Preprocessing of Microarray data II: Filtering and other transformations

(i). Filtering:

Identifying relevant genes, i.e. genes that present interesting variation across conditions.

Main issues/topics:

1. Pragmatic or rigorous (preprocessing or main aim).

2. VERY multiple comparison, want to check/test a very large number of genes… usually few if any replicates available for each condition.

3. Two or many experimental conditions (multivariate).

4. “Computational” approaches.

An important distinction:

- **Preprocessing step**, to reduce the number of genes considered in further analyses: eliminate completely uninteresting genes.

- **Aim in itself**: identifying genes presenting a significant differential expression.

In the first case, criteria and methods can be less stringent and/or rigorous, and one will tend to retain a larger portion of the genes…

Retaining false positives:

In preprocessing data for further statistical analyses, it is better to err towards retaining false positives… unless their number is so large as to obscure patterns of, and relationships among, true positives.

However, if identifying differentially expressed genes is the main (and final) aim of the current data analysis, false positives can be as bad as false negatives:

- wrong conclusions
- expenditure for further experimental validation.

Very general idea:

use a statistic to create a ranking
(i.e. a partial ordering of the genes)

Some obvious examples, but many more are possible and used:

Fold-change: one condition and control, not replicated

- X_{i1} itself if normalized log ratio to a control – spotted, or
- normalized $\log(X_{i1}/ X_{i0})$)

$$t_{i,1} = \frac{\overline{X}_{i,1}}{s_{i,1}}$$

T-ratio: one condition and control, replicated. Add consideration of variability.

$$dt_{i,1,2} = \frac{\overline{X}_{i,2} - \overline{X}_{i,1}}{\sqrt{\left(n_2^{-1} + n_1^{-1}\right)\frac{(n_2 - 1)s_2^2 + (n_1 - 1)s_1^2}{n_2 + n_1 - 1}}}$$

"Two sample" T-ratio: two conditions (possibly common control), replicated. This contains an implicit assumption of equal variability in the two conditions.

$$L^2_{i,1,2\ldots T} = \left\|X_i\right\|^2 \quad \text{or} \quad \left\|X_i - \overline{X}\right\|^2_{S^{-1}}$$

$$l(\lambda)_{i,1,2\ldots T} = (X_i'\lambda)$$

T conditions (possibly common control), not replicated.

$$E_{i,1,2\ldots T} = (\max_{j=1,2\ldots T} X_{ij}) - (\min_{j=1,2\ldots T} X_{ij}) \text{ or } (\max_{j=1,2\ldots T} X_{ij}) / (\min_{j=1,2\ldots T} X_{ij})$$

If preprocessing, create some arbitrary cut-off along the ranking

- in terms of a value for the statistic
- in terms of a number of genes
- in terms of a quantile (a percentage of the genes)

If rigorous, need to employ a testing mechanism: how many top-ranking genes have a "significant" value of the statistic?

VERY multiple testing problem: the same data, often providing very little power (we have very few replicates if any) is used to make a very large number of tests… even if we have null distributions to compute p-values, we need corrections (e.g. Dudoit *et al*., Speed's group at Berkeley)

And what would be the null distributions… distributional assumptions? Asymptotics?

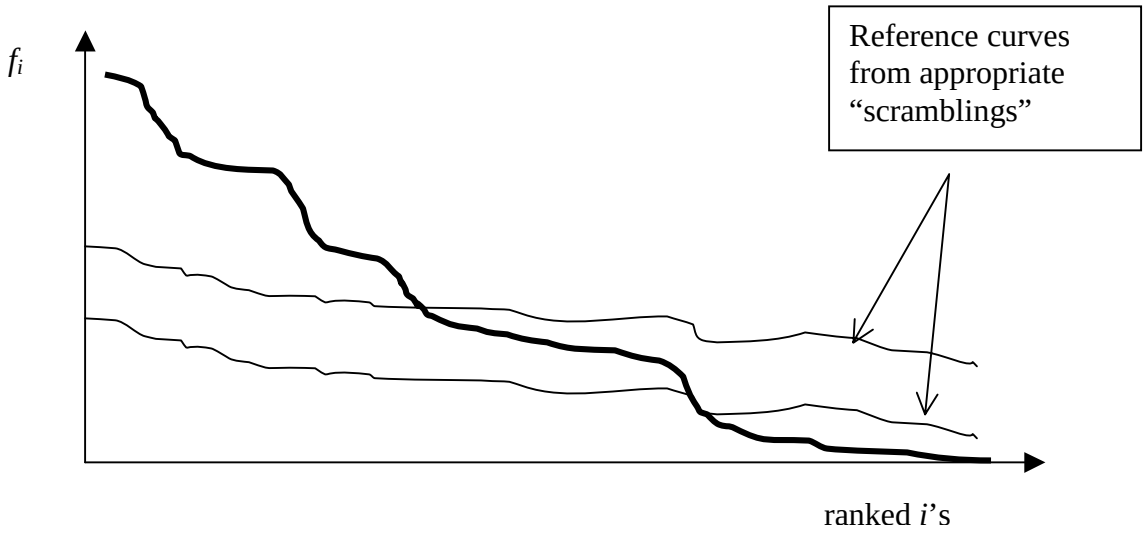
For the two-conditions (or condition and control) case, more sophisticated approaches use the idea of a mixture of two distributions, differentially expressed, and not (seen already in Lee *et al*., also in Efron *et al*., and Pan *et al*.)

When there are several conditions, the statistic and testing ought to be multivariate in nature (e.g. Chiligarian *et al*.)

"Computational" approaches: Instead of referring to distributional assumptions or asymptotics, create a threshold for the ranking statistics by simulating an appropriate null scenario; that is

use randomizations or permutations

to compare the ranking statistic values to a chance background



(ii). Other data transformations:

A. Further improve comparability of measurements across experimental conditions and/or across genes:

Centering and standardizing, by

- experimental condition (replicate), i.e. column in the data matrix, and/or
- "gene", i.e. row in the data matrix

to eliminate location and variation size effects.

Centering and standardization are used in a very large numbers of applications.

B. Further decrease the effect of non-experimental sources of variation:

Quantizing, i.e. discretizing continuous data into (ordered) classes, to eliminate unnecessary "detail", and systematic errors with it.

Note: another way of decreasing effect of non-experimental sources of variation is limiting the analysis to a low-dimensional reconstruction of the data (i.e. an approximation of the expression profiles through a small number of characteristic patterns) that, too, eliminate unnecessary "detail", and systematic errors with it.

(this is one aspect of dimension reduction; next topic)

General questions:

- What is the appropriate "scale" to look at our measurements, given the questions we want to address, and the data analysis methods we want to employ?

- Do we introduce any "spurious structure" in the data by applying certain transformations?

Examples:

(1) Think of the gene profiles as a cloud of N points

$$X_1 X_2 \ldots X_N$$

in T dimensions

- Centering by condition "correlates" the genes' positions by forcing the cloud to be centered at 0:

$$X_1 + X_2 + \ldots X_N = 0$$

- Centering by gene creates a linear constraint; the points are forced to live on a hyperplane:

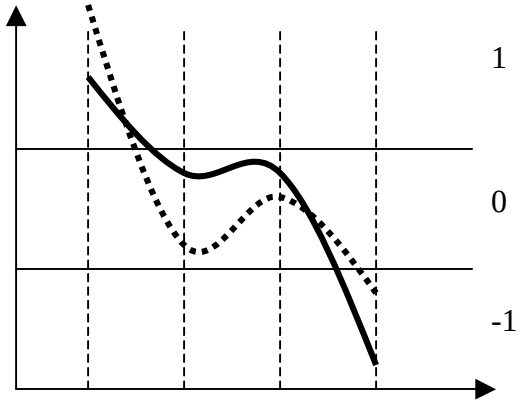
$$X_i \text{ '} \mathbf{1} = 0$$

- Standardizing by gene forces the points to live on a hypersphere:

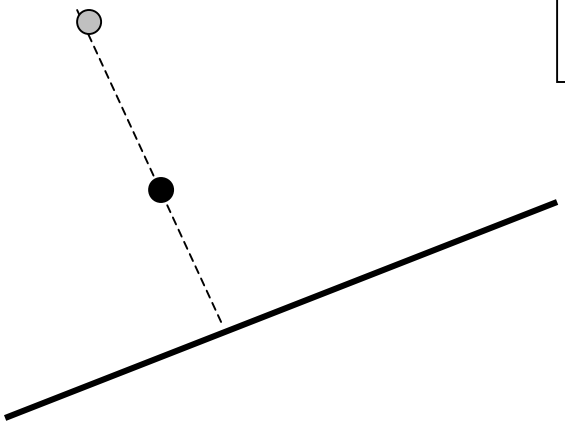
$$\| X_i \|^2 = 1$$

Are we "creating" geometrical structure?

(2) An arbitrary quantization or low-dimensional reconstruction may induce misleading similarities in gene profiles. What is the definition of unnecessary “detail”?



Two profiles are discretized to 1 0 0 –1 . Are they similar?



Two profiles share a 1-dimensional reconstruction. Are they similar?