New methods for the analysis of gene expression arrays

Trevor Hastie & Robert Tibshirani
Stanford University

Joint work with David Botstein, Pat Brown, and others

Email: \{hastie,tibs\}@stat.stanford.edu

http://www-stat.stanford.edu/~\{hastie,tibs\}

Thanks to NSF, NIH.
Gene expression arrays

- Exciting new technology for measuring gene expression of thousands of genes **simultaneously** from a single sample of cells
- first multivariate, quantitative way of measuring gene expression
- a key idea: to find genes, follow around messenger RNA
- also known as DNA microarrays, gene chips—there are a number of different technologies — Affymetrix, Incyte, Brown Lab
Q: How does the microarray compare to the Northern Blot?

A: That’s like comparing Lewis and Clark to a satellite photo!
Grade 9 view of cell biology

mRNA
DNA microarray process

- microarray
- hybridization
- target sample
- reference sample
- measured fluorescence
- log(red/green)

Each sample is labelled with red and green dye.
Micro Array

The Entire Yeast Genome on a Chip
Making the Micro Array

The Biochemistry Lab of Patrick Brown

J. DeRisi
Human tumor data

6830 genes, 64 cell lines


**Statistical challenges**

- Understand patterns in data — how genes and samples are organized
- Look for genes responsible for treatment effects (cautiously).
- Biologists don’t want to miss anything (low type II error). Statisticians worry about Type I error too.
Gene Shaving

- Idea is to look for subsets of the genes showing large variation over the samples.

- Unlike global clustering methods, we hope to find more that one useful organization of the samples.

- The basic method is *unsupervised*, i.e. the samples are unlabeled. Later we discuss partially or fully supervised shaving, in which apriori labels are available for the samples.
Clusters 1-4
Clusters 5–8
Goal

- For a given cluster size $k$, we seek the cluster of genes having highest variance of its column mean, allowing sign changes for each gene.
- Infeasible to do an exhaustive search, so we use a top-down procedure based on principal components.
- Inspired by Fisher & Friedman’s PRIM method for bump hunting.
- We also would like the genes in a cluster to be highly correlated with each other. Hence we use a different measure to choose $k$ (details later).
Principal component shaving

1. Start with the entire expression array $X$, each row centered to have zero mean.

2. Compute the leading principal component of the rows of $X$.

3. Shave off the proportion $\alpha$ (typically 10%) of the rows having smallest inner-product with the leading principal component. Allow a sign flip for each gene.

4. Repeat steps 2 and 3 until only one gene remains.

5. This procedure produces a sequence of nested gene clusters $S_N \supset S_{k_1} \supset S_{k_2} \supset \cdots \supset S_1$ where $S_k$ denotes a cluster of $k$ genes. We estimate the optimal cluster size $\hat{k}$ using the gap statistic described later.

6. Orthogonalize each row of $X$ with respect to $\bar{x}$, the average gene in $S_{\hat{k}}$.

7. Repeat steps 1-5 above with the orthogonalized data, to find the second optimal cluster. This process is continued until a maximum of $M$ clusters are found.
Classification by all 8 clusters

Cluster 5

Cluster 4

Cluster 3

Cluster 2

Cluster 1

Cluster 8

Cluster 7

Cluster 6

0 BREAST
1 CNS
2 COLON
3 LEUKEMIA
4 MELANOMA
5 NSCLC
6 OVARIAN
7 PROSTATE
8 RENAL
9 UNKNOWN
Gap estimate of cluster size

- Let $d_k$ be the percent variance explained by the column means of a cluster of $k$ genes.
- The Gap function is defined by

$$\text{Gap}(k) = d_k - \bar{d}_k^*$$

(1)

where $\bar{d}_k^*$ is the average value of $d_k$ computed on permuted data matrices, obtained by permuting the elements within each row of $X$.

- We then select as the optimal number of genes that value of $k$ producing the largest gap. The idea is that at the value $\hat{k}$, the observed variance is the most ahead of expected.

- We are studying the Gap idea in simpler clustering problems (joint with Guenther Walther)
**Gap estimate: example**

**Variance Ratio plots for Real and Randomized Data**

- **Cluster 1**: Variance Ratio for Real Data decreases more sharply than for Randomized Data.
- **Cluster 2**: Variance Ratio plots show similar patterns for Real and Randomized Data.
- **Cluster 3**: Variance Ratio for Real Data shows a slight increase compared to Randomized Data.
- **Cluster 4**: Variance Ratio for Randomized Data is consistently lower than for Real Data.

**Gap Estimates of Cluster Size**

- **Cluster 1**: Gap Curve shows a peak indicating the optimal cluster size.
- **Cluster 2**: Gap Curve has a less pronounced peak.
- **Cluster 3**: Gap Curve indicates a clear optimal cluster size.
- **Cluster 4**: Gap Curve shows a peak similar to Cluster 3.

**Cluster Size**: The number of clusters identified in each set of data.

**Variance Ratio**: A measure of variability in the data, used to evaluate the stability of the clusters.

**Gap Curve**: A plot that helps in determining the optimal number of clusters by identifying the gap between the total sums of squares for a given number of clusters and the next number of clusters.
Other approaches

for generating small clusters

- Bottom up agglomeration
- Top down shaving to maximize cluster mean variance (column mean shaving)
- Global Principal component thresholding
Comparison on tumor data

Variance and Percent Variance for First Cluster

Cluster Size

Variance

Percent Variance

PC shave
Bottom up
PC threshold
Mean shave
Simulation study

100 genes, 60 samples, data has two orthogonal blocks of 10 genes that separate the samples.

<table>
<thead>
<tr>
<th>Method</th>
<th># correct/100</th>
<th>Ave Variance (s.e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC shaving</td>
<td>100</td>
<td>1.16 (0.01)</td>
</tr>
<tr>
<td>PC thresholding</td>
<td>64</td>
<td>1.06 (0.02)</td>
</tr>
<tr>
<td>Column mean shaving</td>
<td>63</td>
<td>1.10 (0.01)</td>
</tr>
</tbody>
</table>

Can understand this behavior mathematically, in this simple model.
Supervised shaving

- Here we assume samples (columns) have some attributes, like tumor class labels, a response such as survival time, etc.

- Fully supervised: look for cluster for genes whose column average has strong association with column attributes. For tumor classes: the between-class variance

- Partially supervised: use a mixture of total variance and between variance as the criterion

\[
\max_{S_k} [(1 - \alpha) \cdot Var(\bar{x}_{S_k}) + \alpha \cdot J(\bar{x}_{S_k}, y)]
\]

\[0 \leq \alpha \leq 1:\]

- \(\alpha = 0\): no supervision
- \(0 < \alpha < 1\): partial supervision
- \(\alpha = 1\): full supervision
Classification by all 8 clusters

Partially Supervised Shaving

Better class separation — 8% vs 30%
With unsupervised shaving, we need to find $w$ (suitably constrained) to maximize

$$w^T X X^T w$$

Shaving finds $w$ by sequence of principal components computation.

Computational trick: use $\text{SVD}(X)$.

If $J$ is quadratic, then

$$(1 - \alpha) \cdot \text{Var}(\bar{x}_{S_k}) + \alpha \cdot J(\bar{x}_{S_k}, y)$$

is the same as

$$(1 - \alpha)w^T X X^T w + \alpha w^T X A X^T w$$

Computational trick:

$$\text{SVD}(X \cdot ((1 - \alpha)I + \alpha A)^{\frac{1}{2}})$$
Cluster found by partial supervision, α = 0.10

Breast, Leuk, Stomut

Genes joint work with Alizadeh, Brown,
Lymphoma Subset of much larger set of 4673
350 genes, 48 samples — Diffuse Large cell

Lymphoma example

Groups defined by thresholding cluster mean gene at median. Supervision based on score test for Cox proportional-hazards model.
Validation: Permutation Tests

Partially supervised

![Graph showing training set p-value and permutation-based p-value for partially supervised method.]

Cluster size

Partially supervised

![Graph showing training set p-value and permutation-based p-value for partially supervised method.]

Cluster size

Fully supervised

![Graph showing training set p-value and permutation-based p-value for fully supervised method.]

Cluster size

Fully supervised

![Graph showing training set p-value and permutation-based p-value for fully supervised method.]

Cluster size