

Statistical inference for simultaneous clustering of gene 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.

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.
2. Clustering:
 - Genes.
 - Patients.
 - Genes and patients *simultaneously*.
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.

Statistical Framework

1. Lockhart & Winzeler (2000): Review.
2. Hughes *et al.* (2000): Monte-Carlo randomization in binary trees.
3. Dudoit *et al.* (2000,2001): Bootstrap classification.
4. Kerr & Churchill (2001): Bootstrap ANOVA model residuals.
5. Yeung *et al.* (2001): Jackknife clustering results.
6. Getz *et al.* (2001): Permutations to determine stable clusters.
7. van der Laan & Bryan (2000):
 - Statistical framework for subsetting and clustering genes using a deterministic rule $S(\mu, \Sigma)$.
 - Consistency of $\hat{\mu}_n, \hat{\Sigma}_n$ and hence smooth functions $S(\hat{\mu}_n, \hat{\Sigma}_n)$ under $\frac{n}{\log(p)} \rightarrow \infty$.
 - Consistency of the parametric bootstrap for the limiting distribution of $\sqrt{n}(\hat{\mu}_n - \mu, \hat{\Sigma}_n - \Sigma)$ under $\frac{n}{\log(p)} \rightarrow \infty$.

Simultaneous Clustering Parameter

Definition: For a data generating distribution P , define a simultaneous clustering parameter $S(P)$ as a composition of mappings involving clustering of patients *and/or* mappings involving clustering of genes.

Estimation: We can estimate a simultaneous clustering parameter by substituting the empirical distribution P_n for P .

Pollard & van der Laan (2002)

- Extends statistical framework to simultaneous clustering.
- Establishes asymptotic validity of parametric *and* non-parametric bootstrap methods for estimating the distribution of $S(P_n)$.

Subsetting

Approaches to subsetting genes:

1. Threshold maximum correlation with another element $\max_i(\rho_{ij})$ (Butte *et al.*).
2. Threshold proportion of log ratios exceeding a cut-off p_j .
3. Threshold mean log ratios μ_j .
4. Threshold standardized mean log ratios μ_j/σ_j
 - Based on $N(0, 1)$ quantiles.
 - Adjust for multiple comparisons (Bonferroni, step-down).
 - Based on t-statistics from randomly permuted data (Tusher *et al.* and others).
 - Based on null distribution with zero means and the true covariance structure (using parametric $N(0, \hat{\rho})$ or nonparametric bootstrap). Can control family-wise type I error or expected number of type I errors.

Clustering

Algorithms map a $p \times p$ dissimilarity matrix \mathbf{D} into p cluster labels.

Approaches:

- Supervised (COBWEB, SVMs, CART, gene-shaving) vs. Unsupervised
- Model-based (AUTOCLASS, SNOB) vs. Nonparametric
- Partitioning (SOMs, PAM, MASLOC, KMEANS) vs. Hierarchical
 1. Agglomerative (single, complete, and average linkage CLUSTER, AGNES)
 2. Divisive (SOTA, DIANA, TSVQ)
- Graphical approaches (CAST)

Clustering

Simultaneous Clustering: Usually genes and patients clustered separately, but with some two-way visualization methods.

Refs: Tibshirani *et al.* (1999), Getz, Levine and Dommany (2000), and Fellenberg *et al.* (2001).

Nice Properties for Clustering Algorithms:

- General distance metric.
- Robust cluster profiles.
- Sensible ordering (hierarchical).
- Identify parameters of biological interest.

Clustering

Examples:

1. **Partitioning Around Medoids (PAM)**, Kaufman & Rousseeuw (1990).

- Minimizes over the vector of K potential medoids $\sum_j d_1(x_j, M)$.
- Each medoid identifies a cluster.

2. **PAMSIL**, van der Laan, Pollard & Bryan (2001).

- Replaces the PAM criteria function with average silhouette.

$$S_j = \frac{b_j - a_j}{\max(a_j, b_j)},$$

where a_j, b_j are the average dissimilarities of element j with the members of its own cluster and its neighboring cluster, respectively. Average silhouette is the mean over j .

- More “efficient”, identifies small clusters.

Clustering

3. Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH), van der Laan & Pollard (2001).

- Builds a tree of clusters.
- At every level, each cluster is split into two or more smaller clusters.
- Clusters are ordered deterministically based on \mathbf{D} .
- Collapsing steps correct errors.
- Produces a final ordering that improves on current algorithms.

These new clustering algorithms allow us to find small clusters of interesting genes in the presence of many “noisy” genes or a large gene cluster.

Choosing the Number of Clusters

- **Global Criteria:**

1. 30 methods reviewed by Milligan & Cooper, 1985.
2. Phase transitions in simulated annealing (Rose *et al.*, 1990).
3. Graph theory (e.g.: cliques in CAST) (Ben-Dor *et al.*, 1999).
4. Model-based methods (Scott & Symmons, 1971, Roeder, 1994).
5. Average silhouette (Kaufman & Rousseeuw, 1990).

- **Resampling Methods:**

1. Gap statistic (Tibshirani *et al.*, 2000).
2. WADP (Bittner *et al.*, 2000).
3. Clest (Dudoit & Fridlyand, 2001).
4. Bootstrap (van der Laan & Pollard, 2001).

Choosing the Number of Clusters

Problem: Most existing criteria identify global structure only.

Approach: For each of a series of proposed clustering results, apply the clustering routine independently to each of the clusters and evaluate the global criteria to obtain a measure of cluster heterogeneity. Average over clusters. The minimum indicates the clustering result with most homogeneous clusters.

- Any global criteria.
- Any clustering routine.

Illustration:

- Criteria=silhouette,
- Clustering=PAM, HOPACH.

Mean Split Silhouette

Given a clustering with K clusters, consider each cluster $k = 1, \dots, K$ separately.

1. Apply the clustering algorithm to the elements in cluster k .
2. Choose the number of child clusters that maximizes mean silhouette.
3. Call this maximum the split silhouette, SS_k .

Define **Mean Split Silhouette** as:

$$MSS(K) = \frac{1}{K} \sum_{k=1}^K SS_k$$

MSS measures average cluster heterogeneity.

Replace means with medians for a robust version.

Method: Choose the number of clusters K which minimizes $MSS(K)$.

MSS Simulations

- Nested data structures
- Genes, patients and patients within gene clusters
- Different distance metrics
- PAM and HOPACH

Results:

1. Minimum of MSS is at correct number of clusters (i.e.: identifies finer structure in the data)
2. MSS chooses fewer clusters if they overlap
3. MSS can be used to choose 1 cluster
4. MSS is computationally easy

HOPACH with MSS

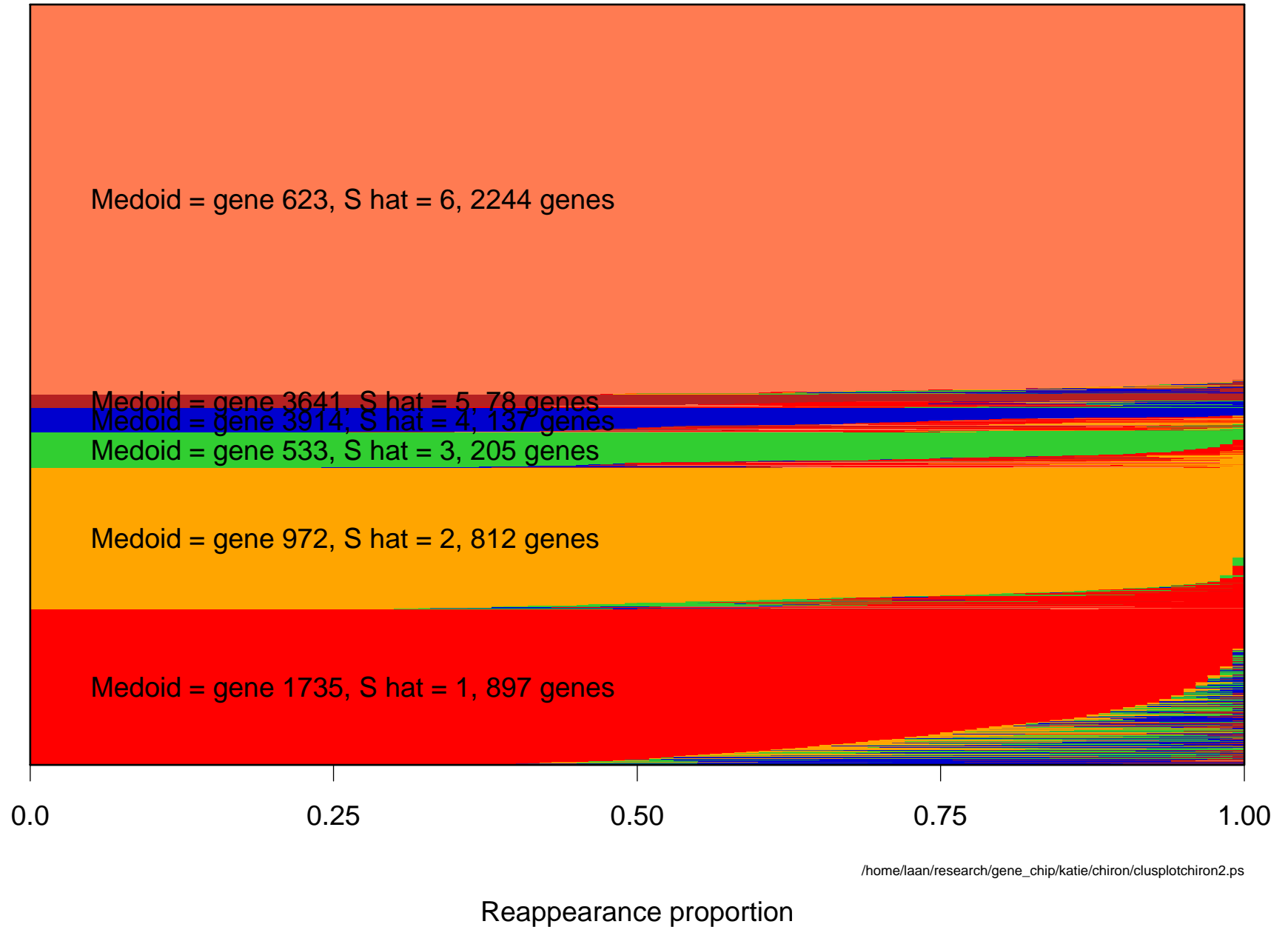
- Use MSS to automate:
 1. Number of clusters in each split
 2. Collapsing steps
 3. Stopping rule
- This HOPACH can be used to identify clusters *or* to produce an ordered list
- Plot the history of MSS as the tree unfolds

Summary

- Simultaneous clustering reveals interesting patterns in gene expression data, for example, subpopulations with different gene expression patterns.
- Statistical inference is possible with gene expression data using the bootstrap and appropriate null distributions.
- New clustering algorithms help us to find biologically meaningful groups of genes and samples and to produce sensible ordered lists.
- Visualization of ordered data and distance matrices reveals the underlying cluster structure.
- MSS criteria identifies the finer structure in gene expression data.

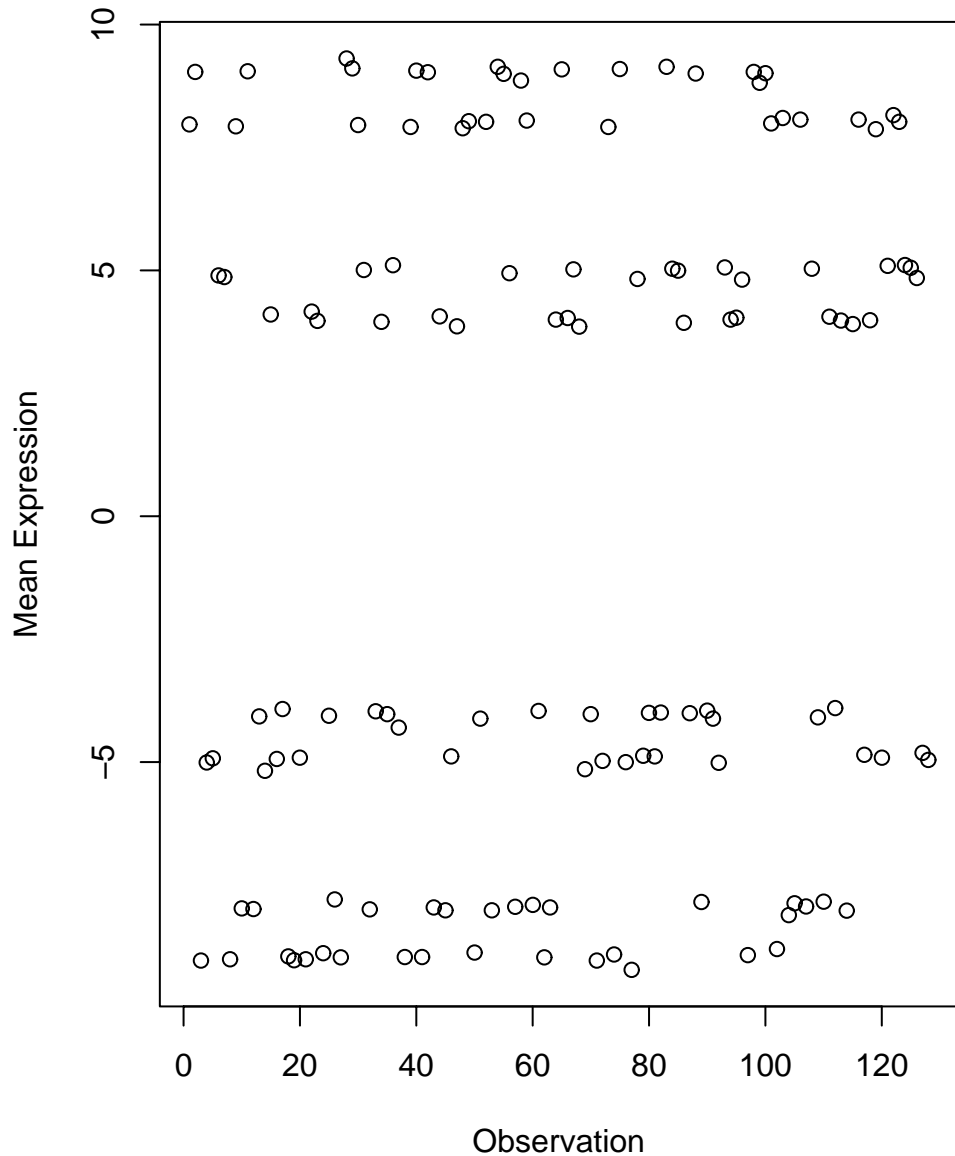
Reappearance proportions and cluster reproducibility

Subset contains 4373 genes.

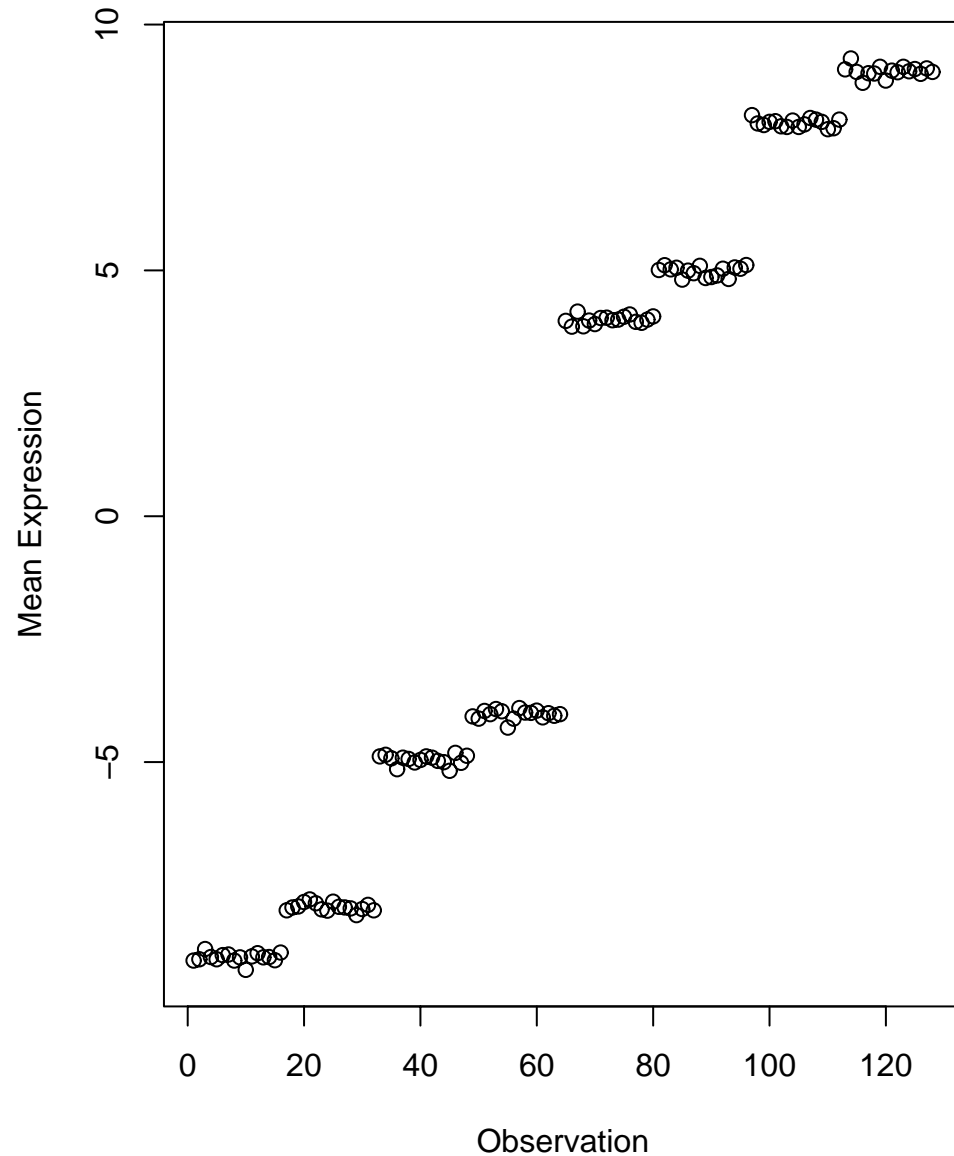


Data from 8 clusters of size 16: $\mu=(-9,-8,-5,-4,4,5,8,9)$, $\sigma=0.1$

Randomized Data

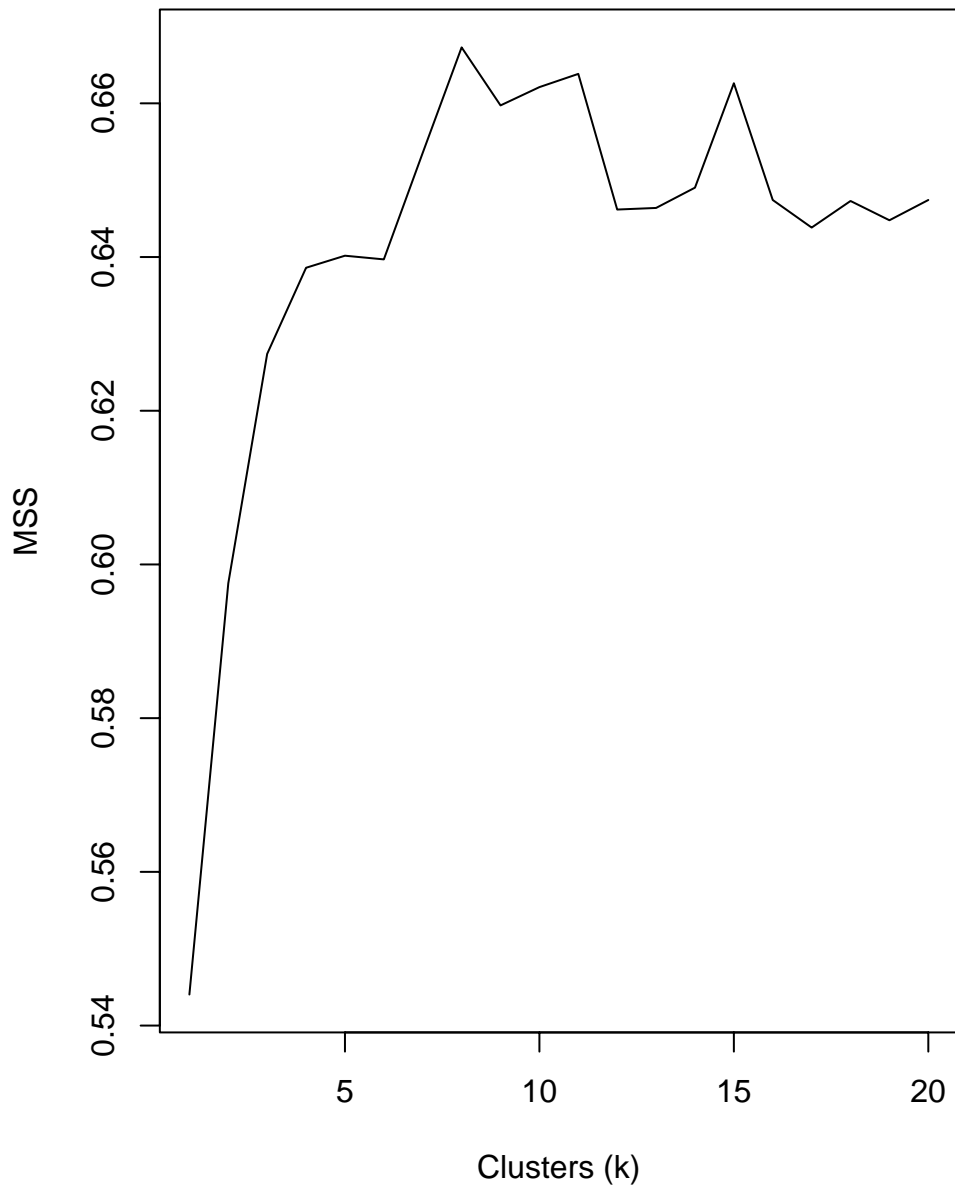


Ordered Data



Univariate normal data: $n=360$, $p=1$, $\mu=0$, $\sigma=0.05$

PAM



HOPACH

