Preprocessing of Microarray data II: Filtering and other transformations

Filtering:

Identifying differentially expressed genes ("Significance testing").

This can be an aim in itself, or a preprocessing step to reduce the number of genes considered in further analyses (of course, in the second case criteria can be less stringent and/or rigorous).

Retaining false positives: especially if one is preprocessing data for further analyses, it is better to err towards retaining false positives, unless their number is so large as to obscure patterns and relationships of true positives.

Multivariate "significance testing" (more than one experimental condition)

Some references:

Dudoit, S., Yang, Y. H, Speed, T. P., and Callow, M. J. "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments". To appear in Statistica Sinica. UC Berkeley Statistics tech. report 2001:
http://www.stat.Berkeley.EDU/users/terry/zarray/Html/matt.html
Storey, J. D., and Tibshirani, R. "Estimating false discovery rates under dependence with applications to DNA microarrays". Stanford tech report 2001:
http://www-stat.stanford.edu/~tibs/lab/publications.html
Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. PNAS 98, 5116-5121. Preprint at:
http://www-stat.stanford.edu/~tibs/lab/publications.html
Efron B., Tibshirani, R., Storey J. D., and Tusher V. (2001). Empirical Bayes analysis of a microarray experiment. Journal of the American Statistical Association 96, 1151-1160.
Pan, W. "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments". To appear in Bioinformatics. Biostatistics, University of MN tech report, 2001:
http://www.biostat.umn.edu/~weip/ge.html
Pan, W., Lin, J. and Le, C. "A mixture model approach to detecting differentially expressed genes with microarray data". Biostatistics, University of MN tech report, 2001:
http://www.biostat.umn.edu/~weip/ge.html
T. Ideker, V. Thorsson, A. F. Siegel, and L. Hood. (2000). Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. Journal of Computational Biology 7 (6) 805-817.
Sapir M. and Churchill, G. A. "Estimating the posterior probability of differential gene expression from microarray data". JAX preprint (2000):
http://www.jax.org/research/churchill/pubs/marina.pdf
Wentian Li, Yaning Yang (2002), "How many genes are needed for a discriminant microarray data analysis?" in Methods of Microarray Data Analysis , eds. SM Lin, KF Johnson (Kluwer Academic), pp. 137-150.
http://linkage.rockefeller.edu/wli/pub/

The general idea of **using a statistic to create a ranking** (partial order of the genes)

e.g. (some obvious, but many more are possible and used)

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

One condition, replicated

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

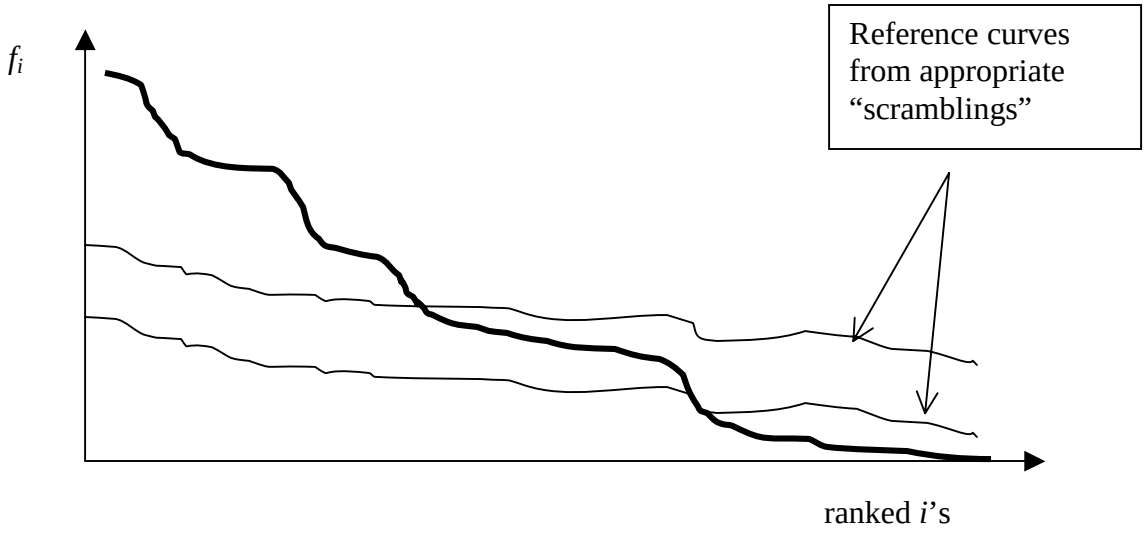
Two conditions, replicated

$$L^2_{i,1,2...T} = \left\| X_i \right\|^2_{S^{-1}}$$

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

T conditions, not replicated

The general idea of **using randomization or permutation** to compare the ranking statistic values to a chance background (instead of referring to distributional assumptions and/or asymptotics)



Replicating:

How many replicates for each experimental condition are needed to achieve a given objective ?
(e.g. reliably detect differential expression).

This issue is also related to design of experiments (see papers by Churchill's group).

Some references:

Lee M.T., Ku F.C., Whitmore G.A., Sklar J. (2000), "Importance of replication for microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations". PNAS 97(18) 9834-9839.
W Pan, J Lin, C Le, "How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach". Biostatistics, University of MN tech report, 2001: http://www.biostat.umn.edu/~weip/ge.html
Zien, Fluck, Lengauer (2002), "Microarrays: How many do you need?", in RECOMB 2002: Proceedings of the Sixth Annual International Conference on Computational Biology (ACM Press). Preprint at: http://cartan.gmd.de/~zien/publications.html

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.

A reference:

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:

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

$$X_1\ X_2 \ldots X_N$$

in *T (TxR)* dimensions

- Centering by condition (replicate) "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:

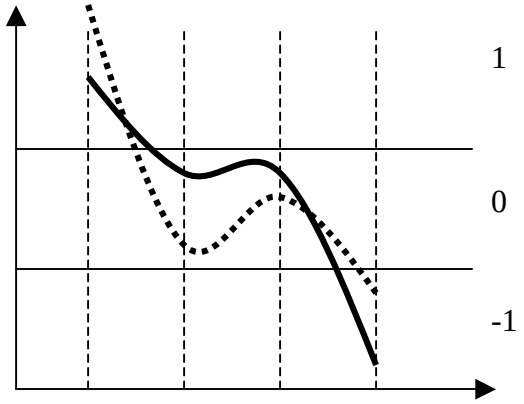
$$X_i \text{'} \ \boldsymbol{1} = 0$$

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

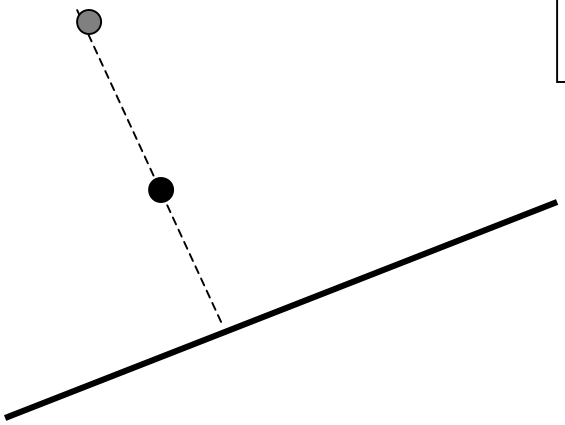
$$\| X_i \|^2 = 1$$

Are we "creating" geometrical structure?

(ii) 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?