Bayesian Estimation and Classification with Incomplete Data Using Mixture Models

Jufen Zhang
J.Zhang@exeter.ac.uk

Department of Computer Science
University of Exeter
Exeter, UK.

Richard Everson
R.M.Everson@exeter.ac.uk

Abstract

Reasoning from data in practical problems is frequently hampered by missing observations. Mixture models provide a powerful general semi-parametric method for modelling densities and have close links to radial basis function neural networks (RBFs). In this paper we extend the Data Augmentation (DA) technique for multiple imputation to Gaussian mixture models to permit fully Bayesian inference of the mixture model parameters and estimation of the missing values. The method is illustrated and compared to imputation using a single normal density on synthetic data and real-world data sets. In addition to a lower mean squared error, mixture models provide valuable information on the potentially multi-modal nature of imputed values, and by modelling the missing data more accurately, so that higher classification rates can be achieved compared with simple mean imputation methods. The DA formalism is extended to a classifier closely related to RBF networks to permit Bayesian classification with incomplete data; the technique is illustrated on synthetic and real-world datasets. This efficient technology enables us to perform Bayesian imputation, parameter estimation and classification simultaneously for data with missing values.

1 INTRODUCTION

Measured data are frequently marred by missing values. When data are plentiful it may be sufficient to discard the incomplete observations, but utilising all the available information for learning and inference is generally important, and it is often necessary to classify or predict an outcome from incomplete predictors. For example, trauma room or intensive care unit medical data, such as blood pressure, heart rate, injury type, etc., collected in extremis are often incomplete, but it is necessary to make predictions from these data.

Many methods for filling in, or imputing, missing values have been developed (see [Little and Rubin, 2002] for a comprehensive treatment); simple methods are to replace missing values by the mean of the observed values or to regress missing values from the observed data. Maximum likelihood learning via the Expectation-Maximisation (EM) algorithm [Dempster et al., 1977], in which the missing observations are regarded as hidden variables, permits inference of missing values and takes account of the additional uncertainty in parameters caused by missing observations [Ghahramani and Jordan, 1994].

In a Bayesian framework the Data Augmentation (DA) algorithm, introduced by Tanner and Wong [1987], is the natural analogue of the EM algorithm; it amounts to Gibbs sampling from the joint posterior distribution of the parameters and the missing values. Since many samples are drawn for the missing variables DA is a multiple imputation (MI) technique [Rubin, 1987]. The DA algorithm has been widely used for missing value imputation under the assumption of a normal model [Schafer, 1997]. In this paper we use data augmentation for the inference of missing values and parameters of mixture models. Mixture models are well known as a flexible semi-parametric density model, capable of modelling a wide range of densities. Diebolt and Robert [1994] developed a Gibbs sampling scheme for sampling from the posterior parameter distribution of uni-dimensional mixture models, which we extend to the multi-dimensional mixtures in order to incorporate DA. Mixture models are closely related to radial basis function (RBF) neural networks. In a similar manner to Tråvén [1991] and Sykacek [2000], for classification problems with missing data we utilise a mixture model with separate mixing coefficients for each class to model class conditional densities. The mixture model structure permits the easy incorporation of DA and allows class membership information to influence the imputation of missing values.

This paper is organized as follows. In section 2 we briefly review the Data Augmentation algorithm and describe how it is applied to the single multivariate normal model which forms the basic building block for the mixture models described in section 3. Illustrative results for mixture models are also presented in section 3, after which the use of mixture models as the basis for classifiers is described and illustrated in section 4. The paper concludes with a brief discussion.

2 DATA AUGMENTATION

We partition each datum $\mathbf{x} = (x_1, \ldots, x_p)^T$ into the observed components $\mathbf{x}_{obs}$ and the (possibly empty) missing components $\mathbf{x}_{mis}$. The goal of Bayesian inference with missing data is to describe the joint posterior distribution $p(\boldsymbol{\theta}, X_{mis} \,|\, X_{obs})$ of the model parameters $\boldsymbol{\theta}$ and the missing values X_{mis} having observed the data X_{obs}. By X_{obs} and X_{mis} we mean all the observed and missing values respectively.

The DA algorithm [Tanner and Wong, 1987] uses Gibbs sampling to draw samples $\{\boldsymbol{\theta}^{(t)}, X_{mis}^{(t)}\}$ from the joint distribution: In the Imputation step missing values are imputed by simulating from the conditional predictive distribution of X_{mis} based on the observed data, $p(X_{mis} \,|\, X_{obs}, \boldsymbol{\theta}^{(t-1)})$. In the Posterior step the data, completed with the simulated $X_{mis}^{(t)}$, is used to draw new parameters $\boldsymbol{\theta}^{(t)}$ from the posterior distribution $p(\boldsymbol{\theta} \,|\, X_{mis}^{(t)}, X_{obs})$. After the usual burn-in period to ensure convergence, the draws $\{\boldsymbol{\theta}^{(t)}, X_{mis}^{(t)}\}$ form a Markov chain whose stationary distribution is $p(\boldsymbol{\theta}, X_{mis} \,|\, X_{obs})$. The imputed samples $X_{mis}^{(t)}$ may be histogrammed to show the posterior distribution of the imputed samples, or $\{\boldsymbol{\theta}^{(t)}, X_{mis}^{(t)}\}$ can be used for Monte Carlo integration for predictive purposes.

2.1 Single multivariate normal

Schafer [1997] provides an extensive discussion, which we summarise here, of the DA algorithm for the imputation of missing values when the data are assumed to be independent and drawn from a normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

I-step. The t-th imputation step consists of making a draw of the missing values given the current parameters $\boldsymbol{\theta}^{(t)} = \{\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}\}$. It is straightforward to show that the distribution of missing values given the observed values is

$$\mathbf{x}_{mis} \sim \mathcal{N}(\boldsymbol{\mu}_{m|o}^{(t)}, \boldsymbol{\Sigma}_{mm|o}^{(t)}) \qquad (1)$$

where, omitting for clarity the (t) superscripts,

$$\boldsymbol{\mu}_{m|o} = \boldsymbol{\mu}_m + \boldsymbol{\Sigma}_{om}^T \boldsymbol{\Sigma}_{oo}^{-1} (\mathbf{x}_{obs} - \boldsymbol{\mu}_o) \qquad (2)$$

and

$$\boldsymbol{\Sigma}_{mm|o} = \boldsymbol{\Sigma}_{mm} - \boldsymbol{\Sigma}_{om}^T \boldsymbol{\Sigma}_{oo}^{-1} \boldsymbol{\Sigma}_{om} \qquad (3)$$

in which the mean and covariance are partitioned in the same manner as $\mathbf{x}$ with m and o subscripts denoting the components corresponding to missing and observed variables:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_o \\ \boldsymbol{\mu}_m \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{oo} & \boldsymbol{\Sigma}_{om} \\ \boldsymbol{\Sigma}_{mo} & \boldsymbol{\Sigma}_{mm} \end{pmatrix} \qquad (4)$$

P-step. The normal-inverted Wishart density is the conjugate prior for the mean and covariance matrix:

$$\boldsymbol{\Sigma} \sim W^{-1}(m, \boldsymbol{\Lambda}), \qquad \boldsymbol{\mu} \,|\, \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}_0, \tau^{-1}\boldsymbol{\Sigma}) \qquad (5)$$

where m, $\boldsymbol{\Lambda}$, $\boldsymbol{\mu}_0$ and $\tau > 0$ are hyperparameters. In the P-step, using the data completed with the draws in the I-step, new means $\boldsymbol{\mu}^{(t+1)}$ and covariance $\boldsymbol{\Sigma}^{(t+1)}$ are drawn from their normal-inverted-Wishart posterior for which the parameters are:

$$m' = m + N \qquad (6)$$

$$\boldsymbol{\Lambda}' = [\boldsymbol{\Lambda}^{-1} + N\mathbf{S} + \frac{\tau N}{\tau + N}(\overline{\mathbf{x}} - \boldsymbol{\mu}_0)(\overline{\mathbf{x}} - \boldsymbol{\mu}_0)^T]^{-1} \qquad (7)$$

$$\boldsymbol{\mu}_0' = \frac{N}{\tau + N}\overline{\mathbf{x}} + \frac{\tau}{\tau + N}\boldsymbol{\mu}_0 \qquad (8)$$

$$\tau' = \tau + N \qquad (9)$$

where N is the number of observations and $\overline{\mathbf{x}}$ and $\mathbf{S}$ are the sample mean and covariance respectively.

2.2 Example: Old Faithful data

As an illustrative example and for contrast with the mixture case, we show the results of imputation on the Old Faithful dataset [Hardle, 1991]. The data consists of 298 observations of 2-dimensional continuous variables: the duration (in minutes) of the eruption, and waiting time (in minutes) before the next eruption. 68 of the observations do not have a value for the waiting time until the next eruption.

Using non-informative priors, we learn posterior distributions for the mean and covariance of the distribution and impute 5000 samples from the posterior distribution for each of the missing values. Figure 1 shows the data, while Figure 2 shows histograms of the imputed values for 4 representative missing values. These results highlight the inadequacies of the single multivariate normal model for these data, which are clearly comprised of two clusters. The result of using a single normal model is rather diffuse unimodal imputations, reflecting the fact that the single Gaussian

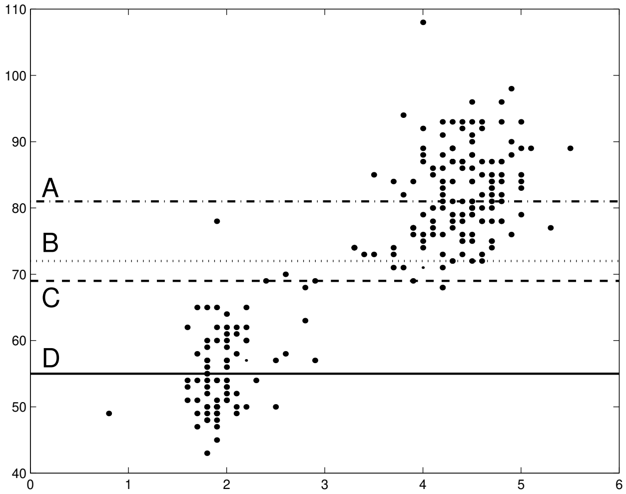

Figure 1: Old Faithful data: Waiting time between eruptions (ordinate) plotted versus eruption duration (abscissa). Dots mark the complete samples. Horizontal lines indicate the waiting time for 4 samples for which the duration is missing.

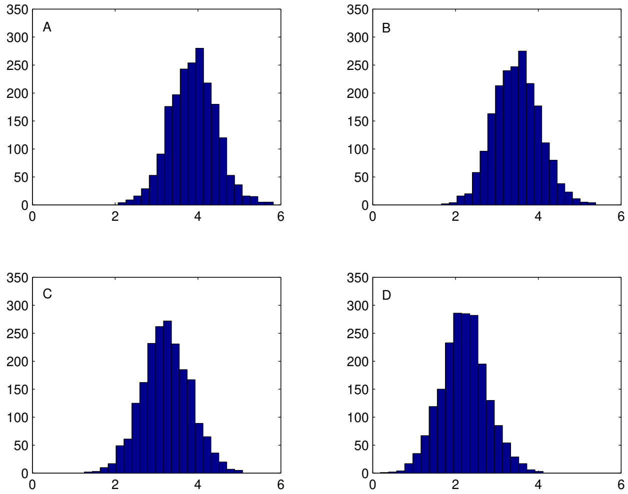

Figure 2: Histograms of the imputed eruption duration values for the 4 missing values indicated by horizontal lines in Figure 1, using a single multivariate normal model. Letters A-D refer to the horizontal lines in Figure 1.

is forced to account for both clusters. Note that the imputed values for points A and D show a bias towards higher values (A) and lower values (D), reflecting the orientation of the posterior covariance matrix, but there is no hint of multi-modality in the imputations for points B and C. These considerations lead us to consider imputation using mixture models.

3 DA FOR MIXTURES

Mixture models provide a very general semi-parametric model for arbitrary densities and they have been extensively investigated [see, for example, Titterington et al., 1985]. Their structure makes them amenable to Gibbs sampling, so that DA for missing values may be naturally incorporated into learning. In addition to unconditional density estimation, they may be simply modified to form classifiers with a structure similar to RBF neural networks. Maximum likelihood inference of mixture model parameter values is frequently carried out via the Expectation-Maximisation (EM) algorithm [Dempster et al., 1977], and Ghahramani and Jordan [1994] showed how to obtain maximum likelihood estimates for missing data within the EM framework. Diebolt and Robert [1994] described MCMC schemes for Bayesian inference of the parameters of unidimensional mixture model with a fixed number of components. Here we extend their work to multidimensional mixtures, permitting simultaneous inference of missing values.

We consider M-component mixture models,

$$p(\mathbf{x}) = \sum_{j=1}^{M} w_j p(\mathbf{x} \,|\, \boldsymbol{\theta}_j), \tag{10}$$

with non-negative mixing coefficients w_j summing to one, and component densities $p(\mathbf{x} \,|\, \boldsymbol{\theta}_j)$. We focus on the common Gaussian mixture model (GMM), so that the parameters of each component density are a mean $\boldsymbol{\mu}_j$ and a covariance matrix $\boldsymbol{\Sigma}_j$. Inference in mixture models is facilitated by the introduction of latent variables $\mathbf{z}_n$, which indicate which mixture component generated the observation $\mathbf{x}_n$; $z_{nj} = 1$ if the jth component generated $\mathbf{x}_n$, otherwise $z_{nj} = 0$.

For Gaussian mixture models, normal-inverse Wishart densities are the natural (conjugate) priors over the parameters of each component:

$$\boldsymbol{\Sigma}_j \sim \mathcal{W}^{-1}(m_j, \boldsymbol{\Lambda}_j) \tag{11}$$

$$\boldsymbol{\mu}_j \,|\, \boldsymbol{\Sigma}_j \sim \mathcal{N}(\boldsymbol{\mu}_{0j}, \tau_j^{-1}\boldsymbol{\Sigma}_j) \tag{12}$$

and we place a symmetric Dirichlet prior over the vector of weights, $\mathbf{w} = (w_1, \ldots, w_M)$:

$$\mathbf{w} \sim \mathcal{D}(\alpha, \ldots, \alpha) \tag{13}$$

With complete data, the Gibbs sampling scheme of Diebolt and Robert [1994] can be extended to multidimensional Gaussian mixture models. The t-th sweep of the sampling comprises draws as follows:

1. **Indicator variables**

$$\mathbf{z}_n^{(t)} \sim \mathcal{M}(h_{n1}^{(t)}, \ldots, h_{nM}^{(t)}) \tag{14}$$

where $\mathcal{M}$ denotes the multinomial-1 density and h_{nj} is the responsibility taken by the jth component for the $\mathbf{x}_n$:

$$h_{nj}^{(t)} = \frac{w_j \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{i=1}^{M} w_i \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \tag{15}$$

where w_j, $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ all refer to the samples from the $(t-1)$-st sweep.

2. **Component parameters** are drawn from their posterior densities, with the 'data' contributing to the jth component indicated by the n for which $z_{nj} = 1$. Expressions for the normal-inverse-Wishart parameters for each of the M components are given by (6) – (9), with N replaced by $N_j = \sum_n z_{nj}$ and

$$\overline{\mathbf{x}}_j = \frac{1}{N_j} \sum_n \mathbf{x}_n \delta_{n z_{nj}} \tag{16}$$

$$\mathbf{S}_j = \frac{1}{N_j} \sum_n (\mathbf{x}_n - \overline{\mathbf{x}}_j)(\mathbf{x}_n - \overline{\mathbf{x}}_j)^T \delta_{n z_{nj}} \tag{17}$$

where $\delta_{n z_{nj}}$ is the Kronecker delta.

3. **Mixing coefficients** are drawn from their posterior densities, which are also Dirichlet:

$$\mathbf{w}' \sim \mathcal{D}(\alpha + N_1, \ldots, \alpha + N_M) \tag{18}$$

These three steps form the P-step of a DA algorithm. In the corresponding I-step, the missing values are filled in by sampling from the component that, at the t-th step takes responsibility for the observation with the missing values. Thus if $\mathbf{x}_n$ has missing values, imputations are made from the conditional Gaussian distribution (1), in which the appropriate $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are the samples from the P-step with j such that $z_{nj} = 1$.

In the majority of the work reported here we have used non-informative priors for the parameters of the component distributions and the weights. However, it can also be helpful to adopt an empirical Bayes approach. Thus one may set weak priors on component means $\boldsymbol{\mu}_j$ to be equal to the overall mean of the data and weak spherical priors on the $\boldsymbol{\Lambda}_j$ so that the covariance of the dataset is shared out equally among the M components. It is also helpful to choose α to be slightly informative ($\alpha > 1$), encoding the belief that the mixture components carry responsibility for equal numbers of data points. Non-informative α may result in insufficient data points being associated with a component, in which case we adopt Diebolt and Robert's [1994] strategy and skip the Gibbs updates parameters of that component on this sweep. We initialise the Gibbs sampling Markov chain with K-means clustering, which itself is bootstrapped either by ignoring records with missing observations or using simple mean imputation. In any case we find that after a few sweeps memory of the initialisation is lost. In the results reported here we have used 500 or 100 burn-in sweeps.

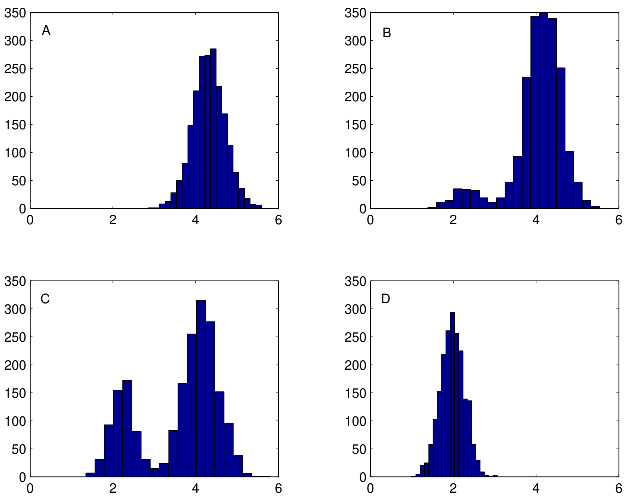

Figure 3: Histograms of the imputed eruption duration values for the 4 missing values indicated by horizontal lines in Figure 1 using a two component mixture model.

3.1 Example: Old Faithful data

We illustrate DA for mixture models on the Old Faithful data previously discussed in section 2.2. As we noted there, it is reasonable to model these data with a two-centre model, and the minimum description length (MDL) criterion [J.Rissanen, 1978] concurs with this. Figure 3 shows histograms of the imputed values for the data points indicated by the horizontal lines A-D. The imputations for points B and C are clearly multi-modal, reflecting the likelihood that the missing value in each case might have been generated by either cluster. In addition, the imputed values for points A and D are less dispersed than the imputations shown in Figure 2 because only one component takes responsibility for each of them.

3.2 Example: Synthetic and Iris data

We also generated a synthetic dataset consisting of 100 observations from a 3-centre mixture model.[1] The efficacy of the mixture model DA algorithm for different proportions of missing data was assessed using this dataset by deleting (at random) a proportion of the values and comparing the mean imputed values with the known original value. This synthesis of an incom-

[1] The model parameters were $\mathbf{w} = (0.5, 0.25, 0.25)$; $\boldsymbol{\mu}_1 = (0, -0.2)^T$, $\boldsymbol{\mu}_2 = (2, 2)^T$, $\boldsymbol{\mu}_3 = (2, -2)^T$; $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.625 & -0.217 \\ -0.217 & 0.875 \end{pmatrix}$; $\boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.224 & -0.137 \\ -0.137 & 0.976 \end{pmatrix}$; $\boldsymbol{\Sigma}_3 = \begin{pmatrix} 0.238 & 0.152 \\ 0.152 & 0.413 \end{pmatrix}$.

%	Mean		1 centre		3 centres	
10%	4.352	1.59	4.582	0.73	3.60	0.87
20%	4.279	1.08	4.387	0.44	3.78	0.74
30%	4.310	0.88	4.360	0.34	3.69	0.64
40%	4.300	0.76	4.425	0.33	3.80	0.53
50%	4.379	0.66	4.440	0.21	4.00	0.57

Table 1: Mean squared error per data point for mean imputation, single centre and three centre imputation. Columns show the MSE and standard deviation over 1000 (mean imputation) and 50 (1 & 3 centres) synthetic datasets.

plete dataset and imputation from it was repeated 50 times for each of the missing proportions, and we report averages over the 50 repeats.

Table 1 compares the mean squared error for mean imputation, imputation using a single Gaussian and mixture model imputation with three centres. It can be seen that the single centre DA is little better than mean imputation in this situation, although there is less variability in its results. Mixture model DA provides better imputation than either of the other methods, although there is significant variability between the synthesised datasets.

We find a similar pattern with incomplete datasets synthesised in the same manner from the famous four-dimensional, Fisher iris dataset [Fisher, 1936; Blake and Merz, 1998], which consists of 150 observations, 50 from each of three different classes. Figure 4 shows that the mean squared error per data point (over 10 repeats) for mixture model DA is slightly lower than for imputation from a single Gaussian.

Although, in both these cases, the MSE between the posterior mean imputed value and the 'true' value is lower for mixture model DA, we emphasise that comparing the mean of the posterior imputed value with the true value is rather insensitive and hides the principal advantage of the method, which is its ability to capture multi-modality in the imputed values.

4 CLASSIFICATION

Although unsupervised density modelling may be a goal in its own right, the task of many machine learning and statistical modelling efforts is to classify new data based on the knowledge of training data comprised of features $\mathbf{x}_n$ and the corresponding classes t_n. In order to take advantage of information residing in incomplete training data as well as to be able to classify incomplete observations, we model class conditional densities using mixtures. The mixture model foundation of these classifiers makes them amenable

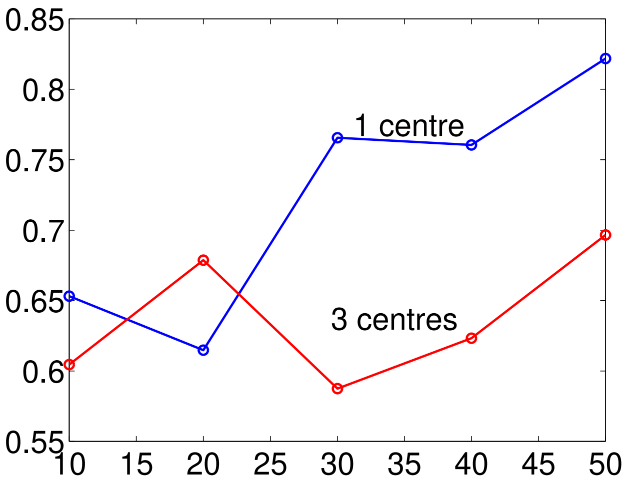

Figure 4: Fisher iris data. Mean squared error versus percentage of records with missing data using one and three centre imputation.

to Gibbs sampling and DA algorithms.

The classifiers are built on an architecture proposed by Tråvén [1991] and used by Sykacek [2000] for input feature selection. Its structure is akin to that RBF neural networks, and we note that Dybowski [1998] has proposed a similar mixture model for maximum likelihood based imputation.

The class conditional probability density for each class $k = 1, \ldots, K$ is modelled as:

$$p(\mathbf{x} \,|\, k) = \sum_{j=1}^{M} w_{kj} p(\mathbf{x} \,|\, \boldsymbol{\theta}_j) \tag{19}$$

Thus the mixture components $p(\mathbf{x} \,|\, \boldsymbol{\theta}_j)$ are common to all classes, but each class conditional probability is a separate linear combination of the components. The non-negative weights w_{kj} satisfy the constraint: $\sum_j w_{jk} = 1$ for each k.

The posterior probability for class k is therefore:

$$p(k \,|\, \mathbf{x}) = \frac{P_k \sum_{j=1}^{M} w_{kj} p(\mathbf{x} \,|\, \boldsymbol{\theta}_j)}{p(\mathbf{x})} \tag{20}$$

where P_k is the prior probability of class k, and $p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x} \,|\, k) P_k$.

Inference for this classifier can be achieved via Gibbs sampling in similar manner to the unsupervised mixture models. In addition to usual normal-inverted-Wishart priors over the component parameters, we place a separate Dirichlet prior, but with common hyper-parameter α, over each of the weight vectors:

$$\mathbf{w}_k \sim \mathcal{D}(\alpha, \ldots, \alpha) \tag{21}$$

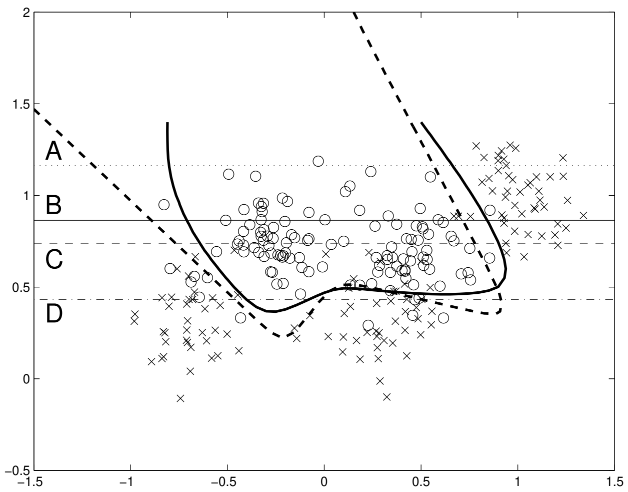

Figure 5: Synthetic two-class data. Circles and crosses mark *complete* observations from the two classes. The bold dashed line indicates the Bayes decision boundary; the bold solid line marks the posterior average decision boundary using data in which 50% of observations had a missing variable. Horizontal lines show the x_2 coordinate of an observation for which the x_1 coordinate is missing.

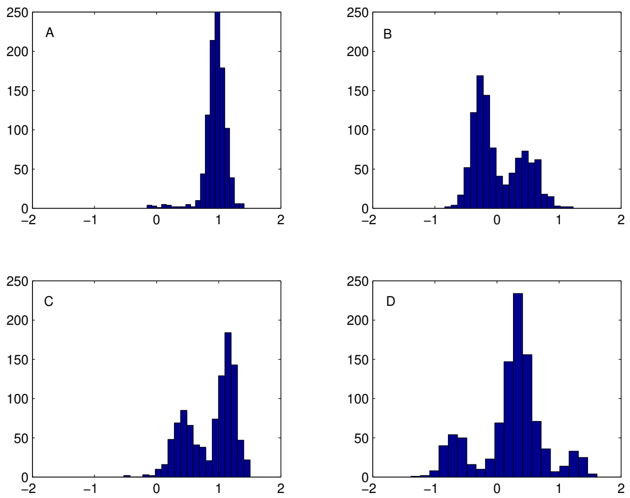

Figure 6: Histograms of posterior imputations of x_1 for data corresponding to the four horizontal A-D lines in Figure 5.

In addition we place Dirichlet priors over the prior class probabilities:

$$\mathbf{P} \sim \mathcal{D}(\delta, \ldots, \delta) \tag{22}$$

Generally the prior counts are set to be non-informative, $\delta = 1$. Let $\eta_1, \ldots \eta_K$ be the counts per class and $\eta_{1k}, \ldots, \eta_{Mk}$ be the number of targets of class k assigned to each component. Then Gibbs sampling for complete data proceeds as outlined for the unconditional mixture model with the following modifications:

- **Indicator variables** for $\mathbf{x}_n$ is drawn from a multinomial distribution based on the target t_n:

$$\mathbf{z}_n \sim \mathcal{M}(h_{n1}, \ldots h_{nM}) \tag{23}$$

  where

$$h_{nj} = \frac{w_{jt_n}\mathcal{N}(\mathbf{x}_n \,|\, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{i=1}^{M} w_{it_n}\mathcal{N}(\mathbf{x}_n \,|\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \tag{24}$$

- **Weights** are drawn from their conditional posterior densities:

$$\mathbf{w}_k \sim \mathcal{D}(\alpha + \eta_{1k}, \ldots, \alpha + \eta_{Mk}) \tag{25}$$

- **Prior proportions** are drawn from their conditional posterior densities:

$$\mathbf{P} \sim \mathcal{D}(\alpha + \eta_1, \ldots, \alpha + \eta_K) \tag{26}$$

- **Component parameters** are drawn from the (conditional) posterior normal-inverse Wishart densities (equations (6) – (9)) with the data contributing to the parameter update determined by the indicators $\mathbf{z}_{nj}$.

- **Class allocations.** Denoting the inferred class of $\mathbf{x}_n$ as y_n, samples are drawn from the conditional posterior class density as:

$$\mathbf{y}_n \sim \mathcal{M}(\pi_{n1}, \ldots, \pi_{nK}) \tag{27}$$

  where

$$\pi_{nk} = \frac{p(\mathbf{x}_n \,|\, \boldsymbol{\Theta}_k)P_k}{\sum_{i=1}^{K} p(\mathbf{x}_n \,|\, \boldsymbol{\Theta}_i)P_i} \tag{28}$$

  where $\boldsymbol{\Theta}_k$ denotes the coefficients on which the k-th class conditional density depends.

Augmenting this classifier to impute missing values is straightforward. The above Gibbs sampling steps form the P-step; the I-step consists of imputing missing values from the centre which (at a particular sweep) takes responsibility for that $(\mathbf{x}_n, t_n)$ pair, that is the centre j for which $z_{nj} = 1$. The responsibility h_{nj} is determined by (24) and we emphasise that the target class t_n plays a central role in determining the responsible component.

4.1 Illustration

We illustrate the performance of the classifier on a two-dimensional synthetic data set [Fieldsend et al., 2003] comprised of 5 Gaussian components, which is a modelled on a dataset used by Ripley [1994] for classifier

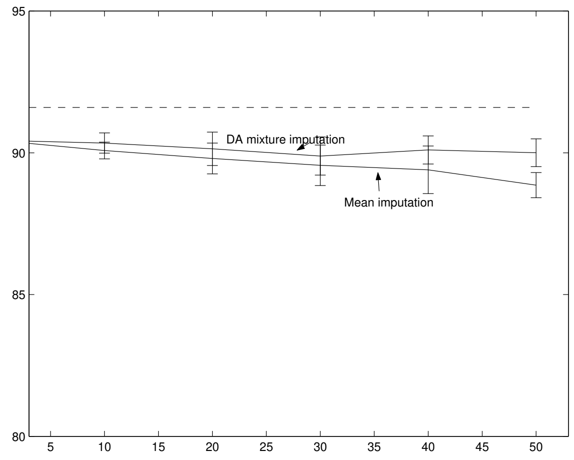

Figure 7: Classification performance for mean and DA mixture imputation plotted against the percentage of missing values for the synthetic data shown in Figure 5. The horizontal line marks the Bayes classification rate.

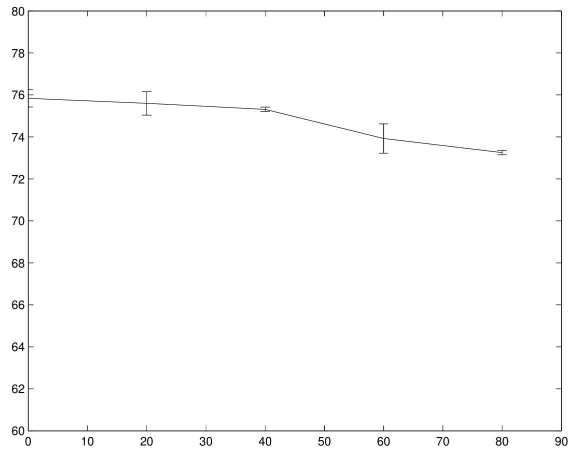

Figure 8: Pima data. Mean classification performance plotted against the percentage of records containing a missing value.

testing.[2] As illustrated in Figure 5, observations generated by two Gaussian components are allocated to one class, while observations generated by the other class are assigned to a second class; the component weights are such that there is equal prior probability of an observation from either class. The Bayes error rate for these data is 9.4%.

Non-informative priors were used during learning on data in which 50% of 250 observations have a coordinate missing; targets for all the data were present. Following burn-in, samples were generated from the joint parameter and missing data distribution, and classifications of an additional 1000 complete testing data examples were made. Figure 5 shows both the Bayes rule boundary and the 0.5 mean posterior probability contour, which reasonably closely approximates the Bayes decision boundary. We note that the principal deviations from the Bayes decision boundary are in regions of low data density and the posterior distribution $p(y \,|\, \mathbf{x}, X_{obs})$ is relatively wide in these regions.

Figure 6 shows histograms of imputations for four example data observations whose x_1 coordinate is missing, as indicated in Figure 5. Observation A belongs to the 'crosses' class and, as the histogram shows the imputed values are concentrated around the cluster centred at $(1, 1)$. In contrast observation B belongs to the 'circles' class and we draw attention to the fact that the imputed values for B are generated almost exclusively from the centres at $(-0.3, 0.7)$ and $(0.4, 0.7)$ with no contribution from the centre at $(1, 1)$ which is responsible for the 'crosses' observations. Without class information the imputations for B would extend to $x_1 > 1$. Observations C and D are more complicated, receiving contributions from several centres. We emphasise again that the mixture model permits multimodal imputations, MCMC methods allow visualisation of the full posterior and class membership provides significant conditioning for the imputations.

Figure 7 shows that the performance of the classifier on 250 observations from this synthetic data using both mean and DA mixture model imputation against the proportion of missing values (deletions at random from the complete data). The classification rate plotted represents the mean classification rate on 1000 test data points repeated over at least 5 different realisations of the incomplete data set. Even when a large proportion of the observations are missing, it is interesting to note that using the mixture model parameters are not only getting better imputation but also are sufficiently well estimated to permit better classification rates compared with mean imputation.

Finally, we show in Figure 8 the results of artificially deleting values from the nine-dimensional Pima dataset.[3] These results were obtained by again deleting, at random, observations from a proportion of the records and using a classifier based on three-centre mixture model (as suggested by MDL). The deletion process and classification process was repeated five times. The classification rates with complete data are comparable with other work [e.g. Sykacek, 2000] and decay

[2] Weights for the five components were $(0.16, 0.17, 0.17, 0.25, 0.25)$; the component means were $\boldsymbol{\mu}_1 = (1, 1)^T$, $\boldsymbol{\mu}_2 = (-0.7, 0.3)^T$, $\boldsymbol{\mu}_3 = (0.3, 0.3)^T$, $\boldsymbol{\mu}_4 = (-0.3, 0.7)^T$, $\boldsymbol{\mu}_5 = (0.4, 0.7)^T$; the covariances were all isotropic: $\boldsymbol{\Sigma}_j = 0.03\mathbf{I}$.

[3] Available from `http://www.stats.ox.ac.uk`.

only slowly as the proportion of records with a missing value becomes large.

5 CONCLUSIONS

In this paper we have shown how the data augmentation mechanism may be simply extended from single multivariate models to general mixture models, and also to a classifier with a structure similar to RBF networks. The mixture model architecture is particularly suitable for Gibbs sampling and makes for straightforward incorporation of data augmentation, which itself can be viewed as a Gibbs sampling step.

Although mixture models have previously been used for imputation in a maximum likelihood framework, the MCMC methodology permits the posterior density of the imputed values to be recovered. As we have shown, these densities are often multimodal, a characteristic that is inevitably missed by point estimates.

The results presented here were obtained using mixture models with full covariance matrices and non-informative priors. Modelling even moderately high-dimensional data with full covariance matrices and improper priors can lead to difficulties as very few observations are associated with each mixture component. Current work is investigating the efficacy of restricting the covariance matrices to be spherical or diagonal, which has the additional benefit of simplifying setting of priors. These simplifications will also facilitate a reversible jump MCMC approach to the data augmentation and classification via mixture models [Richardson and Green, 1997].

Finally, we remark that extending the Bayesian mixture model formulation to the kernel density estimator limit, in which each observation is associated with a mixture component, is a promising avenue for imputation in high dimensional data.

Acknowledgements

We thank Jonathan Fieldsend for his help with this paper. JZ gratefully acknowledges support by the EPSRC, grant GR/R24357/01.

References

C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Society B*, 39:1–38, 1977.

J. Diebolt and C.P. Robert. Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Society B*, 56(2):363–375, 1994.

R. Dybowski. Classification of incomplete feature vectors by radial basis function networks. *Pattern Recognition Letters*, 19(14):1257–1264, 1998.

J.E. Fieldsend, T.C. Bailey, R.M. Everson, W.J. Krzanowski, D. Partridge, and V. Schetinin. Bayesian Inductively Learned Modules for Safety Critical Systems. In *Proceedings of the 35th Syposium on the Interface: Computing Science and Statistics*, Salt Lake City, March 12-15 2003.

R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

Z. Ghahramani and M.I. Jordan. Supervised learning from incomplete data via an EM approach. *Advances in Neural Information Processing Systems*, 6: 120–127, 1994.

W. Hardle. *Smoothing Techniques with Implementation in S*. Springer, New York, 1991.

J.Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.

R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. Wiley series in probability and mathematical statistics. Wiley, New York, 2002.

S. Richardson and P.J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Statist. Society B*, 59: 731–792, 1997.

B.D. Ripley. Neural Networkis and Related Methods for Classification. *J. Roy. Statist. Society B*, 56(3): 409–456, 1994.

D. B Rubin. *Multiple Imputation for Nonresponse in Surveys*. J. Wiley and Sons, New York., 1987.

J.L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman and Hall, 1997.

P. Sykacek. On input selection with reversible jump Markov chain Monte Carlo Sampling. In S.A. Solla, T.K. Leen, and K.-R. M uller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 638–644, 2000.

M.A. Tanner and W.H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398 Theory and Methods), June 1987.

D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical analysis of finite mixture distributions*. Wiley, Chichester, 1985.

H.G.C. Tråvén. A neural network approach to statistical pattern classification by 'semiparametric' estimation of probability density functions. *IEEE Trans. Neural Networks*, 2(3):366–377, 1991.