

**Resampling-based methods for
identification of significant subsets of
genes in expression data**

Katherine S. Pollard & Mark J. van der Laan

Division of Biostatistics, U.C. Berkeley

`www.stat.berkeley.edu/~laan/`

Motivation: Microarray Data

- Observe a matrix X whose columns are n copies of a p -dimensional vector of relative gene expression measurements.
- Each measurement is a ratio, calculated from the intensities of two fluorescently labeled mRNA samples cohybridized to an array spotted with known cDNA sequences.
- Data preprocessing may include background subtraction, normalization, log transformation.
- May also observe an n -dimensional vector Y .

e.g.: Tumor vs. healthy tissues of n cancer patients.

NOTE: Methodology also applies to gene chips, where each element of X is a quantitative expression level rather than a ratio.

Goals

1. Identifying interesting subsets of genes.
 - Dimension reduction
 - Statistical significance
2. Clustering genes and/or samples.
3. Classification and prediction.
4. Defining statistical notions such as parameter, parameter estimate, consistency, and confidence.
5. Assessing the reliability of subsets, clusters, predictors.

Statistical issues are particularly crucial with the high dimensional data structures and relatively small samples of gene expression data!

Approaches to Subsetting

1. Threshold a statistic of interest

- Max correlation with another element: $\max_i(\rho_{ij})$ (Butte *et al.*)
- Mean or difference in means: μ_j or $d_j = (\mu_j^1 - \mu_j^2)$
- Standardized mean or difference in means: μ_j/σ_j or $d_j/se(d_j)$
- Proportion of log ratios exceeding a cut-off: $p_j = \frac{1}{n} \sum_i I(X_{ij} > c)$
- Regression parameter (possibly standardized): β_j or $\beta_j/se(\beta_j)$

2. Choosing the threshold

- Fold changes
- To obtain a fixed sized subset
- Testing approach (error control)
 - Tabled distributions
 - Resampling-based distributions

3. Accounting for multiple comparisons...

Testing Approach to Subsetting

Suppose $X_{p \times n} \sim P(\mu_1, \dots, \mu_p)$, where $\vec{\mu}$ is a real-valued finite dimensional parameter:

- location (means, medians, difference in means)
- covariance (correlation coefficients, distances)
- regression coefficients (linear, logistic, Cox PH)

Consider null hypotheses $H_j^0 : \mu_j = \mu_j^0, j = 1, \dots, p$.

Test H^0 with statistics such as $T_j = \frac{\sqrt{n}(\hat{\mu}_j - \mu_j^0)}{se(\hat{\mu}_j)}$. Reject H_j^0 and include gene j in the subset if $|T_j| \geq c_j$, where c_j depends on the type of error control and the null distribution.

Null Distribution

Define: P^0 to be the projection of P on the null.

e.g.: For a shift experiment, $P^0 = P(\cdot - \vec{\mu})$.

Theorem: Resampling from P^0 gives strong control for any P .

Goal: Estimation of the distribution of \vec{T}_n under P^0 . Note that asymptotically, $\vec{T}_n \sim N(0, \rho)$.

1. Tables (identically distributed? correlation?)
2. Resampling (Westfall & Young, 1993 and Dudoit *et al.*, 2002)
 - Permutations (strong control? break correlation?)
 - Parametric, e.g.: $N(0, \hat{\rho})$
 - Non-parametric Bootstrap
 - Convex-pseudo data (Breiman)
 - $2n$ out of n (Bickel)

Testing Procedures

1. Per comparison (separate inference)
2. Multiple Comparisons (MCPs)
 - single- vs. multi-step
 - more computation gives sharper bound
 - slow for p in the thousands!
 - p-values vs. cut-offs
 - can choose α later with p-values
 - can control error directly (sharp bound, usually less computation) with cut-offs
 - equivalent if both are from resampling distribution
 - why reduce resampling distribution to a p-value?
 - want a subset (i.e.: decision) in gene expression context

Gene-specific cut-offs

Marginally, the statistics $T_{j,n}$ are not identically distributed unless n is very large and pivotal (i.e.: standardized) statistics are used.

What we care about is the simultaneous *probability* $P(T_j \geq c_j)$, not necessarily a common cut-off c . In particular, for some error rate α , we want the *simultaneous* $P(T_j \geq c_j)$ for all j to be at most α .

Method: From an appropriate resampling distribution P^0 , let $c_j = q_j(1 - \tilde{\alpha})$, where $\tilde{\alpha}$ is chosen to control some family error rate at level α and $q_j(\cdot)$ is the quantile for that gene or e.g.: $\max |T|$.

| | Bonferroni/Holm | Šidák | Westfall & Young |
|-------------|------------------------|--------------------------------------|--|
| single-step | α/p | $1 - (1 - \alpha)^{1/p}$ | $q(\alpha)$ of $\max_{l \leq p} T_l $ |
| step-down | $\alpha/(p - r_j + 1)$ | $1 - (1 - \alpha)^{1/(p - r_j + 1)}$ | $q_j(\alpha)$ of $\max_{l \leq r_j} T_l $ |

note: $\{r_j\}$ are order statistics of $\{|T_j|\}$, $(p - r_j + 1) = \text{rank}(|T_j|)$.

Data Analysis

Expression of $p = 13,412$ genes was measured in malignancies of $N = 40$ DLBCL patients, 19 labeled as GC (germinal center) and 21 as Activated (Alizadeh *et al.*, 2000). GC patients have better survival, probably due to slower tumor proliferation rates.

Bootstrap resampling with Bonferroni type cut-offs $q_j(1 - \alpha/p)$, yielded 186 genes differently expressed between the two groups. This compares with only 32 found using the T distribution and a Bonferroni adjustment. None of the genes were the same.

The 186 genes were clustered using the hierarchical algorithm HOPACH with cosine-angle distance. Six distinct clusters were identified.

Summary

- Resampling methods are needed for estimation of P^0 in the gene expression context.
- The correlation matrix needs to be estimated well for strong control. Since we expect some differently expressed genes, strong control is important in this context!
- Good implementations of a variety of resampling methods are needed along with a comparison of different approaches in terms of strong control and power.
- Gene-specific cut-offs $c_j(\alpha)$ should be used, since test statistics are not identically distributed.
- These cut-offs are equivalent to computing p-values from P^0 and using adjustment approaches, but usually require less computation.