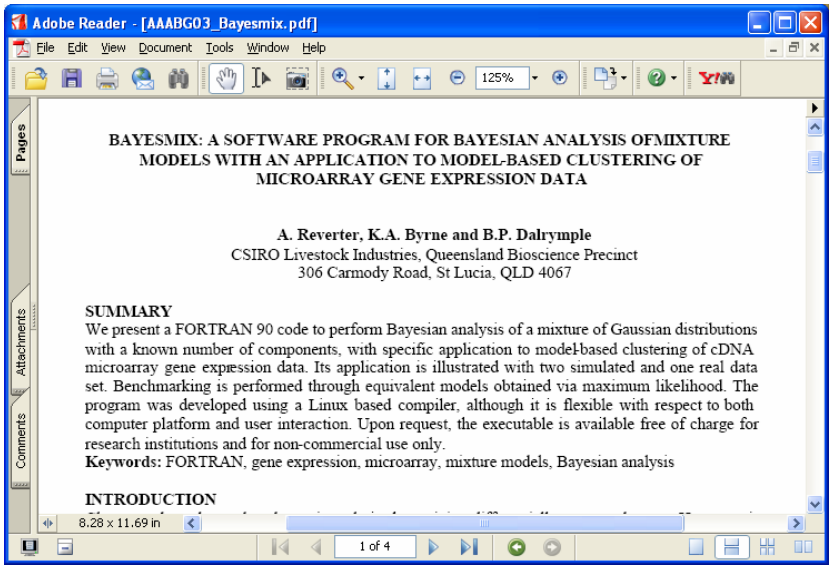
Analysis of (cDNA) Microarray Data:

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

Bayesmix



Armidale Animal Breeding Summer Course, UNE, Feb. 2006

Bayesmix

Contents

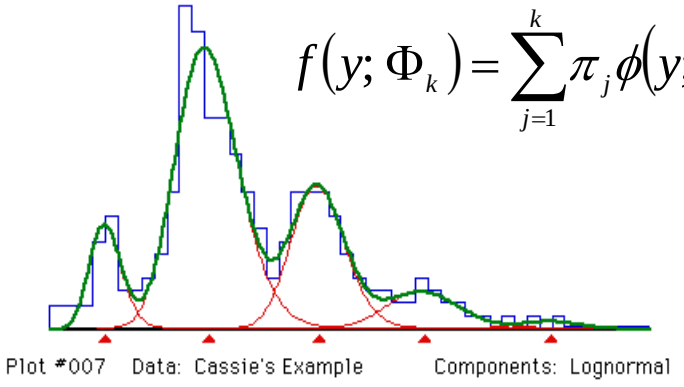
- Introduction:
  - Concept & Analysis possibilities
  - Challenges for microarray
- Technical Concerns:
  - Software development
  - Software comparison
  - Final remarks

A Quantitative Overview to Gene Expression Profiling in Animal Genetics

Bayesmix

Concept & Analysis Possibilities

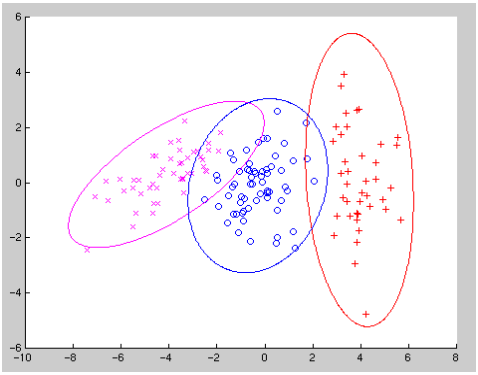
$$f(y;\Phi_k)=\sum_{j=1}^{k}\pi_j\phi(y;\mu_j,V_j)$$



Armidale Animal Breeding Summer Course, UNE, Feb. 2006

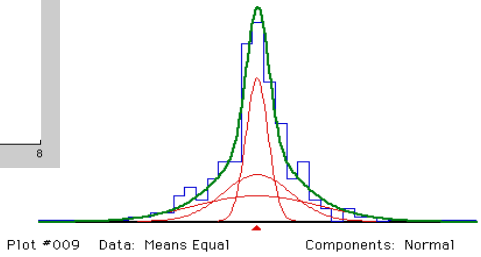
Bayesmix

Challenges for Microarrays



1. Class Comparison
2. Class Discovery

Detecting Outliers



Bayesmix

BAYESMIX: Software development

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

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

$$p(a_i=j\,|\,\pi,k,\theta,d)\,\alpha\;\pi_j\;N(\mu_j,V_j)$$

$$p(\mu_j\,|\,\pi,k,u,\mu_{-j},V_j,d)\,\alpha\;N\left(\left(n_jV_j^{-1}+k\right)^{-1}\left(n_jV_j^{-1}\bar{d}_j+k\xi_j\right),\left(n_jV_j^{-1}+k\right)^{-1}\right)$$

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

$$p(\pi\,|\,k,u,\theta,d)\,\alpha\;\;Dir\;(t+n_1,\cdots,t+n_k)$$

BAYESMIX: Software development

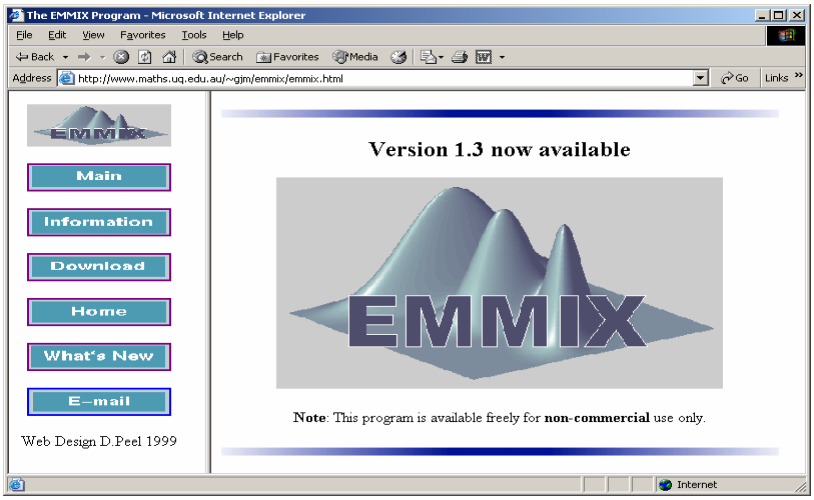
$$f(y;\Phi_k)=\sum_{j=1}^{k}\pi_j\phi(y;\mu_j,V_j)$$

$$AIC=-2\log L(\hat{\Phi}_k)+2\upsilon_k$$

$$BIC=-2\log L(\hat{\Phi}_k)+\upsilon_k\log(n) \qquad \upsilon_k=3k-1$$

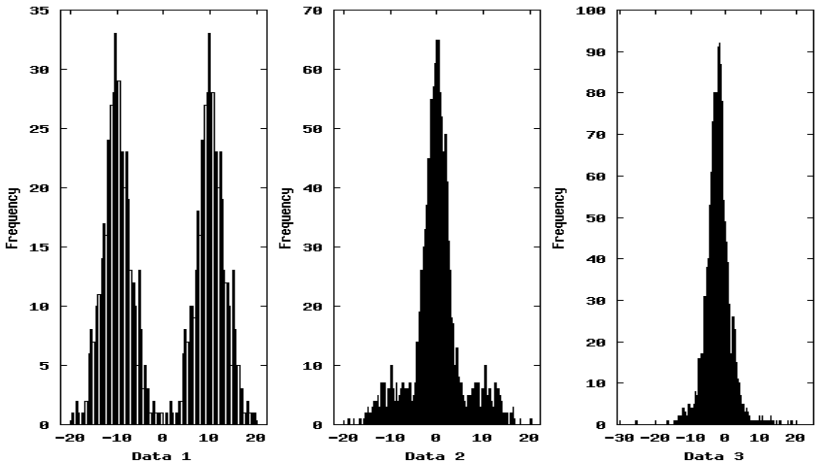
$$\tau_{ij}^{(m)}=\frac{\pi_i^{(m)}\phi(y_j;\mu_i^{(m)},V_i^{(m)})}{f(y_j;\Phi^{(m)})}$$

BAYESMIX: Software comparison (EMMIX)



Bayesmix

BAYESMIX: Software test



N = 1,000 N = 2,000 N = 4,747

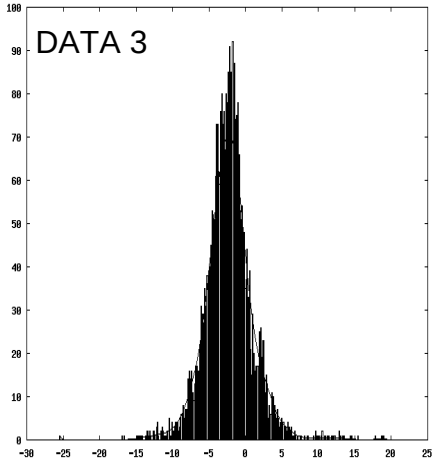
Bayesmix

BAYESMIX: Software test (results)

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

Bayesmix

BAYESMIX: Software test (results)



EMMIX: (logL = -11,864)

0.044 × N(-0.87, 67.46)
+ 0.590 × N(-2.30, 10.42)
+ 0.366 × N(-2.41, 2.32)

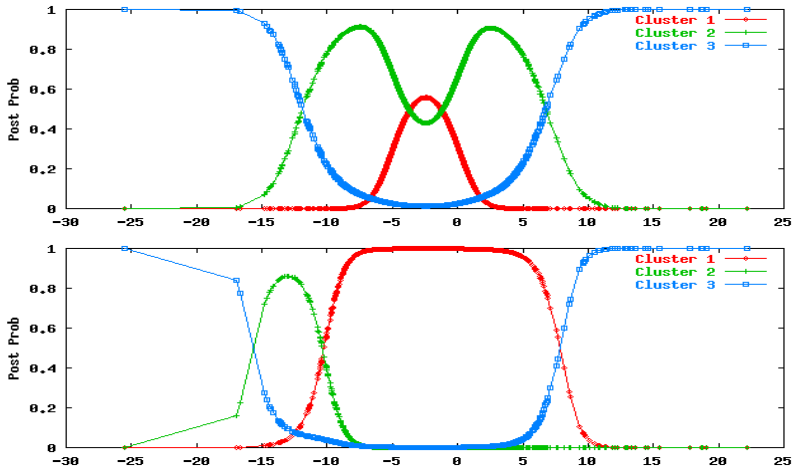
BAYESMIX: (logL = -11,944)

0.008 × N(-1.02, 208.79)
+ 0.981 × N(-2.26, 7.61)
+ 0.011 × N(-11.18, 3.63)

A Quantitative Overview to Gene Expression Profiling in Animal Genetics

Bayesmix

BAYESMIX: Software test (results)



Armidale Animal Breeding Summer Course, UNE, Feb. 2006

BAYESMIX: Software comparison (EMMIX)

		CPU Time (sec.)	
		BAYESMIX	EMMIX
DATA 1	1,000	74	314
DATA 2	2,000	154	2,265
DATA 3	4,747	360	55,359 (15.4hr!)

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