PH296, Section 36

February 25, 2002

Discussion of:

K. Kerr, M. Martin, and G. Churchill. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**(6): 819-837.

S. Dudoit, Y.H. Yang, M. Callow, and T. P. Speed. (2002). Statistical methods for identifying differentially expressed genes in replicated DNA microarray experiments. *Statistica Sinica* **12** (1).

R. Wolfinger, G. Gibson, E. Wolfinger, L. Bennett, H. Hamdadeh, P. Bushel, C. Afshari, and R. Paules. (2001). Assessing gene significance from cDNA microarray expression via mixed models. *Journal of Computational Biology* **8**(6): 625-637.

Issues

- Identification of differentially expressed genes.
- Magnitude of difference for the spotted genes given the sources of variation.
- What level of observation is statistically significant?
- Methods for analyzing data.
- Experimental design, number of replications.

Sources of variation

1. Interesting variation
   - variation in the expression profile for a given gene
   - variation in the expression profile among genes
   - variation in the expression profile due to different treatments
2. Obscuring variation due to
   - sample preparation
   - manufacture of the array
   - hybridization of the sample
   - optical measurements

ANOVA Model

Kerr and Churchill (2000)

$$\log(y_{ijkg}) = \mu + A_i + D_j + T_k + G_g + (AG)_{ig} + (TG)_{kg} + \varepsilon_{ijkg}$$

μ	-	overall average signal (normalization term)
A	-	array (normalization term)
D	-	dye (normalization term)
T	-	treatment (normalization term)
G	-	overall gene effect
(AG)	-	a particular spot on the array
(TG)	-	gene expression attributable to treatments!!!

ε_{ijkg} independent, identically distributed

ANOVA Model - Bootstrap

Kerr and Churchill (2000)

Estimated differences (Latin square design)

$$(\widehat{TG})_{1g_0} - (\widehat{TG})_{2g_0} = \frac{1}{2}\log\left(\frac{y_{111g_0}y_{221g_0}}{y_{122g_0}y_{212g_0}}\right) - \frac{1}{2N}\log\left(\prod_g \frac{y_{111g}y_{221g}}{y_{122g}y_{212g}}\right)$$

- variety × gene interactions are averages of just two observations (no CLT)
- fitted residuals appear heavy-tailed
- Bootstrap: simulated data sets

  $$\log(y_{ijkg})^* = \hat{\mu} + \hat{A}_i + \hat{D}_j + \hat{V}_k + \hat{G}_g + (\widehat{AG})_{ig} + (\widehat{TG})_k g + \varepsilon^*_{ijkg}$$

  where $\varepsilon^*_{ijkg} \sim \sqrt{4N/(N-4)}\hat{F}$ (independently drawn), $\hat{F}$ empirical distribution of original residuals.

- percentile method to obtain 99% confidence intervals for the differences $(\widehat{TG})_{1g_0} - (\widehat{TG})_{2g_0}$. Width=1.61, i.e. estimated fold change of $e^{1.61/2} = 2.24$ is significant at the 0.01 level. (normal confidence interval width = 1.29)

Checking assumptions:

- residuals are identically distributed,
- constant error variance,
- log scale seems appropriate.
- Multiple testing not taken into account.

ANOVA Model - Least squares estimators

Kerr and Churchill (2000)

Objective: Minimize the residual sum of squares, RSS.

$t_{ijkg} = \log(y_{ijkg})$

$$\mathrm{RSS} = \sum_{ijkg} (t_{ijkg} - \mu - A_i - D_j - V_k - G_g - (AG)_{ig} - (TG)_{kg})^2$$

Partial derivatives, constraints lead to

$$(\widehat{TG})_{kg} = t_{..kg} - t_{..k.} - t_{...g} + t_{....}$$

ANOVA Model - Comments

Kerr and Churchill (2000)

- early analyses of microarray data: fold changes to identify genes for the standardized log ratios of the fluorescence intensities.
- "Global" normalization procedures may not be able to remove undesirable experimental effects.
- ANOVA: estimate sources of variation for large data sets.
- A, D, T terms normalize data without preliminary data manipulation.
- no computation of log ratios
- accounts for effects of dyes or variation between samples (experimental design).

- residual distribution nonnormal, but constant error variance: bootstrap approach.
- large number of similar quantities → estimates of highest and lowest effects too extreme.
- multiple testing not taken into account.

Multiple testing

- false positives: genes declared to be differentially expressed which in reality are not
- false negatives: genes truly differentially expressed but not declared as such

Normalization and multiple testing

Dudoit et al. (2002)

X of log intensities $\log_2 R/G$ with k rows (genes), $n = n_1 + n_2$ columns (control, treatment hybridizations).

1. Normalization: $\log_2 R/G \to \log_2 R/G - c_j(A)$,
   $c_j(A) = \texttt{lowess}$ fit to M vs. A plot, jth print-tip.

2. test statistic
$$t_j = \frac{\bar{x}_{2j} - \bar{x}_{1j}}{\sqrt{s_{ij}^2/n_1 + s_{2j}^2/n_2}}$$

3. permutation test statistics $t_1^{(b)}, \ldots, t_k^{(b)}$

4. adjusted p-values to account for multiple hypotheses testing (Westfall and Young)

Normalization - Comments

Dudoit et al. (2000)

- "Global" methods of normalization miss some experimental features
- multiple testing
- ANOVA model by Kerr et al: one main effect for normalization, one error term for all genes
- strong model assumptions? (parametric models (gamma, Gaussian), functional relationships)
- which effects should be included?
- replication, experimental design questions

Effects

- fixed effects: attributable to a finite set of factor levels that occur in the data
- random effects: attributable to a (infinite) set of factor levels, of which a random sample occur in the data

Mixed models: fixed effects and random effects

Benefits: recovery of interblock information

Mixed Models

Wolfinger et al. (2001)

y_{gki} = $\log_2$ of the background corrected measurement from gene g, treatment k, and array i.

1. Normalization model

$$y_{gki} = \mu + T_k + A_i + (TA)_{ki} + \varepsilon_{gki},$$

μ	-	overall mean value,
T	-	main effect for treatments,
A	-	main effect for arrays,
(TA)	-	interaction effect of arrays and treatments,
ε	-	stochastic error.

random effects: $A_i, (TA)_{ki}, \varepsilon_{gki}$ normally distributed random variables, zero means, variance components $\sigma^2_A, \sigma^2_{TA}, \sigma^2_\varepsilon$

2. Gene model

$$r_{gki} = G_g + (GT)_{gk} + (GA)_{gi} + \gamma_{gki},$$

r_{gki}	-	residuals of normalization model
(GA)	-	spot effects

random effects: $(GA)_{gi}, \gamma_{gki}$ normally distributed random variables, zero means, variance components $\sigma^2_{(GA)_g}, \sigma^2_{\gamma_g}$, independent across their indices and with each other.

Restricted Maximum likelihood (REML)

REML: maximize the part of the likelihood which is invariant to the location parameters of the model (i.e. to the fixed effects).

REML takes account of implicit degrees of freedom associated with the fixed effects (ML does not).

For balanced data: Solutions to REML equations = ANOVA estimators

Mixed Models - Comments

Wolfinger et al. (2001)

- replication within and between arrays necessary
- experimental design
- global distributional assumptions too strong
- effects to be included depends on research question
- heterogeneity in the gene models
- false positive rates: cutoff at the Bonferroni value $0.05/(6917 \times 10) = 1e - 6.14$ for experimentwise false positive rate of 0.05.
- missing values, background correction, various designs
- correlation of the residuals: little difference in practice?
- normality on the log scale "usually reasonable."

Power analysis

Wolfinger et al. (2001)

Power - probability of declaring statistical significance when a true difference exists.

power $= 1 - P(\text{false negative})$

- experimental design
- model assumptions
- approximate values for the model parameters
- hypotheses to be tested
- desired false positive rate