

Computationally intensive statistical methods for microarray based drug discovery

Katherine S. Pollard & Mark J. van der Laan

Division of Biostatistics, U.C. Berkeley

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

Motivation: Microarray Data

- We 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.

Example: 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:

- Gene clustering.
- Patient clustering.
- Simultaneous clustering of patients and genes.
- Definitions of important statistical notions such as parameter, parameter estimate, consistency, and confidence.

Tools:

- Mostly exploratory methods (cluster analysis, neural networks)
- Usually applied to genes or patients separately, but with some two-way visualization methods (Tibshirani *et al.*, 1999).

Previous Work: van der Laan & Bryan (2000)

Proposes a statistical framework for clustering genes using a deterministic rule $S(\mu, \Sigma)$. e.g.: Apply PAM to all genes at least 2-fold differentially expressed.

Provides measures and graphs of cluster stability based on the parametric bootstrap using a truncated multivariate normal $N(\hat{\mu}_n, \hat{\Sigma}_n)$.

Establishes consistency of $\hat{\mu}_n, \hat{\Sigma}_n$ and hence smooth functions $S(\hat{\mu}_n, \hat{\Sigma}_n)$ under $\frac{n}{\log(p)} \rightarrow \infty$. Also establishes consistency of the parametric bootstrap for the limiting distribution of $\sqrt{n}(\hat{\mu}_n - \mu, \hat{\Sigma}_n - \Sigma)$ and simple convergence of the bootstrap subset to the true subset, under $\frac{n}{\log(p)} \rightarrow \infty$.

Simultaneous Clustering Parameter

Definition: Define a simultaneous clustering parameter as a composition of a mapping involving clustering of patients *and* a mapping involving clustering of genes.

Clustering Patients: Given k_1 and an m -variate distribution P , let $\Phi_1(P) = (P_j(P), p_j(P) : j = 1, \dots, k_1)$ be an algorithm that maps P into k_1 m -variate distributions P_1, \dots, P_{k_1} and corresponding proportions p_1, \dots, p_{k_1} .

Clustering Genes: Given k_2 and an m -variate distribution Q , let $\Phi_2(Q) = (m_j(Q), S_j(Q), G_j(Q) : j = 1, \dots, k_2)$ be an algorithm that maps Q into k_2 $m_j(Q)$ -variate subdistributions $G_1(Q), \dots, G_{k_2}(Q)$ corresponding with subsets $S_j(Q)$ of $\{1, \dots, m\}$ of sizes m_j .

Then $\Phi_2 \circ \Phi_1(P)$ represents clustering genes within clusters of patients and $\Phi_1 \circ \Phi_2(P)$ represents clustering patients within clusters of genes.

Simultaneous Clustering Parameter (Cont.)

Estimation: We can estimate a simultaneous clustering parameter by substituting the empirical distributions.

Summary Measures:

- **Cluster Membership:** probabilities (fuzzy clustering), assignments (hard clustering), size
- **Cluster Profile:** means, medoids
- **Cluster Strength:** diameter, separation, silhouettes

Many of these measures can be summarized together in a picture. For each cluster of patients, order the gene clusters and then order patients and genes within clusters. Ordering can be based on distance between clusters, silhouette, or some dimension reducing projection (multidimensional scaling, principal components).

Important Points

1. Φ_1 and Φ_2 are defined by what algorithm you want to use (e.g.: clustering with a specified distance metric).
2. One might choose to iterate these compositions, yielding very aggressive search algorithms for finding clustering patterns in the data, like those found using gene shaving and block clustering algorithms.
3. We have formally defined parameters of interest as functions of the data generating distribution. This formalization allows us to investigate issues relating to consistency, sample size, and asymptotic validity of the bootstrap (in the context of $n \ll p$).

Inference with the Bootstrap

Nonparametric: Resample n columns from X with replacement.

Convex: For $d \in \{0, 0.5\}$, choose $\epsilon \in \{0, d\}$. Then use ϵ to form new samples as convex combinations of two randomly sampled columns of X . (Breiman, 1998)

Parametric: Fit a model (e.g.: multivariate normal, mixture of multivariate normals) and generate observations from the fitted distribution.

Apply the simultaneous clustering algorithm to each of B bootstrap samples generated in one of the above ways, keeping track of parameters of interest. The distribution of a parameter is approximated by its empirical distribution over the B samples.

Simulations to Assess the Bootstrap

Simulation 1: Multivariate Normal Data (with diagonal covariance)

Parameter:	0.9 quantile of maximum absolute difference (B=100)		
	Mean	Median	Correlation
N=40	Mean (sd) over 20 repetitions		
True distribution	0.60	0.74	0.75
Nonparametric	0.63 (0.016)	1.01 (0.031)	0.91 (0.014)
Convex d=0.1	0.59 (0.012)	0.95 (0.034)	0.91 (0.013)
Convex d=0.3	0.55 (0.014)	0.86 (0.020)	0.90 (0.011)
Convex d=0.5	0.51 (0.013)	0.78 (0.026)	0.89 (0.013)
Parametric	0.64 (0.015)	0.92 (0.016)	0.82 (0.009)
N=250	Only performed once		
True distribution	0.25	0.30	0.35
Nonparametric	0.26	0.38	0.36
Convex d=0.1	0.23	0.34	0.36
Convex d=0.3	0.21	0.30	0.36
Convex d=0.5	0.20	0.27	0.36
Parametric	0.24	0.36	0.35

Simulation 2: Mixture of 2 Multivariate Normals (diag. covariance)

Supervised Clustering:

- For each patient, compute the distance to each of the known mean vectors.
- Assign that patient to the closest cluster.
- Examine the distributions of various parameters, such as the mixing proportion and cluster means.

B=100	0.9 quantile of maximum absolute difference		0.95 CI
Parameter:	Mean 1	Mean 2	\hat{p}
N=40			
True distribution	1.29	0.82	{0.30,0.33}
Nonparametric	3.35	1.17	{0.14,0.16}
Convex d=0.1	3.24	1.14	{0.13,0.16}
Convex d=0.3	3.07	1.07	{0.15,0.17}
Convex d=0.5	2.76	0.99	{0.14,0.16}
Parametric	3.86	1.21	{0.14,0.16}
N=250			
True distribution	0.49	0.31	{0.30,0.31}
Nonparametric	1.38	0.63	{0.27,0.28}
Convex d=0.1	1.32	0.59	{0.27,0.28}
Convex d=0.3	1.20	0.54	{0.27,0.28}
Convex d=0.5	1.09	0.49	{0.27,0.28}
Parametric	1.38	0.62	{0.27,0.28}

Unsupervised Clustering:

- Apply the PAM algorithm ($k = 2$) to each bootstrap sample.
- Note that in this case the clusters do not necessarily correspond to the original clusters.
- We can nonetheless keep track of parameters that do not depend on cluster number (e.g.: features of the smallest cluster, differences between clusters).

B=100	0.95 CI		
Parameter:	Dist. Between Medoids	Diam. Smaller Clust.	Avg. Silhouette
N=40			
True dist.	{86.90,87.28}	{79.48,79.67}	{0.1268,0.1276}
Nonparametric	{82.10,84.31}	{79.22,82.90}	{0.1121,0.1352}
Convex d=0.1	{75.66,77.68}	{66.93,75.92}	{0.1062,0.1291}
Convex d=0.3	{67.15,69.20}	{58.00,67.89}	{0.1102,0.1295}
Convex d=0.5	{60.41,62.17}	{66.05,71.83}	{0.0823,0.0986}
Parametric	{90.20,90.59}	{79.13,79.45}	{0.1546,0.1557}
N=250			
True dist.	{86.27,86.65}	{80.80,80.90}	{0.1269,0.1272}
Nonparametric	{87.43,87.89}	{80.45,80.54}	{0.1338,0.1342}
Convex d=0.1	{80.09,80.69}	{78.98,79.13}	{0.1326,0.1333}
Convex d=0.3	{68.21,69.46}	{78.26,78.57}	{0.1212,0.1233}
Convex d=0.5	{61.52,63.37}	{77.88,78.20}	{0.0944,0.0996}
Parametric	{86.54,86.95}	{80.75,80.89}	{0.1301,0.1304}

Data Analysis: Golub *et. al.* (1999)

Affymetrix GeneChip measured for each of 38 leukemia patients, 27 with acute myeloid leukemia (AML) and 11 with acute lymphoblastic leukemia (ALL). Good gene expression data available for 5925 genes.

Selected a subset of 2000 genes with greatest across patient variance, clustered patients, then clustered genes within each patient group.

B=100	0.95 CI		
Silhouettes:	Patients	Genes 1	Genes 2
Null 0.95 quantile	0.0224	0.0355	0.0851
Observed Data	0.1037	0.9189	0.9213
Nonparametric	{0.2018,0.2332}	{0.9148,0.9231}	{0.9117,0.9230}
Convex d=0.1	{0.1979,0.2297}	{0.9225,0.9253}	{0.9209,0.9260}
Parametric	{0.0705,0.0874}	{0.9161,0.9193}	{0.9156,0.9188}
\hat{p} :			
Observed Data	0.6842	0.9740	0.9760
Nonparametric	{0.7326,0.7732}	{0.9715,0.9771}	{0.9666,0.9734}
Convex d=0.1	{0.7614,0.8007}	{0.9745,0.9761}	{0.9702,0.9733}
Parametric	{0.6364,0.6747}	{0.9750,0.9763}	{0.9752,0.9765}

Conclusions

1. Simultaneous clustering parameters identify interesting patterns in gene expression data, including genes specifically upregulated in subtypes of acute leukemia.
2. The bootstrap can be used to estimate the variability of simultaneous clustering parameters.
 - Nonparametric and parametric bootstrap are both OK for gene clustering and summary measures of unsupervised patient clustering.
 - Using convex pseudo-data is not an improvement over the nonparametric bootstrap.
 - For small n , the parametric bootstrap may be optimistic about patient clustering.