On BIC's Selection Consistency for Discriminant Analysis

Qiong Zhang and Hansheng Wang

Guanghua School of Management, Peking University

This version: April 23, 2009

Abstract

Linear and quadratic discriminant analysis are two very 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 fulfill this important gap, we show theoretically that the 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: BIC; Discriminant Analysis; Selection Consistency

[†]Both Qiong Zhang and Hansheng Wang (*the corresponding author) are from Department of Business Statistics & Econometrics at Guanghua School of Management, Peking University, Beijing, 100871, P. R. China. (E-mail: {zhangqiong, hansheng}@gsm.pku.edu.cn). This research is supported in part by NSFC (Grant No. 10771006) and also a research grant from Microsoft Research Asia.

1. INTRODUCTION

In supervised classification, the discriminant analysis (both linear and quadratic) is extremely popular for real problems (Friedman, 1989; Tibshirani et al., 2003; Guo et al., 2007). Their popularity is mainly due to their simplicity, interpretability, and also effectiveness. In fact, empirical comparisons do show that a large portion of the prediction accuracy can be easily achieved by those classical methods (Hand, 2006; Clemmensen et al., 2008). Thus, a thorough understanding of the discriminant analysis is fundamentally important.

In contrast to its popularity, very limited has been known about variable selection for discriminant analysis. The basic problem like the definition of relevant (or irrelevant) variables is not immediately straightforward. To appreciate the difficulty, one can consider a standard linear regression model, where irrelevant predictors can be defined easily as those with zero regression coefficients. However, for discriminant analysis, no "regression coefficient" can be defined naturally. Then, how to define a variable to be irrelevant is not immediately straightforward. One way to solve the problem is to define irrelevant variables as those, who cannot provide any additional prediction power, conditional on the existence of the others; see for example Kohavi and John (1997), Raftery and Dean (2006), among others. Based on such a definition and the idea of the 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 theoretically investigated. Then, the primary objective of this article is to fulfill this important gap by supplying a rigorous theoretical proof. Specifically, we show that the BIC in conjunction with a backward elimination procedure is indeed selection consistent. Numerical studies are presented

to confirm our asymptotic theory.

The rest of the article is organized as the follows. The next section introduces the new methodology with 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) with $1 \leq i \leq n$ be the observation collected from the ith subject, where Y_i is the class label taking values in $\{1, 2, \ldots, K\}$ and $X_i = (X_{i1}, \cdots, X_{ip})^\top \in \mathbb{R}^p$ is the associated p-dimensional predictor. Furthermore, 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 $\mu_k = (\mu_{k1}, \cdots, \mu_{kp})^\top \in \mathbb{R}^p$ and covariance $\Sigma_k \in \mathbb{R}^{p\times p}$, where Σ_k is assumed to be positive definite for every $1 \leq k \leq K$. Next, we use a generic notation $\mathcal{S} = \{j_1, \cdots, j_d\}$ to denote a candidate model, which contains $X_{ij_1}, \cdots, X_{ij_d}$ 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 $\mathcal{S}_F = \{1, \cdots, p\}$ is the full model. For an arbitrary p-dimensional vector (e.g.) μ_k, we use $\mu_{k(\mathcal{S})} = (\mu_{kj} : j \in \mathcal{S}) \in \mathbb{R}^{|\mathcal{S}|}$ 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 those truly relevant variables from those redundant ones. Under a linear regression setup, this is typically related to identifying those predictors with nonzero regression coefficients. However, in supervised learning problems, what kind of predictors should be considered as relevant/irrelevant should be carefully defined. To this end, we follow the idea of Kohavi and John (1997), and define an arbitrary set of predictors $\mathcal{S}_I$ to be irrelevant, if it satisfies that 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), which can be defined as the intersection of all $\mathcal{S}_R$ satisfying (2.1) and is denoted by $\mathcal{S}_T$. Following similar argument as in Cook (1998), we can show that $\mathcal{S}_T$ also satisfies (2.1). Consequently, we can refer to $\mathcal{S}_T$ as our uniquely defined true model.

Because the distribution of $X_i|Y_i$ is Gaussian, then for any $\mathcal{S} \supset \mathcal{S}_T$, we must have $\mathcal{S}$ 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), thus the relationship (2.2) is not valid for any $\mathcal{S}$ satisfying $\mathcal{S} \not\supset \mathcal{S}_T$.

2.2. The BIC Criterion

To identify the true model $\mathcal{S}_T$, we assume that we are given a set of candidate models, which are collected by $\mathcal{M}$. Practically, how to generate $\mathcal{M}$ is a very important question, which will be carefully addressed in the next subsection. Following traditional

definition (Raftery and Dean, 2006), we consider a BIC criterion defined as

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

We can then implement the above BIC criterion as the follows. Firstly, we need to evaluate the log likelihood function under a given candidate model $\mathcal{S}$. Secondly, we then evaluate the degrees of freedom under that particular model. We firstly consider the likelihood function, denoted by $\ell(\theta_{(\mathcal{S})})$, where the unknown parameter is

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

Then, consider an arbitrary candidate model $\mathcal{S}$. By (2.2), we have

$$-2\log \ell(\theta_{(\mathcal{S})}) = \sum_{i=1}^{n}\sum_{k=1}^{K} I(Y_i = k)\left\{\Big(X_{i(\mathcal{S})} - \mu_{k(\mathcal{S})}\Big)^{\top}\Sigma_{k(\mathcal{S})}^{-1}\Big(X_{i(\mathcal{S})} - \mu_{k(\mathcal{S})}\Big) + \log\Big|\Sigma_{k(\mathcal{S})}\Big|\right\}$$

$$+\sum_{i=1}^{n}\left\{\Big(X_{i(\mathcal{S}^c)} - \mu_{(\mathcal{S})} - B_{(\mathcal{S})}X_{i(\mathcal{S})}\Big)^{\top}\Sigma_{\varepsilon(\mathcal{S})}^{-1}\Big(X_{i(\mathcal{S}^c)} - \mu_{(\mathcal{S})} - B_{(\mathcal{S})}X_{i(\mathcal{S})}\Big) + \log\Big|\Sigma_{\varepsilon(\mathcal{S})}\Big|\right\}$$

$$+\sum_{i=1}^{n}\sum_{k=1}^{K} I(Y_i = k)\log\pi_k. \tag{2.4}$$

Then, by optimizing (2.4) with respect to $\theta_{(\mathcal{S})}$, we obtain a set of maximum likelihood estimators, which are given by

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

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

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

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

where $n_k = \sum_{i=1}^{n} I(Y_i = k)$, $\tilde{X}_{i(\mathcal{S})} = \Big(1, X_{i(\mathcal{S})}^\top\Big)^\top$ and $\hat{\varepsilon}_{i(\mathcal{S})} = X_{i(\mathcal{S}^c)} - \hat{\mu}_{(\mathcal{S})} - \hat{B}_{(\mathcal{S})}X_{i(\mathcal{S})}$. We then collect those MLEs by $\hat{\theta}_{(\mathcal{S})}$, and evaluate the $-2\times$loglikelihood as

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

where some irrelevant constants are omitted. Next, due to each term in (2.3), one can count the number of parameters needed by such a model specification as

$$df(\mathcal{S}) = K - 1 + K\left\{|\mathcal{S}| + \frac{1}{2}|\mathcal{S}|\Big(|\mathcal{S}| + 1\Big)\right\}$$

$$+\Big(p - |\mathcal{S}|\Big)|\mathcal{S}| + \frac{1}{2}\Big(p - |\mathcal{S}|\Big)\Big(p - |\mathcal{S}| + 1\Big) + \Big(p - |\mathcal{S}|\Big), \tag{2.6}$$

where the first term is due to $\{\pi_k\}$, the second one is due to $\{\mu_{k(\mathcal{S})}, \Sigma_{k(\mathcal{S})}\}$, the third term is due to $B_{(\mathcal{S})}$, and the last two are due 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 always lead to larger degrees of freedom. Combing the results from (2.5) and (2.6), we obtain a BIC criterion as

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

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

2.3. A Backward Algorithm

As we mentioned before, practically how to generate the candidate model set $\mathcal{M}$ is

a very important problem. To this end, we considered here a very standard backward algorithm, and it could be conducted as the follows.

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

Step 2 : (*The Evaluation Step*). In the t-th step$(t > 0)$, we are given $\mathcal{S}_{(t-1)}$, $X_{i(\mathcal{S}_{(t-1)})}$, and $X_{i(\mathcal{S}^c_{(t-1)})}$. Then, we 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*). We iterate Step 2 for p times, which generates a sequence of candidate model $\mathcal{M} = \{\mathcal{S}_{(t)} : 0 \leq t \leq p\}$. Based on $\mathcal{M}$, the best model is then selected as $\hat{\mathcal{S}} = \mathrm{argmin}_{\mathcal{S} \in \mathcal{M}} \mathrm{BIC}(\mathcal{S})$.

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

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

As a cautionary note, we would like to kindly remark that the above theorem only guarantees that the proposed backward elimination procedure will converge to the true model asymptotically. Nevertheless, with a finite dataset, whether it will converge to the model with smallest BIC score is not always 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) are conducted.

Example 1. This example includes a total of seven variables and two classes. The first two variables are relevant ones and are generated from bivariate normal distributions. For the first class, the mean vector and covariance matrix are given by, 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 are respectively $\mu_{2(\mathcal{S}_T)} = (-0.5, 0)^\top$ and $\Sigma_{2(\mathcal{S}_T)} = [1.1, 0.5; 0.5, 0.85] \in \mathbb{R}^{2\times 2}$. Then, the remaining five X_{ij} variables are independently generated from $N(m_j, 1)$, where m_j is generated from $U[0, 1]$.

Example 2. Here we consider another model with a total of fifteen variables and two classes. The first two variables are relevant ones and are generated exactly the same as in Example 1. The next eight variables are irrelevant ones and are all generated from standard normal distribution. The subsequent two variables are also irrelevant ones and are generated from a bivariate normal distribution with mean 0, variance 1, and correlation 0.5. The thirteenth predictor is given by

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

where α_{13} is generated from $U[0, 1]$, β_{13} from $U[0, 10]$, and ε_{i13} from $N(0, 16)$. The fourteenth variable X_{i14} is generated in a similar manner as X_{i13}, however, 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} are also generated in a similar manner as α_{13} and ε_{i13}, while both β_a and β_b are independently generated from $U[0, 1]$.

For a given simulation model and parameter setup (e.g., the sample size n), two independent but identically distributed datasets are generated. The first dataset serves

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 mis-classification 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

as our training data while the second one will be used for testing. We then apply our method (i.e., the BIC criterion with the backward algorithm) to the training data. By doing so, a "best" model can be selected. Subsequently, the "best" model's prediction accuracy (in terms of Mis-classification Error, ME) is evaluated based on the testing

data, via the method of quadratic discriminant analysis (Johnson and Wichern, 2003). For a reliable evaluation, such an experiment is randomly replicated for a total of 100 times. Then, the average value of ME (i.e., Average Mis-classification Error, AME) is computed and reported in Table 1. We next evaluate the BIC method's model selection consistency. To this end, we define a selected model as a correct fit if the selected model (i.e., $\hat{\mathcal{S}}$) is exactly the same as the true model, i.e., $\hat{\mathcal{S}} = \mathcal{S}_T$. Then, the Percentage of the Correct Fit (PCF) across the 100 simulation replications is computed. To better gauge our method's underfitting effect, we define an Average False Negative (AFN) frequency as the average number of relevant variables missed by $\hat{\mathcal{S}}$. To further characterize the overfitting effect, we define an Average False Positive (AFP) frequency as the average number of irrelevant variables been included by $\hat{\mathcal{S}}$. Lastly, the Average Model Size (AMS) of $\hat{\mathcal{S}}$ is also summarized. For comparison propose, the following models are also included. They are, respectively, the FULL model (the model without going through variable selection), the CV model (the model selected by cross-validation in terms of ME), the AIC model (the model selected by the AIC criterion, where the factor $\log n$ in (2.7) is replaced by 2). Lastly, our method is denoted by BIC in Table 1.

According to Table 1, we find that as the sample size n increases, the BIC's PCF value approaches 100% very quickly, which numerically confirms that the BIC criterion (2.7) together with the backward elimination procedure is indeed selection consistent. On the other hand, no similar pattern was observed for other methods, which suggests that those methods are unlikely to be selection consistent. As a consequence, we find that the prediction accuracy of the BIC models are very competitive, particularly in the large sample size situations. It is noteworthy that such a competitive prediction accuracy is achieved with a much smaller average model size, as compared with other competing methods.

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

3.2. The Landsat Satellite Data

To further illustrate the usefulness of the new method, we consider here the Landsat Satellite Data, which 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 already been divided into a training set with 4435 samples and a testing set with 2000 samples. For experiment purpose, we only use 1000 samples (randomly selected from the training data) to estimate and select the model. Based on the selected model, we also evaluate the BIC model's ME on the testing data. For a reliable evaluation, we randomly replicated this experiment for a total of 100 times and then summarized the detailed results in Table 2. As one can see, the models selected by the BIC have both the smallest average model size and mis-classification error.

APPENDIX

Appendix A. A Useful Lemma

The following lemma is useful for proving the new method's selection consistency. Thus, it is formally stated and proved first.

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

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

$$-2n^{-1}\ell(\hat{\theta}_{(\mathcal{S}_1)}) + 2n^{-1}\ell(\hat{\theta}_{(\mathcal{S}_2)}) < -C_{\mathcal{S}_1,\mathcal{S}_2}, \quad \text{if} \quad \mathcal{S}_T \nsubseteq \mathcal{S}_2, \quad \text{and} \quad \mathcal{S}_T \subseteq \mathcal{S}_1, \tag{A.3}$$

where $C_{\mathcal{S}_1,\mathcal{S}_2} > 0$ *is a constant given* $\mathcal{S}_1$ *and* $\mathcal{S}_2$*. In addition,* (A.3) *still holds under the condition that* $\mathcal{S}_1 \subset \mathcal{S}_T$*.*

Proof. Firstly, (A.2) is clear by following Theorem 6.5 in Shao (2003). We then directly going to (A.3). For simplicity, we denote $\mathcal{S}_1 = \{1, 2, \cdots, b\}$ and $\mathcal{S}_2 = \mathcal{S}_1 \backslash \{b\}$. Correspondingly, we denote $X_{i(a)} = (X_{i1}, \cdots, X_{i,b-1})^\top$ and $X_{i(c)} = (X_{i,b+1}, \cdots, X_{ip})^\top$. According to Raftery and Dean (2006) and Jensen inequality, we can easily write the difference of loglikelihood functions between $\mathcal{S}_1$ and $\mathcal{S}_2$,

$$\begin{aligned} &-2n^{-1}\ell(\hat{\theta}_{(\mathcal{S}_1)}) + 2n^{-1}\ell(\hat{\theta}_{(\mathcal{S}_2)}) \\ &= \sum_{k=1}^{K} \hat{\pi}_k \log \hat{\sigma}^2_{k,b|a} - \log \hat{\sigma}^2_{b|a} \leq \log\Big(\sum_{k=1}^{K} \hat{\pi}_k \hat{\sigma}^2_{k,b|a}\Big) - \log \hat{\sigma}^2_{b|a}, \end{aligned} \tag{A.4}$$

where $\hat{\sigma}^2_{k,b|a}$ is the MLE of $\text{var}(X_{ib}|Y_i = k, X_{i(a)})$, $\hat{\sigma}^2_{b|a}$ is the MLE of $\text{var}(X_{ib}|X_{i(a)})$, and the second equation is true if and only if $\hat{\sigma}^2_{k,b|a} = \hat{\sigma}^2_{b|a}$ for each of the $k \in \{1, \cdots, K\}$.

Denote $\tilde{X}_{i(\mathcal{S})} = \Big(1, X^\top_{ia}\Big)^\top$, we can obtain that $\hat{\sigma}^2_{k,b|a} = 1/n_k \sum_{i=1}^{n} I(Y_i = k)(X_{ib} - \hat{\mu}_k - \hat{B}_k X_{i(a)})^2$ and $\hat{\sigma}^2_{b|a} = 1/n \sum_{i=1}^{n}(X_{ib} - \hat{\mu} - \hat{B} X_{i(a)})^2$, where

$$(\hat{\mu}_k, \hat{B}_k) = (\sum_{i=1}^{n} I(Y_i = k) X_{ib} \tilde{X}^\top_{i(a)})(\sum_{i=1}^{n} I(Y_i = k) \tilde{X}_{i(a)} \tilde{X}^\top_{i(a)})^{-1},$$

and $(\hat{\mu}, \hat{B}) = (\sum_{i=1}^{n} X_{ib}\tilde{X}_{i(a)}^{\top})(\sum_{i=1}^{n} \tilde{X}_{i(a)}\tilde{X}_{i(a)}^{\top})^{-1}$. Then we have,

$$\hat{\sigma}_{b|a}^2 = \sum_{k=1}^{K} \hat{\pi}_k \Big\{ \hat{\sigma}_{k,b|a}^2 + 1/n_k \sum_{i=1}^{n} I(Y_i = k)(\hat{\mu}_k - \hat{\mu} + \hat{B}_k X_{i(a)} - \hat{B} X_{i(a)})^2 \Big\}. \tag{A.5}$$

By (A.4) and (A.5), we have

$$\log\left(\sum_{k=1}^{K} \hat{\pi}_k \hat{\sigma}_{k,b|a}^2 \right) - \log \hat{\sigma}_{b|a}^2$$

$$= -\log\left\{ 1 + \frac{1/n \sum_{k=1}^{K} \sum_{i=1}^{n} I(Y_i = k)(\hat{\mu}_k - \hat{\mu} + \hat{B}_k X_{i(a)} - \hat{B} X_{i(a)})^2}{\sum_{k=1}^{K} \hat{\pi}_k \hat{\sigma}_{k,b|a}^2} \right\} \triangleq C_{b|a}.$$

If $\mathcal{S}_T \nsubseteq \mathcal{S}_2$ and $\mathcal{S}_T \subseteq \mathcal{S}_1$, we have $C_{b|a} > 0$, and $C_{b|a}$ is constant given $\mathcal{S}_1$ and $\mathcal{S}_2$. We then have $-2n^{-1}\ell(\hat{\theta}_{(\mathcal{S}_1)}) + 2n^{-1}\ell(\hat{\theta}_{(\mathcal{S}_2)}) < -C_{b|a}$. Similar conclusion can be proved under $\mathcal{S}_1 \subseteq \mathcal{S}_T$. This completes the entire proof of the Lemma 1.

Appendix B. Proof of Theorem 1

Considering the t-th step of the backward algorithm, we assume that $\mathcal{S}_T \subsetneq \mathcal{S}_{(t-1)}$. Let $j_{d_1}, j_{d_2} \in \mathcal{S}_{(t-1)}$, and $j_{d_1} \in \mathcal{S}_T, j_{d_2} \in \mathcal{S}_T^c$. We denote $\mathcal{S}_{d_1} = \mathcal{S}_{(t-1)} \backslash \{j_{d_1}\}$, and $\mathcal{S}_{d_2} = \mathcal{S}_{(t-1)} \backslash \{j_{d_2}\}$. Then, by (2.7) and lemma 1, we obtain

$$n^{-1}\Big\{ \mathrm{BIC}(\mathcal{S}_{d_1}) - \mathrm{BIC}(\mathcal{S}_{(t-1)}) \Big\} > C_{\mathcal{S}_{(t-1)}, \mathcal{S}_{d_1}} - \Big\{ |\mathcal{S}_{(t-1)}|(K-1) + K \Big\} \times \frac{\log n}{n}, \tag{A.6}$$

$$n^{-1}\Big\{ \mathrm{BIC}(\mathcal{S}_{d_2}) - \mathrm{BIC}(\mathcal{S}_{(t-1)}) \Big\} = O_p(n^{-1}) - \Big\{ |\mathcal{S}_{(t-1)}|(K-1) + K \Big\} \times \frac{\log n}{n}. \tag{A.7}$$

By combining (A.6) and (A.7), we can verify that $P\Big\{ \mathrm{BIC}(\mathcal{S}_{d_1}) > \mathrm{BIC}(\mathcal{S}_{d_2}) \Big\} \to 1$ as $n \to \infty$. For the element $d_{(t)}$ which will be eliminated in this step, we have

$$P(d_{(t)} \in \mathcal{S}_T^c) = P\Big\{ \mathrm{BIC}(\mathcal{S}_{d_1}) > \mathrm{BIC}(\mathcal{S}_{d_2}) \Big\}^{(|\mathcal{S}_{(t-1)}| - |\mathcal{S}_T|)|\mathcal{S}_T|} \to 1 \text{ as } n \to \infty.$$

Then, we have $P(\mathcal{S}_T \in \mathcal{M}) = \prod_{t=1}^{p-|\mathcal{S}_T|} P(d_{(t)} \in \mathcal{S}_T^c) \to 1$ as $n \to \infty$. If $\mathcal{S}_{(p-|\mathcal{S}_T|)} = \mathcal{S}_T$, (A.6) also holds when $p - |\mathcal{S}_T| + 1 \leq t \leq p$. Also, in conjunction with (A.7), we have $P\Big\{\text{BIC}(\mathcal{S}_T) < \text{BIC}(\mathcal{S}_{(t)})\Big\} \to 1$ as $n \to \infty$, with $0 \leq t < p - |\mathcal{S}_T|$ and $p - |\mathcal{S}_T| < t \leq p$. 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, NY.

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," *Journal of the American Statistical Association*, 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," *Statistical Science*, 21, 115.

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

Kass, R. E. and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Kass, R. E. and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypothese and Its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 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," *Journal of the American Statistical Association*, 101, 168-178.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

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

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

Smith, A. F. M. and Spiegelhalter, D. J. (1980), "Bayes Factors and Choice Criteria for Linear Models," *Journal of Royal Statistical Society, Series 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," *Statistical Science*, 18, 104–117.