Mem. Gra. Sci. Eng. Shimane Univ.
Series B: Mathematics
48 (2015), pp. 15–26

ASYMPTOTIC THEORY FOR DISCRIMINANT ANALYSIS IN HIGH DIMENSION LOW SAMPLE SIZE

MITSURU TAMATANI

Communicated by Kanta Naito

(Received: March 2, 2015)

Abstract. This paper is based on the author's thesis, "Pattern recognition based on naive canonical correlations in high dimension low sample size". This paper is concerned with discriminant analysis for multi-class problems in a High Dimension Low Sample Size (HDLSS) context. The proposed discrimination method is based on canonical correlations between the predictors and response vector of class label. We investigate the asymptotic behavior of the discrimination method, and evaluate bounds for its misclassification rate.

1. Introduction

Discriminant analysis is a statistical method by which a new data with unknown class label is classified into one of the previously known classes. Fisher's well-known linear discriminant function for K-class problems is the solution to his paradigm: maximize the between-class variance while minimizing the within-class variances. See Fisher [6], and Rao [10] for the general multi-class setting. Especially, for the special case of two-class problems, we assume that the two populations, $\mathcal{C}_1$ and $\mathcal{C}_2$, have multivariate normal distributions which differ in their means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, but share a common covariance matrix Σ. For $\boldsymbol{X}$ from one of the two classes, Fisher's linear discriminant function

$$g(\boldsymbol{X}) = \left(\boldsymbol{X} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right)^T \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{1}$$

assigns $\boldsymbol{X}$ to $\mathcal{C}_1$ if $g(\boldsymbol{X}) > 0$, and to $\mathcal{C}_2$ otherwise. If the random vectors have equal probability of belonging to either of the two classes, then the misclassification rate of (1) is shown to be $\left[\Phi\left(-\Delta^2/2\right) + \left\{1 - \Phi\left(\Delta^2/2\right)\right\}\right]/2$, where Φ is the standard normal distribution function and Δ is the Mahalanobis distance between the two

2000 *Mathematics Subject Classification.* 62H30, 62H20, 62E20,

Key words and phrases. High Dimension Low Sample Size, Canonical Correlations, Consistency, Misclassification, Multi-class Linear Discriminant Analysis.

populations: $\Delta = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}$. A sample version of (1) is

$$\widehat{g}(\boldsymbol{X}) = \left(\boldsymbol{X} - \frac{1}{2}(\widehat{\boldsymbol{\mu}}_1 + \widehat{\boldsymbol{\mu}}_2) \right)^T \widehat{\Sigma}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2), \tag{2}$$

where $\widehat{\boldsymbol{\mu}}_1$, $\widehat{\boldsymbol{\mu}}_2$ and $\widehat{\Sigma}$ are appropriate estimates for the corresponding population parameters of the two classes. (2) can be applied to data sets in the general setting where the sample size n is much larger than the dimension d. However, for problems with a large dimension, it has been shown that Fisher's linear discriminant function performs poorly due to diverging spectra. Further for *High Dimension Low Sample Size* (HDLSS) problems, namely $n \ll d$, it is not possible to directly define Fisher's linear discriminant function since the sample covariance matrix $\widehat{\Sigma}$ in (2) becomes singular when the dimension d exceeds the sample size n.

Several discriminant rules have been proposed for the HDLSS context which overcome the problem of singularity of $\widehat{\Sigma}$ in different ways: Dudoit et al. [4] proposed *diagonal linear discriminant analysis* which only uses the diagonal elements of the sample covariance matrix, Srivastava and Kubokawa [12] proposed a discriminant function based on the Moore-Penrose inverse, Ahn and Marron [1] constructed discriminant function based on the Maximal Data Piling (MDP) direction vectors, and Aoshima and Yata [2] considered a discriminant rule based on second moments in conjunction with geometric representations of high-dimensional data. In this paper, we focus on the 'diagonal' approach of Dudoit et al. [4] which has special appeal since it is conceptually simpler than the competitors.

For two-class problems in a HDLSS setting, Bickel and Levina [3] and Fan and Fan [5] investigated the asymptotic behavior of the naive Bayes rule, and calculated bounds for its misclassification rate. Tamatani, Koch and Naito [13] defined a modification of the canonical correlation matrix that is suitable for classification problems, and studied the asymptotic behavior of the eigenvector and the discriminant direction of the modified canonical correlation matrix in the context of two-class problems. Also, for multi-class problems in a HDLSS context, Tamatani, Naito and Koch [14] proposed a naive Bayes rule in a general multi-class setting and investigated its asymptotic properties for high-dimensional data when both the dimension d and the sample size n grow.

Throughout this paper we focus on HDLSS data from $K(\geq 2)$ classes, that is, we assume that the dimension d of the data is much bigger than the sample size n. Our discriminant approach is based on canonical correlations, and in particular on a modification of the canonical correlation matrix suitable for vector-valued class labels from K classes. In this framework, we replace the covariance matrix by its diagonal counterpart as discussed in Dudoit et al. [4], Bickel and Levina [3] and Fan and Fan [5]. We call such a matrix the *naive canonical correlation matrix*, and observe that this matrix plays an important role in the present theory. The $K-1$ eigenvectors belonging to $K-1$ non-zero eigenvalues of the estimated naive canonical correlation matrix yield discriminant directions which inform our choice of a discriminant function for K classes. For this setting we study the asymptotic behavior of the eigenvectors and associate discriminant directions.

For HDLSS data from K multivariate normal classes, we derive an upper bound for the misclassification rate of the proposed multi-class discriminant function. Our asymptotic results for the misclassification rate are divided into two disjoint types depending on the precise growth rates of d and n. Depending on the two distinct growth rates, we also develop HDLSS asymptotic results for estimators of the eigenvectors and discriminant directions.

The paper is organized as follows. In Section 2, we review a relationship between multi-class discriminant function and canonical correlations for K classes. Further, we focus on the gene expression microarray data, which is a typical example of HDLSS data. For HDLSS problems, we replace Σ by $\mathrm{diag}\Sigma$, and obtain a corresponding criterion, which leads to a natural derivation of a multi-class version of the naive Bayes rule. Section 3 details the asymptotic behavior of the eigenvectors of the estimated naive canonical correlation matrix and the associated discriminant directions in a HDLSS setting under general distributional assumptions. We derive an upper bound for the asymptotic misclassification rate of the proposed multi-class discriminant function under assumptions of normal distribution. In Section 4, we summarize our results. Proofs of all mathematical results are omitted.

2. Discriminant function based on naive canonical correlation

We consider the K-class discrimination problem. Let $\mathcal{C}_\ell$ $(\ell = 1, \ldots, K)$ be d-dimensional populations with different means $\boldsymbol{\mu}_\ell$ and common covariance matrix Σ. For a random vector $\boldsymbol{X}$ from one of the K classes, let π_ℓ be the probability that $\boldsymbol{X}$ belongs to $\mathcal{C}_\ell$. Let $\boldsymbol{Y}$ be the K-dimensional vector-valued class label with ℓth entry 1 if $\boldsymbol{X}$ belongs to $\mathcal{C}_\ell$ and 0 otherwise, so $P(\boldsymbol{Y} = \boldsymbol{e}_\ell) = \pi_\ell$, and $\sum_{\ell=1}^K \pi_\ell = 1$.

In canonical correlation analysis of two vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, with $\boldsymbol{Y}$ the vector-valued labels and $\boldsymbol{\mu} = \sum_{\ell=1}^K \pi_\ell \boldsymbol{\mu}_\ell$, the matrix

$$\widetilde{C} = \Sigma^{-1/2} E\left[(\boldsymbol{X} - \boldsymbol{\mu})\, \boldsymbol{Y}^T\right] \left\{E\left[\boldsymbol{Y}\boldsymbol{Y}^T\right]\right\}^{-1/2} \tag{3}$$

plays an important role. From the definition of $(\boldsymbol{X}, \boldsymbol{Y})$ it follows that $E[\boldsymbol{Y}\boldsymbol{Y}^T] = \Pi$ and $E\left[(\boldsymbol{X} - \boldsymbol{\mu})\boldsymbol{Y}^T\right] = M_0\Pi$ where $\Pi = \mathrm{diag}(\pi_1, \ldots, \pi_K)$ and $M_0 = [\ \boldsymbol{\mu}_1 - \boldsymbol{\mu}\ \cdots\ \boldsymbol{\mu}_K - \boldsymbol{\mu}\ ]$. Using $M = \sum_{\ell=1}^K \pi_\ell (\boldsymbol{\mu}_\ell - \boldsymbol{\mu})(\boldsymbol{\mu}_\ell - \boldsymbol{\mu})^T = M_0 \Pi M_0^T$, we have $\widetilde{C}\widetilde{C}^T = \Sigma^{-1/2} M \Sigma^{-1/2}$. If we put $\widetilde{\boldsymbol{b}} = \Sigma^{-1/2}\widetilde{\boldsymbol{p}}$, where $\widetilde{\boldsymbol{p}}$ is a solution to the eigenvalue problem $\widetilde{C}\widetilde{C}^T\widetilde{\boldsymbol{p}} = \tilde{\lambda}\widetilde{\boldsymbol{p}}$ with $\tilde{\lambda} > 0$, then the vector $\widetilde{\boldsymbol{b}}$ is the maximizer of the criterion

$$\widetilde{J}(\boldsymbol{b}) = \frac{\boldsymbol{b}^T M \boldsymbol{b}}{\boldsymbol{b}^T \Sigma \boldsymbol{b}} \tag{4}$$

over vectors $\boldsymbol{b}$. Note that (4) is nothing other than the criterion which yields Fisher's linear discriminant rule for the multi-class setting. In particular, the rank of $\widetilde{C}\widetilde{C}^T$ is $K-1$, so the $K-1$ eigenvectors $[\ \widetilde{\boldsymbol{p}}_1\ \cdots\ \widetilde{\boldsymbol{p}}_{K-1}\ ]$ belonging to $K-1$ non-zero eigenvalues should be used for constructing the discriminant directions

$$\widetilde{B} \equiv [\ \widetilde{\boldsymbol{b}}_1\ \cdots\ \widetilde{\boldsymbol{b}}_{K-1}\ ] = \Sigma^{-1/2}[\ \widetilde{\boldsymbol{p}}_1\ \cdots\ \widetilde{\boldsymbol{p}}_{K-1}\ ].$$

Using the discriminant directions $\widetilde{B}$, we want to define a discriminant function $\widetilde{g}$ for classifying new observations $\boldsymbol{X}$ whose class is unknown. For $\alpha \in \{1, \ldots, K-1\}$,

put $Z_\alpha(\boldsymbol{X}) = \widetilde{\boldsymbol{b}}_\alpha^T \boldsymbol{X}$, and define the vector $Z(\boldsymbol{X}) = [Z_1(\boldsymbol{X}), \ldots, Z_{K-1}(\boldsymbol{X})]^T = \widetilde{B}^T \boldsymbol{X}$. For $Z(\boldsymbol{X})$ and $\mathcal{C}_\ell$, define the Mahalanobis distance between $Z(\boldsymbol{X})$ and $\mathcal{C}_\ell$ by $\Delta_\ell(Z(\boldsymbol{X})) = \sqrt{(Z(\boldsymbol{X}) - \boldsymbol{\nu}_\ell)^T \Sigma_\ell^{-1} (Z(\boldsymbol{X}) - \boldsymbol{\nu}_\ell)}$, where $\boldsymbol{\nu}_\ell = \widetilde{B}^T \boldsymbol{\mu}_\ell$ and $\Sigma_\ell = \widetilde{B}^T \Sigma \widetilde{B}$. Hence $\Delta_\ell(Z(\boldsymbol{X}))^2$ can be rewritten as

$$\Delta_\ell(Z(\boldsymbol{X}))^2 \;=\; (\boldsymbol{X} - \boldsymbol{\mu}_\ell)^T \, \widetilde{B} (\widetilde{B}^T \Sigma \widetilde{B})^{-1} \widetilde{B}^T \, (\boldsymbol{X} - \boldsymbol{\mu}_\ell) \,. \tag{5}$$

Using (5), we now derive the multi-class linear discriminant function $\widetilde{g}$ as the minimizer of the Mahalanobis distance, and let

$$\widetilde{g}(\boldsymbol{X}) = \underset{\ell \in \{1, \ldots, K\}}{\arg\min} \, \Delta_\ell(Z(\boldsymbol{X}))^2. \tag{6}$$

Note that $\widetilde{B}$ depends on Σ and that (6) reduces to the Fisher's linear discriminant function in the case of 2-class by considering the sign of $\Delta_1(Z(\boldsymbol{X}))^2 - \Delta_2(Z(\boldsymbol{X}))^2$.

2.1. **HDLSS setting.** Gene expression microarray data is a form of high-throughput biological data providing relative measurements of mRNA levels for genes in a biological sample. One important application of gene expression microarray data is the classification of new samples into known classes. In the recent years, many laboratories have collected and analyzed microarray data, and such data have been appeared in public databases or on web sites. Indeed, the data collected from gene expression microarrays consist of thousands or tens of thousands of genes that constitute features: leukemia data (Golub et al. [7], $K = 2,\ n = 73,\ d = 7129$), lung cancer data (Gordon et al. [8], $K = 2,\ n = 184,\ d = 12433$), prostate cancer data (Singh et al. [11], $K = 2,\ n = 134,\ d = 12600$), the small round blue cell tumors (SRBCT) data (Khan et al. [9], $K = 4,\ n = 83,\ d = 2308$), and so on. These data, often called the High Dimension Low Sample Size (HDLSS) data, are characterized with large number of dimensions d and a relatively small number of sample size n, that is, we can write $n \ll d$.

To establish asymptotic theory in this paper, we consider an asymptotic situation in which the dimension d and the sample size n both approach infinity in such a way that d is much faster than n:

$$n = o(d) \quad \text{as} \quad n, d \longrightarrow \infty.$$

2.2. **Difficulties in HDLSS.** The HDLSS data involves serious problems. Let $\boldsymbol{X}_i$ $(i = 1, \ldots, n)$ be *independently and identically distributed* (i.i.d.) d-dimensional random vectors with the covariance matrix Σ. The usual estimates of Σ is defined as

$$\widehat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}})^T, \tag{7}$$

where $\widehat{\boldsymbol{\mu}} = (1/n) \sum_{i=1}^n \boldsymbol{X}_i$ is the mean vector of the data $\{\boldsymbol{X}_i\}_{1 \le i \le n}$. When we assume that the data satisfies $n \ll d$, (7) is a singular matrix because the rank of (7) is at most $n - 1$. Hence, it means that the sample version of (3) is not directly applicable in the case of $n \ll d$.

2.3. **Naive canonical correlations.** In the HDLSS two-class discrimination settings, the sample covariance matrix $\widehat{\Sigma}$ is singular as was discussed in the previous subsection. For a population framework, it therefore does not make sense to define a discriminant function based on the within-class covariance matrix Σ. To overcome such a difficulty, we first require a suitable framework for the population. We define the *naive canonical correlation matrix* C and vectors $\boldsymbol{b}_\alpha$ and $\boldsymbol{p}_\alpha$ by

$$C = D^{-1/2} E\left[(\boldsymbol{X} - \boldsymbol{\mu})\boldsymbol{Y}^T\right] \left\{E[\boldsymbol{Y}\boldsymbol{Y}^T]\right\}^{-1/2},$$

where $D = \mathrm{diag}\Sigma$, $\boldsymbol{b}_\alpha = D^{-1/2}\boldsymbol{p}_\alpha$ and $\boldsymbol{p}_\alpha$ is eigenvector of the matrix CC^T corresponding to the αth largest eigenvalue λ^*_α. Put $P = [\ \boldsymbol{p}_1\ \cdots\ \boldsymbol{p}_{K-1}\ ]$. The discriminant directions

$$B \equiv [\ \boldsymbol{b}_1\ \cdots\ \boldsymbol{b}_{K-1}\ ] = [\ D^{-1/2}\boldsymbol{p}_1\ \cdots\ D^{-1/2}\boldsymbol{p}_{K-1}\ ] = D^{-1/2}P$$

can now be seen to maximise the analogous *naive* criterion

$$J(\boldsymbol{b}) = \frac{\boldsymbol{b}^T M \boldsymbol{b}}{\boldsymbol{b}^T D \boldsymbol{b}}. \tag{8}$$

Note that Σ in (4) has been replaced by the diagonal matrix D in (8).

2.4. **Discriminant function in HDLSS.** The corresponding discriminant function g is therefore

$$g(\boldsymbol{X}) = \underset{\ell\in\{1,\ldots,K\}}{\arg\min}\ (\boldsymbol{X} - \boldsymbol{\mu}_\ell)^T B(B^T D B)^{-1} B^T (\boldsymbol{X} - \boldsymbol{\mu}_\ell)\,. \tag{9}$$

Note that Σ has been replaced by D both in (4) and (5) to yield (8) and (9) respectively.

It is worth noting that for $K = 2$, the discriminant function g reduces to the naive Bayes discriminant function discussed in Bickel and Levina [3].

3. Asymptotic Theory

In this section, we evaluate the asymptotic behavior of the estimators of $\boldsymbol{p}_\alpha$ and $\boldsymbol{b}_\alpha$ in a HDLSS setting under general assumptions about the underlying distributions. In addition, we derive upper bounds for the misclassification rate of multi-class discriminant function under d-dimensional normal populations.

3.1. **Preliminaries.** Consider data $(\boldsymbol{X}_{\ell i}, \boldsymbol{Y}_{\ell i})$ $(\ell = 1, \ldots, K;\ i = 1, \ldots, n_\ell)$, where the independently distributed $\boldsymbol{X}_{\ell i}$ are from K disjoint classes, and the $\boldsymbol{Y}_{\ell i}$ are independent realizations of vector labels

$$\boldsymbol{Y}_{\ell i} = (Y_{\ell i 1}, \ldots, Y_{\ell i K})^T \quad \text{with} \quad Y_{\ell i j} = \begin{cases} 1 & \ell\text{-th component,} \\ 0 & \text{otherwise.} \end{cases}$$

Let X and Y be matrices defined by $X = [\ \boldsymbol{X}_{11}\ \cdots\ \boldsymbol{X}_{Kn_K}\ ]$ and $Y = [\ \boldsymbol{Y}_{11}\ \cdots\ \boldsymbol{Y}_{Kn_K}\ ]$. Then X is of size $d \times n$, and Y is of size $K \times n$, where $n = \sum_{\ell=1}^K n_\ell$.

Next we derive an empirical version of C and its left eigenvectors $\boldsymbol{p}_\alpha$. Define estimators $\widehat{\boldsymbol{\mu}}_\ell$ and $\widehat{\Sigma}$ of $\boldsymbol{\mu}_\ell$ and Σ by $\widehat{\boldsymbol{\mu}}_\ell = (1/n_\ell)\sum_{i=1}^{n_\ell} \boldsymbol{X}_{\ell i}$ and $\widehat{\Sigma} = (1/K)\sum_{\ell=1}^K \widehat{S}_\ell$,

where $\widehat{S}_\ell = \{1/(n_\ell - 1)\} \sum_{i=1}^{n_\ell} (\boldsymbol{X}_{\ell i} - \widehat{\boldsymbol{\mu}}_\ell)(\boldsymbol{X}_{\ell i} - \widehat{\boldsymbol{\mu}}_\ell)^T$. Using the centering matrix, a natural estimator for C is

$$\widehat{C} \;=\; \widehat{D}^{-1/2} \widehat{M}_0 N^{1/2}, \tag{10}$$

where $\widehat{D} = \mathrm{diag}\widehat{\Sigma}$, I_n is the $n \times n$ identity matrix, $\mathbf{1}_n$ is the n-dimensional vector of ones, $\widehat{M}_0 = [\, \widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}} \;\cdots\; \widehat{\boldsymbol{\mu}}_K - \widehat{\boldsymbol{\mu}} \,]$ and $N = \mathrm{diag}(n_1/n, \ldots, n_K/n)$. Hence we obtain the expression $\widehat{C}\widehat{C}^T = \widehat{D}^{-1/2} \widehat{M} \widehat{D}^{-1/2}$, where $\widehat{M} = \widehat{M}_0 N \widehat{M}_0^T$.

Since the rank of $\widehat{C}\widehat{C}^T$ is $K-1$, we use the $K-1$ eigenvectors $\widehat{\boldsymbol{p}}_\alpha$ of $\widehat{C}\widehat{C}^T$ corresponding to the $K-1$ non-zero eigenvalues in the definition of the discrimination directions $\widehat{\boldsymbol{b}}_\alpha$, and put $\widehat{\boldsymbol{b}}_\alpha = \widehat{D}^{-1/2} \widehat{\boldsymbol{p}}_\alpha$ $(\alpha = 1, \ldots, K-1)$. We note that these $\widehat{\boldsymbol{b}}_\alpha$ can be obtained as the maximizers of the function

$$\widehat{J}(\boldsymbol{b}) = \frac{\boldsymbol{b}^T \widehat{M} \boldsymbol{b}}{\boldsymbol{b}^T \widehat{D} \boldsymbol{b}}.$$

Put $\widehat{B} = [\, \widehat{\boldsymbol{b}}_1 \;\cdots\; \widehat{\boldsymbol{b}}_{K-1} \,] = [\, \widehat{D}^{-1/2} \widehat{\boldsymbol{p}}_1 \;\cdots\; \widehat{D}^{-1/2} \widehat{\boldsymbol{p}}_{K-1} \,] \equiv \widehat{D}^{-1/2} \widehat{P}$, then $\widehat{g}$, defined by

$$\widehat{g}(\boldsymbol{X}) = \underset{\ell \in \{1, \ldots, K\}}{\arg\min} \, (\boldsymbol{X} - \widehat{\boldsymbol{\mu}}_\ell)^T \, \widehat{B} (\widehat{B}^T \widehat{D} \widehat{B})^{-1} \widehat{B}^T \, (\boldsymbol{X} - \widehat{\boldsymbol{\mu}}_\ell) \,,$$

is a natural estimator of g in (9), and this $\widehat{g}$ is our proposed discriminant function in the HDLSS multi-class setting. To elucidate the asymptotic behavior of $\widehat{g}$, it is necessary to develop first asymptotics for $\widehat{B}$ as well as $\widehat{P}$ in a HDLSS setting.

Throughout this paper, we make the assumption that the sample size of each of the K classes satisfies $c \leq n_\ell / n$ for some positive constant c and $\ell = 1, \ldots, K$.

In what follows we use the asymptotic notation:

(i) $a_{n,d} = O(b_{n,d})$ to mean that $a_{n,d}/b_{n,d} \to M \in (0, \infty)$ as $n, d \to \infty$.
(ii) $a_{n,d} = o(b_{n,d})$ to mean that $a_{n,d}/b_{n,d} \to 0$ as $n, d \to \infty$.

The definition of o is usually included in that of big O, however we distinguish these cases in this paper.

In order to evaluate the asymptotic behavior of $\widehat{\boldsymbol{p}}_\alpha$, we need the following definition and conditions:

Definition 3.1. Let $\boldsymbol{x} \in \mathbb{R}^d$ be a non-stochastic unit vector, and let $\widehat{\boldsymbol{x}}$, a vector of length one, denote an estimate of $\boldsymbol{x}$ based on the sample of size n . If

$$\widehat{\boldsymbol{x}}^T \boldsymbol{x} \;\xrightarrow{P}\; 1 \quad \text{as} \quad n, d \to \infty,$$

where $\xrightarrow{P}$ refers to convergence in probability, then $\widehat{\boldsymbol{x}}$ is HDLSS consistent with $\boldsymbol{x}$.

CONDITION A. Let $X = [\, \boldsymbol{X}_{11} \;\cdots\; \boldsymbol{X}_{Kn_K} \,]$ be a data matrix from K classes. Each column of X can be written as $\boldsymbol{X}_{\ell i} = \boldsymbol{\mu}_\ell + \boldsymbol{\varepsilon}_{\ell i}$, where $\boldsymbol{\varepsilon}_{\ell i}$ $(\ell = 1, \ldots, K;\ i = 1, \ldots, n_\ell)$ are i.i.d. copies of an underlying random vector $\boldsymbol{\varepsilon}$ with mean 0 and covariance matrix Σ.

CONDITION B. (CRAMÉR'S CONDITION) There exist constants ν_1, ν_2, M_1 and M_2 such that each component of $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_d)^T$ satisfies

$$E[|\varepsilon_j|^m] \leq m! M_1^{m-2} \nu_1 / 2 \quad \text{and} \quad E[|\varepsilon_j^2 - \sigma_{jj}|^m] \leq m! M_2^{m-2} \nu_2 / 2$$

for all $m \in \mathbb{N}$.

CONDITION C. Let $n_\ell \to \infty$ (for $\ell = 1, \ldots, K$), $d \to \infty$, $\log d = o(n)$, $n = o(d)$. There exists a positive sequence C_d depending only on the dimension d such that $d/(nC_d) \to \xi$, where $\xi \geq 0$.

CONDITION D. All eigenvalues λ_α^* of $C^T C$ are simple (so $\lambda_1^* > \cdots > \lambda_K^*$) and satisfy

$$\lambda_\alpha^* = O(C_d) \quad \text{and} \quad \frac{\lambda_\alpha^* - \lambda_{\alpha+1}^*}{C_d} > \xi \quad \text{for} \quad \alpha = 1, \ldots, K-1,$$

and C_d as in CONDITION C.

CONDITION E. As $d \to \infty$, $\boldsymbol{\mu}_\ell^T D^{-1} \boldsymbol{\mu}_\ell = O(C_d)$, and there exists $\delta \in (0,1)$ such that $\boldsymbol{\mu}_\ell^T D^{-1} \boldsymbol{\mu}_k = O(C_d^\delta)$ for all $k, \ell \in \{1, \ldots, K\}$, and C_d as in CONDITION C.

CONDITION F. For $k, \ell \in \{1, \ldots, K\}$, and $k \neq \ell$,

$$\lim_{d\to\infty} \sqrt{\pi_k} \frac{\boldsymbol{\mu}_k^T D^{-1} \boldsymbol{\mu}_k}{C_d} \neq \lim_{d\to\infty} \sqrt{\pi_\ell} \frac{\boldsymbol{\mu}_\ell^T D^{-1} \boldsymbol{\mu}_\ell}{C_d},$$

and C_d as in CONDITION C.

We note that the positive sequence C_d will play an important role in the subsequent discussions since it controls the gap between d and n.

3.2. **Asymptotic results for eigenvectors.** To establish the asymptotic behaviour of the estimators $\widehat{\boldsymbol{p}}_\alpha$, we start with an asymptotic expansion of the matrix $\widehat{C}^T \widehat{C}/C_d$. In what follows, Θ denotes the parameter space for our multi-class setting:

$$\Theta = \left\{ (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \Sigma) \;\middle|\; \begin{array}{l} \min_{k\neq\ell} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)^T D^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell) \geq C_d, \\ \lambda_{\max}(R) \leq b_0, \quad \min_{1\leq j\leq d} \sigma_{jj} > 0 \end{array} \right\},$$

where R is the correlation matrix $R = D^{-1/2} \Sigma D^{-1/2}$, $\lambda_{\max}(R)$ is the largest eigenvalue of R and σ_{jj} is jth diagonal entry of Σ.

Lemma 3.2. *Suppose that* CONDITION A – CONDITION E *hold. Then for all parameters* $\theta \in \Theta$, $\widehat{C}^T \widehat{C}/C_d$ *can be expanded as*

$$\frac{\widehat{C}^T \widehat{C}}{C_d} = \frac{C^T C}{C_d} + \xi (I_K - \Pi^{1/2} \mathbf{1}_K \mathbf{1}_K^T \Pi^{1/2}) + \mathbf{1}_K \mathbf{1}_K^T o_P(1).$$

If $\xi = 0$, which means $d = o(nC_d)$ by CONDITION C, then the above expansion becomes very simple.

To proceed with the theoretical considerations, we need to show that the $K-1$ eigenvalues of $C^T C/C_d + \xi(I_K - \Pi^{1/2} \mathbf{1}_K \mathbf{1}_K^T \Pi^{1/2})$ are simple in the case of $d = O(nC_d)$, that is, $\xi > 0$.

Lemma 3.3. *Let λ_α/C_d be αth largest eigenvalue of $C^TC/C_d + \xi(I_K - \Pi^{1/2}\mathbf{1}_K\mathbf{1}_K^T\Pi^{1/2})$. Suppose that the λ_α/C_d satisfy* Condition D*, and that $d = O(nC_d)$. Then for all parameters $\theta \in \Theta$,*

$$\lambda_\alpha = O(C_d) \quad \textit{and} \quad \frac{\lambda_\alpha}{C_d} > \frac{\lambda_{\alpha+1}}{C_d} \quad \textit{for} \quad \alpha = 1, \dots, K-1.$$

In this paper, eigenvectors have unit length. In addition, we assume that the first entry of each eigenvector is positive. This assumption avoids any ambiguity about the direction of the eigenvector. Using Lemmas 3.2 and 3.3, we can now describe the asymptotic behavior of $\widehat{\boldsymbol{p}}_\alpha$ as follows.

Theorem 3.4. *Suppose that* Condition A – Condition D *hold. Then, for all parameters $\theta \in \Theta$,*

$$\widehat{\boldsymbol{p}}_\alpha^T \frac{\widehat{C}\boldsymbol{\gamma}_\alpha}{||\widehat{C}\boldsymbol{\gamma}_\alpha||} = 1 + o_P(1) \quad \textit{for} \quad \alpha = 1, \dots, K-1,$$

where $\boldsymbol{\gamma}_\alpha$ is eigenvector of $C^TC/C_d + \xi(I_K - \Pi^{1/2}\mathbf{1}_K\mathbf{1}_K^T\Pi^{1/2})$ belonging to the non-zero eigenvalue λ_α.

To gain further insight in the behavior of the $\widehat{\boldsymbol{p}}_\alpha$, we define the vectors

$$\boldsymbol{p}_\alpha = \frac{C\boldsymbol{\gamma}_\alpha}{||C\boldsymbol{\gamma}_\alpha||} \quad \text{for} \quad \alpha = 1, \dots, K-1$$

by referring to Theorem 3.4. Under the assumption that $d = O(nC_d)$, one can show that the vectors $\widehat{\boldsymbol{p}}_\alpha$ are consistent estimators for the $\boldsymbol{p}_\alpha$. We have the following theorem and corollary:

Theorem 3.5. *Suppose that* Condition A – Condition E *hold, and that $d = O(nC_d)$. Moreover, assume that $\lambda_\alpha/C_d \to \kappa_\alpha$ and $\boldsymbol{\gamma}_\alpha^T\Pi^{1/2}\mathbf{1}_K \to \eta_\alpha$. Then for all parameters $\theta \in \Theta$,*

$$\widehat{P}^TP \xrightarrow{P} \left(\frac{\kappa_\beta\delta_{\alpha\beta} - \xi(\delta_{\alpha\beta} - \eta_\alpha\eta_\beta)}{\sqrt{\kappa_\alpha}\sqrt{\kappa_\beta - \xi(1-\eta_\beta^2)}} \right)_{1 \le \alpha,\beta \le K-1},$$

where $\delta_{\alpha\beta}$ is the Kronecker delta-function.

Corollary 3.6. *Suppose that* Condition A – Condition E *hold, and that $d = o(nC_d)$. Then for all parameters $\theta \in \Theta$,*

$$\widehat{P}^TP \xrightarrow{P} I_{K-1}.$$

Corollary 3.6 states that if $d = o(nC_d)$ is satisfied, then $\widehat{\boldsymbol{p}}_\alpha$ is HDLSS consistent with $\boldsymbol{p}_\alpha$. Furthermore, $\widehat{\boldsymbol{p}}_\alpha$ is asymptotically orthogonal to $\boldsymbol{p}_\beta$ for $\alpha \neq \beta$. On the other hand, if $d = O(nC_d)$ is satisfied, then the angle between $\widehat{\boldsymbol{p}}_\alpha$ and $\boldsymbol{p}_\beta$ converges to a particular non-zero angle for all α and β.

3.3. **Asymptotic results for discriminant directions.** Next we turn to the asymptotic behavior of the vectors $\widehat{\boldsymbol{b}}_\alpha$. We define normalized versions of direction vectors for discrimination by

$$\widehat{\boldsymbol{b}}^*_\alpha = \frac{\widehat{D}^{-1/2}\widehat{\boldsymbol{p}}_\alpha}{\sqrt{\widehat{\boldsymbol{p}}_\alpha^T \widehat{D}^{-1}\widehat{\boldsymbol{p}}_\alpha}}, \quad \boldsymbol{b}^*_\alpha = \frac{D^{-1/2}\boldsymbol{p}_\alpha}{\sqrt{\boldsymbol{p}_\alpha^T D^{-1}\boldsymbol{p}_\alpha}} \tag{11}$$

for $\alpha = 1, \ldots, K-1$. Then we have the following theorem and corollary:

Theorem 3.7. *Suppose that* Condition A – Condition E *hold, and that* $d = O(nC_d)$. *Put* $\sigma_{\max} = \max_{1\le j\le d}\sigma_{jj}$ *and* $\sigma_{\min} = \min_{1\le j\le d}\sigma_{jj}$. *Then for all parameters* $\theta \in \Theta$,

$$\begin{aligned}
\widehat{\boldsymbol{b}}^{*T}_\alpha \boldsymbol{b}^*_\beta &\le \frac{\boldsymbol{b}^{*T}_\alpha \boldsymbol{b}^*_\beta \sqrt{\kappa_\alpha - \xi(1-\eta_\alpha^2)}}{\sqrt{\kappa_\alpha - \xi(1-\eta_\alpha^2)\,(1-\sigma_{\min}/\sigma_{\max})}}(1+o_P(1)),\\
\widehat{\boldsymbol{b}}^{*T}_\alpha \boldsymbol{b}^*_\beta &\ge \frac{\boldsymbol{b}^{*T}_\alpha \boldsymbol{b}^*_\beta \sqrt{\kappa_\alpha - \xi(1-\eta_\alpha^2)}}{\sqrt{\kappa_\alpha - \xi(1-\eta_\alpha^2)\,(1-\sigma_{\max}/\sigma_{\min})}}(1+o_P(1)).
\end{aligned}$$

Corollary 3.8. *Let* $\widehat{B}^* = [\ \widehat{\boldsymbol{b}}^*_1\ \cdots\ \widehat{\boldsymbol{b}}^*_{K-1}\ ]$ *and* $B^* = [\ \boldsymbol{b}^*_1\ \cdots\ \boldsymbol{b}^*_{K-1}\ ]$. *Suppose that* Condition A – Condition E *hold, and that* $d = o(nC_d)$. *Then for all parameters* $\theta \in \Theta$,

$$\widehat{B}^{*T}B^* - B^{*T}B^* \xrightarrow{P} O.$$

Theorem 3.7 states that the upper and lower bounds of $\widehat{\boldsymbol{b}}^{*T}_\alpha \boldsymbol{b}^*_\beta$ are determined by the ratio of $\sigma_{\max}$ and $\sigma_{\min}$. For example, if all diagonal elements of Σ are equal, then

$$\widehat{\boldsymbol{b}}^{*T}_\alpha \boldsymbol{b}^*_\alpha - \boldsymbol{b}^{*T}_\alpha \boldsymbol{b}^*_\alpha \sqrt{1 - \frac{\xi}{\kappa_\alpha}(1-\eta_\alpha^2)} \xrightarrow{P} 0.$$

If $d = o(nC_d)$ is satisfied, then the angle between $\widehat{\boldsymbol{b}}^*_\alpha$ and $\boldsymbol{b}^*_\alpha$ converges to 0 in probability, which shows that the $\widehat{\boldsymbol{b}}^*_\alpha$ are HDLSS consistent with the corresponding $\boldsymbol{b}^*_\alpha$. However, $\widehat{\boldsymbol{b}}^*_\alpha$ and $\boldsymbol{b}^*_\beta$ may not necessarily be orthogonal for $\alpha \ne \beta$, since

$$\widehat{\boldsymbol{b}}^{*T}_\alpha \boldsymbol{b}^*_\beta - \frac{\boldsymbol{p}_\alpha^T D^{-1}\boldsymbol{p}_\beta}{\sqrt{\boldsymbol{p}_\alpha^T D^{-1}\boldsymbol{p}_\alpha}\sqrt{\boldsymbol{p}_\beta^T D^{-1}\boldsymbol{p}_\beta}} = o_P(1).$$

3.4. **Upper bound for misclassification rate.** In this section, we study the misclassification rate of our method in a multi-class setting. The misclassification rate for two classes has been investigated in Fan and Fan [5] who derived an upper bound for the misclassification rate in a HDLSS setting.

Specifically, for the results in this section we will assume that Condition A' holds.

Condition A'. Let $X = [\ \boldsymbol{X}_{11}\ \cdots\ \boldsymbol{X}_{Kn_K}\ ]$ be a data matrix from K classes. Each column of X can be written as $\boldsymbol{X}_{\ell i} = \boldsymbol{\mu}_\ell + \boldsymbol{\varepsilon}_{\ell i}$, where $\boldsymbol{\varepsilon}_{\ell i}$ are i.i.d. random vectors

distributed as $N_d(\mathbf{0}, \Sigma)$, d-dimensional normal distribution with mean vector $\mathbf{0}$ and covariance matrix Σ.

In addition to Condition A, Condition A' makes statements about the distribution of X.

Suppose that $\boldsymbol{X}$ belongs to class $\mathcal{C}_k$. The misclassification rate of $\widehat{g}$, an estimate of the discriminant function g in (9), for class $\mathcal{C}_k$ is defined as

$$\begin{aligned} W_k(\widehat{g}, \theta) &= P\left(\widehat{g}(\boldsymbol{X}) \neq k \,|\boldsymbol{X}_{\ell i},\ \ell = 1, \ldots, K;\ i = 1, \ldots, n_\ell\right) \\ &= 1 - \int_{\widehat{\mathscr{D}}_k} \frac{1}{\sqrt{|2\pi\widehat{\Sigma}_k|}} \exp\left(-\frac{1}{2}\boldsymbol{z}^T\widehat{\Sigma}_k^{-1}\boldsymbol{z}\right) d\boldsymbol{z} \\ &\equiv 1 - \Phi_{K-1}\left(\widehat{\mathscr{D}}_k;\ \mathbf{0}, \widehat{\Sigma}_k\right), \end{aligned}$$

where $\widehat{\Sigma}_k$ is the transformed covariance matrix of size $(K-1) \times (K-1)$ which is defined in Tamatani, Naito and Koch [14], and $\widehat{\mathscr{D}}_k$ is the $(K-1)$-dimensional region given by

$$\widehat{\mathscr{D}}_k = \left\{ \boldsymbol{z} \in \mathbb{R}^{K-1} \;\middle|\; z_j < \underline{\hat{d}}_{k\alpha},\ \alpha \in \{1, \ldots, K-1\} \right\},$$

where $\underline{\hat{d}}_{k\alpha} = I(\alpha < k)\hat{d}_{k\alpha} + I(\alpha \geq k)\hat{d}_{k(\alpha+1)}$ and

$$\hat{d}_{k\alpha} = \frac{(\boldsymbol{\mu}_k - (\widehat{\boldsymbol{\mu}}_k + \widehat{\boldsymbol{\mu}}_\alpha)/2)^T \widehat{B}(\widehat{B}^T\widehat{D}\widehat{B})^{-1}\widehat{B}^T(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)}{\sqrt{(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)^T\widehat{B}(\widehat{B}^T\widehat{D}\widehat{B})^{-1}\widehat{B}^T\Sigma\widehat{B}(\widehat{B}^T\widehat{D}\widehat{B})^{-1}\widehat{B}^T(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_\alpha)}}.$$

We note that the region $\widehat{\mathscr{D}}_1$ results in the interval obtained in Theorem 1 in Fan and Fan [5] for their special case of $K = 2$.

Let Θ_k be the parameter space associated with the misclassification rate of $\widehat{g}$ for class $\mathcal{C}_k$:

$$\Theta_k = \left\{ (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \Sigma) \;\middle|\; \begin{array}{l} \min_{\ell \neq k}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_k)^T D^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_k) \geq C_d, \\ \lambda_{\max}(R) \leq b_0, \min_{1 \leq j \leq d} \sigma_{jj} > 0 \end{array} \right\}.$$

In addition to the region $\widehat{\mathscr{D}}_k$ we also require the following region and quantities in Theorem 3.9:

$$\mathscr{D}_{k,O} = \left\{ \boldsymbol{z} \in \mathbb{R}^{K-1} \;\middle|\; z_\alpha < \underline{d}_{k\alpha}(1 + o_P(1)),\ \alpha = 1, \ldots, K-1 \right\},$$

where $\underline{d}_{k\alpha} = I(\alpha < k)d_{k\alpha} + I(\alpha \geq k)d_{k(\alpha+1)}$,

$$d_{k\alpha} = \frac{S_{k\alpha}\Gamma\left[\Gamma^T\left\{C^TC + (d/n)(I_K - \Pi^{1/2}\mathbf{1}_K\mathbf{1}_K^T\Pi^{1/2})\right\}\Gamma\right]^{-1}\Gamma^TQ_{k\alpha}^T}{\sqrt{\lambda_{\max}(R)}\sqrt{Q_{k\alpha}\Gamma\left[\Gamma^T\left\{C^TC + (d/n)(I_K - \Pi^{1/2}\mathbf{1}_K\mathbf{1}_K^T\Pi^{1/2})\right\}\Gamma\right]^{-1}\Gamma^TQ_{k\alpha}^T}},$$

$\Gamma = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{K-1}]$, $S_{k\alpha} = M_{k\alpha}/2 + (d/n)\boldsymbol{s}_{k\alpha}\Pi^{-1/2}$, $Q_{k\alpha} = M_{k\alpha} + (d/n)\boldsymbol{q}_{k\alpha}\Pi^{-1/2}$,

$$\begin{aligned} M_{k\alpha} &= (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\alpha)^T D^{-1/2} C, \\ \boldsymbol{s}_{k\alpha} &= [s_1, \ldots, s_K], \quad s_\ell = \pi_\ell - \frac{1}{2}\{I(\ell = k) + I(\ell = \alpha)\}, \\ \boldsymbol{q}_{k\alpha} &= [q_1, \ldots, q_K], \quad q_\ell = I(\ell = k) - I(\ell = \alpha). \end{aligned}$$

Furthermore, we need the following region in Corollary 3.10:

$$\mathscr{D}_{k,o} = \left\{ \boldsymbol{z} \in \mathbb{R}^{K-1} \;\middle|\; z_\alpha < \underline{d}^*_{k\alpha}(1 + o_P(1)),\ \alpha = 1, \ldots, K-1 \right\},$$

where $\underline{d}^*_{k\alpha} = I(\alpha < k)d^*_{k\alpha} + I(\alpha \geq k)d^*_{k(\alpha+1)}$ and

$$d^*_{k\alpha} = \frac{\sqrt{M_{k\alpha}\Gamma\left(\Gamma^T C^T C \Gamma\right)^{-1}\Gamma^T M^T_{k\alpha}}}{2\sqrt{\lambda_{\max}(R)}}.$$

We have the following theorem and corollary using Theorem 3.4.

Theorem 3.9. *Suppose that* Condition A' *and* Condition B – Condition F *hold, and that* $d = O(nC_d)$. *Then, for all parameters* $\theta \in \Theta_k$,

$$W_k(\widehat{g}, \theta) \leq 1 - \Phi_{K-1}\left(\mathscr{D}_{k,O};\ \mathbf{0}, \widehat{\Sigma}_k\right).$$

Corollary 3.10. *Suppose that* Condition A' *and* Condition B – Condition F *hold, and that* $d = o(nC_d)$. *Then, for all parameters* $\theta \in \Theta_k$,

$$W_k(\widehat{g}, \theta) \leq 1 - \Phi_{K-1}\left(\mathscr{D}_{k,o};\ \mathbf{0}, \widehat{\Sigma}_k\right).$$

Note that Theorem 3.9 and Corollary 3.10 extend Theorem 1 in Fan and Fan [5] to the general multi-class setting considered in this paper.

4. Conclusion

In this paper, we discussed the asymptotic theories of the multi-class linear discriminant function in a HDLSS context. In Section 2, we constructed the linear discriminant function based on naive canonical correlation in the context of multi-class problem. In Section 3, we derived the asymptotic behavior of eigenvectors of the naive canonical correlation matrix corresponding to positive eigenvalues. In the asymptotic theory, both the dimension d and the sample size n grow, and provided d does not grow too fast, we showed that all eigenvectors and discriminant directions are HDLSS consistent.

References

[1] Ahn, J. and Marron, J. S. The maximal data piling direction for discrimination. *Biometrika*, **97** (2010), 254–259.

[2] Aoshima, M. and Yata, K. Two-stage procedures for high-dimensional data. *Sequential analysis*, **30** (2011), 356–399.

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

[4] Dudoit, S., Fridlyand, J. and Speed, T. P. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, **97** (2002), 77–87.
[5] Fan, J. and Fan, Y. High dimensional classification using features annealed independence rules. *Annals of statistics*, **36** (2008), 2605–2637.
[6] Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, **7** (1936), 179–188.
[7] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, **286** (1999), 531–537.
[8] Gordon, G. J., Jensen, R. V., Hsiao, L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J. and Bueno, R. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research*, **62** (2002), 4963–4967.
[9] Khan, J., Wei, J., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C. and Meltzer, P. S. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7** (2001), 673–679.
[10] Rao, C. R. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, **10** (1948), 159–203.
[11] 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. and Sellers, W. R. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, **1** (2002), 203–209.
[12] Srivastava, M. S. and Kubokawa, T. Comparison of discrimination methods for high dimensional data. *J. Japan Statist. Soc*, **37** (2007), 123–134.
[13] Tamatani, M., Koch, I. and Naito, K. Pattern recognition based on canonical correlations in a high dimension low sample size context. *Journal of Multivariate Analysis*, **111** (2012), 350–367.
[14] Tamatani, M., Naito, K. and Koch, I. Multi-class discriminant function based on canonical correlation in high dimension low sample size. *Bulletin of Informatics and Cybernetics*, **45** (2013), 67-101.

Mitsuru Tamatani: Graduate school of science and engineering, Shimane University, Matsue, Shimane, 690-8504, Japan
E-mail address: tamatani@riko.shimane-u.ac.jp