# Parametric Empirical Bayes Methods for Microarrays

3/7/2011





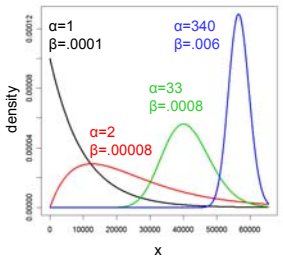
## Parametric Empirical Bayes Methods for Microarrays

- Kendziorski, C. M., Newton, M. A., Lan, H., Gould, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*. **22**, 3899-3914.
- Newton, M. A. and Kendziorski, C. M. (2003). Parametric empirical Bayes methods for microarrays. Chapter 11 of *The Analysis of Gene Expression Data*. Springer. New York.



## The Gamma Distribution

- $X \sim \text{Gamma}(\alpha, \beta)$
- $f(x)= \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1}e^{-\beta x}$ for $x>0$.
- $E(X)=\alpha / \beta$
- $\text{Var}(X)= \alpha / \beta^2$






## A Model for the Data from a Two-Treatment Experiment

- Assume there are J genes indexed by j=1, 2, ..., J.
- Data for gene j is $\mathbf{x}_j = (x_{j1}, x_{j2}, ..., x_{jI})$ where $x_{ji}$ is the normalized measure of expression on the original scale for the $j^{th}$ gene and $i^{th}$ experimental unit.
- Let $s_1$ denote the subset of the indices {1,...,I} corresponding to treatment 1.
- Let $s_2$ denote the subset of the indices {1,...,I} corresponding to treatment 2.



## The Model (continued)

- Assume that each gene is differentially expressed (DE) with an unknown probability p, and equivalently expressed (EE) with probability 1-p.
- If gene j is equivalently expressed, then

$$x_{j1}, x_{j2}, ..., x_{jI} \mid \lambda_j \sim \text{Gamma}(\alpha, \lambda_j) \text{ with mean } \alpha / \lambda_j,$$

where $\lambda_j \sim \text{Gamma}(\alpha_0, v)$



## The Model (continued)

- If gene j is differentially expressed, then

$$\{x_{ji} : i \text{ in } s_1\} \mid \lambda_{j1} \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\alpha, \lambda_{j1}) \text{ with mean } \alpha / \lambda_{j1},$$

where $\lambda_{j1} \sim \text{Gamma}(\alpha_0, v)$, and

$$\{x_{ji} : i \text{ in } s_2\} \mid \lambda_{j2} \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\alpha, \lambda_{j2}) \text{ with mean } \alpha / \lambda_{j2},$$

where $\lambda_{j2} \sim \text{Gamma}(\alpha_0, v)$.

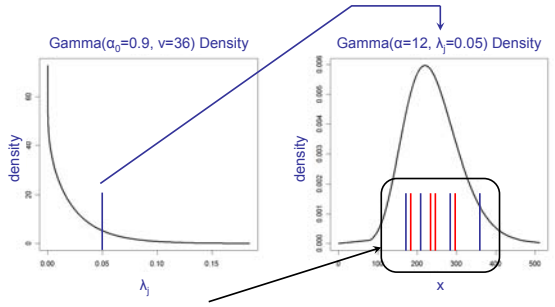
- All random variables are assumed to be independent.
- $p$, $\alpha$, $\alpha_0$, and $v$ are unknown parameters to be estimated from the data.

## An example of how the model is imagined to generate the data for the $j^{th}$ gene.

- Suppose p=0.05, α=12, $\alpha_0$=0.9, and v=36.
- Generate a Bernoulli random variable with success probability 0.05. If the result is a success the gene is DE, otherwise the gene is EE.
- If EE, generate $\lambda_j$ from Gamma($\alpha_0$=0.9, v=36).
- Then generate i.i.d. expression values from Gamma(α=12, $\lambda_j$).

7

## If gene is EE...




Expression values for the $j^{th}$ gene. Trt 1 and Trt 2

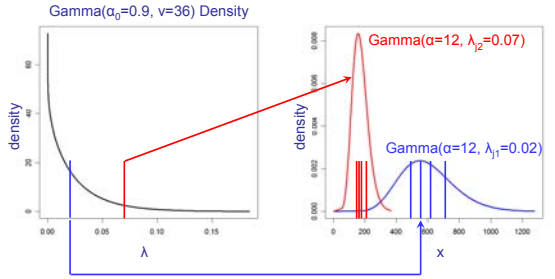
8

## Example Continued

- If the gene is DE, generate $\lambda_{j1}$ and $\lambda_{j2}$ independently from Gamma($\alpha_0$=0.9, v=36) .
- Then generate treatment 1 expression values i.i.d. from Gamma(α=12, $\lambda_{j1}$), and
- generate treatment 2 expression values i.i.d. from Gamma(α=12, $\lambda_{j2}$).

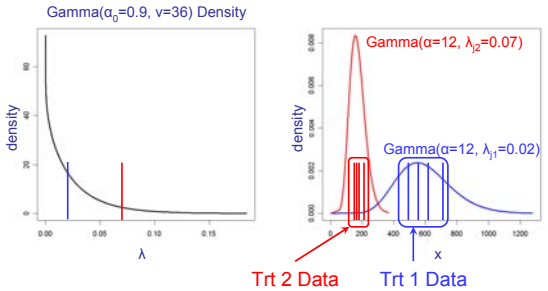
9

## If gene is DE...




10

## If gene is DE...




11

## Joint Density of $\mathbf{x}_j$ for an EE Gene

$$f_{EE}(\boldsymbol{x}_j) = \int_0^\infty f(\boldsymbol{x}_j|\lambda_j)g(\lambda_j)d\lambda_j$$

$$= \int_0^\infty \left(\prod_{i=1}^{I} \frac{\lambda_j^\alpha}{\Gamma(\alpha)} x_{ji}^{\alpha-1} e^{-\lambda_j x_{ji}}\right) g(\lambda_j)d\lambda_j$$

$$= \int_0^\infty \frac{\lambda_j^{I\alpha} \prod_{i=1}^{I} x_{ji}^{\alpha-1} e^{-\lambda_j \sum_{i=1}^{I} x_{ji}}}{\Gamma^I(\alpha)} g(\lambda_j)d\lambda_j$$

$$= \frac{\prod x_{ji}^{\alpha-1}}{\Gamma^I(\alpha)} \int_0^\infty \lambda_j^{I\alpha} e^{-\lambda_j \sum x_{ji}} g(\lambda_j)d\lambda_j$$

$$\left(\text{where } \prod x_{ji} \equiv \prod_{i=1}^{I} x_{ji} \text{ and } \sum x_{ji} \equiv \sum_{i=1}^{I} x_{ji}\right)$$

2

## Joint Density of $\mathbf{x}_j$ for an EE Gene (continued)

$$= \frac{\prod x_{ji}^{\alpha-1}}{\Gamma^I(\alpha)} \int_0^\infty \lambda_j^{I\alpha} e^{-\lambda_j \sum x_{ji}} g(\lambda_j) d\lambda_j$$

$$= \frac{\prod x_{ji}^{\alpha-1}}{\Gamma^I(\alpha)} \int_0^\infty \lambda_j^{I\alpha} e^{-\lambda_j \sum x_{ji}} \frac{\nu^{\alpha_0}}{\Gamma(\alpha_0)} \lambda_j^{\alpha_0-1} e^{-\nu\lambda_j} d\lambda_j$$

$$= \frac{\prod x_{ji}^{\alpha-1} \nu^{\alpha_0}}{\Gamma^I(\alpha)\Gamma(\alpha_0)} \int_0^\infty \lambda_j^{I\alpha+\alpha_0-1} e^{-\lambda_j(\nu+\sum x_{ji})} d\lambda_j$$

$$= \frac{\prod x_{ji}^{\alpha-1} \nu^{\alpha_0} \Gamma(I\alpha+\alpha_0)}{\Gamma^I(\alpha)\Gamma(\alpha_0)(\nu+\sum x_{ji})^{I\alpha+\alpha_0}} \leftarrow f_{EE}(\mathbf{x}_j)$$

13

## Joint Density for a DE Gene

$$f_{DE}(\mathbf{x}_j) = \prod_{k=1}^{2} \frac{\prod_{i \in S_k} x_{ji}^{\alpha-1} \nu^{\alpha_0} \Gamma(I_k\alpha+\alpha_0)}{\Gamma^{I_k}(\alpha)\Gamma(\alpha_0)(\nu+\sum_{i\in s_k} x_{ji})^{I_k\alpha+\alpha_0}}$$

where $I_k$ = the number of treatment k observations.

14

## Marginal Density for Gene j

$$f(\mathbf{x}_j) = p\, f_{DE}(\mathbf{x}_j) + (1\text{-}p)\, f_{EE}(\mathbf{x}_j)$$

## Marginal Likelihood for the Observed Data

$$f(\mathbf{x}_1)\, f(\mathbf{x}_2) \cdots f(\mathbf{x}_J)$$

Use the EM algorithm to find values of p, α, $\alpha_0$, and v that make the log likelihood as large as possible.

15

The posterior probability of differential expression for gene j is obtained by replacing p, α, $\alpha_0$, and v in

$$\frac{p\, f_{DE}(\mathbf{x}_j)}{p\, f_{DE}(\mathbf{x}_j) + (1\text{-}p)\, f_{EE}(\mathbf{x}_j)}$$

with their maximum likelihood estimates.

Software for EBArrays is available at http://www.biostat.wisc.edu/~kendzior.

16

## Extension to Multiple Treatment Groups

- If there are 3 treatment groups, each gene can be classified into 5 categories rather than just the two categories EE and DE:

  a) 1=2=3  b) 1=2≠3  c) 1≠2=3
  d) 1=3≠2  e) 1≠2 , 2≠3, 1≠3.

- Extensions to more than 3 groups can be handled similarly.

17

## Potential Drawbacks

- Coefficient of variation is assumed constant across gene-treatment combinations. This is analogous to assuming constant error variance across all gene-treatment combinations in the analysis of log-scale expression data.

- Between-gene difference are assumed to have the same distribution as within-gene between-treatment differences for differentially expressed genes.

18

## Coefficient of Variation is Constant across Gene-Treatment Combinations

- Coefficient of Variation = CV = sd / mean
- Conditional on the mean for a gene-treatment combination, say $\alpha / \lambda_{jk}$, the CV for the expression data is the CV of Gamma($\alpha$, $\lambda_{jk}$).
- CV of Gamma($\alpha$, $\lambda_{jk}$) is $(\alpha^{1/2}/\lambda_{jk})/(\alpha/\lambda_{jk})=1/\alpha^{1/2}$.
- Note that $\alpha$ is assumed to be the same for all gene-treatment combinations.

19

## Between-gene diffs $\stackrel{d}{=}$ within-gene between-trt diffs

- In real data, differences between genes tend to be larger than treatment differences within a DE gene.
- Thus, assuming the same distribution for both types of differences leads to conservative inferences.

20