2

Effective Methodologies for Statistical Inference on Microarray Studies

Makoto Aoshima and Kazuyoshi Yata
Institute of Mathematics, University of Tsukuba, Ibaraki
Japan

1. Introduction

A common feature of high-dimensional data such as genetic microarrays is that the data dimension is extremely high, however the sample size is relatively small. This type of data is called the high-dimension, low-sample-size (HDLSS) data. Such HDLSS data present with substantial challenges to reconsider existing methods in the multivariate statistical analysis. Unfortunately, it has been known that most conventional methods break down in HDLSS situations and alternative methods are often highly sensitive to the curse of dimensionality.
In this chapter, we present modern statistical methodologies that are very effective to draw statistical inference from HDLSS data. We focus on a series of effective HDLSS methodologies developed by Aoshima and Yata (2011) and Yata and Aoshima (2009, 2010a,b, 2011a,b). We demonstrate how those methodologies perform well and bring a new insight into researches on prostate cancer.
In Section 2, we first consider Principal Component Analysis (PCA) for microarray data to visualize a data structure having tens of thousands of dimension by projecting on a few dimensional PC space. We note that classical PCA cannot sufficiently visualize a latent structure of microarray data because of the curse of dimensionality. We overcome the difficulty with the help of the *cross-data-matrix (CDM) methodology* that was developed by Yata and Aoshima (2010a,b).
Next, in Section 3, we consider an effective clustering for microarray data. We apply the CDM methodology to estimating the principal component (PC) scores. We show that a clustering method given by using a CDM-based first PC score effectively classifies individuals into two groups. We demonstrate accurate clustering by using prostate cancer data given by Singh et al. (2002).
Further, in Section 4, we consider an effective classification for microarray data. We pay special attention to the quadratic-type classification methodology developed by Aoshima and Yata (2011). We give a sample size determination for the classification so that the misclassification rates are controlled by a prespecified upper bound. We examine how the classification methodology performs well by using some microarray data sets.
Finally, in Section 5, we consider a variable selection procedure to select a set of significant variables from microarray data. In most gene expression studies, it is important to select relevant genes for classification so that researchers can identify the smallest possible set of genes that can still achieve good predictive performance. We implement the two-stage

variable selection procedure, developed by Aoshima and Yata (2011), that provides screening of variables in the first stage. We select a significant set of associated variables from among a set of candidate variables in the second stage. We show that the selection procedure assures a high accuracy by eliminating redundant variables. We identify predictive genes to classify patients according to disease outcomes on prostate cancer.

2. PCA for high-dimension, low-sample-size data

Suppose we have a $p \times n$ data matrix $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$ with $p > n$, where $\boldsymbol{x}_k = (x_{1k}, ..., x_{pk})^T$, $k = 1, ..., n$, are independent and identically distributed as a p-dimensional distribution having mean $\boldsymbol{\mu}$ and positive-definite covariance matrix $\boldsymbol{\Sigma}$. The eigen-decomposition of $\boldsymbol{\Sigma}$ is given by $\boldsymbol{\Sigma} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^T$, where $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p (> 0)$ and $\boldsymbol{H} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_p]$ is a matrix of corresponding eigenvectors. Then, $\boldsymbol{Z} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{H}^T(\boldsymbol{X} - [\boldsymbol{\mu}, ..., \boldsymbol{\mu}])$ is considered as a $p \times n$ sphered data matrix from a distribution with zero mean and the identity covariance matrix. Here, we write $\boldsymbol{Z} = [\boldsymbol{z}_1, ..., \boldsymbol{z}_p]^T$ and $\boldsymbol{z}_j = (z_{j1}, ..., z_{jn})^T$, $j = 1, ..., p$. We assume that the fourth moments of each variable in $\boldsymbol{Z}$ are uniformly bounded and $||\boldsymbol{z}_j|| \neq 0$ for $j = 1, ..., p$, where $||\cdot||$ denotes the Euclidean norm. We note that the multivariate distribution assumed here does not have to be a normal distribution, $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and the random variables in $\boldsymbol{Z}$ do not have to be regulated by a ρ-mixing condition. We consider a general setting as follows:

$$\lambda_j = a_j p^{\alpha_j} \ (j = 1, ..., m) \quad \text{and} \quad \lambda_j = c_j \ (j = m+1, ..., p). \tag{1}$$

Here, $a_j (> 0)$, $c_j (> 0)$ and $\alpha_j (\alpha_1 \geq \cdots \geq \alpha_m > 0)$ are unknown constants preserving the ordering that $\lambda_1 \geq \cdots \geq \lambda_p$, and m is an unknown positive integer. The model (1) is an extension of a multi-component model or spiked covariance model given by Johnstone and Lu (2009). This is a quite general model for high-dimensional data. For example, a mixture model given by (6) in Section 3 is one of the examples that have the model (1) as in (7). One would also find the model (1) in a highly-correlated, high-dimensional data analysis such as graphical models, high dimensional regression models, and so on.

Let $\boldsymbol{X}_o = \boldsymbol{X} - [\overline{\boldsymbol{x}}, ..., \overline{\boldsymbol{x}}]$, where $\overline{\boldsymbol{x}} = \sum_{i=1}^n \boldsymbol{x}_i / n$. The sample covariance matrix is given by $\boldsymbol{S} = (n-1)^{-1}\boldsymbol{X}_o\boldsymbol{X}_o^T$ and its dual matrix is defined by $\boldsymbol{S}_D = (n-1)^{-1}\boldsymbol{X}_o^T\boldsymbol{X}_o$. Note that $\boldsymbol{S}_D$ and $\boldsymbol{S}$ share non-zero eigenvalues. Let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_{n-1} (\geq 0)$ be the eigenvalues of $\boldsymbol{S}_D$. Let us write the eigen-decomposition of $\boldsymbol{S}_D$ by $\boldsymbol{S}_D = \sum_{j=1}^{n-1} \hat{\lambda}_j \hat{\boldsymbol{u}}_j \hat{\boldsymbol{u}}_j^T$, where $\hat{\boldsymbol{u}}_j$'s are the corresponding eigenvectors of $\hat{\lambda}_j$ such that $||\hat{\boldsymbol{u}}_j|| = 1$ and $\hat{\boldsymbol{u}}_i^T \hat{\boldsymbol{u}}_j = 0$ $(i \neq j)$.

2.1 Naive PCA in HDLSS situations

Yata and Aoshima (2009) gave sufficient conditions to claim the consistency property for the sample eigenvalues: For $j = 1, ..., m$, it holds that

$$\frac{\hat{\lambda}_j}{\lambda_j} \xrightarrow{p} 1 \tag{2}$$

under the conditions:

(YA-i) $p \to \infty$ and $n \to \infty$ for j such that $\alpha_j > 1$;

(YA-ii) $p \to \infty$ and $p^{2-2\alpha_j}/n \to 0$ for j such that $\alpha_j \in (0, 1]$.

Here, $\xrightarrow{p}$ denotes the convergence in probability. If z_{jk}, $j = 1, ..., p$ $(k = 1, ..., n)$ are independent, the above conditions are improved by the necessary and sufficient conditions as follows:

(YA-i') $p \to \infty$ and $n \to \infty$ for j such that $\alpha_j > 1$;

(YA-ii') $p \to \infty$ and $p^{1-\alpha_j}/n \to 0$ for j such that $\alpha_j \in (0, 1]$.

For the details including the limiting distribution of $\hat{\lambda}_j$, see Yata and Aoshima (2009). If the population distribution is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, one may consider that z_{jk}, $j = 1, ..., p$ $(k = 1, ..., n)$ are independent. When $\alpha_j > 1$, the sample size n is free from p in (YA-i) or (YA-i'). However, when $\alpha_j \in (0, 1]$, one would find difficulty in naive PCA in view of (YA-ii) or (YA-ii') in HDLSS data situations. Let us see a simple case that $p = 10000$, $\lambda_1 = p^{1/2}$ and $\lambda_2 = \cdots = \lambda_p = 1$. Then, we observe from (YA-ii) that it should be $n >> p^{2-2\alpha_1} = p = 10000$. It is somewhat inconvenient for the experimenter to handle PCA in HDLSS data situations.

2.2 Beyond naive PCA

Yata and Aoshima (2010a,b) created an effective methodology called the *cross-data-matrix (CDM) methodology* to handle HDLSS data situations: Let $n_{(1)} = [n/2] + 1$ and $n_{(2)} = n - n_{(1)}$, where $[x]$ denotes the largest integer less than x. Suppose that we have a $p \times n$ data matrix,

$$\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n] = [\boldsymbol{x}_{11}, ..., \boldsymbol{x}_{1n_{(1)}}, \boldsymbol{x}_{21}, ..., \boldsymbol{x}_{2n_{(2)}}]. \tag{3}$$

We define $p \times n_{(i)}$ data matrices, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, by $\boldsymbol{X}_i = [\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_{(i)}}]$, $i = 1, 2$. Note that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent. Let $\boldsymbol{X}_{oi} = \boldsymbol{X}_i - [\overline{\boldsymbol{x}}_i, ..., \overline{\boldsymbol{x}}_i]$, $i = 1, 2$, where $\overline{\boldsymbol{x}}_i = \sum_{j=1}^{n_{(i)}} \boldsymbol{x}_{ij}/n_{(i)}$. We define a cross data matrix by $\boldsymbol{S}_{D(1)} = ((n_{(1)} - 1)(n_{(2)} - 1))^{-1/2} \boldsymbol{X}_{o1}^T \boldsymbol{X}_{o2}$ or $\boldsymbol{S}_{D(2)} = ((n_{(1)} - 1)(n_{(2)} - 1))^{-1/2} \boldsymbol{X}_{o2}^T \boldsymbol{X}_{o1}$ $(= \boldsymbol{S}_{D(1)}^T)$. Note that $\mathrm{rank}(\boldsymbol{S}_{D(1)}) \leq n_{(2)} - 1$. When we consider the singular value decomposition of $\boldsymbol{S}_{D(1)}$, it follows that $\boldsymbol{S}_{D(1)} = \sum_{j=1}^{n_{(2)}-1} \tilde{\lambda}_j \tilde{\boldsymbol{u}}_{j(1)} \tilde{\boldsymbol{u}}_{j(2)}^T$, where $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_{n_{(2)}-1} (\geq 0)$ denote singular values of $\boldsymbol{S}_{D(1)}$, and $\tilde{\boldsymbol{u}}_{j(1)}$ (or $\tilde{\boldsymbol{u}}_{j(2)}$) denotes a unit left- (or right-) singular vector corresponding to $\tilde{\lambda}_j$ $(j = 1, ..., n_{(2)} - 1)$.

[Cross-data-matrix (CDM) methodology]

(Step 1) Define a cross data matrix by $\boldsymbol{S}_{D(1)} = ((n_{(1)} - 1)(n_{(2)} - 1))^{-1/2} \boldsymbol{X}_{o1}^T \boldsymbol{X}_{o2}$.

(Step 2) Calculate the singular values, $\tilde{\lambda}_j$'s, of $\boldsymbol{S}_{D(1)}$ for the estimation of λ_j's.

Note that $\boldsymbol{S}_{D(1)} \boldsymbol{S}_{D(1)}^T = \sum_{j=1}^{n_{(2)}-1} \tilde{\lambda}_j^2 \tilde{\boldsymbol{u}}_{j(1)} \tilde{\boldsymbol{u}}_{j(1)}^T$. Thus one can calculate the singular values, $\tilde{\lambda}_j$'s, by the positive square-root of the eigenvalues of $\boldsymbol{S}_{D(1)} \boldsymbol{S}_{D(1)}^T$. The CDM methodology assures the following properties. For the details, see Yata and Aoshima (2010a,b).

Theorem 2.1. *For $j = 1, ..., m$, it holds that*

$$\frac{\tilde{\lambda}_j}{\lambda_j} \xrightarrow{p} 1 \tag{4}$$

under the conditions:

(i) *$p \to \infty$ and $n \to \infty$ for j such that $\alpha_j > 1/2$;*

(ii) $p \to \infty$ *and* $p^{2-2\alpha_j}/n \to 0$ *for* j *such that* $\alpha_j \in (0, 1/2]$.

Corollary 2.1. *Assume further in Theorem 2.1 that* z_{jk}, $j = 1, ..., p$ $(k = 1, ..., n)$ *are independent. Then, for* $j = 1, ..., m$, *we have (4) under the conditions:*

(i) $p \to \infty$ *and* $n \to \infty$ *for* j *such that* $\alpha_j > 1/2$;

(ii) $p \to \infty$ *and there exists a positive constant* ε_j *satisfying* $p^{1-2\alpha_j}/n < p^{-\varepsilon_j}$ *for* j *such that* $\alpha_j \in (0, 1/2]$.

Theorem 2.2. *Let* $Var(z_{jk}^2) = M_j$ $(< \infty)$ *for* $j = 1, ..., m$ $(k = 1, ..., n)$. *Assume that* λ_j $(j \le m)$ *has multiplicity one. Then, under the conditions (i)-(ii) in Theorem 2.1, it holds for* $j = 1, ..., m$, *that*

$$\sqrt{\frac{n}{M_j}} \left(\frac{\tilde{\lambda}_j}{\lambda_j} - 1 \right) \Rightarrow N(0, 1), \tag{5}$$

where "$\Rightarrow$*" denotes the convergence in distribution and* $N(0, 1)$ *denotes a random variable distributed as the standard normal distribution.*

Corollary 2.2. *Assume further in Theorem 2.2 that* z_{jk}, $j = 1, ..., p$ $(k = 1, ..., n)$ *are independent. Then, for* $j = 1, ..., m$, *we have (5) under the conditions:*

(i) $p \to \infty$ *and* $n \to \infty$ *for* j *such that* $\alpha_j > 1/2$;

(ii) $p \to \infty$ *and* $p^{2-4\alpha_j}/n \to 0$ *for* j *such that* $\alpha_j \in (0, 1/2]$.

Remark 2.1. When the population distribution is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, one has that $M_j = 2$ for $j = 1, ..., p$.

Remark 2.2. The condition (ii) given by Theorem 2.1 (or Theorem 2.2) is a sufficient condition for the case of $\alpha_j \in (0, 1/2]$. If more information is available about the population distribution, the condition (ii) can be relaxed to give consistency under a broader set of (p, n). For example, when the population distribution is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the asymptotic properties are claimed under a broader set of (p, n) given by (ii) of Corollary 2.1 (or Corollary 2.2).

Remark 2.3. In view of Theorem 2.1 compared to (2), the CDM methodology successfully relaxes the condition for the case that $\alpha_j > 1/2$. The conditions given by Theorem 2.1 are not continuous in α_j at $\alpha_j = 1/2$. On the other hand, the conditions given by Corollaries 2.1 and 2.2 are continuous in α_j.
When we apply the CDM methodology, we simply divided $\boldsymbol{X}$ into $\boldsymbol{x}_1, ..., \boldsymbol{x}_{n_{(1)}}$ and $\boldsymbol{x}_{n_{(1)}+1}, ..., \boldsymbol{x}_n$ in (3). In general, there exist ${}_nC_{n_{(1)}}$ ways to divide $\boldsymbol{X}$ into $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. The CDM methodology can be generalized as follows:

[Generalized cross-data-matrix (GCDM) methodology]

(Step 1) Set iteration number T. Set $t = 1$.

(Step 2) Randomly split $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ into $\boldsymbol{X}_1 = [\boldsymbol{x}_{1(1)}, ..., \boldsymbol{x}_{1(n_{(1)})}]$ and $\boldsymbol{X}_2 = [\boldsymbol{x}_{2(1)}, ..., \boldsymbol{x}_{2(n_{(2)})}]$.

(Step 3) Define a cross data matrix by $\boldsymbol{S}_{D(1)t} = ((n_{(1)} - 1)(n_{(2)} - 1))^{-1/2} \boldsymbol{X}_{o1}^T \boldsymbol{X}_{o2}$, where $\boldsymbol{X}_{oi} = \boldsymbol{X}_i - [\overline{\boldsymbol{x}}_i, ..., \overline{\boldsymbol{x}}_i]$, $i = 1, 2$, and $\overline{\boldsymbol{x}}_i = \sum_{j=1}^{n_{(i)}} \boldsymbol{x}_{i(j)}/n_{(i)}$.

(Step 4) Calculate the singular values, $\tilde{\lambda}_{1t} \ge \cdots \ge \tilde{\lambda}_{n_{(2)}-1\,t} (\ge 0)$, of $\boldsymbol{S}_{D(1)t}$.

(Step 5) If $t < T$, put $t = t + 1$ and go to Step 2; otherwise go to Step 6.

(Step 6) Estimate λ_j by $\tilde{\lambda}_{j(T)} = \sum_{t=1}^{T} \tilde{\lambda}_{jt}/T$ for each j.

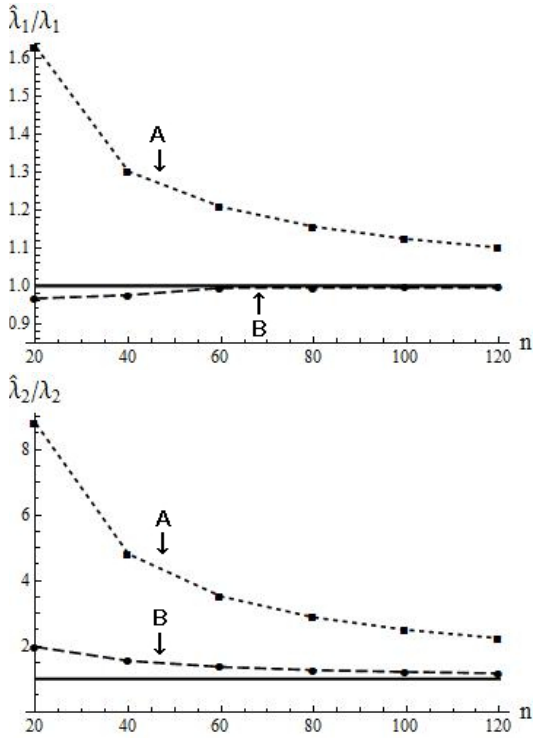

Fig. 1. The behaviors of A: $\hat{\lambda}_j/\lambda_j$ and B: $\tilde{\lambda}_j/\lambda_j$ for the first eigenvalue (upper panel) and second eigenvalue (lower panel) when the samples, of size $n = 20(20)120$, were taken from $N_p(\mathbf{0}, \mathbf{\Sigma})$ with $p = 1600$.

2.3 Performances

We observed that naive PCA requires the sample size n depending on p for $\alpha_i \in (1/2, 1]$ in (2). On the other hand, the CDM methodology allows the experimenter to choose n free from p for the case that $\alpha_i > 1/2$ as in Theorem 2.1 or Corollary 2.1. The CDM methodology might make it possible to give feasible estimation of eigenvalues for HDLSS data with extremely small n compared to p.

We first considered a normal distribution case. Independent pseudorandom normal observations were generated from $N_p(\mathbf{0}, \mathbf{\Sigma})$ with $p = 1600$. We considered $\lambda_1 = p^{2/3}$, $\lambda_2 = p^{1/3}$ and $\lambda_3 = \cdots = \lambda_p = 1$ in (1). We used the sample of size $n = 20(20)120$ to define the data matrix $\boldsymbol{X} : p \times n$ for the calculation of $\boldsymbol{S}_D$ in naive PCA, whereas we divided the sample into $\boldsymbol{X}_1 : p \times n_{(1)}$ and $\boldsymbol{X}_2 : p \times n_{(2)}$ for the calculation of $\boldsymbol{S}_{D(1)}$ in the CDM methodology. The findings were obtained by averaging the outcomes from 1000 $(= R$, say) replications. Under a fixed scenario, suppose that the r-th replication ends with estimates of λ_j, $\hat{\lambda}_{jr}$ and $\tilde{\lambda}_{jr}$ $(r = 1, ..., R)$, given by naive PCA and the CDM methodology. Let us simply write $\hat{\lambda}_j = R^{-1}\sum_{r=1}^{R}\hat{\lambda}_{jr}$ and $\tilde{\lambda}_j = R^{-1}\sum_{r=1}^{R}\tilde{\lambda}_{jr}$. We considered two quantities, A: $\hat{\lambda}_j/\lambda_j$ and B: $\tilde{\lambda}_j/\lambda_j$. Figure 1 shows the behaviors of both A and B for the first two eigenvalues. By observing the behavior of A, naive PCA seems not to give a feasible estimation within

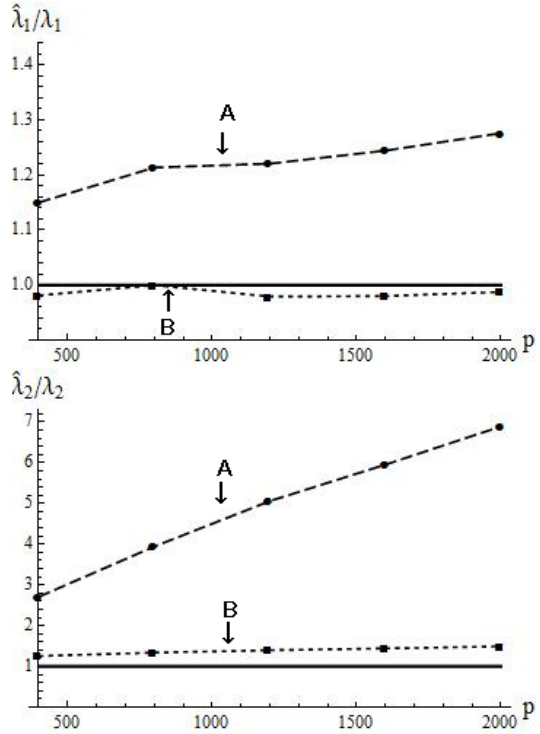

Fig. 2. The behaviors of A: $\hat{\lambda}_j/\lambda_j$ and B: $\tilde{\lambda}_j/\lambda_j$ for the first eigenvalue (upper panel) and second eigenvalue (lower panel) when the samples, of size $n = 60$, were taken from $t_p(\mathbf{0}, \mathbf{\Sigma}, \nu)$ with $\nu = 15$ and $p = 400(400)2000$.

the range of n. The sample size n was not large enough to use the eigenvalues of $\boldsymbol{S}_D$ for such a high-dimensional space. On the other hand, in view of the behavior of B, the CDM methodology gave a reasonable estimation surprisingly well for such HDLSS data sets. The CDM methodology seems to perform excellently as expected theoretically.

Next, we considered a non-normal distribution case. Independent pseudorandom observations were generated from a p-variate t-distribution, $t_p(\mathbf{0}, \mathbf{\Sigma}, \nu)$, with mean zero, covariance matrix $\mathbf{\Sigma}$ and degree of freedom $\nu = 15$. We considered the case that $\lambda_1 = p^{2/3}$, $\lambda_2 = p^{1/3}$ and $\lambda_3 = \cdots = \lambda_p = 1$ in (1) as before. We fixed the sample size as $n = 60$. We set the dimension as $p = 400(400)2000$. Similarly to Figure 1, the findings were obtained by averaging the outcomes from 1000 replications. Figure 2 shows the behaviors of two quantities, A: $\hat{\lambda}_j/\lambda_j$ and B: $\tilde{\lambda}_j/\lambda_j$, for the first two eigenvalues. Again, the CDM methodology seems to perform excellently as expected theoretically. One can observe the consistency of $\tilde{\lambda}_j$ for all $p = 400(400)2000$. We conducted simulation studies for other settings as well and verified the superiority of the CDM methodology to naive PCA in various HDLSS data situations.

3. Clustering for high-dimension, low-sample-size data

Suppose we have a mixture model to classify a data set into two groups. We assume that the observation is sampled with mixing proportions w_j's from two populations, Π_1 and Π_2, and the label of the population is missing. We consider a mixture model whose p.d.f. (or p.f.) is given by

$$f(\boldsymbol{x}) = w_1\pi_1(\boldsymbol{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + w_2\pi_2(\boldsymbol{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2), \tag{6}$$

where w_j's are positive constants such that $w_1 + w_2 = 1$ and $\pi_i(\boldsymbol{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$'s are p-dimensional p.d.f. (or p.f.) of Π_i having mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ be the mean vector and the covariance matrix of the mixture model. Then, we have that $\boldsymbol{\mu} = w_1\boldsymbol{\mu}_1 + w_2\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma} = w_1w_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T + w_1\boldsymbol{\Sigma}_1 + w_2\boldsymbol{\Sigma}_2$. We suppose that $\boldsymbol{x}_k$, $k = 1,...,n$, are independently taken from (6) and define a $p \times n$ data matrix $\boldsymbol{X} = [\boldsymbol{x}_1,...,\boldsymbol{x}_n]$. Let $\Delta = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2$. Let λ_{11} and λ_{21} be the largest eigenvalues of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. We assume that $\Delta = cp^\beta$, where c and β are positive constants. We assume that $\lambda_{11}/\Delta \to 0$ and $\lambda_{21}/\Delta \to 0$ as $p \to \infty$. Then, as for the largest eigenvalue, λ_1, of $\boldsymbol{\Sigma}$ and the corresponding eigenvector, $\boldsymbol{h}_1$, we have that

$$\frac{\lambda_1}{\omega_1\omega_2\Delta} \to 1 \quad \text{and} \quad \text{Angle}(\boldsymbol{h}_1, (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/\Delta^{1/2}) \to 0. \tag{7}$$

We note from (7) that the mixture model given by (6) holds the model (1) about $\boldsymbol{\Sigma}$. Let s_{1k} denote the first principal component (PC) score of $\boldsymbol{x}_k$ $(k = 1,...,n)$. Then, from Yata and Aoshima (2010b), it holds as $p \to \infty$ that

$$\frac{s_{1k}}{\sqrt{\lambda_1}} \xrightarrow{p} \begin{cases} \sqrt{w_2/w_1} & \text{when } \boldsymbol{x}_k \in \Pi_1, \\ -\sqrt{w_1/w_2} & \text{when } \boldsymbol{x}_k \in \Pi_2. \end{cases}$$

Thus one would be able to classify the data set $\{\boldsymbol{x}_1,...,\boldsymbol{x}_n\}$ into two groups if s_{1k} is effectively estimated in HDLSS data situations. In this section hereafter, we borrow symbols from Section 2.

3.1 Effective estimation for PC scores

In general, the j-th PC score of $\boldsymbol{x}_k$ is given by $\boldsymbol{h}_j^T(\boldsymbol{x}_k - \boldsymbol{\mu}) = z_{jk}\sqrt{\lambda_j}$ $(= s_{jk}$, say). Yata and Aoshima (2009) considered a sample eigenvector by $\hat{\boldsymbol{h}}_j = ((n-1)\hat{\lambda}_j)^{-1/2}\boldsymbol{X}_o\hat{\boldsymbol{u}}_j$ and an naive estimator of the j-th PC score of $\boldsymbol{x}_k$ by $\hat{\boldsymbol{h}}_j^T(\boldsymbol{x}_k - \overline{\boldsymbol{x}}) = \hat{u}_{jk}\sqrt{(n-1)\hat{\lambda}_j}$ $(= \hat{s}_{jk}$, say), where $\hat{\boldsymbol{u}}_j^T = (\hat{u}_{j1},...,\hat{u}_{jn})$. Note that $\hat{\boldsymbol{h}}_j$ can be calculated by using a unit-norm eigenvector, $\hat{\boldsymbol{u}}_j$, of $\boldsymbol{S}_D$ whose size is much smaller than $\boldsymbol{S}$ especially for a HDLSS data matrix. Now, we apply the CDM methodology to the PC score in order to improve the naive estimator. Recall that $\tilde{\boldsymbol{u}}_{j(1)}$ (or $\tilde{\boldsymbol{u}}_{j(2)}$) is a unit left- (or right-) singular vector corresponding to the singular value $\tilde{\lambda}_j$ $(j = 1,...,n_{(2)} - 1)$ of $\boldsymbol{S}_{D(1)} = ((n_{(1)} - 1)(n_{(2)} - 1))^{-1/2}\boldsymbol{X}_{o1}^T\boldsymbol{X}_{o2}$.

[CDM methodology for PC scores]

(Step 1) Calculate the singular vectors $\tilde{\boldsymbol{u}}_{j(i)}$'s, $i = 1, 2$, of $\boldsymbol{S}_{D(1)}$.

(Step 2) Adjust the sign of $\tilde{\boldsymbol{u}}_{j(2)}$ by $\tilde{\boldsymbol{u}}_{j(2)} = \mathrm{Sign}(\tilde{\boldsymbol{u}}_{j(1)}^T \boldsymbol{X}_{o1}^T \boldsymbol{X}_{o2} \tilde{\boldsymbol{u}}_{j(2)}) \tilde{\boldsymbol{u}}_{j(2)}$. After the modification, let $\tilde{\boldsymbol{u}}_{j(i)}^T = (\tilde{u}_{j1(i)}, ..., \tilde{u}_{jn_{(i)}(i)})$, $i = 1, 2$.

(Step 3) Calculate $\tilde{s}_{jk(i)} = \tilde{u}_{jk(i)} \sqrt{(n_{(i)} - 1)\tilde{\lambda}_j}$, $k = 1, ..., n_{(i)}$; $i = 1, 2$. Estimate the j-th PC score of $\boldsymbol{x}_k$ by $\tilde{s}_{jk} = \tilde{s}_{jk(1)}$, $k = 1, ..., n_{(1)}$ and $\tilde{s}_{jk+n_{(1)}} = \tilde{s}_{jk(2)}$, $k = 1, ..., n_{(2)}$.

One can calculate the singular vector $\tilde{\boldsymbol{u}}_{j(i)}$'s by the eigenvectors of $\boldsymbol{S}_{D(i)} \boldsymbol{S}_{D(i)}^T$. Let $\mathrm{MSE}(\tilde{s}_j) = n^{-1} \sum_{k=1}^n (\tilde{s}_{jk} - s_{jk})^2$ denote the sample mean-square error of the j-th PC score. Note that $\mathrm{Var}(s_{jk}) = \lambda_j$. Then, Yata and Aoshima (2010b) gave the following properties on the CDM-based PC scores.

Theorem 3.1. *Assume that λ_j $(j \leq m)$ has multiplicity one. Then, it holds that*

$$\frac{MSE(\tilde{s}_j)}{\lambda_j} \xrightarrow{p} 0 \tag{8}$$

under the conditions (i)-(ii) in Theorem 2.1. If z_{jk}, $j = 1, ..., p$ $(k = 1, ..., n)$ are independent, we have (8) under the conditions (i)-(ii) in Corollary 2.1.

Theorem 3.2. *Assume that λ_j $(j \leq m)$ has multiplicity one. Then, for any k $(= 1, ..., n)$, it holds that*

$$\lambda_j^{-1/2} \tilde{s}_{jk} \xrightarrow{p} z_{jk} \tag{9}$$

under the conditions (i)-(ii) of Theorem 2.1. If z_{jk}, $j = 1, ..., p$ $(k = 1, ..., n)$ are independent, we have (9) under the conditions (i)-(ii) of Corollary 2.2.

The CDM-based PC score can be generalized as follows:

[GCDM methodology for PC scores]

(Step 1) Set iteration number T. Set $t = 1$.

(Step 2) Randomly split $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ into $\boldsymbol{X}_1 = [\boldsymbol{x}_{1(1)}, ..., \boldsymbol{x}_{1(n_{(1)})}]$ and $\boldsymbol{X}_2 = [\boldsymbol{x}_{2(1)}, ..., \boldsymbol{x}_{2(n_{(2)})}]$.

(Step 3) Define a cross data matrix by $\boldsymbol{S}_{D(1)t} = ((n_{(1)} - 1)(n_{(2)} - 1))^{-1/2} \boldsymbol{X}_{o1}^T \boldsymbol{X}_{o2}$, where $\boldsymbol{X}_{oi} = \boldsymbol{X}_i - [\overline{\boldsymbol{x}}_i, ..., \overline{\boldsymbol{x}}_i]$, $i = 1, 2$, and $\overline{\boldsymbol{x}}_i = \sum_{j=1}^{n_{(i)}} \boldsymbol{x}_{i(j)} / n_{(i)}$. Calculate the singular values, $\tilde{\lambda}_{1t} \geq \cdots \geq \tilde{\lambda}_{n_{(2)}-1\,t} (\geq 0)$, and the corresponding singular vectors, $\tilde{\boldsymbol{u}}_{j(i)t}$'s, $i = 1, 2$, of $\boldsymbol{S}_{D(1)t}$. If $t = 1$, go to Step 5; otherwise go to Step 4.

(Step 4) Adjust the sign of $\tilde{\boldsymbol{u}}_{j(1)t}$ by $\tilde{\boldsymbol{u}}_{j(1)t} = \mathrm{Sign}(\tilde{\boldsymbol{u}}_{j(1)t}^T \tilde{\boldsymbol{u}}_{j(1)1}) \tilde{\boldsymbol{u}}_{j(1)t}$.

(Step 5) Adjust the sign of $\tilde{\boldsymbol{u}}_{j(2)t}$ by $\tilde{\boldsymbol{u}}_{j(2)t} = \mathrm{Sign}(\tilde{\boldsymbol{u}}_{j(1)t}^T \boldsymbol{X}_{o1}^T \boldsymbol{X}_{o2} \tilde{\boldsymbol{u}}_{j(2)t}) \tilde{\boldsymbol{u}}_{j(2)t}$. After the modification, let $\tilde{\boldsymbol{u}}_{j(i)t}^T = (\tilde{u}_{j(1i)t}, ..., \tilde{u}_{j(n_{(i)}i)t})$, $i = 1, 2$.

(Step 6) Calculate $\tilde{s}_{j(ki)t} = \tilde{u}_{j(ki)t} \sqrt{(n_{(i)} - 1)\tilde{\lambda}_{jt}}$, $k = 1, ..., n_{(i)}$; $i = 1, 2$, and adjust the subscript k of $\tilde{s}_{j(ki)t}$ as $\tilde{s}_{jkt}$ corresponding to x_k.

(Step 7) If $t < T$, put $t = t + 1$ and go to Step 2; otherwise go to Step 8.

(Step 8) Estimate the j-th PC score of $\boldsymbol{x}_k$ by $\tilde{s}_{jk(T)} = \sum_{t=1}^T \tilde{s}_{jkt} / T$ for each j and k.

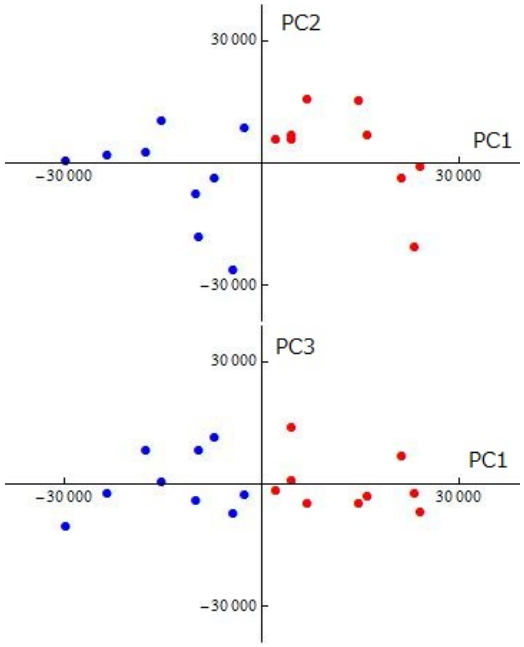

Fig. 3. Scatterplots of PC scores by PC1 and PC2 (upper panel) or PC1 and PC3 (lower panel) by using the GCDM methodology. There are 9 samples from Normal Prostate (blue point) and 9 samples from Prostate Tumors (red point).

3.2 Demonstration

We analyzed gene expression data about prostate cancer given by Singh et al. (2002). Refer to Pochet et al. (2004) for details of the data set. The data set consisted of 12600 $(= p)$ genes and 34 microarrays in which there were 9 samples from Normal Prostate and 25 samples from Prostate Tumors. As for Prostate Tumors, we chose the first 9 samples and set 18 $(= n)$ microarrays in which there were 9 samples from Normal Prostate and 9 samples from Prostate Tumors. We assumed the mixture model given by (6) for the data set. We defined the data matrix by $\boldsymbol{X}$: 12600×18. We set $(n_{(1)}, n_{(2)}) = (9, 9)$ and $T = 1000$. We focused on the first three PC scores. We randomly divided $\boldsymbol{X}$ into $\boldsymbol{X}_1$: 12600×9 and $\boldsymbol{X}_2$: 12600×9, and calculated $\tilde{s}_{jkt}$, $k = 1, ..., 18$, for $j = 1, 2, 3$. According to the GCDM methodology, we repeated this operation $T = 1000$ times and obtained $\tilde{s}_{jk(T)}$, $k = 1, ..., 18$; $j = 1, 2, 3$, as an estimate of the j-th PC score of $\boldsymbol{x}_k$. We also obtained $(\tilde{\lambda}_{1(T)}, \tilde{\lambda}_{2(T)}, \tilde{\lambda}_{3(T)}) = (2.77 \times 10^8, 1.62 \times 10^8, 6.34 \times 10^7)$. Figure 3 gives the scatterplots of the first three PC scores on the (PC1, PC2) plane or the (PC1, PC3) plane. As observed in Figure 3, Normal Prostate (blue point) and Prostate Tumors (red point) seem to be separated clearly. It is obvious especially for the first PC score (PC1) line. All the first PC scores of the samples from Normal Prostate are negative, whereas those from Prostate Tumors are positive. This observation is theoretically supported by the arguments in Section 3.1.

4. Classification for high-dimension, low-sample-size data

Suppose we have independent and p-dimensional populations, Π_i, $i = 1, 2$, having *unknown* mean vector $\boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{ip})^T$ and *unknown* positive-definite covariance matrix $\boldsymbol{\Sigma}_i$ for each i. *We do not assume that* $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. The eigen-decomposition of $\boldsymbol{\Sigma}_i$ $(i = 1, 2)$ is given by $\boldsymbol{\Sigma}_i = \boldsymbol{H}_i \boldsymbol{\Lambda}_i \boldsymbol{H}_i^T$, where $\boldsymbol{\Lambda}_i$ is a diagonal matrix of eigenvalues $\lambda_{i1} \geq \cdots \geq \lambda_{ip} > 0$ and $\boldsymbol{H}_i = [\boldsymbol{h}_{i1}, ..., \boldsymbol{h}_{ip}]$ is an orthogonal matrix of corresponding eigenvectors. Having recorded i.i.d. samples, $\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i}$, from each Π_i, we have a $p \times n_i$ $(p > n_i)$ data matrix $\boldsymbol{X}_i = [\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i}]$, where $\boldsymbol{x}_{ij} = (x_{i1j}, ..., x_{ipj})^T$, $j = 1, ..., n_i$. We assume $n_i \geq 4$, $i = 1, 2$. Then, $\boldsymbol{Z}_i = \boldsymbol{\Lambda}_i^{-1/2} \boldsymbol{H}_i^T (\boldsymbol{X}_i - [\boldsymbol{\mu}_i, ..., \boldsymbol{\mu}_i])$ is considered as a $p \times n_i$ sphered data matrix from a distribution with zero mean and the identity covariance matrix. Here, we write $\boldsymbol{Z}_i = [\boldsymbol{z}_{i1}, ..., \boldsymbol{z}_{in_i}]$ and $\boldsymbol{z}_{ij} = (z_{i1j}, ..., z_{ipj})^T$, $j = 1, ..., n_i$. Note that $E(z_{ijl}^2) = 1$ and $E(z_{ijl} z_{ij'l}) = 0$ for $i = 1, 2$; $j(\neq j') = 1, ..., p$; $l = 1, ..., n_i$. We assume that $\lambda_{ip} > 0$ $(i = 1, 2)$ as $p \to \infty$ and the fourth moments of each variable in $\boldsymbol{Z}_i$ are uniformly bounded. In this section, we assume the following assumption for Π_i's:

(A-i) z_{ijl}, $j = 1, ..., p$, are independent for $i = 1, 2$.

One of the population distributions satisfying (A-i) is $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. We also assume the following condition for $\boldsymbol{\Sigma}_i$'s as necessary:

(A-ii) $\dfrac{\mathrm{tr}(\boldsymbol{\Sigma}_i^t)}{p} < \infty$ $(t = 1, 2)$ and $\dfrac{\mathrm{tr}(\boldsymbol{\Sigma}_i^4)}{p^2} \to 0$ as $p \to \infty$ for $i = 1, 2$.

Remark 4.1. If all λ_{ij}'s are bounded, (A-ii) trivially holds. For a spiked model such as $\lambda_{ij} = a_{ij} p^{\alpha_{ij}}$ $(j = 1, ..., m_i)$ and $\lambda_{ij} = c_{ij}$ $(j = m_i + 1, ..., p)$ with positive constants a_{ij}'s, c_{ij}'s and α_{ij}'s, (A-ii) holds under the condition that $\alpha_{ij} < 1/2$, $j = 1, ..., m_i (< \infty)$; $i = 1, 2$. As an interesting example, (A-ii) holds for $\boldsymbol{\Sigma}_{i'} = c_{i'} (\rho_{i'}^{|i-j|^{q_{i'}}})$, $i' = 1, 2$, where $c_{i'}$'s, $q_{i'}$'s and $\rho_{i'}$'s(< 1) are positive constants.

4.1 Discriminant rule for HDLSS data

Let $\boldsymbol{x}_0$ be an observation vector on an individual belonging to Π_1 or to Π_2. Having recorded $\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i}$ from each Π_i, we estimate $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ by $\overline{\boldsymbol{x}}_{in_i} = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}/n_i$ and $\boldsymbol{S}_{in_i} = \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})^T/(n_i - 1)$. Aoshima and Yata (2011) considered a discriminant rule that classifies $\boldsymbol{x}_0$ into Π_1 if

$$\frac{p||\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1}||^2}{\mathrm{tr}(\boldsymbol{S}_{1n_1})} - \frac{p||\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2}||^2}{\mathrm{tr}(\boldsymbol{S}_{2n_2})} - p \log \left\{ \frac{\mathrm{tr}(\boldsymbol{S}_{2n_2})}{\mathrm{tr}(\boldsymbol{S}_{1n_1})} \right\} - \frac{p}{n_1} + \frac{p}{n_2} + \gamma < 0 \tag{10}$$

and into Π_2 otherwise. Here, $-p/n_1 + p/n_2$ is a bias-correction and γ is a tuning parameter. We denote the error rate of misclassifying an individual from Π_1 (into Π_2) or from Π_2 (into Π_1) by $e(2|1)$ or $e(1|2)$. Let $\Delta = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2$ and $\Delta_{\boldsymbol{\Sigma}_i} = (\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2))^2/\mathrm{tr}(\boldsymbol{\Sigma}_i)$, $i = 1, 2$. Let us write that $\Delta_i = \Delta + \Delta_{\boldsymbol{\Sigma}_i}/2$, $i = 1, 2$, and $\Delta_\star = \min_{i=1,2} \Delta_i$. Aoshima and Yata (2011) gave the following property.

Theorem 4.1. *Assume (A-i)-(A-ii). Under the condition that* $\max_{i=1,2}\{tr(\boldsymbol{\Sigma}_i^2)\}/(\Delta_\star^2 \min_{i=1,2}\{n_i\}) \to 0$ *as* $p \to \infty$, *for the discriminant rule given by (10) with* $\gamma = 0$, *it holds as* $p \to \infty$ *that*

$$e(2|1) \to 0 \quad \textit{and} \quad e(1|2) \to 0. \tag{11}$$

Remark 4.2. Assume (A-i)-(A-ii). Let us consider a case that $\mathrm{tr}(\mathbf{\Sigma}_1)/\mathrm{tr}(\mathbf{\Sigma}_2) \neq 1$ as $p \to \infty$. Then, it follows that $\min_{i=1,2} \Delta_{\mathbf{\Sigma}_i}/p > 0$ as $p \to \infty$. Since it holds $\max_{i=1,2}\{\mathrm{tr}(\mathbf{\Sigma}_i^2)\} / (\Delta_\star^2 \min_{i=1,2}\{n_i\}) \to 0$ as $p \to \infty$, we can claim (11) in the case.

Remark 4.3. Let $n_{i(1)} = [n_i/2] + 1$ and $n_{i(2)} = n_i - n_{i(1)}$ for each Π_i $(i = 1, 2)$. We omit the subscript i for a while. For each Π, split $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ into $\boldsymbol{X}_1 = [\boldsymbol{x}_{11}, ..., \boldsymbol{x}_{1n_{(1)}}]$ and $\boldsymbol{X}_2 = [\boldsymbol{x}_{21}, ..., \boldsymbol{x}_{2n_{(2)}}]$. Let $\boldsymbol{X}_{o1} = \boldsymbol{X}_1 - [\overline{\boldsymbol{x}}_1, ..., \overline{\boldsymbol{x}}_1]$ and $\boldsymbol{X}_{o2} = \boldsymbol{X}_2 - [\overline{\boldsymbol{x}}_2, ..., \overline{\boldsymbol{x}}_2]$, where $\overline{\boldsymbol{x}}_1 = \sum_{j=1}^{n_{(1)}} \boldsymbol{x}_{1j}/n_{(1)}$ and $\overline{\boldsymbol{x}}_2 = \sum_{j=1}^{n_{(2)}} \boldsymbol{x}_{2j}/n_{(2)}$. Define $\boldsymbol{S}_{n(1)} = (n_{(1)} - 1)^{-1}\boldsymbol{X}_{o1}\boldsymbol{X}_{o1}^T$ and $\boldsymbol{S}_{n(2)} = (n_{(2)} - 1)^{-1}\boldsymbol{X}_{o2}\boldsymbol{X}_{o2}^T$. Note that $\mathrm{tr}(\boldsymbol{S}_{n(1)}\boldsymbol{S}_{n(2)}) = \mathrm{tr}(\boldsymbol{S}_{D(1)}\boldsymbol{S}_{D(1)}^T) = \sum_{j=1}^{n_{(2)}-1} \tilde{\lambda}_j^2$. Then, we have that $E(\mathrm{tr}(\boldsymbol{S}_{n(1)}\boldsymbol{S}_{n(2)})) = \mathrm{tr}(\mathbf{\Sigma}^2)$. As for $\mathrm{tr}(\mathbf{\Sigma}_i^2)$, Yata (2010) considered an unbiased estimator, $\mathrm{tr}(\boldsymbol{S}_{in_i(1)}\boldsymbol{S}_{in_i(2)})$, as an application of the CDM methodology given by Yata and Aoshima (2010a,b).

Remark 4.4. We note that $\Delta_\star$ is estimated by

$$||\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}||^2 - \sum_{i=1}^{2} \mathrm{tr}(\boldsymbol{S}_{in_i})/n_i + \frac{|\mathrm{tr}(\boldsymbol{S}_{1n_1}) - \mathrm{tr}(\boldsymbol{S}_{2n_2})|^2}{2\max_{i=1,2}\mathrm{tr}(\boldsymbol{S}_{in_i})} \ (= \widehat{\Delta}_\star, \text{ say}).$$

We analyzed gene expression data given by Armstrong et al. (2001) in which data set consisted of 12582 $(= p)$ genes. We had two populations about leukemia subtypes, i.e., Π_1: acute lymphoblastic leukemia (ALL, 24 samples) and Π_2: acute myeloid leukemia (AML, 28 samples). We set $n_1 = n_2 = 10$. Then, we constructed the discriminant rule given by (10) with $\gamma = 0$. From Remarks 4.3 and 4.4, we calculated $\max_{i=1,2}\{\mathrm{tr}(\boldsymbol{S}_{in_i(1)}\boldsymbol{S}_{in_i(2)})\} = 3.16 \times 10^{19}$ and $\widehat{\Delta}_\star = 2.67 \times 10^{10}$, so that $\max_{i=1,2}\{\mathrm{tr}(\boldsymbol{S}_{in_i(1)}\boldsymbol{S}_{in_i(2)})\}/(\widehat{\Delta}_\star^2 \min_{i=1,2}\{n_i\}) = 0.0044$. Thus, one may conclude that $\max_{i=1,2}\{\mathrm{tr}(\mathbf{\Sigma}_i^2)\}/(\Delta_\star^2 \min_{i=1,2}\{n_i\})$ must be sufficiently small. Hence, from Theorem 4.1, the discriminant rule given by (10) with $\gamma = 0$ was expected to hold (11). In Table 1, we investigated the performance of the discriminant rule by using test data sets consisting of $24 - n_1 = 14$ remaining samples from Π_1 and $28 - n_2 = 18$ remaining samples from Π_2. We observed that the discriminant rule showed $e(1|2) = 0$ and $e(2|1) = 0$ successfully as expected by theory.

	(10) with $\gamma = 0$
1-$\overline{e(2\|1)}$	14/14 (=1.0)
1-$\overline{e(1\|2)}$	18/18 (=1.0)

Table 1. The correct discrimination rates for test data sets consisting of 14 samples from Π_1 and 18 samples from Π_2.

4.2 Sample size determination for classification

One would be interested in designing the discriminant rule given by (10) so as to hold both $e(2|1) \leq \alpha$ and $e(1|2) \leq \beta$ when $\Delta_\star \geq \Delta_L$, where α, $\beta \in (0, 1/2)$ and Δ_L (> 0) are prespecified constants. We assume $\Delta_L = o(p^{1/2})$. Aoshima and Yata (2011) showed the following property.

Theorem 4.2. *Assume that* $tr(\mathbf{\Sigma}_1)/tr(\mathbf{\Sigma}_2) \to 1$ *as* $p \to \infty$. *Let*

$$\omega(\boldsymbol{x}_0) = \frac{p||\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1}||^2}{\mathrm{tr}(\boldsymbol{S}_{1n_1})} - \frac{p||\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2}||^2}{\mathrm{tr}(\boldsymbol{S}_{2n_2})} - p\log\left\{\frac{\mathrm{tr}(\boldsymbol{S}_{2n_2})}{\mathrm{tr}(\boldsymbol{S}_{1n_1})}\right\} - \frac{p}{n_1} + \frac{p}{n_2}.$$

Then, under the regularity conditions, it holds as $p \to \infty$ and $n_1, n_2 \to \infty$ that

$$\frac{\omega(\boldsymbol{x}_0) + \Delta_2(\mathrm{tr}(\boldsymbol{\Sigma}_2)/p)^{-1}}{2\sqrt{(\mathrm{tr}(\boldsymbol{\Sigma}_1)/p)^{-2}\mathrm{tr}(\boldsymbol{\Sigma}_1^2)/n_1 + (\mathrm{tr}(\boldsymbol{\Sigma}_2)/p)^{-2}\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/n_2}} \Rightarrow N(0,1) \quad \text{when } \boldsymbol{x}_0 \in \Pi_1;$$

$$\frac{\omega(\boldsymbol{x}_0) - \Delta_1(\mathrm{tr}(\boldsymbol{\Sigma}_1)/p)^{-1}}{2\sqrt{(\mathrm{tr}(\boldsymbol{\Sigma}_2)/p)^{-2}\mathrm{tr}(\boldsymbol{\Sigma}_2^2)/n_2 + (\mathrm{tr}(\boldsymbol{\Sigma}_1)/p)^{-2}\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/n_1}} \Rightarrow N(0,1) \quad \text{when } \boldsymbol{x}_0 \in \Pi_2.$$

Let $\sigma = \max\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)^{1/2}, \mathrm{tr}(\boldsymbol{\Sigma}_2^2)^{1/2}\}$. We find the sample size for each Π_i $(i = 1, 2)$ as

$$n_i \geq \frac{(z_\alpha + z_\beta)^2\sigma}{\Delta_L^2}\mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/4}\sum_{j=1}^{2}\mathrm{tr}(\boldsymbol{\Sigma}_j^2)^{1/4} \quad (= C_i, \text{ say}), \tag{12}$$

where z_α is the upper α point of $N(0,1)$. Note that $C_i = O(p/\Delta_L^2)$ for $i = 1, 2$, under (A-ii). Thus under $\Delta_L \to \infty$ as $p \to \infty$, it holds that $C_i/p \to 0$ as $p \to \infty$. Then, Aoshima and Yata (2011) gave the following theorem.

Theorem 4.3. *Assume (A-i)-(A-ii). Let $\gamma = (tr(\boldsymbol{S}_{1n_1} + \boldsymbol{S}_{2n_2})/(2p))^{-1}\Delta_L(z_\beta - z_\alpha)/(z_\alpha + z_\beta)$ in (10). Then, under the regularity conditions, for the discriminant rule given by (10) with (12), it holds as $p \to \infty$ that*

$$\limsup e(2|1) \leq \alpha \quad \text{and} \quad \limsup e(1|2) \leq \beta$$

when $\Delta_\star \geq \Delta_L$.

Remark 4.5. One can design Δ_L by using the two sample test given by Aoshima and Yata (2011). Under the regularity conditions, it holds that

$$\frac{||\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}||^2 - \sum_{i=1}^{2}\mathrm{tr}(\boldsymbol{S}_{in_i})/n_i - \Delta}{\sqrt{\widehat{Var}(||\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}||^2)}} \Rightarrow N(0,1)$$

as $p \to \infty$ and $n_i \to \infty$, $i = 1, 2$, where

$$\widehat{Var}(||\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}||^2) = 2\frac{\mathrm{tr}(\boldsymbol{S}_{1n_1(1)}\boldsymbol{S}_{1n_1(2)})}{n_1(n_1-1)} + 2\frac{\mathrm{tr}(\boldsymbol{S}_{2n_2(1)}\boldsymbol{S}_{2n_2(2)})}{n_2(n_2-1)} + 4\frac{\mathrm{tr}(\boldsymbol{S}_{1n_1}\boldsymbol{S}_{2n_2})}{n_1 n_2}.$$

Note that $E(||\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}||^2 - \sum_{i=1}^{2}\mathrm{tr}(\boldsymbol{S}_{in_i})/n_i) = \Delta$. Thus it follows that

$$P\left(\frac{||\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}||^2 - \sum_{i=1}^{2}\mathrm{tr}(\boldsymbol{S}_{in_i})/n_i}{\sqrt{\widehat{Var}(||\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}||^2)}} - z_{\alpha'} \leq \frac{\Delta}{\sqrt{\widehat{Var}(||\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}||^2)}}\right) \to 1 - \alpha'$$

with $\alpha' \in (0, 1/2)$. From the fact that $\Delta_\star \geq \Delta$, we design a lower bound of $\Delta_\star$ by

$$\Delta_L = ||\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}||^2 - \sum_{i=1}^{2}\mathrm{tr}(\boldsymbol{S}_{in_i})/n_i - z_{\alpha'}\sqrt{\widehat{Var}(||\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}||^2)}$$

for sufficiently small α'.

Since $\mathbf{\Sigma}_i$'s are unknown, it is necessary to estimate C_i's in (12) with some pilot samples. We proceed the following two steps:

[Two-stage procedure for classification]

(Step 1) Choose a pilot sample size, $m(\geq 4)$, such as $m/C_i \in (0,1)$, $i = 1,2$, as $p \to \infty$. Take pilot samples of size m from each Π_i and define $\boldsymbol{X}_i = [\boldsymbol{x}_{i1},...,\boldsymbol{x}_{im}]$, $i = 1,2$. Let $m_{(1)} = [m/2]+1$ and $m_{(2)} = m - m_{(1)}$. For each Π_i, divide $\boldsymbol{X}_i$ into $\boldsymbol{X}_i = [\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}]$ with $\boldsymbol{X}_{i1}:\ p \times m_{(1)}$ and $\boldsymbol{X}_{i2}:\ p \times m_{(2)}$, and calculate

$$\begin{aligned} \boldsymbol{S}_{im(1)} &= \frac{(\boldsymbol{X}_{i1} - [\overline{\boldsymbol{x}}_{im_{(1)}},...,\overline{\boldsymbol{x}}_{im_{(1)}}])(\boldsymbol{X}_{i1} - [\overline{\boldsymbol{x}}_{im_{(1)}},...,\overline{\boldsymbol{x}}_{im_{(1)}}])^T}{m_{(1)} - 1} \\ &\text{and} \\ \boldsymbol{S}_{im(2)} &= \frac{(\boldsymbol{X}_{i2} - [\overline{\boldsymbol{x}}_{im_{(2)}},...,\overline{\boldsymbol{x}}_{im_{(2)}}])(\boldsymbol{X}_{i2} - [\overline{\boldsymbol{x}}_{im_{(2)}},...,\overline{\boldsymbol{x}}_{im_{(2)}}])^T}{m_{(2)} - 1}, \end{aligned} \tag{13}$$

where $\overline{\boldsymbol{x}}_{im_{(1)}} = \sum_{j=1}^{m_{(1)}} \boldsymbol{x}_{ij}/m_{(1)}$ and $\overline{\boldsymbol{x}}_{im_{(2)}} = \sum_{j=m_{(1)}+1}^{m} \boldsymbol{x}_{ij}/m_{(2)}$. Define the total sample size for each Π_i by

$$N_i = \max\Big\{m,\ \Big[\frac{(z_\alpha + z_\beta)^2\hat{\sigma}}{\Delta_L^2}\mathrm{tr}(\boldsymbol{S}_{im(1)}\boldsymbol{S}_{im(2)})^{1/4}\sum_{j=1}^{2}\mathrm{tr}(\boldsymbol{S}_{jm(1)}\boldsymbol{S}_{jm(2)})^{1/4}\Big] + 1\Big\}, \tag{14}$$

where $\hat{\sigma} = \max\{\mathrm{tr}(\boldsymbol{S}_{1m(1)}\boldsymbol{S}_{1m(2)})^{1/2}, \mathrm{tr}(\boldsymbol{S}_{2m(1)}\boldsymbol{S}_{2m(2)})^{1/2}\}$.

(Step 2) Take additional samples $\boldsymbol{x}_{ij}$, $j = m+1,...,N_i$, of size $N_i - m$ from each Π_i. By combining the initial samples and the additional samples, calculate $\overline{\boldsymbol{x}}_{iN_i} = \sum_{j=1}^{N_i}\boldsymbol{x}_{ij}/N_i$ and $\boldsymbol{S}_{iN_i} = \sum_{j=1}^{N_i}(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{iN_i})(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{iN_i})^T/(N_i - 1)$, $i = 1,2$. Then, we classify $\boldsymbol{x}_0$ into Π_1 if

$$\frac{p||\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1N_1}||^2}{\mathrm{tr}(\boldsymbol{S}_{1N_1})} - \frac{p||\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2N_2}||^2}{\mathrm{tr}(\boldsymbol{S}_{2N_2})} - p\log\Big\{\frac{\mathrm{tr}(\boldsymbol{S}_{2N_2})}{\mathrm{tr}(\boldsymbol{S}_{1N_1})}\Big\} - \frac{p}{N_1} + \frac{p}{N_2} + \hat{\gamma} < 0 \tag{15}$$

and into Π_2 otherwise, where $\hat{\gamma} = (\mathrm{tr}(\boldsymbol{S}_{1N_1} + \boldsymbol{S}_{2N_2})/(2p))^{-1}\Delta_L(z_\beta - z_\alpha)/(z_\alpha + z_\beta)$.

Aoshima and Yata (2011) gave the following theorem.

Theorem 4.4. *Assume (A-i)-(A-ii). Then, under the regularity conditions, for the discriminant rule given by (15) with (14), it holds as $p \to \infty$ that*

$$\limsup e(2|1) \leq \alpha \quad \text{and} \quad \limsup e(1|2) \leq \beta$$

when $\Delta_\star \geq \Delta_L$.

Remark 4.6. One may take different pilot-sample-sizes, $m_i(\geq 4)$, such as $m_i/C_i \in (0,1)$ as $p \to \infty$ for $i = 1,2$. Then, the assertion in Theorem 4.4 is still claimed.

Remark 4.7. Assume (A-i)-(A-ii). Then, it holds as $p \to \infty$ that $N_i/C_i \xrightarrow{p} 1$ for $i = 1,2$, which are in the HDLSS situation in the sense that $N_i/p \xrightarrow{p} 0$, $i = 1,2$, under $\Delta_L \to \infty$ as $p \to \infty$.

4.3 Demonstration

We analyzed gene expression data given by Chiaretti et al. (2004) in which data set consisted of 12625 $(= p)$ genes and 128 samples. Note that the expression measures were obtained by using the three-step robust multichip average (RMA) preprocessing method. Refer to Pollard et al. (2005) as well for the details. The data set had two tumor cellular subtypes, Π_1: B-cell (95 samples) and Π_2: T-cell (33 samples). We set $\alpha = 0.1$, $\beta = 0.02$ and $m = 6$. Our goal was to construct a discriminant rule ensuring that both $1 - e(2|1) \geq 0.9$ and $1 - e(1|2) \geq 0.98$ when $\Delta_\star \geq \Delta_L$, where Δ_L is designed later.

First, we took the first 6 samples from each Π_i as a pilot sample. According to Remark 4.5, we calculated $||\overline{\boldsymbol{x}}_{1m} - \overline{\boldsymbol{x}}_{2m}||^2 - \sum_{i=1}^{2} \mathrm{tr}(\boldsymbol{S}_{im})/m = 1890$ and $\widehat{Var}(||\overline{\boldsymbol{x}}_{1m} - \overline{\boldsymbol{x}}_{2m}||^2) = 87860$. By setting $\alpha' = 0.01$ so that $z_{\alpha'} = 2.33$, we designed a lower bound of $\Delta_\star$ by

$$\Delta_L = ||\overline{\boldsymbol{x}}_{1m} - \overline{\boldsymbol{x}}_{2m}||^2 - \sum_{i=1}^{2} \mathrm{tr}(\boldsymbol{S}_{im})/m - z_{\alpha'}\sqrt{\widehat{Var}(||\overline{\boldsymbol{x}}_{1m} - \overline{\boldsymbol{x}}_{2m}||^2)} = 1200.$$

According to (14), the total sample size for each Π_i was given by

$$N_1 = \max\left\{6,\ \left[\frac{(z_\alpha + z_\beta)^2\hat{\sigma}}{\Delta_L^2}\mathrm{tr}(\boldsymbol{S}_{1m(1)}\boldsymbol{S}_{1m(2)})^{1/4}\sum_{j=1}^{2}\mathrm{tr}(\boldsymbol{S}_{jm(1)}\boldsymbol{S}_{jm(2)})^{1/4}\right] + 1\right\} = 10,$$

$$N_2 = \max\left\{6,\ \left[\frac{(z_\alpha + z_\beta)^2\hat{\sigma}}{\Delta_L^2}\mathrm{tr}(\boldsymbol{S}_{2m(1)}\boldsymbol{S}_{2m(2)})^{1/4}\sum_{j=1}^{2}\mathrm{tr}(\boldsymbol{S}_{jm(1)}\boldsymbol{S}_{jm(2)})^{1/4}\right] + 1\right\} = 6.$$

So, we took the next 4 $(= N_1 - m)$ samples from Π_1. On the other hand, since $N_2 = m$, we did not take additional samples from Π_2. We had $\hat{\gamma} = (\mathrm{tr}(\boldsymbol{S}_{1N_1} + \boldsymbol{S}_{2N_2})/(2p))^{-1}\Delta_L(z_\beta - z_\alpha)/(z_\alpha + z_\beta) = 58.1$. Then, we constructed the discriminant rule given by (15) ensuring that both $1 - e(2|1) \geq 0.9$ and $1 - e(1|2) \geq 0.98$ when $\Delta_\star \geq 1200$.

We compared the constructed discriminant rule with two other discriminant rules, DLDR and DQDR, that were given by Dudoit et al. (2002) as follows: Diagonal linear discriminant rule (DLDR) classifies $\boldsymbol{x}_0$ into Π_1 if

$$(\boldsymbol{x}_0 - (\overline{\boldsymbol{x}}_{1N_1} + \overline{\boldsymbol{x}}_{2N_2})/2)^T\boldsymbol{S}_{diag}^{-1}(\overline{\boldsymbol{x}}_{2N_2} - \overline{\boldsymbol{x}}_{1N_1}) < 0$$

and into Π_2 otherwise, where $\boldsymbol{S}_{diag} = \mathrm{diag}(s_{1N}, ..., s_{pN})$ having $s_{jN} = \sum_{i=1}^{2}\sum_{l=1}^{N_i}(x_{ijl} - \overline{x}_{ijN_i})^2 / (N_1 + N_2 - 2)$ and $\overline{x}_{ijN_i} = \sum_{l=1}^{N_i} x_{ijl}/N_i$. On the other hand, diagonal quadratic discriminant rule (DQDR) classifies $\boldsymbol{x}_0$ into Π_1 if

$$(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1N_1})^T\boldsymbol{S}_{diag(1)}^{-1}(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1N_1}) - (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2N_2})^T\boldsymbol{S}_{diag(2)}^{-1}(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2N_2}) - \log\left\{\frac{\det(\boldsymbol{S}_{diag(2)})}{\det(\boldsymbol{S}_{diag(1)})}\right\} < 0$$

and into Π_2 otherwise, where $\boldsymbol{S}_{diag(i)} = \mathrm{diag}(s_{(i)1N_i}, ..., s_{(i)pN_i})$ having $s_{(i)jN_i} = \sum_{l=1}^{N_i}(x_{ijl} - \overline{x}_{ijN_i})^2 / (N_i - 1)$. We constructed the three discriminant rules by using the common data sets of sizes $(N_1, N_2) = (10, 6)$. In Table 2, we investigated those performances by using the remaining samples of sizes $(95 - N_1, 33 - N_2) = (85, 27)$ as test data sets. We observed that the discriminant rule given by (15) showed an adequate performance.

	(15)	DLDR	DQDR
1-$\overline{e(2\|1)}$	75/85 (=0.882)	63/85 (=0.741)	76/85 (=0.894)
1-$\overline{e(1\|2)}$	27/27 (=1.0)	24/27 (=0.889)	24/27 (=0.889)

Table 2. The correct discrimination rates, by (15), DLDR and DQDR, for test data sets consisting of 85 samples from Π_1 and 27 samples from Π_2.

5. Variable selection for high-dimension, low-sample-size data

Suppose we have two independent and p-dimensional populations, Π_i, $i = 1, 2$, having *unknown* mean vector $\boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{ip})^T$ and *unknown* positive-definite covariance matrix $\boldsymbol{\Sigma}_i$ for each i. *We do not assume* $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. We consider an effective methodology to select a significant set of associated variables from among high-dimensional data sets. That is, we consider testing the following hypotheses simultaneously:

$$H_{0j} : \mu_{1j} = \mu_{2j} \quad \text{vs.} \quad H_{1j} : \mu_{1j} \neq \mu_{2j} \quad \text{for } j = 1, ..., p. \tag{16}$$

Our interest is to select a set of significant variables such that $\boldsymbol{D} = \{j : \mu_{1j} \neq \mu_{2j}\}$. Assume that $|\boldsymbol{D}| = S$ for some $S \geq 1$, where $|\boldsymbol{D}|$ denotes the number of elements in set $\boldsymbol{D}$. A variable selection procedure $\widehat{\boldsymbol{D}}$ maps the data into subsets of $\{1, ..., p\}$. We are interested in designing $\widehat{\boldsymbol{D}}$ ensuring that both the asymptotic *family-wise error rate* (FWER) is 0, i.e.,

$$P(|\boldsymbol{D}^c \cap \widehat{\boldsymbol{D}}| \neq 0) \to 0, \tag{17}$$

and the asymptotic *average power* (AP) is 1, i.e.,

$$\frac{|\boldsymbol{D} \cap \widehat{\boldsymbol{D}}|}{S} \xrightarrow{p} 1 \quad \text{when } \min_{j \in \boldsymbol{D}} |\mu_{1j} - \mu_{2j}|^2 > \delta, \tag{18}$$

where δ (> 0) is a prespecified constant. When S is bounded $(< \infty)$, one can modify (18) by

$$P(\boldsymbol{D} \subseteq \widehat{\boldsymbol{D}}) \to 1 \quad \text{when } \min_{j \in \boldsymbol{D}} |\mu_{1j} - \mu_{2j}|^2 > \delta.$$

We note that the assertion (18) does not consider the case when $\min_{j \in \boldsymbol{D}} |\mu_{1j} - \mu_{2j}|^2 = \delta$.

5.1 Variable selection procedure for HDLSS data

Let $\sigma_i = \max_{1 \leq j \leq p} \sigma_{(i)j}$ $(i = 1, 2)$, where $\sigma_{(i)j}$, $j = 1, ..., p$, are diagonal elements of $\boldsymbol{\Sigma}_i$. We assume that $\sigma_{(i)j} < \infty$ for $i = 1, 2$; $j = 1, ..., p$, and $E_{\boldsymbol{\theta}}\{\exp(t|x_{ijl} - \mu_{ij}|/\sigma_{(i)j}^{1/2})\} < \infty$, $i = 1, 2$; $j = 1, ..., p$, for some $t > 0$. Then, for testing the hypotheses (16), we take samples,

$\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i}$, of size

$$n_i \geq \frac{(\log p)^{1+\zeta}}{\delta} \tag{19}$$

from each Π_i $(i = 1, 2)$, where $\zeta \in (0, 1]$ is a chosen constant. Let $\boldsymbol{x}_{il} = (x_{i1l}, ..., x_{ipl})^T$, $l = 1, ..., n_i$. Calculate $T_{j(\mathbf{n})} = \overline{x}_{1jn_1} - \overline{x}_{2jn_2}$ for $j = 1, ..., p$, where $\overline{x}_{ijn_i} = \sum_{l=1}^{n_i} x_{ijl}/n_i$ for each Π_i. Then, test the hypothesis for $j = 1, ..., p$, by

$$\text{rejecting } H_{0j} \Longleftrightarrow |T_{j(\mathbf{n})}| > \sqrt{\delta}. \tag{20}$$

Let $\widehat{\boldsymbol{D}} = \{j \mid \text{rejecting } H_{0j}\}$. Then, from Theorem 5.1 given in Aoshima and Yata (2011), we can claim the following theorem.

Theorem 5.1. *The test given by (20) with (19) has as $p \to \infty$ that*

$$\begin{aligned} &P(|\boldsymbol{D}^c \cap \widehat{\boldsymbol{D}}| \neq 0) \to 1\,; \\ &\frac{|\boldsymbol{D} \cap \widehat{\boldsymbol{D}}|}{S} \xrightarrow{p} 1 \quad \textit{when } \min_{j \in \boldsymbol{D}} |\mu_{1j} - \mu_{2j}|^2 > \delta. \end{aligned} \tag{21}$$

One would be interested in a two-stage variable selection procedure so as to provide screening of variables in the first stage. We consider selecting a significant set of associated variables from among a set of candidate variables in the second stage. Aoshima and Yata (2011) proposed the following effective methodology:

[Two-stage variable selection procedure]

(Step 1) Choose a pilot sample size m such as $m = O(\log p)$ and $m \to \infty$ as $p \to \infty$. Take pilot samples $\boldsymbol{x}_{il}$, $l = 1, ..., m$, of size m from each Π_i $(i = 1, 2)$. Calculate $T_{j(m)} = \overline{x}_{1jm} - \overline{x}_{2jm}$ for $j = 1, ..., p$, where $\overline{x}_{ijm} = \sum_{l=1}^{m} x_{ijl}/m$ for each Π_i. Then, provide screening of variables by

$$\widetilde{\boldsymbol{D}} = \{j \mid |T_{j(m)}| > \sqrt{\delta}\} \tag{22}$$

for a set of candidate variables. Let $\widetilde{S} = |\widetilde{\boldsymbol{D}}|$. Define the additional sample size for each Π_i by

$$N = \Big[\frac{\max\{(\log \widetilde{S})^{1+\xi},\ (\log p)^{\varepsilon}\}}{\delta}\Big] + 1, \tag{23}$$

where $\xi \in (0, 1]$ and $\varepsilon \in (0, 1]$ are chosen constants.

(Step 2) Regarding $j \in \widetilde{\boldsymbol{D}}$, take new samples x_{ijl}, $l = m + 1, ..., m + N$, of size N from each Π_i. Calculate $T_{j(N)} = \overline{x}_{1j(N)} - \overline{x}_{2j(N)}$, where $\overline{x}_{ij(N)} = \sum_{l=m+1}^{m+N} x_{ijl}/N$, $j \in \widetilde{\boldsymbol{D}}$ for each Π_i. Then, test the hypothesis by

$$\text{rejecting } H_{0j} \Longleftrightarrow |T_{j(N)}| > \sqrt{\delta} \tag{24}$$

for $j \in \widetilde{\boldsymbol{D}}$, and define

$$\widehat{\boldsymbol{D}} = \{j \in \widetilde{\boldsymbol{D}} \mid \text{rejecting } H_{0j}\}. \tag{25}$$

Select the variables regarding $\widehat{\boldsymbol{D}}$.

From Theorem 5.2 in Aoshima and Yata (2011), we can claim the following theorem.
Theorem 5.2. *The two-stage variable selection procedure, (22) and (25), given by (24) with (23) has (21) as* $p \to \infty$.

5.2 Demonstration

We analyzed the gene expression data of Prostate Cancer that were given by Singh et al. (2002). The data took a pre-processing given by Jeffery et al. (2006). The data set consisted of 12600 ($= p$) genes and two groups, Π_1: Normal Prostate (50 samples) and Π_2: Prostate Tumors (52 samples).

5.2.1 Variable selection procedure

We set $\delta = 1.5$. Our goal was to find variables j's such that $|\mu_{1j} - \mu_{2j}|^2 > 1.5$. We chose the pilot sample size for each Π_i as $m = 18$ ($= O(\log p)$). Then, we took the first 18 samples from each Π_i as pilot samples, which are given in Table 3.

	Π_1: Normal Prostate			Π_2: Prostate Tumors		
$j \backslash l$	1	$\cdots$	18	1	$\cdots$	18
1	6.776	$\cdots$	7.017	6.888	$\cdots$	6.905
$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$
$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$
12600	3.050	$\cdots$	3.612	3.097	$\cdots$	3.549

Table 3. Pilot samples, x_{ijl} ($p = 12600$, $m = 18$)

We considered screening variables by $\widetilde{D} = \{j \mid |\overline{x}_{1jm} - \overline{x}_{2jm}|^2 > 1.5\}$. Then, we obtained a set of candidate variables as $\widetilde{D} = \{192, 198, 200, ..., 12153, 12156, 12432\}$ with $\widetilde{S} = |\widetilde{D}| = 160$. We set $(\xi, \varepsilon) = (1.0, 1.0)$. According to (23), the additional sample size for each Π_i was given by

$$N = \Big\lceil \frac{\max\{(\log \widetilde{S})^{1+\xi},\ (\log p)^{\varepsilon}\}}{\delta} \Big\rceil + 1 = 18.$$

Regarding $j \in \widetilde{\boldsymbol{D}}$, we took additional samples x_{ijl}, $l = m+1, ..., m+N$, of size $N = 18$ from each Π_i, which are given in Table 4.

	Π_1: Normal Prostate			Π_2: Prostate Tumors		
$j \backslash l$	19	$\cdots$	36	19	$\cdots$	36
192	9.859	$\cdots$	8.973	9.338	$\cdots$	10.212
198	8.622	$\cdots$	7.077	6.120	$\cdots$	7.724
$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$
12432	9.884	$\cdots$	9.091	8.00	$\cdots$	9.388

Table 4. Additional samples, $x_{ijl}, j \in \widetilde{\boldsymbol{D}}$ ($\widetilde{S} = 160$, $N = 18$)

We selected significant variables by $\widehat{D} = \{j \in \widetilde{D} \mid \text{rejecting } H_{0j}\} = \{j \in \widetilde{D} \mid |\overline{x}_{1j(N)} - \overline{x}_{2j(N)}|^2 > 1.5\}$ and finally obtained

$$\widehat{D} = \{556, 7412, 8662, 11552\} \tag{26}$$

with $\widehat{S} = |\widehat{D}| = 4$. For $j \in \widehat{\boldsymbol{D}}$, we calculated $\bar{x}_{ijm+N} = \sum_{l=1}^{m+N} x_{ijl}/(m+N)$ for each Π_i and obtained estimates of $\mu_{1j} - \mu_{2j}$ for $j \in \widehat{\boldsymbol{D}}$ as

$$\{\bar{x}_{1jm+N} - \bar{x}_{2jm+N} |\ j \in \widehat{\boldsymbol{D}}\} = \{-1.511, -1.472, -1.79, -2.148\}.$$

The required sample-size in the two-stage variable selection procedure was $m + N = 36$ for each Π_i. On the other hand, the required sample-size in the single variable selection procedure given by (20) with (19) was $n_i \geq (\log p)^{1+\zeta}/\delta = 59.43$ with $\zeta = 1.0$. The two-stage variable selection procedure allows the experimenter to reduce the cost of sampling in the second stage.

5.2.2 Classification after variable selection

In Section 4, we considered a two-stage discriminant procedure in HDLSS data situations. In some cases the experimenter would encounter the situation that the required sample size, N_i, is much larger than the available sample size if $\Delta_\star = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2 + (\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2))^2 / \max_{i=1,2}\{2\mathrm{tr}(\boldsymbol{\Sigma}_i)\}$ is much smaller than $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$'s. In that case, we recommend that the experimenter should consider the classification based only on the selected variables. We selected a set of significant variables by $\widehat{D} = \{556, 7412, 8662, 11552\}$ that was given in (26). We set $n_1 = n_2 = m + N = 36$, where m and N were given in Section 5.2.1. Let us write the selected 4-variable data as $\boldsymbol{x}_{il} = (x_{i556l}, x_{i7412l}, x_{i8662l}, x_{i11552l})^T$, $i = 1, 2$, for the l-th sample. Then, we considered a typical quadratic discriminant rule that classifies $\boldsymbol{x}_0$ into Π_1 if

$$(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1})^T \boldsymbol{S}_{1n_1}^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1}) - \log\left\{\frac{\det(\boldsymbol{S}_{2n_2})}{\det(\boldsymbol{S}_{1n_1})}\right\} < (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2})^T \boldsymbol{S}_{2n_2}^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2}), \qquad (27)$$

and into Π_2 otherwise, where $\boldsymbol{x}_0$ is an observation vector with respect to the 4 variables on an individual belonging to Π_1 or to Π_2, $\overline{\boldsymbol{x}}_{in_i} = \sum_{l=1}^{n_i} \boldsymbol{x}_{il}/n_i$ and $\boldsymbol{S}_{in_i} = \sum_{l=1}^{n_i} (\boldsymbol{x}_{il} - \overline{\boldsymbol{x}}_{in_i})(\boldsymbol{x}_{il} - \overline{\boldsymbol{x}}_{in_i})^T/(n_i - 1)$, $i = 1, 2$.

We compared the discriminant rule given by (27) after variable selection with those given by (10) with $\gamma = 0$, DLDR and DQDR. Note that the three competitors were constructed by using the original (12600-variable) data without variable selection. In Table 5, we investigated those performances by using test data sets consisting of $50 - n_1 = 14$ remaining samples from Π_1 and $52 - n_2 = 16$ remaining samples from Π_2. We observed that the discriminant rule given by (10) with $\gamma = 0$ showed a bad performance for $\boldsymbol{x}_0$ classified into Π_1: Normal Prostate. We inspected the condition of Theorem 4.1 for the data sets and found that $\max_{i=1,2}\{\mathrm{tr}(\boldsymbol{S}_{in_i(1)}\boldsymbol{S}_{in_i(2)})\}/(\widehat{\Delta}_\star^2 \min_{i=1,2}\{n_i\}) = 0.15$ according to Remark 4.4 so that $\max_{i=1,2}\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\}/(\Delta_\star^2 \min_{i=1,2}\{n_i\})$ seems not to be sufficiently small. This may be a reason why Theorem 4.1 is not applicable to the present data sets. On the other hand, we observed that the discriminant rule given by (27) after variable selection showed a good performance when compared to the competitors. We recommend that the experimenter should consider

	(27) after variable selection	(10) with $\gamma = 0$	DLDR	DQDR
1-$\overline{e(2\|1)}$	10/14 (=0.714)	4/14 (=0.286)	4/14 (=0.286)	4/14 (=0.286)
1-$\overline{e(1\|2)}$	15/16 (=0.938)	15/16 (=0.938)	15/16 (=0.938)	15/16 (=0.938)

Table 5. The correct discrimination rates by (27) after variable selection, (10) with $\gamma = 0$, DLDR and DQDR for test data sets consisting of 14 samples from Π_1 and 16 samples from Π_2.

the classification after variable selection if $\widehat{\Delta}_\star$ is not large enough to claim the condition of Theorem 4.1 or to claim the assertion in Theorem 4.4 within the available sample size.

6. Acknowledgments

Research of the first author was partially supported by Grant-in-Aid for Scientific Research (B) and Challenging Exploratory Research, Japan Society for the Promotion of Science (JSPS), under Contract Numbers 22300094 and 23650142. Research of the second author was partially supported by Grant-in-Aid for Young Scientists (B), JSPS, under Contract Number 23740066.

7. References

[1] Aoshima, M. & Yata, K. (2011). Two-stage procedures for high-dimensional data, *Sequential Analysis* (Editor's Special Invited Paper), to appear.

[2] Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R. den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R. & Korsmeyer, S.J. (2001). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia, *Nature Genetics*, Vol. 30, 41-47.

[3] Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J. & Foa, R. (2004). Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival, *Blood*, Vol. 103, 2771-2778.

[4] Dudoit, S., Fridlyand, J. & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of American Statistical Association*, Vol. 97, 77-87.

[5] Jeffery, I.B., Higgins, D.G. & Culhane, A.C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data, *Bioinformatics*, Vol. 7, 359.

[6] Johnstone, I.M. & Lu, A.Y. (2009). On consistency and sparsity for principal components analysis in high dimensions, *Journal of American Statistical Association*, Vol. 104, 682-693.

[7] Pochet, N., De Smet, F., Suykens, J.A. & De Moor, B.L. (2004). Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction, *Bioinformatics*, Vol. 20, 3185-3195.

[8] Pollard, K.S., Dudoit, S. & van der Laan, M.J. (2005). Multiple testing procedures: R multtest package and applications to genomics, In: Gentleman, R., Carey, V., Huber, W., Irizarry, R. & Dudoit, S. (ed.) Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, New York, pp 249-271.

[9] Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R. & Sellers, W.R. (2002). Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, Vol.1(2), 203-209.

[10] Yata, K. (2010). Effective two-stage estimation for a linear function of high-dimensional Gaussian means, *Sequential Analysis*, Vol. 29, 463-482.

[11] Yata, K. & Aoshima, M. (2009). PCA consistency for non-Gaussian data in high dimension, low sample size context, *Communications in Statistics -Theory & Methods*, Special Issue Honoring Zacks, S. (ed. Mukhopadhyay, N.), Vol. 38, 2634-2652.

[12] Yata, K. & Aoshima, M. (2010a). Intrinsic dimensionality estimation of high-dimension, low sample size data with *d*-asymptotics, *Communications in Statistics -Theory* & *Methods*, Special Issue Honoring Akahira, M. (ed. Aoshima, M.), Vol. 39, 1511-1521.
[13] Yata, K. & Aoshima, M. (2010b). Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix, *Journal of Multivariate Analysis*, Vol. 101, 2060-2077.
[14] Yata, K. & Aoshima, M. (2011a). Inference on high-dimensional mean vectors with fewer observations than the dimension, *Methodology and Computing in Applied Probability*, in press.
[15] Yata, K. & Aoshima, M. (2011b). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations, *Journal of Multivariate Analysis*, revised.