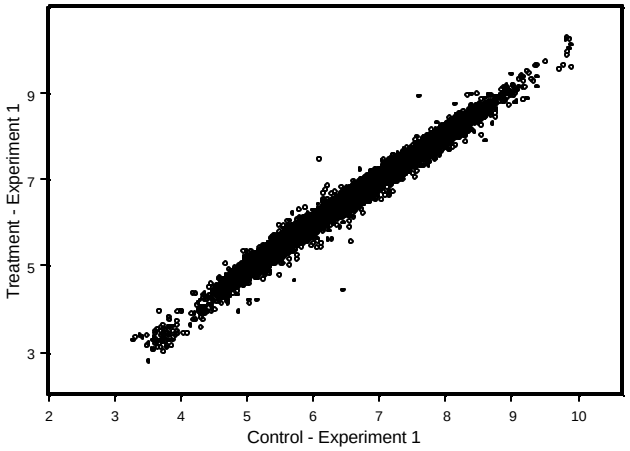
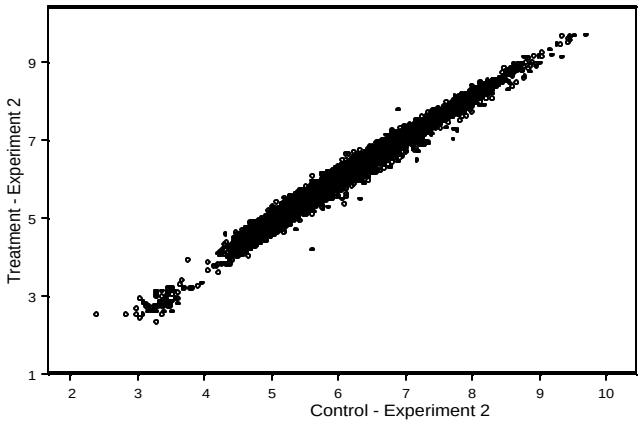
Expression data generated by DNA arrays incorporates different sources of variability present in the process of obtaining fluorescence intensity measurements. When examining expression profiles of thousands of genes at once, certain groups of genes will exhibit some level of similarity purely due to chance. Such spurious results are almost inevitable unless a proper statistical model is applied to assess the statistical significance of the observed patterns. Assessing statistical significance of observed expression patterns means determining the level of similarity that is unlikely to be the result of random fluctuations in observed data.

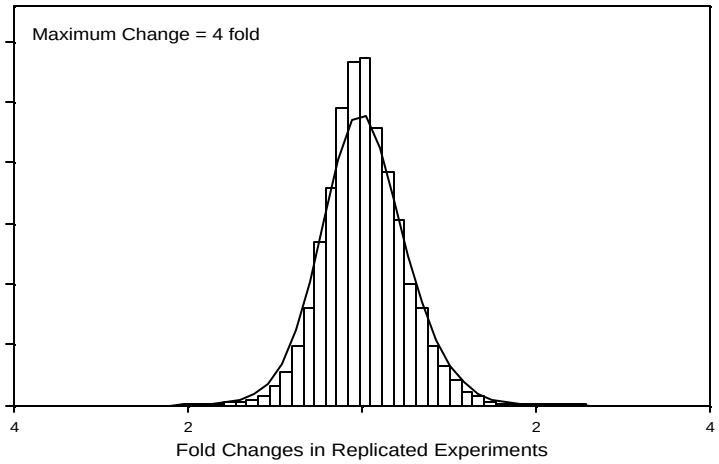
Variability in Microarray Data

Replicate 1

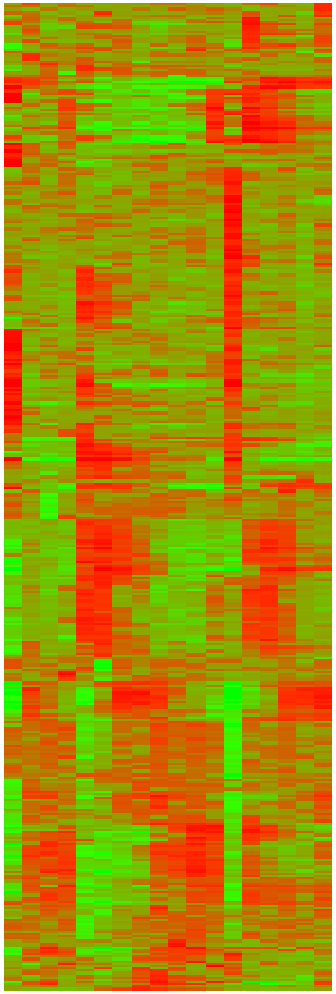


Replicate 2

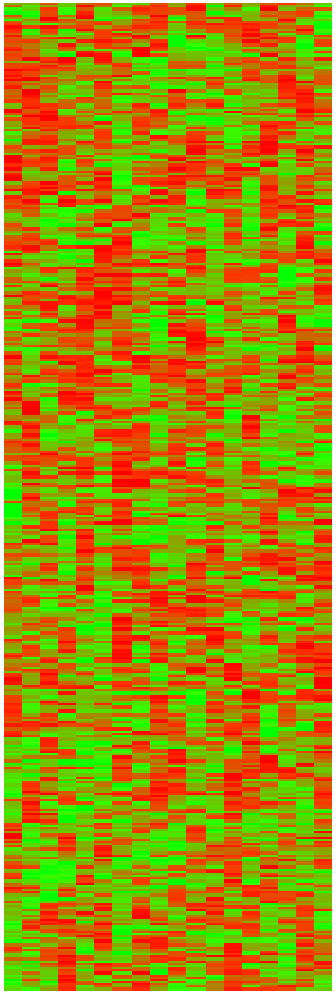




Cell Cycle Data



Randomly Generated Data



- Randomly generated data obviously less structured.
- At which point do we claim that there is any structure in the data?
- What is a statistically significant correlation?
- How many distinct patterns there are in the data?

Finite Mixture Model

x_{ij} = Expression of the i^{th} gene at j^{th} experiment

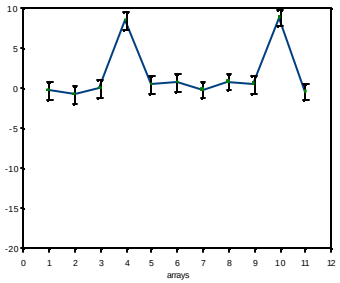
i=1,…,T, j=1,…,M

$\mathbf{x}_i = (x_{j1}, x_{j2}, \ldots, x_{jM})$ = Expression profile of the i^{th} gene

Suppose each expression profile is generated by one of Q "expression patterns"$\leftrightarrow$ Multivariate Normal Random variable with the mean m_k and the covariance matrix $\sigma\mathbf{I}$, k=1,…,Q

c_i=k, if the i^{th} profile is generated by the k^{th} "expression pattern"

$(m_k, \sigma\mathbf{I})$



Model:

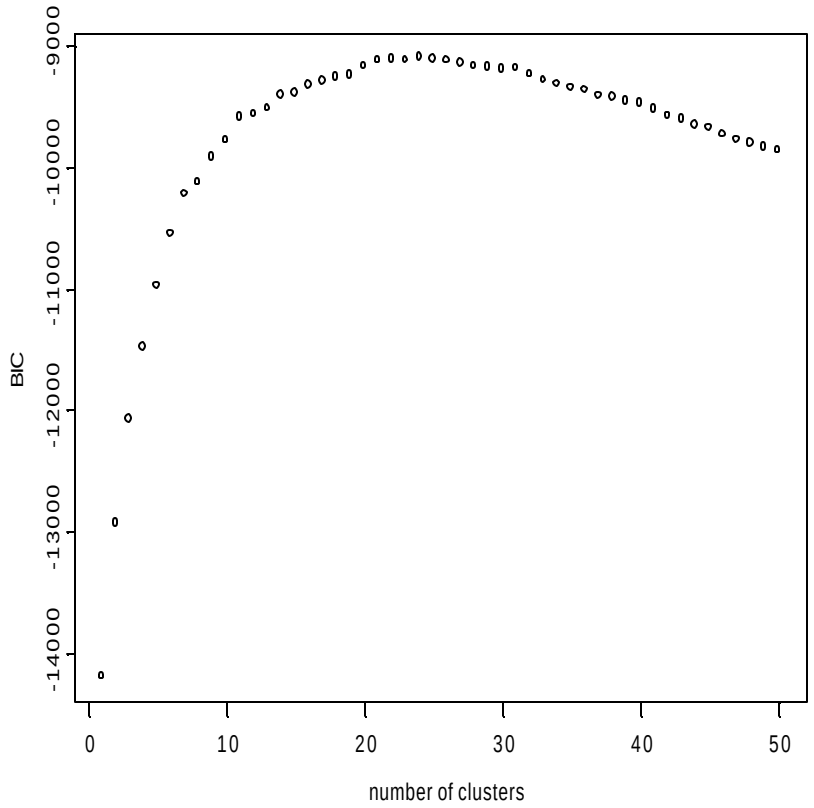
$$\mathbf{x} \sim \sum_{k=1}^{Q} p_k N(\boldsymbol{\mu}_k, s)$$

How many patterns of expression ? - Bayesian Information Criterion

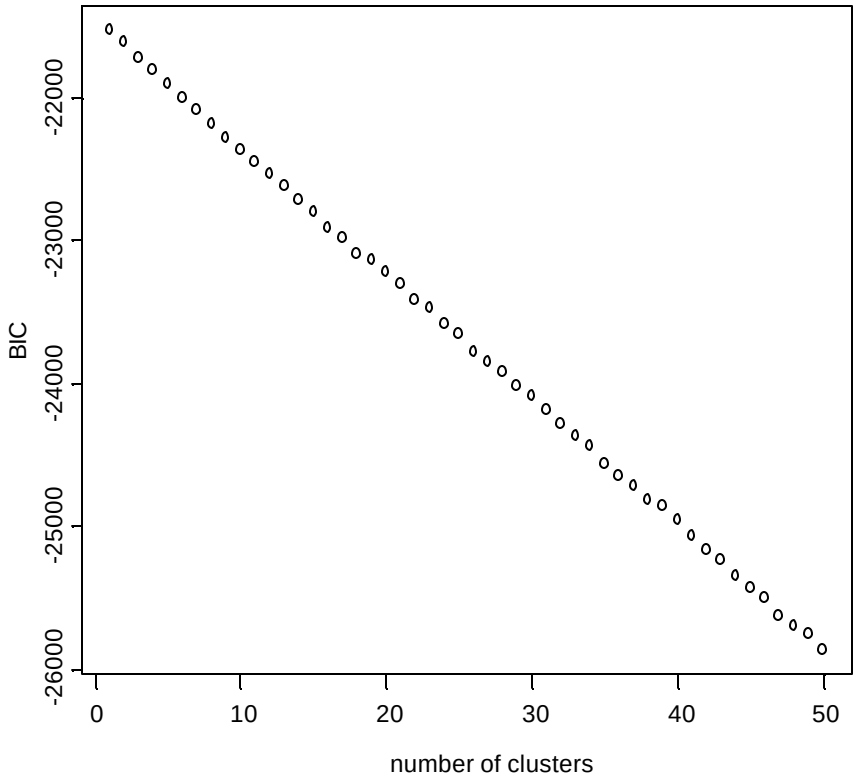
$$BIC = p(\mathbf{x}_1,...,\mathbf{x}_n;\ \hat{p}_1,...,\hat{p}_G,\ \hat{\boldsymbol{\mu}}_1,...,\hat{\boldsymbol{\mu}}_G,\ \hat{s}) - (18\,Q + 1)\log(n)$$

Optimal model « Highest *BIC*

Cell Cycle Data



Random Data



•BIC is an asymptotic criterion – valid for determining the number of mixture components when the number of data points for each mixture component is large – it favors large clusters

•When running a clustering procedure (k-means or SOM for example) with the predicted number of clusters, genes with patterns of expression that don't fit well with any major pattern are forced in some of the cluster

•There is no way to assess the absolute confidence that a profile belongs to a cluster – only relative confidence with respect to membership in other clusters can be estimated

Bayesian Infinite Mixtures

Hierarchical model defines the stochastic procedure that generates gene expression profiles. This model implicitly defines the posterior distribution of the classification variables and consequently of the number of clusters (patterns) in the data Q.

$$p(\mathbf{x}_i \mid c_i = j, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_Q, s_j) = f_N(\mathbf{x}_i \mid \boldsymbol{\mu}_j, s_j^2 \mathbf{I})$$

$$p(c_i) = \prod_{j=1}^{Q} p_j^{I(c_i=j)}$$

$$p(m_j) = f_N(m_j \mid m_x, \sigma^2_x \mathbf{I})$$

$$p(\sigma_j^{-2}) = f_G(\sigma_j^{-2} \mid 1/2,)$$

$$p(\pi_1, \ldots, \pi_Q) = f_D(\pi_1, \ldots, \pi_Q \mid \alpha/Q, \ldots, \alpha/Q)$$

Bayesian Infinite Mixtures

Posterior marginal distribution of classification variables $Q \to \infty$

$$p(c_i = j \mid \mathbf{c}_{-i}, \mathbf{x}_i, \boldsymbol{\mu}_j, s_j^2) = b\frac{n_{-i,j}}{T-1+a} f_N(\mathbf{x}_i \mid \boldsymbol{\mu}_j, s_j^2\mathbf{I})$$

$$p(c_i \neq c_j, j \neq i \mid \mathbf{c}_{-i}, \mathbf{x}_i, \boldsymbol{\mu}_x, s_x^2) = b\frac{a}{T-1+a}\int f_N(\mathbf{x}_i \mid \boldsymbol{\mu}_j, s_j^2\mathbf{I})\, p(\boldsymbol{\mu}_j, s_j^2 \mid \boldsymbol{\mu}_x, s_x^2)\, d\boldsymbol{\mu}_j ds_j^2$$

- $n_{-i,c}$ is the number of expression profiles classified in c, not counting the i^{th} profile
- $\mathbf{c}_{-i}$ is the classification vector for all except the i^{th} profile
- The joint posterior distribution of the classification vector $\mathbf{c}=(c_1,\ldots,c_T)$ is approximated by the sample $\mathbf{c}^1$, $\mathbf{c}^2$, $\mathbf{c}^3$,... generated by a Gibbs sampler utilizing these marginal distributions

Gibbs sampler

•A general procedure for sampling observations from multivariate distributions by iteratively drawing observations from marginal distributions of all components. Under mild condition, the distribution of generated multivariate observations converges to the target distribution.

•**Algorithm**

•*Initialization phase*: The algorithm is started by assuming that all profiles are clustered together

$\mathbf{c}_i^0 = 1$, for all i=1,…,T

$Q_0 = 1$

Pattern parameters $\mathtt{m}_1$ and σ^2_1 are generated as random samples from their prior distributions

•

Iterations: Given parameters after the k^{th} step ($\mathbf{c}^k$, Q_k, $\mathtt{m}_1,\ldots,\mathtt{m}_{Qk}$, σ^2), the $k+1^{st}$ set of parameters is generated by first updating classification variables $\mathbf{c}^{k+1}$

Given $\mathbf{c}^{k+1}$, new $\mathtt{m}_1,\ldots,\mathtt{m}_Q$ and σ^2 are generated according to their posterior marginal distibutions

Whenever the number of profiles in a cluster falls to zero, the cluster is removed from the list

A new cluster is created whenever a $c_i \neq c_j$ for all $i \neq j$ is selected

Identifying patterns from the Gibbs Sampler output

•**Output after G "burn-in"cycles**

$\mathbf{c}^{G+1} = (c_1^{G+1}, c_2^{G+1}, \ldots, c_T^{G+1})$

$\mathbf{c}^{G+2} = (c_1^{G+2}, c_2^{G+2}, \ldots, c_T^{G+2})$

$\mathbf{c}^{G+3} = (c_1^{G+3}, c_2^{G+3}, \ldots, c_T^{G+3})$

…….

•T by T assignment matrix **Z** is created by setting **Z**[i,j] to the number of simulated assignment vectors for which $c_i=c_j$.

•Groups of genes that had common assignments in a large portion of samples define a statistically significant pattern.

•Genes whose profiles show a low association with any other cluster are assumed to be significantly different from all other profiles

Advantages

•Underlying patterns are identifiey after "averaging" over all possible numbers of clusters

•Allows for "fuzzy" assignments and estimations of probabilities that a profile is generated by a particular "expression pattern"

•Circumvents multiple hypotheses testing issues