BAYESMIX: A SOFTWARE PROGRAM FOR BAYESIAN ANALYSIS OFMIXTURE MODELS WITH AN APPLICATION TO MODEL-BASED CLUSTERING OF MICROARRAY GENE EXPRESSION DATA

A. Reverter, K.A. Byrne and B.P. Dalrymple
CSIRO Livestock Industries, Queensland Bioscience Precinct
306 Carmody Road, St Lucia, QLD 4067

SUMMARY
We present a FORTRAN 90 code to perform Bayesian analysis of a mixture of Gaussian distributions with a known number of components, with specific application to model-based clustering of cDNA microarray gene expression data. Its application is illustrated with two simulated and one real data set. Benchmarking is performed through equivalent models obtained via maximum likelihood. The program was developed using a Linux based compiler, although it is flexible with respect to both computer platform and user interaction. Upon request, the executable is available free of charge for research institutions and for non-commercial use only.
Keywords: FORTRAN, gene expression, microarray, mixture models, Bayesian analysis

INTRODUCTION
Cluster analyses have played a major role in determining differentially expressed genes. However, it is not sensible to view microarray data as being drawn from a single distribution. Hence, model-based clustering, via a mixture of distributions, has been identified as a method of choice to identify which genes have differential expression levels. It defines a cluster as a subpopulation with a certain distribution, results are stable and several methods exist to estimate the number of clusters (Yeung *et al.* 2001). In addition, model-based clustering provides an elegant framework to calculate the power of detecting a specified magnitude of change (Rekaya, 2002), as well as to estimate the number of replicates needed for precise inferences (Pan *et al.* 2002). Although extensive literature exists on mixture models (http://www.csse.monash.edu.au/~dld/mixture.modelling.page.html), there is limited software available that is sufficiently efficient for application to the large datasets generated in microarray experiments. The objective of this study is to introduce BAYESMIX, a FORTRAN 90 code, to perform Bayesian analysis of mixture Gaussian models with a specific application to model-based clustering of cDNA microarray data.

MATERIALS AND METHODS
Development. BAYESMIX assumes that each component (or cluster) of the data is generated by an underlying normal distribution. Each data point in $y = y_1$ to y_n is assumed to be an independent draw from a mixture density with k (unknown but finite) components and with density function:

$$f(y;\Phi_k) = \sum_{j=1}^{k} p_j f(y;m_j,V_j) \qquad [1]$$

where $f(y;m_j,V_j)$ denotes the normal density function with mean m_j and (co)variance matrix V_j, and the mixing proportions p_j are constrained to be non-negative and sum to unity. All unknown parameters are represented in Φ_k for a k-component (or k-cluster) mixture model. BAYESMIX

contemplates mixture models with up to five components (or clusters). Following Raftery (1996), the following prior densities are used:

$$V_j^{-1} \propto \Gamma\left(\frac{w_j}{2}, \frac{l_j}{2}\right),\ m_j \propto N\left(x_j, \frac{n_j}{t_j}\right), \text{ and } p \propto Dir(a_1, \cdots, a_k) \quad [2]$$

where $\Gamma(a,b)$ denotes a gamma density with mean a/b and variance a/b^2; $N(a,b)$ is a normal distribution with mean a and variance b; $p = (p_1, \cdots, p_k)$; and $Dir(a_1, \cdots, a_k)$ denotes the Dirichlet distribution with parameter a. Following Richardson and Green (1997), prior hyper-parameters are data-dependent constant chosen so that the prior distribution was relatively flat over the range of values that could be expected. Finally, the Gibbs sampler proceeds by sampling successively from the following conditional distributions:

$$\begin{aligned}
&p(a_i = j | p, k, q, d) \propto p_j\ N(m_j, V_j) \\
&p(m_j | p, k, u, m_{-j}, V_j, d) \propto N\left((n_j V_j^{-1} + k)^{-1} (n_j V_j^{-1} \bar{d}_j + k x_j), (n_j V_j^{-1} + k)^{-1}\right) \\
&p(V_j | p, k, u, m, d) \propto c^{-2}\left(2a + n_j, \left(2n_j + \sum_{i:d_i = j} (d_i - m_j)^2\right)^{-1}\right) \\
&p(p | k, u, q, d) \propto Dir(t + n_1, \cdots, t + n_k)
\end{aligned} \quad [3]$$

where the latent data $a = (a_1, \ldots, a_n)$ is an indicator of the mixture component from which y_i was generated. The Gibbs sampler runs until a Markov chain of length 12,000 is generated, the first 2,000 (burn-in) discarded, and averages from the remaining 10,000 samples used to obtain point estimates. Criteria for model selection include a combination of the likelihood ($\log L$), the Akaike Information Criterion (AIC; Akaike, 1969) and the Bayesian Information Criterion (BIC; Schwartz, 1978):

$$AIC = -2\log L(\hat{\Phi}_k) + 2u_k \text{ , and } BIC = -2\log L(\hat{\Phi}_k) + u_k \log(n) \quad [4]$$

where $u_k = 3k - 1$, the number of independent parameters in Φ_k. Point estimates in Φ_k along with $\log L$, AIC and BIC statistics are given for each model in the standard output. Once each mixture model has been fitted, the probability of each data point to belong to each cluster is given by the posterior probability in:

$$t_{ij}^{(m)} = \frac{p_i^{(m)} f\left(y_j; m_i^{(m)}, V_i^{(m)}\right)}{f(y_j; \Phi^{(m)})} \quad [5]$$

Further, a data point in y_i is classified to a given cluster if its posterior probability is the largest.

Validation. BAYESMIX was tested with two simulated data sets (Data 1 and Data 2) and one real microarray data set (Data 3). Results from BAYESMIX are contrasted with those obtained via maximum likelihood using the EMMIX software (McLachlan *et al.* 1999). This program is available freely from the web (http://www.maths.uq.edu.au/~gjm/emmix/emmix.html). Figure 1 presents the empirical density function of the three data sets. The first simulated data set (Data 1) contained 1,000 records generated from the following mixture of two components:

$$0.50\, N(-10.0, 10.0) + 0.50\, N(10.0, 10.0)$$

Data 1 is an arbitrary example and its analysis aims mostly at illustrating the concept of mixtures of distribution. Data 2 contained 2,000 records from a mixture of three distributions as follow:

$$0.10\,N(-10.0,10.0)+0.80\,N(0.0,5.0)+0.10\,N(10.0,10.0)$$

Data 2 represents a more realistic scenario in gene expression data for which the most extreme records (10% on each side) belong to a distribution with vastly different mean and increased variance compared to the majority, in this case, 80% of observations. Finally, Data 3 belongs to gene expression intensity ratios on 4,747 genes evaluated in Brahman steers fed high and low quality diets. Details of the microarray experiment from which Data 3 arises are given in Reverter *et al.* (2003).

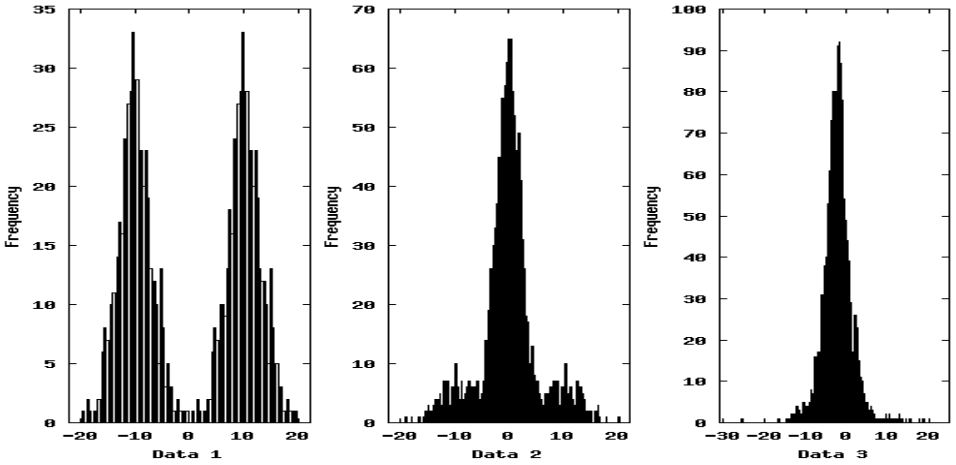

Figure 1. Density distribution for the three data sets explored in this study.

RESULTS AND DISCUSSION

Table 1 presents the clustering solutions obtained from both BAYESMIX and EMMIX and for each data set. For Data 1 and Data 2, there were no differences between software programs in the estimates of mixing proportions, means and variances of each cluster. Also, parameter estimates were equivalent to the true simulated values. The likelihood evaluation was seemingly higher for EMMIX than for BAYESMIX, which corroborates the maximum likelihood condition of the former. However, the CPU time required for BAYESMIX to achieve a solution (1 min 14 sec and 2 min 34 sec, for Data 1 and Data 2, respectively) was 4.24 and 14.71 times quicker than that required by EMMIX for Data 1 and Data 2, respectively. When analysing Data 3, both packages identified three clusters as the mixture model of choice. However, with EMMIX (which took 922 min 39 sec CPU time), a single cluster with the closest to zero mean (-0.87) and the largest variance (67.46) captured the 4.4 % most extreme records corresponding to differentially expressed genes. Instead, after 5 min 59 sec, BAYESMIX resulted in three clusters two of which contained the most extremes observations from each side. Although results from EMMIX were used as a benchmark, it should be acknowledged that estimates of conditional probabilities based on maximum likelihood estimates of parameters, acting as if these were the true values, must be viewed with caution because in mixture

models it is not always clear if asymptotic properties of the estimates hold as soon as one departs from standard settings (Gianola *et al.* 2002).

Table 1. Model-based clustering solutions for each data set

Data	N	Procedure	LogL	Parameter			
				Cluster	%	Mean	Var.
1	1,000	True	-	1	50.0	-10.0	10.0
				2	50.0	10.0	10.0
		ML	-3,233.50	1	50.0	-9.97	9.46
				2	50.0	10.02	9.47
		BAYESMIX	-3,240.42	1	50.0	-9.61	9.59
				2	50.0	9.66	9.61
2	2,000	True	-	1	10.0	-10.0	10.0
				2	80.0	0.0	5.0
				3	10.0	10.0	10.0
		ML	-5,679.40	1	10.5	-9.52	9.93
				2	79.0	0.01	4.55
				3	10.5	9.98	10.34
		BAYESMIX	-5,692.53	1	10.3	-8.80	8.93
				2	79.6	0.04	4.62
				3	10.1	9.20	9.50
3	4,747	ML	-11,863.6	1	4.4	-0.87	67.46
				2	59.0	-2.30	10.42
				3	36.6	-2.41	2.32
		BAYESMIX	-11,943.9	1	0.8	-1.02	208.79
				2	98.1	-2.26	7.61
				3	1.1	-11.18	3.63

REFERENCES

Gianola, D., Rekaya, R., Rosa, G.J.M. and Sanches, A. (2002) *Proc. 7th World Cong. Genet. Appl. Livest. Prod.* **n**=17-02.

McLachlan, G.J., Peel, D., Basford, K.E. and Adams, P. (1999) *J. Stat. Software* **4**:2.

Pan, W., Lin, J. and Le, C.T. (2002) *Genome Biol.* **3(5)**:research0022.1-0022.10.

Raftery, A.E. (1996) In "Markov Chain Monte Carlo in Practice", p. 163, editors W.R. Gilks, S. Richardson and D.J. Spiegelhalter, Chapman & Hall, Suffolk, UK.

Rekaya, R. (2002) *Proc. 7th World Cong. Genet. Appl. Livest. Prod.* **n**=16-12.

Reverter, A., Byrne, K.A., Bruce, H.L., Wang, Y.H., Dalrymple, B.P. and Lehnert, S.A. (2003) *J. Anim. Sci.* (Submitted).

Richardson, S. and Green, P.J. (1997) *J. Royal Stat. Soc.* **59**:731.