Analysis of (cDNA) Microarray Data:

Part V. Mixtures of Distributions

Model-Based Clustering
via
Mixtures of Distribution

Mixtures of Distributions

Definition

- The mixture model assumes that each cluster (or component) of the data is generated by an underlying normal distribution.
- Each of the data in *y* are assumed to be independent observations from a mixture density with *k* (possibly unknown but finite) components and with probability density function:

$$f(y;\Phi_k)=\sum_{i=1}^{k}\pi_i\phi(y;\mu_i,V_i)$$

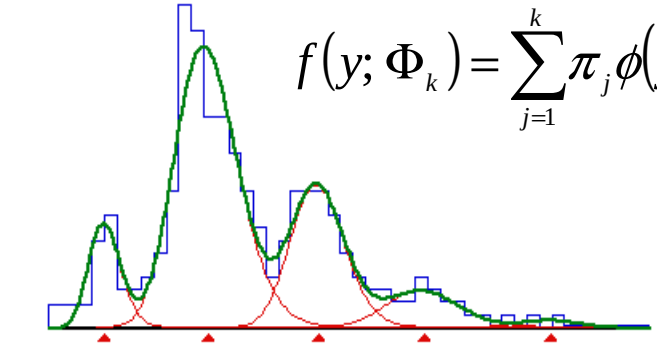
Normal density function

Mixing proportions (add to 1)

Mixtures of Distributions

Introduction

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



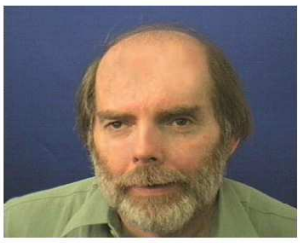
A Quantitative Overview to Gene Expression Profiling

Mixtures of Distributions

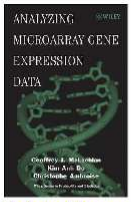
The Guru

http://www.maths.uq.edu.au/~gjm

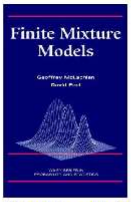
Geoff McLachlan



- **Position**: Professor of Statistics (Personal Chair) in the Department of Mathematics and Professorial Research Fellow in the Institute for Molecular Bioscience
- Email: gjm@maths.uq.edu.au
- **Phone**: +61 7 336 52150
- **Fax**: +61 7 336 51477
- **Postal address**:
- *Department of Mathematics*
  University of Queensland
  St. Lucia, Brisbane, Australia 4072
- **Office in Dept. of Mathematics**: Room 745, Priestley Building (No. 67)
- **Office in IMB**: Room 6.114, Queensland Bioscience Precinct (West Wing)

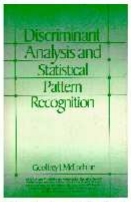


Analyzing Microarray Gene Expression Data by G. J. McLachlan, K.-A. Do, and C. Ambroise

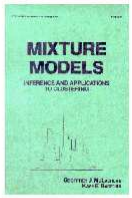


Finite Mixture Models by G. J. McLachlan and D. Peel

The EM Algorithm and Extensions by G. J. McLachlan and T. Krishnan



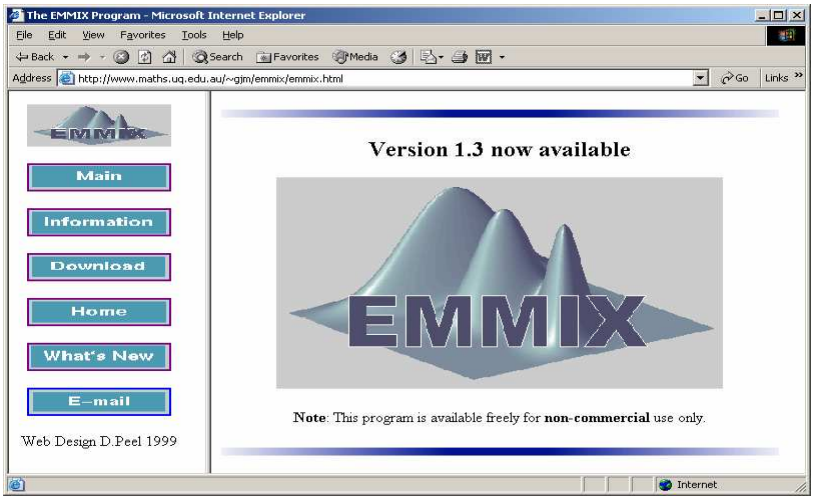
Discriminant Analysis and Statistical Pattern Recognition by G. J. McLachlan



Mixture Models: Inference and Applications to Clustering by G. J. McLachlan and K. E. Basford

Mixtures of Distributions

Software and Resources



Mixtures of Distributions

EM Algorithm

$$f(y;\Phi_k)=\sum_{i=1}^{k}\pi_i\phi(y;\mu_i,V_i)$$

The EM algorithm obtains the maximum likelihood estimate of Φ by iteration. In the (m+1)th iteration, the estimates of the parameters of interest are updated by:

$$\pi_i^{(m+1)}=\sum_{j=1}^{n}\tau_{ij}^{(m)}/n \qquad \mu_i^{(m+1)}=\sum_{j=1}^{n}\tau_{ij}^{(m)}\,y_i/\sum_{j=1}^{n}\tau_{ij}^{(m)}$$

$$V_i^{(m+1)}=\left[\sum_{j=1}^{n}\tau_{ij}^{(m)}\,(y_i-\mu_i^{(m+1)})(y_i-\mu_i^{(m+1)})^T\right]/\sum_{j=1}^{n}\tau_{ij}^{(m)}$$

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

Is the **Posterior Probability** that y_j belongs to the i-th component of the mixture (…with a very elegant link to **False Discovery Rate**).

Mixtures of Distributions

EM Algorithm

$$f(y;\Phi_k)=\sum_{i=1}^{k}\pi_i\phi(y;\mu_i,V_i)$$

- We proceed for k = 1, 2, 3, …, and so on components.
- Criteria for model selection includes the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC):

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

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

Where $\upsilon_k=3k-1$ Is the number of independent parameters in the mixture.

- Alternatively, the distribution of the likelihood ratio test (LRT) can be estimated by bootstrapping and P-values obtained to contrast a model with *k* components against a model with *k + 1* components.

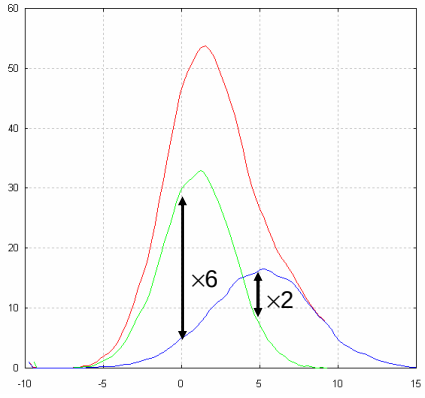
Mixtures of Distributions

Simulation 1

Consider these Distribution

		Records
N(1,5)	…and simulate	10,000
N(5,10)		5,000

The Mixture becomes: $f(y;\hat{\Phi})=\frac{2}{3}\times N(1,5)+\frac{1}{3}\times N(5,10)$



Posterior Prob: $\tau_{ij}=\frac{\pi_i\phi(y_j;\mu_i,V_i)}{f(y_j;\Phi)}$

	Likelihood *N*(1,5)		Likelihood *N*(5,10)
-1	0.120		0.021
0	0.161	×4	0.036
1	0.178		0.056
5	0.036	×3	0.126
7	0.005		0.103

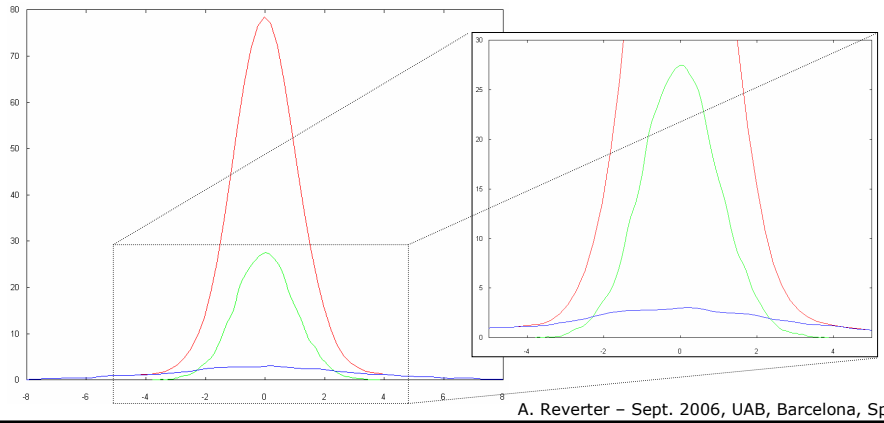
Weighted average (by mixing proportions)

Mixtures of Distributions

Simulation 2

Consider these Distribution		Records	Microarray
$N(0,1)$		9,000	Non-DE Genes
$N(0,10)$	…and simulate	1,000	DE Genes

The Mixture becomes: $f(y;\hat{\Phi}) = 0.9\times N(0,1) + 0.1\times N(0,10)$



Mixtures of Distributions

Simulation 2

1. Simulate: $f(y;\hat{\Phi}) = 0.9\times N(0,1) + 0.1\times N(0,10)$

2. Ask EMMIX to fit mixtures with up to 5 components and…

areverte@bioserver:~/UNE_data/Mixture

```
 ANALYSIS SUMMARY
 ~~~~~~~~~~~~~~~~

-------------------------------------------------------------
| NG | Log Like | -2logLAM |    AIC    |    BIC    |
-------------------------------------------------------------
|  1 |-17510.66 |     0.00 | 35025.33  | 35039.75  |
-------------------------------------------------------------
|  2 |-16383.09 |  2255.14 | 32776.18  | 32812.24  |
-------------------------------------------------------------
|  3 |-16382.27 |     1.64 | 32780.55  | 32838.23  |
-------------------------------------------------------------
|  4 |-16382.14 |     0.27 | 32786.27  | 32865.59  |
-------------------------------------------------------------
|  5 |-16381.06 |     2.16 | 32790.11  | 32891.06  |
-------------------------------------------------------------
[Mixture]$
```

areverte@bioserver:~/UNE_data/Mixture

```
Estimated mean (as a row vector) for component  1
 -0.599250E-02

Estimated mean (as a row vector) for component  2
 -0.102239E-01

Estimated covariance matrix for component    1
  0.993322

Estimated covariance matrix for component    2
   10.8053

Mixing proportion from each component
     0.903  0.097

                                          5119,0-1      7%
```

3. EMMIX model of best fit:

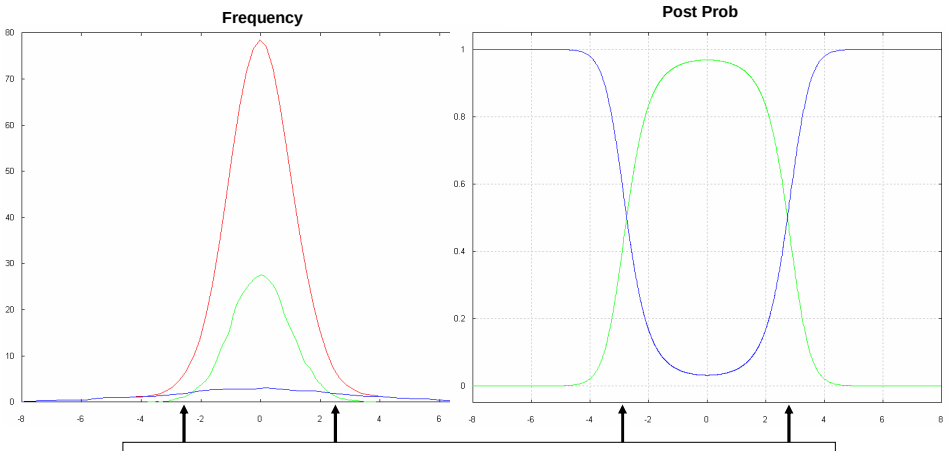
$$f(y;\hat{\Phi}) = 0.903\times N(-0.006, 0.993) + 0.097\times N(-0.010, 10.805)$$

Mixtures of Distributions

Simulation 2

1. Simulate: $f(y;\hat{\Phi}) = 0.9\times N(0,1) + 0.1\times N(0,10)$

3. EMMIX best fit: $f(y;\hat{\Phi}) = 0.903\times N(-0.006, 0.993) + 0.097\times N(-0.010, 10.805)$



Posterior Probabilities are "Decision Function" changing at ± 2.75

Mixtures of Distributions

Linking Posterior Probabilities with False Discovery Rate

CSIRO PUBLISHING
www.publish.csiro.au/journals/ajea
Australian Journal of Experimental Agriculture, 2005, **45**, 859–866

Using mixture models to detect differentially expressed genes

G. J. McLachlan[A,B,C,D], *R. W. Bean*[B], *L. Ben-Tovim Jones*[B] *and J. X. Zhu*[B]

[A]Department of Mathematics, University of Queensland, Qld 4072, Australia.
[B]ARC Centre in Bioinformatics, Institute for Molecular Bioscience,
University of Queensland, Qld 4072, Australia.
[C]ARC Special Research Centre for Functional and Applied Genomics,
University of Queensland, Qld 4072, Australia.
[D]Corresponding author. Email: gjm@maths.uq.edu.au

Abstract. An important and common problem in microarray experiments is the detection of genes that are differentially expressed in a given number of classes. As this problem concerns the selection of significant genes from a large pool of candidate genes, it needs to be carried out within the framework of multiple hypothesis testing. In this paper, we focus on the use of mixture models to handle the multiplicity issue. With this approach, a measure of the local false discovery rate is provided for each gene, and it can be implemented so that the implied global false discovery rate is bounded as with the Benjamini-Hochberg methodology based on tail areas. The latter procedure is too conservative, unless it is modified according to the prior probability that a gene is not differentially expressed. An attractive feature of the mixture model approach is that it provides a framework for the estimation of this probability and its subsequent use in forming a decision rule. The rule can also be formed to take the false negative rate into account.

Additional keywords: multiple hypothesis testing, false discovery rate, Bayes formula, Bayes rule.

Mixtures of Distributions

Linking Posterior Probabilities with False Discovery Rate

expressed. We use a prediction rule approach based on a 2-component mixture model as formulated by Lee *et al.* (2000) and Efron *et al.* (2001). We let G denote the population of genes under consideration. It can be decomposed into G_0 and G_1, where G_0 is the population of genes that are not differentially expressed, and G_1 is the complement of G_0 that is, G_1 contains the genes that are differentially expressed.

given by the 2-component mixture model:

$$f(w_j) = \pi_0 f_0(w_j) + \pi_1 f_1(w_j). \quad (6)$$

Not-DE DE

$$\tau_0(w_j) = \pi_0 f_0(w_j)/f(w_j)\ (j = 1,\ldots N). \quad (7)$$

In this framework, the gene-specific posterior probabilities $\tau_0(w_j)$ provide the basis for optimal statistical inference about differential expression.

The posterior probability $\tau_0(w_j)$ has been termed the local false discovery rate (local FDR) by Efron and Tibshirani (2002). As noted by Efron (2004), it can be viewed as an

$$\hat{\text{Risk}} = (1-c)\hat{\omega}\hat{\text{FDR}} + c(1-\hat{\omega})\hat{\text{FNR}}$$

$$\hat{\text{FDR}} = \sum_{j=1}^{N} \hat{\tau}_0(w_j)\hat{z}_{1j} \Big/ \sum_{j=1}^{N} \hat{z}_{1j}$$

$$\hat{\text{FNR}} = \sum_{j=1}^{N} \hat{\tau}_1(w_j)\hat{z}_{0j} \Big/ \sum_{j=1}^{N} \hat{z}_{0j}$$

Select the N most extreme genes, and FDR is the average posterior probability of not being in the cluster of extremes.

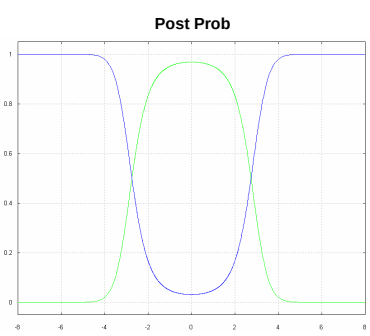
Mixtures of Distributions

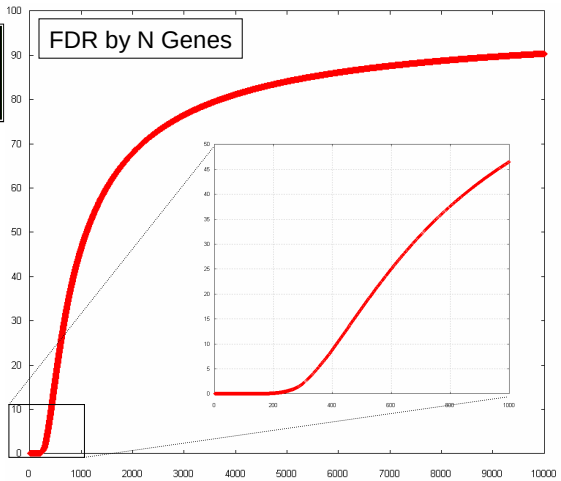
Simulation 2

1. Simulate: $f(y;\hat{\Phi}) = 0.9\times N(0,1) + 0.1\times N(0,10)$

3. EMMIX best fit: $f(y;\hat{\Phi}) = 0.903\times N(-0.006, 0.993) + 0.097\times N(-0.010, 10.805)$

Select the N most extreme genes, and FDR is the average Post Prob of not being in the cluster of extremes.



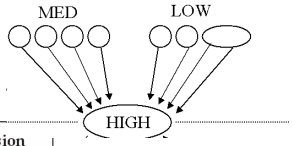


Mixtures of Distributions

Example

"Diets"

(only REFERENCE components of the design)



A mixture model-based cluster analysis of DNA microarray gene expression data on Brahman and Brahman composite steers fed high-, medium-, and low-quality diets[1]

A. Reverter*[2], K. A. Byrne*, H. L. Bruce†, Y. H. Wang*, B. P. Dalrymple*, and S. A. Lehnert*

Cooperative Research Centre for Cattle and Beef Quality,
*CSIRO Livestock Industries, Queensland Bioscience Precinct, St Lucia, Queensland 4067, Australia;
†Food Science Australia, Tingalpa DC, Queensland D 4173, Australia

ABSTRACT: The objective of this study is to explore aspects of the statistical analysis of gene expression response at the muscle tissue level to varying levels of energy and protein in the diet. Eleven Brahman and Brahman composite steers (weighing 302 ± 9.8 kg, on average) were allocated randomly into high- (HIGH), medium- (MED), and low- (LOW) quality forage diets for 27 d. After this period, a biopsy of the longissimus dorsi muscle was taken from each animal and total RNA was extracted to generate the labeled target for microarray experimentation. These targets were hybridized to a complementary DNA (cDNA) microarray of 9,274 probes from cattle muscle and subcutaneous fat cDNA libraries. After 6 hits, 151,904 expression intensity levels of 4,747 genes were analyzed. Emphasis was given to the choice of power transformation of the intensity channel readings and to the consistency of readings within each diet quality group. The statistical approach to isolate differentially expressed genes was based on model-based clustering via a mixture of normal distributions estimated through maximal likelihood. The base-2 logarithm was found to be the optimal power transformation to normalize gene intensity levels. A two-sample *t*-statistic was defined as a measure of possible differential expression. For each of the three diet contrasts, HIGH vs. LOW, HIGH vs. MED, and MED vs. LOW, three clusters were found, two of which contained more than 94% genes with almost no altered gene expression levels, whereas the third cluster contained the remaining genes with a differential expression. Results from the HIGH vs. LOW contrast identified 27 genes with a greater than 95% posterior probability of belonging to the cluster of differentially expressed genes.

Key Words: Beef, Complementary DNA, Gene Expression, Maximum Likelihood, Statistical Analysis

 J. Anim. Sci. 2003. 81:1900–1910

$$y_i^{HvL} = \frac{\overline{g}_i - \overline{r}_i}{\frac{\sigma_{g_i}}{\sqrt{8}} + \frac{\sigma_{r_i}}{\sqrt{8}}}$$

areverte@biose...

```
n = 4747
Mean = -2.275701
Variance = 10.062060
Min = -25.532000
Max = 22.160000
Skewness = 0.602129
Kurtosis = 4.630930
[Keren]$
```

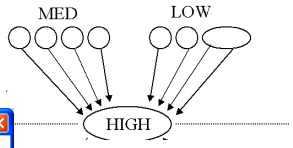

A Quantitative Overview to Gene Expression Profiling

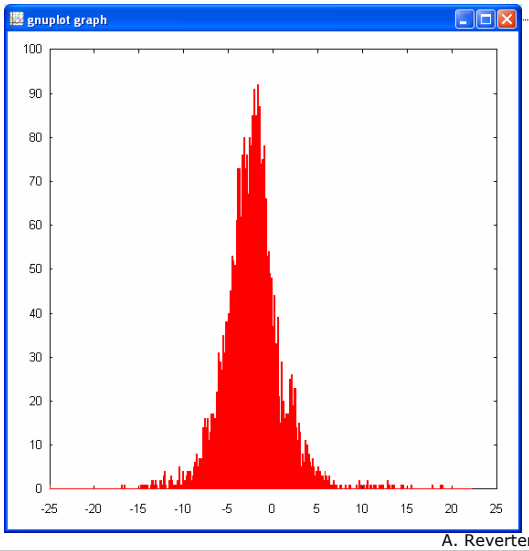
Mixtures of Distributions

Example

"Diets"

(only REFERENCE components of the design)





$$y_i^{HvL} = \frac{\overline{g}_i - \overline{r}_i}{\frac{\sigma_{g_i}}{\sqrt{8}} + \frac{\sigma_{r_i}}{\sqrt{8}}}$$

areverte@biose...

```
n = 4747
Mean = -2.275701
Variance = 10.062060
Min = -25.532000
Max = 22.160000
Skewness = 0.602129
Kurtosis = 4.630930
[Keren]$
```

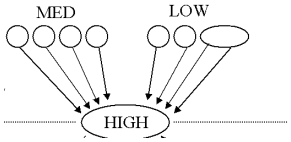
A. Reverter – Sept. 2006, UAB, Barcelona, Spain

Mixtures of Distributions

Example

"Diets"

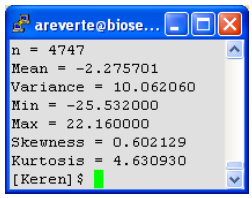
(only REFERENCE components of the design)



$$f(y;\Phi_k)=\sum_{i=1}^{k}\pi_i\phi(y;\mu_i,V_i)$$

$$y_i^{HvL}=\frac{\bar{g}_i-\bar{r}_i}{\frac{\sigma_{g_i}}{\sqrt{8}}+\frac{\sigma_{r_i}}{\sqrt{8}}}$$

$$f(y;\hat{\Phi})=0.044\times N(-0.87,67.46)+$$
$$0.590\times N(-2.30,10.42)+$$
$$0.366\times N(-2.41,\ \ 2.32)$$



A Quantitative Overview to Gene Expression Profiling

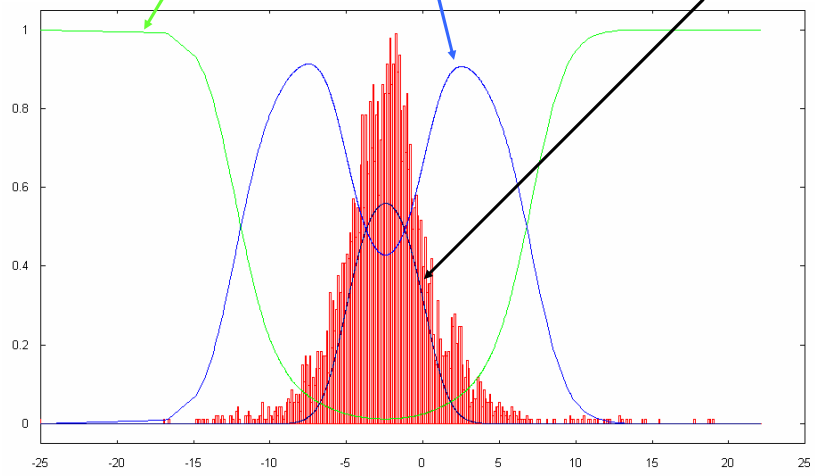
Mixtures of Distributions

Example

"Diets"

(only REFERENCE components of the design)

$$f(y;\hat{\Phi})=0.044\times N(-0.87,67.46)+0.590\times N(-2.30,10.42)+0.366\times N(-2.41,2.32),$$



A. Reverter – Sept. 2006, UAB, Barcelona, Spain

Mixtures of Distributions

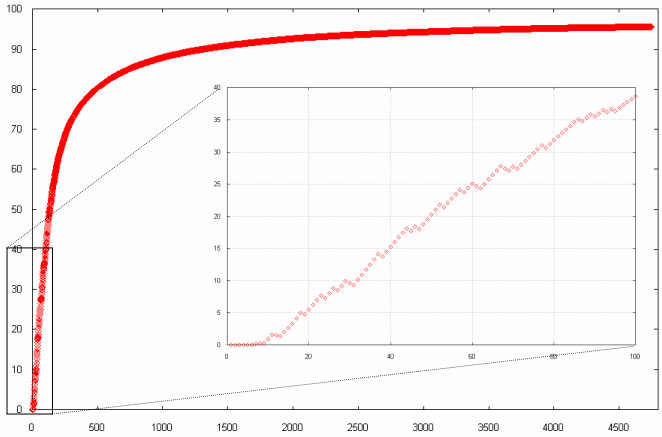
Example

"Diets"

(only REFERENCE components of the design)

$$f(y;\hat{\Phi}) = 0.044 \times N(-0.87, 67.46) + 0.590 \times N(-2.30, 10.42) + 0.366 \times N(-2.41, 2.32),$$

FDR by N Genes



In Reverter et al. '03 (JAS 81:1900), 27 genes were reported as having a PP > 0.95 of being in the extreme cluster.

Now, we can assess that these 27 genes include a FDR < 10%.