The Canadian Journal of Statistics
Vol. 27, No. 2, 1999, Pages 299–313
La revue canadienne de statistique

Prior elicitation in the classification problem

Craig A. COOLEY[†] and Steven N. MacEACHERN

The Ohio State University

Key words and phrases: Unclassified observations, mixture models, prior odds, Gibbs sampling, inconsistency, discriminant analysis.
AMS 1991 subject classifications: 62H30, 62F12, 62F15, 62H12, 62A15.

ABSTRACT

Results are developed concerning the asymptotic behaviour of the Bayes classification rule as the number of unclassified observations grows without bound. It is shown that unclassified observations serve only to estimate the individual population parameters in an unlabeled sense and do not provide information about the labels that are attached to the populations. Prior construction is approached through investigation of prior odds over regions of the joint parameter space (across all populations) deemed likely to contain the true joint parameter vector. It is shown that consideration of these prior odds can lead to more robust *a posteriori* classification of individual observations.

RÉSUMÉ

Les auteurs s'intéressent au comportement de la règle de classification bayésienne à mesure qu'augmente le nombre d'observations non classifiées. Ils démontrent que ces observations peuvent servir à estimer certains des paramètres propres aux classes, mais qu'elles ne fournissent pas d'information permettant d'identifier ces classes. Pour construire la loi a priori, les auteurs proposent d'évaluer la probabilité relative que le vecteur de paramètres caractérisant les classes appartienne à telle ou telle région de l'espace paramétrique conjoint. Ils expliquent pourquoi cette façon de procéder permet d'accroître la robustesse de la règle de classification a posteriori des observations.

1. INTRODUCTION

Consider the multigroup classification problem in which it is desired to classify an object as belonging to one of K populations or classes based upon observed values of d predictors, $\mathbf{X} \in \mathbb{R}^d$. Under 0-1 loss, the Bayes rule classifies the object into the population with largest posterior probability, traditionally calculated using parametric models for each population and a *training set* consisting of class membership and predictors for N objects. Under mild regularity conditions, as the size of the training set from each population increases without bound, the Bayes rule collapses to the optimal rule in which the parameters of the model for each population are completely known. However, it is often the case that the cost of classifying with certainty is high while the cost of data collection is low, resulting in a wealth of unclassified data and a scarcity of

[†] On 14 June 1996, after this paper had been submitted, Craig A. Cooley was killed by a hit-and-run driver. Craig was planning to defend his doctoral dissertation in August 1996 and had accepted a faculty position at Carleton College in Minnesota. Craig was an outstanding student, a researcher with a promising future, an excellent teacher, and a wonderful colleague. This work is but one portion of Craig's Ph.D. dissertation. It is respectfully dedicated to his memory.

classified data. In such a situation, it seems unwise to ignore the information contained in the unclassified data. McLachlan (1975, 1977) and O'Neill (1978) explore the utility of unclassified data in a classical setting; Lavine and West (1992) present a Bayesian classification method that incorporates information from unclassified data.

In this paper, we develop results concerning the asymptotic behaviour of the Bayes rule as the size of the unclassified sample increases without bound. Our results can be seen to be direct consequences of the asymptotic theory developed by Berk (1966) and hold under extremely mild regularity conditions. We apply the results to the task of prior elicitation, where the prior is based on asymptotic considerations.

In Section 2, we introduce notation and regularity conditions and state Berk's result. We also develop asymptotic results concerning unclassified samples and provide an example of a prior distribution that results in a lack of convergence of the Bayes rule. Section 3 provides a Bayesian classification analysis through the use of Monte Carlo Markov-chain methods, with attention given to prior elicitation. Conclusions are drawn in Section 4.

2. NOTATION AND THEORETICAL RESULTS

We adopt the following notation for use throughout this paper. Let $\{G_{\boldsymbol\eta}(\cdot), \boldsymbol\eta \in \Theta\}$ be a family of distribution functions indexed by the finite-dimensional parameter $\boldsymbol\eta$, each distribution having density $g(\cdot|\boldsymbol\eta)$ with respect to some dominating σ-finite measure $\mu(\cdot)$. Let the density of observations from the ith population, the ith *class conditional density*, be given by $g(\cdot|\boldsymbol\eta_i)$, and let the true value of $\boldsymbol\eta_i$ be denoted $\boldsymbol\eta_{0i} \in \Theta$. Let $G_{\boldsymbol\eta_i}$ denote the corresponding distribution. Denote the joint parameter vector space by $\Theta^K = \{(\boldsymbol\eta_1, \ldots, \boldsymbol\eta_K)|\boldsymbol\eta_i \in \Theta, i = 1, \ldots, K\}$, and denote the true joint parameter value by $\boldsymbol\theta_0 = (\boldsymbol\eta_{01}, \ldots, \boldsymbol\eta_{0K})$. Let $\alpha_1, \ldots, \alpha_K$, $\sum \alpha_i = 1$, be the *a priori* probabilities of group membership, so that an unclassified observation has mixture density

$$f(\mathbf{x}|\boldsymbol\theta) = \sum_{i=1}^{K} \alpha_i g(\mathbf{x}|\boldsymbol\eta_i). \tag{1}$$

Let $F_{\boldsymbol\theta_0}$ denote the true distribution function of an unclassified observation, and let $P_{\boldsymbol\theta_0}$ denote the corresponding probability measure. Let

$$H^i(\mathbf{x}|\boldsymbol\eta) = \log \frac{g(\mathbf{x}|\boldsymbol\eta)}{g(\mathbf{x}|\boldsymbol\eta_{0i})}, \qquad \bar{H}_p^i(\boldsymbol\eta) = \sum_{j=1}^{p} H^i(\mathbf{x}_j|\boldsymbol\eta)/p,$$

$$H(\mathbf{x}|\boldsymbol\theta) = \log \frac{f(\mathbf{x}|\boldsymbol\theta)}{f(\mathbf{x}|\boldsymbol\theta_0)}, \qquad \bar{H}_p(\boldsymbol\theta) = \sum_{j=1}^{p} H(\mathbf{x}_j|\boldsymbol\theta)/p,$$

$$A_0 = \{\boldsymbol\theta \,|\mathcal{E}\, H(\mathbf{X}|\boldsymbol\theta) = \mathcal{E}\, H(\mathbf{X}|\boldsymbol\theta_0)\},$$

where p is an arbitrary integer. Here and throughout, we adopt the usual notation in which $\mathbf{X}$ denotes a random variable and $\mathbf{x}$ its realization. Finally, let $\Pi(\cdot)$ denote the prior probability measure on Θ^K. We assume all classified data have been absorbed into the prior, so that unclassified observations form the totality of the data.

Berk (1966) shows that, under mild regularity conditions on the mixture likelihood $f(\cdot|\boldsymbol\theta)$ and the prior distribution $\Pi(\cdot)$, the posterior distribution for $\boldsymbol\theta$ given an unclassified sample concentrates, almost surely $F_{\boldsymbol\theta_0}$, in any open set U such that $A_0 \subset U$.

RESULT 1. (Berk). *Suppose* $f(\mathbf{x}|\boldsymbol{\theta})$ *satisfies the conditions*

(B1) $f(\mathbf{x}|\boldsymbol{\theta})$ *is jointly measurable in* $(\mathbf{x},\boldsymbol{\theta})$, *and* $f(\mathbf{x}|\boldsymbol{\theta})$ *is continuous in* $\boldsymbol{\theta}$ *at all* $\boldsymbol{\theta} \in \Theta$ $(a.s.\, F_{\boldsymbol{\theta}_0})$.

(B2) *For any* $r^* \in \mathbb{R}$, *there exists an integer p and a cocompact subset D of* Θ^K *(i.e.,* $\Theta^K - D$ *is compact) such that*

$$\mathcal{E} \sup_{\boldsymbol{\theta} \in D} \tilde{H}_p(\boldsymbol{\theta}) \leq r^*.$$

Then, for any open set $U \in \Theta^K$ *such that* $A_0 \subset U$,

$$\Pi(U|\mathbf{X}_1,\ldots,\mathbf{X}_N) \overset{N\to\infty}{\longrightarrow} 1 \qquad (a.s.\; F_{\boldsymbol{\theta}_0}),$$

where $\Pi(\cdot|\mathbf{X}_1,\ldots,\mathbf{X}_N)$ *is the posterior probability measure for* $\boldsymbol{\theta}$ *given an unclassified sample of size* N.

Note that condition (B2) differs slightly in statement from the corresponding condition found in Berk. The additional conditions stated by Berk are necessary only for the case of an incorrectly specified model. Condition (B2) can be difficult to check for mixture likelihoods. We shall show that it is sufficient to check a similar set of conditions on the individual components of the mixture.

For the moment, we assume that the *a priori* probabilities α_i, $i = 1,\ldots,K$, are known and fixed. If no two of the *a priori* probabilities are equal, and if $\sum_{i=1}^K \alpha_i G_{\boldsymbol{\eta}_{0i}}$ cannot be expressed as $\sum_{i=1}^K \alpha_i G_{\boldsymbol{\eta}_i}$ for any set $\boldsymbol{\eta}_i$, $i = 1,\ldots,K$, other than $\boldsymbol{\eta}_{01},\ldots,\boldsymbol{\eta}_{0K}$, then $f(\mathbf{x}|\boldsymbol{\theta}) = f(\mathbf{x}|\boldsymbol{\theta}_0)$ (a.s. $F_{\boldsymbol{\theta}_0}$) if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. [If $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, there is a set of positive probability for which $f(\mathbf{x}|\boldsymbol{\theta}) \neq f(\mathbf{x}|\boldsymbol{\theta}_0)$.] If $\alpha_i = 1/K$ for all $i = 1,\ldots,K$, then

$$A_0 = \Theta_0 \equiv \{(\boldsymbol{\eta}_{0\beta_1},\ldots,\boldsymbol{\eta}_{0\beta_K})|(\beta_1,\ldots,\beta_K) \text{ permutes } (1,\ldots,K)\}. \tag{2}$$

That is, the posterior distribution concentrates in the union of any $K!$ open balls surrounding the $K!$ permutations of $\boldsymbol{\theta}_0$. The intuitively reasonable implication for practical problems is that, in general, an infinite amount of unclassified data serves to pin down, in an unordered sense, the locations of the K true parameter values $\boldsymbol{\eta}_{01},\ldots,\boldsymbol{\eta}_{0K}$, but does not aid in determining to which population each of these true parameter values are attached. This kind of "labelling" knowledge is provided by the prior distribution, which might include information obtained from preclassified data.

Alternatively, suppose that the *a priori* probabilities are unknown and are regarded as parameters in the model. Then the generalized parameter space is $\Theta^K \times [0,1]^{K-1} \equiv \Omega$. Since

$$H(\mathbf{x}|\boldsymbol{\theta},\boldsymbol{\alpha}) = \log \frac{\sum_{j=1}^K \alpha_j g(\mathbf{x}|\boldsymbol{\eta}_j)}{\sum_{j=1}^K \alpha_j g(\mathbf{x}|\boldsymbol{\eta}_{0j})},$$

it is easy to see that

$$A_0 = \bigcup_{\boldsymbol{\beta} \in B} \{(\boldsymbol{\eta}_{0\beta_1},\ldots,\boldsymbol{\eta}_{0\beta_K},\alpha_{\beta_1},\ldots,\alpha_{\beta_K})\},$$

where B is the set of permutations of $(1,\ldots,K)$. Then A_0 is the asymptotic carrier set provided the prior distribution places positive mass in each of the $K!$ neighbourhoods of Ω containing the various permutations of $(\boldsymbol{\theta}_0,\boldsymbol{\alpha})$.

The following results are proved through applications of Berk's result to the mixture likelihoods (1) that arise when the data are unclassified. Throughout, it is assumed that the *a priori* probabilities of group membership are fixed and equal. To facilitate checking conditions (B1) and (B2) for a mixture density, we show that it is enough to check similar conditions on the components of the mixture. A proof is given in the Appendix.

LEMMA 1. *Suppose:*

(A1) *For each* $i = 1,\ldots,K$, $g(\mathbf{x}|\boldsymbol{\eta})$ *is jointly measurable in* $(\mathbf{x},\boldsymbol{\eta})$, *and* $g(\mathbf{x}|\boldsymbol{\eta})$ *is continuous in* $\boldsymbol{\eta}$ *at all* $\boldsymbol{\eta} \in \Theta$ *(a.s.,* $G_{\boldsymbol{\eta}_{0i}}$*).*

(A2) *For each* $i = 1,\ldots,K$, *for every* $r \in \mathbb{R}$, *there exists an integer* p_i *and a cocompact subset* D_i *of* Θ *(i.e.,* $D_i^c \equiv \Theta - D_i$ *is compact) such that*

$$\mathcal{E} \sup_{\boldsymbol{\eta}\in D_i} \bar{H}^i_{p_i}(\boldsymbol{\eta}) \leq r.$$

(A3) *There exists* M *such that*

$$\mathcal{E} \sup_{\Theta} H^i(\mathbf{X}|\boldsymbol{\eta}) < M \qquad \text{for all} \quad i = 1,\ldots,K.$$

Then conditions (B1) *and* (B2) *of Result 1 are satisfied for the mixture density* $f(\mathbf{x}|\boldsymbol{\theta})$.

Result 2 shows that, if the prior distribution on the joint parameter space Θ^K has a density that is continuous and positive at one or more points in Θ_0 [given in (2)], then the posterior mass at each $\boldsymbol{\theta}' \in \Theta_0$ converges to the prior mass at $\boldsymbol{\theta}'$ relative to the prior mass of Θ_0. In the result, $N_\epsilon(\boldsymbol{\theta})$ represents an ϵ-neighbourhood of $\boldsymbol{\theta}$. A proof is provided in the Appendix.

RESULT 2. *Assume the a priori probabilities of class membership are equal, and suppose the class conditional densities satisfy conditions* (A1) *through* (A3) *of Lemma 1. Suppose the prior probability measure* Π *on* Θ^K *is dominated by Lebesgue measure and has continuous density* $\pi(\cdot)$ *such that* $\pi(\boldsymbol{\theta}) > 0$ *for some* $\boldsymbol{\theta} \in \Theta_0$. *Then, for any* $\boldsymbol{\theta}' \in \Theta_0$ *and* $\epsilon > 0$ *such that* $\bigcap_{\Theta_0} N_\epsilon(\boldsymbol{\theta}) = \emptyset$,

$$\Pi(N_\epsilon(\boldsymbol{\theta}')|\mathbf{X}_1,\ldots,\mathbf{X}_n) \longrightarrow \frac{\pi(\boldsymbol{\theta}')}{\sum_{\Theta_0} \pi(\boldsymbol{\theta})} \qquad (a.s.\ F_{\boldsymbol{\theta}_0}).$$

If the prior places positive mass on at least one member of Θ_0, then Berk's result can be strengthened with the addition of an extra regularity condition governing the behaviour of the log likelihood ratio:

(A4) The log likelihood ratio

$$p\bar{H}_p(\boldsymbol{\theta}) = \log \prod_{i=1}^{p} \frac{f(\mathbf{x}_i|\boldsymbol{\theta})}{f(\mathbf{x}_i|\boldsymbol{\theta}_0)}$$

is uniformly bounded in probability. That is, for any $\epsilon > 0$, there exists $M > 0$ and an integer N such that

$$p > N \quad \Rightarrow \quad P_{\boldsymbol{\theta}_0}\left\{\sup_{\Theta^K} p\bar{H}_p(\boldsymbol{\theta}) > M\right\} < \epsilon.$$

Specifically, it can be shown that, under (A1) through (A4), the posterior mass assigned to Θ_0 tends to 1 (a.s. F_{θ_0}); see Lemma 2 in the Appendix. As a direct consequence of this asymptotic concentration of the posterior on a finite subset of Θ, we obtain a result for "nice" discrete priors that is analogous to Result 2. A proof of this result is contained in Cooley (1996).

RESULT 3. *Assume the a priori probabilities of class membership are equal, and suppose the class conditional densities satisfy conditions* (A1) *through* (A4). *Denote the prior probability measure by* $\Pi(\cdot)$, *and suppose* $\Pi(\Theta_0) > 0$. *Then, for every* $\boldsymbol{\theta}' \in \Theta_0$,

$$\Pi(\{\boldsymbol{\theta}'\}|\mathbf{X}_1 \ldots, \mathbf{X}_p) \longrightarrow \frac{\Pi(\{\boldsymbol{\theta}'\})}{\Pi(\Theta_0)} \qquad (a.s.\ F_{\theta_0}).$$

It is important to note that, despite Results 2 and 3, priors do exist for which the posterior probability of any one neighbourhood corresponding to a particular permutation of $\boldsymbol{\theta}_0$ never converges, but continues to fluctuate, even as $n \longrightarrow \infty$. We outline one such example below.

We consider the two-population, unidimensional problem with equal *a priori* probabilities, so that the mixture likelihood for an unclassified observation is

$$f(\mathbf{x}|\boldsymbol{\theta}) = \tfrac{1}{2}\{g(\mathbf{x}|\theta_1) + g(\mathbf{x}|\theta_2)\}.$$

We assume that this mixture satisfies conditions necessary for asymptotic normality of the MLE for $\boldsymbol{\theta}$ (see, for example, Lehmann 1983). We take $\boldsymbol{\theta}_0 \in \Theta_U^2 = \{(\theta_1, \theta_2)|\theta_1 < \theta_2\}$ (the upper half plane), and we distinguish between a point $\boldsymbol{\theta}^U$ and its reflection by writing $\boldsymbol{\theta}^U = (\theta_1, \theta_2)$ and $\boldsymbol{\theta}^L = (\theta_2, \theta_1)$. The likelihoods of the points $\boldsymbol{\theta}^U$ and $\boldsymbol{\theta}^L$ are equal.

We place a prior distribution on Θ that assigns mass to the countable set of points $\boldsymbol{\theta}_0^U + 2^{-2k}\mathbf{1}\epsilon^*$ and $\boldsymbol{\theta}_0^L + 2^{-2k+1}\mathbf{1}\epsilon^*$, $k = 1, 2, \ldots$, where $\mathbf{1} = (1, 1)^{\mathrm{T}}$. The discrete prior distribution assigns mass 2^{-2k} and 2^{-2k+1}, respectively, to each of these points. This prior distribution, in the neighbourhood of Θ_0, admits a rescaling. If the point furthest from Θ_0 is removed, the remainder of the prior follows the above description with ϵ^* replaced by $\epsilon^*/2$, and with the upper and lower half planes reversed.

For large n, the likelihood is locally approximately normal. The likelihood based on $4n$ observations is also locally approximately normal, but with half the scale. We note that the rescaling of the likelihood matches the rescaling of the prior, so that the change from a sample of size n to a sample of size $4n$ results in an effective *reversal* of the prior labeling. As a consequence, the posterior probability assigned to $N_\epsilon(\boldsymbol{\theta}_0^U)$, with ϵ small, does not converge as $n \longrightarrow \infty$. Technical details for this example appear in Cooley and MacEachern (1996).

The point of the example is to disprove the notion that the posterior distribution "stabilizes" as the sample size grows, regardless of the prior distribution. In this example, the fluctuations arise from the instability of $\Pi(\boldsymbol{\theta}^U)/\Pi(\boldsymbol{\theta}^L)$ near $\boldsymbol{\theta}_0$. The fluctuations are not purely random, but are governed by the prior distribution, $\boldsymbol{\theta}_0$ and the sample size.

This example shows that, in the case of unclassified data, some care in the choice of prior distribution on Θ^K must be exercised to avoid asymptotic instability in the classification rule. On the other hand, "standard" priors that are known to yield asymptotically stable posteriors do not necessarily reflect prior information. In Section 3, we investigate an alternative method of prior construction that is based upon the asymptotic results given above. Throughout the section, we rely on the connection between probabilities and odds, noting that in this setting asymptotic stability of the posterior odds is equivalent to

TABLE 1: Group means and standard errors for crab data.

	Mean (SE)		
Group	FL	RW	BD
Blue males	14.8 (0.453)	11.7 (0.299)	13.4 (0.453)
Blue females	13.3 (0.372)	12.1 (0.345)	11.8 (0.389)
Orange males	16.6 (0.497)	12.3 (0.311)	15.3 (0.499)
Orange females	17.6 (0.421)	14.8 (0.332)	15.6 (0.389)
Blues combined	14.05 (0.328)	11.9 (0.261)	12.6 (0.315)
Orange combined	17.1 (0.301)	13.55 (0.228)	15.45 (0.307)

asymptotic stability of the posterior probabilities. We present an analysis of a nontrivial example in which the standard elicitation methods result in implausible solutions.

3. EXAMPLE OF PRIOR CONSTRUCTION

The standard approach to prior specification in the classification problem is to take independent noninformative priors for each of the classes (Titterington *et al.* 1985). However, when the classified data contain relatively little information about one of the classes or when previously collected data hold a tenuous connection with the current unclassified observations, modeling of prior information plays an essential role. The previous section's results highlight a key feature of even well behaved prior distributions: since the posterior odds of various labelings are asymptotically equal to the prior odds of the labelings, the prior must assign plausible odds in regions of the parameter space that receive large prior probability. In this section we synthesize standard techniques of prior construction with the additional concern that the prior odds remain relatively stable where the prior assigns the bulk of its mass. We illustrate prior construction on the well-known *Leptograpsus* crab data. For a general discussion of robustness concerns in Bayesian settings, see Berger (1994).

Campbell and Mahon (1974) performed a classical linear discriminant analysis on data collected on blue and orange forms of the rock crab *Leptograpsus variegatus* in an attempt to categorize a given crab into one of the two color groups based on five carapace characteristics. Linear discriminant analysis on the five characteristics very effectively discriminates between the four gender × color groups, and so, to make discrimination more challenging, we focus only on three of the characteristics: width of carapace frontal lip (FL), carapace rear width (RW), and body depth (BD). The complete data set is currently available by anonymous ftp at `markov.stats.ox.ac.uk:pub/neural/ASI`.Ripley (1994) presents some nonlinear approaches to analysis of these data.

A total of 200 observations were collected, 50 from each of the four gender × color subgroups. To study the issue of prior construction when the data consist of unclassified observations, we alter the data set. We split the data by gender into two parts, imagining that the male crabs constitute classified data collected some time previously and that the female data are unclassified. The problem is then to classify female crabs as either blue or orange, based upon the inexpensive carapace measurements. Means and standard errors for the four groups as well as for orange and blue crabs combined across gender groups are given in Table 1. Scatterplots of the data are shown in Figure 1, where the 1's correspond to blue crabs. The plots reveal traces of heteroscedasticity; the log transform was applied to the original data with minimal effect on the results of this section. We therefore present the analysis on the original scale.

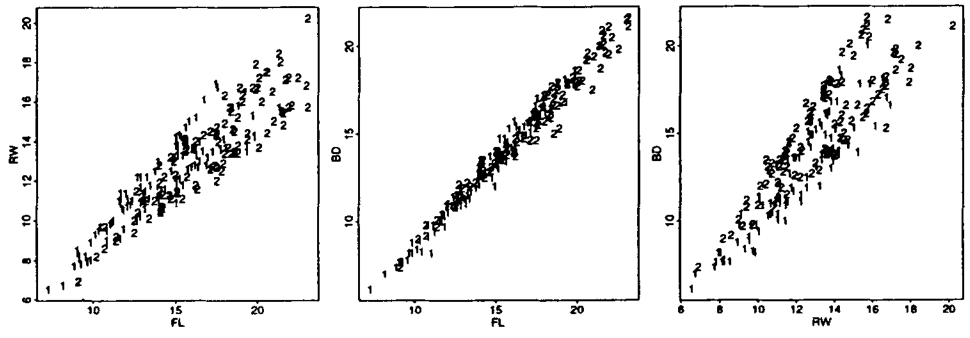

Figure 1: Scatterplots of crab data.

We adopt normal likelihoods,

$$g(\mathbf{x}|Z = i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \propto |\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\}, \qquad i = 1,2,$$

where Z indicates class membership and group 1 corresponds to the blue crabs. A plot of the data in the first two linear discriminant directions reveals a modest amount of separation.

We construct our initial prior distribution for the means and covariances of the female crabs by examining the male crab data and assuming the male crabs are similar in size and shape to the female crabs. We assume independence between the joint mean vector $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^{\mathrm{T}}, \boldsymbol{\mu}_2^{\mathrm{T}})^{\mathrm{T}}$ and the joint covariance matrix $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$.

As the prior for $\boldsymbol{\Sigma}$, we specify independent Wishart distributions for $\boldsymbol{\Sigma}_i^{-1}$,

$$\boldsymbol{\Sigma}_i^{-1} \sim \mathcal{W}\left(\tfrac{1}{3}\,\mathbf{S}_i^{-1}, 3\right), \qquad i = 1,2,$$

where $\mathbf{S}_i$ is the sample covariance matrix for the ith color group within the male crabs. Under this prior, $\mathcal{E}\,\boldsymbol{\Sigma}_i^{-1} = \mathbf{S}_i^{-1}$, and $\nu = 3$ is the smallest integral choice for the degree-of-freedom parameter that gives a proper prior density for $\boldsymbol{\Sigma}_i$. It is important to have a proper prior distribution because use of an improper prior would lead to the usual troubles when comparing models of differing dimension: With only unclassified data at our disposal, we would end up classifying all female crabs to one color group.

For the means, we begin with the assumption of total independence between the color groups and take

$$\boldsymbol{\mu} \sim \mathbf{N}\left(\begin{pmatrix}\bar{\mathbf{x}}_1 \\ \bar{\mathbf{x}}_2\end{pmatrix}, \boldsymbol{\Sigma}^{(a)} \equiv \frac{1}{k}\begin{bmatrix}\mathbf{S}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_2\end{bmatrix}\right), \tag{3}$$

where $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ are the sample means from blue and orange male crabs, respectively. We take $k = 3$, claiming an equal amount (or lack) of information to that regarding the covariance structure, to prevent the prior from swamping the relevant information contained in the female crab data; information from the female data is useful only for estimation of $(\boldsymbol{\mu}_1^*, \boldsymbol{\Sigma}_1^*, \boldsymbol{\mu}_2^*, \boldsymbol{\Sigma}_2^*)$, the unordered version of $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, while the information from the male crab data is useful for determining to which color group $(\boldsymbol{\mu}_1^*, \boldsymbol{\Sigma}_1^*)$ corresponds.

As a second stage of the prior construction, we consider the possibility of a difference in carapace size between male and female crabs. Since the current prior is centered at the blue and orange male crab sample means, we must specify a prior distribution for the possible male-to-female location shift. Relying on intuition, it seems reasonable that the amount of shift in one characteristic should be positively correlated with the amount of shift in another, and that the differences between blue male and female crabs should be similar to the differences between orange male and female crabs. Assuming no *a priori* knowledge of which gender should be larger, we use a normal distribution centered at 0 to model the location shift.

The new model for the means is

$$\boldsymbol{\mu} = \begin{pmatrix} \bar{\mathbf{x}}_1 \\ \bar{\mathbf{x}}_2 \end{pmatrix} + \frac{1}{\sqrt{k}}(\boldsymbol{\beta} + \sqrt{c_1}\,\boldsymbol{\delta}_1 + \sqrt{c_2}\,\boldsymbol{\delta}_2 + \sqrt{c_3}\,\boldsymbol{\delta}_3),$$

where

$$\boldsymbol{\beta} \sim \mathbf{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{S}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_2 \end{bmatrix}\right)$$

corresponds to the previous choice of prior that assumed total independence between color groups, and where $\sqrt{c_i}\,\boldsymbol{\delta}_i$ are terms that model the location shift from male crab means to female crab means as follows:

(i)

$$\boldsymbol{\delta}_1 \sim \mathbf{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{V} & \mathbf{V} \\ \mathbf{V} & \mathbf{V} \end{bmatrix}\right),$$

where

$$\mathbf{V} = \begin{bmatrix} v_1 & \sqrt{v_1 v_2} & \sqrt{v_1 v_3} \\ & v_2 & \sqrt{v_2 v_3} \\ & & v_3 \end{bmatrix},$$

represents a location shift that is perfectly correlated within and between color groups.

(ii)

$$\boldsymbol{\delta}_2 \sim \mathbf{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{V}_0 & \mathbf{V}_0 \\ \mathbf{V}_0 & \mathbf{V}_0 \end{bmatrix}\right),$$

where

$$\mathbf{V}_0 = \begin{bmatrix} v_1 & 0 & 0 \\ & v_2 & 0 \\ & & v_3 \end{bmatrix},$$

represents a location shift that is perfectly correlated between color groups, but independent between characteristics.

(iii)

$$\boldsymbol{\delta}_3 \sim \mathbf{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{V} \end{bmatrix}\right),$$

where $\mathbf{V}$ is as in (i), represents a location shift that is perfectly correlated between characteristics, but independent between color groups.

Thus we have that

$$\boldsymbol{\mu} \sim \mathbf{N}\left(\begin{pmatrix} \bar{\mathbf{x}}_1 \\ \bar{\mathbf{x}}_2 \end{pmatrix}, \boldsymbol{\Sigma}^{(b)} \equiv \left(\frac{1}{k}\begin{bmatrix} \mathbf{S}_1 + (c_1 + c_3)\mathbf{V} + c_2\mathbf{V}_0 & c_1\mathbf{V} + c_2\mathbf{V}_0 \\ c_1\mathbf{V} + c_2\mathbf{V}_0 & \mathbf{S}_2 + (c_1 + c_3)\mathbf{V} + c_2\mathbf{V}_0 \end{bmatrix}\right)\right).$$

Since the covariance matrices in (i) and (ii) are singular, (i), (ii) and (iii) are regarded only as a heuristic means to obtain a model for the location shift. The parameters v_1, v_2 and v_3 reflect the amount of variation in the location shift from males to females and must be elicited. We constrain $c_1 + c_2 + c_3$ to equal 1, so that c_1, c_2 and c_3 can be interpreted as the proportions of contribution of each kind of shift to the overall shift in location. These quantities must also be elicited. Alternatively, hyperpriors may be placed on the v_i's and c_i's. We choose $\mathbf{c}^{\mathsf{T}} = (0.6, 0.1, 0.3)$ to reflect our belief that there is a small degree of independence between the male-to-female location shifts for the different characteristics and color groups. We rely on an empirical approach to estimate the variances of the shifts in the three components by assuming that they are roughly the same as the variances of the carapace characteristics, which we estimate by the diagonal of the pooled sample covariance matrix obtained from the male crab data: $\mathbf{v}^{\mathsf{T}} = (11.995, 4.668, 12.210)$. This completes the second stage of our prior specification.

The third stage of our prior specification consists of adjustment of the prior through investigation of the prior odds discussed in Section 2. Since the results of Section 2 suggest that with a large number of unclassified training data, the posterior for $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ will place most of its mass within a small neighbourhood of the two reflections of the truth, where the amount of mass given at each reflection depends upon prior odds at the truth. We therefore propose to check the reasonableness of the prior by checking the prior odds over a region of the parameter space that seems likely to contain the truth. If the odds over this region are unreasonable or do not coincide with *a priori* beliefs, then adjustments to the prior are necessary.

Considering, for the moment, the case of general covariance matrix $(\hat{\boldsymbol{\Sigma}})$ for the mean parameter $\boldsymbol{\mu}$ in the prior given by (3), we have

$$\begin{aligned}
&\pi(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \\
&\qquad \propto \exp[-\tfrac{3}{2}\{(\boldsymbol{\mu} - \bar{\mathbf{x}})^{\mathsf{T}} \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{x}}) + \mathrm{tr}(\mathbf{S}_1 \boldsymbol{\Sigma}_1^{-1} + \mathbf{S}_2 \boldsymbol{\Sigma}_2^{-1})\}] \\
&\qquad \times |\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2|^{\frac{1}{2}}.
\end{aligned}$$

The prior odds of $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ to $(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ are given by

$$\begin{aligned}
R(\boldsymbol{\theta}) &= \frac{\pi(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)}{\pi(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)} \\
&= \exp[-\tfrac{3}{2}\{(\boldsymbol{\mu} - \bar{\mathbf{x}})^{\mathsf{T}} \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{x}}) - (\boldsymbol{\mu}' - \bar{\mathbf{x}})^{\mathsf{T}} \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\mu}' - \bar{\mathbf{x}})\} \\
&\qquad - \tfrac{3}{2}\,\mathrm{tr}\{(\mathbf{S}_1 - \mathbf{S}_2)(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\}], \qquad (4)
\end{aligned}$$

where $\boldsymbol{\mu}' = (\boldsymbol{\mu}_2^{\mathsf{T}}, \boldsymbol{\mu}_1^{\mathsf{T}})^{\mathsf{T}}$.

We can obtain a heuristic estimate of the amount of the contribution to the prior odds of each of $\boldsymbol{\mu}$ and $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ by setting each of them in turn equal to some value that lies along the general hyperplane $\{\boldsymbol{\eta} | (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = (\boldsymbol{\eta}_2, \boldsymbol{\eta}_1)\}$. Thus, to gain some insight into the relative importance of each parameter in the prior odds, we compute the odds at the three Θ^2-space points

(i)

$$\boldsymbol{\mu} = \begin{pmatrix} \bar{\mathbf{x}}_1 \\ \bar{\mathbf{x}}_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_1 = \mathbf{S}_1, \qquad \boldsymbol{\Sigma}_2 = \mathbf{S}_2,$$

(ii)

$$\boldsymbol{\mu} = \begin{pmatrix} \bar{\mathbf{x}}_1 \\ \bar{\mathbf{x}}_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{S},$$

where $\mathbf{S}$ is the pooled sample covariance matrix of the male crabs, and

TABLE 2: Prior odds for three parameter points.

Mean	Covariances	Prior odds
$\bar{\mathbf{x}}$	$\Sigma_1 = \mathbf{S}_1,\ \Sigma_2 = \mathbf{S}_2$	12,642
$\bar{\mathbf{x}}$	$\Sigma_1 = \Sigma_2 = \mathbf{S}$	3,666
$\begin{pmatrix}\frac{1}{2}(\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2)\\ \frac{1}{2}(\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2)\end{pmatrix}$	$\Sigma_1 = \mathbf{S}_1,\ \Sigma_2 = \mathbf{S}_2$	3.449

TABLE 3: Minimum and maximum prior odds for three choices of t.

t	$\alpha_{\min}$	$\alpha_{\max}$	$R_{\min}$	$R_{\max}$
1	0.7532	1.2468	219.9481	7.27 E + 05
4	0.5063	1.4937	3.8268	4.18 E + 07
9	0.2595	1.7405	0.0666	2.40 E + 09

(iii)
$$\boldsymbol{\mu} = \begin{pmatrix}\frac{1}{2}(\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2)\\ \frac{1}{2}(\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2)\end{pmatrix}, \qquad \Sigma_1 = \mathbf{S}_1, \qquad \Sigma_2 = \mathbf{S}_2$$

for the prior $\boldsymbol{\mu} \sim \mathbf{N}(\bar{\mathbf{x}}, \Sigma^{(b)})$. To compute the prior odds, we use the expression (4) and set $\hat{\Sigma} = \Sigma^{(b)}$. The prior odds are given in Table 2. The table shows that the vast majority of the contribution to the prior odds comes from the mean parameter. We therefore concentrate on this portion of the parameter space and set (Σ_1, Σ_2) equal to $(\mathbf{S}_1, \mathbf{S}_2)$ for the remainder of the prior odds analysis.

We also see from Table 2 that the prior odds at the prior mean for $\boldsymbol{\mu}$ is inordinately large, even when the covariance parameters are chosen to be equal. It is of considerable concern that the prior odds might show large fluctuations over relatively small regions of the parameter space. For this reason, we consider minimization and maximization of the prior odds over a portion of the parameter space considered likely to contain the true value of $\boldsymbol{\mu}$. We focus attention on the most recent version of the prior $[\boldsymbol{\mu} \sim \mathbf{N}(\bar{\mathbf{x}}, \Sigma^{(b)})]$ and consider the region

$$\Gamma_t = \{\boldsymbol{\mu}, \Sigma_1, \Sigma_2 | (\boldsymbol{\mu} - \bar{\mathbf{x}})^{\mathrm{T}} \Sigma^{(b)-1} (\boldsymbol{\mu} - \bar{\mathbf{x}}) \leq t,\ \Sigma_1 = \mathbf{S}_1,\ \Sigma_2 = \mathbf{S}_2\}.$$

That is, we consider all points in the "mean portion" of the parameter space that lie within $\sqrt{t}$ standard deviations of the prior mean $\bar{\mathbf{x}}$. It can be shown that the minimum and maximum prior odds over Γ_t lie at the two intersections of the region Γ_t and the hyperplane $\boldsymbol{\mu} = \alpha\bar{\mathbf{x}} + (1 - \alpha)\bar{\mathbf{x}}'$, $\alpha \in (-\infty, \infty)$, where $\bar{\mathbf{x}}' = (\bar{\mathbf{x}}_2^{\mathrm{T}}, \bar{\mathbf{x}}_1^{\mathrm{T}})^{\mathrm{T}}$. Solving for α under three choices of t ($t \in \{1, 4, 9\}$) and computing the prior odds at those points gives Table 3.

It is clear from the table that the current version of the prior results in excessive, erratic behaviour of the prior odds, even for points within 1 standard deviation of the prior mean. Therefore, the final stage of prior construction is to *flatten* the prior so as to force the prior odds at each point in Γ_t to lie within an acceptable range.

A standard flattening method replaces the multinormal prior for $\boldsymbol{\mu}$ with a multivariate-T prior with location parameter $\bar{\mathbf{x}}$, scale matrix $\Sigma^{(b)}$ and some number of degrees of freedom. For T priors with degrees-of-freedom parameter ranging from 1 to 5, the

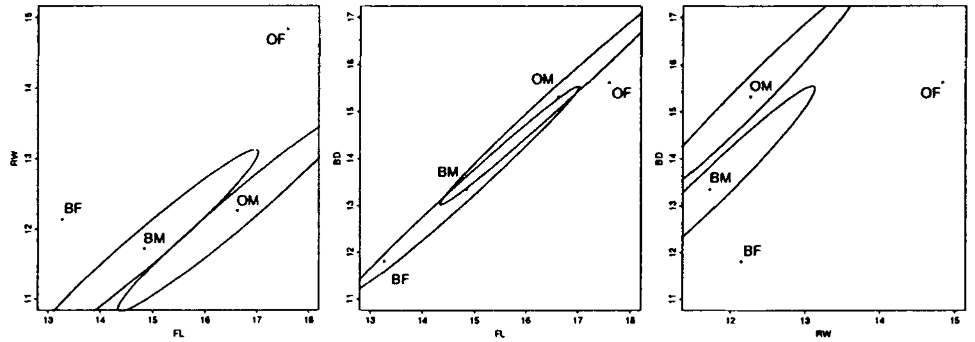

FIGURE 2: Contours of a prior with unstable prior odds ratio.

maximum prior odds over the ellipsoid corresponding to $t = 1$ ranges from 1120 to 2308. As the degrees of freedom increase, the T prior tends toward the normal prior with the prior odds reflecting this trend. Hence, replacing the normal prior with a multivariate-T does not solve the problem of erratic odds.

As an alternative, we return to the normal prior and consider reducing the degree of correlation between components of the mean by injecting a small amount of additional variability into the prior. Specifically, we specify a new covariance matrix for $\boldsymbol{\mu}$,

$$\boldsymbol{\Sigma}^{(c)} = \boldsymbol{\Sigma}^{(b)} + \gamma \operatorname{diag}(\boldsymbol{\Sigma}^{(b)}).$$

We then find limits on γ that guarantee $r_1 < \min_{\Gamma_t} R(\boldsymbol{\theta}) < \max_{\Gamma_t} R(\boldsymbol{\theta}) < r_2$ where r_1 and r_2 are specified bounds. It can be shown that

$$r_1 < R(\boldsymbol{\theta}) < r_2 \quad \Longleftrightarrow \quad \gamma_1 < \gamma < \gamma_2$$

for some γ_1 and γ_2. As $\gamma \longrightarrow \infty$, the prior odds converges to that of the case where the two color-group mean vectors are taken to be equal (row 3 of Table 2). In our case, since 3.449 seems a quite reasonable value for the prior odds, γ_2 is taken to be ∞. Obtaining the lower bound γ_1 requires numerical methods; the size of γ_1 depends upon the bounds r_1 and r_2 and upon the size of the region over which we wish to stabilize the prior odds. For the choices $r_1 = 0.01$ and $r_2 = 100$, the numerical solutions for γ_1 corresponding to $t = 1$, 4 and 9 are $\gamma_1 = 0.0381$, 0.0611 and 0.0883, respectively. Figures 2 and 3 display two-dimensional contours of the prior for $\boldsymbol{\mu}$ before and after the addition of the extra variation. The points labelled BM, OM, BF and OF are the respective means of the blue males, orange males, blue females and orange females. Figure 3 was produced using $\gamma = 0.0611$. The contours displayed in the plots correspond to regions including points within $\frac{1}{2}$ standard deviation of the blue and orange male sample means.

The end goal of the analysis is to estimate $P\{Z_i = 1|\mathbf{X}_i = \mathbf{x}_i\}$, where $\mathbf{X}_i$ is the feature vector for the ith female crab. The prior distribution for Z_i is Bernoulli(α), with $\alpha = \frac{1}{2}$. Since the integrals necessary to compute the predictive densities cannot be expressed in closed form, a Gibbs sampler (Gelfand and Smith 1990) was applied to obtain a sample of the relevant probabilities. The Gibbs sampler had a burn-in phase of 1000 iterations and subsequently sampled 10,000 *a posteriori* blue crab membership probabilities for each of the 100 female observations. Each female observation was then classified to blue population if the mean of the generated membership probabilities exceeded 0.5. If the

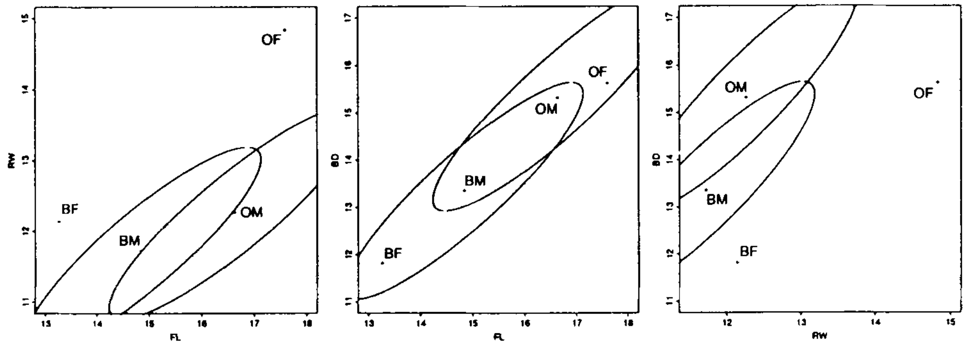

FIGURE 3: Contours of a prior with stable prior odds ratio.

TABLE 4: Error rates for two choices of prior.

Prior covariance	Color group	No. incorrect (proportion)	No. of clear classifications	No. of incorrect clear classifications
$\Sigma^{(b)}$	Both	75 (0.75)	54	46
	Blue	36 (0.72)	22	20
	Orange	39 (0.78)	32	26
$\Sigma^{(c)}$	Both	23 (0.23)	0	—
	Blue	10 (0.20)	0	—
	Orange	13 (0.26)	0	—

mean of the generated membership probabilities did not exceed 0.5, the observation was classified to the orange population. The program was written in C, used IMSL random variate generation and matrix manipulation routines, and was run on an HP 715/64 workstation, timing out at approximately 2 min 37 s.

Overall error rates and error rates by color group appear in Table 4 for both choices of prior covariance matrix, $\Sigma^{(b)}$ and $\Sigma^{(c)}$. The number of incorrect classifications is much smaller under the robust prior. The fourth column of the table gives the number of observations for which the *a posteriori* probability of blue crab membership was either at least 0.8 or at most 0.2; we refer to these observations as *clear classifications*. The fifth column gives the number of misclassified clear classifications. Note the dramatic difference in the number of clear classifications for the two priors. Extreme *a posteriori* group membership probabilities contradict our uncertainty about the labelling scheme; the "robust" prior serves to stabilize these *a posteriori* probabilities, forcing them to agree more closely with our lack of labeling information. Under the robust prior, the minimum and maximum *a posteriori* probabilities of blue crab membership were 0.37 and 0.61, respectively. The number of clear classifications allocated incorrectly under the $\Sigma^{(b)}$ prior is particularly disturbing and illustrates the need for a more robust prior.

4. CONCLUSION

We advocate, as a key step in prior elicitation, consideration of the asymptotic behaviour of the posterior distribution under various parameter values. In the classification problem, an important asymptotic allows the size of the unclassified sample to grow while keeping the size of the classified sample fixed. This asymptotic highlights the features of the posterior for which the unclassified sample conveys little or no information: namely, the labelling of the K classes. Since no amount of unclassified data will fully determine the class labels, the analyst must take special care to create a prior that provides acceptable large-sample inference for these class labels.

The analytical results of Section 2 describe conditions under which a prior distribution will produce a posterior distribution, with convergent odds as $n \longrightarrow \infty$, of the various class labelings. Roughly, the prior distribution should have locally stable odds of the various class labellings. The example of Section 3 shows how these results can influence the elicitation of a prior distribution. In the example, we develop a methodology for evaluation of the prior distribution when working with normal likelihoods, and also show how the prior distribution can be modified to produce acceptable large sample inference. The final choice of a prior distribution, in this instance, leads to a much superior classification rule.

APPENDIX

Proof of Lemma 1. (B1) is obvious from (A1). For (B2), let r^* be arbitrary, and begin by noting that, for any integer p,

$$\begin{aligned}\mathcal{E}_{\boldsymbol{\eta}_{0i}} \sup_{\Theta} \bar{H}_p^i(\boldsymbol{\eta}) &\leq \frac{1}{p}\sum_{j=1}^{p} \mathcal{E}_{\boldsymbol{\eta}_{0i}} \sup_{\Theta} H^i(\mathbf{x}_j|\boldsymbol{\eta}) \\ &\leq M,\end{aligned}$$

where the second inequality follows from condition (A3).

Choose $r < \min\{0, r^* - KM + \sum_i \log \alpha_i\}$ with M given by (A3), and let $D_1, \ldots, D_K$ be the cocompact sets given by (A2). Let

$$D^c = \bigcup_{\boldsymbol{\beta} \in B} D^c_{\beta_1} \times \cdots \times D^c_{\beta_K},$$

where

$$D^c_{\beta_1} \times \cdots \times D^c_{\beta_K} = \{(u_1, \ldots, u_K) | u_1 \in D^c_{\beta_1}, \ldots,\ u_K \in D^c_{\beta_K}\}$$

and $B = \{\text{all permutations of } (1, 2, \ldots, K)\}$. Clearly, D is cocompact. Choose $p > Kp^*$, where $p^* = \max_i\{p_i\}$, with the p_i's given by (A2). We now write

$$\mathcal{E}_{\boldsymbol{\theta}_0} \sup_{D} \bar{H}_p(\boldsymbol{\theta}) = \sum_{S} \mathcal{E}_{\boldsymbol{\theta}_0}\left(\sup_{D} \bar{H}_p(\boldsymbol{\theta}) \middle| Z_1, \ldots, Z_p \right) P\{Z_1, \ldots, Z_p\},$$

where $Z_1, \ldots, Z_p$ are the class membership indicators ($Z_i = j$ if $\mathbf{X}_i$ originates from population j), and S is the collection of all possible values of $\mathbf{Z} = (Z_1, \ldots, Z_p)$. Let $T_i = \{j | Z_j = i,\ j = 1, \ldots, p\}$ and $t_i = \# T_i$ (so that T_i indexes the observations from population i and $\sum_i t_i = p$). Then

$$
\begin{aligned}
&\mathcal{E}_{\boldsymbol{\theta}_0} \sup_D \bar{H}_p(\boldsymbol{\theta}) \\
&= \sum_S \int \sup_D \frac{1}{p} \left(\sum_{i=1}^K \sum_{j \in T_i} \log \frac{\sum_{l=1}^K \alpha_l g(\mathbf{x}_j|\boldsymbol{\eta}_l)}{\sum_{l=1}^K \alpha_l g(\mathbf{x}_j|\boldsymbol{\eta}_{0l})} \right) \prod_{m=1}^K \prod_{n \in T_m} [g(\mathbf{x}_n|\boldsymbol{\eta}_{0m})\, \mathrm{d}\mathbf{x}_n] \\
&\quad \times P\{Z_1, \ldots, Z_p\} \\
&\le \sum_S \sum_{i=1}^K \int \sup_D \frac{1}{t_i} \left(\sum_{j \in T_i} \log \frac{\sum_{l=1}^K \alpha_l g(\mathbf{x}_j|\boldsymbol{\eta}_l)}{\alpha_i g(\mathbf{x}_j|\boldsymbol{\eta}_{0i})} \right) \prod_{m=1}^K \prod_{n \in T_m} [g(\mathbf{x}_n|\boldsymbol{\eta}_{0m})\, \mathrm{d}\mathbf{x}_n] \\
&\quad \times P\{Z_1, \ldots, Z_p\} \\
&\le \sum_S \sum_{i=1}^K \int \sup_{D_i} \frac{1}{t_i} \left(\sum_{j \in T_i} \log \frac{g(\mathbf{x}_j|\boldsymbol{\eta}_i)}{\alpha_i g(\mathbf{x}_j|\boldsymbol{\eta}_{0i})} \right) \prod_{m=1}^K \prod_{n \in T_m} \{g(\mathbf{x}_n|\boldsymbol{\eta}_{0m})\, \mathrm{d}\mathbf{x}_n\} \\
&\quad \times P\{Z_1, \ldots, Z_p\} \\
&= \sum_S \sum_{i=1}^K \left(\mathcal{E}_{\boldsymbol{\eta}_{0i}} \sup_{D_i} \bar{H}_{t_i}^i(\boldsymbol{\eta}) - \log \alpha_i \right) P\{Z_1, \ldots, Z_p\} \\
&= \sum_S \sum_{i=1}^K \left(\mathcal{E}_{\boldsymbol{\eta}_{0i}} \sup_{D_i} \bar{H}_{t_i}^i(\boldsymbol{\eta})(1\{t_i \ge p_i\} + 1\{t_i < p_i\}) \right) P\{Z_1, \ldots, Z_p\} \\
&\quad - \sum_{i=1}^K \log(\alpha_i).
\end{aligned}
$$

Now, since $\sum_i t_i = p > K \max_i\{p_i\}$, at least one of the t_i must be greater than $\max_i\{p_i\}$. Since $r < 0$, we therefore have that, for p, D and r as chosen above,

$$
\mathcal{E}_{\boldsymbol{\theta}_0} \sup_D \bar{H}_p(\boldsymbol{\theta}) \le r + KM - \sum_{i=1}^K \log \alpha_i < r^*. \quad \square
$$

Proof of Result 2. Let ϵ be small enough that $\bigcap_{\Theta_0} N_\epsilon(\boldsymbol{\theta}) = \emptyset$. Denoting $\Pi_p(\cdot) \equiv \Pi(\cdot|\mathbf{X}_1, \ldots, \mathbf{X}_p)$, $f_p(\mathbf{x}|\boldsymbol{\theta}) \equiv f(\mathbf{x}_1, \ldots, \mathbf{x}_p|\boldsymbol{\theta})$, and $N_\epsilon^U(\boldsymbol{\theta}_0) \equiv \bigcup_{\Theta_0} N_\epsilon(\boldsymbol{\theta})$, Berk's result gives

$$
\Pi_p(N_\epsilon(\boldsymbol{\theta}')) = \frac{\Pi_p(N_\epsilon(\boldsymbol{\theta}'))}{\Pi_p(N_\epsilon^U(\boldsymbol{\theta}_0))} + o(1) \qquad \text{for all} \quad \boldsymbol{\theta}' \in \Theta_0,
$$

where Θ_0 is given in (2) and where $o(1) \overset{p \to \infty}{\longrightarrow} 0$ (a.s. $F_{\boldsymbol{\theta}_0}$). Since $\pi(\cdot)$ is continuous, for every $\boldsymbol{\theta} \in \Theta$ and $\epsilon > 0$ there exists $\delta(\epsilon)$ such that

$$
\pi(\boldsymbol{\theta}^*) \in (\pi(\boldsymbol{\theta}) - \delta(\epsilon), \pi(\boldsymbol{\theta}) + \delta(\epsilon)) \qquad \text{for all} \quad \boldsymbol{\theta}^* \in N_\epsilon(\boldsymbol{\theta}).
$$

Therefore, since the likelihood is symmetric in neighbourhoods around the various reflections of $\boldsymbol{\theta}_0$,

$$
\begin{aligned}
\Pi_p(N_\epsilon(\boldsymbol{\theta}')) &\ge \frac{\{\pi(\boldsymbol{\theta}') - \delta(\epsilon)\} \int_{N_\epsilon(\boldsymbol{\theta}')} f_p(\mathbf{x}|\boldsymbol{\theta}^*)\, \mathrm{d}\boldsymbol{\theta}^*}{\sum_{\Theta_0} \{\pi(\boldsymbol{\theta}) + \delta(\epsilon)\} \int_{N_\epsilon(\boldsymbol{\theta})} f_p(\mathbf{x}|\boldsymbol{\theta}^*)\, \mathrm{d}\boldsymbol{\theta}^*} + o(1) \\
&= \frac{\pi(\boldsymbol{\theta}') - \delta(\epsilon)}{\sum_{\Theta_0} \pi(\boldsymbol{\theta}) + K!\delta(\epsilon)} + o(1).
\end{aligned}
$$

Similarly,

$$\Pi_p(N_\epsilon(\boldsymbol{\theta}')) \leq \frac{\pi(\boldsymbol{\theta}') + \delta(\epsilon)}{\sum_{\Theta_0} \pi(\boldsymbol{\theta}) - K!\delta(\epsilon)} + o(1).$$

Hence, for any $\gamma > 0$, there is a small enough ϵ such that

$$P_{\boldsymbol{\theta}_0}\left\{\lim_{p\to\infty} \Pi_p(N_\epsilon(\boldsymbol{\theta}')) \in \left(\frac{\pi(\boldsymbol{\theta}')}{\sum_{\Theta_0} \pi(\boldsymbol{\theta})} - \gamma,\ \frac{\pi(\boldsymbol{\theta}')}{\sum_{\Theta_0} \pi(\boldsymbol{\theta})} + \gamma\right)\right\} = 1.$$

The result now follows from

$$\Pi_p(N_\epsilon(\boldsymbol{\theta}')) = \Pi_p(N_\epsilon(\boldsymbol{\theta}')) + o(1). \quad \square$$

Lemma 2. *Assume the a priori probabilities of class membership are equal, and suppose the class conditional densities satisfy conditions* (A1) *through* (A4). *Denote the prior probability measure by* $\Pi(\cdot)$, *and suppose* $\Pi(\Theta_0) > 0$. *Then*

$$\Pi(\Theta_0 | \mathbf{X}_1, \ldots, \mathbf{X}_p) \to 1 \qquad (a.s.\ F_{\boldsymbol{\theta}_0}),$$

where Θ_0 *is given in (2).*

ACKNOWLEDGEMENTS

We thank the associate editor and referees for comments that improved the presentation of the paper.

REFERENCES

Berger, J.O. (1994). An overview of robust Bayesian analysis. *Test*, 3, 5–124.

Berk, R.H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Statist.*, 37, 51–58.

Campbell, N.A., and Mahon, R.J. (1974). A multivariate study of variation in two species of rock crab of the genus *Leptograpsus*. *Austral. J. Zool.*, 22, 417–425.

Cooley, C.A. (1996). Bayesian and nonparametric models in the classification problem, Ph.D. Dissertation, The Ohio State University.

Cooley, C.A., and MacEachern, S.N. (1996). Prior elicitation in the classification Problem. Technical Report 574, Department of Statistics, The Ohio State University.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, 85, 398–409.

Lavine, M., and West, M. (1992). A Bayesian method for classification and discrimination. *Canad. J. Statist.*, 20, 451–461.

Lehmann, E.L. (1983). *Theory of Point Estimation*. Wiley.

McLachlan, G.J. (1975). Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *J. Amer. Statist. Assoc.*, 70, 365–369.

McLachlan, G.J. (1977). Estimating the linear discriminant function from initial samples containing a small number of unclassified observations. *J. Amer. Statist. Assoc.*, 72, 403–406.

O'Neill, T.J. (1978). Normal discrimination with unclassified observations. *J. Amer. Statist. Assoc.*, 73, 821–826.

Ripley, B.D. (1994). Flexible non-linear approaches to classification. *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*, ASI Proc., Subser. F, Computer and Systems Sciences (V. Cherkassky, J.H. Friedman, and H. Wechsler, *eds.*), Springer-Verlag, 105–126.

Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley.

Received 27 March 1996
Revised 21 May 1997
Accepted 18 July 1997

Department of Statistics
Ohio State University
Columbus, Ohio 43210-1247
U.S.A.