Journal of the Korean
Statistical Society
Vol. 26, No. 4, 1997

A Predictive Two-Group Multinormal Classification Rule Accounting for Model Uncertainty †

Hea-Jung Kim [1]

ABSTRACT

A new predictive classification rule for assigning future cases into one of two multivariate normal populations(with unknown normal-mixture model) is considered. The development involves calculation of posterior probability of each possible normal-mixture model via a default Bayesian test criterion, called intrinsic Bayes factor, and suggests predictive distribution for future cases to be classified that accounts for model uncertainty by weighting the effect of each model by its posterior probability. In this paper, our interest is focused on constructing the classification rule that takes care of uncertainty about the types of covariance matrices(homogeneity/ heterogeneity) involved in the model. For the constructed rule, a Monte Carlo simulation study demonstrates routine application and notes benefits over traditional predictive classification rule by Geisser(1982).

Key Words : Intrinsic Bayes factor; Predictive classification; Model uncertainty; Normal-mixture model; Posterior probability; Monte Carlo simulation.

†The author wishes to acknowledge the financial support of the Korea Research Foundation made in program year of 1997.

[1]Department of Statistics, Dongguk University, Seoul 100-715, Korea.

1. INTRODUCTION

Many practical situations require the assignment of individual elements of unknown origin to one or more populations on the basis of the values of several characteristics. The objective of classification analysis is to construct a well-defined rule using available data which can be used for assigning new objects. Widespread prevalence of the classification problem in many fields has seen the development of a plethora of new approaches for classification analysis. See, for example, Anderson(1984) and Rencher(1995) for classical approach, and Geisser(1982) and Lavin and West(1992) for Bayesian approach.

Among them, two major approaches, namely estimative and predictive methods are well accepted and commonly used. Practical differences of them are illustrated by Aitchison and Donsmore(1975). Aitchison, et al.(1977) compared the two methods, and then advocated the use of predictive methods when the population distribution can be transformed to multinormality. They also suggested the use of heteroscedastic predictive method when there is a high possibility that the covariance matrices may differ appreciably across the populations. Therefore, it is common practice that, before getting into multinormal predictive classification, the first step is to conduct test of homogeneity in the covariance matrices to get a single model; mixed-normal model with homogeneous(or heterogeneous) covariances. Then, based on the test result, we usually proceed to get classification rule conditionally on the selected model.

However, as stated in Kass and Raftery(1995), any statistical analysis that selects a single model and then makes inference conditionally on that model fails to take into account fully of uncertainty involved in model selection so that it may well underestimate the uncertainty associated with quantities of interest. Classification analysis also bears this problem. Thus it is our view that we need a new approach more attuned to the investigator's query: To what extent does each model considered has probability of being fitted model. The query can be answered, at least in principle, if one adopts a Bayesian approach and calculates the posterior probabilities of the competing models, which follow directly from a default Bayes factor (named intrinsic Bayes factor). A composite inference can then be made that takes account of model uncertainty in a simple and formally justifiable way. In Section 2, we review this approach which makes use of the idea of the intrinsic Bayes factor introduced by Berger and Pericchi(1996), and in Section 3 we derive a predictive classification rule which accounts for the possibilities of homogeneous/heterogeneous covariance matrices across two multivariate

normal populations. Section 4 examines the performance of the suggested rule and notes some merits over traditional predictive classification rule by Geisser(1982). Finally, Section 5 includes some concluding remarks.

2. PREDICTIVE APPROACH ACCOUNTING FOR MODEL UNCERTAINTY

Suppose we have data D(say, training sample), assumed to have arisen under one of several alternative models $M_1, \ldots, M_J$ having probability densities $p(D|\theta_i, M_i)$, under $M_i, i = 1, \ldots, J$, where parameter vectors are unknown and are of dimension k_i. Given a prior distribution $\pi(\theta_i|M_i)$ for the parameter of each model, together with prior probability p_i of each model being true, the data produce the posterior probability of M_i being true as

$$p(M_i|D) = \frac{p(D|M_i)p_i}{\sum_{j=1}^{J} p(D|M_j)p_j}, \quad i = 1, \ldots, J, \tag{2.1}$$

where the densities $p(D|M_i)$ are obtained by integrating over the parameter space, so that

$$p(D|M_i) = \int p(D|\theta_i, M_i)\pi(\theta_i|M_i)d\theta_i, \tag{2.2}$$

The above equation is called the marginal or predictive density of D under M_i. The Bayes factor(cf. Jeffreys, 1961) for M_i against $M_{i'}$ is defined by

$$B_{ii'} = \frac{p(D|M_i)}{p(D|M_{i'})} = \frac{\int p(D|\theta_i, M_i)\pi(\theta_i|M_i)d\theta_i}{\int p(D|\theta_{i'}, M_{i'})\pi(\theta_{i'}|M_{i'})d\theta_{i'}}, \quad i \neq i'. \tag{2.3}$$

The Bayes factor denotes the ratio of the posterior odds of M_i to its prior odds, regardless of the value of the prior odds. Thus $B_{ii'}$ can be viewed as the weighted likelihood ratio of M_i to $M_{i'}$ and hence can be solely in terms of comparative support of the data for the two models(cf. Kass and Raftery, 1995). The posterior probability (2.1) that M_i is true is then expressed in terms of the Bayes factors:

$$p(M_i|D) = \left(\sum_{i'=1}^{J} \frac{p_{i'}}{p_i} B_{i'i} \right)^{-1}, \quad i = 1, \ldots, J, \tag{2.4}$$

where $B_{i'i} = 1/B_{ii'}$.

The posterior model probabilities given by (2.4) lead to solutions of prediction that take account of model uncertainty. The ultimate goal of classification analysis is prediction of a new observation Z, under the model M_i and for given data D, has density $g(Z|D, \theta_i, M_i)$. Then the predictive density of Z, given D is

$$g(Z|D) = \sum_{i=1}^{J} f(Z|D, M_i)p(M_i|D), \tag{2.5}$$

where $f(Z|D, M_i) = \int g(Z|D, \theta_i, M_i)p(\theta_i|D, M_i)d\theta_i$, and $p(\theta_i|D, M_i)$ is posterior density of θ_i given D and model M_i. This accounts for the uncertainty about a true model by weighting the conditional predictive densities according to the posterior probabilities of the alternative models.

Computing $B_{ii'}$ in equation (2.3) requires specification of the priors, $\pi(\theta_i|M_i)$ and $\pi(\theta_{i'}|M_{i'})$. Often in Bayesian analysis, especially in model selection, one can use noninformative priors such as the uniform prior, the Jeffreys(1961) prior, and the reference prior by Berger and Bernardo(1992). It is well known that the difficulty with using the noninformative priors is that the priors are typically improper and hence are defined up to arbitrary constants c's. Hence the Bayes factor $B_{ii'}$ will be defined up to $c_i/c_{i'}$, which is itself arbitrary. A common solution to this problem is to construct a default Bayes factor which uses part of the data as a sub-training sample to eliminate the constant. Formal developments of the idea can be found in work of Gelfand, Dey, and Chang(1992) and Berger and Pericchi(1996). If we let $D(\ell)$ and $\pi^N(\theta_i|M_i)$ be the part of data(sub-training sample) and improper priors to be so used, respectively. The idea is that $D(\ell)$ will be used to convert the $\pi^N(\theta_i|M_i)$ to proper posterior distributions

$$\pi^N(\theta_i|D(\ell), M_i) \propto p(D(\ell)|\theta_i, M_i)\pi^N(\theta_i|M_i). \tag{2.6}$$

The idea is to then compute the Bayes factors with the remainder of the data, using the $\pi^N(\theta_i|D(\ell), M_i)$ as priors. Denoting the remaining data by $D(-\ell)$, the Bayes factor $B_{ii'}(D(\ell))$ so obtained can be expressed as

$$\begin{aligned} B_{ii'}(D(\ell)) &= \frac{\int p(D(-\ell)|\theta_i, D(\ell), M_i)\pi^N(\theta_i|D(\ell), M_i)d\theta_i}{\int p(D(-\ell)|\theta_{i'}, D(\ell), M_{i'})\pi^N(\theta_{i'}|D(\ell), M_{i'})d\theta_{i'}} \\ &= B^N_{ii'}B^N_{i'i}(D(\ell)), \end{aligned} \tag{2.7}$$

where

$$B^N_{i'i}(D(\ell)) = \frac{\int p(D(\ell)|\theta_{i'}, M_{i'})\pi^N(\theta_{i'}|M_{i'})d\theta_{i'}}{\int p(D(\ell)|\theta_i, M_i)\pi^N(\theta_i|M_i)d\theta_i},$$

and $B^N_{ii'}$ denotes the Bayes factor (2.3) obtained by the noninformative priors $\pi^N(\theta_i|M_i)$ and $\pi^N(\theta_{i'}|M_{i'})$.

Clearly, (2.7) removes the arbitrariness in the choice of constant multiples of the improper priors; the arbitrary ratio $(c_i/c_{i'})$ that multiplies $B^N_{ii'}$ would be cancelled by the ratio $(c_{i'}/c_i)$ that would multiply $B^N_{i'i}(D(\ell))$, only if it is finite. Using the Bayes factor (2.7), Berger and Pericchi(1996) introduced the default Bayes factor which is called intrinsic Bayes factor under the following definitions.

Definition 1. The training sample, $D(\ell)$, is said to be proper if

$$0 < \int p(D(\ell)|\theta_i, M_i)\pi^N(\theta_i|M_i)d\theta_i < \infty$$

for all M_i and minimal if it is proper and no subset is proper.

Definition 2. Let $\mathbf{D}_T = \{D(1), D(2), \ldots, D(L)\}$ denote the set of all minimal training samples, $D(\ell)$. The averages of the $B_{ii'}(D(\ell))$ over all $D(\ell) \in \mathbf{D}_T$ defined by

$$B^{AI}_{ii'} = \frac{1}{L}\sum_{\ell=1}^{L} B_{ii'}(D(\ell)), \text{ and } B^{GI}_{ii'} = \left(\prod_{\ell=1}^{L} B_{ii'}(D(\ell))\right)^{1/L}$$

are called intrinsic Bayes factors(IBF). Specifically, the former is called the arithmetic IBF and the latter is called the geometric IBF. These are commonly denoted as $B^I_{ii'}$.

The IBF eliminates the instability due to dependence of $B^N_{i'i}(D(\ell))$ on the choice of the minimal training sample. See Berger and Pericchi (1996) for the properties of the IBF. Several variants of the IBF are also suggested by Berger and Pericchi (1996). These (including the IBF) are applicable for general situations (nested, nonnested, and even irregular problems) and they are shown to be corresponding to actual Bayes factors, at least asymptotically.

3. POSTERIOR PROBABILITY FOR CLASSIFICATION MODEL

Suppose we have two populations Π_1 and Π_2 each specified by a classification model M_i, $i = 1, 2$, where M_i defines the distribution of each population distribution $\Pi_k \sim N_p(\mu_k, \Sigma_k)$, $k = 1, 2$, where parameters are unknown. Let our interest of model comparison be homogeneity(or heterogeneity) of the

covariance matrices between two multivariate normal populations, so that model specification may be

$$M_1 : \Sigma_1 = \Sigma_2 (= \Sigma); \quad M_2 : \Sigma_1 \neq \Sigma_2. \tag{3.1}$$

Let $X_1(k), X_2(k), \ldots, X_{N_k}(k)$ denote independent p variate sample of size N_k from Π_k with distribution $N_p(\mu_k, \Sigma_k)$, $k = 1, 2$, and let denote the two samples as D(training sample). Then if we define

$$\bar{X}(k) = \sum_{j=1}^{N_k} X_j(k)/N_k, \; V_k = \sum_{j=1}^{N_k} (X_j(k) - \bar{X}(k))(X_j(k) - \bar{X}(k))',$$

and $S_k = V_k/(N_k - 1)$, under the model M_2, the joint probability density of D is given by

$$p(D|\mu_1, \mu_2, \Sigma_1, \Sigma_2, M_i) = \prod_{k=1}^{2} (2\pi)^{-N_k p/2} |\Sigma_k|^{-N_k/2} \exp\{-\frac{1}{2} tr[\Sigma_k^{-1} C_k]\}, \tag{3.2}$$

where $C_k = (N_k - 1)S_k + N_k(\mu_k - \bar{X}(k))(\mu_k - \bar{X}(k))'$. Setting $\Sigma_1 = \Sigma_2 = \Sigma$ in equation (3.2), we get the joint probability density conditionally on M_1.

It is to be noted that in classification applications our interest focuses primarily on a statement concerning the relative probability that an observation belongs to one or another of the population, and not about of making probability statement about where a parameter lies. Thus we shall use a particular convenient prior density to reflect an initial diffuseness or vagueness about the unknown parameters. Since the arbitrary constants can be removed by the default Bayes factor (2.7), without loss of generality, we can consider respective standard vague prior densities(Jeffreys diffuse priors) of M_1 and M_2 with constant multiples c_1 and c_2:

$$\pi^N(\mu_1, \mu_2, \Sigma | M_1) = c_1 |\Sigma|^{-(p+1)/2}$$

and

$$\pi^N(\mu_1, \mu_2, \Sigma_1, \Sigma_2 | M_2) = c_2 \prod_{k=1}^{2} |\Sigma_k|^{-(p+1)/2}. \tag{3.3}$$

Lemma 1. Under the noninformative priors (3.3), respective marginal densities under M_1 and M_2 are

$$p^N(D|M_1) = c_1 (2\pi)^{-(N^*-2)p/2} (N_1 N_2)^{-p/2} \Delta(1) |(N^* - 2)S_p|^{-\frac{(N^*-2)}{2}}, \tag{3.4}$$

$$p^N(D|M_2) = c_2 \prod_{k=1}^{2} (2\pi)^{-(N_k-1)p/2} N_k^{-p/2} \Delta(2) |(N_k - 1)S_k|^{-(N_k-1)/2}, \tag{3.5}$$

where $\Delta(1) = 2^{p(N^*-2)/2}\pi^{p(p-1)/4}\prod_{j=1}^{p}\Gamma\{(N^*-1-j)/2\}$, $N^* = N_1 + N_2$ and $\Delta(2) = 2^{p(N_k-1)/2}\pi^{p(p-1)/4}\prod_{j=1}^{p}\Gamma\{(N_k-j)/2\}$.

Proof. From the definition of the marginal density (2.2) and the likelihood (3.2),

$$\begin{aligned} p^N(D|M_1) &= \int p(D|\mu_1,\mu_2,\Sigma,M_1)\pi^N(\mu_1,\mu_2,\Sigma|M_1)\prod_{k=1}^{2}d\mu_k d\Sigma \qquad (3.6)\\ &= c_1\int(2\pi)^{-N^*p/2}|\Sigma|^{-(N^*+p+1)/2}\exp\{-\frac{1}{2}tr\Sigma^{-1}[(N^*-2)S_p \\ &+ \sum_{k=1}^{2}N_k(\mu_k-\bar{X}(k))(\mu_k-\bar{X}(k))']\}\prod_{k=1}^{2}d\mu_k d\Sigma, \end{aligned}$$

where $N^* = N_1 + N_2$ and $S_p = (V_1 + V_2)/(N^* - 2)$ is the pooled sample covariance matrix.

We obtain the marginal density $p^N(D|M_1)$ in the following way. Integrate (3.6) with respect to μ_k's using multivariate normal distribution. This gives

$$c_1(2\pi)^{-(N^*-2)p/2}(N_1N_2)^{-p/2}|\Sigma|^{-(N^*+p-1)/2}\exp\left\{-\frac{1}{2}tr[\Sigma^{-1}(N^*-2)S_p]\right\}, \qquad (3.7)$$

and hence the desired marginal density (3.4) can be found by integrating with respect to Σ, using the inverted Wishart normalizing constant. Similar integrations with respect to μ_k and Σ_k, $k = 1, 2$, for the expression below yield (3.5).

$$\begin{aligned} p^N(D|M_2) &= \int p(D|\mu_1,\mu_2,\Sigma_1,\Sigma_2,M_2)\pi^N(\mu_1,\mu_2,\Sigma_1,\Sigma_2|M_2)\prod_{k=1}^{2}d\mu_k\prod_{k=1}^{2}d\Sigma_k \\ &= c_2\int\prod_{k=1}^{2}(2\pi)^{-N_kp/2}|\Sigma_k|^{-(N_k+p+1)/2}\exp\{-\frac{1}{2}tr\Sigma_k^{-1}[(N_k-1)S_k \\ &+ N_k(\mu_k-\bar{X}(k))(\mu_k-\bar{X}(k))']\}\prod_{k=1}^{2}d\mu_k\prod_{k=1}^{2}d\Sigma_k. \qquad (3.8) \end{aligned}$$

The above marginal densities obtained from the diffuse priors give the following theorems.

Lemma 2. A sub-training sample, $D(\ell)$, consisting of $p+1$ distinct observations from each of population$(\Pi_k, k = 1, 2)$, is proper and minimal for the marginal densities of M_1 and M_2.

Proof. For the derivation of the marginal densities (3.4) and (3.5), the sample covariance matrix $S_k = V_k/(N_k - 1)$ of k-th multivariate normal population is

involved (see, equations (3.7) and (3.8)). Therefore, it is necessary to secure $|S_k| \neq 0$, and $|S_p| \neq 0$ for the derivation, and hence $N_k \geq p+1$ for $k = 1, 2$. If this condition is met, we can see that $0 < p^N(D|M_i) < \infty$ for both M_1 and M_2.

Using the sub-training sample of size $p+1$ from each population, we calculate the default Bayes factor in (2.7) which removes the arbitrariness in the choice of constant multiples of the improper priors (3.3).

Theorem 1. Let underlying classification model be M_1 as defined in (3.1), and let alternative model to be compared be M_2. Then the default Bayes factor (2.7) for M_1 against M_2 based upon a minimal training sample $D(\ell)$ and the improper prior (3.3) is given by

$$\begin{aligned} B_{12}(D(\ell)) &= B_{12}^N B_{21}^N(D(\ell)) \qquad (3.9)\\ &= \frac{\prod_{k=1}^2 |(N_k-1)S_k|^{(N_k-1)/2}}{|(N^*-2)S_p|^{(N^*-2)/2}} \times \frac{|(2p)S_p^*|^p}{\prod_{k=1}^2 |pS_k^*|^{p/2}} \\ &\times \frac{\prod_{j=1}^p \Gamma\{(N^*-1-j)/2\}}{\prod_{k=1}^2 \prod_{j=1}^p \Gamma\{(N_k-j)/2\}} \times \frac{\prod_{k=1}^2 \prod_{j=1}^p \Gamma\{(p-j+1)/2\}}{\prod_{j=1}^p \Gamma\{(2p-j+1)/2\}}, \end{aligned}$$

where S_k^* and S_p^* denote respective quantities of the unbiased sample covariance and the pooled sample covariance matrix obtained from a minimal training sample $D(\ell)$ of size $2(p+1)$ consisting of two sets of $p+1$ observations obtained from each population Π_k, $k = 1, 2$.

Proof. From (2.7), we see that

$$\begin{aligned} B_{12}(D(\ell)) &= B_{12}^N B_{21}^N(D(\ell)), \\ &= \frac{p^N(D|M_1)}{p^N(D|M_2)} \frac{p^N(D(\ell)|M_2)}{p^N(D(\ell)|M_1)}. \qquad (3.10) \end{aligned}$$

Lemma 2 implies that $p^N(D(\ell)|M_i)$ is the marginal likelihood of M_i, $i = 1, 2$, obtained solely from the sub-training sample, $D(\ell)$ of size $2(p+1)$ consisting of two sets of $p+1$ observations observed from each population Π_k, $k = 1, 2$. Modifying the marginal densities in Lemma 1 accordingly for $p^N(D(\ell)|M_1)$ and $p^N(D(\ell)|M_2)$, we see that Lemma 1 evaluates (3.10) in the expression of (3.9).

The intrinsic Bayes factor(IBF) obtained from $B_{12}(D(\ell))$ (denoting B_{12}^I) can be an alternative to the Bayes factor using the natural conjugate priors. The former has merit that, when sample size is large it circumvents the problem of estimating the hyperparameters involved in the latter.

4. PREDICTIVE CLASSIFICATION RULE

As asserted in Berger and Pericchi (1996, p.120), B_{12}^I can be used for the model selection criterion, and let the model M_1 be selected as the best supported by the data. Then, regardless of the value of the B_{12}^I (big or almost equal to 1), it has been common practice to construct predictive classification rule conditionally on M_1. However, in our view, rather than constructing classification rule under M_1, it is natural to construct a classification rule which not only takes good care of the situation but also keeps all models in the analysis, accounting for model uncertainty by weighting the effect of each model. The effect of each model, $M_i, i = 1, 2$, can be weighted by its posterior:

$$p(M_i|D) = \left(\sum_{i'=1}^{2} \frac{p_{i'}}{p_i} B_{i'i}^I \right)^{-1}, \quad i = 1, 2, \tag{4.1}$$

where p_i is the prior probability of the model M_i. If p_i is known, the posterior probability of M_i can be obtained from (4.1). If p_i is unknown default choices of p_i will often be used. In many situations the obvious choice is $p_i = 1/2$(see, Berger and Pericchi(1996), for other default choices).

Lemma 3. Suppose there exist underlying models M_1 and M_2 in (3.1), for the two populations classification, and suppose the improper prior densities (3.3) are assumed for respective model parameters. Then, accounting for the model uncertainty, we obtain predictive density of a new observation Z to be assigned to one of the two populations, Π_k, $k = 1, 2$, given by

$$g(Z|D, \Pi_k) = \sum_{i=1}^{2} f(Z|D, \Pi_k, M_i) p(M_i|D), \tag{4.2}$$

where

$$f(Z|D, \Pi_k, M_1) = St_p\{N^* - p - 1, \bar{X}(k), (1 + 1/N_k)/(N^* - p - 1)(V_1 + V_2)\}$$

and

$$f(Z|D, \Pi_k, M_2) = St_p\{N_k - p, \bar{X}(k), (1 + 1/N_k)/(N_k - p)V_k\},$$

for $k = 1, 2$. Here $St_p\{a, b, c\}$ denotes a p-dimensional variate t density function with location parameter b, scale parameter c, and degrees of freedom a, defined on R^p by the density at Z (cf. Press, 1982).

Proof. Given the posterior probabilities $p(M_i|D)$, $i = 1, 2$, the uncertainty of each model can be accounted for by weighting the effect of each model

by its posterior probability. This leads to the unconditional(unconditional on M_i) predictive density of Z. See Geisser(1982) for the derivation of the conditional predictive densities $f(Z|D, \Pi_k, M_i)$, $i = 1, 2$.

Suppose we use the unconditional predictive density $g(Z|D, \Pi_k)$, $k = 1, 2$, for the classification, and suppose a prior probability of belonging to k-th population Π_k is π_k, $k = 1, 2$, and $\sum_{k=1}^{2} \pi_k = 1$, the risk incurred in classifying an individual with measurement vector Z as $\Pi_{\hat{k}}$ is

$$R(\Pi_{\hat{k}}|Z, D) = \frac{\sum_{k=1}^{2} L(\Pi_k, \Pi_{\hat{k}}) g(Z|D, \Pi_k)\pi_k}{\sum_{k=1}^{2} g(Z|D, \Pi_k)\pi_k}, \tag{4.3}$$

where $L(\Pi_k, \Pi_{\hat{k}})$, $k = 1, 2$, is the cost or loss associated with the classification error. Let assume the special but commonly used loss function;

$$L(\Pi_k, \Pi_{\hat{k}}) = 1 - \delta(\Pi_k, \Pi_{\hat{k}}), \tag{4.4}$$

where $\delta(\Pi_k, \Pi_{\hat{k}}) = 1$ if $\Pi_k = \Pi_{\hat{k}}$, otherwise it is zero. Then we have the following theorem.

Theorem 2. A Bayes rule for the predictive classification, accounting for the model uncertainty about M_1 and M_2, is to classify a new observation Z into Π_1 if Z belongs to the classification region R_1.

$$R_1 : \pi_1 g(Z|D, \Pi_1) > \pi_2 g(Z|D, \Pi_2). \tag{4.5}$$

Otherwise classify it into Π_2.

Proof. The classification risk incurred in classifying an object with measurement vector Z as $\hat{k}$ is (4.3). This can be minimized by choosing $\hat{k}$ that minimizes the numerator in (4.3). Minimizing the numerator, we have the rule which chooses $\hat{k}$ such that

$$g(Z|D, \Pi_{\hat{k}})\pi_{\hat{k}} = \text{Max } g(Z|D, \Pi_k)\pi_k, \; k = 1, 2.$$

This leads to the classification region R_1.

The rule resulting from choosing $\hat{k}$ to minimize $R(\Pi_{\hat{k}}|Z, D)$ in (4.3) is known as the Bayes rule, and it achieves minimal classification risk among all possible rules based on the unconditional predictive density. The mixing proportion(the prior probability of Π_k, π_k, can be found in various ways. Besides the so-called little knowledge estimate which set $\pi_1 = \pi_2 = 1/2$, a decision theoretic estimate by Anderson(1984), Bayesian estimate(cf. Lavin

and West, 1992), and an information theoretic estimate by Kim(1995, 1996) are available.

5. NUMERICAL RESULTS

The goal of this section is to study the effectiveness of the suggested predictive classification rule (denoted by NEW) in Theorem 2 and to identify some situations where one would (and would not) expect improvement with NEW. The performance of NEW is compared with the classical predictive classification rule by Geisser(1982) which does not account for the model uncertainty. The comparison between the two rules is conducted in terms of correct classification rate (1- Error rate) estimates evaluated by the validation sample method (cf. Rencher, 1995).

We used computer simulation to calculate the desired correct classification rates of the two rules, Pr(NEW) and Pr(Geisser), under the default prior values $p_1 = p_2 = 1/2$ and $\pi_1 = \pi_2 = 1/2$. Our SAS/IML program generated couple of samples (training sample(D_i) and validation sample(V_i)) from each population Π_i, $i = 1, 2$, formed NEW and Geisser's predictive classification rules with given training samples D_i of each size N_i, $i = 1, 2$, so that $D = \{D_1, D_2\}$. Then the validation samples V_i of each size N_i are used to evaluate correct classification rates of the two rules. The correct classification rate is determined by the proportion of correct classified in the validation samples. See Rencher(1995, p.337) for the merits of the validation sample method.

In constructing NEW, the unconditional predictive density in (4.2) needs the posterior probabilities of M_i, $i = 1, 2$, that are calculated from (4.1). The B_{12}^I and the posterior probabilities in (4.1) are obtained in the following scheme:

Step 1. Set up a set of all possible minimal training samples (equivalently sub-training samples) $\{D(\ell)\}$), $\ell = 1, \ldots, L$; $L = {}_{N_1}C_{p+1} \times {}_{N_2}C_{p+1}$, where $D(\ell)$ denotes $2(p+1)$ observations selected from the training samples, D_1 and D_2, consisting of $p+1$ observations from D_1 and $p+1$ from D_2.

Step 2. For each set of the minimal training sample $D(\ell)$ and the remaining sample $D(-\ell)$, calculate $B_{12}(D(\ell))$ in (16) to obtain B_{12}^I using the geometric IBF (cf. Definition 2):

$$B_{12}^I = \left(\prod_{\ell=1}^{L} B_{12}(D(\ell)) \right)^{1/L} \quad \text{and} \quad B_{21}^I = 1/B_{12}^I.$$

In our simulation, we use 100 randomly selected minimal training samples ($L = 100$). These are sufficient for controlling the standard error associated with the average of $B_{12}(D(\ell))$'s.

Step 3. Calculate $p(M_i|D)$, $i = 1, 2$, from (4.1).

The reason for using the geometric IBF for obtaining B_{12}^I is that it has the nice property of symmetry, i.e. $B_{12}^I = 1/B_{21}^I$, while the arithmetic IBF does not have the property (cf. Berger and Pericchi 1996). To highlight the property of NEW that accounts for the model uncertainty in (3.1), without loss of generality, we consider the following case for each set of multivariate normal population density parameters:

(i) Population 1 density: $\phi_p(\mu_1, \Sigma_1)$, where j-th component of μ_1 is given by $\mu_{1j} = 0$; $j = 1, \ldots, p$, and $\Sigma_1 = I_p$. Here $\phi_p(\mu_1, \Sigma_1)$ denotes the density of $N_p(\mu_1, \Sigma_1)$.

(ii) Population 2 (mixed) density:

$$\alpha\phi_p(\mu_2, \Sigma_1) + (1 - \alpha)\phi_p(\mu_2, \Sigma_2), \quad 0 < \alpha < 1,$$

where j-th component of μ_2 is given by $\mu_{2j} = \beta(-1)^{j+1}$ and $\Sigma_2 = \text{Diag}(d_{2j})$, a diagonal matrix with j-th diagonal element $d_{2j} = j/(p-1) + .5$, $j = 1, \ldots, p$.

Under the two population densities D_i and V_i were generated. Especially for D_2 and V_2, a rejection sampling technique(cf. Morgan, 1984) using uniform distribution($U(0, 1)$) is adopted for generating the mixed samples according to the mixed proportion $0 < \alpha < 1$.

Table I, summarizing the results of the simulation with $N_1 = N_2 = 30$, presents the average correct classification rates (Pr(NEW) and Pr(Geisser)) over the 200 replications for NEW and classical predictive classification rule. Also presented are the posterior probability, $Pr(M_1|D)$, of M_1 defined in (3.1) and their standard deviations in the parentheses. For a reference, the probability value of Box's M test(cf. Rencher, 1995, P.282) for testing null hypothesis $H : \Sigma_1 = \Sigma_2$ is also given in the table. The simulation results with other values of N_1 and N_2 revealed the same implications as Table 1, thus we eliminated them in the tabulation.

The table shows that, except for asterisked cases, NEW achieves slightly higher correct classification rate than Geisser's classical predictive classification rule does. Asterisked cases were mainly occurred to some cases where $\alpha = 1/6, 5/6$. This implies that as $\alpha \to 0$ or 1 NEW may not improve the correct classification rate, but at least gives a predictive classification rule which accounts for the model uncertainty. As would be expected, like Geisser's rule,

Pr(NEW) increases as the dimension of the measurement space increases, more dramatically increasing for the larger value of β(relating to the distances between two populations). Moreover, it is noted from the table that the posterior probability of M_1 obtained by (4.1), $p(M_1|D)$, is consistent with the probability value of Box's M test. The above results imply that NEW safely incorporates the model uncertainty in the classification analysis and improves the correct classification rate when uncertainty between the two models M_1 and M_2 is high.

TABLE I. Correct Classification Rates($N_1 = N_2 = 30$)

		β=1		β=2			
p	$(\alpha N_2, (1-\alpha)N_2)$	Pr(Geisser)	Pr(NEW)	Pr(Geisser)	Pr(NEW)	$Pr(M_1\|D)$	p-value
2	(5, 25)	76.183	76.017*	91.650	91.416*	.381	.008
	S.D.	(5.999)	(6.149)	(3.748)	(3.642)	(.379)	(.023)
	(10, 20)	74.183	75.017	90.633	91.117	.653	.054
	S.D.	(6.279)	(6.139)	(3.624)	(3.683)	(.359)	(.098)
	(15, 15)	73.283	73.550	90.483	90.817	.819	.143
	S.D.	(5.896)	(6.084)	(3.756)	(3.755)	(.291)	(.133)
	(20, 10)	74.233	74.267	90.700	90.783	.913	.261
	S.D.	(6.120)	(6.073)	(4.016)	(4.063)	(.191)	(.282)
	(25, 5)	74.267	74.217*	91.383	91.383	.978	.409
	S.D.	(6.331)	(6.393)	(3.964)	(3.971)	(.048)	(.323)
3	(5, 25)	83.200	84.550	96.200	96.667	.321	.007
	S.D.	(4.155)	(4.481)	(2.247)	(2.260)	(.434)	(.025)
	(10, 20)	82.183	84.550	96.200	96.667	.435	.004
	S.D.	(4.622)	(4.787)	(2.556)	(2.628)	(.303)	(.119)
	(15, 15)	81.433	81.667	95.817	95.983	.937	.122
	S.D.	(4.594)	(4.636)	(2.778)	(2.747)	(.198)	(.182)
	(20, 10)	81.016	81.133	95.850	95.883	.983	.249
	S.D.	(4.543)	(4.475)	(2.797)	(2.817)	(.096)	(.268)
	(25, 5)	80.933	80.933	95.583	95.583	.999	.386
	S.D.	(5.266)	(5.266)	(2.836)	(2.836)	(.001)	(.288)
4	(5, 25)	83.083	84.332	97.200	97.050*	.297	.003
	S.D.	(4.478)	(4.482)	(1.908)	(2.182)	(.423)	(.014)
	(10, 20)	82.167	82.500	96.617	96.717	.351	.027
	S.D.	(4.560)	(4.615)	(2.421)	(2.339)	(.106)	(.082)
	(15, 15)	81.833	81.917	96.501	96.551	.703	.086
	S.D.	(4.794)	(4.781)	(2.327)	(2.303)	(.262)	(.149)
	(20, 10)	82.283	83.367	96.717	96.821	.993	.189
	S.D.	(4.682)	(4.693)	(2.177)	(2.102)	(.006)	(.234)
	(25, 5)	82.466	82.417*	96.817	96.833	.997	.327
	S.D.	(4.764)	(4.868)	(2.298)	(2.241)	(.026)	(.278)
5	(5, 25)	86.350	86.633	98.633	98.400*	.133	.003
	S.D.	(4.567)	(4.494)	(1.578)	(1.657)	(.319)	(.016)
	(10, 20)	85.283	85.383	98.167	98.183	.276	.023
	S.D.	(4.526)	(4.500)	(1.693)	(1.716)	(.177)	(.078)
	(15, 15)	84.667	84.887	98.300	98.333	.721	.088
	S.D.	(4.507)	(4.601)	(1.724)	(.909)	(.213)	(.159)
	(20, 10)	84.916	84.916	98.216	98.216	.999	.196
	S.D.	(4.428)	(4.428)	(1.695)	(1.695)	(.000)	(.229)
	(25, 5)	84.916	84.916	98.133	98.133	1.00	.344
	S.D.	(4.351)	(4.351)	(1.985)	(1.985)	(.000)	(.283)

6. CONCLUDING REMARKS

We have considered the problem of developing a predictive classification rule that accounts for uncertainty of the normal mixture model for two population classification analysis. As an alternative to the usual predictive classification rule by Geisser which does not take care of the model uncertainty, a new predictive classification rule is proposed. The rule is designed to take care of the model uncertainty by incorporating the posterior probabilities of the models considered. The approach has at least two advantages too. When the model uncertainty is high, the suggested predictive classification rule yields better correct classification rate. Moreover, unlike the model selection criteria based on asymptotic sampling theory (such as Box's M test and AIC), the posterior probabilities of the models to be compared are not only exact, but provide a way of incorporating external information into the evaluation of evidence about a hypothesis.

The choice of a noninformative priors is reasonable when there is no prior information available for the normal mixture model parameters. Informative priors could also be used. The use of a multivariate normal prior for the mean parameter vectors and inverted Wishart for the covariance parameters would result in similar predictive classification rule as in the paper. The study pertaining to the performance the predictive classification rule obtained by the informative priors is not unimportant and left as a future study of interest.

REFERENCES

(1) Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis,* Cambridge University Press.

(2) Aitchison, J., Habbema, J. D. F., and Kay, J. W. (1977). A critical comparison of two methods of statistical discrimination, *Applied Statistics,* Vol. 26, 15-25.

(3) Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis,* John Wiley and Sons.

(4) Berger, J. O. and Bernardo, J. M. (1992). On the development of reference prior method, in *Bayesian Statistics IV,* Eds. by Bernardo, J. M. et al., Oxford University Press.

(5) Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction, *Journal of the American Statistical Association,* Vol. 91, 109-122.

(6) Geisser, S. (1982). Predictive discrimination, *Handbook of Statistics 2,* Eds. by Krishaiah, P. R. and Kanal, L. N., North Holland Publishing Co., New York.

(7) Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distribution with implementations via sampling based methods (with discussion), *Bayesian Statistics IV,* Eds. by Bernardo, J. M. et al., Oxford University Press.

(8) Jefferys, H. (1961). *Theory of Probability,* Oxford University Press.

(9) Kass, R. E. and Raftery, A. E. (1995). Bayes factors, *Journal of the American Statistical Association,* Vol. 90, 773-795.

(10) Kim, H. J. (1995). On a balanced quadratic classification rule, *Communications in Statistics: Theory and Methods,* Vol. 24, 607-623.

(11) Kim, H. J. (1996). On a constrained optimal rule for classification with unknown prior individual group membership, *Journal of Multivariate Analysis,* Vol. 59(2), 166-186.

(12) Lavin, M. and West, M. (1992). A Bayesian method for classification and discrimination, *The Canadian Journal of Statistics,* Vol.20, 451-461.

(13) Morgan, B. J. T. (1984). *Elements of Simulation,* Chapman and Hall Co., London.

(14) Press, S. J. (1982). *Applied Multivariate Analysis,* Krieger Publishing Co., Florida.

(15) Rencher, A. C. (1995). *Methods of Multivariate Analysis,* John Wiley and Sons Co., New York.