The Penalized Plaid Bi-clustering Model

BIRS 2011

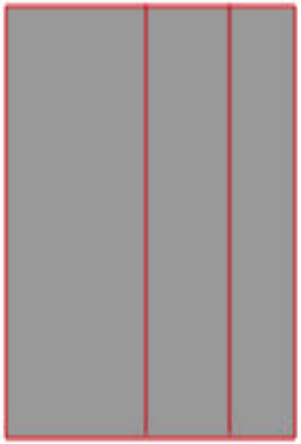
Département de mathématiques et de statistique
Université de Montréal

Thierry Chekouo T. & Alejandro Murua

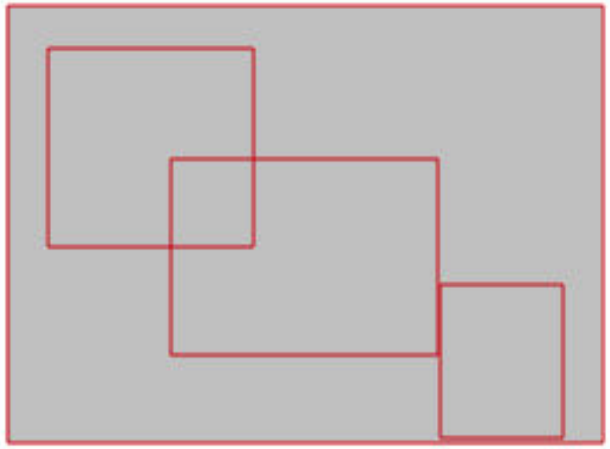
Outline

Row clusters

Columns clusters



Biclusters



$\rho_{ik} = 1$ if the row (gene) i belongs to cluster (or bicluster) k.

$$\sum_{k=1}^{K} \rho_{ik} = 1 \qquad \sum_{k=1}^{K} \kappa_{jk} = 1$$

$$\sum_{k=1}^{K} \kappa_{jk} = K \qquad \sum_{k=1}^{K} \rho_{ik} = K$$

$\sum_{k=1}^{K} \rho_{ik}\kappa_{jk} \geq 1$ (overlapping)

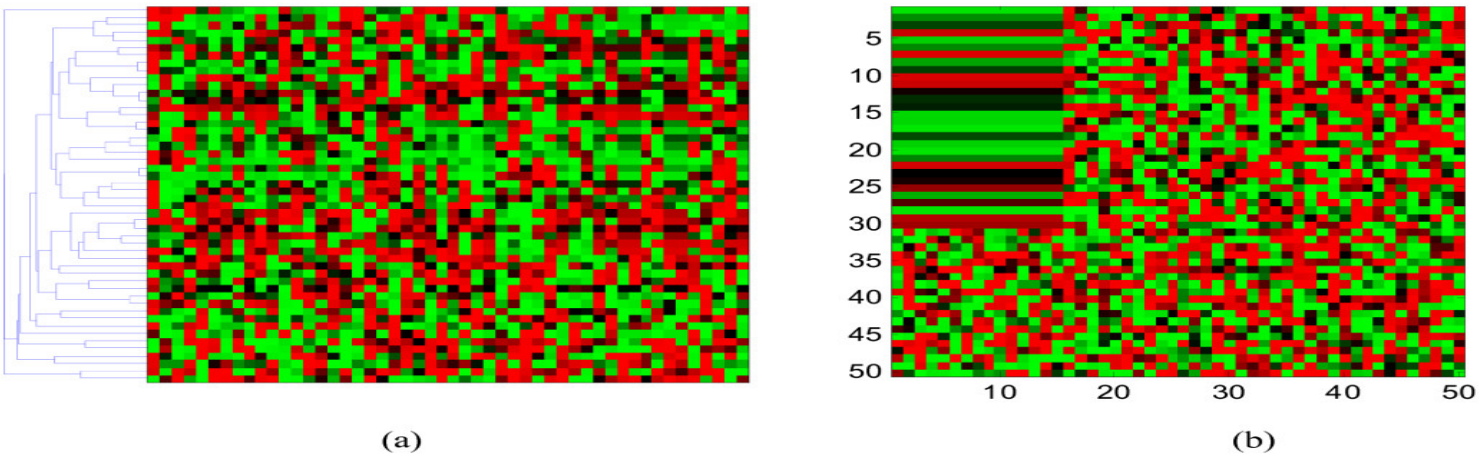
$K \geq \sum_{k=1}^{K} \kappa_{jk} \geq 1$

$K \geq \sum_{k=1}^{K} \rho_{ik} \geq 1$

Clustering versus Biclustering

- A gene or a condition may participate in multiple biological pathways

Gan et al. BMC Bioinformatics 2008 9 :209



- Not all the columns (or rows) participate
- the rows effects

Goal

- To put into a principled (Bayesian) framework the main ideas behind the most popular biclustering algorithms.
- To develop a statistical model for biclustering (of gene expression data).
  the penalized plaid model
- To be able to estimate the biclusters and their number in a principled manner.
  DIC, AIC and BIC criteria

- Cheng and Church(2000) used a greedy research of K biclusters ; the elements obtained during the previous bicluster are replaced by random numbers into the data matrix.
- Lazzeroni and Owen(2002) introduced *plaid model* where the value of an element in the data matrix is viewed as a sum of terms called *layers* (biclusters) ; the labels are assumed continuous ;
- Turner et al (2004) improved plaid model with no relaxation of the membership (i.e. labels are kept discrete).
- Zhang (2010) introduces a Bayesian plaid model : biclusters are estimated sequentially through ICM.

Biclustering as a regression :

$$y_{ij}\Big|\rho,\kappa,\Theta \sim^{i.i.d} Normal\left(\mu_0\gamma_{ij}+\sum_{k=1}^{K}(\mu_k+\alpha_{ik}+\beta_{jk})\rho_{ik}\kappa_{jk},\sigma^2(\rho_i,\kappa_j)\right)$$

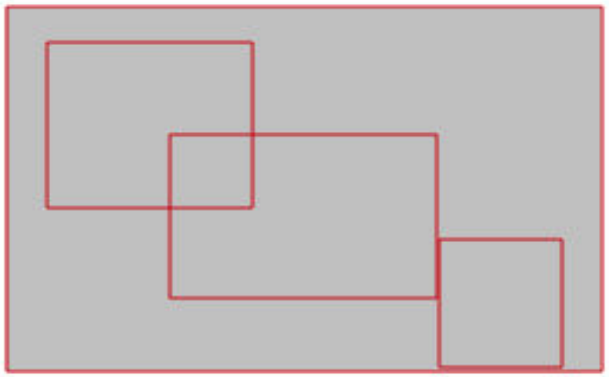
- K=number of biclusters ;
- ρ_{ik} and κ_{jk} are the bicluster membership labels ;
- $\gamma_{ij}=\prod_k(1-\rho_{ik}\kappa_{jk})$ is the label of the zero-bicluster ;
- α_{ik} and β_{jk} are the effects of bicluster k upon the row i and column j ;
- μ_0 is the mean of the zero-bicluster (background).

- If $\sum_k \rho_{ik}\kappa_{jk} \leq 1$ and $\sigma^2(\rho_i,\kappa_j)=\sigma_k^2$, we have a non-overlapping bicluster (mixture) model as in Cheng and Church(2000).
- If $\gamma_{ij}+\sum_k \rho_{ik}\kappa_{jk} \geq 1$, and $\sigma^2(\rho_i,\kappa_j)=\sigma^2$, we have the plaid (regression) model of Lazzeroni and Owen (2002).

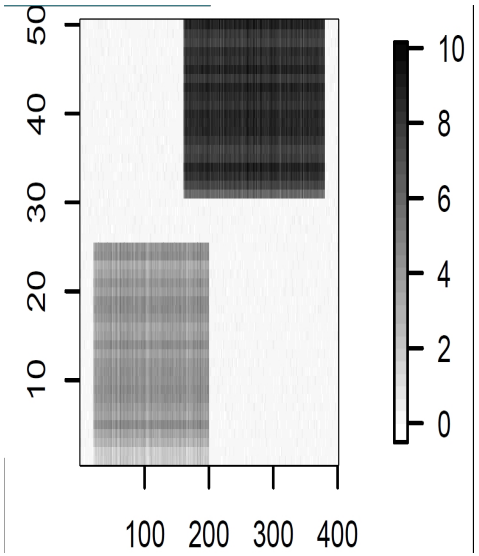
Biclustering may be seen as a mixture model :

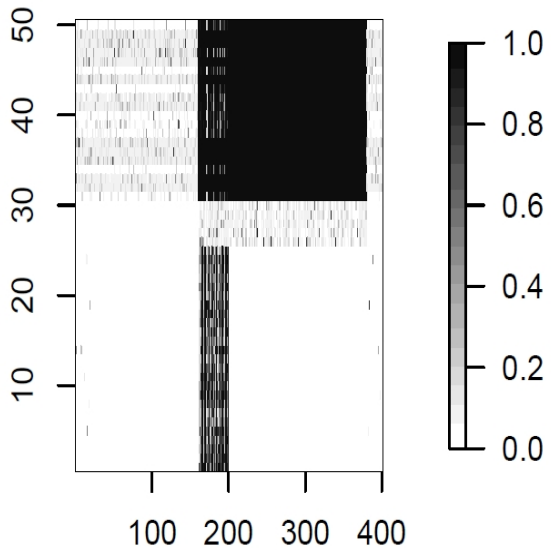
$$\frac{1}{\sigma_0}\phi\left(\frac{y_{ij}-\mu_0}{\sigma_0}\right)p(\gamma_{ij}=1)+\sum_{\substack{\rho_i,\kappa_j\\ \gamma_{ij}=0}}\frac{1}{\sigma(\rho_i,\kappa_j)}\phi\left(\frac{y_{ij}-\mu(\rho_i,\kappa_j)}{\sigma(\rho_i,\kappa_j)}\right)p(\rho_i,\kappa_j),$$

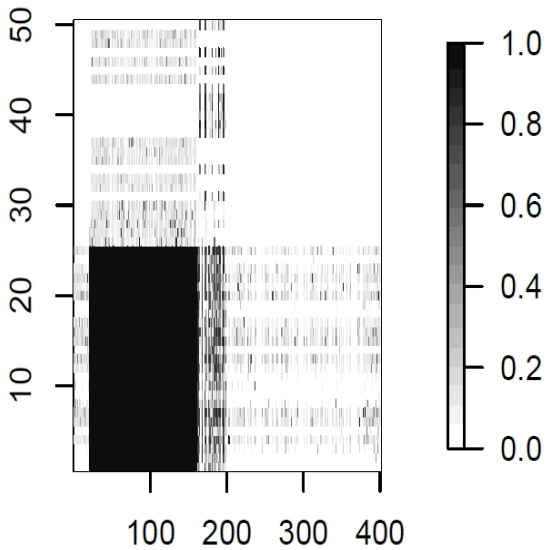
where $\mu(\rho_i,\kappa_j)=\sum_{k=1}^{K}\rho_{ik}\kappa_{jk}\,(\mu_k+\alpha_{ik}+\beta_{jk})$.



- We simulated data with two no-overlapping biclusters
- We used EM algorithm to estimate the probabilities
- $p(\rho_{ik} = 1|y) \geq p(\rho_{ik}\kappa_{jk} = 1|y)$ for each j.







Labels prior distribution

$$\pi((\rho_i, \kappa_j)|\lambda) \propto \exp\left\{-\lambda \left| 1 - \gamma_{ij} - \sum_{k=1}^{K} \rho_{ik}\kappa_{jk} \right|\right\}$$

- $\lambda \geq 0$ controls the amount of biclustering overlapping.
- if $\lambda = 0$, the labels are uniformly distributed (plaid model).
- if $\lambda \to \infty$ (non-overlapping bicluster model).

Labels

In general,

$$p(\rho_{ik}|y,\Theta_{-\rho_{ik}}) \propto A_{ik}^{\rho_{ik}} B_{ik}^{-\rho_{ik}} C_{ik,\rho_{ik}}^{\rho_{ik}} \pi(\rho_{ik})$$

where $z_{ijk} = y_{ij} - \sum_{k' \neq k} \rho_{ik'} \kappa_{jk'} (\mu_{k'} + \alpha_{ik'} + \beta_{jk'})$

$$A_{ik} = \exp\left\{-\frac{1}{2} \sum_{j \in J_k} \frac{(z_{ijk} - \mu_k - \alpha_{ik} - \beta_{jk})^2}{\sigma^2(\rho_i, \kappa_j)}\right\} \prod_{j \in J_k} \left(\sigma^2(\rho_i, \kappa_j)\right)^{-\frac{\gamma_{ijk}}{2}},$$

$$B_{ik} = \exp\left\{-\frac{1}{2\sigma_0^2} \sum_{j \in J_k} \gamma_{ijk} (y_{ij} - \mu_0)^2\right\} \left(\sigma_0^2\right)^{-\sum_{j \in J_k} \frac{\gamma_{ijk}}{2}},$$

$$C_{ik,\rho_{ik}} = \exp\left\{\frac{1}{2} \sum_{j \in J_k} (1 - \gamma_{ijk}) \frac{z_{ijk}^2}{\sigma^2(\rho_i, \kappa_j)}\right\}.$$

Labels (continuation)

- non-overlapping bicluster model : $\rho_{ik} = 1$ is favored if

$$\prod_{j \in J_k} \frac{1}{\sigma_k} \phi\left(\frac{y_{ij} - \theta_{ijk}}{\sigma_k}\right) > \prod_{j \in J_k} \frac{1}{\sigma_0} \phi\left(\frac{y_{ij} - \mu_0}{\sigma_0}\right) \exp\left(\lambda \sum_{j \in J_k} (1 - \gamma_{ijk})\right)$$

- penalized plaid model : $\rho_{ik} = 1$ is favored if

$$\prod_{j \in J_k} \phi\left(\frac{z_{ijk} - \theta_{ijk}}{\sigma}\right) > \prod_{j \in J_k} \left[\phi\left(\frac{y_{ij} - \mu_0}{\sigma}\right)\right]^{1-\gamma_{ijk}} \left[\phi\left(\frac{z_{ijk}}{\sigma}\right)\right]^{\gamma_{ijk}} \exp\left(\lambda \sum_{j \in J_k} (1 - \gamma_{ijk})\right)$$

The full conditional for κ_{jk}'s are found in a similar way by symmetry.

Metropolis-Hasting

For the non-overlapping bicluster model, we define proposal distribution as : $\rho'_{ik} = 1$ is favored if

$$\prod_{j \in J_k} \frac{1}{\sigma_0} \phi\left(\frac{y_{ij} - \mu_0}{\sigma_0}\right) < \frac{1}{(2\pi\sigma_k)^{|J_k|}} \exp(-|J_k|/2)$$

- remove row i from bicluster k if
$$\frac{1}{|J_k|} \sum_{j \in J_k} (y_{ij} - \mu_k - \alpha_{ik} - \beta_{jk})^2 > \sigma_k^2$$
- include row i in bicluster k if :
$$\frac{1}{|J_k|} \sum_{j \in J_k} (y_{ij} - \mu_k - \alpha_{ik} - \beta_{jk})^2 < \sigma_k^2$$
- If parameters are estimated by ICM (Iterated Conditional Modes) and the priors are diffuse $\longrightarrow$ generalization of Cheng and Church's algorithm.

The penalty parameter

If λ is made a parameter, then we suppose that $\pi(\lambda) \propto \lambda^{\alpha-1} \exp\{-\beta\lambda\}$.

$$p(\lambda|\rho,\kappa,y) \propto \prod_{ij} (Z_{ij}(\lambda))^{-1} \lambda^{\alpha-1} \exp\left(-\lambda(\beta + \sum_{i,j} |1 - \gamma_{ij} - \sum_{k} \rho_{ik}\kappa_{jk}|)\right)$$

where

$$\begin{aligned} Z_{ij}(\lambda) &= \sum_{k} \sum_{\rho_{ik},\kappa_{jk}} \exp(-\lambda \sum_{i,j} |1 - \gamma_{ij} - \sum_{k} \rho_{ik}\kappa_{jk}|)) \\ &= \exp(\lambda) \left\{ \exp(-\lambda) + (1 + \exp(-\lambda))^{K} - 1 \right\} \end{aligned}$$

We propose a modified *Deviance Information Criterion* (DIC) (Spiegelhalter et al., 2002) suited for biclustering (Celeux et al., 2006) $\longrightarrow$ model needs a *focus* parameter.
Let $\Theta = (\alpha, \beta, \mu, \sigma^2)$.

$$DIC_m = -4E_\Theta\left[\log(E_{\rho,\kappa}p(y|\Theta,\rho,\kappa)|y)|y\right] + 2\log(E_{\rho,\kappa}p(y|\tilde{\Theta}_m,\rho,\kappa)|y),$$

where $\tilde{\Theta}_m$ is maximum a posteriori (MAP) estimator of Θ $\longrightarrow$ yields a positive *effective dimension*

$$p_m(\tilde{\Theta}_m) = -2E_\Theta\left[\log(E_{\rho,\kappa}p(y|\Theta,\rho,\kappa)|y)|y\right] + 2\log(E_{\rho,\kappa}p(y|\tilde{\Theta}_m,\rho,\kappa)|y).$$

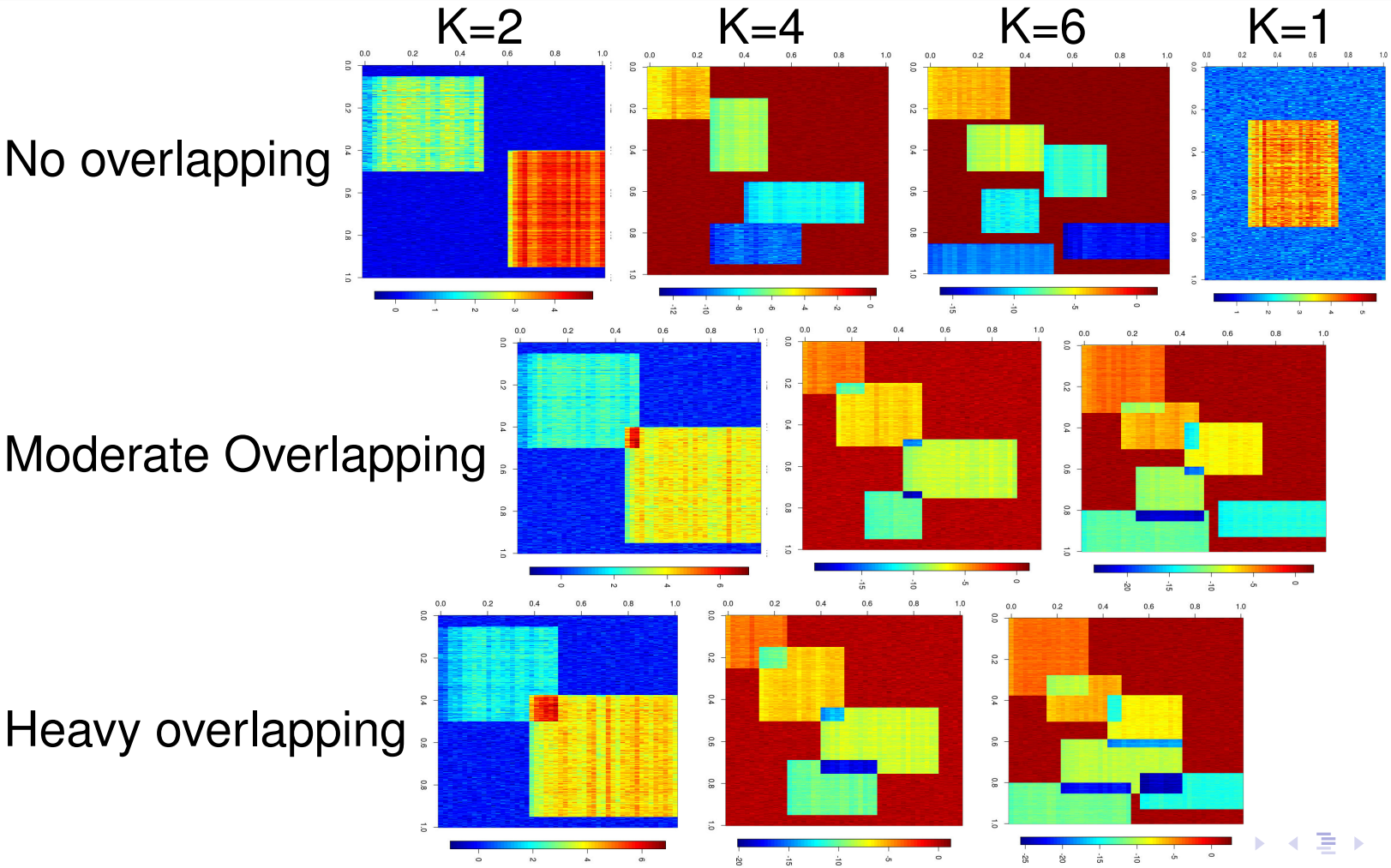
$$DIC_c = -2E_{\Theta,\rho,\kappa}\left[\log p(y|\Theta,\rho,\kappa)|y\right] + p_c(\tilde{\Theta},\tilde{\rho},\tilde{\kappa}).$$

where $(\tilde{\Theta},\tilde{\rho},\tilde{\kappa})$ is the MAP estimator of (Θ,ρ,κ), and

$$p_c(\tilde{\Theta},\tilde{\rho},\tilde{\kappa}) = -2E_{\Theta,\rho,\kappa}\left[\log p(y|\Theta,\rho,\kappa)|y\right] + 2\log p(y|\tilde{\Theta},\tilde{\rho},\tilde{\kappa}),$$

is the corresponding effective dimension.

Simulated data



Evaluating biclustering results

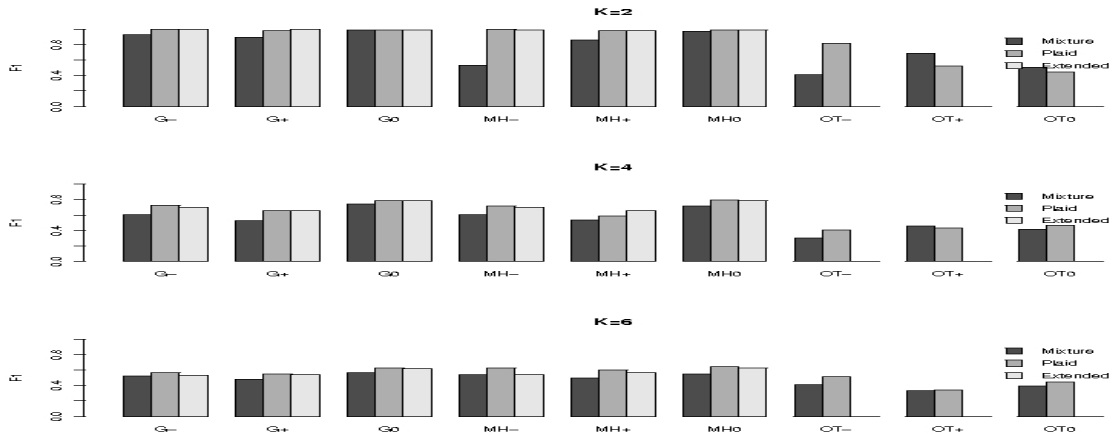
Let :

- M_1 is the set of estimated biclusters
- M_2 is the set of known biclusters
- recall= $\frac{|A \cap B|}{|B|}$, precision= $\frac{|A \cap B|}{|A|}$.
- $F_1(A,B) = \left(\frac{1}{2}(\frac{1}{\text{recall}} + \frac{1}{\text{precision}})\right)^{-1} = 2\frac{|A \cap B|}{|A|+|B|}$

$$S(M_1, M_2) = \frac{1}{|M_1|} \sum_{A \in M_1} \max_{B \in M_2} F_1(A,B)$$

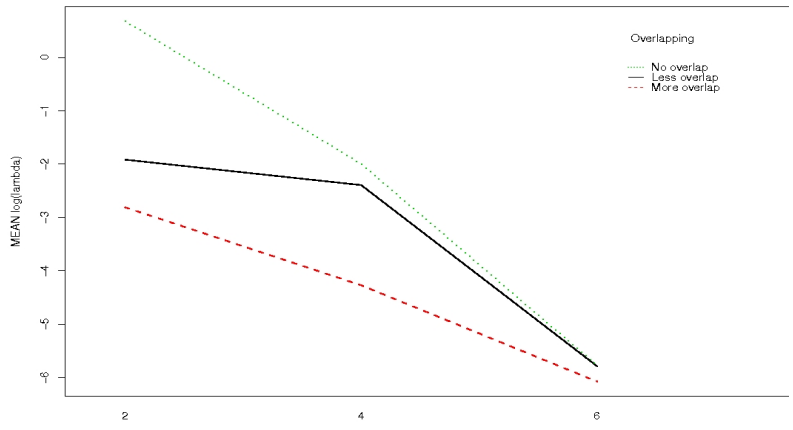
$S(M_1, M_2)$ measures the overall relevance of biclustering M_1.

Results



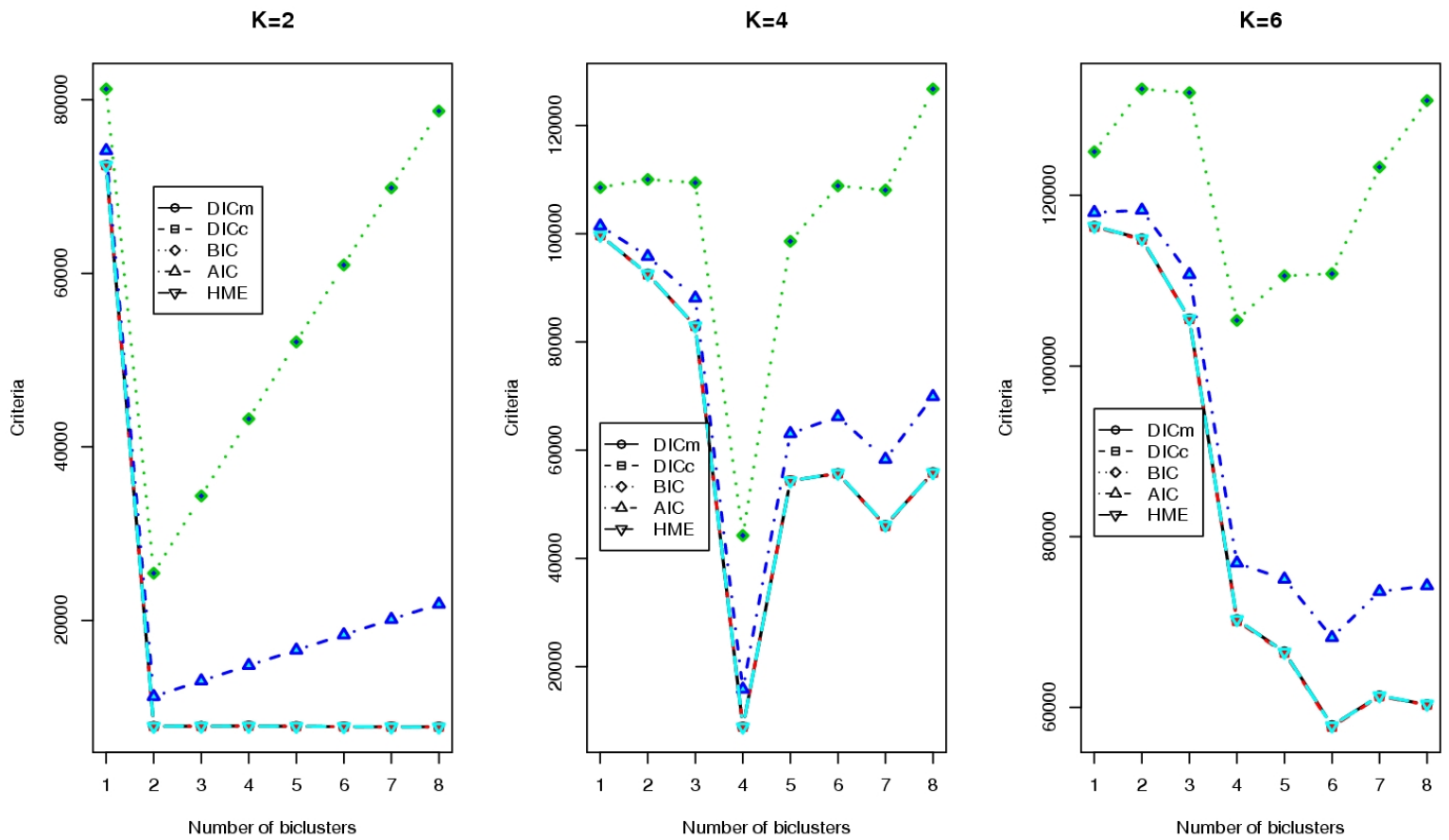
- Gibbs sampler and Metropolis-Hastings give better results than ICM-like algorithms.
- Plaid and penalized plaid models are the best (multiple comparison)
- Biclusters without overlapping are the easiest to estimate

Results



- Penalty parameter λ is a measure of complexity of the data.

Model selection



The Yeast Cycle Data

- This data set (Eisen et al, 1998) shows the fluctuation of the log-expression levels of 2467 genes over ten experimentals series comprising 79 time-points.

  It was obtained for five experimental conditions : the diauxic shift, mitotic cell division cycle, sporulation, temperature shock, and reducing shock. The data is available at http ://genome-www.stanford.edu/clustering/.

- This data set has been analysed by several researchers in the literature (Eisen et al., 1998 ; Katsuhisa and Hiroyuki, 2001 ; Lazzeroni and Owen, 2002 ; Chu et al., 1998).

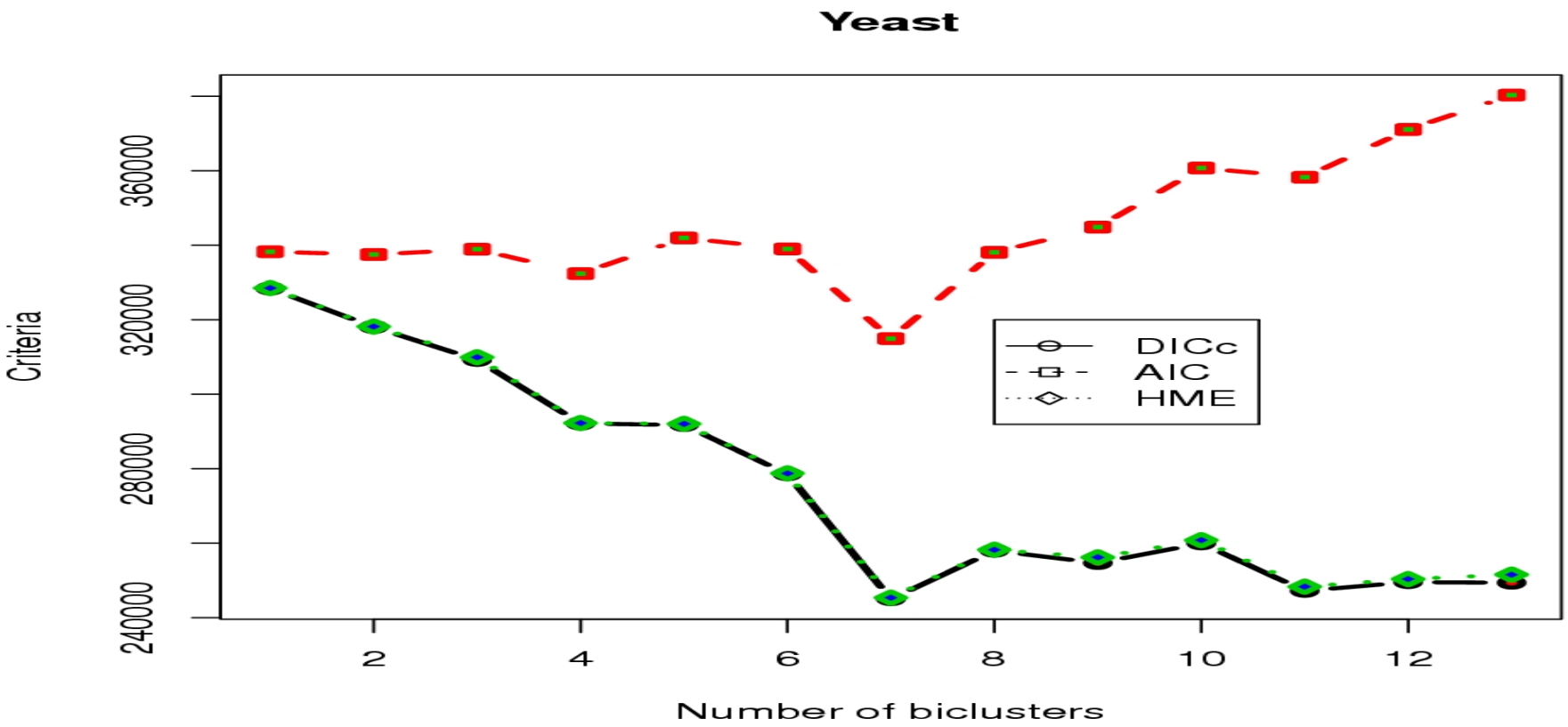
The GO project (Ashburner et al, 2000) : major bioinformatics initiative to standardize the representation of gene attributes across species and databases

- three structured controlled vocabularies or *ontologies* that describe gene products in terms of their associated biological processes, cellular components, and molecular functions.
- discover under or over-represented annotations on biclusters... difficult to assess, since the large number of genes makes it likely to find spurious but significant relations between GO terms and biclusters just by chance.
- The Bioconductor project (Falcon and Gentleman, 2007) uses the Fisher exact test to measure such relations.

  In order to control the rate of false positive relations, we employed a Bonferroni correction

Results

DIC, AIC and HME for the Yeast cell cycle data



Results

About 20% of the genes and 19% of conditions were in a single bicluster and only 8% of genes were in the zero-bicluster. There were 62% of overlapped genes in the data and 67% of overlapped conditions.

Bicluster	conditions	genes	gene annoted
0	11	194	194
1	42	1782	1777
2	48	662	659
3	4	993	992
4	5	568	567
5	13	805	802
6	8	458	456
7	26	835	834

-Biclusters 3, 4 and 6 contain only experimental conditions from sporulation.
-They play different biological roles ;
-Lazzeroni and Owen (2002) detected three similar biclusters.

Conclusions

- Penalized plaid model incorporates a penalty paramater, λ, that controls (or measures) the amount of bicluster overlapping.
  $\lambda = 0 \longrightarrow$ original plaid model of Lazzeroni and Owen (2002)
  $\lambda \uparrow +\infty \longrightarrow$ homogenous-variance version of the non-overlapping model
- We have proposed both a Gibbs sampler and a Metropolis-Hastings algorithm to estimate the parameters.
- We defined a DIC criterion that seems suitable for the biclustering problem.
- We have shown that although the Biclustering problem may be studied as a mixture model, the commonly used (soft) EM-algorithm for mixtures does not seem appropriate. Instead, an ICM-like or hard-EM algorithm appears to be more suitable.

- We note that most of the underlying algorithms for biclustering reported in the literature may be justified using hard-EM or ICM. However, we have shown through our simulations that the results derived from our MCMC implementation of the models are far better than the original Cheng and Church and plaid model algorithms.
- We applied our penalized plaid model to the yeast cell cycle data of Eisen et al. (1998). We found seven biclusters in the data as indicated by our conditional DIC model selection criterion. Among the seven biclusters, we obtained the main biclusters found previously in the literature. We showed that these biclusters are very different as indicated by their diverse biological roles obtained using GO (Falcon and Gentleman, 2007) annotations.

Thank you for your attention