

Parametric Empirical Bayes Methods for Microarrays

4/6/2009

Copyright © 2009 Dan Nettleton

1

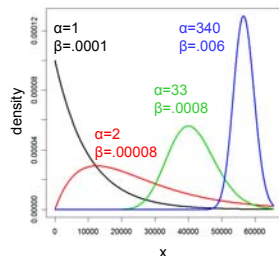
Parametric Empirical Bayes Methods for Microarrays

- 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.

2

The Gamma Distribution

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



3

A Model for the Data from a Two-Treatment Experiment

- Assume there are J genes indexed by $j=1, 2, \dots, J$.
- Data for gene j is $\mathbf{x}_j = (x_{j1}, x_{j2}, \dots, x_{ji})$ 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 $\{1, \dots, J\}$ corresponding to treatment 1.
- Let s_2 denote the subset of the indices $\{1, \dots, J\}$ corresponding to treatment 2.

4

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

$$x_{j1}, x_{j2}, \dots, x_{ji} \stackrel{\text{i.i.d.}}{\sim} \text{Gamma}(\alpha, \lambda_j) \text{ with mean } \alpha / \lambda_j,$$

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

5

The Model (continued)

- If gene j is differentially expressed, then

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

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

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

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

- All random variables are assumed to be independent.
- p, α, α_0 , and ν are unknown parameters to be estimated from the data.

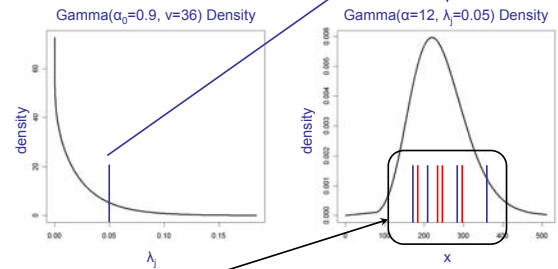
6

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

- Suppose $p=0.05$, $\alpha=12$, $\alpha_0=0.9$, 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 λ_j from $\text{Gamma}(\alpha_0=0.9, v=36)$.
- Then generate i.i.d. expression values from $\text{Gamma}(\alpha=12, \lambda_j)$.

7

If gene is EE...



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

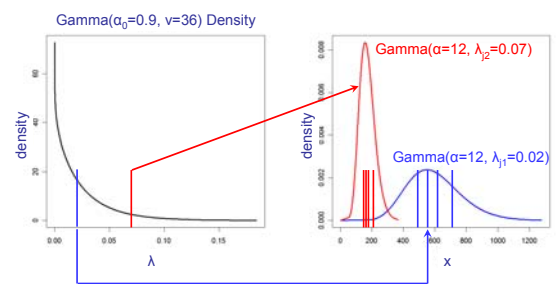
8

Example Continued

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

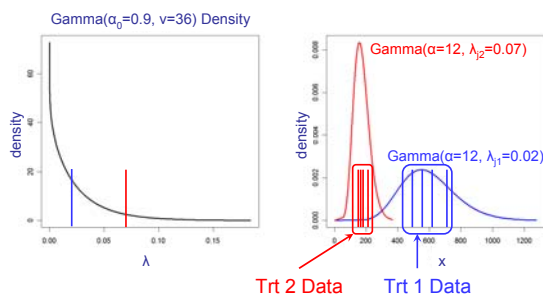
9

If gene is DE...



10

If gene is DE...



11

Coefficient of Variation is Constant across Gene-Treatment Combinations

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

12

Marginal Density for Gene j

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

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 , α_0 , and v that make the log likelihood as large as possible.

13

The posterior probability of differential expression for gene j is obtained by replacing p , α , α_0 , and v in

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

with their maximum likelihood estimates.

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

14

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.

15