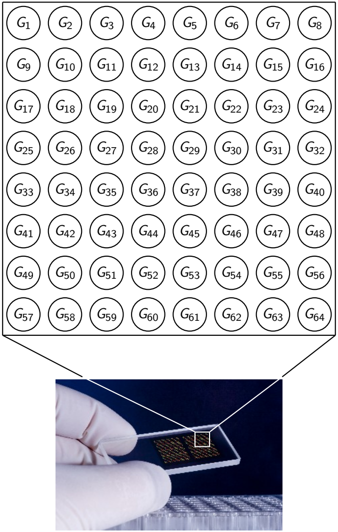
Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments

by Gordon K. Smyth
(as interpreted by Aaron J. Baraff)

STAT 572 Final Talk

May 20, 2014

Microarray Data

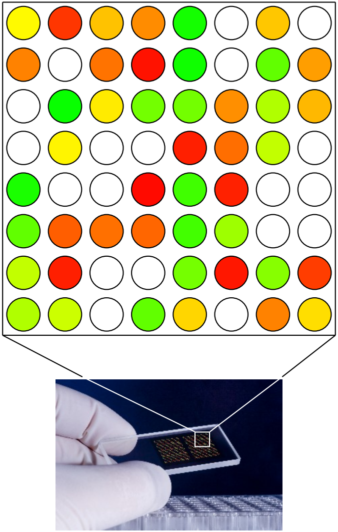


G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8
G_9	G_{10}	G_{11}	G_{12}	G_{13}	G_{14}	G_{15}	G_{16}
G_{17}	G_{18}	G_{19}	G_{20}	G_{21}	G_{22}	G_{23}	G_{24}
G_{25}	G_{26}	G_{27}	G_{28}	G_{29}	G_{30}	G_{31}	G_{32}
G_{33}	G_{34}	G_{35}	G_{36}	G_{37}	G_{38}	G_{39}	G_{40}
G_{41}	G_{42}	G_{43}	G_{44}	G_{45}	G_{46}	G_{47}	G_{48}
G_{49}	G_{50}	G_{51}	G_{52}	G_{53}	G_{54}	G_{55}	G_{56}
G_{57}	G_{58}	G_{59}	G_{60}	G_{61}	G_{62}	G_{63}	G_{64}

Measure expression level across large numbers of genes simultaneously

- Genes express by producing mRNA $\rightarrow$ translated into proteins
- $\sim$20,000 protein-coding genes in humans
- Microarray chip contains cDNA for a different gene at each spot
- Sample cDNA hybridizes with cDNA on chip

Microarray Data



Two-color:

- cDNA from two samples dyed red and green
- Response is log-ratio of intensity

$$y_g = \log_2 \frac{R_g}{G_g}$$

- Relative expressions only (fold changes)

Single-channel:

- cDNA from a single dyed sample
- Absolute expressions

Assumptions

Sample of n microarrays:

- Response vector $\mathbf{y}_g = (y_{g1}, \ldots, y_{gn})^T$ for each gene $g = 1, \ldots, G$
- Assume the linear model

$$\mathrm{E}(\mathbf{y}_g) = X\boldsymbol{\beta}_g, \quad \text{and} \quad \mathrm{Var}(\mathbf{y}_g) = W_g\sigma_g^2$$

  for known design matrix X and weight matrix W_g

- Assume estimates have distributions

$$\hat{\boldsymbol{\beta}}_g|\boldsymbol{\beta}_g, \sigma_g^2 \sim N(\boldsymbol{\beta}_g, V_g\sigma_g^2) \quad \text{and} \quad s_g^2|\sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi^2_{d_g}$$

  independent across all genes

(Note: no assumption that $\mathbf{y}_g$ is normal or model is fit by OLS)

Problems

Under $H_0 : \beta_{gj} = 0$, we have

$$t_{gj} = \frac{\hat{\beta}_{gj}}{s_g \sqrt{v_{gj}}} \sim t_{d_g}$$

Problem #1: Since *n* is often low, test statistics have high variance, leading to many false positives

Solution #1: Share variance information across all genes to improve estimates for σ_g^2

Problem #2: Too many genes! - multiple comparison methods assume independence across genes

Solution #2: Instead of inference, think of *p*-values as statistics used to rank genes

Bayesian Estimation

Assume prior distributions on β_{gj} and σ_g^{-2}:

$$\sigma_g^{-2} \sim \frac{1}{d_0 s_0^2} \chi^2_{d_0}$$

$$\Pr(\beta_{gj} \neq 0) = p_j$$

$$\beta_{gj} | \sigma_g^2, \beta_{gj} \neq 0 \sim N(0, v_{0j}\sigma_g^2)$$

with hyperparameters s_0^2, d_0, p_j, and v_{0j}

Through conjugacy, we get posterior distribution:

$$\sigma_g^{-2} | s_g^2 \sim \frac{1}{d_g s_g^2 + d_0 s_0^2} \chi^2_{d_g + d_0}$$

Bayesian Estimation

Now estimate σ_g^2 using the posterior mean

$$\tilde{s}_g^2 = \frac{1}{\mathrm{E}(\sigma_g^{-2} | s_g^2)} = \frac{d_g s_g^2 + d_0 s_0^2}{d_g + d_0}$$

From this, we get the **moderated** *t*-statistic

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}}$$

(Note: $\tilde{t} \to t$ as $d_0 \to 0$, and $\tilde{t} \to c\hat{\beta}$ as $d_0 \to \infty$)

Marginal Distributions

Under $H_0 : \beta_{gj} = 0$, we have

$$\begin{aligned} p(\tilde{t}_{gj}, s_g^2 | \beta_{gj} = 0) &= \tilde{s}_g v_g p(\hat{\beta}_{gj}, s_g^2 | \beta_{gj} = 0) \\ &= \tilde{s}_g v_g \int p(\hat{\beta}_{gj} | \sigma_g^{-2}, \beta_{gj} = 0) p(s_g^2 | \sigma_g^{-2}) \pi(\sigma_g^{-2}) d\sigma_g^{-2} \\ &= [\text{pdf for } t_{d_g + d_0}] \times [\text{pdf for } s_0^2 F_{d_g, d_0}] \end{aligned}$$

Therefore,

$$\tilde{t}_{gj} \sim t_{d_g + d_0} \quad \text{and} \quad s_g^2 \sim s_0^2 F_{d_g, d_0}$$

and they are independent

Estimation of Hyperparameters

Want to use s_g^2 across all genes to estimate s_0^2 and d_0

Let $z_g = \log s_g^2$ (Fisher's z):

- $\mathrm{E}(z_g) = \log s_0^2 + \psi(d_g/2) - \psi(d_0/2) + \log(d_0/d_g)$
- $\mathrm{Var}(z_g) = \psi'(d_g/2) + \psi'(d_0/2)$

Method of moments! Solve:

$$\psi'(d_0/2) = \frac{1}{G}\sum_{g=1}^{G}[(z_g - \bar{z})^2 - \psi'(d_g/2)]$$

$$\log s_0^2 = \frac{1}{G}\sum_{g=1}^{G}[z_g - \psi(d_g/2) + \psi(d_0/2) - \log(d_0/d_g)]$$

Simulation Study - Setup

Data sets simulated under the assumed model:

$$\begin{aligned}
\hat{\beta}_g | \beta_g, \sigma_g^2 &\sim N(\beta_g, v_g \sigma_g^2) \\
s_g^2 | \sigma_g^2 &\sim \sigma_g^2 \chi_{d_g}^2 / d_g \\
\beta_g | \sigma_g^2, \beta_g \neq 0 &\sim N(0, v_0 \sigma_g^2) \\
\sigma_g^{-2} &\sim \chi_{d_0}^2 / (d_0 s_0^2)
\end{aligned}$$

Using the parameters:

- $G = 15{,}000$ (300 differentially expressed)
- $d_g = 4,\ v_g = 1/3,\ v_0 = 2,\ s_0^2 = 4$
- $d_0 = 1, 4, 1000,$ more to less variable

Simulation Study - Setup

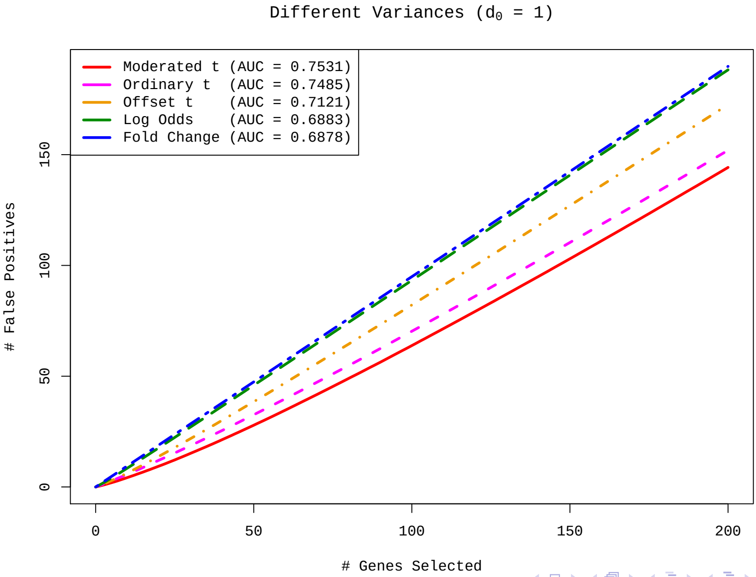
The following statistics were compared:

1. **Fold Change**: $M_g = \hat{\beta}_g$
2. **Ordinary t** - Student (1908): $t_g = \frac{\hat{\beta}_g}{s_g \sqrt{v_g}}$
3. **Offset t** - Efron et al (2001): $t_g^* = \frac{\hat{\beta}_g}{(s_g + s_{0.9})\sqrt{v_g}}$
4. **Log Odds** - Lönnstedt and Speed (2002):

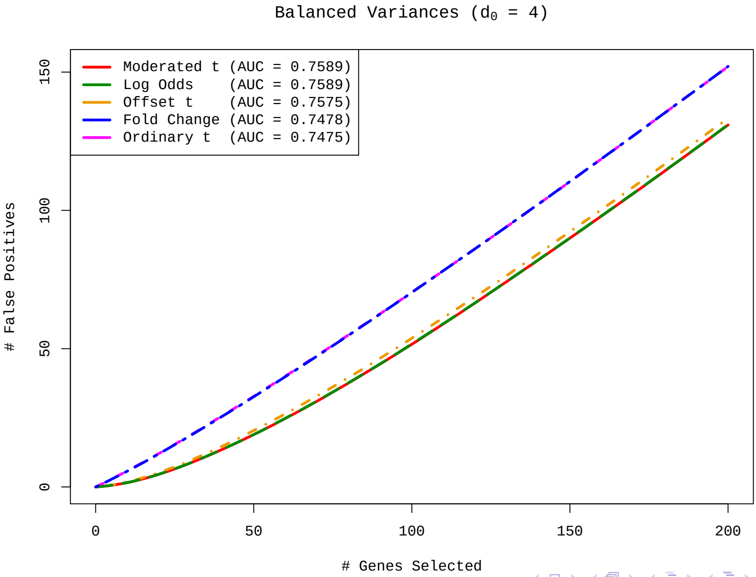
$$B_g = \log \frac{\mathrm{P}(\beta_g \neq 0 | \hat{\beta}_g, s_1^2, \ldots, s_G^2)}{\mathrm{P}(\beta_g = 0 | \hat{\beta}_g, s_1^2, \ldots, s_G^2)}$$

5. **Moderated t** - My paper!: $t_g = \frac{\hat{\beta}_g}{\tilde{s}_g \sqrt{v_g}}$

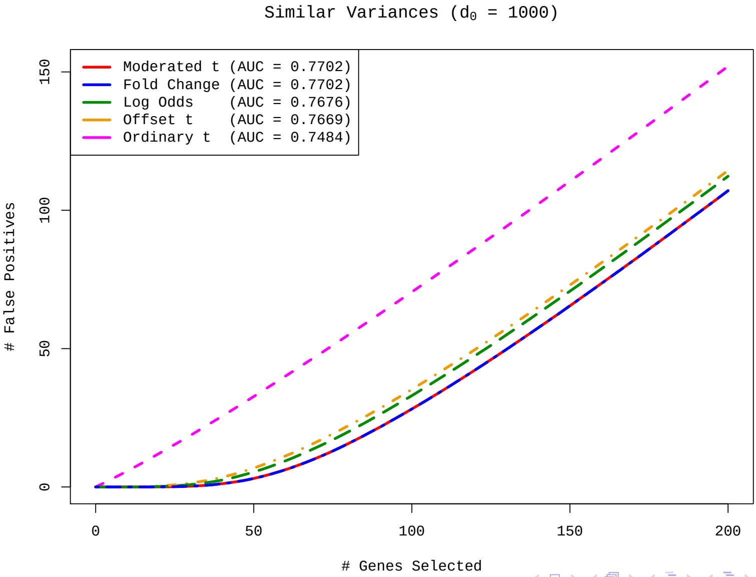
Simulation Study - Results



Simulation Study - Results



Simulation Study - Results



Simulation Study - Redux

Unfair to simulate data from the assumed hierarchical model?

Everything the same except:

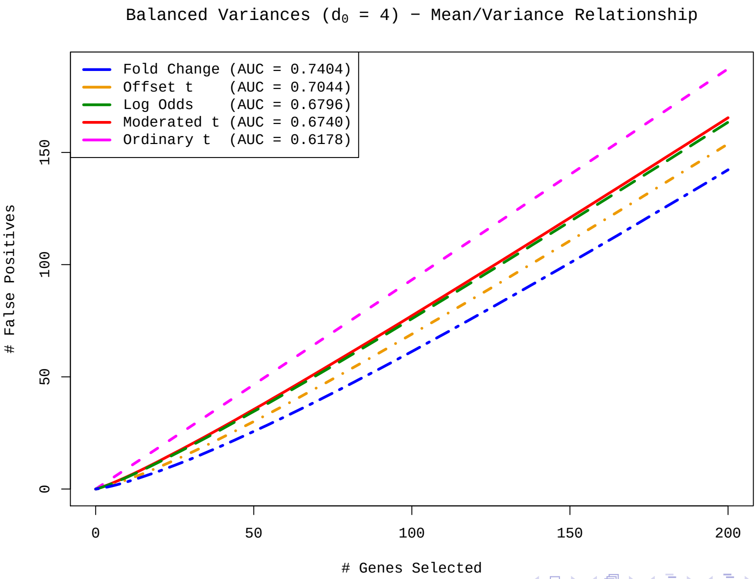
1. Relationship between mean and variance

$$\hat{\beta}_g|\beta_g,\sigma_g^2 \sim N(\beta_g, v_g(1+|\beta_g|)\sigma_g^2)$$
$$s_g^2|\sigma_g^2 \sim (1+|\beta_g|)\sigma_g^2\chi_{d_g}^2/d_g$$

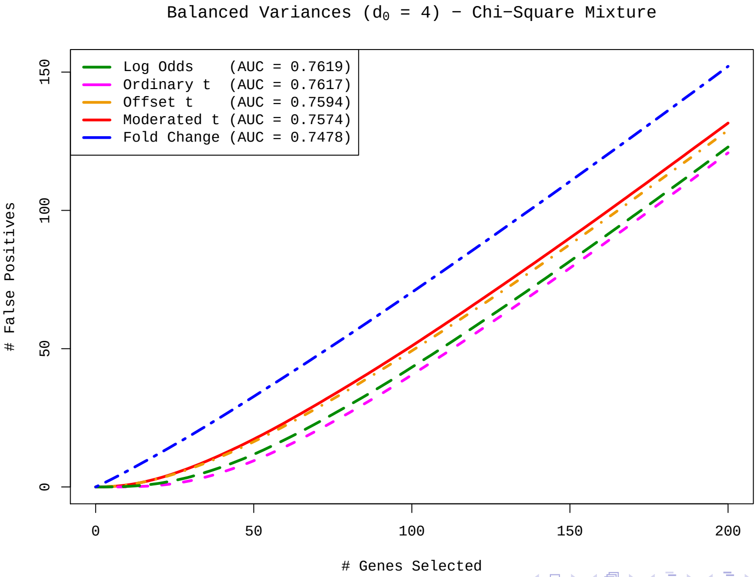
2. Chi-square mixture for variance

$$s_g^2|\sigma_g^2 \sim \frac{\sigma_g^2}{3}\left(\frac{\chi_1^2}{1}+\frac{\chi_4^2}{4}+\frac{\chi_{1000}^2}{1000}\right)$$

Simulation Study - Redux



Simulation Study - Redux



Model Expansion

Can I extend the methods in the paper to handle the mean/variance model?

Recall:

$$
\begin{aligned}
\hat{\beta}_g | \beta_g, \sigma_g^2 &\sim N(\beta_g, v_g(1 + |\beta_g|)\sigma_g^2) \\
s_g^2 | \sigma_g^2 &\sim (1 + |\beta_g|)\sigma_g^2 \chi^2_{d_g} / d_g \\
\Pr(\beta_g \neq 0) &= p \\
\beta_g | \sigma_g^2, \beta_g \neq 0 &\sim N(0, v_0 \sigma_g^2) \\
\sigma_g^{-2} &\sim \chi^2_{d_0} / (d_0 s_0^2)
\end{aligned}
$$

Model Expansion

Now we have

$$\sigma_g^{-2} | \beta_g, \hat{\beta}_g, s_g^2 \sim \left(\frac{d_g s_g^2}{1 + |\beta_g|} + d_0 s_0^2 + \frac{(\hat{\beta}_g - \beta_g)^2}{v_g (1 + |\beta_g|)} \right)^{-1} \chi^2_{d_g + d_0 + 1}$$

Can no longer estimate σ_g^2 from hyperparameters alone.

Instead, consider the posterior probability of differential expression:

$$\begin{aligned}
\Pr(\beta_g = 0 | \hat{\beta}_g, s_g^2, \sigma_g^2) &\propto (1 - p) \cdot p(\hat{\beta}_g, s_g^2 | \beta_g = 0, \sigma_g^2) \\
\Pr(\beta_g \neq 0 | \hat{\beta}_g, s_g^2, \sigma_g^2) &\propto p \cdot p(\hat{\beta}_g, s_g^2 | \beta_g \neq 0, \sigma_g^2) \\
&= p \cdot \int p(\hat{\beta}_g | \beta_g, \sigma_g^2) \cdot p(s_g^2 | \beta_g, \sigma_g^2) \cdot \pi(\beta_g | \sigma_g^2) d\beta_g
\end{aligned}$$

Model Expansion

Empirical Bayes can now be performed using an EM algorithm:

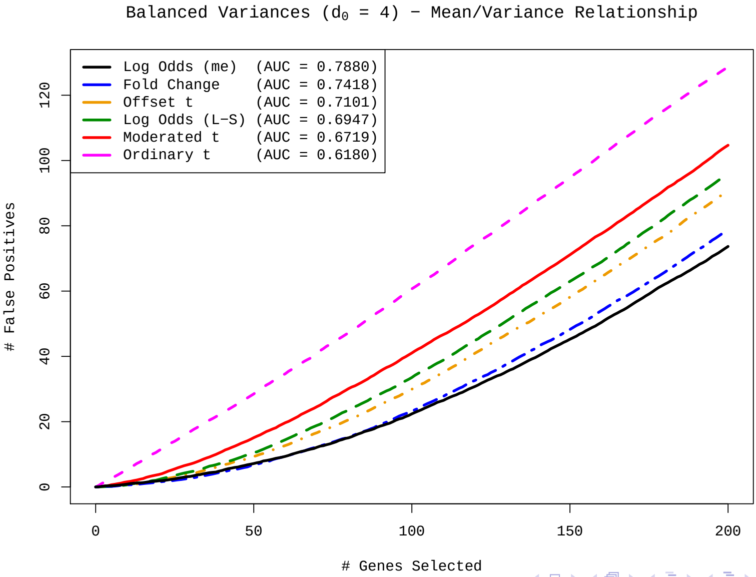
1. E-step - Estimate $\beta_g | \beta_g \neq 0$, σ_g^2, and $Z_g = 1_{\beta_g \neq 0}$ using MCMC
2. M-step - Estimate hyperparameters s_0^2, d_0, p, and v_0 by maximizing $\pi(\beta_g, \sigma_g^2 | s_0^2, d_0, p, v_0)$

Result: posterior log-odds

$$B_g = \log \left(\frac{\Pr(\beta_g \neq 0 | \hat{\beta}_g, s_g^2, \sigma_g^2)}{\Pr(\beta_g = 0 | \hat{\beta}_g, s_g^2, \sigma_g^2)} \right)$$

(Note: Each EM iteration has an MCMC, and each MCMC iteration has a numerical integration)

Simulation Study - Model Expansion



Conclusions

- Empirical Bayes method provides a way to share information across many genes
- Broad use across general microarray experiment designs (as well as other *-omics* experiments)
- Doesn't really solve the problem of performing inference, but allows for classification
- Simulation studies show that the method works well ...
- ... so long as the model is correctly specified
- Method can be modified for other models ... but it isn't pretty
- Can the method be modified to be robust under model misspecification?

All done!

(Chances are slim that anyone will see this slide because I have probably been cut off for time by now.)