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

Gordon Smyth
Walter and Eliza Hall Institute

Designs → Linear Models

\[ y \log_2(R) - \log_2(G) = B - A \]

\[
\begin{pmatrix}
    y_1 \\
    y_2 \\
    y_3
\end{pmatrix} =
\begin{pmatrix}
    1 & 0 \\
    1 & -1 \\
    1 & 1
\end{pmatrix}
\begin{pmatrix}
    \beta_1 \\
    \beta_2
\end{pmatrix}
\]

\[ \beta = B - A \]

\[ \beta_1 = A - \text{Ref} \]

\[ \beta_2 = B - A \]

\[ \beta_3 = C - A \]

Linear Model Estimates

Obtain a linear model for each gene \( g \)

\[
E(y_{gj}) = X\hat{\beta}_{gj} \\
\text{var}(y_{gj}) = W^{-1}_g \sigma^2
\]

Estimate model by robust regression, least squares or generalized least squares to get

coefficients \( \hat{\beta}_{gj} \)

standard deviations \( s_g \)

standard errors \( \text{se}(\hat{\beta}_{gj})^2 = c_g s^2 \)

Parallel Inference for Genes

- 10,000-40,000 linear models
- Curse of dimensionality:
  Need to adjust for multiple testing, e.g., control family-wise error rate (FWE) or false discovery rate (FDR)
- Boon of parallelism:
  Can borrow information from one gene to another

Hierarchical Model

Normal Model

\[ \hat{\beta}_o \sim N(\beta_o, \sigma^2) \]

\[ P(\beta_o \neq 0) = p \]

\[ \beta_o | \beta_o \neq 0 \sim N(0, c_o \sigma^2) \]

\[ s^2 = \sigma^2 x^2_{d_0} \]

Reparameterization of Lönnstedt and Speed 2002

Normality, independence assumptions are wrong but convenient, resulting methods are useful
Posterior Statistics

Posterior variance estimators
\[ \hat{s}_g^2 = \frac{s^2_d + s^2_0}{d + d_0} \]

Moderated t-statistics
\[ \tilde{t}_{gj} = \frac{\hat{s}_g \sqrt{C_{gj}}}{\hat{s}_g^2} \]

Eliminates large t-statistics merely from very small \( s \)

Marginal Distributions

The marginal distributions of the sample variances and moderated t-statistics are mutually independent
\[ s_g^2 \sim s_0^2 F_{d,d_0} \]
\[ \tilde{t}_g \sim \begin{cases} t_{d_0+d} & \text{with prob } 1 - p \\ \frac{1}{\sqrt{1 + \epsilon_g / c} t_{d_0+d}} & \text{with prob } p \end{cases} \]

Degrees of freedom add!

Known result?

Estimating Prior Parameters

Marginal moments of log \( s^2 \) lead to estimators of \( s_0 \) and \( d_0 \):
Estimate \( d_0 \) by solving
\[ \psi'(d_0 / 2) = \text{mean} \left\{ n_s^c - \psi'(d_s / 2) \right\} \]
where
\[ \epsilon_s = \log s^c_s - \psi'(d_s / 2) + \log(d_s / 2) \]
Finally
\[ s_0^2 = \exp \left\{ \tilde{t} + \psi'(d_s / 2) - \log(d_s / 2) \right\} \]

Shrinkage of Standard Deviations

The data decides whether \( \tilde{t}_g \) should be closer to \( t_{\text{pooled}} \) or to \( t_g \)

Simulations

\( \sigma^2 \) similar
\( \sigma^2 \) very different

Scorecard Controls
Posterior Odds

Posterior probability of differential expression for any gene is

\[
p(\beta \neq 0 | \beta, s^2) = p \left( \frac{c}{c + c_0} \right)^{1/2} \left( 1 - \frac{c^2 + d + d_0}{c + c_0} \right)^{1/2 + d_0} \]

Monotonic function of \( t^2 \) for constant \( d \)

Reparameterization of Lönnstedt and Speed 2002

Quantile Estimation of \( c_0 \)

Let \( r \) be rank of \( |\hat{I}_g| \) in descending order, and let \( F(\cdot) \) be the distribution function of the \( t \)-distribution. Can estimate \( c_0 \) by equating empirical to theoretical quantities:

\[
2 \left| t \right| \frac{\sqrt{c + c_0}}{1 - \sqrt{c + c_0}} \left( 1 - \frac{c}{c + c_0} \right) \left( 1 - \frac{c^2 + d + d_0}{c + c_0} \right)^{1/2} \left( 1 - \frac{c^2 + d + d_0}{c + c_0} \right)^{1/2 + d_0} = \frac{r - 0.5}{n}
\]

Get overall estimator of \( c_0 \) by averaging the individual estimators from the top \( \rho/2 \) proportion of the \( |\hat{I}_g| \)

Duplicate spots

- Replicate spots of each gene on same array, assume duplicates at regular spacing
- Assume spatial component of correlation between duplicates is same for each gene
- Estimate spatial correlation from consensus estimator across genes

Posterior F-tests

If

\[
\beta_y = 0
\]

then

\[
\frac{\beta_y^T X^T W X \beta_y}{s_y^2} \sim F_{t, d + k, n}
\]

Non-null prior on \( \beta \) doesn't enter

F-Tests as Classification Problem
Stemmed Heat Diagrams