Linear models for microarray data analysis

Mikhail Dozmorov
Fall 2017

Differential expression in microarrays studies

General framework for differential expression

- Linear models
- Model the expression of each gene as a linear function of explanatory variables (Groups, Treatments, Combinations of groups and treatments, Etc…)

$$y = X\beta + \epsilon$$

- y - vector of observed data
- X - design matrix
- β - vector of parameters to estimate

Example of a design matrix

Normal sample x 2

Cancer Sample x 2

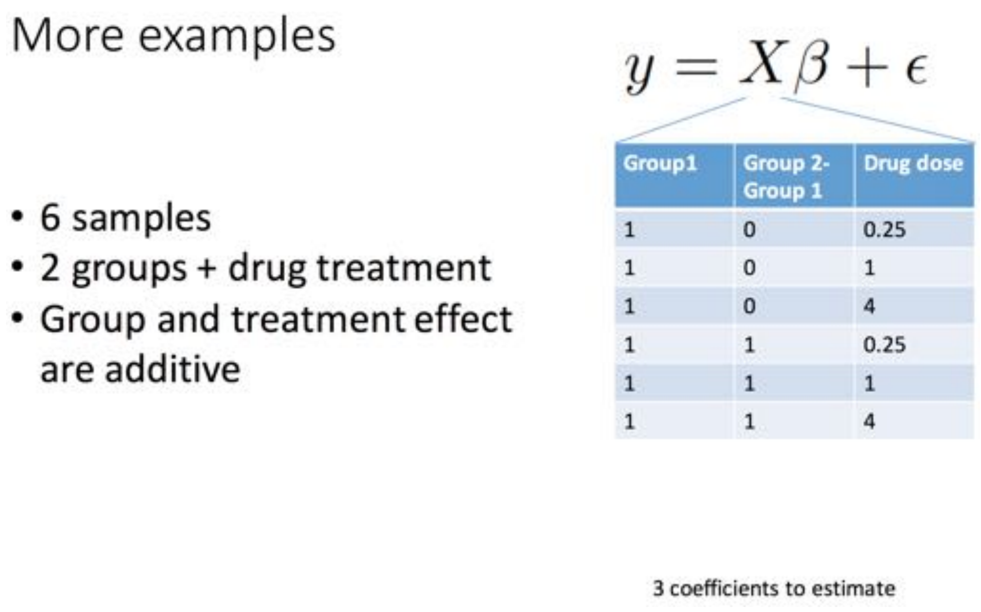
$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

β_1 = normal log-expression

β_2 = cancer – wt

$E[y_1]=E[y_2]=\beta_1$ $E[y_3]=E[y_4]= \beta_1+ \beta_2$

Example of a design matrix



More examples

- 6 samples
- 2 groups + drug treatment
- Group and treatment effect are additive

$$y = X\beta + \epsilon$$

Group1	Group 2- Group 1	Drug dose
1	0	0.25
1	0	1
1	0	4
1	1	0.25
1	1	1
1	1	4

3 coefficients to estimate

Linear model parameter estimation

Model is specified – how do we find the coefficients? $y = X\beta + \epsilon$

- Minimize squared error $\epsilon'\epsilon = (Y - X\beta)'(Y - X\beta)$
- Take derivative $\frac{d}{d\beta}((Y - X\beta)'(Y - X\beta)) = -2X'(Y - X\beta)$
- Set to 0, $-2X'(Y - X\beta) = 0$
- Solve $X'Y = (X'X)\beta$ $\beta = (X'X)^{-1}X'Y$

Hypothesis testing

- Significance of coefficients is tested with a T-test

β can be a vector. We can test the significance of any one coefficient β_i via a T-test

$$t_{score} = \frac{\hat{\beta} - \beta_0}{SE_{\hat{\beta}}}$$

$$t_{score} = \frac{(\hat{\beta} - \beta_0)\sqrt{n-2}}{\sqrt{SSR/\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$SSR = \sum_{i=1}^{n} \hat{\epsilon}^2$ - sum of squares of residuals, depends on the whole model

Linear models and covariates

- Linear models are useful for including nuisance variables - technical factors
- Variables that have an effect on measurements but are not themselves of interest (e.g. sample storage time)
- Incorporating storage time gives smaller residuals and thus larger T-stats for the coefficient of interest

Bayesian-type methods

- We have lots of genes. Gene i has mean μ_i and variance σ_i^2
- Bayesian methods assume that the means and variances come from known distributions (the priors)
- Empirical Bayes methods assume that the means and variances have distributions that are estimated from the data
- "Moderated" methods use test statistics that can be viewed as approximations to Empitical Bayes methods, but are justified by other statistical theory

Empirical Bayes and Moderated methods

Primarily focus of the distribution of the variances

- **SAM** - "moderated t-test" method adds a constant based on a quantile of the distribution of the S^2 over all the genes. Also uses permutation tests
- **LIMMA** - more formal empirical Bayes t-test analysis
    - Results in replacing gene variances by a weighted average of the gene variance and the mean variance of all genes
    - Leads to t-tests with Student's t-distribution

Power, false discovery

The t-statistics will be larger when

- The difference between the means is larger
- The variances are smaller
- The n and m are larger

The only item under our direct control is the sample size

Power, false discovery

For fixed p-value at which we declare statistical significance, increasing sample size:

- *Increases power* (probability of rejecting when alternative is true)
- *Reduces FDR* (percentage of false rejections)
- *Reduces FNR* (percentage of failures to reject when alternative is true)

We also improve power by:

- Good experimental design
- Choice of statistical methodology

Bayes, empirical Bayes and moderated methods are *more powerful* than classical methods

Bayesian reasoning: short intro

- Synthesize prior knowledge and evidence
- Main theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Simple derivation

$$P(A \text{ and } B) = P(A|B)P(B) = P(B|A)P(A)$$

Classical example

- Duchenne Muscular Dystrophy (DMD) can be regarded as a simple recessive sex-linked disease caused by a mutated X chromosome (X).
- An XY male expresses the disease, whereas an XX female is a carrier but does not express the disease
- Suppose neither of a woman's parents expresses the disease, but her brother does. Then the woman's mother must be a carrier, and the woman herself therefore may be a carrier
- $P(C) = 1/2$ - *prior*
- What is the probability she is a carrier if she has a healthy son? - *observation*

Classical example

- $P(C) = 1/2$

$$p(C|h.s.) = \frac{p(h.s.|C)p(C)}{p(h.s.)} = \frac{p(h.s.|C)p(C)}{p(h.s.|C)p(C) + p(h.s.|\bar{C})p(\bar{C})}$$

$$p(C|h.s.) = \frac{(1/2) * (1//2)}{(1/2) * (1/2) + 1 * (1/2)} = \frac{1}{3}$$

- Incorporate evidence into strong prior belief

Bayesian approach to statistics

- Naïve approach: estimate the parameters from observation only
- Bayesian approach: have some prior expectation
- Prior expectation for gene expression: Gene-specific variance comes from an underlying variance distribution

Bayesian approach to statistics

- Bayesian statistical analyses:
  - Begin with 'prior' distributions describing beliefs about the values of parameters in statistical models prior to analysis of the data at hand
  - Requires specification of these parameters
  - 'Empirical Bayes' methods use the data at hand to guide prior parameter specification
  - Use all the data to define priors, compute posteriors of individual estimates

Limma method

- Generalized the hierarchical model of Lonnstedt and Speed (2002) into a practical approach for general microarray experiments.
- The model borrows information across genes to smooth out variances and uses posterior variances in a classical t-test setting.
- Completely data-dependent and uses empirical Bayes approach to estimate hyper parameters

Limma method

Smyth et al. (2004) Statistical Applications in Genetics and Molecular Biology

- Uses a Bayesian hierarchical model in multiple regression setting.
- Borrows in formation from all genes to estimate gene specific variances.
- As a result, variance estimates will be "shrunk" toward the mean of all variances. So very small variance scenarios will be alleviated.
- Implemented in Bioconductor package "limma".

http://bioinf.wehi.edu.au/limma/, https://bioconductor.org/packages/release/bioc/html/limma.html

Limma

- Linear models
  - can be used to compare two or more groups
  - can be used for multifactorial designs
    - e.g. genotype and treatment
- Uses empirical Bayes analysis to improve power in small sample sizes
  - Models gene-level error variances $\{\sigma_1^2, \ldots, \sigma_m^2\}$ with a scaled inverse χ^2
  - borrowing information across genes

Limma method

The sample variance for each gene, given σ^2 is assumed to follow a scaled Chi-square distribution with d_g degree of freedom

$$S_g^2|\sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi_{d_g}^2$$

The unknown residual variances σ_g^2 are allowed to vary across genes by assuming scaled inverse Chi-square prior distribution

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 * S_0^2}\chi_{d_0}^2$$

where d_0 and S_0^2 are the hyperparameters for the degrees of freedom and variance, respectively.

Moderated t-Statistics

The posterior variance S_g^{limma} is a combination of an estimate obtained from the prior distribution S_0^2 and the pooled variance S_g^2

$$S_{g\_limma}^2 = \frac{d_0 S_0^2 + d_g S_g^2}{d_0 + d_g}$$

where d_0 and d_g are, respectively, prior and empirical degrees of freedom

- Including a prior distribution of variances has the effect of borrowing information from all genes to aid with inference about individual genes

Moderated t-Statistics

- Limma, Moderated t-statistics, described in (Gordon K. Smyth, "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments" Statistical Applications in Genetics and Molecular Biology 3 (2004) http://www.statsci.org/smyth/pubs/ebayes.pdf)

$$t_g^{limma} = \frac{\bar{y_{g1.}} - \bar{y_{g2.}}}{S_g^{limma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $S^2_{g\_limma}$ is the posterior variance.

Limma

- **design matrix**
  - defines which conditions arrays belong to
  - rows: arrays; columns: coefficients
- **contrast matrix**
  - specifies which comparisons you would like to make between the RNA samples
  - for very simple experiments, you may not need a contrast matrix

More complicated models

- So far we only consider 2 group experiments
- Many other possibilities
    - Factorial: two groups each has two treatments - Are treatment effects different across groups?
    - Continuous variables: dosage of a drug
    - Continuous discrete variables
        - 2 groups, 3 drug doses — do the drugs affect the groups differently?
- limma on a time course, https://github.com/jennybc/stat540_2014/blob/master/seminars/seminar0

Limma method

- Lönnstedt, Ingrid, and Terry Speed. “REPLICATED MICROARRAY DATA.” Statistica Sinica 12, no. 1 (2002): 31–46. http://www.jstor.org/stable/24307034. - Empirical Bayes method for analyzing microarray replicates. Issues with simple approaches, proposed B statistics - a Bayes log posterior odds with two hyperparameters in the inverse gamma prior for the variances, and a hyperparameter in the normal prior of the nonzero means. Appendix - detailed definitions, derivations, and solutions.
- Smyth, Gordon K. “Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments.” Statistical Applications in Genetics and Molecular Biology 3 (2004): Article3. doi:10.2202/1544-6115.1027. PDF - Linear models for differential analysis, moderated t-statistics via shrinkage of sample variance. Empirical estimation of Bayesian prior variance distribution and shrinkage hyperparameters.
- Phipson, Belinda, Stanley Lee, Ian J. Majewski, Warren S. Alexander, and Gordon K. Smyth. “Robust Hyperparameter Estimation Protects against Hypervariable Genes and Improves Power to Detect Differential Expression.” The Annals of Applied Statistics 10, no. 2 (June 2016): 946–63. doi:10.1214/16-AOAS920. PDF - An extension of differential analysis using linear modeling and empirical Bayes by windsorizing outliers in estimating sample distribution.

Extensions of Limma method

- Sartor, Maureen A., Craig R. Tomlinson, Scott C. Wesselkamper, Siva Sivaganesan, George D. Leikauf, and Mario Medvedovic. “Intensity-Based Hierarchical Bayes Method Improves Testing for Differentially Expressed Genes in Microarray Experiments.” BMC Bioinformatics 7 (December 19, 2006): 538. https://doi.org/10.1186/1471-2105-7-538. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-538 - Intensity-Based Moderated T-statistic (IBMT). Empirical Bayes approach allowing for the relationship between variance and gene signal intensity (estimated with loess). Brief description of previous methods (Smyth, Cyber-T). Details of Smyth hierarchical model and moderated t-statistics, estimation of hyperparameters with implementation of variance-signal. Software at http://eh3.uc.edu/ibmt/.

- Lianbo Yu et al., “Fully Moderated T-Statistic for Small Sample Size Gene Expression Arrays,” Statistical Applications in Genetics and Molecular Biology 10, no. 1 (September 15, 2011), https://doi.org/10.2202/1544-6115.1701. https://www.degruyter.com/view/j/sagmb.2011.10.issue-1/1544-6115.1701/1544-6115.1701.xml - Third implementation of moderated t-statistics. First is Smyth 2004 model assuming d_{0g} and s_{0g}^2 constant, second is IBMT (intensity-based moderated t) Sartor 2006 allows varying s_{0g}^2 with gene expression, third is the present FMT (fully moderated t) model allowing varying d_{0g} and s_{0g}^2. Description of Smyth hierarchical model, estimation of hyperparameters. Adjusted