Authors' Response

Makoto Aoshima[1] **and Kazuyoshi Yata**[2]
[1,2]Institute of Mathematics, University of Tsukuba, Ibaraki, Japan

Abstract: In this paper, we respond to the comments made by the ten discussants on "Two-Stage Procedures for High-Dimensional Data". We also give some new results along with their brief explanations.

Keywords: Classification; Confidence region; Cross-data-matrix methodology; HDLSS; Robustness; Sample size determination; Two-sample test; Variable selection.

Subject Classifications: 62L10, 62H10; 60F05.

1. OVERALL VIEW

First of all, we would like to thank the Editor, Professor Mukhopadhyay, for inviting the article "Two-Stage Procedures for High-Dimensional Data" and for organizing the discussions. We are very grateful to Professors Ahn, Chen, Ing, Lai, Lee, Mukhopadhyay, Panchapakesan, Qin, Solanky and Takada for their inspiring and insightful contributions in discussing the article. They have reflected upon deep theory and interesting applications.

In Section 2, we discuss assumptions (A-i) to (A-v) according to Prof. Solanky's suggestions. We address the robustness issues of the proposed two-stage procedures.

In Section 3, we consider a confidence region with a given diameter for high-dimensional data according to Prof. Takada's suggestions. We introduce a result obtained by Yata and Aoshima (2011a).

In Section 4, we include a new result about two-sample test for high-dimensional data according to Prof. Mukhopadhyay's suggestions. We propose a modified two-sample test procedure in order to improve the power of the original test procedure. We compare the original test procedure with the modified test procedure in terms of accuracy and required sample size.

In Section 5, we discuss the efficiency of the estimators of $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ according to Prof. Qin's suggestions. We emphasize that the estimator induced by the *cross-data-matrix methodology* is quite robust, simple, and computationally efficient.

In Section 6, we explain about how to use the two-stage classification procedure according to Profs. Ahn and Lee's suggestions. We give practical guidelines concerning accuracy and required sample size.

Finally, in Section 7, we discuss a high-dimensional variable selection problem according to Profs. Chen and Panchapakesan's and Profs. Ing and Lai's suggestions.

Address correspondence to Makoto Aoshima, Institute of Mathematics, University of Tsukuba, Ibaraki 305-8571, Japan; Fax: +81-298-53-6501; E-mail: aoshima@math.tsukuba.ac.jp

2. DISCUSSION ON THE ASSMUMPTIONS

Prof. Solanky made specific comments on assumptions (A-i) to (A-v):

(I) How to handle the issue of increasing number of unknown parameters in the asymptotics as $p \to \infty$?

(II) How good are the asymptotic results when p is not too large? Say, when $p = 50$ in the cases considered in the Table 1.

(III) How rigid are assumptions (A-i) to (A-v)? How would a practitioner verify whether or not these assumptions are true in a particular case?

(IV) What would happen if some of these assumptions are not true?

(V) Would a fine-tuned purely sequential procedure along the lines of Mukhopadhyay and Datta (1995), which does not rely on any such lower bounds, perform better?

As for (I), we used the *cross-data-matrix methodology* created by Yata and Aoshima (2010a,b). With the help of this methodology, we could handle the issue effectively in high-dimensional situations. Recently, Aoshima and Yata (2011) proposed a method called the *generalized cross-data-matrix methodology* that is based on resampling and gave a variety of illustrations using microarray data.

As for (II) and (III), we investigated the performance of the two-stage procedure when $p = 50$ and 100. We set $\delta = 2.5$ and $\alpha = 0.05$. Along the lines of Section 2.3, independent pseudorandom normal observations were generated for $\pi_i : N_p(\mathbf{0}, \boldsymbol{\Sigma}_i)$, $i = 1, 2$, where $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$ and $\boldsymbol{B}$ is defined by (2.9) in Section 2.3. Finally, we had $\bar{n}_1 = 19.05$ $(\bar{n}_1 - C_1 = -0.98)$, $\bar{n}_2 = 19.23$ $(\bar{n}_2 - C_2 = -0.8)$ and $\overline{P} = 0.915$ when $p = 50$ $(m = 7)$, and $\bar{n}_1 = 28.51$ $(\bar{n}_1 - C_1 = -0.18)$, $\bar{n}_2 = 28.50$ $(\bar{n}_2 - C_2 = -0.19)$ and $\overline{P} = 0.942$ when $p = 100$ $(m = 10)$. We emphasize that the asymptotic results depend not only on the magnitude of p and n_i's but also on the assumption that $\mathrm{tr}(\boldsymbol{\Sigma}_i^4)/p^2 \to 0$ as $p \to \infty$ in (A-iv). Note that one can verify whether (A-iv) is true or not very easily by using the *cross-data-matrix methodology* (Yata and Aoshima, 2010a,b) or the *noise-reduction methodology* (Yata and Aoshima, 2011b). See Remark 1.1 of Section 1 for the verification of the assumptions.

As for (IV), let us consider the case that the assumptions are not true. For example, for the inference in Section 2, we can handle the problem as follows: We simply change $z_{\alpha/2}$ to $1/\sqrt{\alpha}$ in the sample size determination given by (2.4). Then, from Chebyshev's inequality, it holds as $p \to \infty$ that

$$P_{\boldsymbol{\theta}}(|||\boldsymbol{T}_{\mathbf{n}} - \boldsymbol{\mu}||^2 - \widehat{\Sigma}_{\mathbf{n}}| \geq \delta) \leq Var_{\boldsymbol{\theta}}(||\boldsymbol{T}_{\mathbf{n}} - \boldsymbol{\mu}||^2 - \widehat{\Sigma}_{\mathbf{n}})/\delta^2 \leq \alpha + o(1).$$

Thus we can claim Theorem 2.3 without (A-i) to (A-v). Similarly, by changing $z_{\alpha/2}$ to $1/\sqrt{\alpha}$ in (2.7), we can claim Theorem 2.4 under mild assumptions that $E(z_{ijl}^2 z_{isl}^2) = 1$ and $E(z_{ijl}z_{isl}z_{itl}z_{iul}) = 0$, $j \neq s, t, u$ for $i = 1, ..., k$, but without (A-i) to (A-v). We emphasize that the results in the article are quite robust by introducing a slight modification in the sample size determination for each inference.

Finally, as for (V), we emphasize that the proposed two-stage procedures are quite robust for the misidentification of lower bounds. See Remark 2.3 of Section 2 for the details. In addition, if the experimenter is concerned with the cost of each sampling, the two-stage procedure would be the most likely candidate in a real world. We conclude this section by quoting Profs. Ahn and Lee's encouraging comment: "*Based on the limited empirical study on the classification that we carried out, at least their classification method turns out to be quite robust to the choice of $\Delta_\star$.*"

3. CONFIDENCE REGION WITH A GIVEN DIAMETER FOR HIGH-DIMENSIONAL DATA

Prof. Takada showed that the confidence region with a given diameter is also available for high-dimensional data. For simplicity he assumed that $\boldsymbol{x}_i$'s are distributed as $N_p(\boldsymbol{\mu}, \sigma^2\boldsymbol{I}_p)$. For given $W(>0)$, let

$$\boldsymbol{R}_n = \{\boldsymbol{\mu} \in R^p : ||\overline{\boldsymbol{x}}_n - \boldsymbol{\mu}||^2 \le W\},$$

where $\overline{\boldsymbol{x}}_n = \sum_{i=1}^n \boldsymbol{x}_i/n$. If the sample size n is determined by

$$n \ge \frac{\chi_p^2(\alpha)\sigma^2}{W} \quad (= C, \text{ say}), \tag{3.1}$$

one has that $P(\boldsymbol{\mu} \in \boldsymbol{R}_n) \ge 1 - \alpha$ for given $\alpha \in (0, 1)$. Here, $\chi_p^2(\alpha)$ denotes the upper α point of a chi-square distribution with p degrees of freedom. Then, he showed that if W is chosen such that $W = W(p)$ and $W \to \infty$ as $p \to \infty$, it holds that $C = o(p)$ as $p \to \infty$.

Recently, we have given a result related to Prof. Takada's concern: Suppose there are independent and p-variate populations, π_i, $i = 1, ..., k$, having mean vector, $\boldsymbol{\mu}_i$, and covariance matrix, $\boldsymbol{\Sigma}_i\ (> \boldsymbol{O})$, for each π_i. Let $\boldsymbol{\mu} = \sum_{i=1}^k b_i\boldsymbol{\mu}_i$, where b_i's are known and nonzero scalars. Let $\boldsymbol{T}_{\mathbf{n}} = \sum_{i=1}^k b_i\overline{\boldsymbol{x}}_{in_i}$, where $\boldsymbol{n} = (n_1, ..., n_k)$ and $\overline{\boldsymbol{x}}_{in_i} = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}/n_i$. Define a confidence region for $||\boldsymbol{T}_{\mathbf{n}} - \boldsymbol{\mu}||^2$ by

$$\boldsymbol{R}_{\mathbf{n},W} = \{\boldsymbol{\mu} \in R^p : ||\boldsymbol{T}_{\mathbf{n}} - \boldsymbol{\mu}||^2 \le W\}.$$

Our goal is to construct $\boldsymbol{R}_{\mathbf{n},W}$ satisfying

$$P_{\boldsymbol{\theta}}(\boldsymbol{\mu} \in \boldsymbol{R}_{\mathbf{n},W}) \ge 1 - \alpha \tag{3.2}$$

for given $W\ (> 0)$ and $\alpha \in (0, 1/2)$. As we mentioned at the beginning of Section 2, $\boldsymbol{R}_{\mathbf{n},W}$ satisfying (3.2) is not available for a given and fixed $W(> 0)$ in the HDLSS context. Thus we assume that $W = W(p) \to \infty$ as $p \to \infty$ and $W/\min_{1\le i\le k} \mathrm{tr}(\boldsymbol{\Sigma}_i) = o(1)$. We find the sample size for each π_i as

$$n_i \ge \frac{1}{W}|b_i|\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_i)}\sum_{j=1}^k |b_j|\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_j)} + \frac{z_\alpha\sqrt{2}}{W}|b_i|\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_i)}\sum_{j=1}^k |b_j|\sqrt{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_j^2)}{\mathrm{tr}(\boldsymbol{\Sigma}_j)}} \quad (= \tilde{C}_i, \text{ say}), \tag{3.3}$$

where z_α is the upper α point of $N(0, 1)$. Note that $\tilde{C}_i = W^{-1}|b_i|\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_i)}\sum_{j=1}^k |b_j|\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_j)} + o(p/W) = o(p)$ under (A-iv). Then, we have the following theorem.

Theorem 3.1 (Yata and Aoshima, 2011a). *Assume (A-iv) and either (A-ii) or (A-iii) with (A-v). For n_i's satisfying (3.3), it holds as $p \to \infty$ that*

$$\liminf P_{\boldsymbol{\theta}}(\boldsymbol{\mu} \in \boldsymbol{R}_{\mathbf{n},W}) \ge 1 - \alpha.$$

See Yata and Aoshima (2011a) for the details and further discussions. When $k=1$, $b_i=1$ and $\boldsymbol{\Sigma}_i=\sigma^2\boldsymbol{I}_p$, note that $\tilde{C}_i=(p+z_\alpha\sqrt{2p})\sigma^2/W$ from (3.3). Then, $\tilde{C}_i$ is asymptotically equivalent to C, given by (3.1), when $p\to\infty$.

4. TWO-SAMPLE TEST FOR HIGH-DIMENSIONAL DATA

Prof. Mukhopadhyay gave some comments on the accuracy of Tables 2 and 3:

(I) Why is it that we are seeing more than usual incidences of over/undershooting of the set targets in Tables 2 and 3 compared with what we find in Table1?

(II) It may be possible to somehow fine-tune the associated procedures so that the asymptotics may possibly kick in earlier (that is, for smaller C). If this may be achieved satisfactorily, then I believe that the proposed methodologies will become more apt for practical implementations.

As for (I) about overshooting the target in Table 3, it is quite natural because the classification procedure was given for $\Delta_L(<\Delta_\star)$ to claim the accuracy such that $e(2|1)<\alpha$ and $e(1|2)<\beta$. In the simulation, we set a common Δ_L for brevity that was much smaller than some $\Delta_\star$'s on the safe side. Thus it is reasonable to overshoot the targets in Table 3. As for undershooting the target in Table 2, it occurs because p was not large enough to reach the targets in simulations. However, as for (II), we can consider fine-tuning so that the target is claimed for earlier p. We modify the sample size determination by (3.4) in Section 3.1 as follows:

$$\begin{aligned} n_i \geq &\frac{(z_\alpha+z_\beta)\sqrt{2}}{\Delta_L}\mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/4}\sum_{j=1}^{2}\mathrm{tr}(\boldsymbol{\Sigma}_j^2)^{1/4} \\ &+\frac{2(z_\alpha+z_\beta)^2}{\Delta_L^2}\mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/4}\sum_{j=1}^{2}\frac{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_j(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{\mathrm{tr}(\boldsymbol{\Sigma}_j^2)^{1/4}} \quad (=C_i,\ \text{say}). \end{aligned} \tag{4.1}$$

Note that $C_i=O(p^{1/2}/\Delta)=o(p)$ under (A-iv). Then, we can claim Theorem 3.2. It holds from (3.3) in Section 3.1 that

$$Var_{\boldsymbol{\theta}}(\widetilde{T}_{\mathbf{n}})=\left(\sum_{i=1}^{2}\frac{2}{n_i(n_i-1)}\mathrm{tr}(\boldsymbol{\Sigma}_i^2)+\frac{4}{n_1n_2}\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\right)(1+O(u^2))$$

for n_i's satisfying (4.1), where $u=\max_{i=1,2}\{n_i(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_i(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)/\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\}$. Since one had for n_i's satisfying (3.4) in Section 3.1 that

$$Var_{\boldsymbol{\theta}}(\widetilde{T}_{\mathbf{n}})=\left(\sum_{i=1}^{2}\frac{2}{n_i(n_i-1)}\mathrm{tr}(\boldsymbol{\Sigma}_i^2)+\frac{4}{n_1n_2}\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\right)(1+O(u)),$$

the power of the test procedure is possibly improved by the modified sample size determination. We proceed the following two steps:

[Modified two-sample test procedure]

1. According to (3.7) in Section 3.2, take pilot samples of size m from each π_i. Let $m_1 = [m/2]+1$ and $m_2 = m - m_1$. Then, calculate $\overline{\boldsymbol{x}}_{im_1}$, $\overline{\boldsymbol{x}}_{im_2}$, $\overline{\boldsymbol{x}}_{im}$, $\boldsymbol{S}_{im(1)}$, $\boldsymbol{S}_{im(2)}$ and $\boldsymbol{S}_{im}$ for each π_i according to (1.2) in Section 1. Let

$$U_i = \big\{(\overline{\boldsymbol{x}}_{im_1} - \overline{\boldsymbol{x}}_{jm})^T \boldsymbol{S}_{im(2)}(\overline{\boldsymbol{x}}_{im_1} - \overline{\boldsymbol{x}}_{jm}) + (\overline{\boldsymbol{x}}_{im_2} - \overline{\boldsymbol{x}}_{jm})^T \boldsymbol{S}_{im(1)}(\overline{\boldsymbol{x}}_{im_2} - \overline{\boldsymbol{x}}_{jm})\big\}/2 \\ - \frac{\mathrm{tr}(\boldsymbol{S}_{im(1)}\boldsymbol{S}_{im(2)})m}{2m_1m_2} - \frac{\mathrm{tr}((\boldsymbol{S}_{im(1)}+\boldsymbol{S}_{im(2)})\boldsymbol{S}_{jm})}{2m} \tag{4.2}$$

with $j(\neq i)$. Here, U_i was given by Yata and Aoshima (2011a) as an unbiased estimator of $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_i(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, i.e., $E_{\boldsymbol{\theta}}(U_i) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_i(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Then, define the total sample size for each π_i by

$$N_i = \max\Big\{m,\ \Big[\frac{(z_\alpha + z_\beta)\sqrt{2}}{\Delta_L}\mathrm{tr}(\boldsymbol{S}_{im(1)}\boldsymbol{S}_{im(2)})^{1/4}\sum_{j=1}^{2}\mathrm{tr}(\boldsymbol{S}_{jm(1)}\boldsymbol{S}_{jm(2)})^{1/4} \\ + \max\Big\{\frac{2(z_\alpha+z_\beta)^2}{\Delta_L^2}\mathrm{tr}(\boldsymbol{S}_{im(1)}\boldsymbol{S}_{im(2)})^{1/4}\sum_{j=1}^{2}\frac{U_j}{\mathrm{tr}(\boldsymbol{S}_{jm(1)}\boldsymbol{S}_{jm(2)})^{1/4}},\ 0\Big\}\Big]+1\Big\}. \tag{4.3}$$

Let $\boldsymbol{N} = (N_1, N_2)$.

2. Take additional samples of size $N_i - m$ from each π_i. By combining the initial samples and the additional samples, calculate $\tilde{T}_{\mathbf{N}}$ according to (3.2) in Section 3.1. Then, test the hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ vs. $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ by

$$\text{rejecting } H_0 \Longleftrightarrow \tilde{T}_{\mathbf{N}} > \frac{\Delta_L z_\alpha}{z_\alpha + z_\beta}. \tag{4.4}$$

We have the following theorem.

Theorem 4.1. *Assume either (A-ii) or (A-iii) with (A-v). Assume that* $\max_{1\le i\le 2}\{tr(\boldsymbol{\Sigma}_i^4)\} = O(p\Delta_L^2)$. *For the test given by (4.4) with (4.3), it holds as* $p \to \infty$ *that*

$$\limsup \mathit{size} \le \alpha \quad \mathit{and} \quad \liminf \mathit{poewer}(\Delta_L) \ge 1-\beta,$$

where power(Δ_L) is the power when $\Delta = \Delta_L$.

Proof. In a way similar to the proof of Theorem 3.3 in Yata and Aoshima (2011a), we obtain the result. We omit the details for brevity. □

Remark 4.1. Contrary to Theorem 3.4 in Section 3.2, we cannot claim the asymptotic second-order efficiency for (4.3).

In order to compare the performance of the test procedure given by (4.4) with the original one in Table 2, we conducted simulation studies for earlier p in the same setup as in Section 3.3. Our goal was to construct a test with size $\alpha = 0.05$ and power no less than $1-\beta = 0.9$ when $\Delta \ge 10$. We obtained m=7, 19 and 27 from (3.7) in Section 3.2 for p=100, 800 and 1600, respectively. In view of the following table, we observed that the test procedure given by (4.4) improved performances about the targets, α and $1-\beta$, for earlier p. However, $\overline{n} - C$ was much larger than that for (3.8) in Section 3.2. In other words, the original test procedure in Section 3.2 gives good performances not only about the targets, α and $1-\beta$, for larger p but also about the asymptotic second-order efficiency, $\overline{n} - C$.

Table 1. Required sample size and the size and power by (4.4) with (4.3).

C	$\overline{n}$	$\overline{n}-C$	$Var(n/C)$	$\overline{\alpha}$	$s(\overline{\alpha})$	$1-\overline{\beta}$	$s(\overline{\beta})$
When $p = 100$: m=7 for $\Delta = 0$							
26.51	45.36	18.85	1.565	0.033	0.00397		
11.92	20.60	8.68	1.685				
14.59	24.76	10.17	1.593				
When $p = 100$: m=7 for $\Delta = 10$							
38.94	52.37	13.44	0.955			0.891	0.00698
17.50	23.74	6.23	1.049				
21.44	28.64	7.20	0.957				
When $p = 800$: m=19 for $\Delta = 0$							
75.89	101.78	25.89	0.305	0.030	0.00378		
34.11	45.85	11.74	0.315				
41.78	55.92	14.15	0.301				
When $p = 800$: m=19 for $\Delta = 10$							
87.25	110.64	23.38	0.293			0.917	0.00617
39.22	49.91	10.69	0.302				
48.03	60.72	12.69	0.290				
When $p = 1600$: m=27 for $\Delta = 0$							
107.42	137.20	29.78	0.187	0.033	0.00399		
48.28	61.75	13.47	0.191				
59.14	75.45	16.31	0.186				
When $p = 1600$: m=27 for $\Delta = 10$							
118.71	145.82	27.12	0.199			0.916	0.00620
53.36	65.61	12.25	0.201				
65.35	80.22	14.87	0.199				

5. ESTIMATION OF $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$

Prof. Qin gave some comments on the estimators of $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$. She compared two estimators: (1) $\mathrm{tr}(\boldsymbol{S}_{in_i(1)}\boldsymbol{S}_{in_i(2)})$ used in our paper, that was given by Yata (2010) as an application of the *cross-data-matrix methodology* created by Yata and Aoshima (2010a,b); (2) $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_i^2})$ given by Chen and Qin (2010), that was discussed in Remark 3.2 of Section 3.1. She claimed that $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_i^2})$ outperformed $\mathrm{tr}(\boldsymbol{S}_{in_i(1)}\boldsymbol{S}_{in_i(2)})$ under the assumption that $||\boldsymbol{\mu}_2||^2 = 0$ or $||\boldsymbol{\mu}_2||^2$ is sufficiently small in her Monte Carlo simulation under $||\boldsymbol{\mu}_1||^2 = 0$. We agree to that $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_i^2})$ slightly outperformed $\mathrm{tr}(\boldsymbol{S}_{in_i(1)}\boldsymbol{S}_{in_i(2)})$ in the cases. However, it seems that there is not much significant difference between them when we observe her simulation results.

Let us consider an easy example such as $p = 1000$ and $||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2 = 0$ by setting $\boldsymbol{\mu}_i = (1, ..., 1)^T$, $\boldsymbol{\Sigma}_i = \boldsymbol{I}_p$ and $n_i = 10$ for $i = 1, 2$. Then, from Remark 3.2 of Section 3.1, we have that $E_{\boldsymbol{\theta}}(\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_i^2})) = \mathrm{tr}(\boldsymbol{\Sigma}_i^2) + \boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i\boldsymbol{\mu}_i/(n_i-2) = p+p/(n_i-2) = 1125 > \mathrm{tr}(\boldsymbol{\Sigma}_i^2) = 1000$. It seems that the bias of $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_i^2})$ is not something that can be ignored.

In a real data analysis, it is quite possible to handle a situation such as $||\boldsymbol{\mu}_i||^2 = O(p)$. We emphasize that $\mathrm{tr}(\boldsymbol{S}_{in_i(1)}\boldsymbol{S}_{in_i(2)})$ is always unbiased such that $E_{\boldsymbol{\theta}}(\mathrm{tr}(\boldsymbol{S}_{in_i(1)}\boldsymbol{S}_{in_i(2)})) = \mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ without the assumption that $||\boldsymbol{\mu}_i||^2 = 0$ or $||\boldsymbol{\mu}_i||^2$ is sufficiently small. Moreover, $\mathrm{tr}(\boldsymbol{S}_{in_i(1)}\boldsymbol{S}_{in_i(2)})$ is quite simple and computationally efficient as Prof. Qin kindly mentioned.

6. HOW TO USE THE TWO-STAGE CLASSIFICATION PROCEDURE

Profs. Ahn and Lee gave some comments on the two-stage classification procedure in Section 4. They implemented the procedure for two microarray data sets: Leukemia ($n = 72$, $n_1 = 47$, $n_2 = 25$, $p = 7129$) in Golub et al. (1999) and Colon cancer data ($n = 62$, $n_1 = 22$, $n_2 = 40$, $p = 2000$) in Alon et al. (1999). The comments are summarized as follows:

(I) They suggested that it might be desirable to use different values of $\tau_\star$ or Δ_L for different classes.

(II) From Figure 2 in their comment, they reported that $e(1|2)$ was a lot higher than the targeted level for Colon cancer data.

(III) They were also concerned about a case that the required sample size is larger than what one currently has in the data.

As for (I), in the two-stage classification procedure, 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 assertions in Theorems 4.3 and 4.4 can be still claimed.

As for (II), we have to emphasize that the assertions in Theorems 4.2 and 4.3 can be claimed under $\Delta_\star \geq \Delta_L$. Their setting of Δ_L seems to be larger than $\Delta_\star$. In a real data analysis, it is crucial to fix Δ_L as careful as possible. One may estimate $\Delta_\star$ by

$$||\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}||^2 - \sum_{i=1}^{2}\frac{\mathrm{tr}(\boldsymbol{S}_{in_i})}{n_i} + \frac{|\mathrm{tr}(\boldsymbol{S}_{1n_1}) - \mathrm{tr}(\boldsymbol{S}_{2n_2})|^2}{2\max_{i=1,2}\mathrm{tr}(\boldsymbol{S}_{in_i})} \ (= \widehat{\Delta}_{\star,\mathbf{n}},\ \text{say}). \tag{6.1}$$

By using $\widehat{\Delta}_{\star,\mathbf{n}}$, let us provide an elastic two-stage procedure:

[Elastic two-stage classification procedure]

1. Choose m (>4) such as $m/C_{i\star} \in (0,1)$, $i=1,2$, as $p \to \infty$, where $C_{i\star}$ is defined by C_i when $\Delta_L = \Delta_\star$. We assume $\Delta_\star = o(p^{1/2}) > 0$. Take pilot samples of size m from each π_i and calculate $\widehat{\Delta}_{\star,\mathbf{m}}$ according to (6.1), where $\mathbf{m} = (m,m)$. Then, we note that $\widehat{\Delta}_{\star,\mathbf{m}}/\Delta_\star = 1 + o_p(1)$ as $p \to \infty$ under (A-iv) and either (A-ii) or (A-iii). Let $\Delta_L = \widehat{\Delta}_{\star,\mathbf{m}}$. Define the total sample size for each π_i by (4.6) in Section 4.2.

2. Follow the second step in the two-stage classification procedure.

Let us consider the Colon cancer data having $n_1 = 22$ and $n_2 = 40$ samples. We set $m = 10$. First, we took the first 10 samples as a pilot sample from each π_i. According to (6.1), we calculated $\widehat{\Delta}_{\star,\mathbf{m}} = 4.45 \times 10^7$. Along the lines of the elastic two-stage classification procedure, we set $\Delta_L = 4.45 \times 10^7$. We also calculated $\mathrm{tr}(\boldsymbol{S}_{1m(1)}\boldsymbol{S}_{1m(2)}) = 1.001 \times 10^{16}$ and $\mathrm{tr}(\boldsymbol{S}_{2m(1)}\boldsymbol{S}_{2m(2)}) = 1.09 \times 10^{16}$. When we set $\alpha = \beta = 0.1$, we had $N_1 = 71$ and $N_2 = 72$. The required sample sizes were larger than the available sample sizes, i.e., $N_1 > 22$ and $N_2 > 40$.

As for (III), when the experimenter encounters such a case in which the required sample size is larger than the available sample size, one may tune α and β to slightly upper. For example, when we set $\alpha = \beta = 0.2$ for the Colon cancer data, we had $N_1 = 31$ and $N_2 = 31$. We considered iterating to randomly choose 10 samples from each π_i and calculate N_1 and N_2 for $\alpha = \beta = 0.2$. We repeated the operation 100 times and obtained the average of (N_1, N_2) as $(28.16, 33.73)$. It seems that one cannot always construct the classification rule, given by (4.7) in Section 4.2, for α, $\beta \le 0.2$ as long as the available samples are of sizes $(n_1, n_2) = (22, 40)$. If the experimenter is concerned with an original setting of (α, β), we would like to recommend that one should consider the classification after variable selection given by Section 7.

7. DISCUSSIONS ON HIGH-DIMENSIONAL VARIABLE SELECTION

Profs. Chen and Panchapakesan gave some comments on “large p, small n” multinomial problems. They suggested that it is important to consider a parallel development of asymptotic theory for a multinomial distribution where p is the number of categories and n is the sample size. They also suggested that asymptotic theory for large p could be useful to obtain good approximate results for some inferential problems involving a large number of rare species and parliamentary type election situations in which we commonly find many minor party candidates and several independent candidates.

We understand that the problems they raised are hot and challenging topics. We have several ideas to approach to those topics. The second problem is related to some inferential problems involving a larger number of nuisance cells. One of the starting points would be Aoshima et al. (2003). We would like to handle the “large p, small n” multinomial problems in future research.

Profs. Ing and Lai introduced a fixed-width confidence interval for prediction in conjunction with variable selection for high-dimensional sparse linear regression models as $p \to \infty$. They mentioned that it is well known that fully sequential procedures for fixed-width confidence intervals have advantages over two-stage counterparts.

Although we cannot simply compare our two-stage procedures with their fully sequential procedure that handles a different problem in a different setup, it will not be easy to say something about superiority or inferiority in high-dimensional data situations. Some new criteria would be expected to evaluate in terms of (i) time, (ii) cost, and (iii) easy to use for the implementation of the procedure. In addition, if the experimenter is concerned with the cost of each sampling, the two-stage procedure would be a likely candidate in a real world.

ACKNOWLEDGMENTS

We would like to thank the Editor-in-Chief, Prof. Nitis Mukhopadhyay, for giving us this opportunity to make responses to the discussants' valuable comments. 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.

REFERENCES

Alon, U., Barkai, N., Notterman, D., Gish, K., Mack, S., and Levine, J. (1999). Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays, *Proceedings of National Academy of Sciences of United States of America*, 96: 6745-6750.

Aoshima, M., Chen, P., and Panchapakesan, S. (2003). Sequential Procedures for Selecting the Most Probable Multinomial Cell When a Nuisance Cell is Present, *Communications in Statistics-Theory* & *Methods* 32: 893-906.

Aoshima, M. and Yata, K. (2011). Effective Methodologies for Statistical Inference on Microarray Studies, submitted.

Chen, S. X. and Qin, Y.-L. (2010). A Two-Sample Test for High-Dimensional Data with Applications to Gene-Set Testing, *Annals of Statistics* 38: 808-835.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuti, M. A., and Bloomeld, C. D. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science* 286: 531-537.

Mukhopadhyay, N. and Datta, S. (1995). On Fine-Tuning a Purely Sequential Procedure and Associated Second-Order Properties, *Sankhya, Series A* 57: 100-117.

Yata, K. (2010). Effective Two-Stage Estimation for a Linear Function of High-Dimensional Gaussian Means, *Sequential Analysis* 29: 463-482.

Yata, K. and Aoshima, M. (2010a). Intrinsic Dimensionality Estimation of High-Dimension, Low Sample Size Data with d-Asymptotics, *Communications in Statistics-Theory* & *Methods*, Special Issue Honoring M. Akahira, M. Aoshima ed., 39: 1511-1521.

Yata, K. and Aoshima, M. (2010b). Effective PCA for High-Dimension, Low-Sample-Size Data with Singular Value Decomposition of Cross Data Matrix, *Journal of Multivariate Analysis* 101: 2060-2077.

Yata, K. and Aoshima, M. (2011a). Inference on High-Dimensional Mean Vectors with Fewer Observations Than the Dimension. *Methodology and Computing in Applied Probability*, in press.

Yata, K. and Aoshima, M. (2011b). Effective PCA for High-Dimension, Low-Sample-Size Data with Noise Reduction via Geometric Representations, *Journal of Multivariate Analysis*, revised.