

**Model selection and fitting for
Empirical Bayes analysis of
microarray data**

by B. Caffo and D. Liu and G. Parmigiani
Johns Hopkins University
Department of Biostatistics
bcaffo@jhsph.edu

Goals of the presentation

- Present a conjugate hierarchical model useful in analyzing normalized output from microarray experiments
- Discuss benefits/pitfalls of assuming conjugacy
- Embed the conjugate model into a larger class of models

Diagnose departures from conjugacy

Alternative model should conjugacy not be supported by the data

- Fitting

Some example Affymetrix data

| Group A | | | Group B | | |
|---------|------|-----|---------|------|-----|
| 7.4 | 7.6 | 7.3 | 7.6 | 7.7 | 7.5 |
| 5.0 | 5.4 | 5.3 | 5.3 | 5.5 | 4.9 |
| 4.7 | 4.8 | 4.8 | 4.8 | 5.0 | 4.6 |
| 6.5 | 6.8 | 6.6 | 6.7 | 6.7 | 6.5 |
| 7.2 | 7.4 | 7.0 | 7.1 | 7.4 | 7.3 |
| 8.8 | 9.1 | 8.6 | 8.7 | 9.0 | 8.9 |
| 4.9 | 4.8 | 5.0 | 4.8 | 5.2 | 4.8 |
| 7.4 | 7.7 | 7.5 | 7.5 | 7.8 | 7.4 |
| 8.7 | 9.1 | 8.7 | 9.0 | 9.2 | 8.6 |
| 9.9 | 10.2 | 9.9 | 9.9 | 10.1 | 9.9 |

Reference: Irizarry et al. (2001 Biostatistics)

Proposed model

- Model $Y_g \sim \text{MN}(X\beta_g, \theta_g I)$
- $g = 1, \dots, G$ where $G = \text{number of genes}$
- For our example

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}$$

- Note we assume X does not depend on g so that we preclude the possibility of studying gene \times gene interactions

Hierarchical model

Recall $Y_g | \beta_g \theta_g \sim \text{MN}(X \beta_g, \theta_g I)$

- Specify a conjugate hierarchical model with

$$\beta_g | \theta_g \sim \text{MN}(\mu, F \theta_g)$$

and

$$\theta_g \sim \text{IG}(\nu, \tau)$$

- Obtain EB estimates of ν , τ , μ and F using EM

Closed form “E” step

“M” step for F and μ also has a closed form

$$Y_g | \beta_g \theta_g \sim \text{MN}(X \beta_g, \theta_g I)$$

$$\beta_g | \theta_g \sim \text{MN}(\mu, F \theta_g)$$

$$\theta_g \sim \text{IG}(\nu, \tau)$$

Potential problems with conjugate model

- Dependence between β_g and θ_g
 - Independence between β_g / θ_g and θ_g
- Marginal distribution for β_g is conservatively diffuse (t)
- Simple generalization assumes

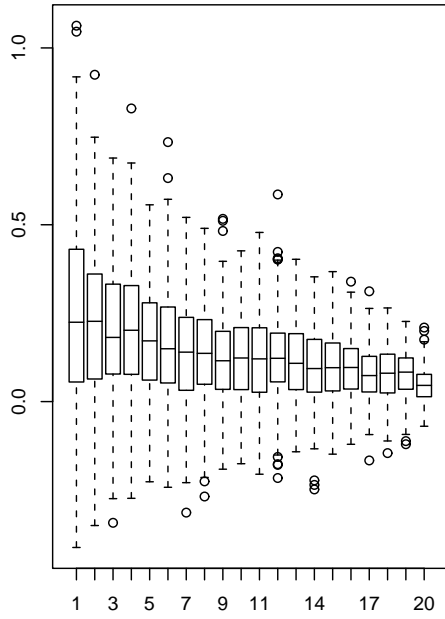
$$\beta_g | \theta_g \sim \text{MN}(\mu, F \theta_g^\delta)$$

$\delta = 1$ conjugate model

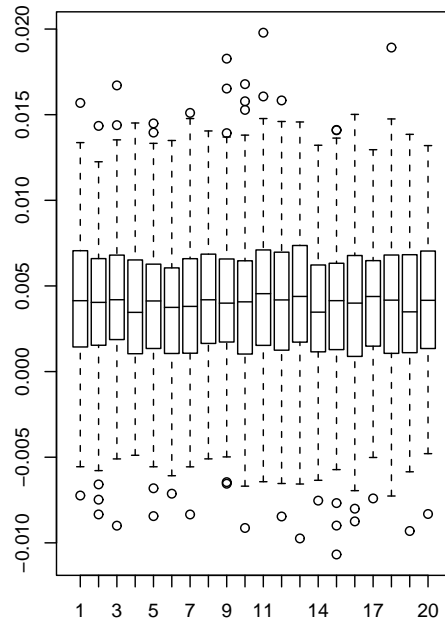
$\delta = 0$ independence

Role of δ

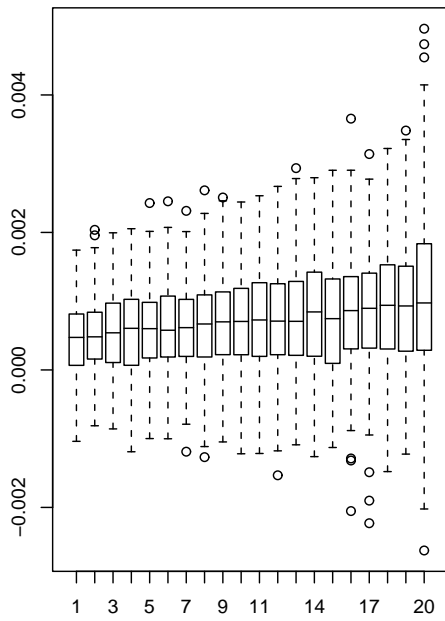
delta = -2



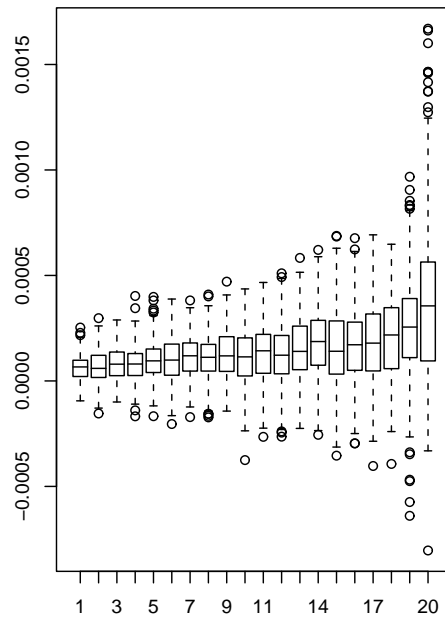
delta = 0



delta = 1

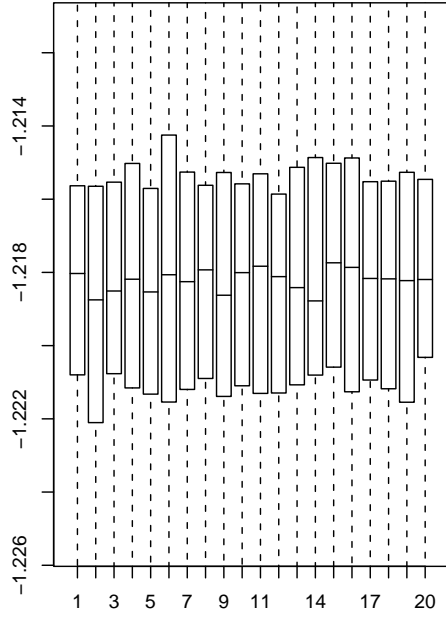


delta = 2

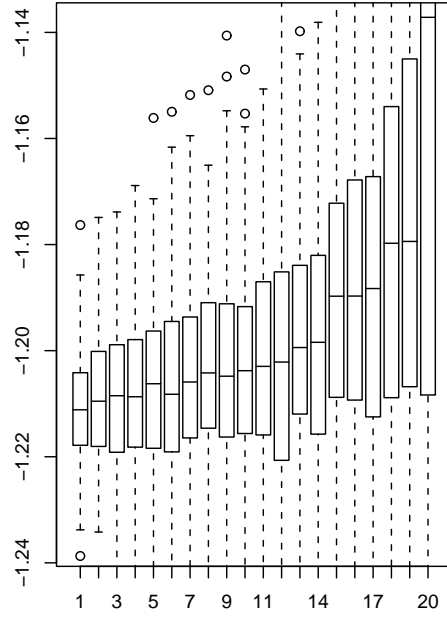


Role of δ

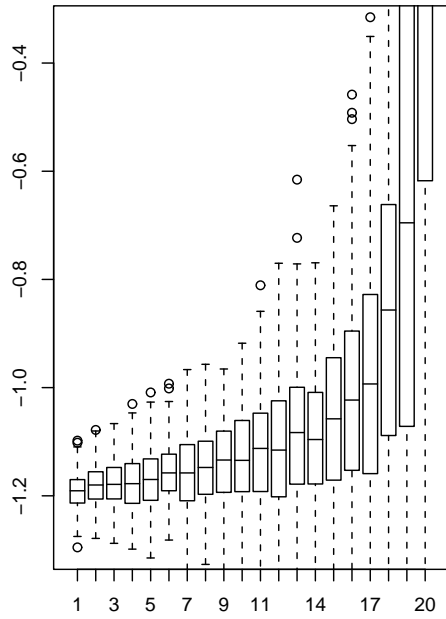
delta = 0



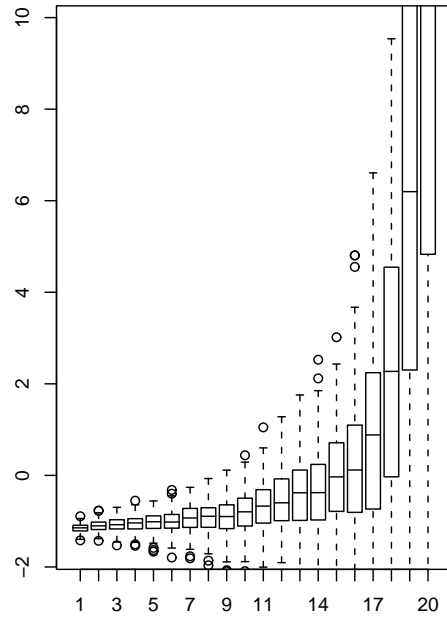
delta = 0.5



delta = 1



delta = 1.5



Fitting via EM

- Profile likelihood for δ

Repeatedly fix values of δ and calculate estimates of F , μ , ν and τ

Fit via EM

- EM algorithm treats β_g and θ_g as missing data
- Let E^* denote expectation conditional on y_g and the current parameter estimates
- It turns out that the next values for μ and F are

$$\mu^{(t)} = \sum_{g=1}^G E^*[\beta_g]/G$$

and

$$F^{(t)} = \sum_{g=1}^G E^*[(\beta_g - \mu^{(t)})(\beta_g - \mu^{(t)})'\theta_g^{-\delta}]/G$$

Calculating ν and τ

- Next value of ν and τ maximize

$$\nu \log \tau - \log \Gamma(\nu) - \frac{\nu}{G} \sum_{g=1}^G E_t^*[\log \theta_g] - \frac{\tau}{G} \sum_{g=1}^G E_t^*[\theta_g^{-1}]$$

Efficient maximization technique can be obtained by modifying methods for maximizing gamma likelihoods (Johnson and Kotz CUD2)

Details on the E-step

- For general $\delta \neq 1$, all of the required expectations are intractable
- Need *fast* (and accurate) ways to approximate these expectations
- Via suitable transformations, each of these expectations can be written in the form

$$\frac{\int_0^\infty f(u)h(u)u^\alpha e^{-u} du}{\int_0^\infty h(u)u^\alpha e^{-u} du}$$

- Accurate approximation to such integrals is given by *Gauss-Laguerre* quadrature (Abramowitz and Stegun).

Gauss-Laguerre quadrature

- Approximate

$$\int_0^{\infty} f(u)h(u)u^{\alpha}e^{-u}du$$

with

$$\sum_{i=1}^n f(n_i)h(n_i)w_i$$

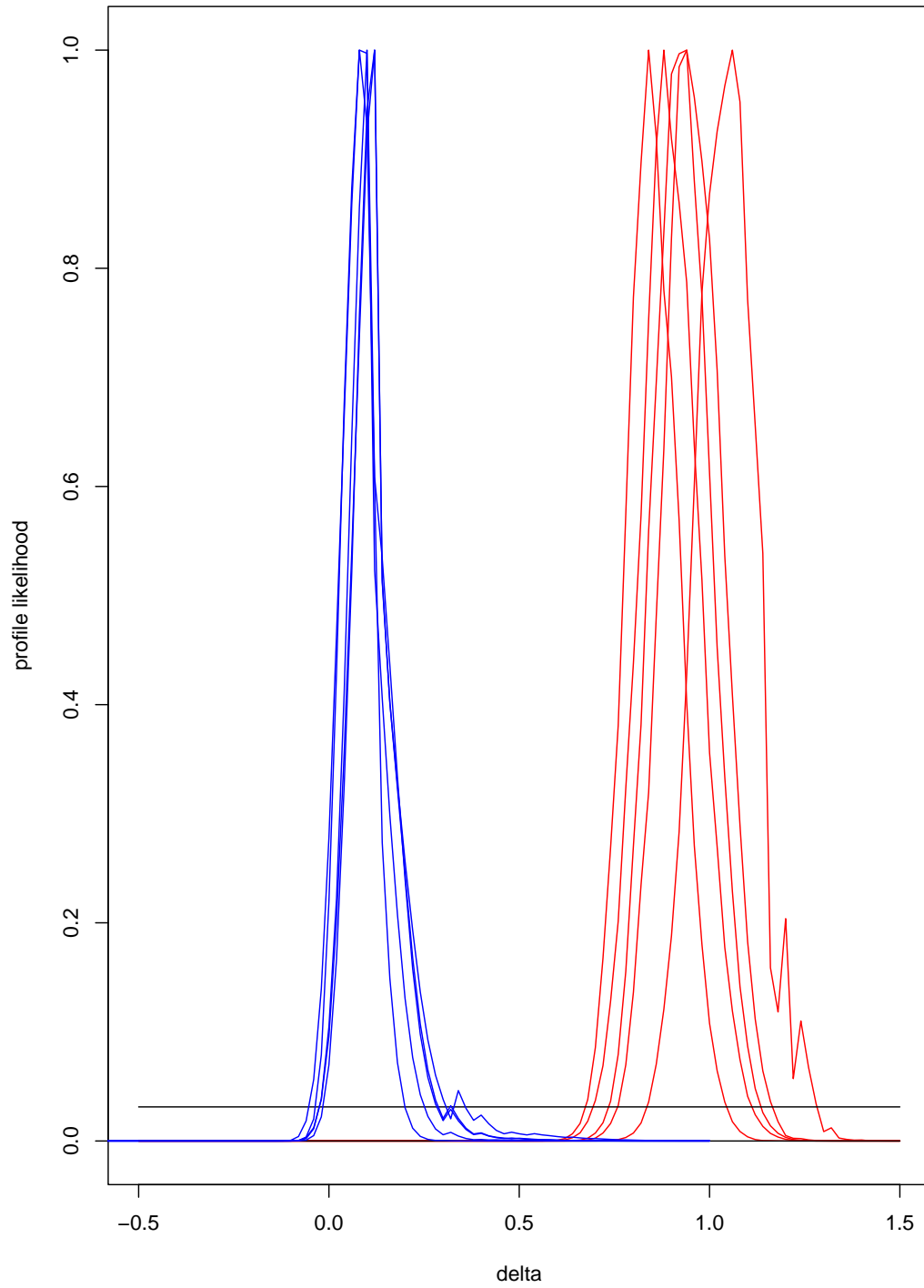
for **weights** w_i and **nodes** n_i .

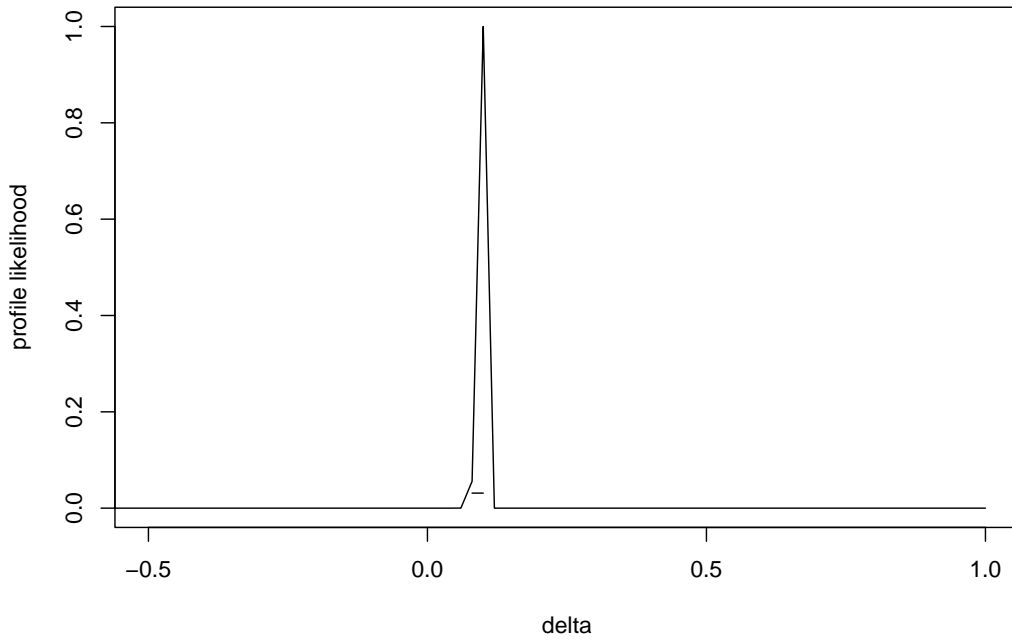
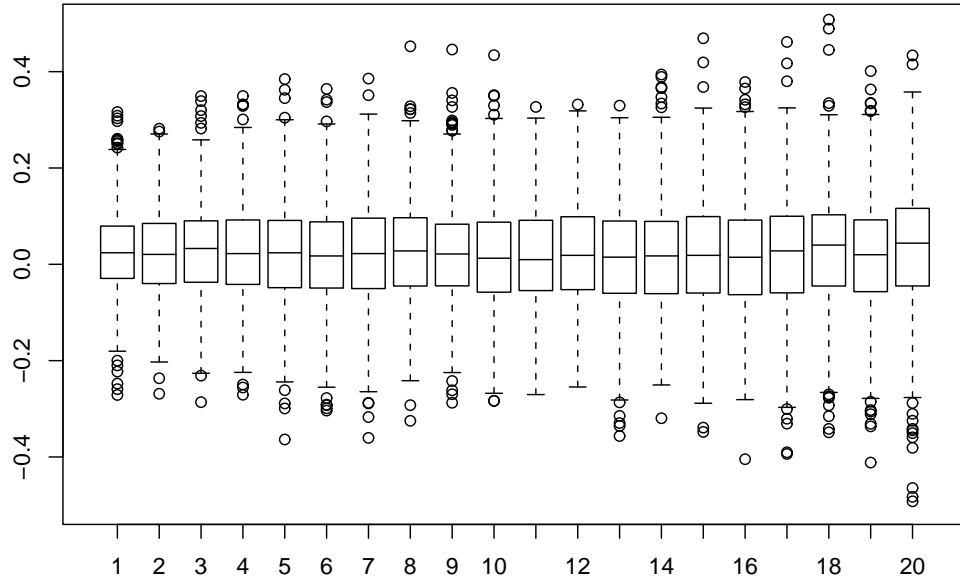
- Weights and nodes are constructed so that the approximation is exact if h is a polynomial of degree n or lower
- Fast algorithms exist for calculating w_i and n_i (Press et al. *Numerical Recipes in C*)
- To make h closer to constant, I multiply and divide the integrand by the h corresponding to the conjugate model

Details

- EM
 - ▶ EM performed this way requires exactly 1 pass through the gene index per “E” step
 - ▶ No need to sum over the entire gene index
 - ▶ Working with the sufficient statistics, $y'_g y_g$ and $X' y_g$ only, saves a lot of computing time
- Calculate the profile likelihood
 - ▶ Start at the conjugate model $\delta = 1$
 - ▶ At the next δ use the fitted values from the previous δ as starting values

Profile likelihoods for δ for simulated data





Concluding remarks

- It is difficult to diagnose conjugacy via plots
- The proposed model can diagnose conjugacy and offers a solution should conjugacy fail
- EM using Gauss-Laguerre integration is an effective way to fit the model