Analysis of (cDNA) Microarray Data:
Part VI. Bayesmix

A software program for Bayesian analysis of mixture models
with an application to model-based clustering
of microarray gene expression data

A Quantitative Overview to Gene Expression Profiling in Animal Genetics

A. Reverter, K.A. Byrne and B.P. Dallymply
CSIRO Livestock Industries, Queensland Bioscience Precinct
300 Cammody 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

Armidale Animal Breeding Summer Course, UNE, Feb. 2006
Contents

- **Introduction:**
  - Concept & Analysis possibilities
  - Challenges for microarray

- **Technical Concerns:**
  - Software development
  - Software comparison
  - Final remarks

Concept & Analysis Possibilities

\[ f(y; \Phi_k) = \sum_{j=1}^{k} \pi_j \phi(y; \mu_j, V_j) \]
Challenges for Microarrays

1. Class Comparison
2. Class Discovery

Detecting Outliers

Bayesmix

A Quantitative Overview to Gene Expression Profiling in Animal Genetics

BAYESMIX: Software development

\[ f(y; \Phi_k) = \sum_{j=1}^{k} \pi_j f(y; \mu_j, V_j) \]

\[ p(a_i = j | \pi, k, \theta, d) \propto \pi_j N(\mu_j, V_j) \]

\[ p(\mu_j | \pi, k, u, \mu_j, V_j, d) \propto N\left( n_j V_j^{-1} + k \right) \left( n_j V_j^{-1} d_j + k \xi_j \right) \]

\[ p(V_j | \pi, k, u, \mu, d) \propto \chi^{-2} \left( 2\alpha + n_j, \left( 2\nu_j + \sum_{i,j} (d_i - \mu_j)^2 \right)^{-1} \right) \]

\[ p(\pi | k, u, \theta, d) \propto Dir \left( t + n_1, \cdots, t + n_k \right) \]

FORTRAN 90
Up to 5 Components
Gibbs Sampling
Chain Length 12,000

Armidale Animal Breeding Summer Course, UNE, Feb. 2006
Bayesmix

BAYESMIX: Software development

\[ f(y; \Phi_k) = \sum_{j=1}^{k} \pi_j \phi(y; \mu_j, V_j) \]

\[ AIC = -2 \log L(\hat{\Phi}_1) + 2\nu_k \]
\[ BIC = -2 \log L(\hat{\Phi}_1) + \nu_k \log(n) \quad \nu_k = 3k - 1 \]

\[ \tau_{ij}^{(m)} = \frac{\pi_i^{(m)} \phi(y_j; \mu_i^{(m)}, V_i^{(m)})}{f(y_j; \Phi^{(m)})} \]

BAYESMIX: Software comparison (EMMIX)

Version 1.3 now available

Note: This program is available freely for non-commercial use only
A Quantitative Overview to Gene Expression Profiling in Animal Genetics

Bayesmix

BAYESMIX: Software test

<table>
<thead>
<tr>
<th>Data</th>
<th>N</th>
<th>Procedure</th>
<th>LogL</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cluster</td>
<td>%</td>
</tr>
<tr>
<td>1</td>
<td>1,000</td>
<td>True</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ML</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>BAYESMIX</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2,000</td>
<td>True</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ML</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>BAYESMIX</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>BAYESMIX</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ML</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>BAYESMIX</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>BAYESMIX</td>
<td>3</td>
</tr>
</tbody>
</table>

A Quantitative Overview to Gene Expression Profiling in Animal Genetics

Bayesmix

BAYESMIX: Software test (results)
DATA 3

EMMIX: (logL = -11,864)

\[ 0.044 \times N(-0.87, 67.46) + 0.590 \times N(-2.30, 10.42) + 0.366 \times N(-2.41, 2.32) \]

BAYESMIX: (logL = -11,944)

\[ 0.008 \times N(-1.02, 208.79) + 0.981 \times N(-2.26, 7.61) + 0.011 \times N(-11.18, 3.63) \]
BAYESMIX: Software comparison (EMMIX)

<table>
<thead>
<tr>
<th></th>
<th>BAYESMIX</th>
<th>EMMIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA 1</td>
<td>1,000</td>
<td>74</td>
</tr>
<tr>
<td>DATA 2</td>
<td>2,000</td>
<td>154</td>
</tr>
<tr>
<td>DATA 3</td>
<td>4,747</td>
<td>360</td>
</tr>
</tbody>
</table>

NB: EMMIX can be modified to make it faster

Conclusions

- BAYESMIX works
- Some features require further development:
  - Flexibility in Chain Length (CODA)
  - Unknown Number of Components
  - Multivariate
- EMMIX is a far more complete software
- Mixtures have other applications (eg. Selective Genotyping)
- Both softwares are available ……and will be used in this course