High-dimensional Consistencies of Efficient Screening Methods Based on Information and MER Criteria in Discriminant Analysis

Yasunori Fujikoshi

Emeritus Professor of Hiroshima University

(This is a joint work with T. Sakurai)

Outline

1. Two Groups Discriminant Analysis
2. Model Selection Methods
3. Efficient Screening Methods
4. Consistency of Selection Methods
5. Simulation Results

Two-Groups Discriminant Analysis

p**-variate**: $\boldsymbol{x} = (x_1, \ldots, x_p)'$

$$\boldsymbol{x} \mid \Pi^{(i)} \sim \mathrm{N}(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma}), \ \ i = 1, 2$$

Samples:

$$\boldsymbol{x}_1^{(i)}, \ldots, \boldsymbol{x}_{n_i}^{(i)}, \quad i = 1, \ldots, n_i; \ i = 1, 2$$

Mean Vectors, Sample Covariance Matrix, Matrix due to Within Groups,

$$\bar{\boldsymbol{x}}^{(i)}, \quad \mathsf{S}, \quad \mathsf{W} = (n-2)\mathsf{S}$$

Linear Discriminant Fuction:

$$(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})'\mathsf{S}^{-1}\left\{\boldsymbol{x} - \frac{1}{2}(\bar{\boldsymbol{x}}^{(1)} + \bar{\boldsymbol{x}}^{(2)})\right\}$$

Main Purposes in Discriminant Analysis

Data Analysis:

(1) Classification of new observations x

(2) Interpretation of discriminant function

Theoretical Interests:

(1) How to find a "best" subvector of x

(2) Study for properties of variable selection methods

Notation for subsets of variables:

$$\boldsymbol{x}_j = \{x_{j_1}, \ldots, x_{j_{p_j}}\} \Longleftrightarrow j = \{j_1, \ldots, j_{p_j}\} \subset \omega = \{1, \ldots, p\}$$

Asymptotic Framework

A0: Large-Sample AF

$$p; \text{ fix}, \ n_1 \to \infty, \ n_2 \to \infty$$

$$n_i/n \to k_i, \quad 0 < k_i < 1, \quad i = 1, 2$$

A1: High-Dimensional AF

$$p, n_1, n_2 \to \infty$$

$$p/n \to c \in (0, 1)$$

$$n_i/n \to k_i \quad 0 < k_i < 1, \quad i = 1, 2$$

Motivation for High-Dimensional Asymptotic Framework

(1) Real Data:

$(p, n) = (30, 100), \ (50, 200)$, etc.

(2) Data set after reduced by naieve criteria

Original microarray data; $p = 3571, n_1 = 47, \ n_2 = 25$

By applying the BW (Between-Within) Criterion:

$p = 20, \ n_1 = 47, \ n_2 = 25.$ Dudoit et al. (2002)

Variable Selection Models

Coefficient Vector of Linear Discriminant Function:

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) = (\beta_1, \ldots, \beta_p)'$$

Variable Selection Models:

$$M_j : \beta_i \neq 0 \text{ for } i \in j, \quad \beta_i = 0 \text{ for } i \notin j$$

No Additional Information Models:

$$\widetilde{M}_j : \beta_i = 0 \text{ for } i \notin j$$

Information Criterion IC_d

AIC:

$$\begin{aligned}\mathrm{A}_j &= \mathrm{AIC}_j - \mathrm{AIC}_\omega \\ &= n \log \left\{ 1 + \frac{g^2 (D_\omega^2 - D_j^2)}{n - 2 + g^2 D_j^2} \right\} - 2(p_\omega - p_j)\end{aligned}$$

D_j and D_ω: Mahalanobis Distances based on x_j and x

Information Criterion:

$$\mathrm{IC}_{d,j} = n \log \left\{ 1 + \frac{g^2 (D_\omega^2 - D_j^2)}{n - 2 + g^2 D_j^2} \right\} - d(p_\omega - p_j)$$

AIC; $d = 2$ BIC; $d = \log n$

ER (Error Rate) Criterion - LS

Expected Error Rate by Linear Discriminant Function W

$$\mathrm{ER} = \frac{1}{2} P(W < 0 | \Pi^{(1)}) + \frac{1}{2} P(W > 0 | \Pi^{(2)})$$

LS-Estimator for ER based on $\boldsymbol{x}_j$**:** $\Phi(\tilde{G}_j)$

$$\tilde{G}_j = -\frac{1}{2} D_j + \frac{1}{2}(p_j - 1)\frac{1}{D_j}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) + \frac{1}{32(n-2)} D_j \left\{4p_j(4p_j - 1) - D_j^2\right\}$$

(McLachlan (1976,1980))

ER (Error Rate) Criterion - HD

Expected Error Rate by Linear Discriminant Function W

$$\mathrm{ER} = \frac{1}{2}P(W < c_0|\Pi^{(1)}) + \frac{1}{2}P(W > c_0|\Pi^{(2)})$$

$$c_0 = \frac{1}{2}\frac{n}{n-p}\left(\frac{p}{n_2} - \frac{p}{n_1}\right)$$

HD-Estimator for ER based on $\boldsymbol{x}_j$: $\Phi(G_j)$

$$G_j = -\frac{1}{2}\tilde{\Delta}_j^2\{\tilde{\Delta}_j^2 + g^{-2}(p_j - 2)\}^{-1/2}\{(n - p_j)/(n-1)\}^{1/2}$$

$$\tilde{\Delta}_j^2 = \{(n - p_j - 1)/(n-2)\}\, D_j^2 - g^{-2}(p_j - 2)$$

$$g^2 = (n_1 n_2)/n$$

(Hyodo and Kubokawa (2014), Yamada, Sakurai and Fujikoshi (2017))

Large-Sample Results

1. AIC **is not consistent**
2. BIC **is consistent**
3. AIC **and** ERC **are asymptotically equivalent**

Efficient Screening (ES) Methods , Based on IC

Subset and Subvector

$$\omega(i) = \omega \cap \{i\}^c \Leftrightarrow \boldsymbol{x}_{\omega(i)}, \qquad \omega = \{1, 2, \ldots, p\}$$

(ES) Methods based on IC_d (Zho et al (1986), etc.)

$$\mathrm{IC}_{d,\omega(i)}(= \mathrm{IC}_{d,\omega(i)} - \mathrm{IC}_{d,\omega}) > 0 \Rightarrow \text{ Select } x_i$$

$$\mathrm{IC}_{d,\omega(i)}(= \mathrm{IC}_{d,\omega(i)} - \mathrm{IC}_{d,\omega}) < 0 \Rightarrow \text{ Delete } x_i$$

Computations:

$$\mathrm{IC}_d : 2^p, \quad \mathrm{ES}_{\mathrm{ICd}} : p$$

Main Theorem I

(1) The dimension of j_* is finite

(2) For any $i \in j_*$, $\lim \Delta^2_{\{i\}\cdot\omega(i)}/\Delta^2 > 0$

(3) $d = \sqrt{n}$

$\Rightarrow \mathrm{ES}_{\mathrm{IC}_{\sqrt{n}}}$ is consistent

Simulation Experiments for $\mathrm{ES}_{\mathrm{IC}\sqrt{n}}$

Set-up

$$j_* = \{1, 2, 3\}$$
$$\boldsymbol{\Sigma} = \mathbf{I}_p$$
$$\boldsymbol{\mu}^{(1)} = (1, 1, 1, 0, \ldots, 0)$$
$$\boldsymbol{\mu}^{(2)} = (-1, -1, -1, 0, \ldots, 0)$$

Selected Probabilities Table for d=2

n_1	n_2	p	$d=2$ Under	True	Over
50	50	5	0.00	0.68	0.32
100	100	5	0.00	0.69	0.31
200	200	5	0.00	0.71	0.29
50	50	25	0.00	0.01	0.99
100	100	50	0.00	0.00	1.00
200	200	100	0.00	0.00	1.00
50	50	50	0.00	0.00	1.00
100	100	100	0.00	0.00	1.00
200	200	200	0.00	0.00	1.00

Selected Probabilities Table for d=log n and d=n^(1/2)

$d = \log n$			$d = \sqrt{n}$		
Under	True	Over	Under	True	Over
0.01	0.92	0.08	0.04	0.95	0.01
0.00	0.95	0.05	0.00	1.00	0.00
0.00	0.97	0.03	0.00	1.00	0.00
0.01	0.28	0.72	0.06	0.82	0.12
0.00	0.14	0.86	0.00	0.94	0.06
0.00	0.05	0.95	0.00	0.99	0.01
0.02	0.01	0.97	0.13	0.33	0.54
0.00	0.00	1.00	0.01	0.53	0.46
0.00	0.00	1.00	0.00	0.75	0.25

Efficient Screening (ES) Methods , Based on IC

$$Bc_j = N \log\left\{1 + \frac{\frac{N_1 N_2}{N}\left(D^2_{\omega} - D^2_{\omega(j)}\right)}{N - 2 + \frac{N_1 N_2}{N} D^2_{\omega(j)}}\right\} - \log N \left(1 + \frac{N}{N - p}\right)$$

Selected Probabilities Table for Bcj

n_1	n_2	p	Bc_j Under	True	Over
50	50	5	0.04	0.96	0.01
100	100	5	0.00	1.00	0.00
200	200	5	0.00	1.00	0.00
50	50	25	0.09	0.81	0.10
100	100	50	0.00	0.89	0.11
200	200	100	0.00	0.88	0.12
50	50	50	0.29	0.49	0.23
100	100	100	0.01	0.64	0.35
200	200	200	0.00	0.63	0.37

Lemma

p**-variate:** $\boldsymbol{x} = (\boldsymbol{x}_1', \boldsymbol{x}_2)'$, $\boldsymbol{x}_i : p_i \times 1$

D **and** D_1**; Mahalanobis distances based on** $\boldsymbol{x}$ **and** $\boldsymbol{x}_1$

$$(1) D_1^2 = (n-2)g^{-2}R, \quad R = \frac{\chi^2_{p_1}(g^2\Delta_1^2)}{\chi^2_{n-p_1-1}}$$

$$(2) D_{2\cdot 1}^2 = (n-2)\frac{\chi^2_{p_2}\left(g^2\Delta^2_{2\cdot 1}\cdot\frac{1}{1+R}\right)}{\chi^2_{n-p-1}} \times (1+R)$$

$$(3) \frac{g^2(D^2 - D_1^2)}{n-2+g^2D_1^2} = \frac{\chi^2_{p_2}(g^2\Delta^2_{2\cdot 1}(1+R)^{-1})}{\chi^2_{n-p-1}}$$

Here, $\chi^2_{p_1}$, $\chi^2_{n-p_1-1}$, $\chi^2_{p_2}$ and χ^2_{n-p-1} are independent. $g^2 = n_1n_2/n$.

Efficient Screening (ES) Methods Based on ER

$$G_j = \frac{-\frac{1}{2}\tilde{\Delta}_j^2}{\sqrt{\left\{\tilde{\Delta}_j^2 + \frac{n(p_j-2)}{n_1 n_2}\right\}\frac{n-1}{n-p_j}}}$$

$$\tilde{\Delta}_j^2 = \frac{n-p_j-1}{n-2} D_j^2 - \frac{n(p_j-2)}{n_1 n_2}$$

$$D_j^2 = (\bar{\boldsymbol{x}}_{1j} - \bar{\boldsymbol{x}}_{2j})' \mathbf{S}_j^{-1} (\bar{\boldsymbol{x}}_{1j} - \bar{\boldsymbol{x}}_{2j})$$

$$\mathbf{S}_j = \frac{1}{n-2} \sum_{i=1}^{2} \sum_{k=1}^{n_i} (\boldsymbol{x}_{ik,j} - \bar{\boldsymbol{x}}_{kj})(\boldsymbol{x}_{ik,j} - \bar{\boldsymbol{x}}_{kj})'$$

Efficient Screening (ES) Methods Based on ER

Method; $\mathrm{ES}_{\mathrm{ER,d}}$

$$\mathrm{G}_{\omega(i)} - \mathrm{G}_{\omega} > d \Rightarrow \text{ Select } x_i$$

$$\mathrm{G}_{\omega(i)} - \mathrm{G}_{\omega} < d \Rightarrow \text{ Delete } x_i$$

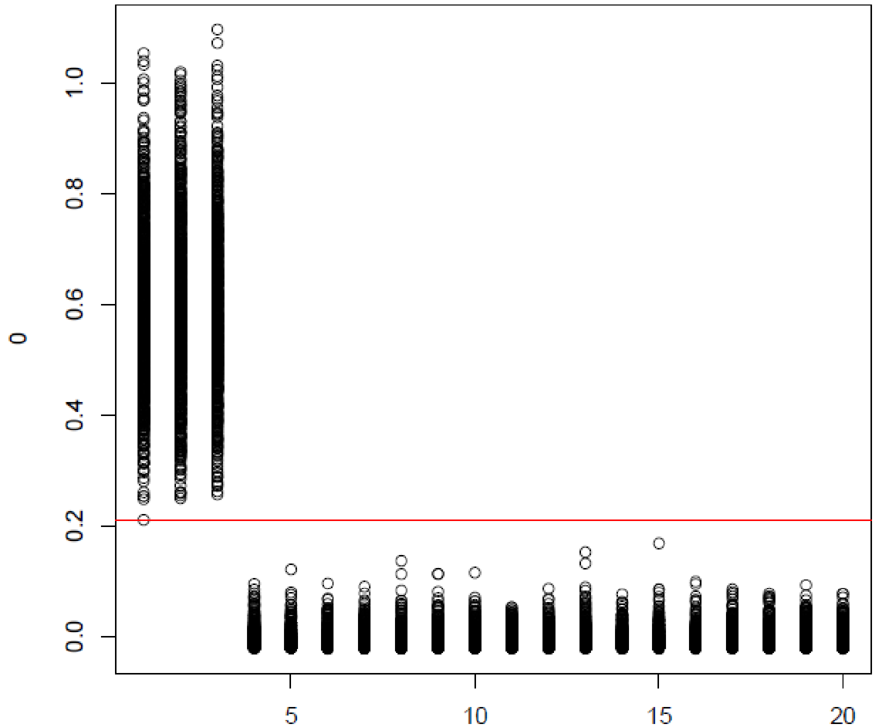
Some Candidates for d:

$$d_1 = 0, \qquad d_2 = \frac{1}{\sqrt{n}}\sqrt{\frac{p}{n}}|G_\omega|$$

Plots of $G_{\omega(j)} - G_{\omega}$

$\omega = \{1, 2, 3, \ldots, p\}$, $p = 20$, $n_1 = n_2 = 100$, $\Sigma = \mathbf{I}_p$

$\boldsymbol{\mu}^{(1)} = (2, 2, 2, 0, \ldots, 0)'$, $\boldsymbol{\mu}^{(2)} = (-2, -2, -2, 0, \ldots, 0)'$



Simulation Experiments for $\mathrm{ES}_{\mathrm{ER,d}}$

Set-up

$$j_* = \{1, 2, 3\}$$
$$\boldsymbol{\Sigma} = \mathbf{I}_p$$
$$\boldsymbol{\mu}^{(1)} = (1, 1, 1, 0, \ldots, 0)$$
$$\boldsymbol{\mu}^{(2)} = (-1, -1, -1, 0, \ldots, 0)$$

Selected Probabilities (ESER) Table for $d=d_1, d_2$

			$d = d_1$			$d = d_2$		
n_1	n_2	p	Under	True	Over	Under	True	Over
50	50	5	0.00	0.69	0.31	0.01	0.94	0.05
100	100	5	0.00	0.70	0.30	0.00	0.96	0.04
200	200	5	0.00	0.71	0.29	0.00	0.96	0.04
50	50	25	0.00	0.03	0.97	0.05	0.81	0.14
100	100	50	0.00	0.00	1.00	0.00	0.90	0.10
200	200	100	0.00	0.00	1.00	0.00	0.96	0.04
50	50	50	0.02	0.00	0.99	0.23	0.52	0.25
100	100	100	0.00	0.00	1.00	0.02	0.71	0.27
200	200	200	0.00	0.00	1.00	0.00	0.86	0.14

Future Subjects

(1) Consistency Properties of $\mathrm{ES}_{\mathrm{IC},d}$
under HD-Asymptotic Framework when :

$$j_*;\ \text{infinite},\qquad \Delta^2 = \mathrm{O}(n^\gamma)$$

(2) CHoice of d **in** $\mathrm{ES}_{\mathrm{ER},d}$
under HD-Asymptotic Framework:

(3) Theoretical study of variable selection methods
under $p >> n$

Thank you for your attention