Statistica Sinica **21** (2011), 731-740

ON BIC'S SELECTION CONSISTENCY FOR DISCRIMINANT ANALYSIS

Qiong Zhang and Hansheng Wang

University of Wisconsin, Madison and Peking University

Abstract: Linear and quadratic discriminant analysis are two useful classification methods for which the problem of variable selection is of fundamental importance. To this end, a BIC-type selection criterion (Schwarz (1978)) was recently studied by Raftery and Dean (2006). Despite its usefulness, the BIC's selection consistency (Shao (1997)) was not investigated. To fill this gap, we show that BIC, in conjunction with a backward elimination procedure, is indeed selection consistent. To confirm our asymptotic theory, a number of numerical studies are presented.

Key words and phrases: BIC, discriminant analysis, selection consistency.

1. Introduction

In supervised classification, discriminant analysis (both linear and quadratic) is extremely popular (Friedman (1989); Tibshirani et al. (2003); Guo, Hastie, and Tibshirani (2007)). Its popularity is mainly due to simplicity, interpretability, and effectiveness. In fact, empirical comparisons show that good prediction accuracy can be easily achieved by these methods (Hand (2006); Clemmensen, Hastie, and Ersbøll (2008)). Thus, a thorough understanding of discriminant analysis is important.

At the same time, very little is known about variable selection for discriminant analysis. The problem of relevant (or irrelevant) variables is not straightforward. To appreciate the difficulty, consider that in a standard linear regression model, irrelevant predictors can be taken as those with zero regression coefficients. However, for discriminant analysis, no "regression coefficient" can be defined naturally. Here, irrelevance of a variable is not clear. One can define irrelevant variables as those that provides no additional prediction power, conditional on the existence of the others; see for example Kohavi and John (1997), Raftery and Dean (2006), among others. Then, bringing in Bayes factors (Smith and Spiegelhalter (1980); Kass and Raftery (1995); Kass and Wasserman (1995); Efron and Gous (2001)), a BIC-type criterion (Schwarz (1978)) was recently studied by Raftery and Dean (2006). Despite its usefulness, the BIC's selection consistency (Shao (1997)) was not investigated. The primary objective of this

article is to fill this gap. Specifically, we show that BIC, in conjunction with a backward elimination procedure, is selection consistent. Numerical studies are presented to confirm our asymptotic theory.

The rest of the article is organized as follows. Section 2 introduces the methodology, giving both computational details and theoretical properties. Numerical studies are presented in Section 3.

2. The Methodology

2.1. Model and notations

Let (Y_i, X_i), $1 \leq i \leq n$, be the observation collected from the ith subject, where Y_i is the class label taking values in $\{1, \ldots, K\}$, and $X_i = (X_{i1}, \ldots, X_{ip})^\top$ is the associated p-dimensional predictor. We assume that $P(Y_i = k) = \pi_k > 0$ for every $1 \leq k \leq K$, and $X_i|Y_i = k$ follows a multivariate normal distribution with mean $\mu_k = (\mu_{k1}, \ldots, \mu_{kp})^\top \in \mathbb{R}^p$ and covariance $\Sigma_k \in \mathbb{R}^{p\times p}$, Σ_k is positive definite for every $1 \leq k \leq K$. Let $\mathcal{S} = \{j_1, \ldots, j_d\}$ denote a candidate model that contains $X_{ij_1}, \ldots, X_{ij_d}$ as relevant predictors. We denote its size by $|\mathcal{S}| = d$ and its complement by $\mathcal{S}^c = \mathcal{S}_F \backslash \mathcal{S}$, where $\mathcal{S}_F = \{1, \ldots, p\}$ is the full model. For an arbitrary p-dimensional vector μ_k, we write $\mu_{k(\mathcal{S})} = (\mu_{kj} : j \in \mathcal{S}) \in \mathbb{R}^{|\mathcal{S}|}$ to denote its subvector corresponding to the candidate model $\mathcal{S}$. Similarly, $\Sigma_{k(\mathcal{S})}$ denote Σ_k's submatrix corresponding to $\mathcal{S}$.

The objective of variable selection is to differentiate relevant variables from redundant ones. For this, we follow the idea of Kohavi and John (1997), and take a set of predictors $\mathcal{S}_I$ to be irrelevant if the distribution of $X_{i(\mathcal{S}_I)}|Y_i, X_{i(\mathcal{S}_R)}$ is the same as that of $X_{i(\mathcal{S}_I)}|X_{i(\mathcal{S}_R)}$, where $\mathcal{S}_R = \mathcal{S}_I^c$. Under this assumption, one can easily verify that

$$P\Big(Y_i = k\Big|X_i\Big) = P\Big(Y_i = k\Big|X_{i(\mathcal{S}_R)}\Big), \tag{2.1}$$

which implies that the model $\mathcal{S}_R$ by itself is sufficient for predicting the class label Y_i. Obviously, there exist more than one model $\mathcal{S}_R$ satisfying (2.1), e.g., $\mathcal{S}_R = \mathcal{S}_F$. However, we are only interested in the "smallest" model satisfying (2.1), defined as the intersection of all $\mathcal{S}_R$ satisfying (2.1), and denoted by $\mathcal{S}_T$. Following an argument of Cook (1998), we can show that $\mathcal{S}_T$ also satisfies (2.1). We refer to $\mathcal{S}_T$ as the true model.

Because $X_i|Y_i$ is Gaussian, any $\mathcal{S} \supset \mathcal{S}_T$ satisfies (2.1). Consequently, for each k, we have $X_{i(\mathcal{S})}|Y_i = k \sim N(\mu_{k(\mathcal{S})}, \Sigma_{k(\mathcal{S})})$, $\Sigma_{k(\mathcal{S})} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ is a positive definite matrix, and

$$X_{i(\mathcal{S}^c)}|X_{i(\mathcal{S})}, Y_i = X_{i(\mathcal{S}^c)}|X_{i(\mathcal{S})} \sim N\Big(\mu_{(\mathcal{S})} + B_{(\mathcal{S})}X_{i(\mathcal{S})}, \Sigma_{\varepsilon(\mathcal{S})}\Big) \tag{2.2}$$

for some $\mu_{(\mathcal{S})} \in \mathbb{R}^{p-|\mathcal{S}|}$, $B_{(\mathcal{S})} \in \mathbb{R}^{(p-|\mathcal{S}|)\times|\mathcal{S}|}$, and $\Sigma_{\varepsilon(\mathcal{S})} \in \mathbb{R}^{(p-|\mathcal{S}|)\times(p-|\mathcal{S}|)}$, where $\Sigma_{\varepsilon(\mathcal{S})}$ is a positive definite matrix. Moreover, because $\mathcal{S}_T$ is the "smallest" model satisfying (2.1), (2.2) is not valid for any $\mathcal{S} \not\supset \mathcal{S}_T$.

2.2. The BIC criterion

To identify the true model $\mathcal{S}_T$, we assume that we are given a set of candidate models $\mathcal{M}$. The choice of $\mathcal{M}$ is an important question that will be addressed in the next subsection. We consider the BIC criterion

$$\text{BIC} = -2 \times \log \text{likelihood} + \text{degrees of freedom} \times \log n,$$

and proceed as follows. Write the likelihood function as $\ell(\theta_{(\mathcal{S})})$, where the unknown parameter is

$$\theta_{(\mathcal{S})} = \Big\{(\mu_{(\mathcal{S})}, B_{(\mathcal{S})}, \Sigma_{\varepsilon(\mathcal{S})}) \text{ and } (\pi_k, \mu_{k(\mathcal{S})}, \Sigma_{k(\mathcal{S})}) \text{ with } 1 \le k \le K\Big\}. \quad (2.3)$$

For a candidate model $\mathcal{S}$,

$$\begin{aligned}
&-2\log \ell(\theta_{(\mathcal{S})}) \\
&\quad = \sum_{i=1}^{n}\sum_{k=1}^{K} I(Y_i = k)\Big\{\Big(X_{i(\mathcal{S})} - \mu_{k(\mathcal{S})}\Big)^{\top}\Sigma_{k(\mathcal{S})}^{-1}\Big(X_{i(\mathcal{S})} - \mu_{k(\mathcal{S})}\Big) + \log\Big|\Sigma_{k(\mathcal{S})}\Big|\Big\} \\
&\qquad + \sum_{i=1}^{n}\Big\{\Big(X_{i(\mathcal{S}^c)} - \mu_{(\mathcal{S})} - B_{(\mathcal{S})}X_{i(\mathcal{S})}\Big)^{\top}\Sigma_{\varepsilon(\mathcal{S})}^{-1}\Big(X_{i(\mathcal{S}^c)} - \mu_{(\mathcal{S})} - B_{(\mathcal{S})}X_{i(\mathcal{S})}\Big) \\
&\qquad + \log\Big|\Sigma_{\varepsilon(\mathcal{S})}\Big|\Big\} + \sum_{i=1}^{n}\sum_{k=1}^{K} I(Y_i = k)\log\pi_k. \qquad (2.4)
\end{aligned}$$

By optimizing (2.4) with respect to $\theta_{(\mathcal{S})}$, we obtain the maximum likelihood estimators

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = k), \quad \hat{\mu}_{k(\mathcal{S})} = \frac{1}{n_k}\sum_{i=1}^{n} X_{i(\mathcal{S})} I(Y_i = k),$$

$$\hat{\Sigma}_{k(\mathcal{S})} = \frac{1}{n_k}\sum_{i=1}^{n} X_{i(\mathcal{S})}X_{i(\mathcal{S})}^{\top} I(Y_i = k) - \hat{\mu}_{k(\mathcal{S})}\hat{\mu}_{k(\mathcal{S})}^{\top},$$

$$\Big(\hat{\mu}_{(\mathcal{S})}, \hat{B}_{(\mathcal{S})}\Big) = \Big\{\frac{1}{n}\sum_{i=1}^{n} X_{i(\mathcal{S}^c)}\tilde{X}_{i(\mathcal{S})}^{\top}\Big\}\Big\{\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_{i(\mathcal{S})}\tilde{X}_{i(\mathcal{S})}^{\top}\Big\}^{-1},$$

and

$$\hat{\Sigma}_{\varepsilon(\mathcal{S})} = n^{-1}\sum_{i=1}^{n}\hat{\varepsilon}_{i(\mathcal{S})}\hat{\varepsilon}_{i(\mathcal{S})}^{\top},$$

where $n_k = \sum_{i=1}^{n} I(Y_i = k)$, $\tilde{X}_{i(\mathcal{S})} = \left(1, X_{i(\mathcal{S})}^\top\right)^\top$, and $\hat{\varepsilon}_{i(\mathcal{S})} = X_{i(\mathcal{S}^c)} - \hat{\mu}_{(\mathcal{S})} - \hat{B}_{(\mathcal{S})} X_{i(\mathcal{S})}$. Denoting these MLEs by $\hat{\theta}_{(\mathcal{S})}$,

$$-2\ell(\hat{\theta}_{(\mathcal{S})}) = n\left\{ \sum_{k=1}^{K} \hat{\pi}_k \log\left|\hat{\Sigma}_{k(\mathcal{S})}\right| + \log\left|\hat{\Sigma}_{\varepsilon(\mathcal{S})}\right| \right\}, \tag{2.5}$$

where some irrelevant constants are omitted. Then the number of parameters needed for such a model specification is

$$\begin{aligned} df(\mathcal{S}) = K - 1 + K\left\{ |\mathcal{S}| + \frac{1}{2}|\mathcal{S}|\Big(|\mathcal{S}| + 1\Big) \right\} \\ + \Big(p - |\mathcal{S}|\Big)|\mathcal{S}| + \frac{1}{2}\Big(p - |\mathcal{S}|\Big)\Big(p - |\mathcal{S}| + 1\Big) + \Big(p - |\mathcal{S}|\Big), \end{aligned} \tag{2.6}$$

where the first term is due to $\{\pi_k\}$, the second to $\{\mu_{k(\mathcal{S})}, \Sigma_{k(\mathcal{S})}\}$, the third to $B_{(\mathcal{S})}$, and the last two to $\Sigma_{\varepsilon(\mathcal{S})}$ and $\mu_{(\mathcal{S})}$. One can verify that $df(\mathcal{S})$ is a monotonically increasing function in $|\mathcal{S}|$. Thus, larger candidate models lead to larger degrees of freedom. Combing the results from (2.5) and (2.6), we have

$$\mathrm{BIC}(\mathcal{S}) = -2\log \ell(\hat{\theta}_{(\mathcal{S})}) + df(\mathcal{S}) \times \log n. \tag{2.7}$$

Thereafter, the best model can be selected as $\hat{\mathcal{S}} = \mathrm{argmin}_{\mathcal{S}\in\mathcal{M}}\mathrm{BIC}(\mathcal{S})$.

2.3. A backward algorithm

The generation of the candidate model set $\mathcal{M}$ is very important. To this end, we consider here a standard backward algorithm conducted as follows.

Step 1: (*The Initialization Step*). Set $\mathcal{S}_{(0)} = \mathcal{S}_F$, and the relevant $X_{i(\mathcal{S}_{(0)})} = X_i$, $X_{i(\mathcal{S}^c_{(0)})} = \varnothing$, $1 \le i \le n$. Calculate $\mathrm{BIC}(\mathcal{S}_{(0)})$.

Step 2: (*The Evaluation Step*). In the t-th step($t > 0$), given $\mathcal{S}_{(t-1)}$, $X_{i(\mathcal{S}_{(t-1)})}$, and $X_{i(\mathcal{S}^c_{(t-1)})}$, compute $d_{(t)} = \mathrm{argmin}_{j\in\mathcal{S}_{(t-1)}}\mathrm{BIC}(\mathcal{S}_{(t-1)}\backslash\{j\})$ and update $\mathcal{S}_{(t)} = \mathcal{S}_{(t-1)}\backslash\{d_{(t)}\}$.

Step 3: (*The Selection Step*). Iterate Step 2 p times, generating a sequence of candidate models $\mathcal{M} = \{\mathcal{S}_{(t)} : 0 \le t \le p\}$. Based on $\mathcal{M}$, the best model is $\hat{\mathcal{S}} = \mathrm{argmin}_{\mathcal{S}\in\mathcal{M}}\mathrm{BIC}(\mathcal{S})$.

We show in the following theorem (the proof is in Appendix B) that, with probability tending to one, $\hat{\mathcal{S}} = \mathcal{S}_T$. Thus, the BIC criterion together with this backward algorithm is indeed selection consistent.

Theorem 1. *Under* (2.2), $P(\hat{\mathcal{S}} = \mathcal{S}_T) \to 1$.

As a cautionary note, we remark that one has only a guarantee that the proposed backward elimination procedure converges to the true model asymptotically; with a finite dataset, whether one gets convergence to the model with smallest BIC score is not guaranteed.

3. Numerical Experiments

3.1. Simulation studies

To evaluate the finite sample performances of the proposed method, two simulation experiments borrowed from Raftery and Dean (2006) were conducted.

Example 1. There were seven variables and two classes; the first two variables were relevant and were generated from bivariate normal distributions. For the first class, the mean vector and covariance matrix were, respectively, $\mu_{1(\mathcal{S}_T)} = (2.5, -1.0)^\top$ and $\Sigma_{2(\mathcal{S}_T)} = [1, 0; 0, 1] \in \mathbb{R}^{2\times 2}$; for the second class, they were, respectively, $\mu_{2(\mathcal{S}_T)} = (-0.5, 0)^\top$ and $\Sigma_{2(\mathcal{S}_T)} = [1.1, 0.5; 0.5, 0.85] \in \mathbb{R}^{2\times 2}$. The remaining five X_{ij} variables were independently generated as $N(m_j, 1)$, where m_j was $U[0, 1]$.

Example 2. There were fifteen variables and two classes; the first two variables were relevant, and generated as in Example 1. The next eight variables were irrelevant and generated from the standard normal distribution; the next two variables were also irrelevant and were generated from a bivariate normal distribution with mean 0, variance 1, and correlation 0.5. The thirteenth predictor was

$$X_{i13} = \alpha_{13} + \beta_{13} X_{i1} + \varepsilon_{i13}, \tag{3.1}$$

where α_{13} was generated from $U[0, 1]$, β_{13} from $U[0, 10]$, and ε_{i13} from $N(0, 16)$. The fourteenth variable X_{i14} was generated in a similar manner as X_{i13}. But with X_{i1} in (3.1) replaced by X_{i2}. Lastly, $X_{i15} = \alpha_{15} + \beta_a X_{i1} + \beta_b X_{i2} + \varepsilon_{i15}$, where α_{15} and ε_{i15} were generated in a similar manner as α_{13} and ε_{i13}, while both β_a and β_b were indepently $U[0, 1]$.

For a given simulation model and parameter setup (e.g., the sample size n), two independent but identically distributed datasets were generated. The first dataset served as the training data while the second one was used for testing. We then applied the BIC criterion with the backward algorithm to the training data. By doing so, a "best" model was selected. Subsequently, the "best" model's prediction accuracy (in terms of mis-classification error, ME) was evaluated based on the testing data, via the method of quadratic discriminant analysis (Johnson and Wichern (2003)). For a reliable evaluation, the experiment was replicated 100 times, the average value of ME, AME, computed and reported in Table 1.

We next evaluated the BIC method's model selection consistency. To this end, we took a selected model to be correct if $\hat{\mathcal{S}} = \mathcal{S}_T$, and the percentage of

Table 1. Detailed results for the two simulation examples. n: the sample size; FULL: the quadratic discriminant analysis without variable selection; CV: the model selected by cross-validation in terms of minimal misclassification error; AIC: the model selected by the AIC; BIC: the model selected by the BIC. PCF: the percentage of the correct fits; AFN: the average false negatives; AFP: the average false positives; AME: the average mis-classification error; AMS: the average model size;

Example	n	Selection Method	PCF (%)	AFN	AFP	AME(%)	AMS
1	75	FULL	–	–	–	6.73	7.00
		CV	17	0.06	1.72	5.23	3.66
		AIC	67	0.02	0.37	4.36	2.35
		BIC	85	0.14	0.01	4.40	1.87
	100	FULL	–	–	–	5.82	7.00
		CV	24	0.05	1.80	5.15	3.75
		AIC	74	0.00	0.30	4.39	2.30
		BIC	93	0.06	0.01	4.24	1.95
	150	FULL	–	–	–	5.25	7.00
		CV	14	0.00	2.17	4.64	4.17
		AIC	78	0.00	0.27	4.35	2.27
		BIC	99	0.01	0.00	4.25	1.99
2	75	FULL	–	–	–	16.37	15.00
		CV	11	0.26	2.61	6.95	4.35
		AIC	37	0.23	1.38	5.76	3.15
		BIC	67	0.33	0.14	5.17	1.81
	100	FULL	–	–	–	12.42	15.00
		CV	9	0.18	3.11	6.61	4.93
		AIC	49	0.16	0.86	5.21	2.70
		BIC	79	0.21	0.15	4.73	1.94
	150	FULL	–	–	–	8.59	15.00
		CV	18	0.03	2.96	5.51	4.93
		AIC	57	0.03	0.71	4.71	2.68
		BIC	95	0.03	0.05	4.59	2.02

the correct fit (PCF) across 100 replications was computed. To better gauge our method's underfitting effect, took the average false negative (AFN) frequency as the average number of the relevant variables missed by $\hat{\mathcal{S}}$; to characterize the overfitting effect, we took the average false positive (AFP) frequency as the average number of irrelevant variables included in $\hat{\mathcal{S}}$. Lastly, the average model size (AMS) of $\hat{\mathcal{S}}$ was also summarized. For comparison proposes, we considered the FULL model (the model without going through variable selection), the CV model (the model selected by cross-validation in terms of ME), and the AIC model (the model selected by the AIC criterion, where the factor $\log n$ in (2.7) is replaced by 2).

Table 2. The detailed analysis results for the Landsat Satellite data based on 100 simulation replications. FULL: the quadratic discriminant analysis without variable selection; CV: the model selected by cross-validation in terms of minimal mis-classification error; AIC: the model selected by the AIC; BIC: the model selected by the BIC. AME: the average mis-classification error; AMS: the average model size.

Selection Methods	AME(%)	AMS
FULL	17.90	36.00
CV	16.48	13.41
AIC	16.66	24.92
BIC	16.36	12.01

According to Table 1, as n increases, the BIC's PCF value approaches 100% very quickly, numerically confirming that the BIC criterion (2.7) with the backward elimination procedure is selection consistent. No similar pattern was observed for other methods. As a consequence, we find that the prediction accuracy of the BIC models to be very competitive, particularly in large sample size situations. It is noteworthy that this prediction accuracy was achieved with a much smaller average model size than with competing methods.

3.2. The Landsat satellite data

To further illustrate the usefulness of our method, we considered the Landsat Satellite Data that is publicly available at the UCI Machine Learning Repository; see `http://www.ics.uci.edu/~mlearn/`. The database consists of the multi-spectral values of pixels in a satellite image. The sample contains a total of 6 different classes and has 36 predictive variables. The original dataset has been divided into a training set with 4,435 samples and a testing set with 2,000 samples. We used 1,000 samples (randomly selected from the training data) to estimate and select the model. Based on the selected model, we evaluated the BIC model's ME on the testing data. We replicated this experiment 100 times and summarized the results in Table 2. As one can see, the models selected by BIC have both the smallest average model size and the smallest misclassification error.

Acknowledgement

The author is grateful to the Editor, the Associate Editor and referees for the helpful comments and advices. This research is supported in part by a NSFC grant (No. 10771006).

Appendix

Appendix A. A useful lemma

Lemma 1. *Assuming that* $\mathcal{S}_2 \subset \mathcal{S}_1$ *and* $|\mathcal{S}_1 \backslash \mathcal{S}_2| = 1$,

$$-2n^{-1}\ell(\hat{\theta}_{(\mathcal{S}_1)}) + 2n^{-1}\ell(\hat{\theta}_{(\mathcal{S}_2)}) = O_p(n^{-1}) \quad \text{if} \quad \mathcal{S}_T \subseteq \mathcal{S}_2 \subseteq \mathcal{S}_1, \tag{A.1}$$

$$-2n^{-1}\ell(\hat{\theta}_{(\mathcal{S}_1)}) + 2n^{-1}\ell(\hat{\theta}_{(\mathcal{S}_2)}) = -C_{\mathcal{S}_1,\mathcal{S}_2} + O_p(n^{-1}) \quad \text{if} \quad |\mathcal{S}_T \backslash \mathcal{S}_2| \neq 0, \tag{A.2}$$

where $C_{\mathcal{S}_1,\mathcal{S}_2} \geq 0$ *is a constant given* $\mathcal{S}_1$ *and* $\mathcal{S}_2$. *In addition,* $C_{\mathcal{S}_1,\mathcal{S}_2} > 0$ *holds if* $\mathcal{S}_T \subseteq \mathcal{S}_1$.

Proof. First, (A.1) is made clear by following Theorem 6.5 on page 432 in Shao (2003). Consider (A.2). For simplicity, write $\mathcal{S}_1 = \{1, \ldots, b\}$ and $\mathcal{S}_2 = \mathcal{S}_1 \backslash \{b\}$, $X_{i(a)} = (X_{i1}, \ldots, X_{i,b-1})^\top$ and $X_{i(c)} = (X_{i,b+1}, \ldots, X_{ip})^\top$. According to Raftery and Dean (2006), we can write

$$-2n^{-1}\ell(\hat{\theta}_{(\mathcal{S}_1)}) + 2n^{-1}\ell(\hat{\theta}_{(\mathcal{S}_2)}) = \sum_{k=1}^{K} \hat{\pi}_k \log \hat{\sigma}^2_{k,b|a} - \log \hat{\sigma}^2_{b|a},$$

where $\sigma^2_{k,b|a} = \mathrm{var}(X_{ib}|Y_i = k, X_{i(a)})$ and $\sigma^2_{b|a} = \mathrm{var}(X_{ib}|X_{i(a)})$. Moreover, $\hat{\sigma}^2_{k,b|a}$ and $\hat{\sigma}^2_{b|a}$ are the corresponding MLE. Then we have

$$\begin{aligned}
&\sum_{k=1}^{K} \hat{\pi}_k \log \hat{\sigma}^2_{k,b|a} - \log \hat{\sigma}^2_{b|a} \to_p \sum_{k=1}^{K} \pi_k \log \sigma^2_{k,b|a} - \log \sigma^2_{b|a} \\
&\quad = \sum_{k=1}^{K} \pi_k \log \sigma^2_{k,b|a} - \log\Big(\sum_{k=1}^{K} \pi_k \sigma^2_{k,b|a}\Big) + \log\Big(\sum_{k=1}^{K} \pi_k \sigma^2_{k,b|a}\Big) - \log \sigma^2_{b|a} \\
&\quad \triangleq -C_{\mathcal{S}_1,\mathcal{S}_2},
\end{aligned}$$

where $C_{\mathcal{S}_1,\mathcal{S}_2}$ is constant given $\mathcal{S}_1$ and $\mathcal{S}_2$. According to Jensen Inequality, $\sum_{k=1}^{K} \pi_k \log \sigma^2_{k,b|a} \leq \log\Big(\sum_{k=1}^{K} \pi_k \sigma^2_{k,b|a}\Big)$. Also, since $\sigma^2_{b|a} = \sum_{k=1}^{K} \pi_k \Big\{\sigma^2_{k,b|a} + \big\{E(X_{ib}|Y_i = k, X_{i(a)}) - E(X_{ib}|X_{i(a)})\big\}^2\Big\}$, $\log\Big(\sum_{k=1}^{K} \pi_k \sigma^2_{k,b|a}\Big) \leq \log \sigma^2_{b|a}$. Consequently, $C_{\mathcal{S}_1,\mathcal{S}_2} \geq 0$. Moreover, if $\mathcal{S}_T \subseteq \mathcal{S}_1$ with $|\mathcal{S}_T \backslash \mathcal{S}_2| \neq 0$, we should have either $\sigma^2_{k_1,b|a} \neq \sigma^2_{k_2,b|a}$ for some $k_1 \neq k_2$, or $E(X_{ib}|Y_i = k, X_{i(a)}) \neq E(X_{ib}|X_{ia})$ for some k, which further leads to either $\sum_{k=1}^{K} \pi_k \log \sigma^2_{k,b|a} - \log\Big(\sum_{k=1}^{K} \pi_k \sigma^2_{k,b|a}\Big) < 0$ or $\log\Big(\sum_{k=1}^{K} \pi_k \sigma^2_{k,b|a}\Big) - \log \sigma^2_{b|a} < 0$ should be true. Then $C_{\mathcal{S}_1,\mathcal{S}_2} > 0$ holds. This completes the proof of Lemma 1.

Appendix B. Proof of Theorem 1

At the t-th step of the backward algorithm. Assume $\mathcal{S}_T \subseteq \mathcal{S}_{(t-1)}$ and $|\mathcal{S}_{(t-1)}| - |\mathcal{S}_T| > 0$. Let $j_{d_1}, j_{d_2} \in \mathcal{S}_{(t-1)}$ such that $j_{d_1} \in \mathcal{S}_T, j_{d_2} \in \mathcal{S}_T^c$. Write $\mathcal{S}_{d_1} = \mathcal{S}_{(t-1)} \backslash \{j_{d_1}\}$, and $\mathcal{S}_{d_2} = \mathcal{S}_{(t-1)} \backslash \{j_{d_2}\}$. Then, by (2.7) and Lemma 1

$$n^{-1}\Big\{\mathrm{BIC}(\mathcal{S}_{d_1}) - \mathrm{BIC}(\mathcal{S}_{(t-1)})\Big\} = C_{\mathcal{S}_{(t-1)},\mathcal{S}_{d_1}} + O_p(n^{-1}) - df_1 \times \frac{\log n}{n}, \quad \text{(A.3)}$$

$$n^{-1}\Big\{\mathrm{BIC}(\mathcal{S}_{d_2}) - \mathrm{BIC}(\mathcal{S}_{(t-1)})\Big\} = O_p(n^{-1}) - df_1 \times \frac{\log n}{n}, \quad \text{(A.4)}$$

where $C_{\mathcal{S}_{(t-1)},\mathcal{S}_{d_1}} > 0$ and $df_1 = (|\mathcal{S}_{(t-1)}| + 1) \times (K-1)$. By combining (A.3) and (A.4), we can verify that $P\Big\{\mathrm{BIC}(\mathcal{S}_{d_1}) > \mathrm{BIC}(\mathcal{S}_{d_2})\Big\} \to 1$ as $n \to \infty$. Consequently, with probability tending to one, we will not eliminate any relevant variable in the tth step. Thus, with probability tending to 1, we must have $\mathcal{S}_T \subset \mathcal{S}_{(t)}$ as long as $|\mathcal{S}_{(t-1)}| - |\mathcal{S}_T| > 0$. This further implies that

$$P\Big\{\mathcal{S}_T \in \mathcal{M}\Big\} \to 1 \ \text{ as } \ n \to \infty. \quad \text{(A.5)}$$

By (A.5), we know that, with probability tending to 1, the true model must be included in candidate model set $\mathcal{M}$. Next, we will show that BIC-criterion will indeed identify the true model consistently. To this end, we consider an arbitrary candidate model $\mathcal{S}_{(t)} \in \mathcal{M}$, but $\mathcal{S}_{(t)} \neq \mathcal{S}_T$. Due to the natural of backward algorithm, we know that, $\mathcal{S}_{(t)}$ satisfies either $\mathcal{S}_{(t)} \supset \mathcal{S}_T$ or $\mathcal{S}_{(t)} \subset \mathcal{S}_T$. Then by Lemma 1, we have

$$n^{-1}\Big\{\mathrm{BIC}(\mathcal{S}_{(t)}) - \mathrm{BIC}(\mathcal{S}_T)\Big\} = \sum_{l=t+1}^{p-|\mathcal{S}_T|} C_{\mathcal{S}_{(l)},\mathcal{S}_{(l-1)}} - df_2 \times \frac{\log n}{n}, \text{ if } \mathcal{S}_{(t)} \subset \mathcal{S}_T, \quad \text{(A.6)}$$

$$n^{-1}\Big\{\mathrm{BIC}(\mathcal{S}_T) - \mathrm{BIC}(\mathcal{S}_{(t)})\Big\} = O_p(n^{-1}) - df_2 \times \frac{\log n}{n}, \text{ if } \mathcal{S}_{(t)} \supset \mathcal{S}_T, \quad \text{(A.7)}$$

where $\sum_{l=t+1}^{p-|\mathcal{S}_T|} C_{\mathcal{S}_{(l)},\mathcal{S}_{(l-1)}} \geq C_{\mathcal{S}_T,\mathcal{S}_{(p-|\mathcal{S}_T|-1)}} > 0$ and $df_2 = |df(\mathcal{S}_{(t)}) - df(\mathcal{S}_T)|$. By (A.6) and (A.7), we have $\mathrm{BIC}(\mathcal{S}_T) < \mathrm{BIC}(\mathcal{S}_{(t)})$ with probability tending to 1, regardless of whether $\mathcal{S}_{(t)} \subset \mathcal{S}_T$ or $\mathcal{S}_{(t)} \supset \mathcal{S}_T$. This completes the proof.

References

Clemmensen, L., Hastie, T. and Ersbøll (2008). Sparse discriminant analysis. Technical Report, Department of Statistics, Stanford University.

Cook, R. D. (1998). *Regression Graphics*. John Wiley, New York.

Efron, B. and Gous, A. (2001). Scales of evidence for model selection: Fisher versus Jeffreys. *IMS Lecture Notes* **38**, 209-249.

Friedman, J. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.* **84**, 165-175.

Guo, Y., Hastie, T. and Tibshirani, R. (2007). Regularized discriminant analysis and its application in microarrays *Biostatistics*. **1**, 86-100.

Hand, D. J. (2006). Classifier technology and the illusion of the progress. *Statist. Sci.* **21**, 115.

Johnson, R. A. and Wichern, D. W. (2003). *Applied Multivariate Statistical Analysis*. 5th Edition. Pearson Education.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors *J. Amer. Statist. Assoc.*. **90**, 773-795.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypothese and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.* **90**, 928-934.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence* **97**, 273-324.

Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering. *J. Amer. Statist. Assoc.* **101**, 168-178.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Shao, J. (1997). An asymptotic theory for linear model selection (with discussion). *Statist. Sinica* **7**, 221-264.

Shao, J. (2003). *Mathematical Statistics*. 2nd edition. Springer, New York.

Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models *J. Roy. Statist. Soc. Ser. B* **42**, 213-220.

Tibshirani, R., Hastie, T., Narashimhan, B. and Chu, G. (2003). Class prediction by nearest shrunken centriods with applications to DNA microarrays. *Statist. Sci.* **18**, 104-117.

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA.

E-mail: qzhang@stat.wisc.edu

Guanghua School of Management, Peking University, Beijing, 100871, P.R. China.

E-mail: hansheng@gsm.pku.edu.cn

(Received October 2008; accepted October 2009)