Bayesian Approach to Mixture Models for Discrimination

Keith Copsey and Andrew Webb

DERA Malvern, St. Andrews Road, Malvern, WR14 3PS.
kcopsey@signal.dera.gov.uk, webb@signal.dera.gov.uk

Abstract. This paper develops a Bayesian mixture model approach to discrimination. The specific problem considered is the classification of mobile targets, from Inverse Synthetic Aperture Radar images. However, the algorithm developed is relevant to the generic classification problem. We model the data measurements from each target as a mixture distribution. A Bayesian formalism is adopted, and we obtain posterior distributions for the parameters of our mixture models. The distributions obtained are too complicated for direct analytical use in a classifier, so a Markov chain Monte Carlo (MCMC) algorithm is used to provide samples from the distributions. These samples are then used to make classifications of future data.

Keywords. Bayesian inference, Discrimination, Inverse Synthetic Aperture Radar, Markov chain Monte Carlo, Mixture models, Target recognition.

1 Introduction

This paper describes a Bayesian mixture model approach to discrimination[14]. The generic discrimination problem that we consider is one where we are given a set of training data consisting of class labelled measurements (and possibly some unlabelled measurements), and then want to assign a previously unseen object to one of the classes, on the basis of the measurements made of that object. Specifically, in this paper, we illustrate the approach by considering automatic target recognition (ATR) of Inverse Synthetic Aperture Radar (ISAR) images from 3 main classes of mobile targets.

The mixture model approach to ATR aims to initially provide, for measurement data x, and classes j, estimates of the class-conditional probability densities of the data, $p(x|j)$. We can then produce estimates for the posterior probabilities of class membership, $p(j|x)$, using Bayes' theorem, $p(j|x) \propto p(x|j)p(j)$, where $p(j)$ are the prior class probabilities.

Estimating the posterior probabilities of class membership offers a number of advantages over producing class membership decisions only. These advantages include giving a measure of confidence for our class predictions, and the ready ability to combine the probabilities with additional information, such as intelligence reports. Furthermore, since after classifying a target we need to decide upon a course of action, we can incorporate the probabilities into a multilevel

F.J. Ferri et al. (Eds.): SSPR&SPR 2000, LNCS 1876, pp. 491–500, 2000.

model that reflects the whole decision making process. For instance, by considering the expected posterior loss of decisions, we can take into account the different costs involved in making a classification.

Estimation of the class-conditional probability densities, $p(x|j)$, is complicated by the high-dimensionality of our radar target data. Non-parametric methods of density estimation, such as kernel-based methods, would require unrealistically large amounts of training data for accurate density estimates[13], while parametric methods, such as a simple Gaussian classifier, might impose a specific form on the density that is too rigid for the problem. Mixture models provide a compromise between the two methods; attempting to provide enough flexibility to accurately estimate the densities, but imposing enough structure that we can train with realistic amounts of data.

Further motivations for the mixture model approach[4,15] in this application arise from the following observations:

1. The probability density function of the radar returns for a single target can be expressed as an integral over the angle of illumination of a conditional density of a simple form (*e.g.* Gaussian, gamma), with the mixture distribution arising as the approximation of this integral by a finite sum.
2. Additional effects, such as robustness to offsets in position, and amplitude scaling, can be readily incorporated into such a model.

Previous work[15] uses the Expectation-Maximisation (EM) algorithm to estimate the parameters of class distributions that are gamma mixture models. Hastie and Tibshirani[9] use an EM-algorithm on class distributions that are Gaussian mixture models, making the assumption of a common covariance matrix across all the mixture components. Laskey[10] formulates a Bayesian approach to modelling classes as mixtures of Gaussian distributions, but uses the EM algorithm to estimate the maximum *a-posteriori* parameter values only. The approach presented here is a generalisation of the work of Lavine and West[11], who look at a Bayesian approach to classification, where each class is distributed as a single multivariate Gaussian.

2 The Bayesian Mixture Model Approach

2.1 Introduction and Notation

We consider classification of an object into one of J distinct classes, on the basis of a d-dimensional data measurement of that object. The probability density function for the d-dimensional data, x, can be written as:

$$p(x) = \sum_{j=1}^{J} \theta_j p(x|j) \ , \tag{1}$$

where $\theta = (\theta_1, \ldots, \theta_J)$ is a vector of the prior classification probabilities for each class, with components satisfying $\sum_{j=1}^{J} \theta_j = 1$, and $p(x|j)$ is the class-conditional probability density for data from class j.

The class-conditional densities are also modelled by mixture distributions, with the j-th class having R_j components, which we refer to as subclasses:

$$p(x|j) = \sum_{r=1}^{R_j} \pi_{j,r} p(x|j,r) \ . \tag{2}$$

$\pi_j = (\pi_{j,1} \ldots, \pi_{j,R_j})$ represents the prior subclass probabilities within class j; *i.e.* $\pi_{j,r}$ is the mixing probability for the r-th subclass of the j-th class, satisfying $\sum_{r=1}^{R_j} \pi_{j,r} = 1$. We denote the complete set by $\pi = \{\pi_j, 1 \leq j \leq J\}$.

The distribution $p(x|j,r)$ represents the probability density of the data within a given subclass r, of a class j. We make the initial assumption that we have independence between the components of the data vector x, conditioned on the class and subclass. For ISAR images this corresponds to making an assumption of independence between the Radar Cross-Section fluctuations of any pair of pixels[6]. Note that this independence assumption for each component does not extend to an independence assumption for the mixture distribution as a whole. We take Gaussian forms for these distributions, with means $\mu_{j,r,l}$ and variances $\sigma^2_{j,r,l}$, where $l = 1, \ldots, d$. For a given class and subclass these are represented by the vectors $\mu_{j,r}$ and $\Sigma_{j,r}$. The sets of all means and variances are represented by μ and Σ, respectively.

We have n observed d-dimensional independent training data samples $y = \{y_1, \ldots, y_n\}$, and introduce two types of classification variable for these data. The overall class allocation variables are denoted $Z = (Z_1, \ldots, Z_n)$, and the subclass classification variables, $z = (z_1, \ldots, z_n)$. These are such that $(Z_i = j, z_i = r)$ implies that the observation indexed by i is modelled to be drawn from subclass r of class j. Z_i is known for our labelled training data, but unknown otherwise. z_i will always be unknown, and is physically unimportant.

2.2 Our Approach

The Bayesian approach to estimating the parameters of mixture models offers a number of advantages over methods based on maximum likelihood, such as the EM algorithm. Not least is the elimination of the problem of unboundedness of the likelihood function, that is frequently ignored in maximum likelihood techniques. There are also the standard arguments in favour of Bayesian techniques, such as the ability to cope with additional prior information, perhaps elicited from expert knowledge, and the production of confidence intervals for the parameters estimated. There is also the potential for using hyper-parameters in our prior distributions for the mixture model parameters, to account for differences between training and test data, such as variations in the vehicle fit, or different types of vehicle from the same generic class.

In the work presented here, the number of components to use in each class mixture distribution, has still to be addressed in a full Bayesian manner. At the moment we hold the number of subclasses fixed throughout the algorithm. Reversible jump Markov chain Monte Carlo techniques[12] would be extremely complicated due to the high-dimensionality of our data.

2.3 Model Details

Prior distributions. The complete prior distribution for the mixture model parameters is:

$$p(\mu, \Sigma, \theta, \pi) = p(\mu, \Sigma)p(\theta) \prod_{j=1}^{J} p(\pi_j) \ . \tag{3}$$

As well as making an assumption of independence between the components that make up the vectors $\mu_{j,r}$ and $\Sigma_{j,r}$, we also make the assumption that $(\mu_{j,r}, \Sigma_{j,r})$ are mutually independent over all classes and subclasses. The components $\mu_{j,r,l}$ and $\sigma^2_{j,r,l}$ are given independent normal-inverse gamma priors:

$$(\mu_{j,r,l}|\sigma^2_{j,r,l}) \sim N\left(m_{j,r,l,0}, \sigma^2_{j,r,l}/h_{j,r,l,0}\right) , \quad \sigma^2_{j,r,l} \sim 1/\mathrm{Ga}(\nu_{j,r,l,0}, V_{j,r,l,0}) \ , \tag{4}$$

for fixed means $m_{j,r,l,0}$, precision parameters $h_{j,r,l,0}$, degrees of freedom $\nu_{j,r,l,0}$ and scale parameters $V_{j,r,l,0}$. The inverse gamma distribution is parameterised so that the expectation is $V_{j,r,l,0}/(\nu_{j,r,l,0} - 1)$.

The values of these hyper-parameters are partially chosen with the aid of the training data, and for our specific application also the known angles of illumination for the labelled data, giving a combination of priors from expert knowledge, and data dependent priors.

We further assume independence of the priors for θ and π_j, $j = 1, \ldots, J$. For both cases we take Dirichlet priors, with $\theta \sim D(a_0)$, where $a_0 = (a_{1,0}, \ldots, a_{J,0})$, and $\pi_j \sim D(b_{j,0})$, where $b_{j,0} = (b_{j,1,0}, \ldots, b_{j,R_j,0})$. The hyper-parameters a_0 and $b_{j,0}$ are held fixed.

The posterior distribution. The likelihood function for the problem is written:

$$p(y|\mu, \Sigma, \theta, \pi) = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{J} \theta_j \sum_{r=1}^{R_j} \pi_{j,r} \prod_{l=1}^{d} N(y_{i,l}; \mu_{j,r,l}, \sigma^2_{j,r,l}) \right\} \ . \tag{5}$$

Bayes' rule gives the following relationship between the posterior, prior and likelihood:

$$p(\mu, \Sigma, \theta, \pi|y) \propto p(y|\mu, \Sigma, \theta, \pi)p(\mu, \Sigma, \theta, \pi) \ , \tag{6}$$

which due to the multiplication of summations in the likelihood function, gives a posterior distribution on which exact analytical inference cannot be made.

In particular, calculation of the normalisation constant is computationally infeasible, as are calculations of various statistics of interest, such as the means and variances of the parameters. To maintain a full Bayesian approach to the problem, we propose to draw samples from the posterior distribution. Since we cannot sample directly from the distribution, we use a Markov chain Monte Carlo (MCMC) algorithm[7], known as a Gibbs sampler[2].

3 MCMC Algorithm

3.1 General

The Gibbs sampler is a means of sampling from a distribution that is too complicated for direct sampling, but which is such that we can split the variables into conditional distributions, each of which can be sampled from. To make use of the Gibbs sampler in our problem, we extend our set of random variables $(\mu, \Sigma, \pi, \theta)$ to include the allocation variables (Z, z), and sample from the posterior distribution $p(\mu, \Sigma, \pi, \theta, Z, z|y)$. To do this we divide into three distinct groupings, (μ, Σ), (π, θ) and (Z, z), obtaining posterior probabilities for each group, that are conditional on the other two groups.

3.2 Conditional Distributions

The mixture components. Given the allocation variables (Z, z), we can make use of the fact that the data y consist of classified independent samples from the $k = R_1 + \cdots + R_J$ subclasses, giving:

$$p(\mu, \Sigma|y, \theta, \pi, Z, z) = p(\mu, \Sigma|y, Z, z) \ . \tag{7}$$

We define $G_{j,r} = \{i|(Z_i = j, z_i = r)\}$, the set of indices of data elements that have been assigned to subclass r of class j, and $g_{j,r}$ to be the cardinality of $G_{j,r}$. We also define $\bar{y}_{j,r,l} = \frac{1}{g_{j,r}} \sum_{i \in G_{j,r}} y_{i,l}$, and $S_{j,r,l} = \sum_{i \in G_{j,r}} (y_{i,l} - \bar{y}_{j,r,l})^2$. Our independent normal-inverse gamma priors then give rise to independent normal-inverse gamma posterior distributions[4,5]:

$$\mu_{j,r,l}|(\sigma^2_{j,r,l}, y, Z, z) \sim N\left(m_{j,r,l}, \sigma^2_{j,r,l}/h_{j,r,l}\right) \ , \tag{8}$$

and:

$$\sigma^2_{j,r,l}|(y, Z, z) \sim 1/\text{Ga}(\nu_{j,r,l}, V_{j,r,l}) \ , \tag{9}$$

where:

$$\begin{aligned} h_{j,r,l} &= h_{j,r,l,0} + g_{j,r} \ , \\ m_{j,r,l} &= (h_{j,r,l,0} m_{j,r,l,0} + g_{j,r} \bar{y}_{j,r,l})/h_{j,r,l} \ , \\ \nu_{j,r,l} &= \nu_{j,r,l,0} + g_{j,r}/2 \ , \\ V_{j,r,l} &= V_{j,r,l,0} + S_{j,r,l}/2 + h_{j,r,l,0} g_{j,r} (\bar{y}_{j,r,l} - m_{j,r,l,0})^2/(2h_{j,r,l}) \ . \end{aligned} \tag{10}$$

The allocation probabilities. Given the allocation variables (Z, z), the class and subclass allocation probabilities, (θ, π), will be independent of (y, μ, Σ). Thus we have:

$$p(\theta, \pi|y, \mu, \Sigma, Z, z) = p(\theta|Z) p(\pi|Z, z) \ . \tag{11}$$

For the class allocation probabilities, defining $g_j = \sum_{r=1}^{R_j} g_{j,r}$, we have:

$$\theta|Z \sim D(a), \quad \text{where } a = (a_1, \ldots, a_J), \text{ with } a_j = g_j + a_{j,0} \ . \tag{12}$$

For the subclass allocation probabilities we obtain the following independent distributions, for $j = 1, \ldots, J$:

$$\pi_j|(Z, z) \sim D(b_j) \quad \text{where } b_j = (b_{j,1}, \ldots, b_{j,R_j}), \text{ with } b_{j,r} = g_{j,r} + b_{j,r,0} \ . \tag{13}$$

The allocation variables. Since the data y_i are conditionally independent given $(\mu, \Sigma, \theta, \pi)$, the pairs of allocation variables (Z_i, z_i) are conditionally independent given $(y, \mu, \Sigma, \theta, \pi)$, so:

$$p(Z, z|y, \mu, \Sigma, \theta, \pi) = \prod_{i=1}^{n} \{p(z_i|Z_i, y_i, \mu, \Sigma, \theta, \pi)p(Z_i|y_i, \mu, \Sigma, \theta, \pi)\} \ . \tag{14}$$

If Z_i is unknown for the i-th data vector, we have:

$$p(Z_i = j|y_i, \mu, \Sigma, \theta, \pi) \propto \theta_j \sum_{r=1}^{R_j} \pi_{j,r} p(y_i|\mu_{j,r}, \Sigma_{j,r}) \ , \tag{15}$$

For the subclass allocation variable z_i, we have:

$$p(z_i = r|Z_i = j, y_i, \mu, \Sigma, \theta, \pi) \propto \pi_{j,r} p(y_i|\mu_{j,r}, \Sigma_{j,r}) \ . \tag{16}$$

3.3 Algorithm Specifics

To start our algorithm, we take initial allocation vectors, $Z^{(0)} = (Z_1^{(0)}, \ldots, Z_n^{(0)})$, and $z^{(0)} = (z_1^{(0)}, \ldots, z_n^{(0)})$. Some or all of the elements of vector Z are actually known, from our labelled training data, in which case we use these known values. We describe the algorithm in terms of the i-th iteration, which updates the set of parameters and allocation variables $(\mu^{(i-1)}, \Sigma^{(i-1)}, \theta^{(i-1)}, \pi^{(i-1)}, Z^{(i-1)}, z^{(i-1)})$, from the end of the $(i-1)$-th iteration, to $(\mu^{(i)}, \Sigma^{(i)}, \theta^{(i)}, \pi^{(i)}, Z^{(i)}, z^{(i)})$:

1. Draw a sample $(\mu^{(i)}, \Sigma^{(i)})$ from $p(\mu, \Sigma|y, Z^{(i-1)}, z^{(i-1)})$ using (9) and (8). This gives an updated set of parameters for the subclass distributions.
2. Sample $(\theta^{(i)}, \pi^{(i)})$ from $p(\theta, \pi|y, \mu^{(i)}, \Sigma^{(i)}, Z^{(i-1)}, z^{(i-1)})$, using (12) and (13). This gives an updated set of the class and subclass allocation probabilities.
3. Sample $(Z^{(i)}, z^{(i)})$ from $p(Z, z|y, \mu^{(i)}, \Sigma^{(i)}, \theta^{(i)}, \pi^{(i)})$, using (15) and (16), giving an updated set of class and subclass allocation variables. Note that for our class labelled data, we set the class allocation variables to their known values, and only re-estimate the corresponding subclass allocation variables.

After an initial burn-in period, during which the generated Markov chain reaches equilibrium, the set of parameters $(\mu^{(i)}, \Sigma^{(i)}, \theta^{(i)}, \pi^{(i)}, Z^{(i)}, z^{(i)})$, can be regarded as dependent samples from the posterior distribution $p(\mu, \Sigma, \theta, \pi, Z, z|y)$. To obtain approximately independent samples we leave a gap, known as the decorrelation gap, between successive samples (*i.e.* we only retain a sample every l-th iteration of the algorithm, where l is an integer greater than one). If we are only concerned with ergodic averages, we actually obtain better variances if we do not sub-sample the output of our Markov chain. However, if storage of the samples is an issue, we may like to leave a decorrelation gap, so that we can be sure to explore the full space of the distribution, without having to keep thousands of samples.

3.4 Classification of the Observations

The training data. We now obtain formulae for the posterior classification probabilities of the training data (*i.e.* already observed measurements, y), remembering that some of this training data may be unlabelled in class. For notational ease we let D denote the combination of our data measurements y, and any known class allocations for this data.

Rather than approximating the posterior distributions directly, as follows:

$$p(\hat{Z}_i = j|D) \approx \frac{1}{N}\sum_{m=1}^{N} \mathrm{I}(Z_i^{(m)} = j) \ , \tag{17}$$

where N is the number of MCMC samples, we use Rao-Blackwellisation[3] to provide more efficient estimates; by using an approximation to the posterior marginalised distribution $p(Z|D)$, based on our MCMC sampled values:

$$p(Z|D) \approx \frac{1}{N}\sum_{s=1}^{N} p(Z|y, \mu^{(s)}, \Sigma^{(s)}, \theta^{(s)}, \pi^{(s)}) \ , \tag{18}$$

where using (15) we have:

$$p(Z_i = j|y_i, \mu^{(s)}, \Sigma^{(s)}, \theta^{(s)}, \pi^{(s)}) \propto \theta_j^{(s)} \sum_{r=1}^{R_j} \pi_{j,r}^{(s)} p(y_i|\mu_{j,r}^{(s)}, \Sigma_{j,r}^{(s)}) \ . \tag{19}$$

Future observations. We now consider classifying a previously unseen observation, y_f, by looking at the posterior probabilities:

$$P(Z_f = j|D, y_f) \propto P(Z_f = j|D)p(y_f|D, Z_f = j) \ , \tag{20}$$

where:

$$P(Z_f = j|D) = E(\theta_j|D) \approx \frac{1}{N}\sum_{s=1}^{N} E(\theta_j|D, Z^{(s)}) \ , \tag{21}$$

and:

$$p(y_f|D, Z_f = j) \approx \frac{1}{N}\sum_{s=1}^{N} p(y_f|D, Z_f = j, Z^{(s)}, z^{(s)}) \ . \tag{22}$$

If we have reason to believe that the spread of future data between the different classes is likely to be different to that in the training data, we can replace the expressions for $P(Z_f = j|D)$ in (21) with our modified beliefs. We write $p(y_f|D, Z_f = j, Z^{(s)}, z^{(s)})$ as a mixture distribution:

$$\begin{aligned} p(y_f|D, Z_f = j, Z^{(s)}, z^{(s)}) = \sum_{r=1}^{R_j} \{ & p(z_f = r|D, Z_f = j, Z^{(s)}, z^{(s)}) \\ & \times p(y_f|D, Z_f = j, z_f = r, Z^{(s)}, z^{(s)})\} \ , \end{aligned} \tag{23}$$

where:

$$p(z_f = r|D, Z_f = j, Z^{(s)}, z^{(s)}) = E(\pi_{j,r}|D, Z^{(s)}, z^{(s)}) \ , \tag{24}$$

and $p(y_f|D, Z_f = j, z_f = r, Z^{(s)}, z^{(s)})$ is the predictive density, for data drawn from subclass r of class j, using component distributions determined by the MCMC sample outputs, $(Z^{(s)}, z^{(s)})$. Some calculations[4], show that this predictive density is given by a product of independent Student-t distributions, with the l-th component having $2\nu^{(s)}_{j,r,l}$ degrees of freedom, location parameter $m^{(s)}_{j,r,l}$, and scale parameter $\sqrt{V^{(s)}_{j,r,l}(h^{(s)}_{j,r,l} + 1)/(h^{(s)}_{j,r,l}\nu^{(s)}_{j,r,l})}$. The parameters being defined in (10) using allocation variables $(Z^{(s)}, z^{(s)})$.

4 Experimental Results

We illustrate the use of the algorithm on real ISAR data (see Fig. 1), consisting of images of vehicles from 3 main types of battlefield target, which we denote by classes 1, 2 and 3. Our training data consist of approximately equal amounts of images from each of the 3 classes (about 2000 per class), collected over single complete rotations of the vehicles, at a constant depression angle. Our test data consist of 6 sets of approximately 400 ISAR images, collected from single complete rotations of 6 vehicles. Of these, datasets B-hd and B-er are the vehicle from dataset B imaged at a higher depression angle and with the engines running respectively, while the remaining sets correspond to different vehicles within the same generic class type. Unfortunately, we do not have an independent set of test data for class 2, negating the possibility of obtaining a meaningful single measure of performance from our test data.

Our ISAR data is, after some initial pre-processing, 38 pixels in range by 26 pixels in cross-range, giving an overall dimensionality of 988. To reduce this down to a slightly more manageable level, a principal components analysis[14] has been conducted on the data, and we actually use only the first 35 linear principal components of each image vector. The algorithm has been run with 12 subclasses per class, to draw 1000 samples with a decorrelation gap of 10 iterations, after a burn-in period of 10000 iterations.

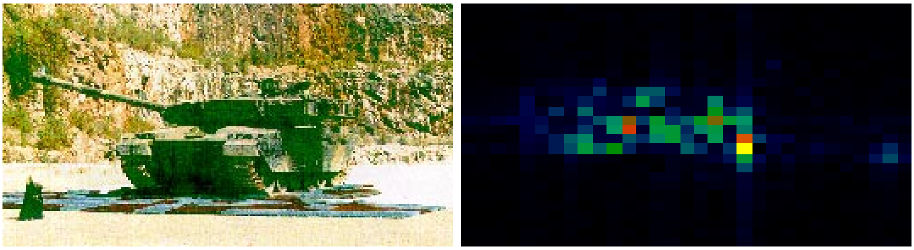

Fig. 1. Typical vehicle and ISAR image from our data set.

Table 1. Classification rates for training datasets.

Data set	True class	Predicted class		
		1	2	3
B	1	99.1	0.6	0.3
A	2	0.6	98.2	1.1
C	3	0.0	1.1	98.8

Table 2. Classification rates for test datasets.

Data set	True class	Predicted class		
		1	2	3
B-hd	1	62.6	22.7	14.8
B-er	1	98.6	1.2	0.2
D	1	71.1	12.5	16.5
E	1	83.6	1.7	14.7
F	3	7.9	55.2	36.8
G	3	37.6	26.9	35.5

Table 1 documents the classification rates for the training data, and Table 2 for the test data. In both cases the classifications have been made by taking the class which gives the largest posterior probability. A full set of results for our algorithm, including a treatment of uncertainty in position of the vehicle in the image, along with a comparison with other classification techniques[14], is given in [4]. The limited sets of results given here, show that we have been able to train the classifier well on the training data (greater than 98% classifier accuracy), and show the performance extending well to test data, when the vehicle and imaging conditions are similar. However the extension to classifying different vehicles to the same generic class, proves problematical, as illustrated by data sets F and G in particular. In part this is due to there sometimes being more similarity between two vehicles from different classes, than there is between two vehicles within the same class. The comparisons in [4] show the Bayesian mixture model technique to compare favourably with other classifiers, including a mixture model classifier based on the EM-algorithm.

It should be noted, however, that classification rate does not give a true indication of the overall performance of a classifier[8]. Not least is the fact that it treats all misclassifications with equal weight. Where we have overlap between classes, we will never be able to obtain perfect classification, and a good classifier would indicate uncertainty between the classes for data in the overlapping region. Thus rather than a classification rate, an assessment of the accuracy of our estimates of the posterior probabilities is desirable. However, such an assessment is extremely difficult for real data. As well as class overlap there is also the issue of the possibly different misclassification costs involved[1].

5 Summary and Discussion

In this paper we have developed a Bayesian mixture model approach to discrimination. We have modelled the class-conditional densities as Gaussian mixtures, and conducted a Bayesian analysis under the assumption of constant model order of the mixtures. The use of the algorithm has been demonstrated on real datasets, consisting of ISAR images of mobile targets. On this data we have attempted the difficult task of classifying vehicles into generic classes, rather than specific examples of a class. Future work will need to address issues such as whether this is actually feasible, the number of components per mixture (a model selection problem), alternative methods for assessing the performance, and the use of different component distributions in our mixture models, such as gamma distributions. However, this work has established that the technique is a viable and sound method for discrimination when test conditions are similar (but not necessarily identical) to the training conditions.

6 Acknowledgments

This research was sponsored by the UK MOD Corporate Research programme TG10.

References

[1] N.M. Adams and D.J. Hand. Comparing classifiers when the misclassification costs are uncertain. *Pattern Recognition*, 7:1139–1148, 1999.

[2] G. Casella and E. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[3] G. Casella and C.P. Robert. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.

[4] K.D. Copsey and A.R. Webb. Bayesian approach to mixture models for target recognition. *In preparation*, 2000.

[5] M.H. DeGroot. *Optimal Statistical Decisions.* McGraw-Hill, 1970.

[6] W. Denton and A. Britton. The classification of vehicles in MMW imagery. *Proceedings of Symposium on Non-cooperative airborne target identification using RADAR, NATO SCI Panel*, 1998.

[7] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov chain Monte Carlo in Practice.* Chapman and Hall, 1996.

[8] D.J. Hand. *Construction and Assessment of Classification Rules.* John Wiley, Chichester, 1997.

[9] T. Hastie and R. Tibshirani. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society, Series B*, 58:155–176, 1996.

[10] K.B. Laskey. A Bayesian approach to clustering and classification. *Proceedings of IEEE Int. Conf. Syst. Man and Cybernetics, Charlottesville, Va*, pages 179–183, October 1991.

[11] M. Lavine and M. West. A Bayesian method for classification and discrimination. *Canadian Journal of Statistics*, 20:451–461, 1992.

[12] S. Richardson and P.J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, 59:731–792, 1997.

[13] B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, 1986.

[14] A.R. Webb. *Statistical Pattern Recognition.* Arnold, London, 1999.

[15] A.R. Webb. Gamma mixture models for target recognition. *Pattern Recognition*, to appear, 2000.