performance of $\tilde{\boldsymbol{\mu}}(\hat{r}_{\text{opt}})$ remains the same. This indicates that the Lindley-type estimator is robust to the shift of the grand mean. In the remainder of the article, the estimator (6) will be adopted to estimate the mean value unless otherwise specified.

3 IMPROVED DIAGONAL DISCRIMINANT ANALYSIS

As mentioned in Section 1, class prediction with high-dimensional data is an important problem in high dimensional data analysis. The objective of class prediction is to assign a new observation to one of the K classes based on its given profile. For ease of notation, we define the class labels to be integers ranging from 1 to K with n_k observations belonging to class k,

$$\boldsymbol{X}_{k,1}, \ldots, \boldsymbol{X}_{k,n_k} \overset{\text{i.i.d.}}{\sim} \text{MVN}_G(\boldsymbol{\mu}_k, \Sigma_k), \quad k = 1, \ldots, K,$$

where $\boldsymbol{\mu}_k$ and Σ_k are the mean value and covariance matrix of the G-dimensional multivariate normal distribution for class

k, respectively. Let $N=n_1+\cdots+n_K$ be the total number of observations.

For high-dimensional data such as microarrays, it is common that the dimension is much larger than the sample size. As a consequence, traditional classification methods such as the linear discriminant analysis are likely to be inapplicable to such high-dimensional data directly. For instance, the sample covariance matrix is singular when G is larger than N. To overcome the singularity problem, Dudoit *et al.* (2002) introduced DLDA and DQDA that ignored the correlations among genes. Specifically, under the assumption that $\Sigma_k=\Sigma$ for all k, DLDA classifies a new observation $\boldsymbol{y}=(y_1,\ldots,y_G)$ to class k which minimizes the following discriminant score

$$\begin{aligned}\hat{d}_k(\boldsymbol{y}) &= (\boldsymbol{y}-\bar{\boldsymbol{X}}_k)^T[\mathrm{diag}(\hat{\Sigma})]^{-1}(\boldsymbol{y}-\bar{\boldsymbol{X}}_k)-2\log\hat{\pi}_k \qquad (7)\\ &= \sum_{i=1}^{G}(y_i-X_{ki})^2/\hat{\sigma}_i^2-2\log\hat{\pi}_k, \qquad k=1,\ldots,K,\end{aligned}$$

where $\boldsymbol{X}_k=\sum_{j=1}^{n_k}\boldsymbol{X}_{k,j}/n_k=(X_{k1},\ldots,X_{kG})$ is the sample mean of class k, $\hat{\Sigma}=\sum_{k=1}^{K}(n_k-1)\hat{\Sigma}_k/(N-K)=\mathrm{diag}(\hat{\sigma}_1^2,\ldots,\hat{\sigma}_G^2)$ is the pooled sample covariance matrix with $\hat{\Sigma}_k=\sum_{j=1}^{n_k}(\boldsymbol{X}_{k,j}-\boldsymbol{X}_k)(\boldsymbol{X}_{k,j}-\boldsymbol{X}_k)^T/(n_k-1)$, and $\hat{\pi}_k=n_k/N$ are the prior probabilities of which the next new observation is coming from class k. DLDA is also called a 'naive Bayes' classifier as it arises in a Bayesian setting (Bickel and Levina, 2004). It has been widely adopted to analyze high-dimensional data in the real sciences. See for example Speed (2003), Noushath *et al.* (2006), Asyali *et al.* (2006) and Heilemann and Schuhr (2008), among others.

Note that DLDA uses the sample mean. We propose here a modified version of DLDA by replacing the sample mean $\bar{\boldsymbol{X}}_k$ in formula (7) by the proposed optimal shrinkage estimates $\tilde{\boldsymbol{\mu}}_k(\hat{r}_{\mathrm{opt}})$ for each class k. We then classify a new observation $\boldsymbol{y}$ to class $\mathrm{argmin}_k\tilde{d}_k^S(\boldsymbol{y})$, where

$$\tilde{d}_k^S(\boldsymbol{y})=(\boldsymbol{y}-\tilde{\boldsymbol{\mu}}_k(\hat{r}_{\mathrm{opt}}))^T[\mathrm{diag}(\hat{\Sigma})]^{-1}(\boldsymbol{y}-\tilde{\boldsymbol{\mu}}_k(\hat{r}_{\mathrm{opt}}))-2\log\hat{\pi}_k. \quad (8)$$

We refer to it as shrinkage-mean-based DLDA (SmDLDA). Similarly, we can propose a shrinkage-mean-based version for DQDA as well. The behavior of the proposed SmDLDA will be studied in the next section under various scenarios.

3.1 Simulated study

In this section, we conduct simulations to compare the performance of SmDLDA with DLDA. We consider a binary classification with simulate data from multivariate normal distributions MVN($\boldsymbol{\mu}_1,\Sigma$) and MVN($\boldsymbol{\mu}_2,\Sigma$), respectively.

The first simulation study considers an identity covariance matrix, i.e. $\Sigma=I_G$. To differentiate the two classes, without loss of generality we assume the first d, $0\le d\le G$, components of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the same and the left ones are different. Specifically, we let $\boldsymbol{\mu}_1=\{0,\ldots,0,\mu_{1,d+1},\ldots,\mu_{1G}\}$ and $\boldsymbol{\mu}_2=-\boldsymbol{\mu}_1$, where $\{\mu_{1,d+1},\ldots,\mu_{1G}\}$ is a random sample of size $G-d$ from the uniform distribution $U(0,0.5)$. We consider two choices of G (50 and 200), and for each G, we consider six different d values at 0, $0.1\times G$, …, $0.5\times G$, respectively. For each simulation, we generate a training set of size n_k for each class k under the simulation setting described above, and a test set of size $5n_k$ under the same setting to assess the misclassification rate. The overall misclassification rate is calculated by the percentage of misclassified observations over the total number of observations in the test sets. Finally, we repeat the procedure 1000 times and report the average misclassification rates in Figure 1 for $n_1=n_2=10$ and 20, respectively. It is evident that SmDLDA has a smaller misclassification rate than DLDA in all settings.

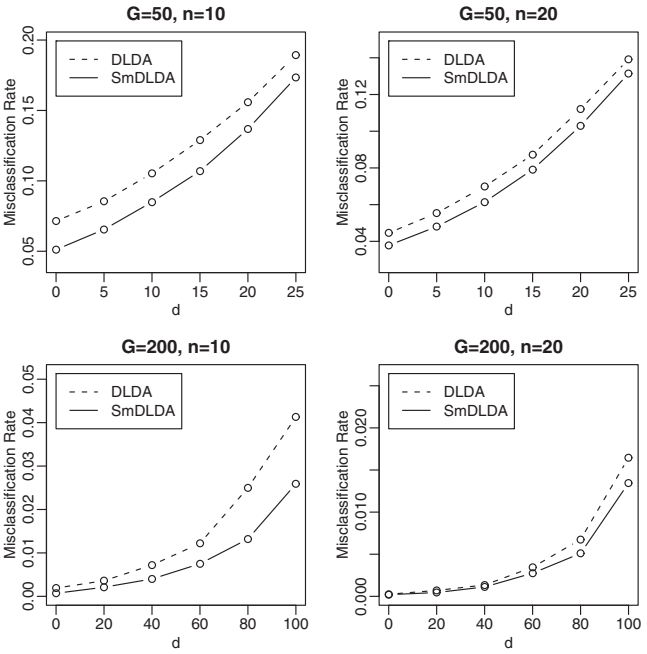

Fig. 1. Plots of the average misclassification rates for DLDA and SmDLDA when the observations are independent.

To evaluate and compare the performance of DLDA and SmDLDA under more realistic situations, we conduct here another simulation study in the case where the observations are correlated. Similarly as in Guo *et al.* (2007) and Pang *et al.* (2009), we use the following block diagonal structure to mimic the true covariance matrix,

$$\Sigma=\begin{pmatrix}\Sigma_\rho & 0 & \cdots & \cdots & \vdots\\ 0 & \Sigma_{-\rho} & 0 & \cdots & \vdots\\ \vdots & 0 & \Sigma_\rho & 0 & \vdots\\ \vdots & \vdots & 0 & \Sigma_{-\rho} & \vdots\\ \cdots & \cdots & \cdots & \cdots & \cdots\end{pmatrix}_{G\times G},$$

where each diagonal block has the following auto-regressive structure

$$\Sigma_\rho=\begin{pmatrix}1 & \rho & \cdots & \rho^9\\ \rho & 1 & \cdots & \rho^8\\ \vdots & \vdots & \ddots & \vdots\\ \rho^9 & \cdots & \rho & 1\end{pmatrix}_{10\times 10}.$$

We consider five different values of the correlation coefficient ρ, ranging between 0, 0.2, 0.4, 0.6 and 0.8, to represent different levels of dependence. Note that $\rho=0$ corresponds to the independent situation in the previous simulation. All other settings are the same as before except that we set $d=0.1\times G$ in this new simulation. We repeat the procedure 1000 times and report the average

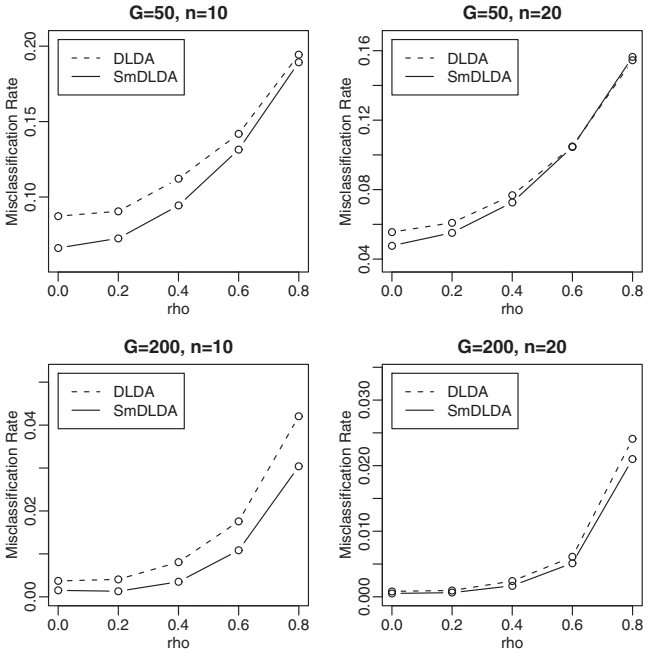

Fig. 2. Plots of the average misclassification rates for DLDA and SmDLDA when the observations are correlated.

misclassification rates in Figure 2 for both DLDA and SmDLDA. Once again, SmDLDA outperforms DLDA in most situations. The comparison result is more evident when the correlation coefficient ρ is not large.

Though we have restricted to a balanced binary classification with $n_1 = n_2$ due to the page limitation, extensive simulations (not shown) indicate that the above comparative conclusions remain the same for unbalanced designs as well as for other simulation settings, including the multiclass comparison problems.

4 APPLICATION TO LEUKEMIA DATA

In this section, we apply the proposed discriminant rule, SmDLDA, to the leukemia data of Golub *et al.* (1999). The dataset is available in the website of www.bioconductor.org. By following the same pre-processing steps (thresholding, filtering and logarithm transformation) as described in Dudoit *et al.* (2002), we end up with a gene expression dataset with a total of 3571 genes for 47 acute lymphoblastic leukemia (ALL) patients and 25 acute myeloid leukemia (AML) patients. We further standardize the dataset so that each array has mean zero and variance one across genes.

Note that in general, the shrinkage methods work well for dense data rather than sparse data. Thus, to better reveal the practical performance of SmDLDA we perform a preliminary screen of informative features before the case study. Different screen methods are available in the literature so as to identify biologically significant gene functional groups or pathways via Gene Ontology annotations (Pan, 2006; Tai and Pan, 2007). In this study for simplicity, we will not carry out the screen by integrating the real biological knowledge, but instead, will perform it based on the ratio of the between-group to within-group sums of squares as in Dudoit *et al.* (2002). Note that the gene selection can also be based on other proposals, see for

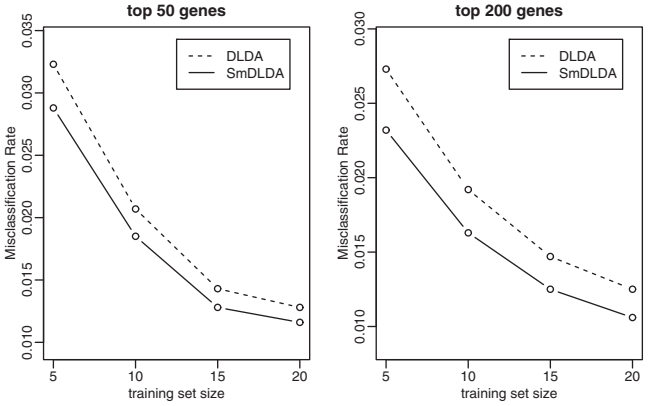

Fig. 3. Plots of the average misclassification rates for DLDA and SmDLDA using leukemia data.

example, Bayesian variable selection (Lee *et al.*, 2003), analysis of variance (Draghici *et al.*, 2003) and independent component analysis (Calò *et al.*, 2005). Let ALL be class 1 and AML be class 2. The ratio for gene j is given as

$$\mathrm{BW}(j) = \frac{\sum_{k=1}^{2}\sum_{i=1}^{n_k}(\bar{X}_{k.j} - \bar{X}_{..j})^2}{\sum_{k=1}^{2}\sum_{i=1}^{n_k}(X_{kij} - \bar{X}_{..j})^2},$$

where $\bar{X}_{..j}$ is the averaged expression values across all samples and $\bar{X}_{k.j}$ is that across samples belonging to class k. We select the top G genes (50 and 200) with the largest BW ratios for further study.

To assess the misclassification rates for both SmDLDA and DLDA, we randomly divide the total 72 samples into training sets and test sets. We let the training set size for each class ranging from 5, 10, 15 to 20, respectively. The remaining samples are then used as the test sets. Recall that the improvement of shrinkage is inversely proportional to the mean heterogeneity (Section 4). To better improve the performance of the shrinkage-based rule, we propose an adaptive procedure that aims to reduce the possibly large level of mean heterogeneity that may appear in real data. Specifically, we shrink the distances between the class centroids and the overall centroids rather than to shrink the class centroids directly. This is a similar idea as in Tibshirani *et al.* (2003). Finally, for each setting, we repeat the procedure 1000 times and report the average misclassification rates in Figure 3. Similarly as in the simulation studies, it is evident again that SmDLDA outperforms DLDA in all settings.

5 DISCUSSION

In this article, we proposed an optimal shrinkage estimator for the mean value under the 'large G small n' scenario. We then applied the proposed shrinkage estimator to high-dimensional classification problem by constructing a shrinkage-based diagonal discriminant rule. Its improvement over the original competitor was demonstrated through both simulations and real data analysis.

Though the independence assumption in this article is popular in the literature (Bickel and Levina, 2004; Dudoit *et al.*, 2002; Hwang *et al.*, 2009; Tong and Wang, 2007), it is unlikely to be true in practice and so certain remedy might be necessary for a further improvement when additional information is available. Langaas

et al. (2005) suggested that the clumpy dependence is a likely form of dependence, where the clumpy dependence means that the genes are dependent within groups and independent among groups. Inspired by that, one natural extension would be to propose new shrinkage estimators for the mean value under the clumpy dependence structure. To avoid the singularity problem, we might need to assume that the largest group size is not larger than the number of samples. Another future work is to examine if the proposed optimal shrinkage estimator has any good in its own right, or if it can be further improved by its positive-part estimator.

Finally, we note that the proposed SmDLDA in this article is a shrinkage-mean-only-based DLDA, whereas in Pang *et al*. (2009) the authors proposed a shrinkage-variance-only-based DLDA. As both the mean and variance estimations are crucial in the statistical analysis, further research might be needed to develop new classification rules that shrink both the mean value and the variance. Possible approaches can be either by plugging-in the existing shrinkage estimators, respectively, or by proposing new shrinkage estimators for the mean value and variances simultaneously.

ACKNOWLEDGEMENTS

The authors are grateful to the editor, the associate editor and three referees for their constructive comments and suggestions that have led to a substantial improvement in the article. The authors are also grateful to Professor J.T. Gene Hwang for helpful discussions.

Funding: Hong Kong RGC grant (HKBU202711) and Hong Kong Baptist University FRG grant (FRG2/10-11/020) to T.T.; AFAR research grant and National Institutes of Health grant (P50HG002790) to L.C; National Institutes of Health grant (GM59507) and NSF grant (DMS0714817) to H.Z.

Conflict of Interest: none declared.

REFERENCES

Assani,I. (1997) Strong laws for weighted sums of independent identically distributed random variables. *Duke Math. J.*, **88**, 217–246.

Asyali,M.H. *et al*. (2006) Gene expression profile classification: a review. *Curr. Bioinformatics*, **1**, 55–73.

Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.

Baranchik,A.J. (1970) A family of minimax estimators of the mean of a multivariate normal distribution. *Ann. Math. Stat.*, **41**, 642–645.

Berger,J.O. and Bock,M.E. (1976) Combining independent normal mean estimation problems with unknown variances. *Ann. Stat.*, **4**, 642–648.

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

Calò,D.G. *et al*. (2005) Variable selection in classification problems: a strategy based on independent component analysis. In Vichi,M. *et al.* (eds) *New Developments in Classification and Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, Berlin, pp. 21–30.

Cui,X. *et al*. (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, **6**, 59–75.

Dettling,M. (2005) Bagboosting for tumor classification with gene expression data. *Bioinformatics*, **20**, 3583–3593.

Draghici,S. *et al*. (2003) Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays. *Bioinformatics*, **19**, 1348–1359.

Dudoit,S. *et al*. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.

Efron,B. and Morris,C. (1973) Stein's estimation rule and its competitors - an empirical Bayes approach. *J. Am. Stat. Assoc.*, **68**, 117–130.

Fourdrinier,D. *et al*. (2003) Robust shrinkage estimation for elliptically symmetric distributions with unknown covariance matrix. *J. Multivar. Anal.*, **85**, 24–39.

Gelman,A. *et al*. (2004) *Bayesian Data Analysis*, 2nd edn. Chapman and Hall, London.

Gleser,L.J. (1986) Minimax estimators of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. *Ann. Stat.*, **14**, 1625–1633.

Golub,T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Guo,Y. *et al.* (2007) Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, **8**, 86–100.

Hausser,J. and Strimmer,K. (2009) Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn. Res.*, **10**, 1469–1484.

Heilemann,U. and Schuhr,R. (2008) On the evolution of german business cycles 1958–2004. *J. Econ. Stat.*, **228**, 84–109.

Huang,S. *et al*. (2010) Bias-corrected diagonal discriminant rules for high-dimensional classification. *Biometrics*, **66**, 1096–1106.

Hwang,J.T.G. and Liu,P. (2010) Optimal tests shrinking both means and variances applicable to microarray data analysis. *Stat. Appl. Genet. Mol. Biol.*, **9**, 36.

Hwang,J.T.G. *et al*. (2009) Empirical Bayes confidence intervals shrinking both means and variances. *J. R. Stat. Soc. Ser. B*, **71**, 265–285.

James,W. and Stein,C. (1961) Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Stat. Probab.*, **1**, 361–379.

Langaas,M. *et al*. (2005) Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. Ser. B*, **67**, 555–572.

Lee,K.E. *et al*. (2003) Gene Selection: a Bayesian variable selection approach. *Bioinformatics*, **19**, 90–97.

Lee,J.W. *et al*. (2005) An extensive comparison of recent classification tools applied to microarray data. *Comput. Stat. Data Anal.*, **48**, 869–885.

Lin,P.E. and Tsai,H.L. (1973) Generalized Bayes minimax estimators of the multivariate normal mean with unknown covariance matrix. *Ann. Stat.*, **1**, 142–145.

Lindley,D.V. (1962) Discussion of professor Stein's paper: confidence sets for the mean of a multivariate normal distribution. *J. R. Stat. Soc. Ser. B*, **24**, 285–287.

Montazeri,Z. *et al*. (2010) Shrinkage estimation of effect sizes as an alternative to hypothesis testing followed by estimation in high-dimensional biology: Applications to differential gene expression. *Stat. Appl. Genet. Mol. Biol.*, **9**, 23.

Noushath,S. *et al*. (2006) Diagonal Fisher linear discriminant analysis for efficient face recognition. *Neurocomputing*, **69**, 1711–1716.

Opgen-Rhein,R. and Strimmer,K. (2007) Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat. Appl. Genet. Mol. Biol.*, **6**, 9.

Pan,W. (2006) Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics*, **22**, 795–801.

Pang,H. *et al.* (2009) Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data. *Biometrics*, **65**, 1021–1029.

Pang,H. *et al*. (2010) Analyzing breast cancer microarrays from african americans using shrinkage-based discriminant analysis. *Hum. Genomics*, **5**, 5–16.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiment. *Stat. Appl. Genet. Mol. Biol.*, **3**, 1.

Speed,R. (2003) *Statistical Analysis of Gene Expression Microarray Data*. Chapman and Hall, London.

Storey,J.D. and Tibshirani,R. (2003) SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In Parmigiani,G. *et al.* (eds) *The Analysis of Gene Expression Data: Methods and Software*. Springer, New York.

Tai,F. and Pan,W. (2007) Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. *Bioinformatics*, **23**, 3170–3177.

Tibshirani,R. *et al*. (2003) Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Stat. Sci.*, **18**, 104–117.

Tong,T. and Wang,Y. (2007) Optimal shrinkage estimation of variances with applications to microarray data analysis. *J. Am. Stat. Assoc.*, **102**, 113–122.

Wang,Y. *et al*. (2009) Variance estimation in the analysis of microarray data. *J. R. Stat. Soc. Ser. B*, **71**, 425–445.

Wright,G.W. and Simon,R.M. (2003) A random variance model for detection of differential gene expression in small microarray experiments, *Bioinformatics*, **19**, 2448–2455.

APPENDIX A

A1. DERIVATION OF FORMULA (4)

Noting that $\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}=(\bar{\boldsymbol{X}}-\boldsymbol{\mu})-r\bar{\boldsymbol{X}}/\|\bar{\boldsymbol{X}}\|_S^2$, the risk function is

$$\begin{aligned}
R(\hat{\boldsymbol{\mu}},\boldsymbol{\mu},D) &= \frac{n}{G}E(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^T D^{-1}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) \\
&= \frac{n}{G}E\Big[(\boldsymbol{X}-\boldsymbol{\mu})^T D^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \\
&\quad +\frac{r^2}{(\|\bar{\boldsymbol{X}}\|_S^2)^2}\boldsymbol{X}^T D^{-1}\boldsymbol{X} \\
&\quad -\frac{r}{\|\bar{\boldsymbol{X}}\|_S^2}(\bar{\boldsymbol{X}}-\boldsymbol{\mu})^T D^{-1}\bar{\boldsymbol{X}} \\
&\quad -\frac{r}{\|\bar{\boldsymbol{X}}\|_S^2}\bar{\boldsymbol{X}}^T D^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu})\Big] \\
&= 1+\frac{nr^2}{G}E\left[\frac{\|\bar{\boldsymbol{X}}\|_D^2}{(\|\bar{\boldsymbol{X}}\|_S^2)^2}\right] \\
&\quad -\frac{2nr}{G}\sum_{i=1}^{G}E\left[\frac{1}{\|\bar{\boldsymbol{X}}\|_S^2}\frac{\bar{X}_i(\bar{X}_i-\mu_i)}{\sigma_i^2}\right].
\end{aligned}$$

By Stein formula, $\mathrm{E}g(X)(X-\mu)=\sigma^2\mathrm{E}g'(X)$ where $X\sim N(\mu,\sigma^2)$, we have

$$\begin{aligned}
E\left[\frac{1}{\|\bar{\boldsymbol{X}}\|_S^2}\frac{X_i(X_i-\mu_i)}{\sigma_i^2}\right] &= \frac{1}{n}E\left[\left(\frac{X_i}{\|\bar{\boldsymbol{X}}\|_S^2}\right)\frac{X_i-\mu_i}{\sigma_i^2/n}\right] \\
&= \frac{1}{n}E\left[\frac{\partial}{\partial\bar{X}_i}\left(\frac{\bar{X}_i}{\|\bar{\boldsymbol{X}}\|_S^2}\right)\right] \\
&= \frac{1}{n}E\left[\frac{1}{\|\bar{\boldsymbol{X}}\|_S^2}-\frac{2\bar{X}_i^2}{(\|\boldsymbol{X}\|_S^2)^2 s_i^2}\right].
\end{aligned}$$

This leads to

$$R(\hat{\boldsymbol{\mu}},\boldsymbol{\mu},D)=\frac{nr^2}{G}E\left[\frac{\|\bar{\boldsymbol{X}}\|_D^2}{(\|\bar{\boldsymbol{X}}\|_S^2)^2}\right]-\frac{2r(G-2)}{G}E\left(\frac{1}{\|\bar{\boldsymbol{X}}\|_S^2}\right).$$

Finally, by minimizing the above quantity, we have

$$r_{\mathrm{opt}}=\frac{(G-2)E(1/\|\bar{\boldsymbol{X}}\|_S^2)}{nE\left[\|\bar{\boldsymbol{X}}\|_D^2/(\|\bar{\boldsymbol{X}}\|_S^2)^2\right]}.$$

APPENDIX B

B1. PROOF OF LEMMA

Noting that $f(x)=1/x$ is a continuous function in $(0,\infty)$, it suffices to show that $\|\bar{\boldsymbol{X}}\|_S^2/\|\bar{\boldsymbol{X}}\|_D^2\xrightarrow{a.s.}\rho_n$ as $G\to\infty$. By Assani (1997), the following strong law holds under the condition $\sup_m(N_m/m)<\infty$ almost surely,

$$\frac{1}{A_G}\sum_{i=1}^{G}a_i[Z_i-\mathrm{E}(Z_i)]\xrightarrow{a.s.}0,\qquad \text{as}\quad G\to\infty.$$

Further, we have $\sum_{i=1}^{G}a_iZ_i/A_G\xrightarrow{a.s.}\mathrm{E}Z_1=1/(n-3)$ for any $n\geq 4$. This proves the lemma by noting that $A_G=\|\bar{\boldsymbol{X}}\|_D^2$ and $\|\bar{\boldsymbol{X}}\|_S^2=(n-1)\sum_{i=1}^{G}a_iZ_i$.