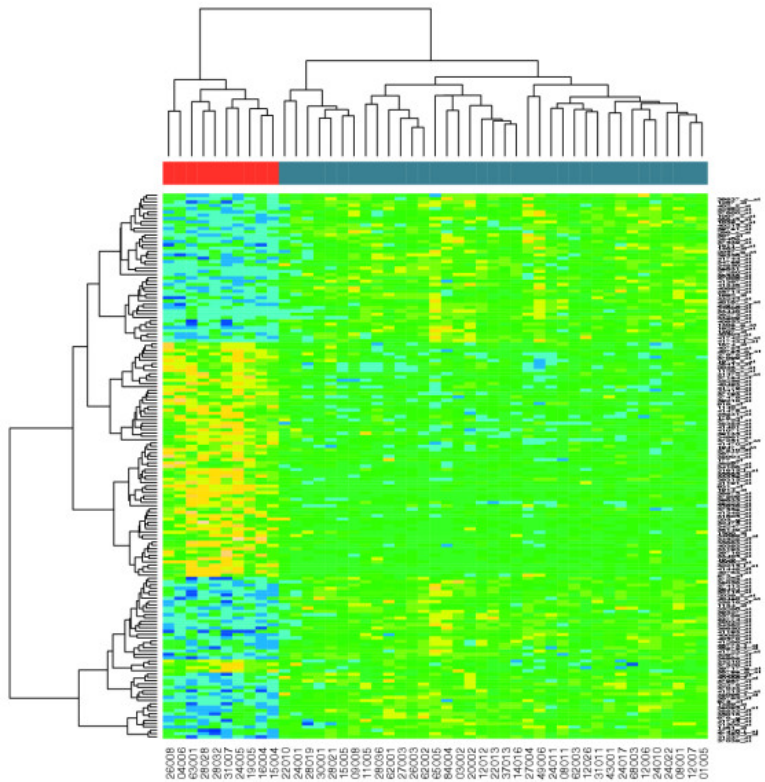
A penalized matrix decomposition, with application to sparse hierarchical clustering

Daniela Witten
Department of Statistics
Stanford University

October 20, 2009

Hierarchical clustering

There has been a resurgence of interest in hierarchical clustering in the field of genomics.



Clustering when $p \gg n$

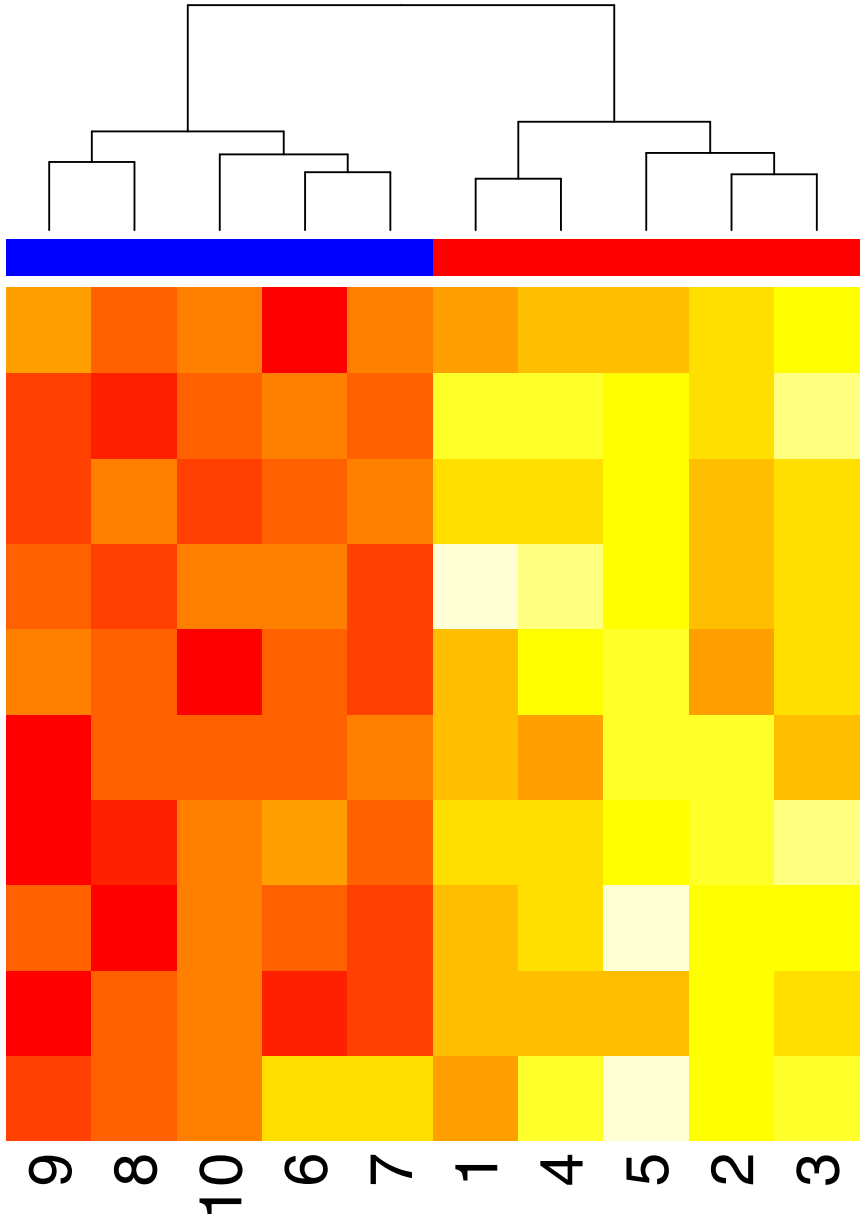
Suppose we wish to cluster n observations on p features, where $p \gg n$.

If the true underlying classes are defined on only a subset of the features, then the presence of noise features can obscure this signal.

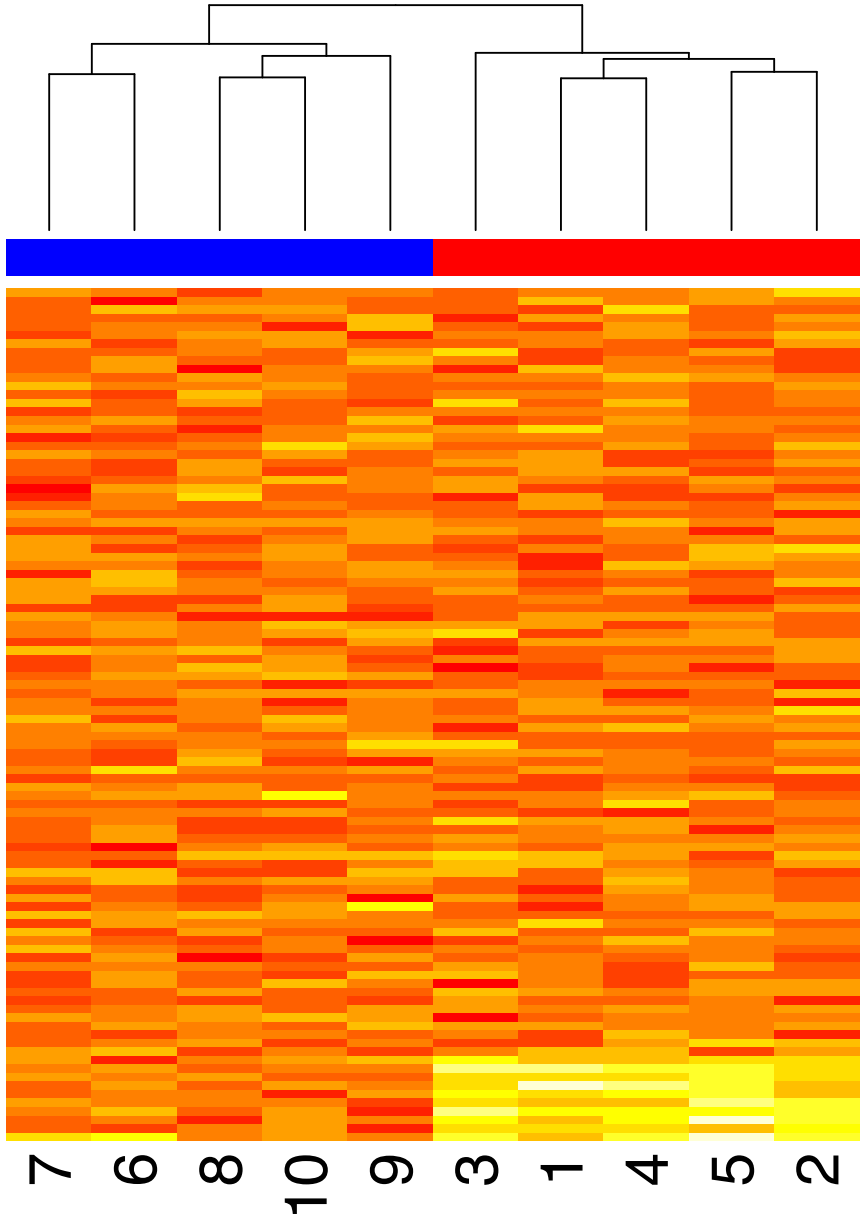
Example

A simple example with 10 observations; 2 classes are defined on 10 important features.

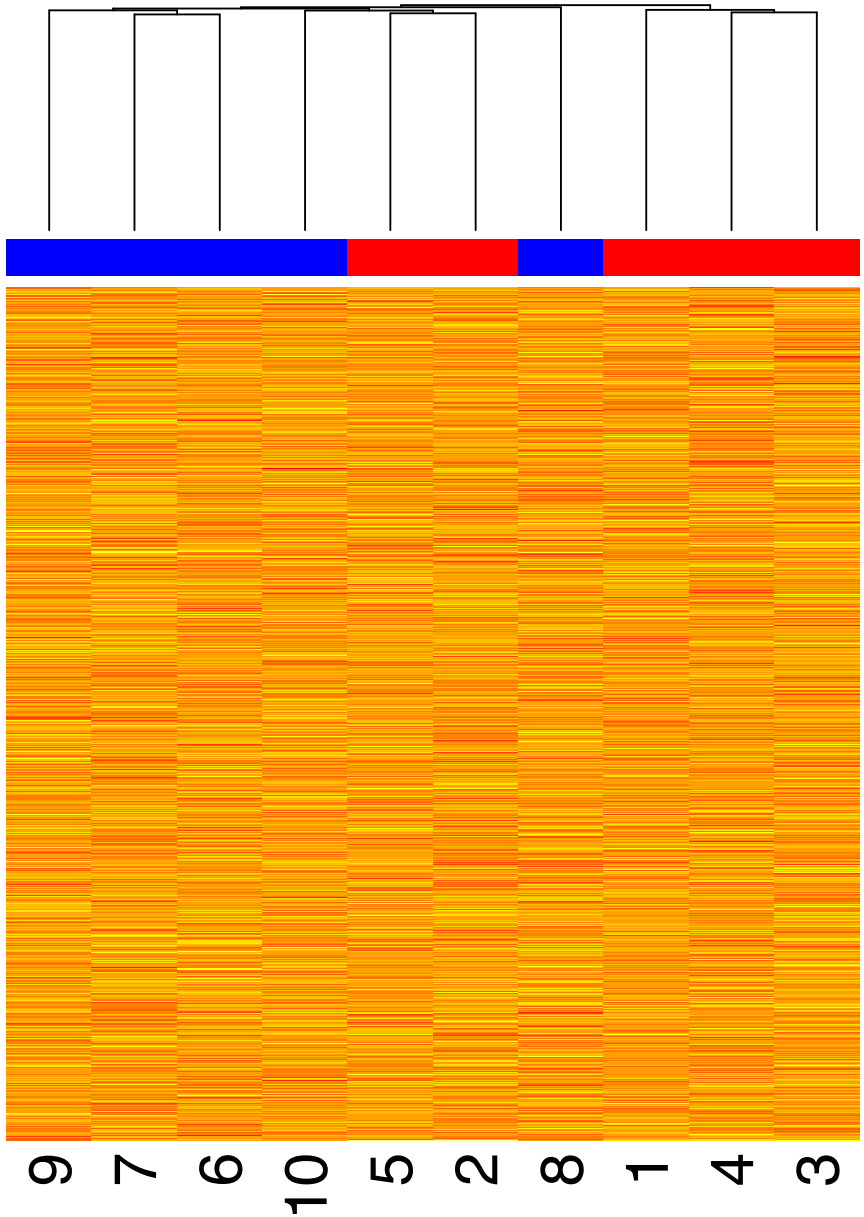
Example: 10 important features; 10 features total



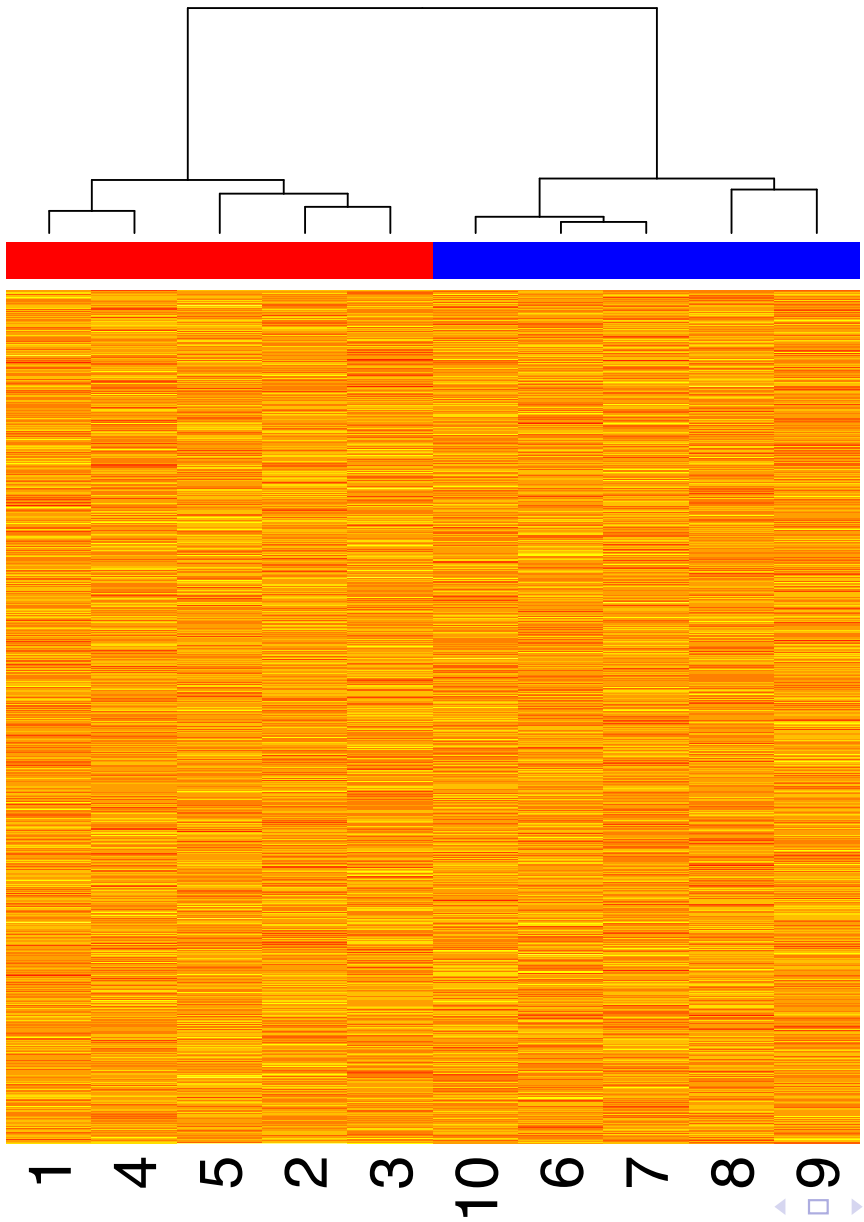
Example: 10 important features; 500 features total



Example: 10 important features; 5000 features total



Sparse hierarchical clustering results: 10 important features; 5000 features total



Sparse Clustering

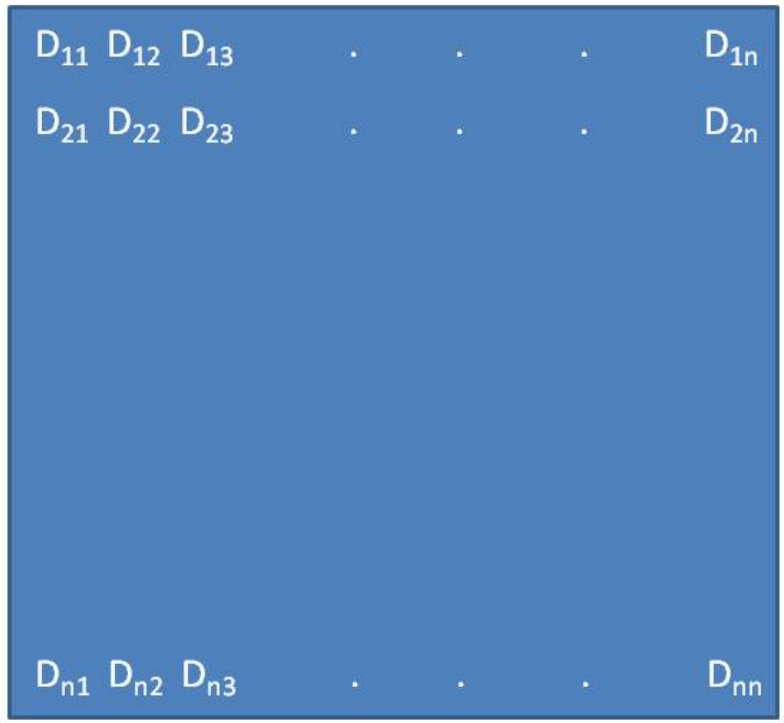
We want a method to hierarchically cluster observations based on a small subset of the features; we will call this sparse hierarchical clustering.

We want an automated way to

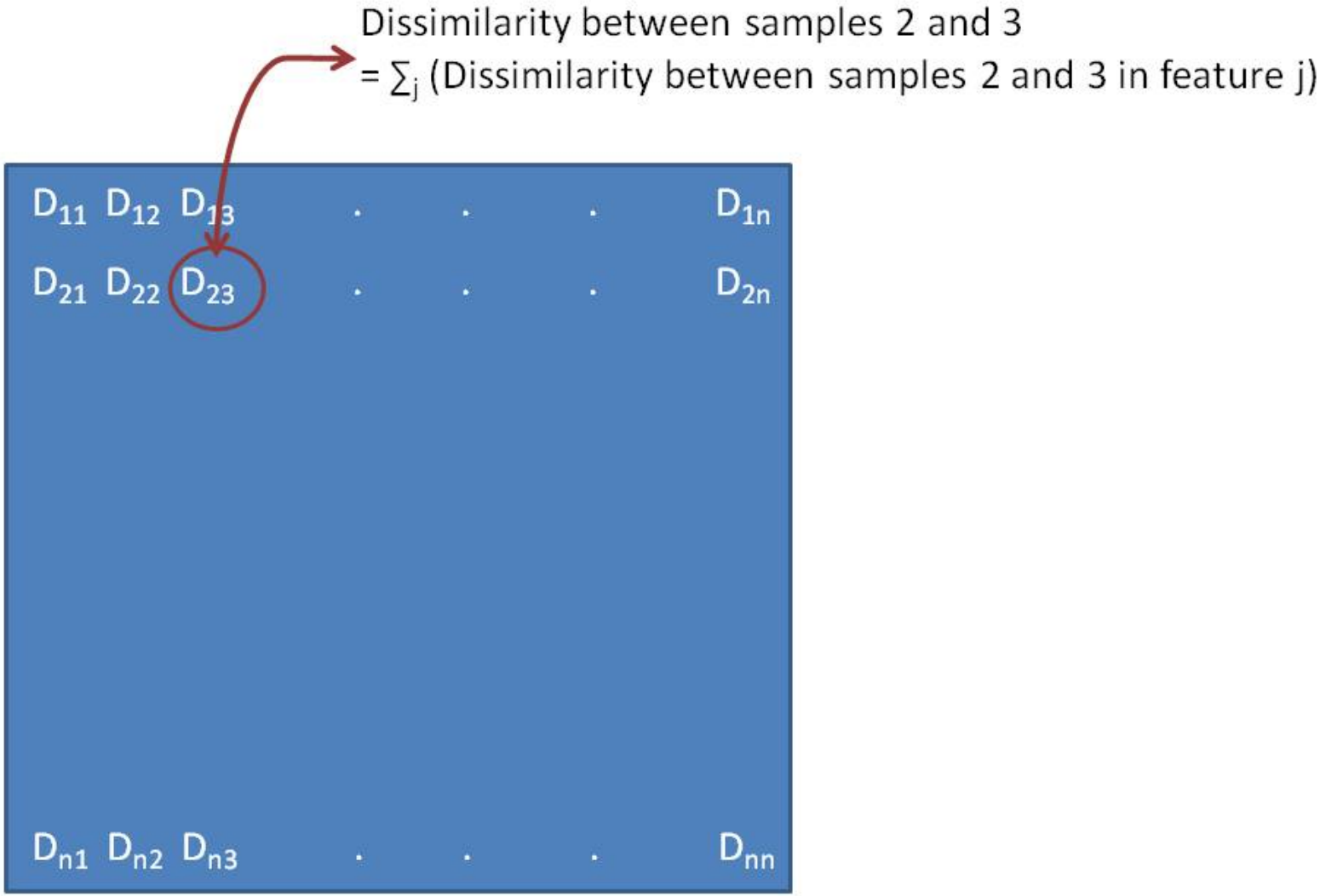
- find a subset of features to use in the clustering, and
- obtain a more accurate or interesting clustering using that subset of features.

Assumption: We assume that the dissimilarity measure used is additive in the features: $D_{i,i'} = \sum_{j=1}^{p} d_{i,i',j}$

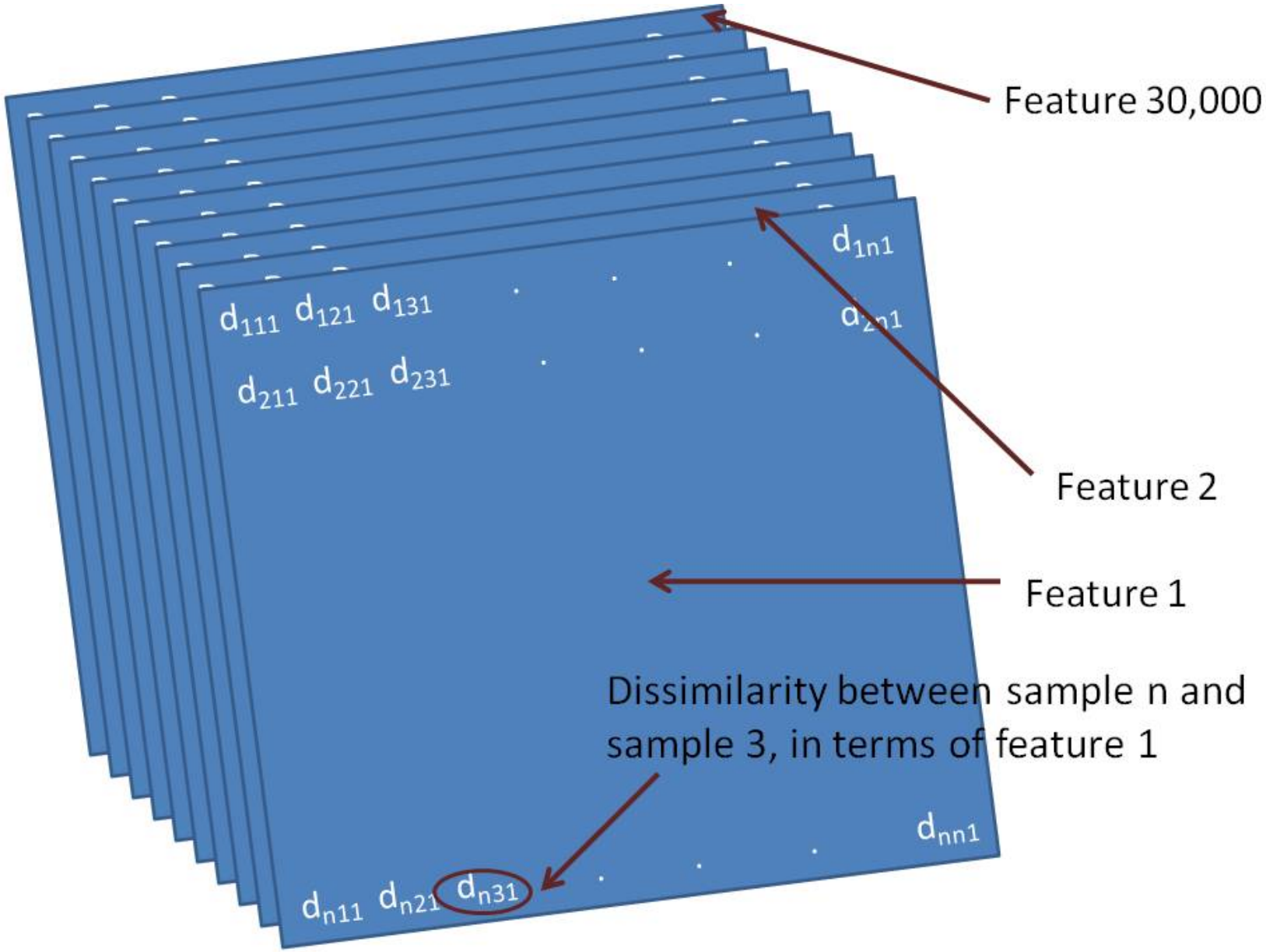
Dissimilarity matrix for the n observations



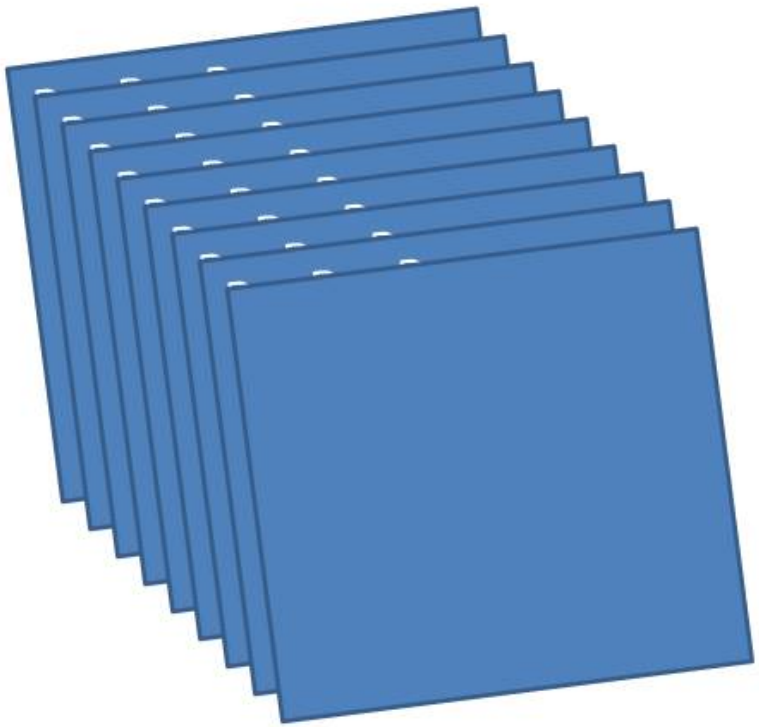
Dissimilarity matrix for the n observations



Dissimilarity matrix is a sum of dissimilarity matrices over the features



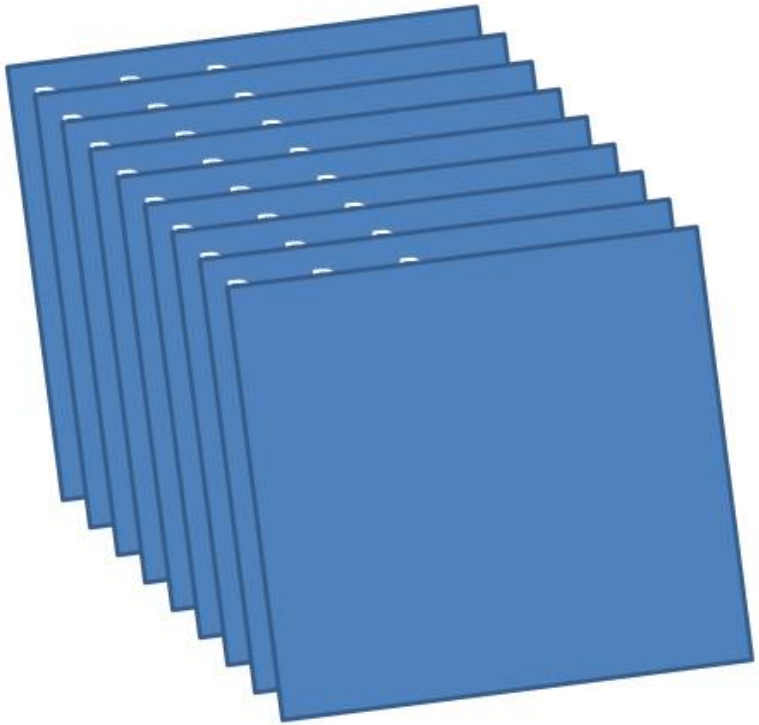
Hierarchical clustering sums the dissimilarity matrices for the features



Addition of dissimilarity matrix for each feature gives total dissimilarity matrix

$$\sum_j d_{ii'j} = D_{ii'}$$

Weighted sum of the dissimilarity matrices for the features



We take a weighted sum of the dissimilarity matrices for each feature, where the features with dissimilarity matrices that agree with each other are given large weights.

$$\sum_j w_j d_{ii'j} = D_{ii'}$$

Sparse hierarchical clustering and the PMD

Let $\mathbf{D}$ denote the $n^2 \times p$ matrix for which column j is the feature-wise dissimilarity matrix for feature j.

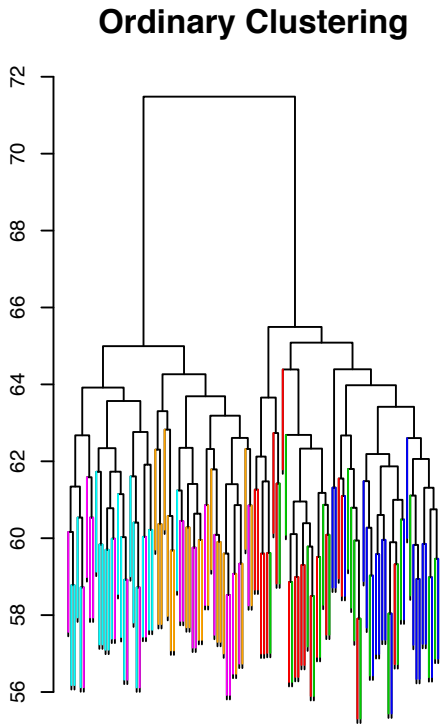
Then, suppose we apply the PMD to $\mathbf{D}$:

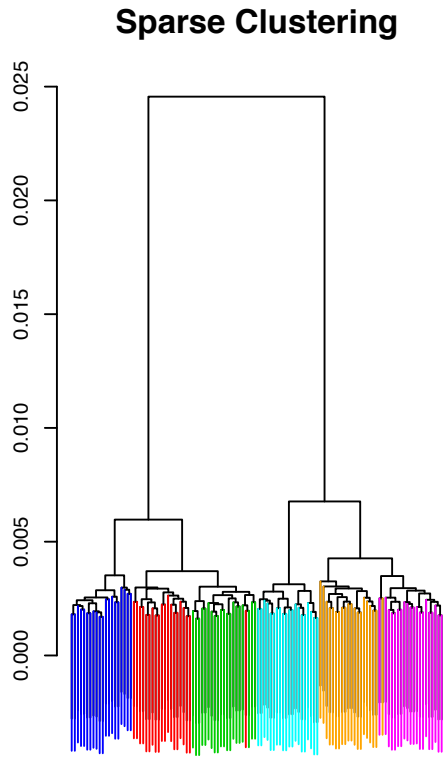
$$\underset{\mathbf{u},\mathbf{w}}{\text{maximize}}\, \mathbf{u}^T \mathbf{D}\mathbf{w} \text{ subject to } ||\mathbf{u}||_2 \leq 1, ||\mathbf{w}||_2 \leq 1, \sum_j w_j \leq s, w_j \geq 0$$

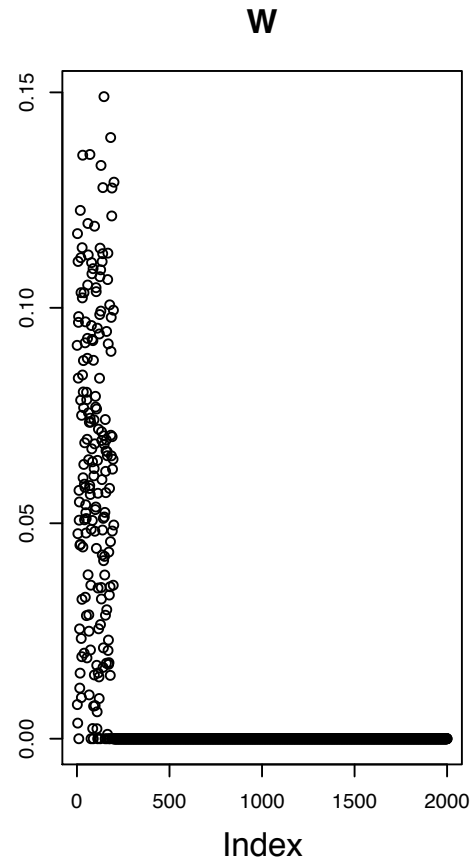
Then, w_j is a weight on the dissimilarity matrix for feature j. If we re-arrange the elements of $\mathbf{Dw}$ into a $n \times n$ matrix, then performing hierarchical clustering on this re-weighted dissimilarity matrix gives sparse hierarchical clustering.

Sparse hierarchical clustering in action

A simulated example with 6 classes defined on 100 signal features; 2000 features in total.







An important breast cancer paper

Nature (2000) **406**:747-752.

letters to nature

Molecular portraits of human breast tumours

Charles M. Perou*†, Therese Sørlie†‡, Michael B. Eisen*, Matt van de Rijn§, Stefanie S. Jeffrey‖, Christian A. Rees*, Jonathan R. Pollack¶, Douglas T. Ross¶, Hilde Johnsen‡, Lars A. Akslen#, Øystein Fluge☆, Alexander Pergamenschikov*, Cheryl Williams*, Shirley X. Zhu§, Per E. Lønning*, Anne-Lise Børresen-Dale‡, Patrick O. Brown¶†† & David Botstein*

* *Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA*

‡ *Department of Genetics, The Norwegian Radium Hospital, N-0310 Montebello Oslo, Norway*

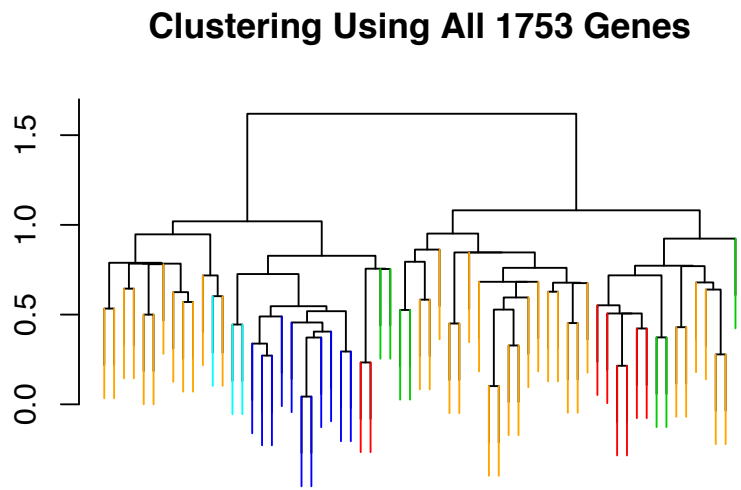
§ *Department of Pathology, Stanford University School of Medicine, Stanford, California 94305, USA*

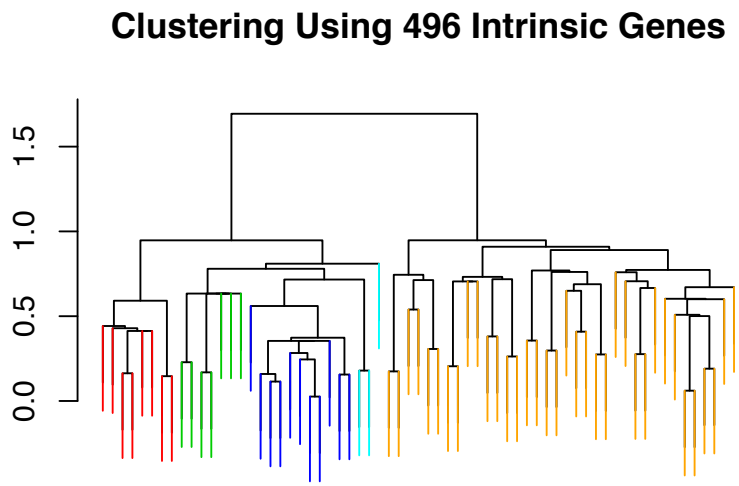
‖ *Department of Surgery, Stanford University School of Medicine, Stanford, California 94305, USA*

Breast cancer data

- 65 breast tumor samples for which gene expression data is available. Some samples are replicates from the same tumor (before and after chemo).
- Clustered based on full set of 1753 genes first.
- Clustered based on 496 intrinsic genes for which the variation between different tumors is large relative to the variation within a tumor.
- Based on the intrinsic gene clustering, determined that 62 of 65 tumors fall into one of four classes: normal-breast-like, basal-like, ER+, Erb-B2+.

Clustering results: normal-breast-like, basal-like, ER+, Erb-B2+





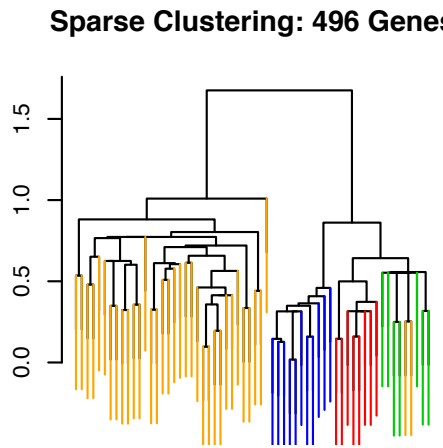
Sparse clustering

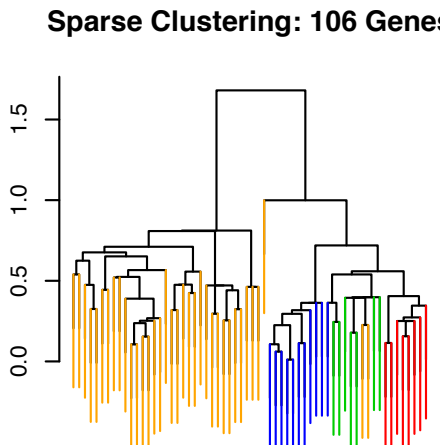
We wonder: If we sparsely cluster the observations using all of the genes, can we identify the four classes successfully?

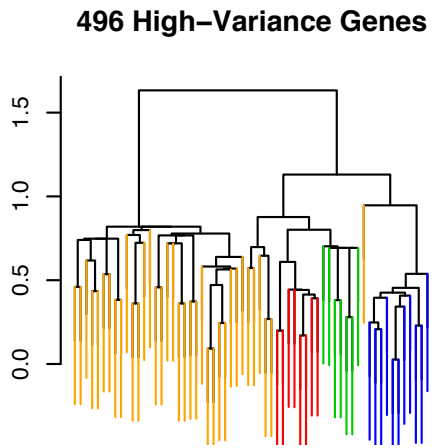
Three types of clustering:

1. Sparse hierarchical clustering of all 1753 genes, with the tuning parameter chosen to yield 496 genes.
2. Sparse hierarchical clustering of all 1753 genes, with the tuning parameter chosen by the gap statistic.
3. Standard hierarchical clustering using the 496 genes with highest marginal variance.

normal-breast-like, basal-like, ER+, Erb-B2+







Genes with high weights

#	Gene	Weight
1	S100 CALCIUM-BINDING PROTEIN A8 (CALGRANULIN A)	0.223
2	SECRETED FRIZZLED-RELATED PROTEIN 1	0.2126
3	ESTROGEN RECEPTOR 1	0.2076
4	KERATIN 17	0.1627
5	HUMAN REARRANGED IMMUNOGLOBULIN LAMBDA	0.1568
6	CYTOCHROME P450, SUBFAMILY IIA	0.155
7	APOLIPOPROTEIN D	0.1509
8	LACTOTRANSFERRIN	0.1471
9	ESTROGEN RECEPTOR 1	0.1405
10	134783	0.14
11	HEPATOCYTE NUCLEAR FACTOR 3, ALPHA	0.1332
12	HUMAN REARRANGED IMMUNOGLOBULIN LAMBDA LIGHT	0.1309
13	FATTY ACID BINDING PROTEIN 4, ADIPOCYTE	0.1292
14	CERULOPLASMIN (FERROXIDASE)	0.126
15	HUMAN SECRETORY PROTEIN (P1.B) MRNA	0.1208
16	NON-SPECIFIC CROSS REACTING ANTIGEN	0.1199
17	LIPOPROTEIN LIPASE	0.1123
18	IMMUNOGLOBULIN LAMBDA LIGHT CHAIN	0.112
19	CRYSTALLIN, ALPHA B	0.1108
20	FATTY ACID BINDING PROTEIN 4, ADIPOCYTE	0.11
21	PLEIOTROPHIN (HEPARIN BINDING GROWTH FACTOR 8)	0.1099
22	85660	0.1077
23	ESTS, HIGHLY SIMILAR TO PROBABLE ATAXIA-TELANGIECTASIA	0.1071
24	V-FOS FBJ MURINE OSTEOSARCOMA VIRAL ONCOGENE HOMOLOG	0.1056
25	EPIDIDYMIS-SPECIFIC, WHEY-ACIDIC PROTEIN TYPE	0.1013
26	ALDO-KETO REDUCTASE FAMILY 1, MEMBER C1	0.1007

Conclusions

- Clustering methods are very sensitive to the set of features used.
- In high dimensions, we may not want to simply use all of the features that happen to be available.
- Objective methods are required for selecting the features for use in clustering.
- This proposal can be applied to K-means clustering, hierarchical clustering, K-medoids clustering, and more.
- Unsupervised learning when $p \gg n$: need better ways to select tuning parameters and validate results obtained.
- R package `sparcl: Sparse Clustering`.
- Witten and Tibshirani (2010) 'A framework for feature selection in clustering', *JASA (T & M)* **105(490)**: 713-726.

References

1. Chin et al. (2006) 'Genomic and transcriptional aberrations linked to breast cancer pathophysiologies', *Cancer Cell* **10**: 529-541.
2. Perou et al. (2000) 'Molecular portraits of human breast tumours', *Nature* **406**: 747-752.
3. Shen and Huang (2008) 'Sparse principal component analysis via regularized low rank matrix approximation', *Journal of Multivariate Analysis* **6**: 1015-1034.
4. Witten, Tibshirani, and Hastie (2009) 'A penalized matrix decomposition, with applications to canonical correlation analysis and principal components', *Biostatistics* **10(3)**: 515-534.
5. Witten and Tibshirani (2009) 'A framework for feature selection in clustering', *Submitted*.