A Bayesian method for classification and discrimination

By

MICHAEL LAVINE and MIKE WEST

Institute of Statistics and Decision Sciences

Duke University, Durham, NC 27706, USA.

SUMMARY

We discuss Bayesian analyses of traditional normal mixture models for classification and discrimination. The development involves application of an iterative resampling approach to Monte Carlo inference, commonly called Gibbs sampling, and demonstrates routine application. We stress the benefits of exact analyses over traditional classification and discrimination techniques, including the ease with which such analyses may be peformed in a quite general setting, with possibly several normal mixture components having different covariance matrices, the computation of exact posterior classification probabilities for observed data and for future cases to be classified, and posterior distributions for these probabilities that allow for assessment of second-level uncertainties in classification.

Some key words: Bayesian computations; Classification; Discrimination; Mixtures of normals; Posterior sampling

Both authors were supported in part by NSF grant DMS-8903842. The second author was also supported by NSF grant DMS-9024793.

1. INTRODUCTION

Binder (1978) describes a general class of normal mixture models, and discusses some ingredients of Bayesian approaches to classification, clustering and discrimination using such models. Hartigan (1975, Chapter 5), McClachlan and Basford (1988), and Titterington, Smith and Makov (1985) discuss approaches to inference, Bayesian and non-Bayesian, in similar models. These analyses, and most others to date, depend on various forms of analytic or numerical approximation to inferences due to the mathematical difficulties induced by the complexity of likelihood functions for model parameters. Recent developments in Monte Carlo analysis using iterative resampling schemes, as in Gelfand and Smith (1990), for example, now provide for the relevant calculations to be performed. This is demonstrated and illustrated here.

2. NORMAL MIXTURE MODELS

Suppose data y_j, $(j = 1, 2, \ldots, ; y_j \in \Re^p)$ are modelled as exchangeably distributed according to a discrete mixture of a known number k of multivariate normal components. Specifically, $(y_j|\pi) \sim \mathcal{N}(\mu_i; \Sigma_i)$ with probability θ_i, for each $i = 1, \ldots, k$, for some mean vectors $\mu = (\mu_i; i = 1, \ldots, k)$, variance matrices $\Sigma = (\Sigma_i; i = 1, \ldots, k)$, and classification probabilities in the $k-$vector $\theta = (\theta_1, \ldots, \theta_k)$; here π represents all parameters $\pi = (\mu, \Sigma, \theta)$. Introduce classification variables z_j, where $z_j = i$ implies that y_j is drawn from component i of the mixture, or classified into group i. Thus knowledge of z_j revises $p(y_j|\pi)$ to the single normal component $(y_j|z_j = i, \pi) \sim \mathcal{N}(\mu_i; \Sigma_i)$. Additionally, $(z_j|\theta)$ are conditionally independent with $P(z_j = i|\theta) = \theta_i$. We are concerned with problems of inference about the model parameters π, the classification quantities z_j, and the classification of future cases. Inferences will be based on observing a sample of the y_j with, typically, only a fraction of the corresponding classification quantities z_j observed. Here we specify a class of prior distributions and detail some structure of the resulting

posteriors.

We generalise Binder (1978, Section 3.3) in defining a conditionally conjugate prior for π. We assume (μ_i, Σ_i) to be mutually independent over groups $i = 1, \ldots, k$, with normal-inverse Wishart priors. The notation and structure for such priors follows West and Harrison (1989, Section 15.4.3) and is briefly detailed in an appendix here. We assume $(\mu_i|\Sigma_i) \sim \mathcal{N}(m_{i,0}; \Sigma_i/h_{i,0})$, for some means $m_{i,0}$ and precision parameters $h_{i,0} > 0$, and take the margin for Σ_i as the inverse Wishart distribution with $v_{i,0} > 0$ degrees of freedom and scale matrix $V_{i,0}$, denoted by $\Sigma_i \sim \mathcal{W}^{-1}(v_{i,0}, V_{i,0})$, as in appendix. Finally, we assume θ to be independent of (μ, Σ), with a Dirichlet prior, $\theta \sim \mathcal{D}(a_0)$ where $a_0 = (a_{1,0}, \ldots, a_{k,0})$; the prior mean vector is $E(\theta) = a_0/A_0$, where $A_0 = a_{1,0} + \ldots + a_{k,0}$.

Conder now a set of observations $y = (y_j; j = 1, \ldots, n)$ for some integer n, writing $z = (z_j; j = 1, \ldots, n)$. Under the specified model, the joint distribution of (y, z, π) has the following component conditional distributions.

(1) $p(\mu, \Sigma|y, z, \theta)$.
Fixing z implies the data are classified as k independent normal samples, and the analysis is standard, as in De Groot (1970, Section 9.10). Prior independence leads to posterior independence of the (μ_i, Σ_i) over groups i, with normal-inverse Wishart posteriors defined as follows. Let $G_i = \{j|z_j = i\}$, the index set for observations in group i, and $g_i = \#G_i$, so that $n = g_1 + \ldots + g_k$. For each group, the sufficient statistics are the mean vectors $\bar{y}_i = g_i^{-1} \sum y_j$, and the matrices of sums of squares and cross-products $S_i = \sum (y_j - \bar{y}_i)(y_j - \bar{y}_i)'$, where each sum is over $j \in G_i$. Then $p(\mu_i, \Sigma_i|y, z, \theta) = p(\mu_i, \Sigma_i|y, z)$ has components $(\mu_i|\Sigma_i, y, z) \sim \mathcal{N}(m_i; \Sigma_i/h_i)$ and $(\Sigma_i|y, z) \sim \mathcal{W}^{-1}(v_i, V_i)$ with $h_i = h_{i,0} + g_i$, $m_i = (h_{i,0}m_{i,0} + g_i\bar{y}_i)/h_i$, $v_i = v_{i,0} + g_i$ and $V_i = V_{i,0} + S_i + (\bar{y}_i - m_i)(\bar{y}_i - m_i)'g_i h_{i,0}/h_i$.

We note a minor modification of these results to apply when the reference prior $p(\mu_i, \Sigma_i) \propto |\Sigma_i|^{-(p+1)/2}$ is used in place of the normal-inverse

Wishart priors above. Then, assuming $g_i > p+1$, the posteriors $p(\mu_i, \Sigma_i|y, z)$ are as above, but now with $h_i = g_i$, $m_i = \bar{y}_i$, $v_i = g_i - p$ and $V_i = S_i$.

(2) $p(\theta|y, z, \mu, \Sigma)$.
Given z, θ is conditionally independent of (y, μ, Σ), and has the Dirichlet posterior $(\theta|z) \sim \mathcal{D}(a)$, where $a = (a_1, \ldots, a_k)$ with $a_i = a_{i,0} + g_i$; the posterior means are $E(\theta_i|z) = a_i/A$ where $A = a_1 + \ldots + a_k$.

(3) $p(z|y, \pi)$.
Given y and $\pi = (\mu, \Sigma, \theta)$, the z_j are conditionally independent. For each $j = 1, \ldots, n$,

$$P(z_j = i|y, \pi) \propto \theta_i p(y_j|\mu_i, \Sigma_i, z_j = i), \qquad (i = 1, \ldots, k),$$

and summing to unity over $i = 1, \ldots, k$. Here $p(y_j|\mu_i, \Sigma_i, z_j = i)$ is just the normal density function for group i, with mean vector μ_i and variance matrix Σ_i, evaluated at the point y_j.

(4) As a corollary to (1) and (2), we may easily obtain the marginal posteriors for $(\mu_i|y, z)$ and the predictive distributions for new cases drawn from any specified group. Using results and notation from the appendix, the margin for μ_i is $(\mu_i|y, z) \sim \mathcal{T}_{v_i}(m_i; V_i/(h_i v_i))$, with density given in equation (A1) of the appendix. In predicting a future observation drawn from group i, say $(y_f|\mu_i, \Sigma_i, z_f = i) \sim \mathcal{N}(\mu_i; \Sigma_i)$, the predictive distribution is $(y_f|y, z, z_f = i) \sim \mathcal{T}_{v_i}(m_i; Q_i)$, where z_f is the classification indicator for y_f and $Q_i = V_i(1+h_i)/(h_i v_i)$; the density function is given in equation (A2) of appendix. If z_f is unknown, the unconditional predictive distribution is just the mixture $p(y_f|y, z) = \sum p(y_f|y, z, z_f = i)a_i/A$.

A general framework supposes that we may observe a training sample of some t perfectly classified cases, and a further u unclassified cases. Thus we assume we are to observe data $y_{(T)} = (y_1, \ldots, y_t)$ together with classification

indicators $z_{(T)} = (z_1, \ldots, z_t)$, and then $y_{(U)} = (y_{t+1}, \ldots, y_{t+u})$, and we will process these two datasets sequentially, $(y_{(T)}, z_{(T)})$ followed by $y_{(U)}$. We will then proceed to inference about the model parameters π and the classification quantities $z_{(U)} = (z_j; j = t+1, \ldots, t+u)$ and also to predictive classification of future cases.

Consider first processing the training sample. The prior is conjugate and the analysis is standard, since the data are perfectly classified into normal components, the quantities $z_{(T)} = (z_1, \ldots, z_t)$ being known. The components of analysis are just as described under (1) and (2) above, with $n = t$ classified observations. Following the analysis, we are left with independent normal-inverse Wishart posteriors $p(\mu_i, \Sigma_i | y_{(T)}, z_{(T)})$, and the Dirichlet posterior $p(\theta | z_{(T)})$. The structure of the joint posterior for $\pi = (\mu, \Sigma, \theta)$ given $(y_{(T)}, z_{(T)})$ is just that of the prior, with the defining parameters appropriately updated.

Consider now the unclassified sample $y_{(U)}$. We know that $(y_{(U)}, z_{(U)})$ is conditionally independent of $(y_{(T)}, z_{(T)})$ given the parameters π, and so points (1) – (3) above apply to determine various components of the posterior $p(\mu, \Sigma, \theta, z_{(U)} | y_{(T)}, z_{(T)}, y_{(U)})$. This involves simply replacing the prior for π throughout by the similarly structured distribution $p(\pi | y_{(T)}, z_{(T)})$, just obtained, that summarises the revised state of information about the parameters based on the training sample. Now marginal posteriors for (μ, Σ), for example, are difficult to compute since $z_{(U)}$ is uncertain. It is at this point that Monte Carlo analysis using iterative resampling from the conditional posteriors defined under points (1) – (3) is useful.

3. SAMPLING THE POSTERIOR

We now identify the posterior $p(\pi | y_{(T)}, z_{(T)})$ as the prior in points (1) – (2) above, the unclassified sample $(y_{(U)}, z_{(U)})$ as the data (y, z) to be processed, with the sample size u replacing n. For notational convenience, write D as the known data information $D = (y_{(T)}, z_{(T)}, y_{(U)})$. Note that the fol-

lowing sampling exercise is computationally straightforward.

(a) Given $z_{(U)}$, we may draw a sample from the posterior $p(\mu, \Sigma|D, z_{(U)}) = \prod p(\mu_i, \Sigma_i|D, z_{(U)})$, the product over $i = 1, \ldots, k$ independent normal-inverse Wishart components defined as in (1). Convenient and efficient algorithms for simulating inverse Wishart distributions are given in Anderson (1984, p247 and pp254-255).

(b) Also given $z_{(U)}$, it is trivial to sample θ from the conditional Dirichlet posterior $p(\theta|D, z_{(U)})$ defined as in (2).

(c) Given $\pi = (\mu, \Sigma, \theta)$, it is similarly trivial to sample from the posterior $p(z_{(U)}|D, \pi)$ defined as in (3).

Based on these observations, an iterative resampling technique, as in Gelfand and Smith (1990), for example, provides for an approximate draw from the joint posterior $p(\pi, z_{(U)}|D)$ to be obtained as follows. Start with an assigned value for the initial classification vector $z_{(U)}$. Proceed through (a) and (b) to sample π from the conditional posterior based on this value of $z_{(U)}$. At (c), use this sampled value of π to determine $p(z_{(U)}|D, \pi)$ and sample from this distribution to get a new value for $z_{(U)}$. Return to (a) and repeat, iterating through this cycle repeatedly to update the values of π and $z_{(U)}$. With sufficient iteration, this process leads to 'final' values $(\pi, z_{(U)})$ that form an approximate draw from the joint posterior $p(\pi, z_{(U)}|D)$. Replicating the process provides for an approximate random sample to be drawn from the posterior, forming the basis of a Monte Carlo analysis.

A suitable starting value for the vector $z_{(U)}$ is given by initially classifying the data $y_{(U)}$ into groups $i = 1, \ldots, k$ according to their individual pre-posterior classification probabilities $P(z_j = i|y_{(T)}, z_{(T)}, y_j)$, for each $j = t + 1, \ldots, t + u$. These are easily computed via $P(z_j = i|y_{(T)}, z_{(T)}, y_j) \propto P(z_j = i|y_{(T)}, z_{(T)})p(y_j|y_{(T)}, z_{(T)}, z_j = i)$. The first term here is just $E(\theta_i|y_{(T)}, z_{(T)})$, the i^{th} element of the mean vector of the Dirichlet posterior $p(\theta|y_{(T)}, z_{(T)})$, from point (2) above. The second term is just the value at y_j of the density of

the multivariate T distribution for predicting new cases in group i, given under point (4) above. A referee suggests an alternative for determining starting values: select z_{t+1} as before and, for $j = t+2, \ldots, t+u$ select z_j based on the probabilities $P(z_j = i|y_{(T)}, z_{(T)}, y_{t+1}, \ldots, y_{j-1}, z_{t+1}, \ldots, z_{j-1}, y_j)$. We don't know which alternative is better.

Suppose this procedure is followed to produce a sample of size N from the posterior, denoted by $(\pi(r), z_{(U)}(r); r = 1, \ldots, N)$, say. Monte Carlo inference about the elements of π and $z_{(U)}$ may be based directly on the sampled values, or more efficiently on refined approximations to the marginal posteriors determined as follows. Simply recall that, were $z_{(U)}$ known, inference about elements of π would be based on standard normal theory. Also, were π known, then inference about $z_{(U)}$ would be simple too, based on the conditional probabilities $P(z_j = i|y, \pi)$ defined in item (3) of Section 2. Then the Monte Carlo approximations to $p(\pi|D)$ and $p(z_{(U)}|D)$ are simply the mixtures

$$\begin{aligned} p(\pi|D) &\approx N^{-1} \sum_{r=1}^{N} p(\pi|D, z_{(U)}(r)), \\ p(z_{(U)}|D) &\approx N^{-1} \sum_{r=1}^{N} p(z_{(U)}|D, \pi(r)). \end{aligned} \tag{1}$$

(i) The first equation in (1) has a margin for μ_i, $(i = 1, \ldots, k)$, that is a mixture of conditional T posteriors, easily evaluated and summarised. Similarly, inference about Σ_i will be based on a mixture of inverse Wisharts.

(ii) Of particular interest in discrimination and classification are the posterior probabilities $P(z_j = i|D)$, for each $i = t+1, \ldots, t+u$. The second equation in (1) directly gives Monte Carlo estimates of these quantities.

(iii) Consider prediction of further observations. Suppose that such an observation y_f is known to come from component i of the mixture; thus, if z_f is the classification indicator for y_f, we require the density function $p(y_f|D, z_f = i)$. Now (1) applies to give a mixture of T distributions,

each component $p(y_f|D, z_f = i, z_{(U)}(r))$ identified as described in point (4) of Section 2.

(iv) If z_f is unknown, the density $p(y_f|D) \approx N^{-1} \sum p(y_f|D, z_{(U)}(r))$ forms the basis for prediction of y_f. The mixture components here are easily given by

$$p(y_f|D, z_{(U)}) = \sum_{i=1}^{k} P(z_f = i|D, z_{(U)}) p(y_f|D, z_{(U)}, z_f = i);$$

the probability forming the first term of the summand here is evaluated as $P(z_f = i|D, z_{(U)}) = E(\theta_i|D, z_{(U)})$, and the second term is just the density in (iii).

(v) In attempting to classify y_f when z_f is unknown, we are interested in the posterior probabilities

$$P(z_f = i|D, y_f) \propto P(z_f = i|D) p(y_f|D, z_f = i). \tag{2}$$

The first term here is simply approximated, using (1), as $P(z_j = i|D) = E(\theta_i|D) \approx N^{-1} \sum E(\theta_i|D, z_{(U)}(r))$, the sum over $r = 1, \ldots, N$, of course. The second term is evaluated as in (iii).

In connection with classification and discrimination in points (ii) and (v), note that the classification probabilites are dependent and that the sampling based calculations allow for assessment of the dependence. Neighbouring points will, with high probability, belong to the same group. To focus discussion, consider the simple example of one dimensional data coming from a mixture of just two normal distributions with known variances of unity. Suppose also that, based on training data, the posteriors for group means are $\mu_i \sim \mathcal{N}((-1)^i 2; 1)$, and the posterior Dirichlet for θ has $a = (25, 25)$, with $E(\theta_i) = 0.5$ for $i = 1, 2$. Suppose two unclassified cases are observed at zero. Easy calculations show that $P(z_j = i|D) = 0.5$ for each i. However, it can also be shown that $P(z_1 = z_2|D) = 0.7$. This feature will arise, quite

generally, in considering an observation or a y_f that has neighbouring points that are uncertainly classified, and particularly when those points are influential in updating the posterior distributions of moments of any component normal with which they are identified in conditioning.

4. ILLUSTRATION

Illustration is based on a two dimensional, three component version of a waveform recognition problem developed in Breiman, Friedman, Olshen and Stone (1984, Section 2.6.2). Bivariate observations are generated from a $k = 3$, equally weighted component mixture of non-normal distributions, as follows. Define matrices

$$C_1 = \begin{pmatrix} 5 & 1 \\ 3 & 5 \end{pmatrix}, \quad C_2 = \begin{pmatrix} 0 & 1 \\ 1 & 5 \end{pmatrix}, \quad \text{and} \quad C_1 = \begin{pmatrix} 5 & 0 \\ 3 & 1 \end{pmatrix}.$$

An observation from component i of the mixture is generated according to

$$y_j = \begin{pmatrix} y_{j1} \\ y_{j2} \end{pmatrix} = C_i \begin{pmatrix} w_j \\ 1 - w_j \end{pmatrix} + \begin{pmatrix} \epsilon_{j1} \\ \epsilon_{j2} \end{pmatrix},$$

where w_j is uniform over the unit interval, and the ϵ_{jr} are independent $\mathcal{N}(0; 1/2)$ quantities. The w_j and ϵ_{jr} are also independent over observations j.

An initial training sample of size $t = 15$ drawn from this model appears in Figure 1(a); there are just 3 cases from component 1, 5 from component two, and 7 from component 3. This forms the basis of initial analysis using independent reference priors for the group moments (μ_i, Σ_i), and a reference prior for θ with Dirichlet parameter $a_0 = (0, 0, 0)$. Predictive distributions based on the training data appear in Figure 2. In these graphs, and in Figure 4, density contours plotted determine approximate 25, 50 and 75% regions. Figure 2 displays these contours for each of the three bivariate T densities, $p(y_f|y_{(T)}, z_{(T)}, z_f = i)$, and also the unconditional predictive density $p(y_f|y_{(T)}, z_{(T)})$, just the mixture of the three T densities as noted

in point (4) of Section 2. The paucity of training data from component 1, in particular, leads to a rather diffuse distribution for that component, with $v_1 = g_1 - p = 3 - 2 = 1$ degree of freedom; this is reflected in the mixture which has a mode corresponding to each component 2 and 3, but not 1. It is also straightforward to compute the posterior classification probabilities for a future case y_f,

$$P(z_f = i|y_{(T)}, z_{(T)}, y_f) \propto E(\theta_i|y_{(T)}, z_{(T)})p(y_f|y_{(T)}, z_{(T)}, z_f = i), \quad (i = 1, 2, 3).$$

Figure 3(a) displays a discrimination function based on these probabilities. In the dark region $P(z_f = 3|y_{(T)}, z_{(T)}, y_f)$ is the largest of the three probabilities; the shaded and white regions correspond to components 2 and 1 respectively. Figures 3(b), 3(c) and 3(d) displays contours of the classification probabilities. In Figure 3(b), for example, the four regions from dark to white are where $P(z_f = 1|y_{(T)}, z_{(T)}, y_f) > .9$, $.5 < P(z_f = 1|y_{(T)}, z_{(T)}, y_f) \leq .9$, $.05 < P(z_f = 1|y_{(T)}, z_{(T)}, y_f) \leq .5$, and $P(z_f = 1|y_{(T)}, z_{(T)}, y_f) \leq .05$, respectively. The probability on component i naturally decreases as we move away from the region of the mode of $p(y_f|y_{(T)}, z_{(T)}, z_f = i)$. One point of interest, however, which is quite general, is that this probability again increases eventually in some directions. For example, $P(z_f = 1|y_{(T)}, z_{(T)}, y_f)$ decreases as y_f moves away from the mode near $(2, 4)'$, but eventually increases again as y_f moves either North-West or South-East. Generally, once we are removed from the central region where the component densities vary, as displayed, the component T density that is most diffuse in any direction will eventually dominate.

So far the computations are standard, analytically derived. Now consider further, unclassified data. Figure 1(b) displays a further set $y_{(U)}$ of $u = 300$ unclassified cases. The iterative resampling analysis of Section 3 is performed to give inferences partially summarised in Figures 4 and 5. The Monte Carlo analysis has sample size $N = 500$, each sampling exer-

cise based on 50 resampling iterations. Figure 4 provides predictive densities $p(y_f|D, z_{(U)}, z_f = i)$, $(i = 1, 2, 3)$, and $p(y_f|D, z_{(U)})$ as described under points (iii) and (iv) of Section 3, neatly summarising the data analysis in predictive terms. Posterior classification probabilities for future cases may be evaluated as in point (v) of Section 3, and again plotted as functions of y_f; the discriminant function and contours of these probabilities appear in Figure 5. The superiority of such plots over the usual linear or quadratic discrimination rules are clear. However, if desired, analogues of such rules may be deduced if a classification loss function is imposed, since Monte Carlo approximations to the posterior expectations required to evaluate expected losses may be easily computed. Further inferences may be easily derived from the results of this analysis. Posterior inference for the parameters of the component normals are commonly of interest, and the ingredients for such additional computations are available, though are not pursued further here.

The computations were performed using C and Fortran routines running on DECstations under Ultrix 4.0. On a DECstation 2100, the resampling computations reported in this example were timed as follows. For the complete 500 samples and with 50 iterations each, the complete analysis of 300 unclassified cases to produce all the required outputs was timed at around 75 minutes cpu time. This code was not optimised in any way. The time will increase roughly in proportion to numbers of unclassified cases. Of course, additional effort is needed to deduce contour plots of predictive densities and classification probabilities, and for further posterior analysis.

APPENDIX

The notation and density functions for the normal-inverse Wishart distribution are as follows. For any $p-$vector μ and $p \times p$ variance matrix Σ we have $(\mu|\Sigma) \sim \mathcal{N}(m; \Sigma/h)$ for some precision parameter $h > 0$. Also, $\Sigma \sim \mathcal{W}^{-1}(v, V)$, the inverse Wishart distribution with density function $p(\Sigma) = c(p, v)|V|^{(p+v-1)/2}|\Sigma|^{-(p+v/2)}\exp((-0.5)\text{trace}(\Sigma^{-1}V))$, for some constant $c(p, v)$. Here $v > 0$ is the degrees of freedom and V is a variance matrix such that $E(\Sigma) = V/(v-2)$, for $v > 2$. The marginal multivariate T distribution for μ has v degrees of freedom, and scale matrix $M = V/(hv)$, with density function

$$p(\mu) = C(p, v)|M|^{-1/2}\{1 + (\mu - m)'M^{-1}(\mu - m)/v\}^{-(p+v)/2}, \qquad (A1)$$

where $C(p, v) = \Gamma((p+v)/2)(v\pi)^{-p/2}/\Gamma(v/2)$. By way of notation, we write $\mu \sim \mathcal{T}_v(m; M)$, the dimension p being implicit.

If $(y|\mu, \Sigma) \sim \mathcal{N}(\mu; \Sigma)$, then the predictive distribution for y is a similar T distribution but with increased spread, namely $Y \sim \mathcal{T}_v(m; Q)$ with $Q = V(1+h)/(hv)$. The density function is simply

$$p(y) = C(p, v)|Q|^{-1/2}\{1 + (y - m)'Q^{-1}(y - m)/v\}^{-(p+v)/2}. \qquad (A2)$$

REFERENCES

Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis* (2nd Edn.) New York:Wiley.

Binder, D.A. (1978). Bayesian cluster analysis. *Biometrika* **65**, 31-38.

Breiman, L., Friedman, J.K., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees*. California:Wadsworth.

De Groot, M.H. (1970). *Optimal Statistical Decisions*. New York:McGraw-Hill.

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

Hartigan, J.A. (1975). *Clustering Algorithms*. New York:Wiley.

McClachan G.J., and Basford, K.E. (1988). *Mixture Models: Inference and applications to clustering*. New York: Marcel Dekker.

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

West, M., and Harrison, P.J. (1989). *Bayesian Forecasting and Dynamic Models*. New York:Springer-Verlag.

TITLES AND LEGENDS FOR FIGURES

Figure 1.

Data $y_f = (y1, y2)$: (a) 15 classified cases marked as 1,2, and 3 according to group, and (b) 300 unclassified cases.

Figure 2.

Contours of densities $p(y_f|y_{(T)}, z_{(T)})$ and $p(y_f|y_{(T)}, z_{(T)}, z_f = i)$ for components $i = 1, 2$ and 3. The contours represent 25%, 50% and 75% probability regions, the latter being the outermost contour in each case.

Figure 3.

(a): Modal probabilities. The black, grey, and white regions are where components 3, 2, and 1 are favoured, respectively.
(b), (c), (d): Contours of $P(z_f = i|y_{(T)}, z_{(T)}, y_f)$, for components $i = 1, 2$ and 3. The black, dark grey, light grey and white regions are where $P(z_f = i|y_{(T)}, z_{(T)}, y_f) > 0.90$, $0.50 < P(z_f = i|y_{(T)}, z_{(T)}, y_f) \leq 0.90$, $0.05 < P(z_f = i|y_{(T)}, z_{(T)}, y_f) \leq 0.50$, and $0.05 \geq P(z_f = i|y_{(T)}, z_{(T)}, y_f)$, respectively.

Figure 4.

Contours of densities $p(y_f|D, z_{(U)})$ and $p(y_f|D, z_{(U)}, z_f = i)$ for components $i = 1, 2$ and 3. The contours represent 25%, 50% and 75% probability regions, the latter being the outermost contour in each case.

Figure 5.

(a): Modal probabilities. The black, grey, and white regions are where components 3, 2, and 1 are favoured, respectively.

(b), (c), (d): Contours of $P(z_f = i|y_{(T)}, z_{(T)}, y_{(U)}, y_f)$, for components $i = 1, 2$ and 3. The black, dark grey, light grey and white regions are where $P(z_f = i|y_{(T)}, z_{(T)}, y_{(U)}, y_f) > 0.90$, $0.50 < P(z_f = i|y_{(T)}, z_{(T)}, y_{(U)}, y_f) \leq 0.90$, $0.05 < P(z_f = i|y_{(T)}, z_{(T)}, y_{(U)}, y_f) \leq 0.50$, and $0.05 \geq P(z_f = i|y_{(T)}, z_{(T)}, y_{(U)}, y_f)$, respectively.

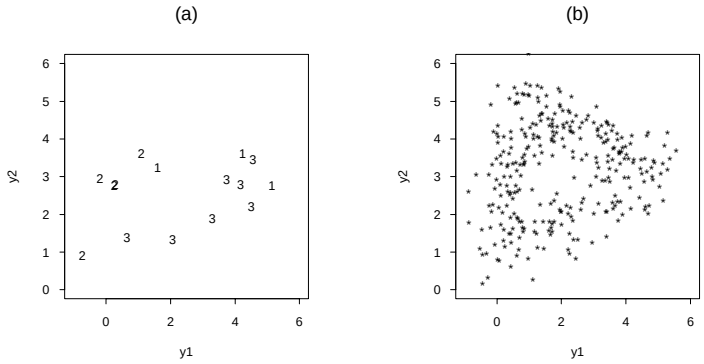

Figure 1

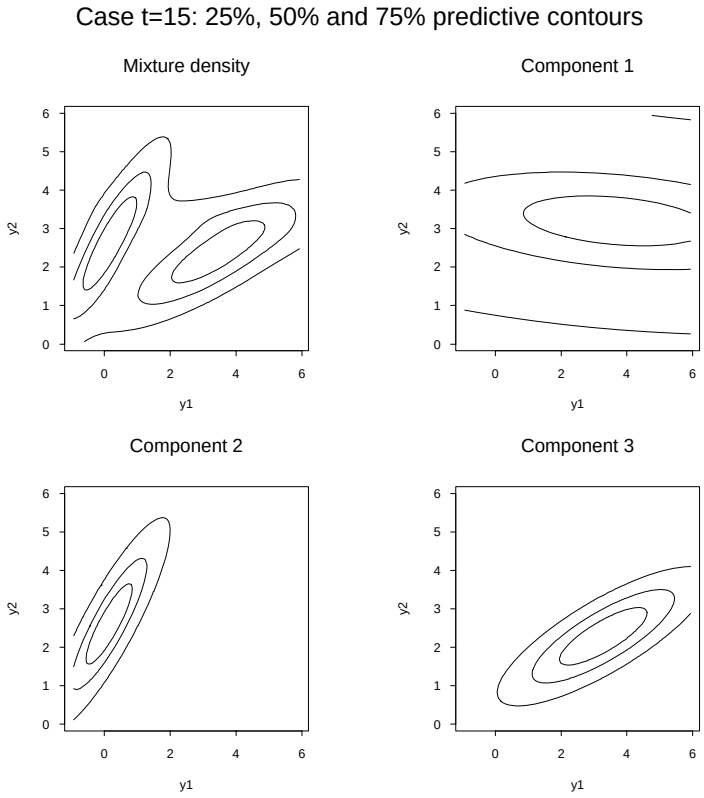
Case t=15: 25%, 50% and 75% predictive contours
Mixture density
Component 1
Component 2
Component 3
y1
y2

Figure 2

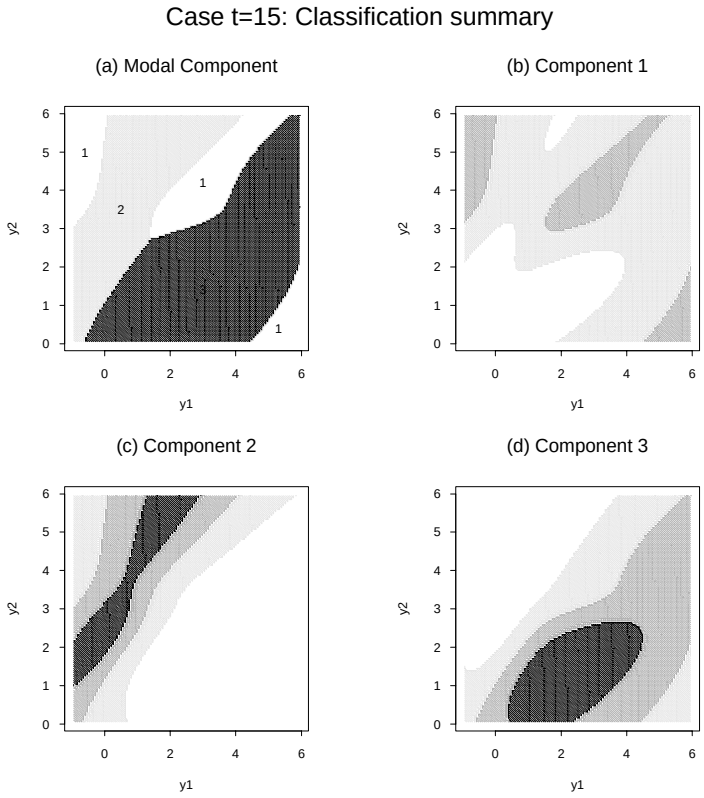
Case t=15: Classification summary
(a) Modal Component
(b) Component 1
(c) Component 2
(d) Component 3
y1
y2

Figure 3

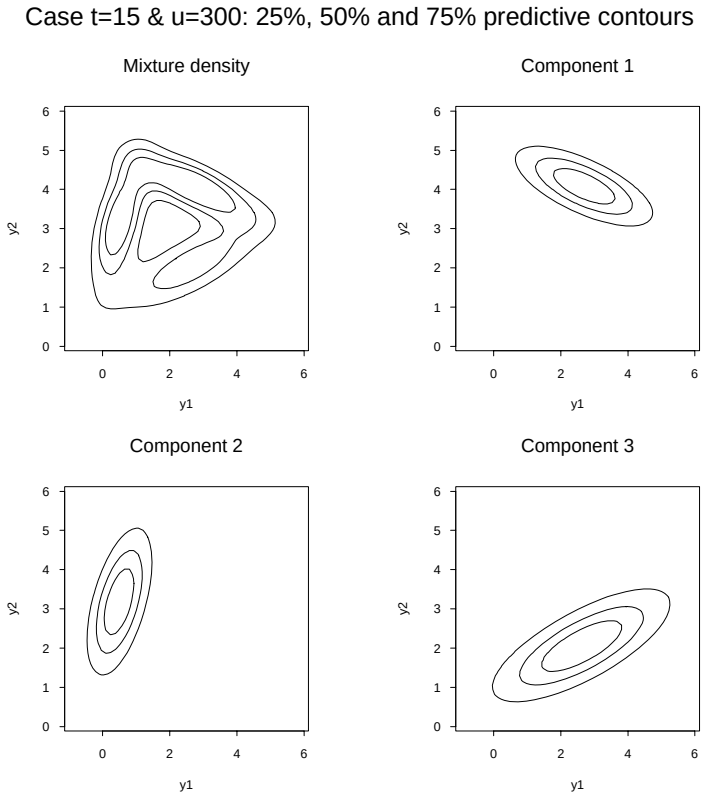
Case t=15 & u=300: 25%, 50% and 75% predictive contours
Mixture density
Component 1
Component 2
Component 3
y1
y2
0
1
2
3
4
5
6

Figure 4

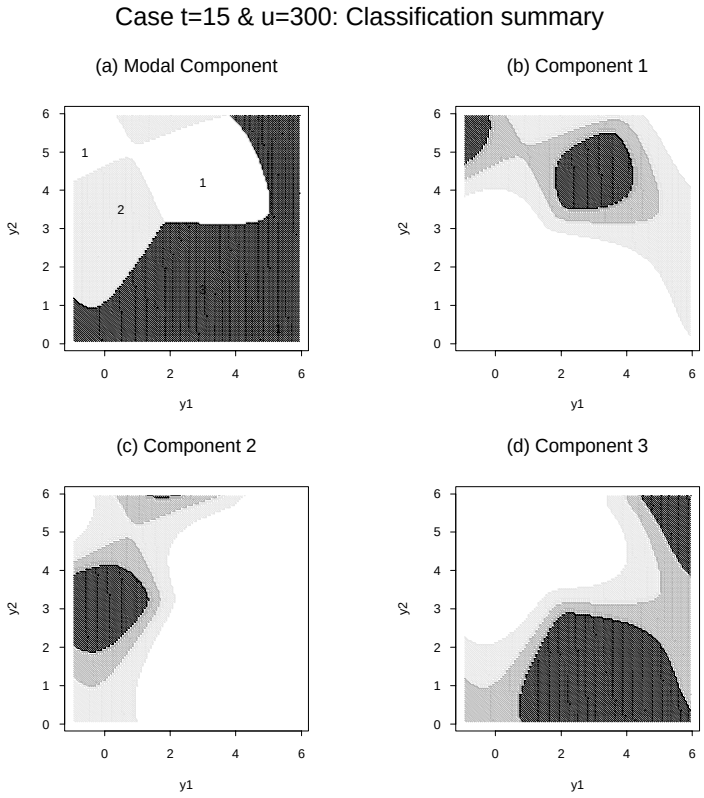
Case t=15 & u=300: Classification summary
(a) Modal Component
(b) Component 1
(c) Component 2
(d) Component 3
y1
y2
1
2
3

Figure 5