Simultaneous Inference for Multiple Testing and Clustering via Dirichlet Process Mixture Models

David B. Dahl

Department of Statistics
Texas A&M University

Marina Vannucci, Michael Newton, & Qianxing Mo

3^{rd} Lehmann Symposium
Rice University
May 16, 2007

Michael Newton

Quincy Mo

Marina Vannucci

- D. B. Dahl, M. A. Newton (200?), *Multiple Hypothesis Testing by Clustering Treatment Effects of Correlated Objects*, Journal of the American Statistical Association, accepted.
- D. B. Dahl (2006), *Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model*, in “Bayesian Inference for Gene Expression and Proteomics,” Kim-Anh Do, Peter Müller, Marina Vannucci (Eds.), Cambridge University Press.
- D. B. Dahl, Q. Mo, M. Vannucci (200?), *Simultaneous Inference for Multiple Testing and Clustering via a Dirichlet Process Mixture Model*, Statistical Modelling: An International Journal, accepted.

Outline

1. Motivation
2. Z-scores Demonstration
3. BEMMA for Differential Expression – Dahl, Newton (200?)
4. BEMMA for Clustering – Dahl (200?)
5. SIMTAC for DE and Clustering – Dahl, Mo, Vannucci (200?)

Outline

1. Motivation
2. Z-scores Demonstration
3. BEMMA for Differential Expression – Dahl, Newton (200?)
4. BEMMA for Clustering – Dahl (200?)
5. SIMTAC for DE and Clustering – Dahl, Mo, Vannucci (200?)

Two Statistical Tasks

- Multiple hypothesis testing:
  - Goal: Detect shift in marginal distribution of gene expression.
  - Statistical dependence among genes is a nuisance parameter.

Two Statistical Tasks

- Multiple hypothesis testing:
  - Goal: Detect shift in marginal distribution of gene expression.
  - Statistical dependence among genes is a nuisance parameter.
- Clustering:
  - Goal: Group genes that are highly correlated.
  - Correlation may reflect underlying biological factors of interest.

Two Statistical Tasks

- Multiple hypothesis testing:
  - Goal: Detect shift in marginal distribution of gene expression.
  - Statistical dependence among genes is a nuisance parameter.
- Clustering:
  - Goal: Group genes that are highly correlated.
  - Correlation may reflect underlying biological factors of interest.
- We propose a hybrid methodology...

Main Idea

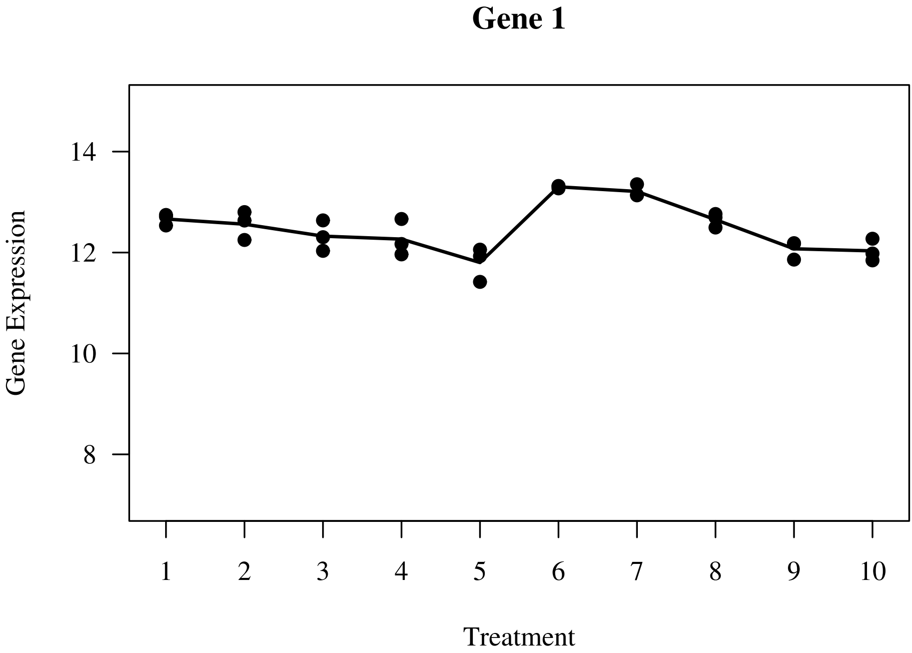
Simultaneously infer clustering & test for differential expression

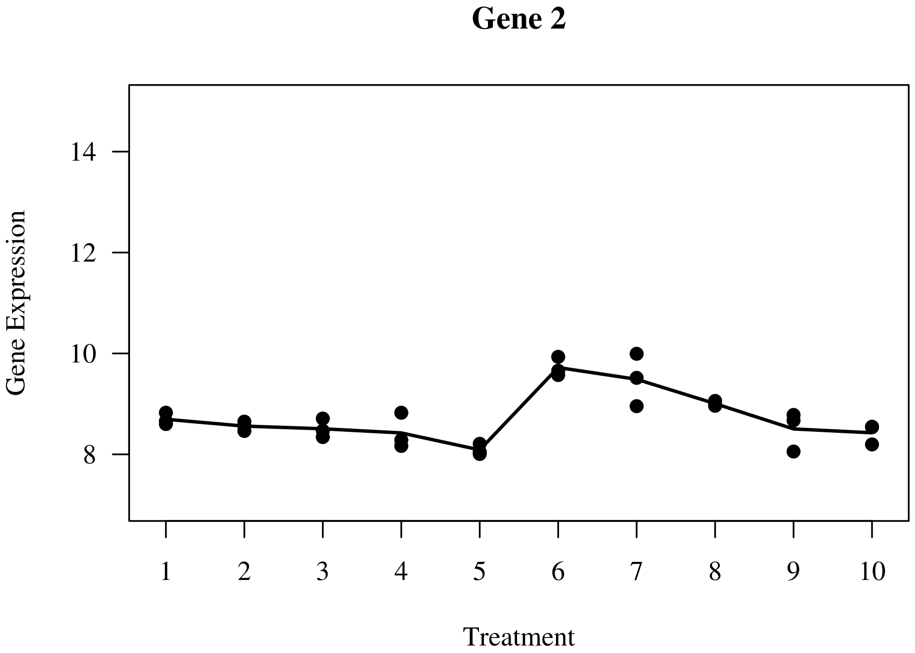
Two Statistical Tasks

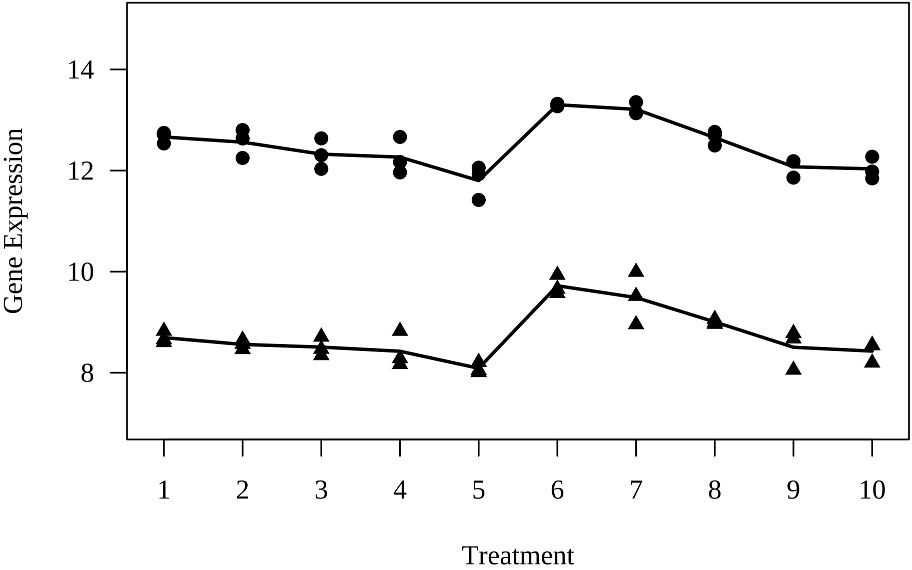
- Multiple hypothesis testing:
  - Goal: Detect shift in marginal distribution of gene expression.
  - Statistical dependence among genes is a nuisance parameter.
- Clustering:
  - Goal: Group genes that are highly correlated.
  - Correlation may reflect underlying biological factors of interest.
- We propose a hybrid methodology...

> **Main Idea**
>
> **Simultaneously infer clustering & test for differential expression**

- Other work: Storey (2007), Yuan & Kendziorski (2006), Tibshirani & Wasserman (2006)


Gene 1
Gene Expression
14
12
10
8
1
2
3
4
5
6
7
8
9
10
Treatment


Gene 2
Gene Expression
14
12
10
8
1
2
3
4
5
6
7
8
9
10
Treatment


Gene Expression
14
12
10
8
1
2
3
4
5
6
7
8
9
10
Treatment

Outline

1. Motivation
2. Z-scores Demonstration
3. BEMMA for Differential Expression – Dahl, Newton (200?)
4. BEMMA for Clustering – Dahl (200?)
5. SIMTAC for DE and Clustering – Dahl, Mo, Vannucci (200?)

Elementary Setup

- Parameters $\theta_1, \ldots, \theta_n$ for *n* observations.
- Hypotheses:
  - $H_{0i} : \theta_i = 0$, vs.
  - $H_{ai} : \theta_i > 0$
- Test statistics $Z_1, \ldots, Z_n$ are independent and

$$Z_i \sim N(\theta_i, 1)$$

Method I

- Standard Z-test in which H_{0i} is rejected if $Z_i > z^*$, where z^* is chosen to achieve the desired size.
- The test has power:

$$1 - \Phi(z^* - \theta_i)$$

where $\Phi(x)$ is the standard normal distribution function evaluated at x.

Method II

- Assumes a known clustering: $c_{ij} = \mathrm{I}\{\theta_i = \theta_j\}$.
- Test statistic:

$$S_i = Z_i + \sum_{i \neq j} c_{ij} Z_j.$$

- The test has power:

$$1 - \Phi(z^* - \sqrt{n^{(i)}}\theta_i)$$

  where $n^{(i)} = \sum_{j=1}^{n} c_{ij}$
- Method 2 is never less powerful than Method 1.

Method III

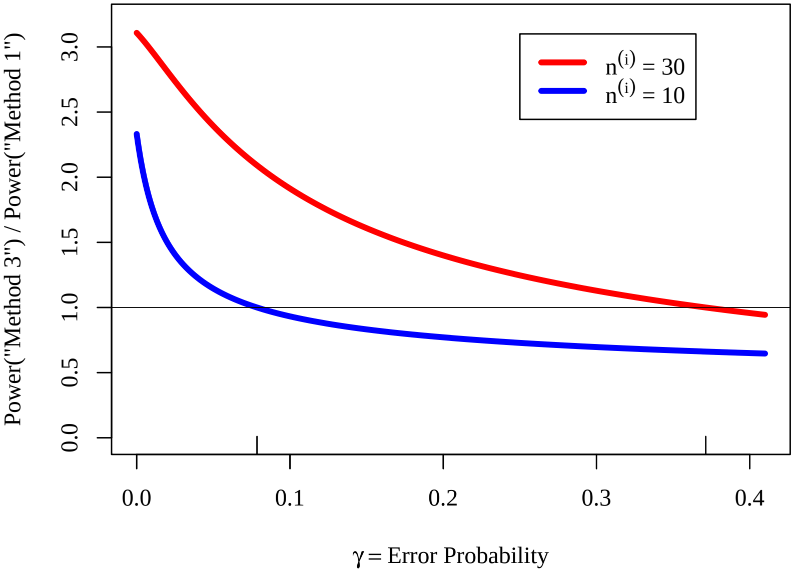
- Clustering indicators c_{ij}'s are estimated:

$$\hat{c}_{ij} = \begin{cases} c_{ij} & \text{with probability } 1 - \gamma \\ 1 - c_{ij} & \text{with probability } \gamma, \end{cases}$$

- γ is the error rate of clustering.
- Take Method II, but replace c_{ij} with $\hat{c}_{ij}$ to form $\hat{S}_i$.
- Under an assumption about the distribution of $\theta_1, \ldots, \theta_n$, the test has power:

$$1 - \Phi(z^* - k\theta_i)$$

  where k is a constant involving γ, $n^{(i)}$, etc.


Power("Method 3") / Power("Method 1")
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.1
0.2
0.3
0.4
γ = Error Probability
$n^{(i)}$ = 30
$n^{(i)}$ = 10

Outline

1. Motivation

2. Z-scores Demonstration

3. BEMMA for Differential Expression – Dahl, Newton (200?)

4. BEMMA for Clustering – Dahl (200?)

5. SIMTAC for DE and Clustering – Dahl, Mo, Vannucci (200?)

BEMMA for Diff. Expression – Dahl, Newton (200?)

- Bayesian Effects Model for Microarrays (BEMMA):
    - Conjugate Dirichlet process mixture (DPM) model.
    - Identifies differentially expressed genes by borrowing strength from genes likely to have the same parameters.
    - Averages over clustering uncertainty.

BEMMA for Diff. Expression – Dahl, Newton (200?)

- Bayesian Effects Model for Microarrays (BEMMA):
  - Conjugate Dirichlet process mixture (DPM) model.
  - Identifies differentially expressed genes by borrowing strength from genes likely to have the same parameters.
  - Averages over clustering uncertainty.
- Sampling model:

$$y_{gtr} \mid \mu_g, \tau_{gt}, \lambda_g \sim N(y_{gtr} \mid \mu_g + \tau_{gt}, \lambda_g),$$

where *r* is replicate, *t* is treatment, and *g* is gene.

BEMMA for Diff. Expression – Dahl, Newton (200?)

- Bayesian Effects Model for Microarrays (BEMMA):
  - Conjugate Dirichlet process mixture (DPM) model.
  - Identifies differentially expressed genes by borrowing strength from genes likely to have the same parameters.
  - Averages over clustering uncertainty.
- Sampling model:

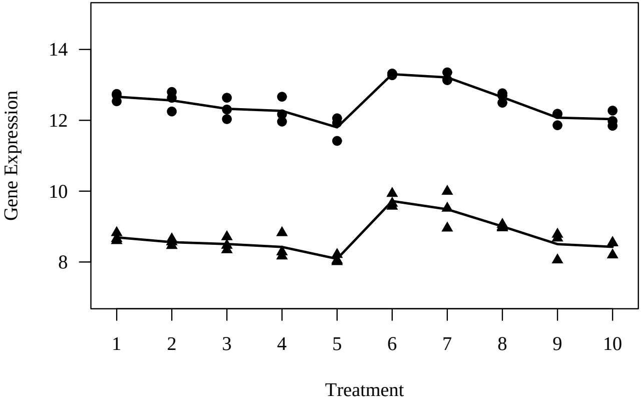
$$y_{gtr} \mid \mu_g, \tau_{gt}, \lambda_g \sim N(y_{gtr} \mid \mu_g + \tau_{gt}, \lambda_g),$$

  where r is replicate, t is treatment, and g is gene.
- Genes g and g' come from the same cluster iff:

$$(\tau_{g1}, \ldots, \tau_{gT}, \lambda_g) = (\tau_{g'1}, \ldots, \tau_{g'T}, \lambda_{g'})$$

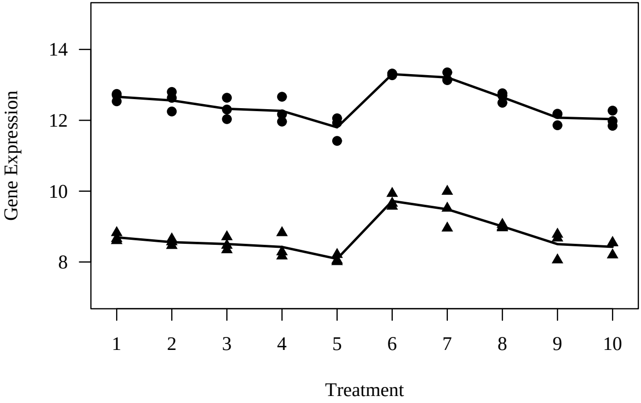
Nuisance Parameters

- Gene-specific means $\mu_1, \ldots, \mu_G$ are not related to differential expression or clustering.



Nuisance Parameters

- Gene-specific means $\mu_1, \ldots, \mu_G$ are not related to differential expression or clustering.



- Let $\boldsymbol{d}_g$ be a vector whose elements are $y_{gtr} - \overline{y}_{g1}$ for $t \geq 2$.

Model

- Sampling distribution:

$$\boldsymbol{d}_g \mid \boldsymbol{\tau}_g, \lambda_g \sim N_N(\boldsymbol{d}_g \mid \mathbf{X}\boldsymbol{\tau}_g, \lambda_g \mathbf{M}),$$

where $\boldsymbol{\tau}_g = (\tau_{g2}, \ldots, \tau_{gT})$, $\mathbf{M} = (\,\mathbf{I} + \frac{1}{R_1}\mathbf{J}\,)^{-1}$, and $\mathbf{X}$ is a design matrix picking off the appropriate element of $\boldsymbol{\tau}_g$.

Model

- Sampling distribution:

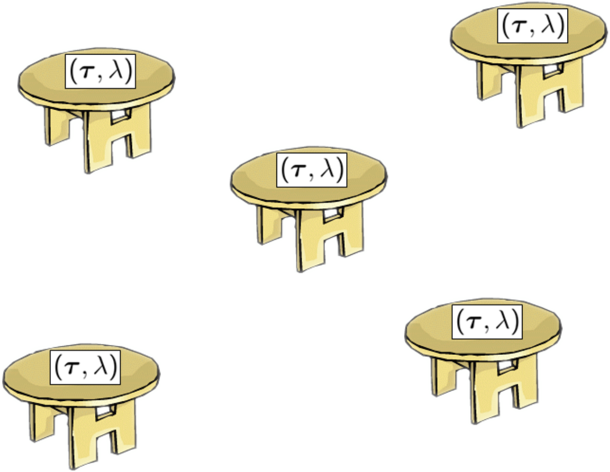
$$\boldsymbol{d}_g \mid \boldsymbol{\tau}_g, \lambda_g \sim N_N(\boldsymbol{d}_g \mid \mathbf{X}\boldsymbol{\tau}_g, \lambda_g \mathbf{M}),$$

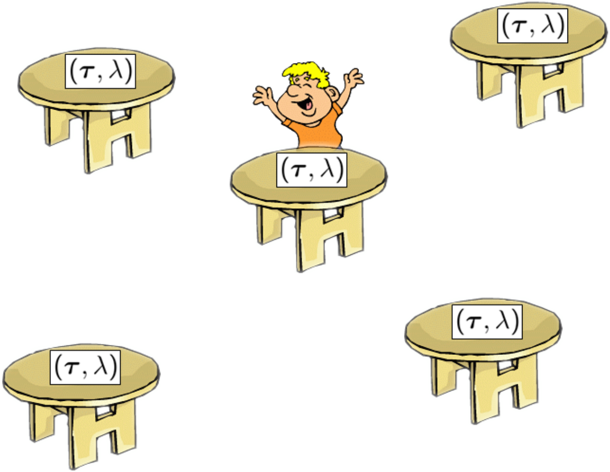
  where $\boldsymbol{\tau}_g = (\tau_{g2}, \ldots, \tau_{gT})$, $\mathbf{M} = (\,\mathbf{I} + \frac{1}{R_1}\mathbf{J}\,)^{-1}$, and $\mathbf{X}$ is a design matrix picking off the appropriate element of $\boldsymbol{\tau}_g$.

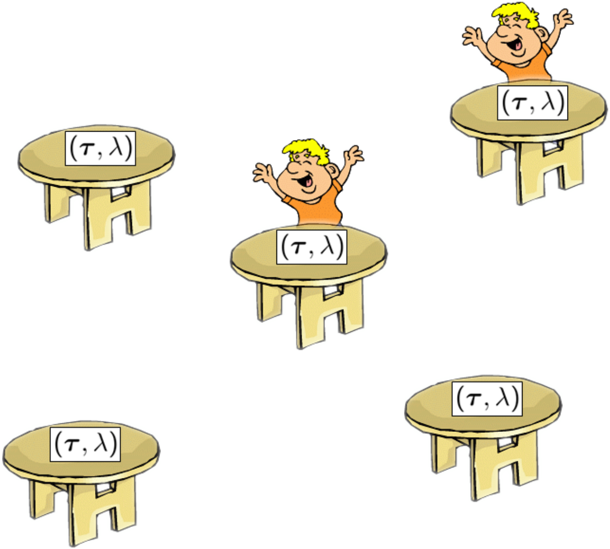
- Clustering based on $(\boldsymbol{\tau}, \lambda)$ via a Dirichlet process prior:

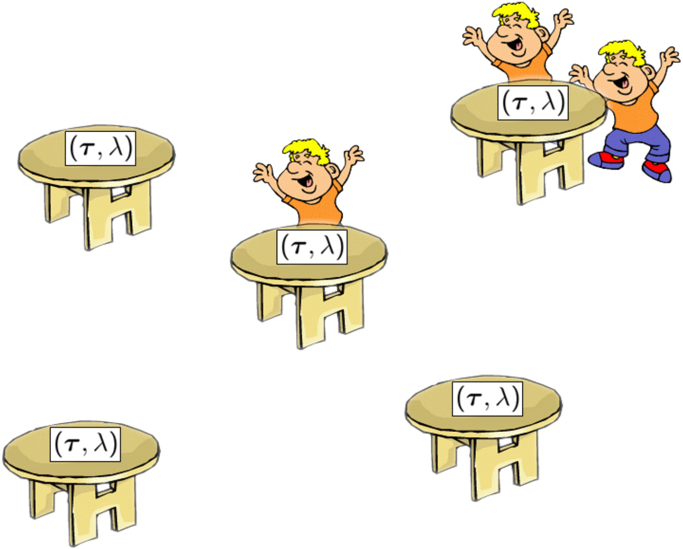
$$\begin{aligned} (\boldsymbol{\tau}_g, \lambda_g) \mid F &\sim F \\ F &\sim DP(\alpha F_0), \end{aligned}$$

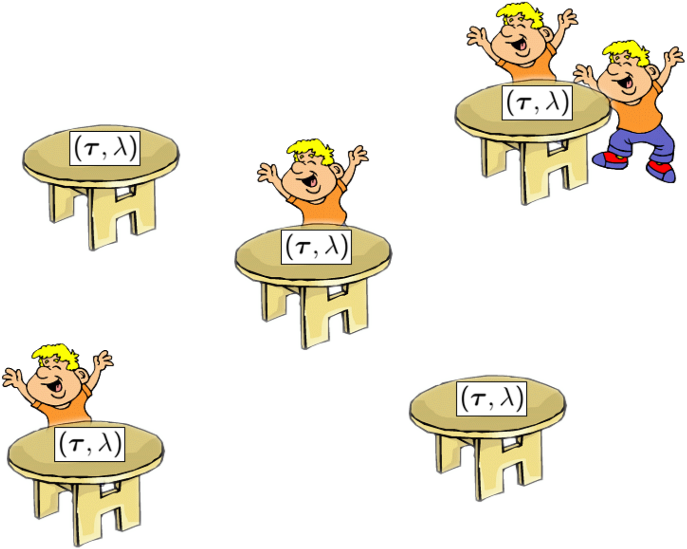
  where F_0 is conjugate to the likelihood.

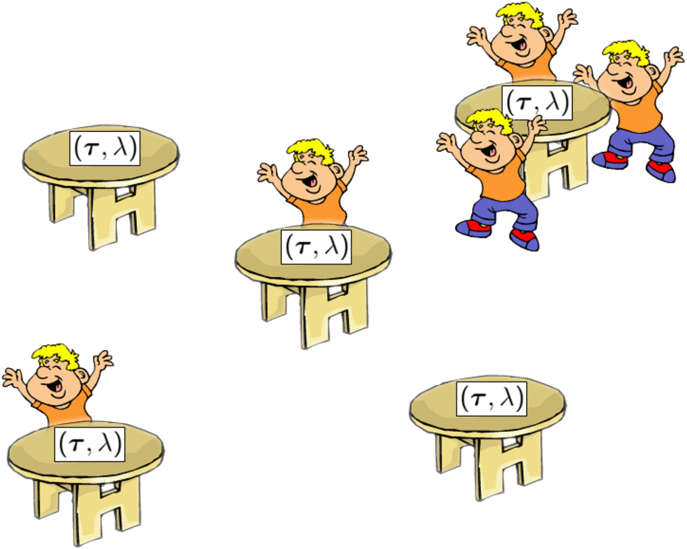

$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$

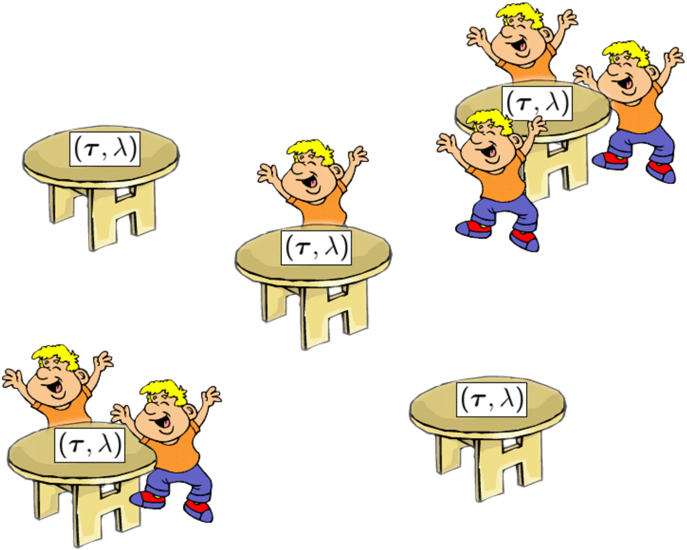

$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$


$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$


$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$


$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$
$(\boldsymbol{\tau}, \lambda)$


(τ, λ)
(τ, λ)
(τ, λ)
(τ, λ)
(τ, λ)


(τ, λ)
(τ, λ)
(τ, λ)
(τ, λ)
(τ, λ)

Model Fitting

- The $\boldsymbol{\tau}$'s and λ's may be integrated away, leaving only the clustering of the G genes.
- Sample from posterior clustering distribution using MCMC.
  - Gibbs of MacEachern (1994) and Neal (1992)
  - Merge-Split of Jain & Neal (2004)
  - Merge-Split of Dahl (2003)
- After MCMC, it's easy to sample $\boldsymbol{\tau}$'s and λ's given clustering.

Inference on Differential Expression

- Define a univariate parameter q_g that encodes the hypothesis of interest.
- For example, the global F-test in one-way ANOVA setting is analogous to:

$$q_g = \sum_{t=2}^{T} \tau_{gt}^2$$

- Estimate q_g under squared-error loss by computing its expection with respect to $p(q_g \mid d_1, \ldots, d_G)$.
- Rank genes for evidence for differential expression using the estimates $\hat{q}_1, \ldots, \hat{q}_G$.

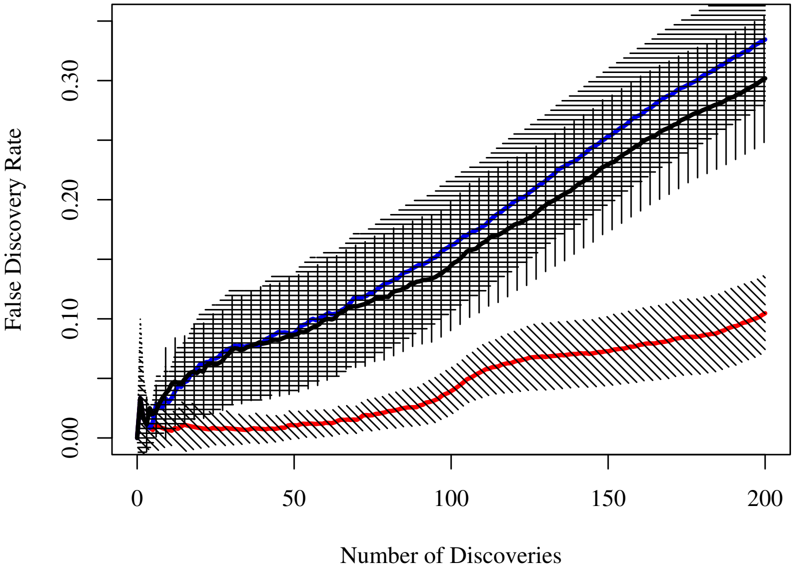
Simulation Study

- Some other methods for differential expression:
  - EBarrays (Kendziorski, Newton, et al., 2003)
  - LIMMA (Smyth 2004)
- Comparison based on proportion of false discoveries.

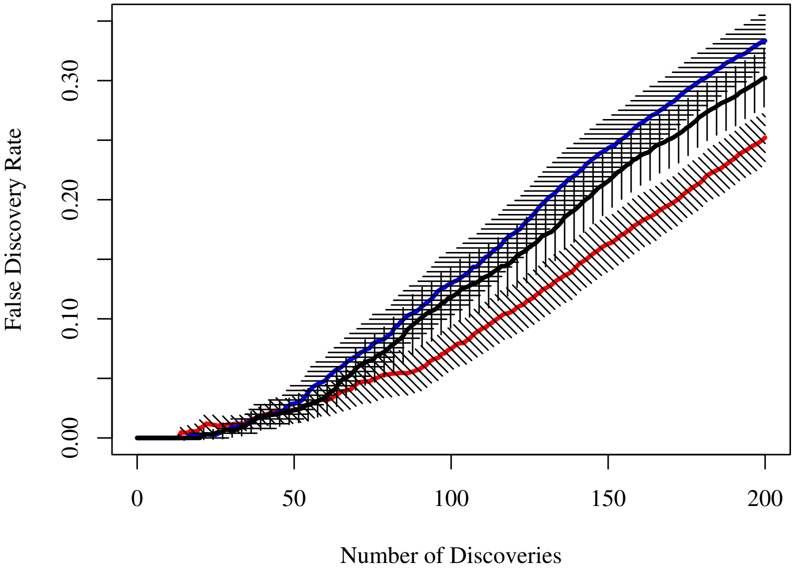
Simulation Study

- Some other methods for differential expression:
  - EBarrays (Kendziorski, Newton, et al., 2003)
  - LIMMA (Smyth 2004)
- Comparison based on proportion of false discoveries.
- Simulate datasets:
  - Time-course experiment:
    - Three time points
    - Two treatment conditions
    - 300 of 1,200 genes are differentially expressed.
  - Interest lies in genes that are differentially expressed at one or more time points.
  - Four levels of clustering:
    - Heavy Clustering: 12 clusters of 100 genes per cluster.
    - Moderate Clustering: 60 clusters of 20 genes per cluster.
    - Weak Clustering: 240 clusters of 5 genes per cluster.
    - No Clustering.

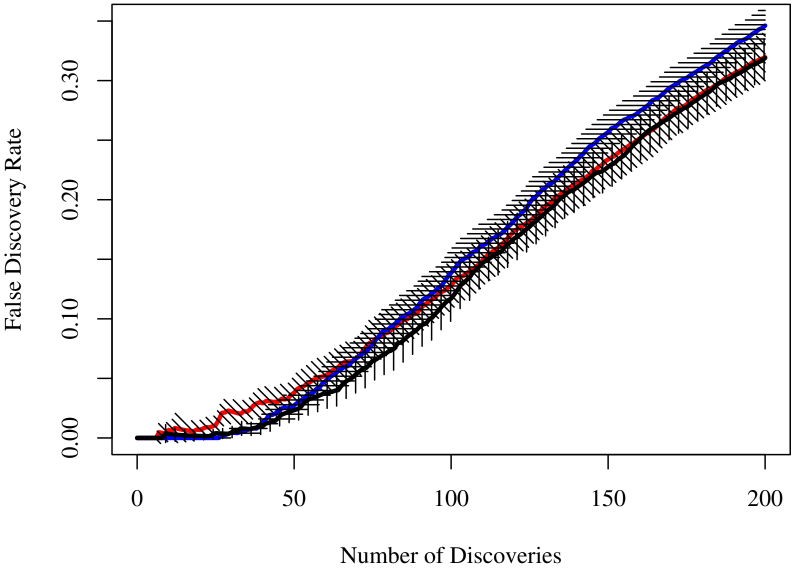
Heavy Clustering



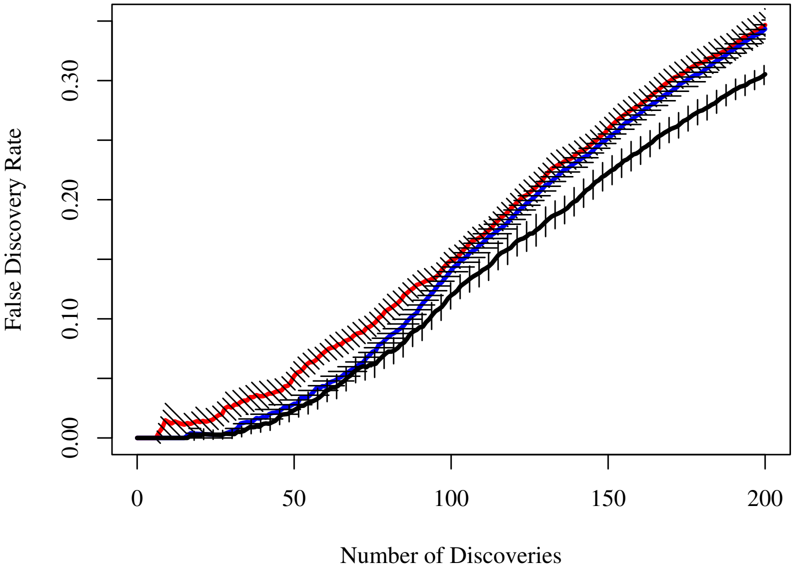
Moderate Clustering



Weak Clustering



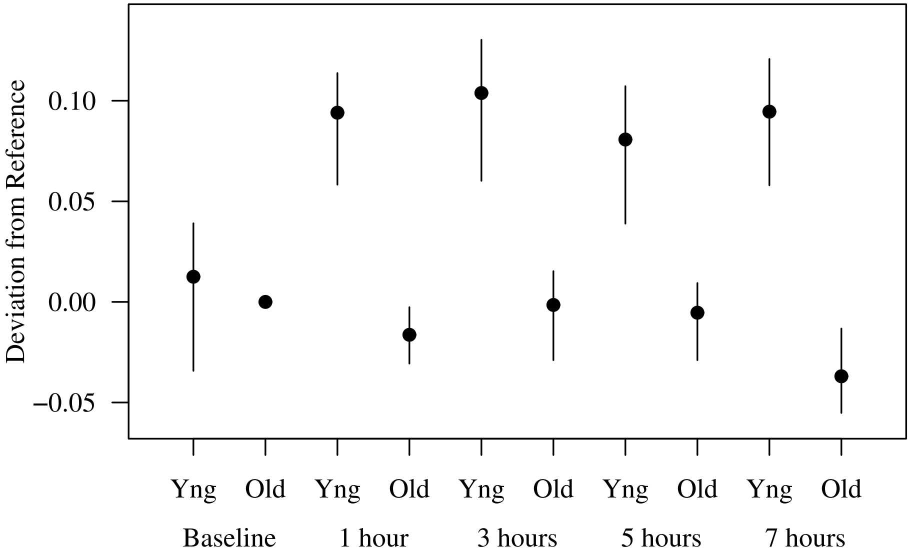
No Clustering



Paraquat Experiment

- Old and young mice treated with paraquat injection.
- Sacrifice as baseline or 1, 3, 5, or 7 hours after injection.
- Three replicates per treatment.
- 10,043 probe sets on Affymetrix MG-U74A arrays.
- Background correction and normalization using RMA (Irizarry et al., 2003).
- Biologists are interested in genes whose expression between old and young is similar at baseline and very different at one hour.

Estimated Treatment Effects for Probe Set 92885_at



Outline

1 Motivation

2 Z-scores Demonstration

3 BEMMA for Differential Expression – Dahl, Newton (200?)

4 BEMMA for Clustering – Dahl (200?)

5 SIMTAC for DE and Clustering – Dahl, Mo, Vannucci (200?)

Inference on Clustering – Dahl (200?)

- MCMC sampling algorithm produces B clusterings $\pi^{(1)}, \ldots, \pi^{(B)}$ from the posterior clustering distribution.
- Point estimation methods:
  - Maximum *a posteriori* (MAP) clustering
  - Medvedovic & Sivaganesan (2002): hierarchical clustering using pairwise probabilities
  - Dahl (2006): stochastic search to minimize posterior expected loss from Binder (1978)
  - Lau & Green (200?): heuristic to minimize posterior expected loss from Binder (1978)

Inference on Clustering – Dahl (200?)

- MCMC sampling algorithm produces B clusterings $\pi^{(1)}, \ldots, \pi^{(B)}$ from the posterior clustering distribution.
- Point estimation methods:
  - Maximum *a posteriori* (MAP) clustering
  - Medvedovic & Sivaganesan (2002): hierarchical clustering using pairwise probabilities
  - Dahl (2006): stochastic search to minimize posterior expected loss from Binder (1978)
  - Lau & Green (200?): heuristic to minimize posterior expected loss from Binder (1978)
- Selects the observed clustering closest to the matrix of pairwise probabilities in terms of squared distances:

$$\pi^{\mathrm{LS}} = \underset{\pi \in \{\pi^{(1)}, \ldots, \pi^{(B)}\}}{\arg\min} \sum_{i=1}^{G} \sum_{j=1}^{G} (\delta_{i,j}(\pi) - \hat{p}_{i,j})^2 \qquad (1)$$

Heavy Clustering

Degree of Clustering	Clustering Method	Adjusted Rand Index w/ 95% C.I.	
Heavy	MCLUST	0.413	(0.380, 0.447)
	BEMMA(least-squares)	0.402	(0.373, 0.431)
	BEMMA(map)	0.390	(0.362, 0.419)
	HCLUST(effects,average)	0.277	(0.247, 0.308)
	HCLUST(effects,complete)	0.260	(0.242, 0.279)
	HCLUST(correlation,complete)	0.162	(0.144, 0.180)
	HCLUST(correlation,average)	0.156	(0.141, 0.172)

Table: Adjusted Rand Index for BEMMA and Other Methods. Large values of the adjusted Rand index indicate better agreement between the estimated and true clustering.

Moderate Clustering

Degree of Clustering	Clustering Method	Adjusted Rand Index w/ 95% C.I.	
Moderate	BEMMA(least-squares)	0.154	(0.146, 0.163)
	MCLUST	0.144	(0.136, 0.152)
	BEMMA(map)	0.127	(0.119, 0.135)
	HCLUST(effects,complete)	0.117	(0.111, 0.123)
	HCLUST(effects,average)	0.101	(0.095, 0.107)
	HCLUST(correlation,average)	0.079	(0.075, 0.083)
	HCLUST(correlation,complete)	0.073	(0.068, 0.078)

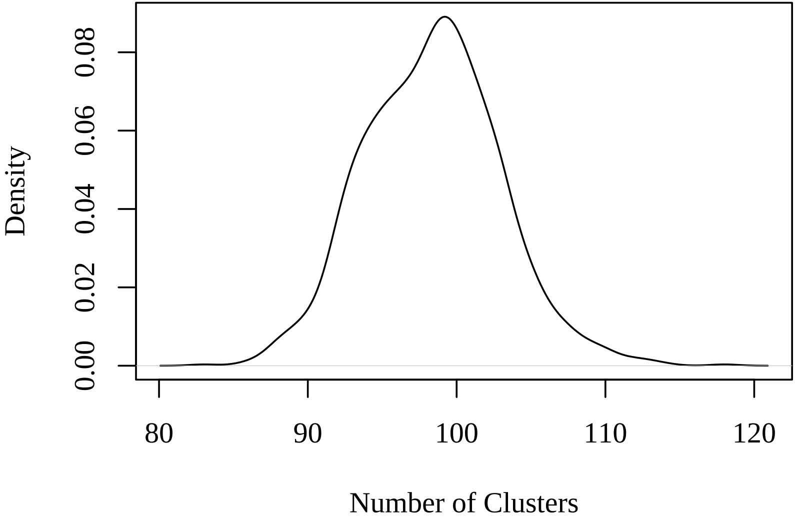
Table: Adjusted Rand Index for BEMMA and Other Methods. Large values of the adjusted Rand index indicate better agreement between the estimated and true clustering.

Weak Clustering

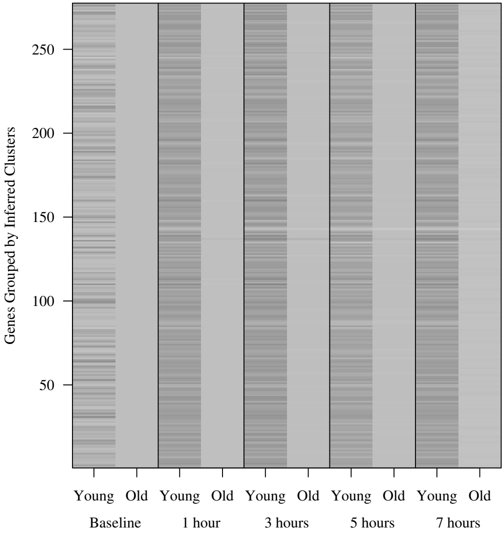
Degree of Clustering	Clustering Method	Adjusted Rand Index w/ 95% C.I.	
Weak	MCLUST	0.050	(0.048, 0.052)
	HCLUST(effects,complete)	0.045	(0.043, 0.048)
	BEMMA(least-squares)	0.042	(0.040, 0.043)
	HCLUST(effects,average)	0.037	(0.035, 0.038)
	BEMMA(map)	0.031	(0.030, 0.033)
	HCLUST(correlation,average)	0.029	(0.027, 0.030)
	HCLUST(correlation,complete)	0.027	(0.025, 0.029)

Table: Adjusted Rand Index for BEMMA and Other Methods. Large values of the adjusted Rand index indicate better agreement between the estimated and true clustering.

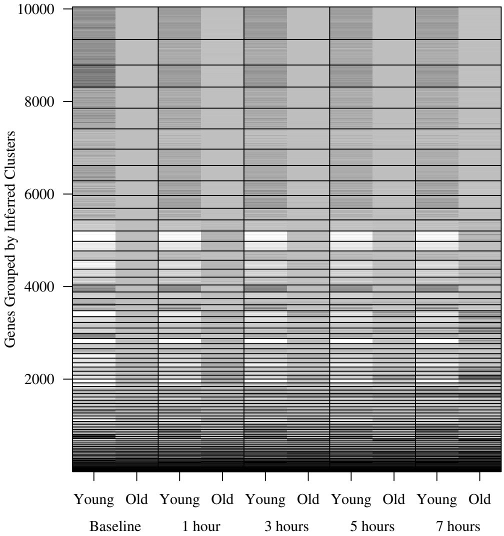
Posterior Distribution of the Number of Clusters



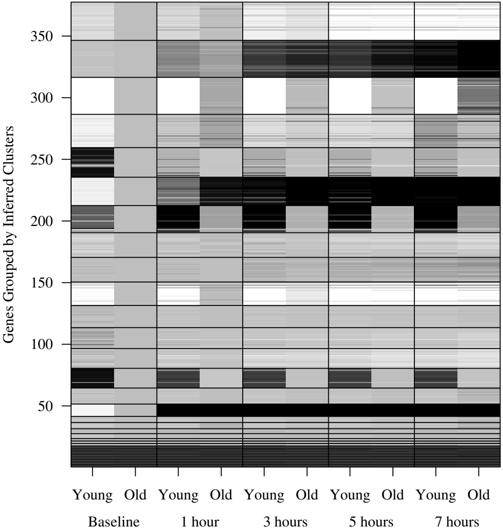
Effects Intensity Plot of Cluster of 92885_at



Effects Intensity Plot for All Clusters



Effects Intensity Plot for Smallest Clusters



Outline

1. Motivation

2. Z-scores Demonstration

3. BEMMA for Differential Expression – Dahl, Newton (200?)

4. BEMMA for Clustering – Dahl (200?)

5. SIMTAC for DE and Clustering – Dahl, Mo, Vannucci (200?)

SIMTAC – Dahl, Mo, Vannucci (200?)

- Simultaneous Inference for Multiple Testing and Clustering (SIMTAC): Extension of BEMMA
  - Separates clustering of regression coefficients from accommodation of heteroscedasticity
  - Wider class of experimental designs
  - No need to specify an arbitrary reference treatment
  - Nonconjugate Dirichlet process mixture (DPM) model
- Sampling distribution:

$$\boldsymbol{d}_g \mid \mu_g, \boldsymbol{\beta}_g, \lambda_g \sim \mathrm{N}_K\left(\boldsymbol{d}_g \mid \mu_g \boldsymbol{j} + \mathbf{X}\boldsymbol{\beta}_g, \lambda_g \mathbf{M}\right),$$

- Prior distribution:

$$\mu_g \sim \mathrm{N}\left(\mu_g \mid m_\mu, p_\mu\right)$$

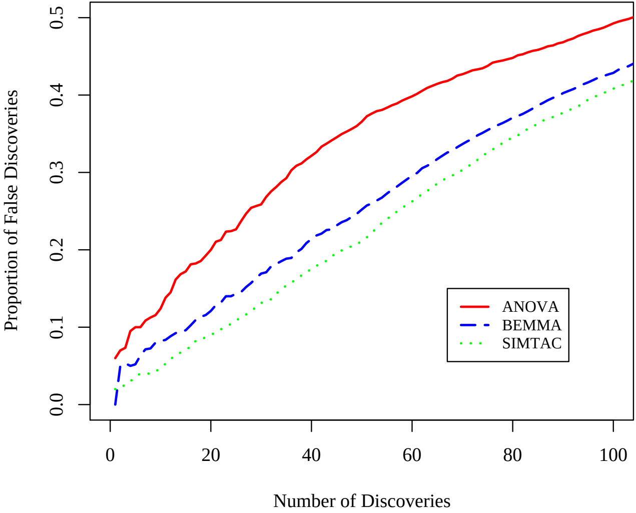
$$\begin{aligned} \boldsymbol{\beta}_g \mid G_{\boldsymbol{\beta}} &\sim G_{\boldsymbol{\beta}} \\ G_{\boldsymbol{\beta}} &\sim \mathrm{DP}\left(\alpha_{\boldsymbol{\beta}} G^{\star}_{\boldsymbol{\beta}}\right) \end{aligned} \qquad \begin{aligned} \lambda_g \mid G_\lambda &\sim G_\lambda \\ G_\lambda &\sim \mathrm{DP}\left(\alpha_\lambda G^{\star}_\lambda\right) \end{aligned}$$

Simulation Study

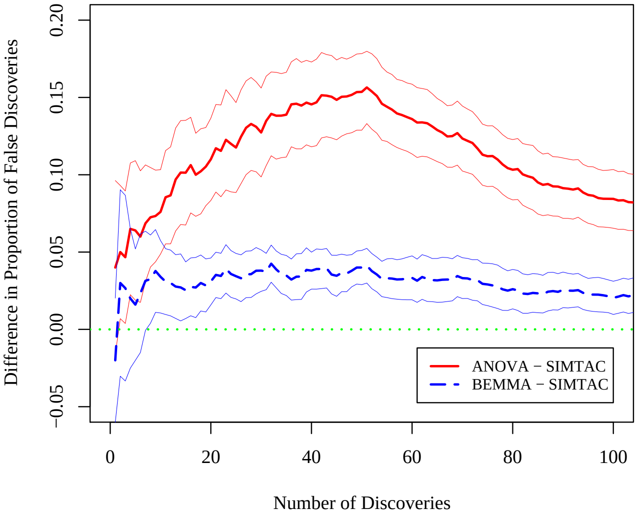
Size of Each Cluster	Relationship of Regression Coefficients Encoding Equivalent and Differential Expression			Number of Clusters with this Configuration
	Time Point 1	Time Point 2	Time Point 3	
120	$\beta_{g,3} = 0$	$\beta_{g,4} = \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	1
40	$\beta_{g,3} = 0$	$\beta_{g,4} = \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	2
40	$\beta_{g,3} = 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	1
15	$\beta_{g,3} = 0$	$\beta_{g,4} = \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	6
15	$\beta_{g,3} = 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	1
15	$\beta_{g,3} = 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} \neq \beta_{g,2}$	1
5	$\beta_{g,3} = 0$	$\beta_{g,4} = \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	19
5	$\beta_{g,3} = 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	2
5	$\beta_{g,3} = 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} \neq \beta_{g,2}$	2
5	$\beta_{g,3} \neq 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} \neq \beta_{g,2}$	1
2	$\beta_{g,3} = 0$	$\beta_{g,4} = \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	48
2	$\beta_{g,3} = 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	4
2	$\beta_{g,3} = 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} \neq \beta_{g,2}$	4
2	$\beta_{g,3} \neq 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} \neq \beta_{g,2}$	4
1	$\beta_{g,3} = 0$	$\beta_{g,4} = \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	95
1	$\beta_{g,3} = 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	5
1	$\beta_{g,3} = 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} \neq \beta_{g,2}$	5
1	$\beta_{g,3} \neq 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} \neq \beta_{g,2}$	5
1	$\beta_{g,3} = 0$	$\beta_{g,4} = \beta_{g,1}$	$\beta_{g,5} \neq \beta_{g,2}$	5
1	$\beta_{g,3} \neq 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	5

Table: **Clusters in a Synthetic Dataset.** For the 216 clusters in each synthetic dataset, this table shows the relationship among and the cluster sizes for the regression coefficients.

Proportion of False Discoveries



Difference in Proportion of False Discoveries



Summary

- Dependence can be exploited to improve power in multiple testing.
- Dirichlet process mixture (DPM) models provide a powerful machinery to accomplish simultaneous inference on clustering and multiple hypothesis testing.
- BEMMA:
  - Under weak clustering, BEMMA performs as well as its peers.
  - Under heavier clustering, BEMMA performs substantially better.
  - BEMMA has been successfully applied to a replicated microarray study with 10,000+ probesets and 10 treatment conditions.
- SIMTAC:
  - Improvemed implementation of the the idea.
  - Simulation results are encouraging... now applying to local data.
- Least-squares clustering:
  - Convenient and conceptually appealing procedure for point estimation of clustering.