Linear Models and Empirical Bayes Methods for Microarray Data Analysis

Alex Sánchez

Department d'Estadística
Universitat de Barcelona

May 24, 2017

Outline

1. Introduction

2. Linear Models for Microarray Data
   - Linear models
   - Example 2: A study on antibiotic resistance
   - Example 3: A factorial experiment
   - Estimation and inference with the linear model
   - Strength and weakness

3. Linear Models and Empirical Bayes for Microarrays
   - Overview of G.Smyth's approach
   - Lönnsted and Speed's B–statistic
   - The moderated-t

4. Implementation and Examples

Overview of the presentation

- This presentation treats two complementary aspects
    - The use of a *general linear model* approach to analyze microarray data, specifically to select differentially expressed genes in statistically designed microarray experiments.
    - The enhancement suggested by Smyth (2004) to solve some weaknesses of this approach when applied to microarray data.
- Along the presentation some examples are introduced. The experimental design is presented but the analysis is referred to the limma user's guide.

Outline

1. Introduction

2. Linear Models for Microarray Data
   - Linear models
   - Example 2: A study on antibiotic resistance
   - Example 3: A factorial experiment
   - Estimation and inference with the linear model
   - Strength and weakness

3. Linear Models and Empirical Bayes for Microarrays
   - Overview of G.Smyth's approach
   - Lönnsted and Speed's B–statistic
   - The moderated-t

4. Implementation and Examples

What is a Linear Model?

- The linear model (Faraway, 2004) is a general frame for modelling and data analysis in statistics.
- Consists of defining *linear* relation between observed values and experimental conditions.
- If some assumptions on the data are true one can...
  - Obtain *good* estimators for the model parameters and their standard errors.
  - (With some extra conditions) make inference about the experiment.
- Regression and Analysis of the Variance can be both formulated as special cases of the linear models.

How to use Linear Models to analyze microarray data

- The application of linear models can be seen as a multi–step sequential process.
  1. Start by specifying the design of the experiment: which samples are allocated to which conditions.
  2. (Re–)Write a linear model for this design in the form of $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, where **X** is the design matrix.
  3. If needed re–state the questions to answer as *linear contrasts* on the parameters of the model.
  4. Once the model is specified apply the general theory to estimate the parameters and the contrasts and,
  5. If the appropriate validity conditions hold, perform inference on the model parameters based on the estimates.
- This process will be illustrated in the examples that follow.

Outline

1. Introduction

2. Linear Models for Microarray Data
   - Linear models
   - Example 2: A study on antibiotic resistance
   - Example 3: A factorial experiment
   - Estimation and inference with the linear model
   - Strength and weakness

3. Linear Models and Empirical Bayes for Microarrays
   - Overview of G.Smyth's approach
   - Lönnsted and Speed's B–statistic
   - The moderated-t

4. Implementation and Examples

Example 2: A study on antibiotic resistances

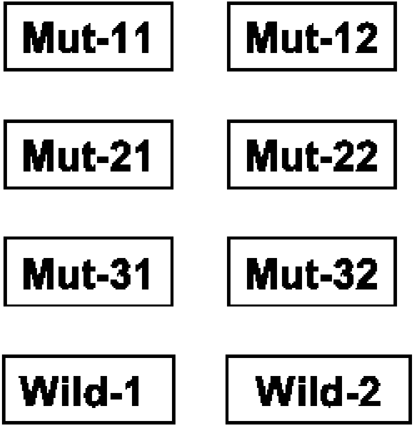
A k-samples problem

- IncHI plasmids encode multiple–antibiotic resistance in *S. enterica*.
- Plasmid R27, the “wild type” is thermosensitive for transfer.
- Some mutant phenotypes associated to both chromosomal *hha* and *hns* participate in different metabolic processes of interest in termoregulated conjugation.
- The goal of the experiment is *to find genes which are differentially expressed in three different mutant types*, say M_1, M_2 and M_3.

Example 2: Possible design strategies

- This experiment might be implemented differently depending on the type of chips used (two or one colour) and on which comparisons are of higher interest.
  - Using two colour–slides
    - A *reference design*: Hybridize each Mutant (M_i) vs. Wild type (W).
    - A *loop design*: Hybridize each mutant to each other in a double loop that includes dye-swapping.
  - Using one colour slides: hybridize mutants and wild types separately.

Design representation (3)



- Allows for direct comparison of
  - Mutant vs Wild and
  - Mutant vs Mutant.
- Number of parameters to estimate=4.
- All comparisons can be made efficiently.

Linear model for one colour arrays design I

Model, $\mathbf{y} = \mathbf{X}\alpha + \varepsilon$, and contrasts $\mathbf{C^1}'\beta$, $\mathbf{C^2}'\beta$

- Model parameters:

$$\alpha_1 = \mathbf{E}(logM_1),\ \alpha_2 = \mathbf{E}(logM_2),\ \alpha_3 = \mathbf{E}(logM_3), \alpha_4 = \mathbf{E}(logW).$$

- *Contrasts*: Two possible sets of interesting comparisons.
  1. Comparison between mutant types ($\mathbf{C^1}'\beta$)

$$\begin{aligned} \beta_1^1 &= \alpha_1 - \alpha_2, \\ \beta_2^1 &= \alpha_3 - \alpha_2, \\ \beta_3^1 &= \alpha_2 - \alpha_3. \end{aligned}$$

  2. Comparison between each mutant and the wild type ($\mathbf{C^2}'\beta$)

$$\begin{aligned} \beta_1^2 &= \alpha_4 - \alpha_1, \\ \beta_2^2 &= \alpha_3 - \alpha_1, \\ \beta_3^2 &= \alpha_2 - \alpha_1. \end{aligned}$$

Linear model for one colour arrays design II

Model, $\mathbf{y} = \mathbf{X}\alpha + \varepsilon$, and contrasts $\mathbf{C^1}'\beta$, $\mathbf{C^2}'\beta$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{\text{Design Matrix,}\mathbf{X}} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \end{pmatrix}$$

Linear model for one colour arrays design III

Model, $\mathbf{y} = \mathbf{X}\alpha + \varepsilon$, and contrasts $\mathbf{C^1}'\beta$, $\mathbf{C^2}'\beta$

$$\begin{pmatrix} \beta_1^1 \\ \beta_2^1 \\ \beta_3^1 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 \end{pmatrix}}_{\text{Contrast Matrix,}\mathbf{C^1}} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix}.$$

$$\begin{pmatrix} \beta_1^2 \\ \beta_2^2 \\ \beta_3^2 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}}_{\text{Contrast Matrix,}\mathbf{C^2}} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix}.$$

Outline

1. Introduction

2. Linear Models for Microarray Data
   - Linear models
   - Example 2: A study on antibiotic resistance
   - Example 3: A factorial experiment
   - Estimation and inference with the linear model
   - Strength and weakness

3. Linear Models and Empirical Bayes for Microarrays
   - Overview of G.Smyth's approach
   - Lönnsted and Speed's B–statistic
   - The moderated-t

4. Implementation and Examples

Example 3: Estrogen experiment

Description

- Goal: to study the effect of estrogen on the genes in ER+ breast cancer cells over time.
- After serum starvation of eight samples, four samples exposed to estrogen, and mRNA abundance measured after 10 hours (2 samples) and 48 hours (other two).
- Remaining four samples left untreated, and mRNA abundance measured similarly (10 hours for two samples, 48 hours for the other two).
- Experiment with 2x2 factorial design: two factors (estrogen and time), each at two levels (present or absent,10 hours or 48 hours).

Example 3: Experimental design

Slide	Estrogen	Time
1	Absent	10
2	Absent	10
3	Present	10
4	Present	10
5	Absent	48
6	Absent	48
7	Present	48
8	Present	48

- One channel microarrays (Affymetrix) used.
- Each condition replicated twice.
- Specific questions to answer:
  - Estrogen effect after 10 hours.
  - Estrogen effect after 48 hours.
  - Time effect when no estrogen applied.

Example 3: Linear model

- This experiment admits different parametrizations
  - Separate factors with 2 levels each for estrogen (Abs/Pres), time (10h/48h) and interaction:

    $$Y_{ijk} = \underbrace{\alpha_i}_{Estrogen} + \underbrace{\beta_j}_{Time} + \underbrace{\gamma_{ij}}_{interaction} + \varepsilon_{ijk}, \; i = 1,2, \; j = 1,2, \; k = 1,2$$

    This first parametrization seems more natural but it is more complicated to rely on it to answer the questions posed.
  - One combinate factor with 4 levels (*Abs.10h, Abs.48h, Pres.10h, Pres.48h*)

    $$Y_{ij} = \alpha i + \varepsilon_{ij}, \quad i = 1,...,4, \; j = 1,2.$$

    This parametrization seems more rigid but it is better adapted to answer the questions posed.
- The second parametrization is adopted here.

Linear model for factorial design (1)

- Model parameters:

$$
\begin{aligned}
\alpha_1 &= \mathbf{E}(logAbs.10h),\ \alpha_2 = \mathbf{E}(logAbs.48h),\\
\alpha_3 &= \mathbf{E}(logPres.10h),\ \alpha_4 = \mathbf{E}(logPres.48h).
\end{aligned}
$$

- *Contrasts*: Interesting questions are straightforward.

$$
\begin{aligned}
\beta_1^1 &= \alpha_3 - \alpha_1, && \text{Estrogen effect after 10 hours}\\
\beta_2^1 &= \alpha_4 - \alpha_2, && \text{Estrogen effect after 48 hours}\\
\beta_3^1 &= \alpha_2 - \alpha_1, && \text{Time effect in absence of estrogen}
\end{aligned}
$$

Linear models for one colour arrays design (2) I

Model, $\mathbf{y} = \mathbf{X}\alpha + \varepsilon$, and contrasts $\mathbf{C}'\beta$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{\text{Design Matrix,}\mathbf{X}} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \end{pmatrix}$$

Linear models for one colour arrays design (2) II

Model, $\mathbf{y} = \mathbf{X}\alpha + \varepsilon$, and contrasts $\mathbf{C}'\beta$

$$\begin{pmatrix} \beta_1^1 \\ \beta_2^1 \\ \beta_3^1 \end{pmatrix} = \underbrace{\begin{pmatrix} -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ -1 & 1 & 0 & 0 \end{pmatrix}}_{\text{Contrast Matrix,}\mathbf{C}} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix}.$$

Outline

1 Introduction

2 Linear Models for Microarray Data
- Linear models
- Example 2: A study on antibiotic resistance
- Example 3: A factorial experiment
- Estimation and inference with the linear model
- Strength and weakness

3 Linear Models and Empirical Bayes for Microarrays
- Overview of G.Smyth's approach
- Lönnsted and Speed's B–statistic
- The moderated-t

4 Implementation and Examples

Estimation and inference I

- Having expressed the experiment as a linear model:

$$\mathbf{E}(\mathbf{y_g}) = \mathbf{X}\alpha_{\mathbf{g}}, \quad \mathrm{var}(y_g) = W_g \sigma_g,$$

  allows to use *standard linear model theory* to obtain ....
  - Parameter estimates: $\hat{\alpha}_g (\approx \alpha)$.
  - Standard deviation estimates: $\hat{\sigma}_g = s_g (\approx \sigma)$.
  - Standard error estimates:$\widehat{\mathrm{var} \hat{\alpha}_g} = V_g \, s_g^2$.
- These estimates are the basis to perform inference about α i.e. test $H_0 : \ \alpha = 0$?, based on the fact that:

$$t_{gj} = \frac{\alpha_{gj}}{s_g \sqrt{V_{gj}}} \sim \text{Student distribution.}$$

  Similar result holds for $\alpha_1 - \alpha_2$.
- The estimation and inferential procedures do not depend on which parametrization has been adopted, although different numerical values may be, of course, obtained.

Outline

1. Introduction
2. Linear Models for Microarray Data
   - Linear models
   - Example 2: A study on antibiotic resistance
   - Example 3: A factorial experiment
   - Estimation and inference with the linear model
   - Strength and weakness
3. Linear Models and Empirical Bayes for Microarrays
   - Overview of G.Smyth's approach
   - Lönnsted and Speed's B–statistic
   - The moderated-t
4. Implementation and Examples

Strength and Weakness of Linear Models

- Linear model approach is flexible and powerful
  - Can be adapted to many different and complex situations.
  - Always yields good (*BLUE*) estimates.
  - If assumptions are true it provides a basis for inference.
- However...
  - If assumptions don't hold conclusions are not to be trusted.
  - Even if they hold they may be affected by small sample sizes, so that high variances estimates may yield non significant t-values.
- The methodology developed by Smyth (2004) based upon results of Lönsted & Speed (2002) addresses how to deal with these weaknesses.

Strength and Weakness of Linear Models

- Linear model approach is flexible and powerful
  - Can be adapted to many different and complex situations.
  - Always yields good (*BLUE*) estimates.
  - If assumptions are true it provides a basis for inference.
- However...
  - If assumptions don't hold conclusions are not to be trusted.
  - Even if they hold they may be affected by small sample sizes, so that high variances estimates may yield non significant t-values.
- The methodology developed by Smyth (2004) based upon results of Lönsted & Speed (2002) addresses how to deal with these weaknesses.

Strength and Weakness of Linear Models

- Linear model approach is flexible and powerful
  - Can be adapted to many different and complex situations.
  - Always yields good (*BLUE*) estimates.
  - If assumptions are true it provides a basis for inference.
- However...
  - If assumptions don't hold conclusions are not to be trusted.
  - Even if they hold they may be affected by small sample sizes, so that high variances estimates may yield non significant t-values.
- The methodology developed by Smyth (2004) based upon results of Lönsted & Speed (2002) addresses how to deal with these weaknesses.

Outline

1. Introduction

2. Linear Models for Microarray Data
   - Linear models
   - Example 2: A study on antibiotic resistance
   - Example 3: A factorial experiment
   - Estimation and inference with the linear model
   - Strength and weakness

3. Linear Models and Empirical Bayes for Microarrays
   - Overview of G.Smyth's approach
   - Lönnsted and Speed's B–statistic
   - The moderated-t

4. Implementation and Examples

General overview I

- Smyth (2004) considers the problem of identifying genes which are differentially expressed across specified conditions in designed microarray experiments.
- He addresses the fact that
    - the variability of the expression values differs between genes, but
    - the parallel nature of the inference in microarrays allows some possibilities for borrowing information from the ensemble of genes which can assist in inference about each gene individually.

General overview II

Smyth (2004) develops in 3 steps the hierarchical model of Lönnstedt and Speed (2002) into a practical approach.

- The first step is to re–state it in the context of general linear models.
- The second step is to derive consistent, closed form estimators for the hyperparameters. These estimators have robust behavior even for small numbers of arrays.
- The third step is to re–formulate the posterior odds statistic in terms of a moderated t-statistic.

Outline

1. Introduction

2. Linear Models for Microarray Data
   - Linear models
   - Example 2: A study on antibiotic resistance
   - Example 3: A factorial experiment
   - Estimation and inference with the linear model
   - Strength and weakness

3. Linear Models and Empirical Bayes for Microarrays
   - Overview of G.Smyth's approach
   - Lönnsted and Speed's B–statistic
   - The moderated-t

4. Implementation and Examples

The B–statistic

- L-S (2002) addressed the problem of improving usual measures of differential expression such as $M = log(R/G)$ or $t = M \left/ (s/\sqrt{n}) \right.$.
    - Other attempts: Tibshirani et al. SAM's Statistic.
- They rely on the (log) ratio of two probabilities: *the probability of the gene being expressed* vs. *the probability of not being expressed*.
    - This is a common approach in clinical studies or genetics and is called a (log) odds-ratio or LODS.

$$B = \log \frac{P[\text{Affected}|M_{ij}]}{P[\text{Not Affected}|M_{ij}]},$$

gene=i ($i = 1...N$), replication=j ($j = 1, ..., n$).

Empirical Bayes approach

- Assume that the mean and variance of log-ratios for each gene follow *a priori* fixed distributions.
- Combine the information from all the genes to estimate their parameters.
- Use the Bayesian method to derive an expression of B which combines both the information of each gene and the information obtained from all the genes in a *posterior* log–odds–ratio.

$$B_g = \text{const} + \log\left(\frac{\frac{2a}{n} + s^2 + M_g^2}{\frac{2a}{n} + s^2 + \frac{M_g^2}{1+nc}}\right)$$

Pro's and Con's of B.

$$B_g = \text{const}+$$

$$\log\left(\frac{\frac{2a}{n}+s^2+M_g^2}{\frac{2a}{n}+s^2+\frac{M_g^2}{1+nc}}\right)$$

- Useful to rank genes ...
  - B_g increases with M_g,
  - If M_g is small, a ensures that the ratio will not be expanded by a very small variance.
  - $B \approx M_g/s_g$ for large n.
- However...
  - Still no $p - values$.
  - Depends on many parameters.

Outline

1. Introduction

2. Linear Models for Microarray Data
   - Linear models
   - Example 2: A study on antibiotic resistance
   - Example 3: A factorial experiment
   - Estimation and inference with the linear model
   - Strength and weakness

3. Linear Models and Empirical Bayes for Microarrays
   - Overview of G.Smyth's approach
   - Lönnsted and Speed's B–statistic
   - The moderated-t

4. Implementation and Examples

Hierarchical model

- A hierarchichal model is introduced to describe how the unknown coefficients β_{gj} and unknown variances $\sigma_g$vary across genes.
- This is done adopting a Bayesian approach that puts prior distributions for these sets of parameters.
-

Normal Model	Priors
$\hat{\beta}_{gj} \sim N(\beta_{gj}, v_{gj}\sigma_g^2)$ $s_g^2 \sim \sigma_g^2 \chi^2_{d_g}$	$P(\beta_{gj} \neq 0) = p$ $\beta_{gj}\vert\sigma_g^2, \beta_{gj} \neq 0 \sim N(0, v_{0j}\sigma_g^2)$ $\sigma_g^2 \sim s_0^2 \left(\chi^2_{d_0}/d_0\right)^{-1}$

Hierarchical model

- A hierarchichal model is introduced to describe how the unknown coefficients β_{gj} and unknown variances $\sigma_g$vary across genes.
- This is done adopting a Bayesian approach that puts prior distributions for these sets of parameters.

Normal Model	Priors
$\hat{\beta}_{gj} \sim N(\beta_{gj}, v_{gj}\sigma_g^2)$ $s_g^2 \sim \sigma_g^2 \chi^2_{d_g}$	$P(\beta_{gj} \neq 0) = p$ $\beta_{gj} \mid \sigma_g^2, \beta_{gj} \neq 0 \sim N(0, v_{0j}\sigma_g^2)$ $\sigma_g^2 \sim s_0^2 \left(\chi^2_{d_0}/d_0\right)^{-1}$

Hierarchical model

- A hierarchical model is introduced to describe how the unknown coefficients β_{gj} and unknown variances $\sigma_g$vary across genes.
- This is done adopting a Bayesian approach that puts prior distributions for these sets of parameters.
-

Normal Model	Priors
$\hat{\beta}_{gj} \sim N(\beta_{gj}, v_{gj}\sigma_g^2)$ $s_g^2 \sim \sigma_g^2\chi^2_{d_g}$	$P(\beta_{gj} \neq 0) = p$ $\beta_{gj}\|\sigma_g^2, \beta_{gj} \neq 0 \sim N(0, v_{0j}\sigma_g^2)$ $\sigma_g^2 \sim s_0^2 \left(\chi^2_{d_0}/d_0\right)^{-1}$

Posterior statistics

- Posterior variance estimators

$$\tilde{s}_g^2 = E(\sigma_g^2 | s_g^2) = \frac{d_o s_o^2 + d_g s_g^2}{d_0 + d_g}$$

The posterior values shrink the observed variances towards the prior values with the degree of shrinkage depending on the relative sizes of the observed and prior degrees of freedom

- The moderated $t-$statistic is:

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

- This distributional result assumes d_0 and s_0 to be given values. In practice they need to be estimated from the data

Implementation and Examples

- This approach has become very popular between microarray users mainly due to the fact that it is implemented in an excellently well documented Bioconductor package: `limma`.
- The limma user guide (available after installation) contains the analysis of the Swirl and the estrogen data as well as many other examples.

Summary

- Linear models provide a flexible and powerful approach to modelling and analyzing microarray experiments.
- The hierarchical model presented gives moderated statistics that help to borrow the information across genes to compensate for the usually small number of replicates.
- The programs `Limma`, `LimmaGUI` and `LimmaAffyGUI` allow a direct application of these approaches.

References

- Faraway, J. (2004) *Linear models with R* Chapman and Hall (CRC) (*preliminary version freely available in CRAN*).
- Lönnstedt, I. and Speed, T. (2002) *Replicated Microarray Data* Statistica Sinica 12(2002), 31–46.
- Smyth G.K. (2004) *Linear models and empirical bayes methods for assessing diferential expression in microarray experiments*. Statistical Applications in Genetics and Molecular Biology, 3:Article 3, 2004.