

A Bayesian Method for Partially Paired Gene Expression Data

Fei Liu¹ Woncheol Jang¹ Feng Liang¹

¹Institute of Statistics and Decision Sciences
Duke University

SAMSI 2006 Summer Program
on Multiplicity and Reproducibility in Scientific Studies

Abstract

Microarray technology is often applied in control studies where measurements are simultaneously taken for thousands of gene expressions on matched pairs. For completely paired data, Bayesian methods provide a natural solution with small sample size. However, the Gene expression data often have missing values which causes part of the observations unmatched. For the small sample size as in the context of gene expression data, excluding those unpaired observations for the analysis may lead to significant information loss. We generalize the Bayesian method for partially paired gene expressions, which provides a natural way to incorporate the information from the unmatched observations for a better estimation.

The Data

- Paired experiment:
 - Suppose we conduct a certain treatment on n patients.
 - Isolate RNA from each sample and sent to the Microarray facility
- Paired and unpaired data
 - 3 samples failed, so data are from 9 samples.
 - 3 paired data, 2 from **GFP-**, and 1 from **GFP+**.

Statistical Model

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

$$\begin{array}{ll} X_{1j}, \dots, X_{nj}; & X_{1j}^*, \dots, X_{n_{1j}}^*; \\ Y_{1j}, \dots, Y_{nj}; & Y_{1j}^*, \dots, Y_{n_{2j}}^*. \end{array}$$

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

- $(X_i, Y_i)^T \sim N(\mu, \Sigma)$ where $\mu = (\mu_j, \mu_j + \delta_j)^T$ and

$$\Sigma_j = \sigma^2 \begin{pmatrix} 1 & \rho_j \\ \rho_j & 1 \end{pmatrix}.$$

- $X_{ij}^* \sim N(\mu_j, \sigma^2)$ and $Y_{ij}^* \sim N(\mu_j + \delta_j, \sigma^2)$

Mixture Model

- The primary interest here is to test the difference in gene expression δ_j .
- δ_j is modeled by a mixture of normal and point mass at 0:

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

where p denotes the probability of being differentially expressed (DE).

- Latent variable $\gamma_j = 1$ if j th gene is DE; otherwise 0.

Priors and Posteriors

- Priors distributions for $(\boldsymbol{\mu}, \sigma^2, \boldsymbol{\rho}, p, \tau^2)$ are specified as follows:

$$\pi(\boldsymbol{\mu}, \sigma^2, \boldsymbol{\rho}) \propto \frac{1}{\sigma^2} \prod_{j=1}^J \frac{1}{(1 - \rho_j^2)^2} \quad (*)$$

$$\pi(\tau^2 \mid \sigma^2) \propto \frac{1}{\sigma^2} \left(1 + \frac{\tau^2}{\sigma^2}\right)^{-2}$$

$$\pi(p) \propto p^{\alpha-1} (1-p)^{\beta-1} \equiv \text{Beta}(\alpha, \beta)$$

- Though improper, $(*)$ is a reference prior, and it can be shown to have proper posterior.
- We are interested in the posterior $P(\gamma_j = 1 \mid \text{Data})$.

Gibbs Sampling

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

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

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

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

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

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

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

- No closed forms for $(\tau^2, \boldsymbol{\rho})$. Use Metropolis-Hastings algorithm to sample $(\tau^2 \mid \Theta_{-\tau^2}, \text{Data})$, and $(\rho_j \mid \Theta_{-\boldsymbol{\rho}}, \text{Data})$.

Simulation Study I

Data is exactly generated from the aforementioned Bayesian Hierarchical model by adding predetermined mean shifts (simulated from Normal) to samples from group II.

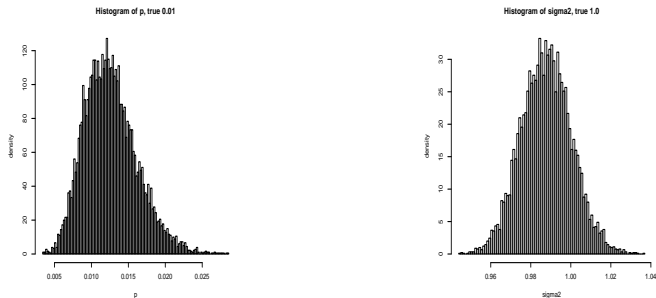


Figure: Left panel is the posterior distribution for p (true is 0.01) and Right panel is for σ^2 (true is 1.0).

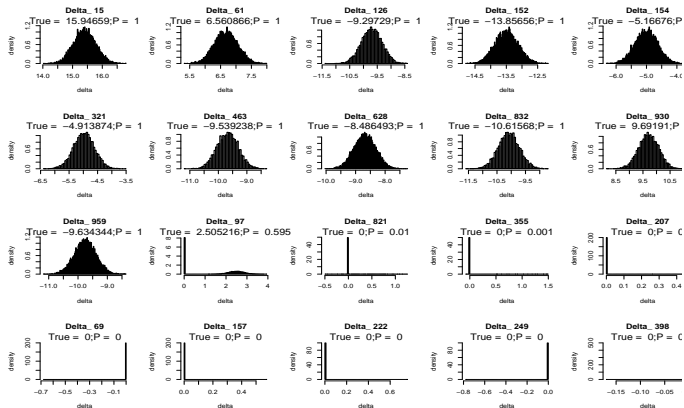
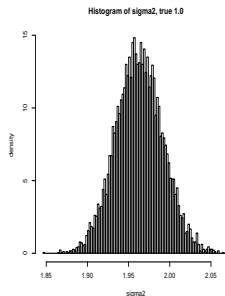
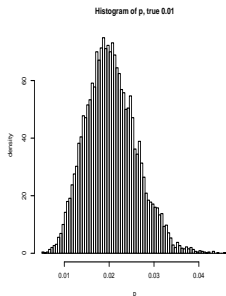


Figure: posterior distributions for the top 20 differentially expressed δ . We get no false discovery comparing with the truth

Simulation Study II

- To learn the robustness of our analysis, we simulate data from Bivariate t-distribution, and adding predetermined mean shifts (simulated from Normal) to samples from group II.



- Experimental results showed that we had false discoveries comparing with the truth.

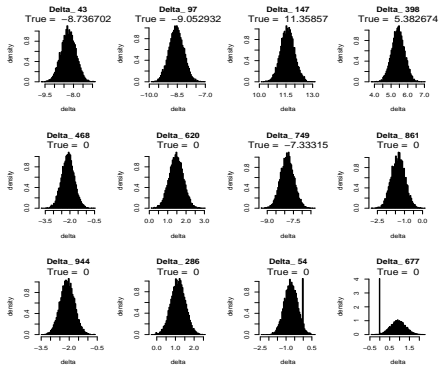


Figure: posterior distributions for the top 12 differentially expressed δ .
 Note - we are getting false discoveries comparing with the truth

Results of the HSC Data

- After screening, reduce 32841 genes to 8763 genes.
- 120 DE genes are identified with $P(\gamma_j = 1 \mid \text{data}) > 0.5$; among them 84 genes are very significant with $P(\gamma_j = 1 \mid \text{data}) > 0.9$.

