

A Bayesian Method for Partially Paired High Dimensional Data

Fei Liu, Feng Liang, Woncheol Jang

Institute of Statistics and Decision Sciences
Duke University

SAMSI, Summer 2006

Outline

- ▶ Bayesian methods have been developed for paired high dimensional data such as gene expression data.
- ▶ For partially paired data, however, excluding those unpaired observations for the analysis may lead to significant information loss.
- ▶ Using test statistics with FDR control is a possible solution.
- ▶ We provides a generalized Bayesian method for partially paired high dimensional data.

Statistical Model

- ▶ The data for j th gene are arranged as:

$$\begin{array}{l} X_{1j}, \dots, X_{nj}; \quad X_{1j}^*, \dots, X_{n_1j}^*; \\ Y_{1j}, \dots, Y_{nj}; \quad Y_{1j}^*, \dots, Y_{n_2j}^*. \end{array}$$

(X_{ij}, Y_{ij}) : paired gene expressions. X_{ij}^*, Y_{ij}^* : unpaired observations.

- ▶ $(X_{ij}, Y_{ij})^T \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$

$$\boldsymbol{\mu}_j = \begin{pmatrix} \mu_j \\ \mu_j + \delta_j \end{pmatrix}, \boldsymbol{\Sigma}_j = \begin{pmatrix} \sigma_1^2 & \rho_j \sigma_1 \sigma_2 \\ \rho_j \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}.$$

- ▶ For the incomplete data

$$X_{ij}^* \sim N(\mu_j, \sigma_1^2), Y_{ij}^* \sim N(\mu_j + \delta_j, \sigma_2^2)$$

Review the FDR Control

	Accept	Reject	Total
True Null	U	V (False positive)	m_0
Untrue Null	T (False negative)	S	$m - m_0$
Total	W	R	m

$$\text{FDR} = \mathbb{E}(V/R \mid R \neq 0)$$

Benjamini and Hochberg(1995) procedure to control FDR at q^* :

$$\begin{pmatrix} H_1 & H_2 & \dots & H_m \\ p_1 & p_2 & \dots & p_m \end{pmatrix} \implies \begin{pmatrix} H_{(1)} & H_{(2)} & \dots & H_{(m)} \\ p_{(1)} & p_{(2)} & \dots & p_{(m)} \end{pmatrix}$$

$k = \max(i, p_{(j)} \leq \frac{j}{m} q^*, \text{ for all } j \leq i)$, Reject $H_{(1):(k)}$.

Test Statistics for Partially Paired Data in One Dimensional Space

- ▶ Lin and Stivers (1974) use the test statistic when n is small and $|\rho| \leq 0.5$:

$$T = \frac{\bar{x}_1^{(n+n_1)} - \bar{x}_2^{(n+n_2)}}{\sqrt{\frac{1}{n+n_1} + \frac{1}{n+n_2} - \frac{2nr}{(n+n_1)(n+n_2)} \sqrt{(a_{11}^* + b_{22})/(N-2)}}$$

where $T \sim t_{N-4}$ approximately, and $N = n + n_1 + n_2$.

- ▶ Another test statistic is given by the mixed effect model:

$$Z_{ik} = \mu + \alpha_j + \beta_k + \epsilon_{ik}, \text{ where } \beta_k \sim N(0, \sigma_\beta^2), \epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$$

Perform ANOVA to test the fixed effect $\alpha_j = 0$.

Scott and Berger (2003)

Noticing the built-in penalty (“Ockham’s razor effect”) of the Bayesian method, Scott and Berger (2003) propose A Bayesian Hierarchical model for multiple comparisons,

Observe $\mathbf{x} = (x_1, \dots, x_M)$:

$$x_j \sim \mathbf{N}(\mu_j, \sigma^2)$$

$$\gamma_j = 1 - \delta(\mu_j = 0)$$

$$f(\mathbf{x} \mid \sigma^2, \gamma, \mu) = \prod_{j=1}^M \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_j - \gamma_j\mu_j)^2}{2\sigma^2}\right)$$

$$\mu_j \sim \mathbf{N}(0, V)$$

$$\pi(V, \sigma^2) \propto (V + \sigma^2)^{-2}$$

$$\gamma_j \sim \text{Bernoulli}(p)$$

$$\pi(p) \sim \text{Beta}(\alpha, \beta)$$

EBarrays Method

Kendziorski et. al (2004) propose Parametric Empirical Bayes Method to account for replicating arrays (multiple conditions as well).

Observe $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_J)$, where $\mathbf{x}_j = (x_{j1}, x_{j2}, \dots, x_{jl})$

- ▶ If gene j is not differentially expressed ($\delta_j = 0$),

$$f_0(\mathbf{x}_j) = \int \left(\prod_{i=1}^l f_{obs}(x_{ji} | \mu) \right) \pi(\mu) d_\mu$$

- ▶ If gene j is differentially expressed ($\delta_j \neq 0$),

$$f_1(\mathbf{x}_j) = f_0(\mathbf{x}_{j1}) f_0(\mathbf{x}_{j2})$$

- ▶ Data is marginally distributed: $pf_1(\mathbf{x}_j) + (1 - p)f_0(\mathbf{x}_j)$
- ▶ By Bayes' rule, posterior probability of $\delta_j \neq 0$ is

$$\frac{pf_1(\mathbf{x}_j)}{pf_1(\mathbf{x}_j) + (1 - p)f_0(\mathbf{x}_j)}$$

Mixture Prior

- ▶ Our primary interest is:

$$H_0 : \delta_j = 0$$

- ▶ We propose a mixture distribution for δ_j , i.e.,

$$\pi(\delta_j | p, \tau^2) = p\phi(\delta_j/\tau) + (1 - p)I_{\{0\}}(\delta_j),$$

p : probability of being differentially expressed.

γ_j : Latent variables. Set to 1 if the j th gene is differentially expressed; otherwise 0. Interest $P(\gamma_j = 1 | \text{Data})$.

Priors and Posteriors

- ▶ Priors distributions for $(\mu, \sigma^2, \rho, \tau^2)$ are:

$$\begin{aligned}\pi(\mu_j) &\propto 1 \\ \pi(\tau^2 | \sigma^2) &\propto \frac{1}{\sigma^2} \left(1 + \frac{\tau^2}{\sigma^2}\right)^{-2} \\ \pi(\rho) &\propto \rho^{\alpha-1} (1 - \rho)^{\beta-1} \equiv \text{Beta}(\alpha, \beta)\end{aligned}$$

- ▶ Improper prior distributions for ρ

$$\pi_1(\rho_j) \propto \frac{1}{(1 - \rho_j^2)}, \quad \pi_2(\rho_j) \propto \frac{1}{(1 - \rho_j^2)^2}$$

π_1 and π_2 are both can be shown to have proper posteriors.

Bayarri(1981) shows that π_1 avoids the “Jeffrey-Lindley” paradox.

Gibbs Sampling

$$\Theta = (\boldsymbol{\mu}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\rho}, \sigma^2, \boldsymbol{\tau}^2, \boldsymbol{\rho}), \text{Data} = (\mathbf{x}, \mathbf{y}, \mathbf{x}^*, \mathbf{y}^*)$$

Closed forms for sampling $\boldsymbol{\mu}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \sigma^2$:

$$(\mu_j \mid \Theta_{-\mu}, \text{Data}) \sim N(m_j^{(\mu)}, \sigma_j^{(\mu)})$$

$$(\gamma_j \mid \Theta_{-(\boldsymbol{\gamma} \cup \boldsymbol{\delta})}, \text{Data}) \sim \text{Bernoulli}(p_j^{(\gamma)})$$

$$(\delta_j \mid \Theta_{-\delta}, \text{Data}) \sim N(m_j^{(\delta)}, \sigma_j^{(\delta)})$$

$$(\rho \mid \Theta_{-\rho}, \text{Data}) \sim \text{Beta} \left(\alpha + \sum_{j=1}^J \gamma_j, \beta + J - \sum_{j=1}^J \gamma_j \right)$$

$$(\sigma^2 \mid \Theta_{-\sigma^2}, \text{Data}) \sim \text{IG} \left(J \left(n + \frac{n_1 + n_2}{2} \right), \eta \right)$$

No closed forms for $(\boldsymbol{\tau}^2, \boldsymbol{\rho})$.

Simulation Study with Normal Distributions

Simulate the data with 1000 genes, 5 paired, 2 unpaired control, 2 unpaired treatment, and

$$\rho = 0.1, \tau^2 = 100, \sigma^2 = 1.0, p = 0.01.$$

	False Positive	False Negative
FDR - T test	0/9	2/991
FDR - random effect	1/11	1/989
Bayesian Model	0/10	1/990

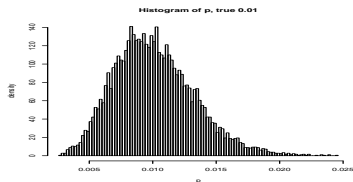


Figure: Posterior distribution of p (true is 0.01)

Simulation Study with Normal Distributions (Cont...)

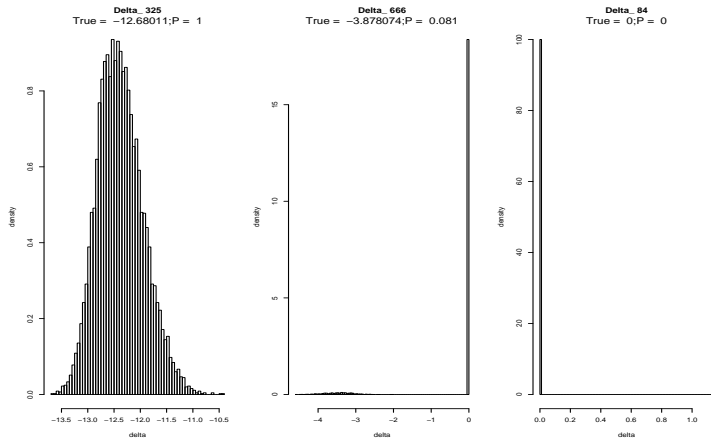


Figure: Posterior distribution for δ 's

Simulation Study with t Distributions

Simulate the data with 1000 genes, 9 samples (5 pairs, 2 unpaired control, 2 unpaired treatment)

Data = Bivariate T_4 with mean 0 and $\Sigma = \begin{pmatrix} 0.1 & 1 \\ 1 & 0.1 \end{pmatrix}$
 $+ \mu + \delta$

$\mu \sim U(-0.01, 0.01)$

$\delta_i \mid \delta_i \neq 0 \sim N(0, \tau^2 = 100)$;

$P(\delta_i \neq 0) = 0.01$

	False Positive	False Negative
FDR - T test	1/7	3/993
FDR - random effect	6/13	2/987
Bayesian Model	6/13	2/987

Simulation Study with t Distributions (Cont...)

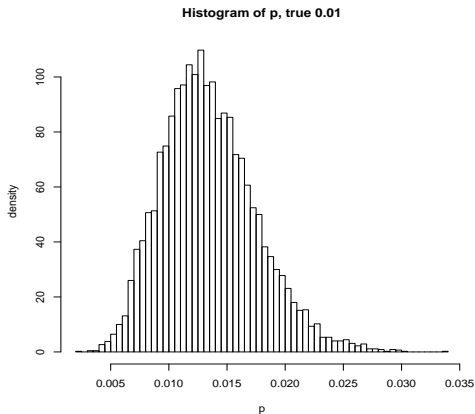


Figure: Posterior distribution for p

Simulation Study with t Distributions (Cont...)

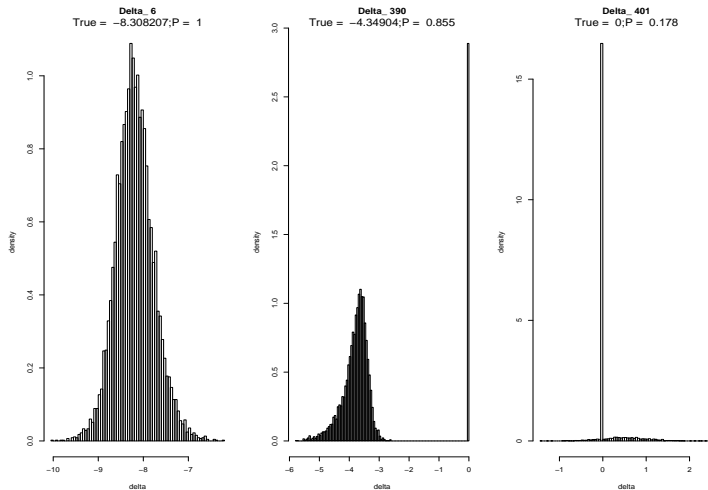


Figure: Posterior distribution for δ 's

Future work

- ▶ Apply the method to gene expression data.
- ▶ Use different ρ to achieve different thresholding.
- ▶ EBarrays method with random effects e_k .

Observe $\mathbf{X}_{jk} = (\mathbf{X}_{jk1}, \mathbf{X}_{jk2})$ for gene j and sample k .

- If gene j is not differentially expressed,

$$f_0(\mathbf{X}_{jk}) = \int \prod_i \left(\int \prod_k f_{obs}(X_{jki} | \mu + e_k) \pi(\mu) \pi(e_k) d_{e_k} \right) d_{\mu}$$

- If gene j is differentially expressed,

$$f_0(\mathbf{X}_{jk}) = \int \prod_i \left(\int \prod_k f_{obs}(X_{jki} | \mu_i + e_k) \pi(\mu_i) \pi(e_k) d_{e_k} \right) d_{\mu_i}$$