

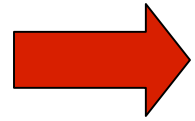
Principal component analysis (PCA) for clustering gene expression data

Ka Yee Yeung

Walter L. Ruzzo

Bioinformatics, v17 #9 (2001) pp 763-774

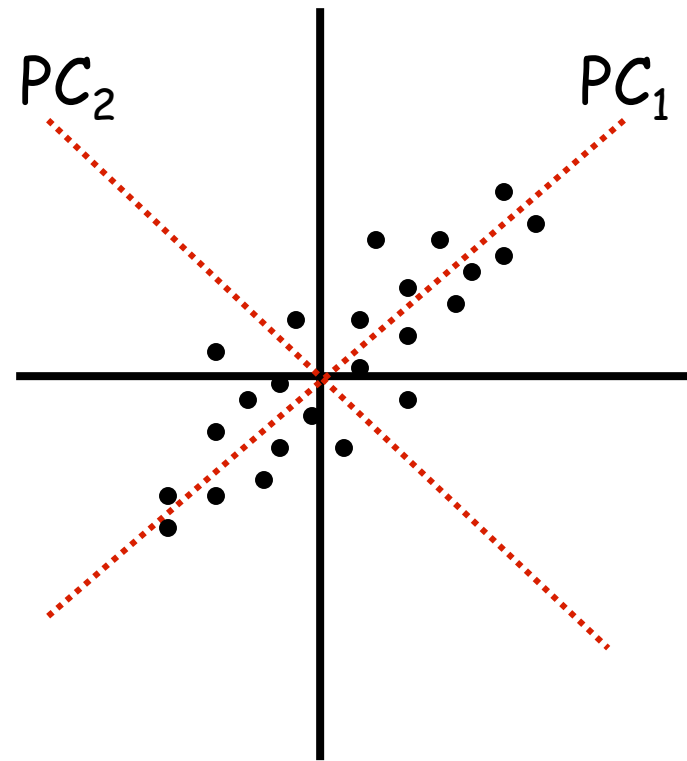
Outline of talk



- Background and motivation
- Design of our empirical study
- Results
- Summary and Conclusions

Principal Component Analysis (PCA)

- Reduce dimensionality
- Retain as much variation as possible
- Linear transformation of the original variables
- Principal components (PC's) are uncorrelated and ordered



Definition of PC's

- FIRST principle component - the direction which maximizes variability of the data when projected on that axis
- Second PC - the direction, among those orthogonal to the first, maximizing variability
- ...
- Fact: They are eigenvectors of $A^T A$; eigenvalues are the variances

Motivation

- Chu et al. [1998] identified 7 clusters using Eisen *et al.*'s CLUSTER software (hierarchical centroid-link) on the yeast sporulation data set.
- Raychaudhuri *et al.* [2000] applied PCA to the sporulation data, and claimed that the data showed a unimodal distribution in the space of the first 2 PC's.

PC's in Sporulation Data

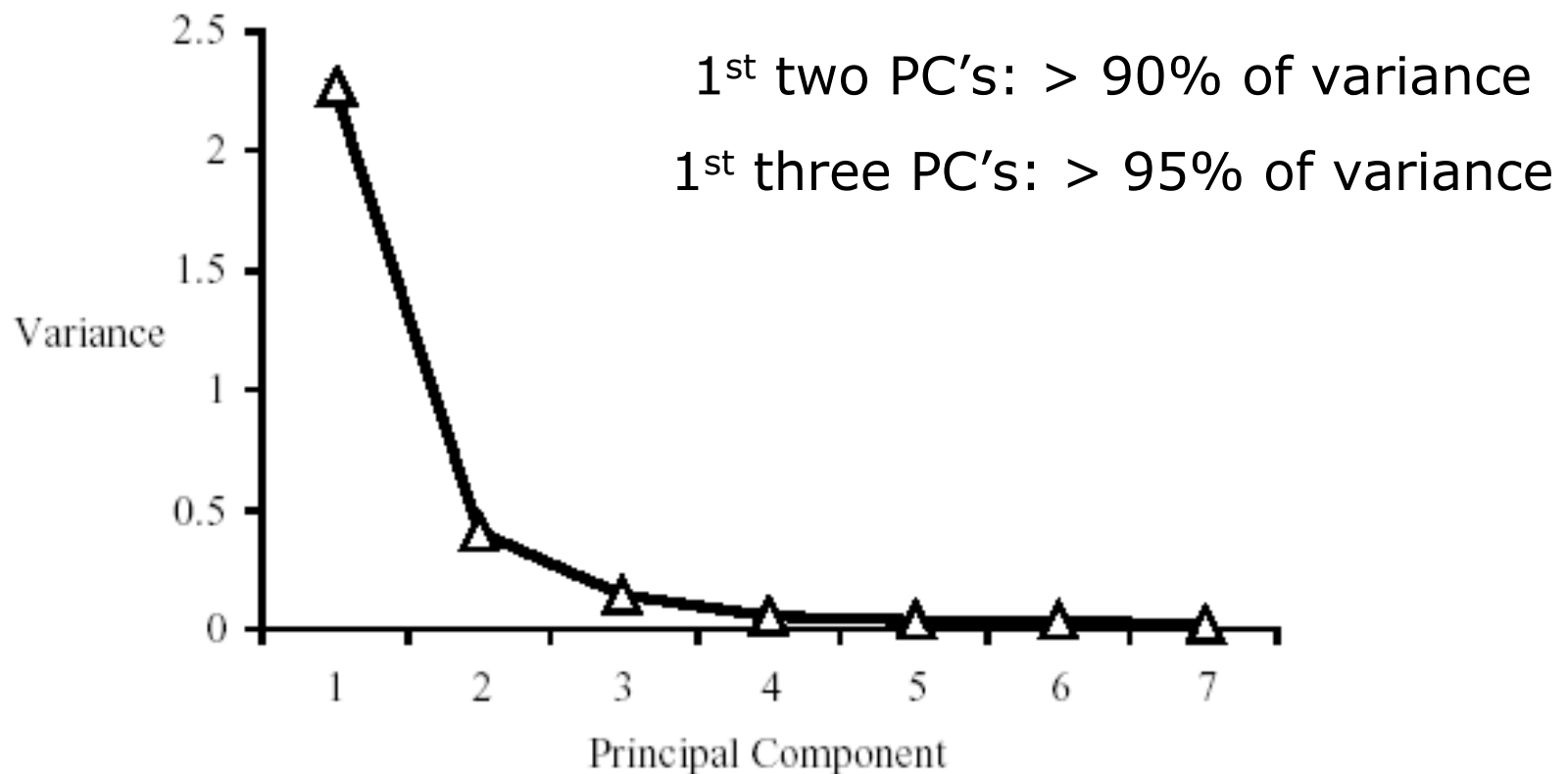
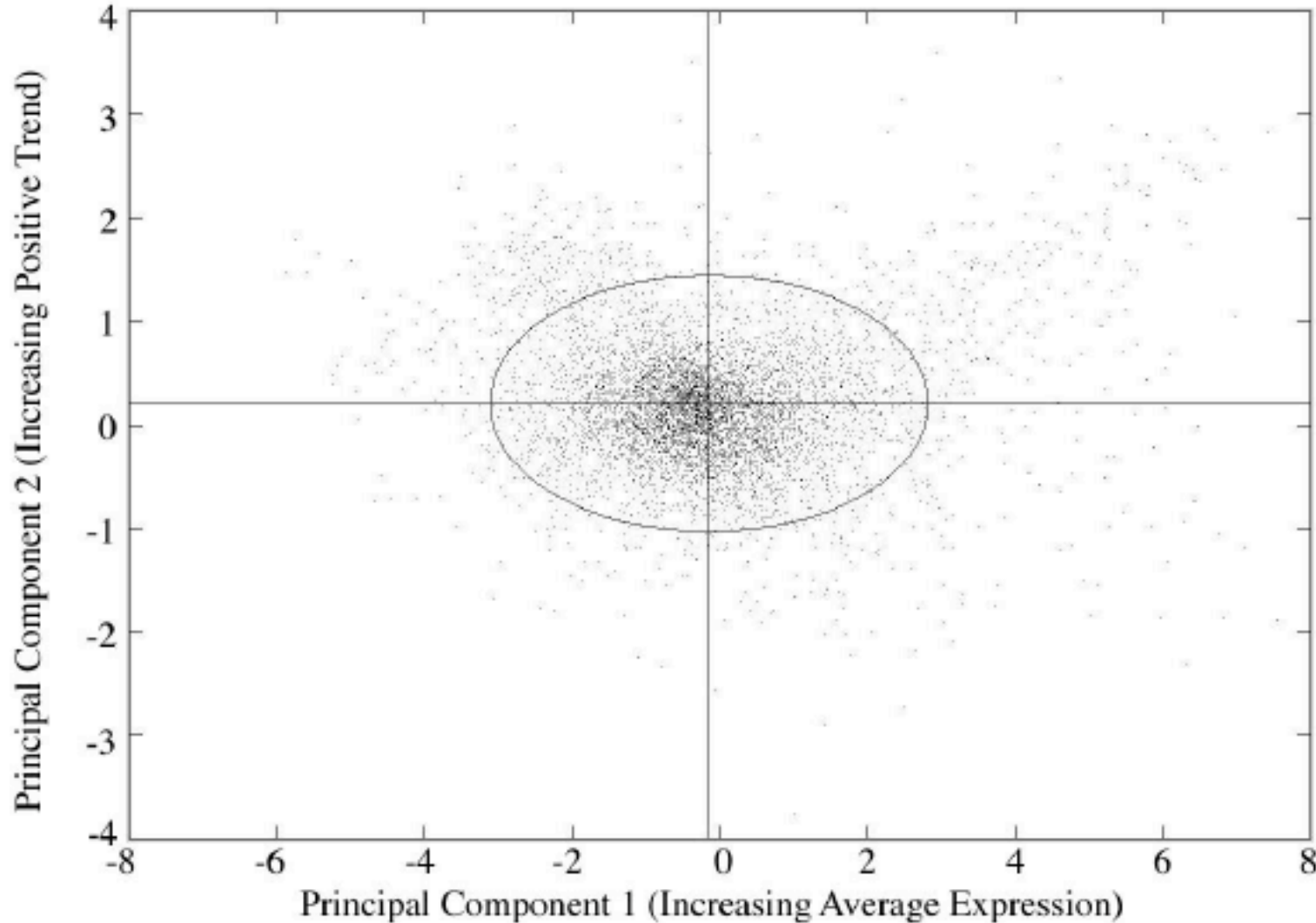


Figure 1. Plot of eigenvalues of the principal components. Most of the variance in the sporulation data set is contained in the first two principal components.



“The unimodal distribution of expression in the most informative two dimensions suggests the genes do not fall into well-defined clusters.”

-- Raychaudhuri et al.

Figure 3. The rotated and dimensionally reduced expression data. All yeast genes are plotted on to the first and second principal components. The first principal component is a measure of total average expression, the second is a measure of increasing expression with respect to time. The ellipse at the center contains 95% of the genes.

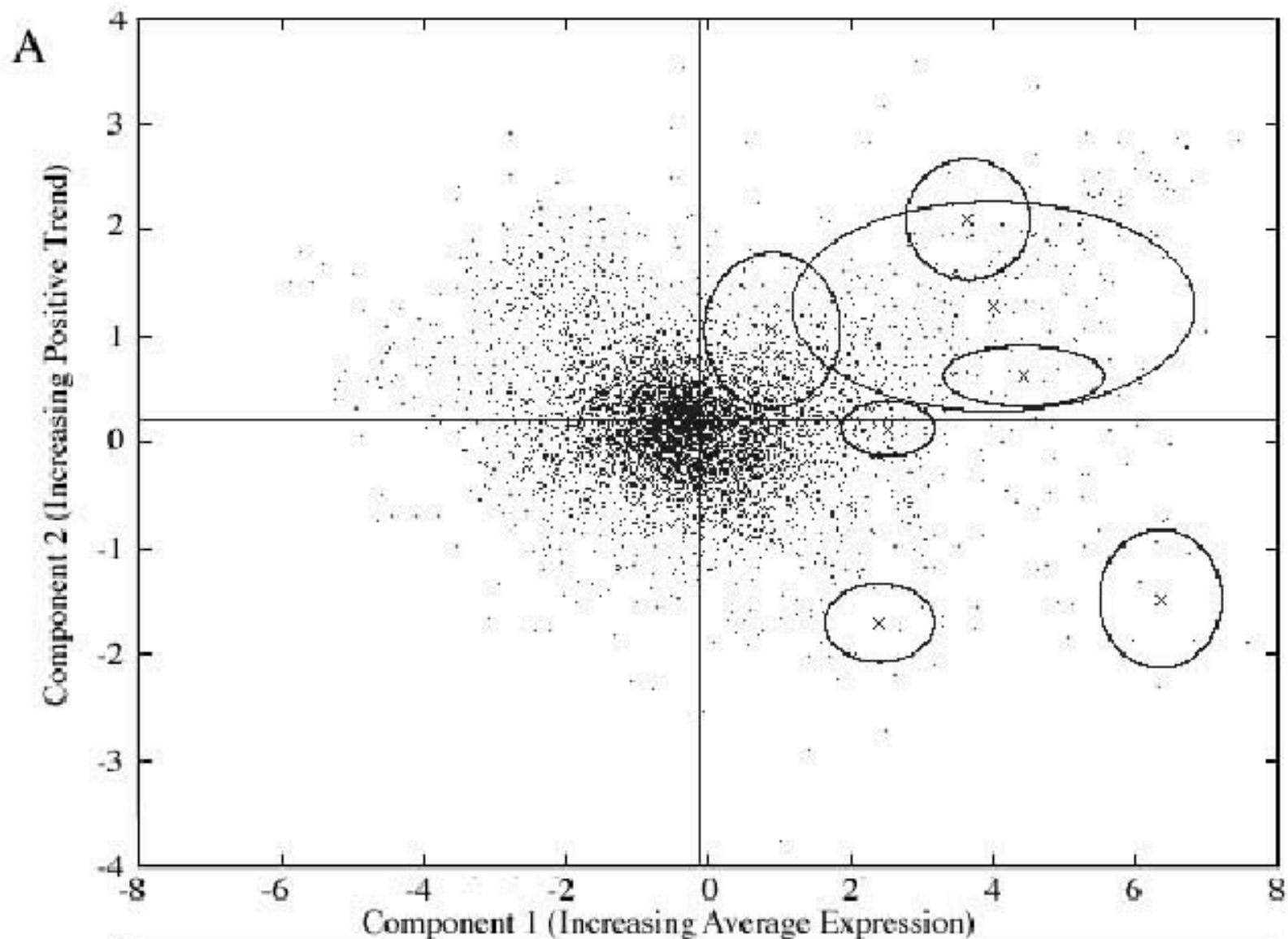


Figure 4. A. All genes plotted with respect to first and second principal components. Ellipses represent clusters identified in the original publication of the sporulation data. Ellipses are drawn to include 68% of the genes in the cluster. B. Ellipses are labelled using labels reported by the original investigators (Chu et al. 1998) and drawn to include 95% of genes in the cluster.

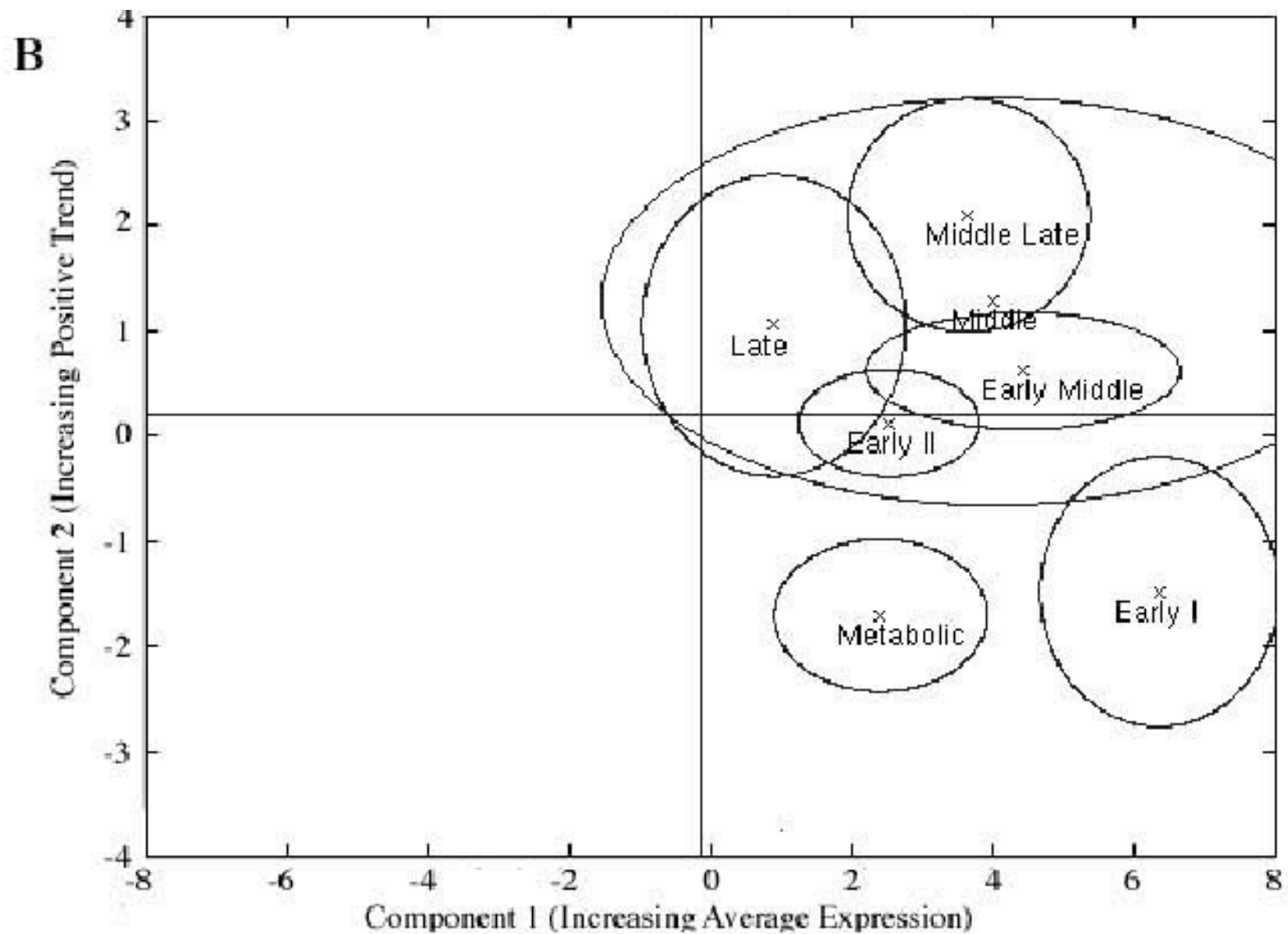
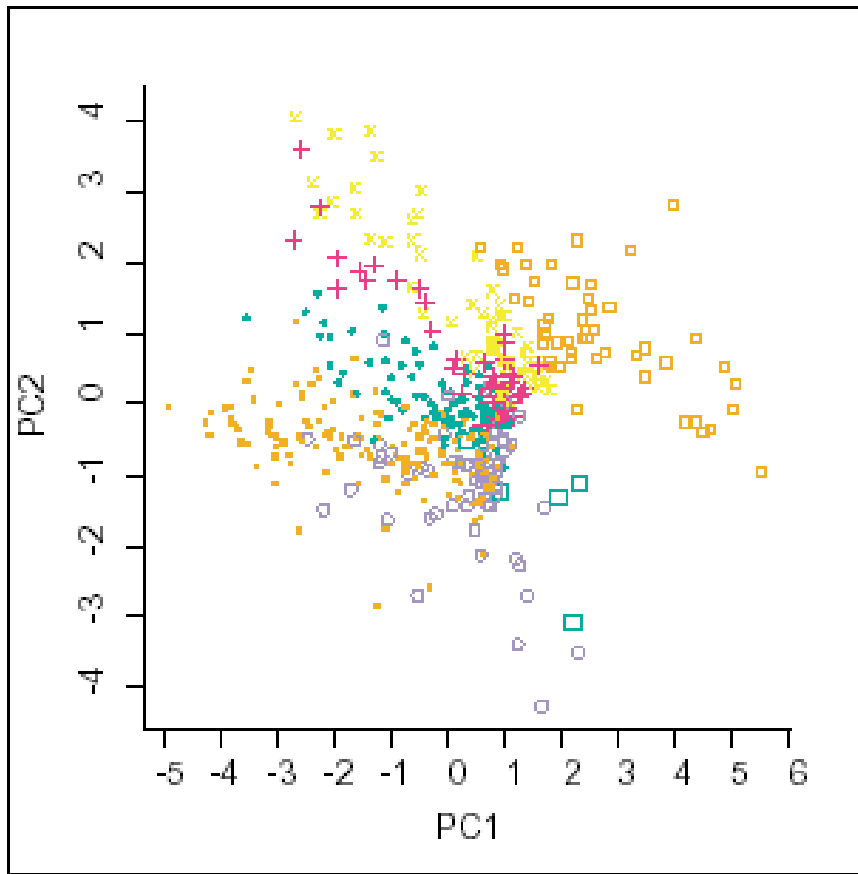
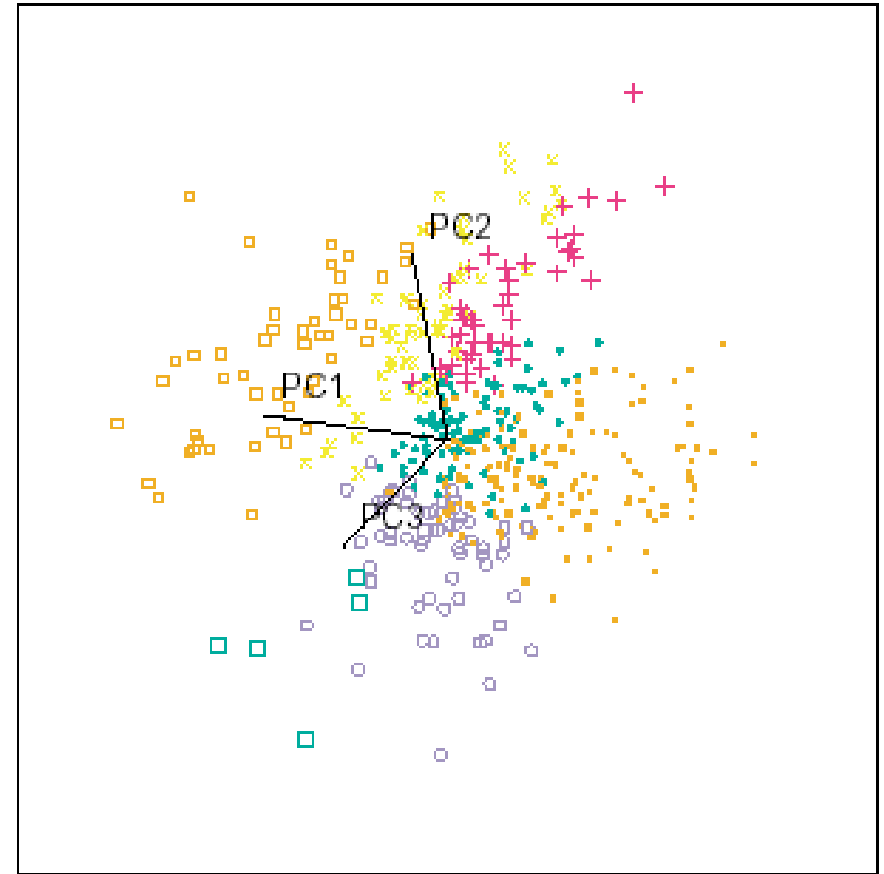


Figure 4. A. All genes plotted with respect to first and second principal components. Ellipses represent clusters identified in the original publication of the sporulation data. Ellipses are drawn to include 68% of the genes in the cluster. B. Ellipses are labelled using labels reported by the original investigators (Chu et al. 1998) and drawn to include 95% of genes in the cluster.



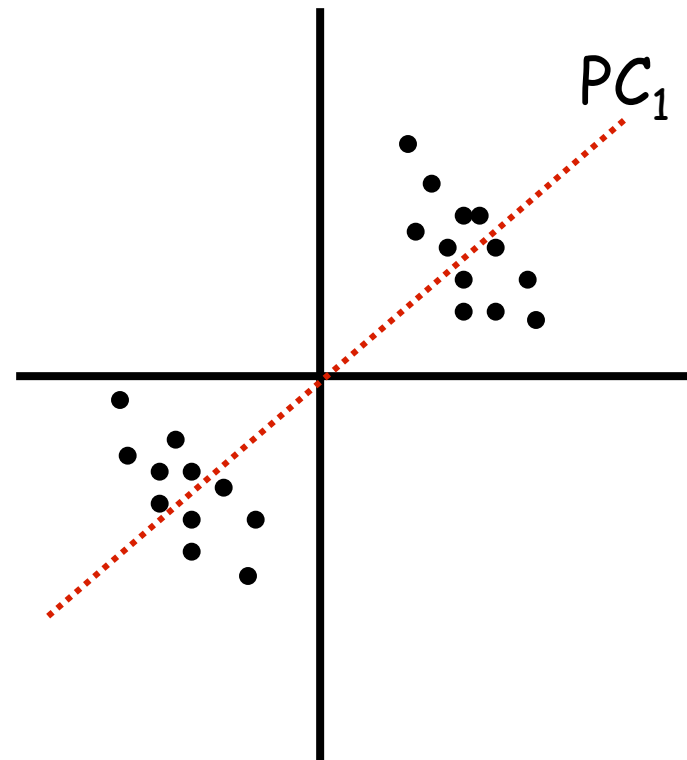
(a) In the subspace of the first 2 PC's



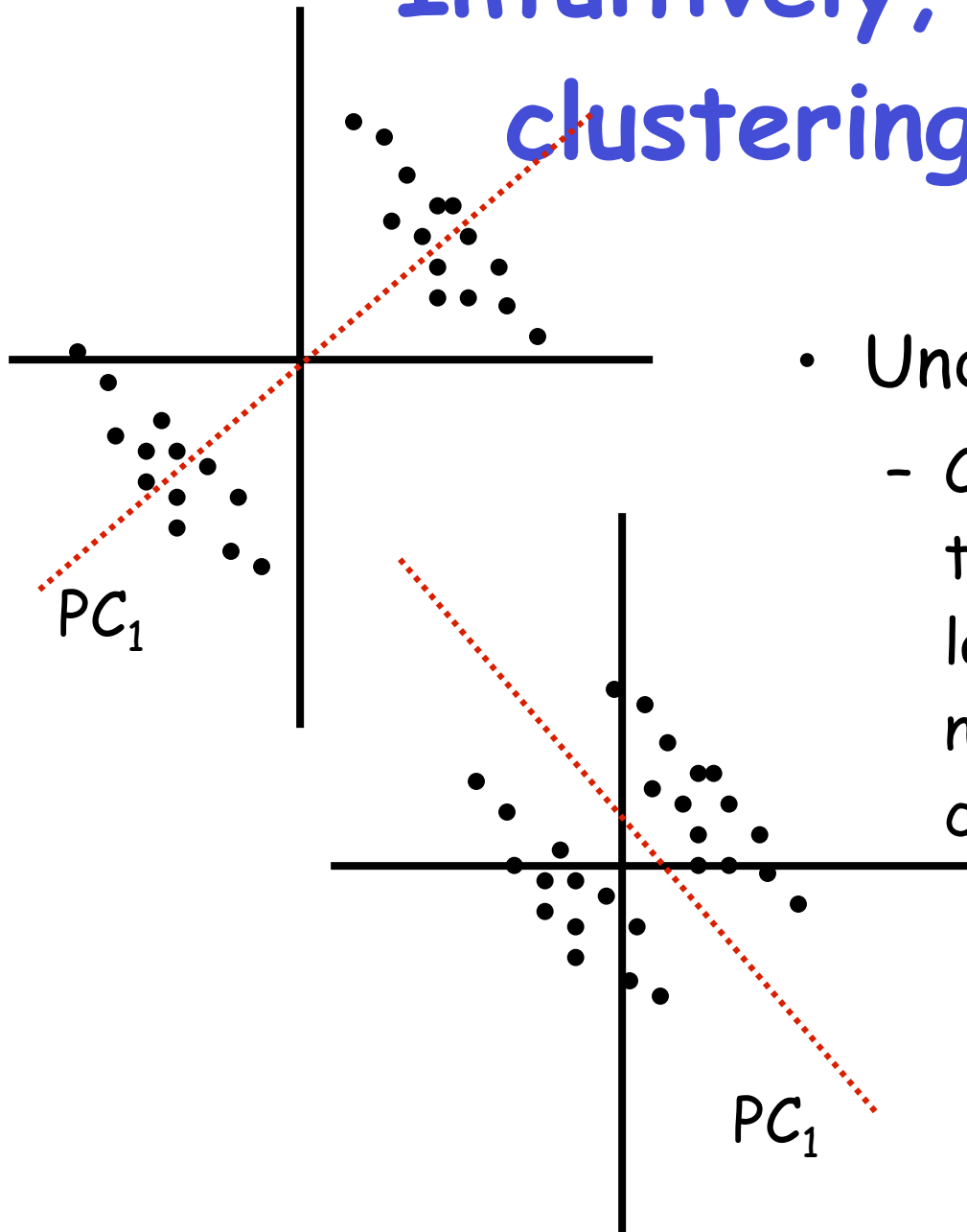
(b) In the subspace of the first 3 PC's

PCA and clustering

- Euclidean distance:
 - using all p variables, Euclidean distance between a pair of genes unchanged after PCA [Jolliffe 1986]
 - using m variables ($m < p$) \implies approximation
- Correlation coefficient
 - no general relationship before and after PCA

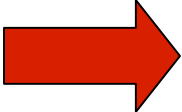


Intuitively, PCA helps clustering. But...



- Under some assumptions,
 - Chang[1983] showed that the set of PC's with the largest eigenvalues does not necessarily capture cluster structure info

Outline of talk

- Background and motivation
-  • Design of our empirical study
- Results
- Summary and Conclusions

Our empirical study

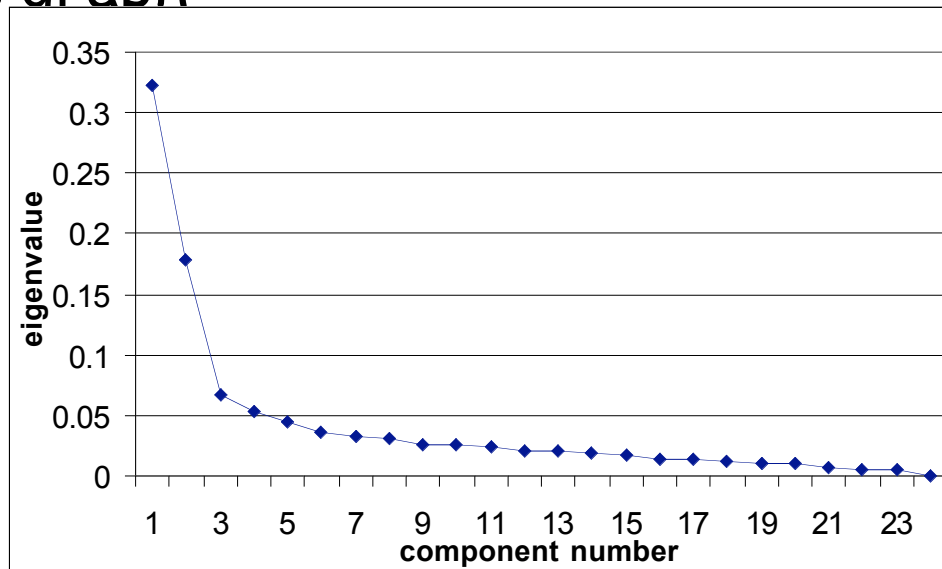
- Goal: Compare the clustering results with and without PCA to an external criterion:
 - expression data set with external criterion
 - synthetic data sets
 - methodology to compare to an external criterion
 - clustering algorithms
 - similarity metrics

Ovary data (Michel Schummer)

- Randomly selected cDNA's on membrane arrays
- Subset of data:
 - 235 clones
 - 24 experiments (7 from normal tissues, 4 from blood samples, 13 from ovarian cancers)
- 235 clones correspond to 4 genes (sizes 58, 88, 57, 32)
- The four genes form the 4 classes (external criterion)

PCA on ovary data

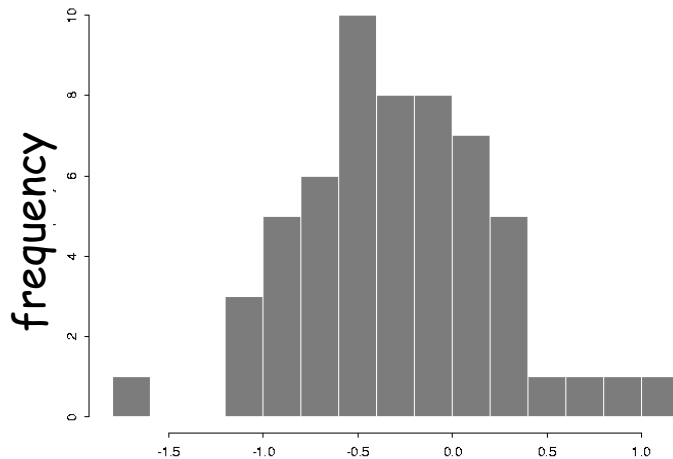
- Number of PC's to adequately represent the data:
 - 14 PC's cover 90% of the variation
 - scree graph



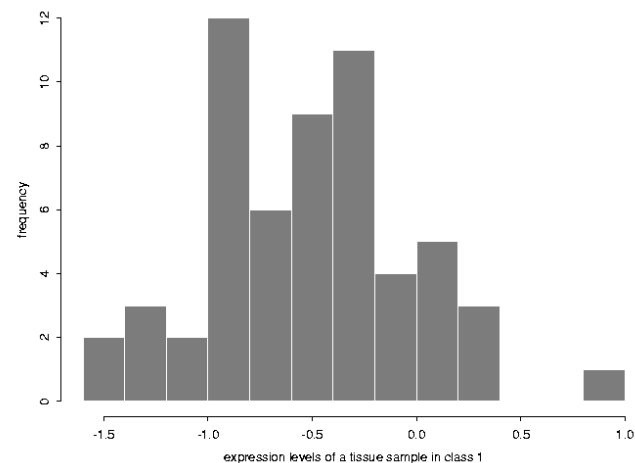
Synthetic data sets (1)

- Mixture of normal distributions
 - Compute the mean vector and covariance matrix for each class in the ovary data
 - Generate a random mixture of normal distributions using the mean vectors, covariance matrices, and size distributions from the ovary data

Histogram of a normal class



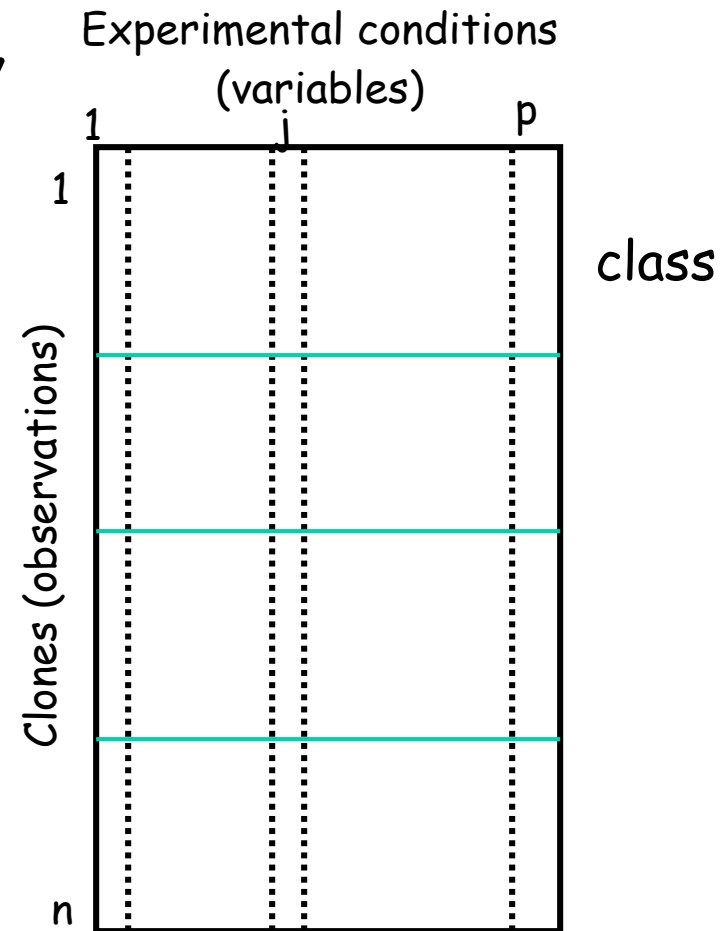
Histogram of a tumor class



Expression level

Synthetic data sets (2)

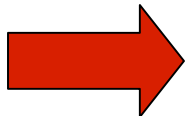
- Randomly permuted ovary data
 - Random sample (with replacement) the expression levels in the same class
 - Empirical distribution preserved
 - But covariance matrix not preserved



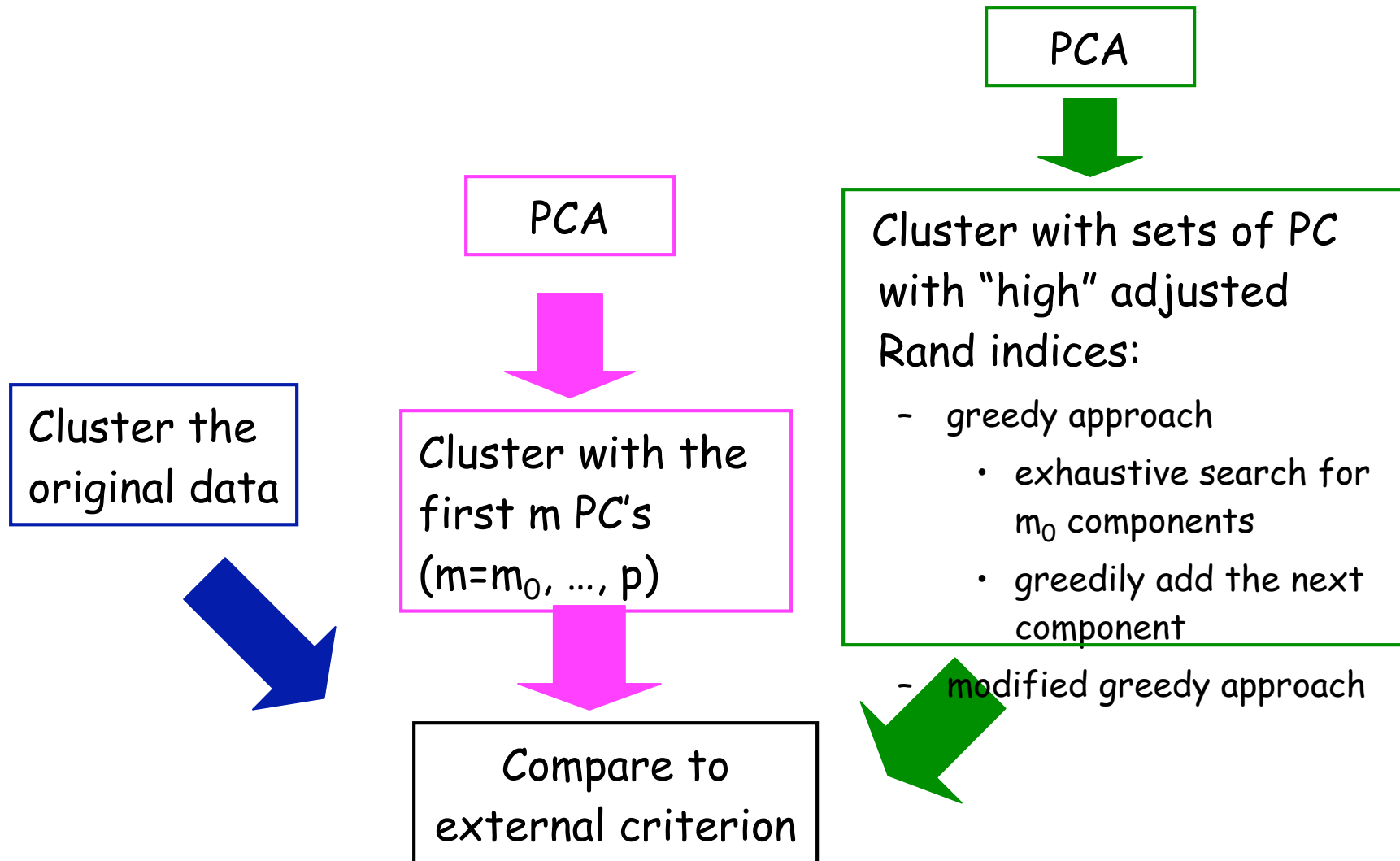
Clustering algorithms and similarity metrics

- **CAST** [Ben-Dor and Yakhini 1999] with correlation
 - build one cluster at a time
 - add or remove genes from clusters based on similarity to the genes in the current cluster
- **k-means** with correlation and Euclidean distance
 - initialized with hierarchical average-link

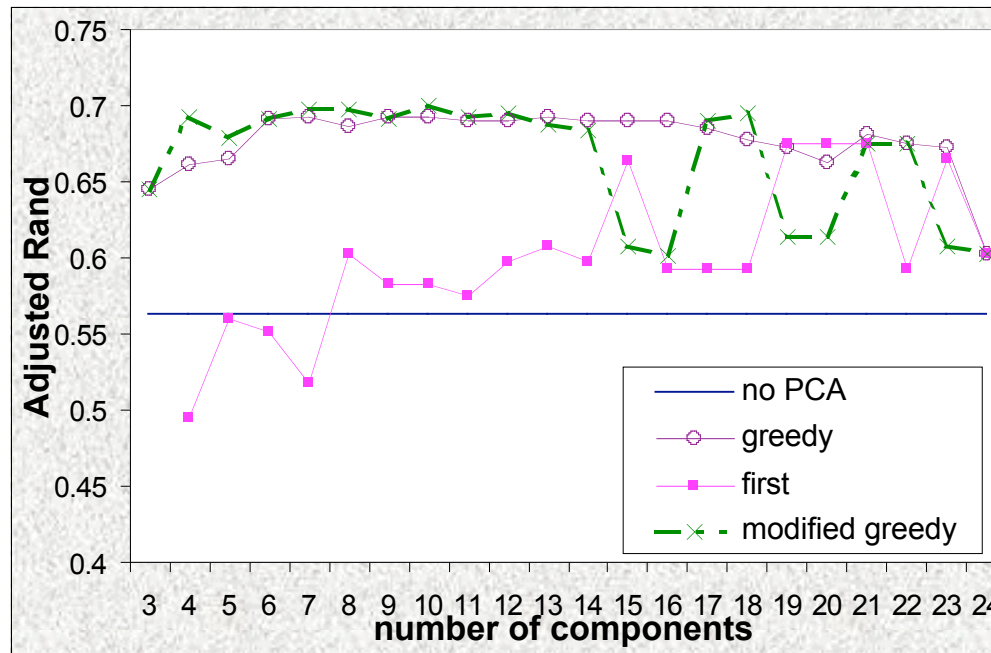
Outline of talk

- Background and motivation
- Design of our empirical study
-  • Results
- Summary and Conclusions

Our approach



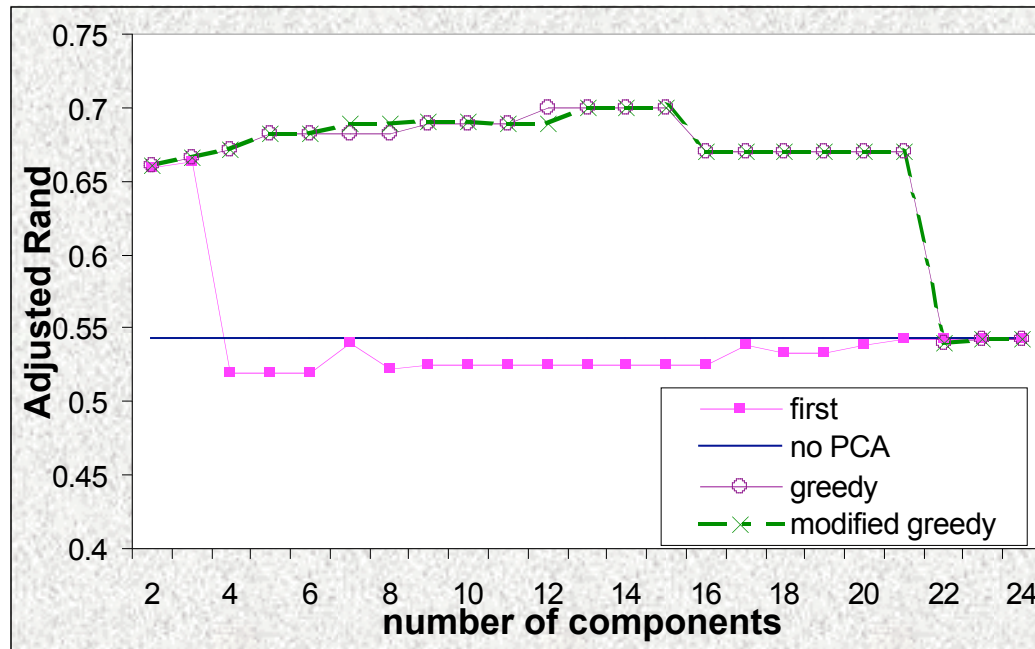
Results: ovary data k-means with correlation



- Adjusted Rand index for the first m ($m \geq 7$) PC's higher than without PCA
- adjusted Rand index with all 24 PC's higher than without PCA

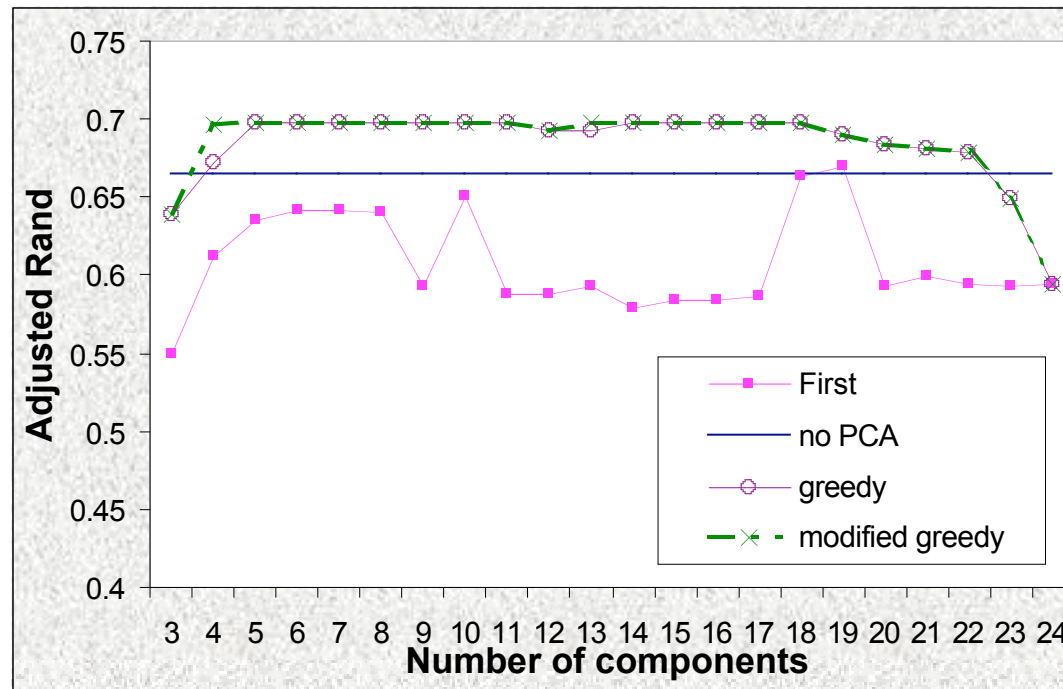
Results: ovary data

k-means with Euclidean distance



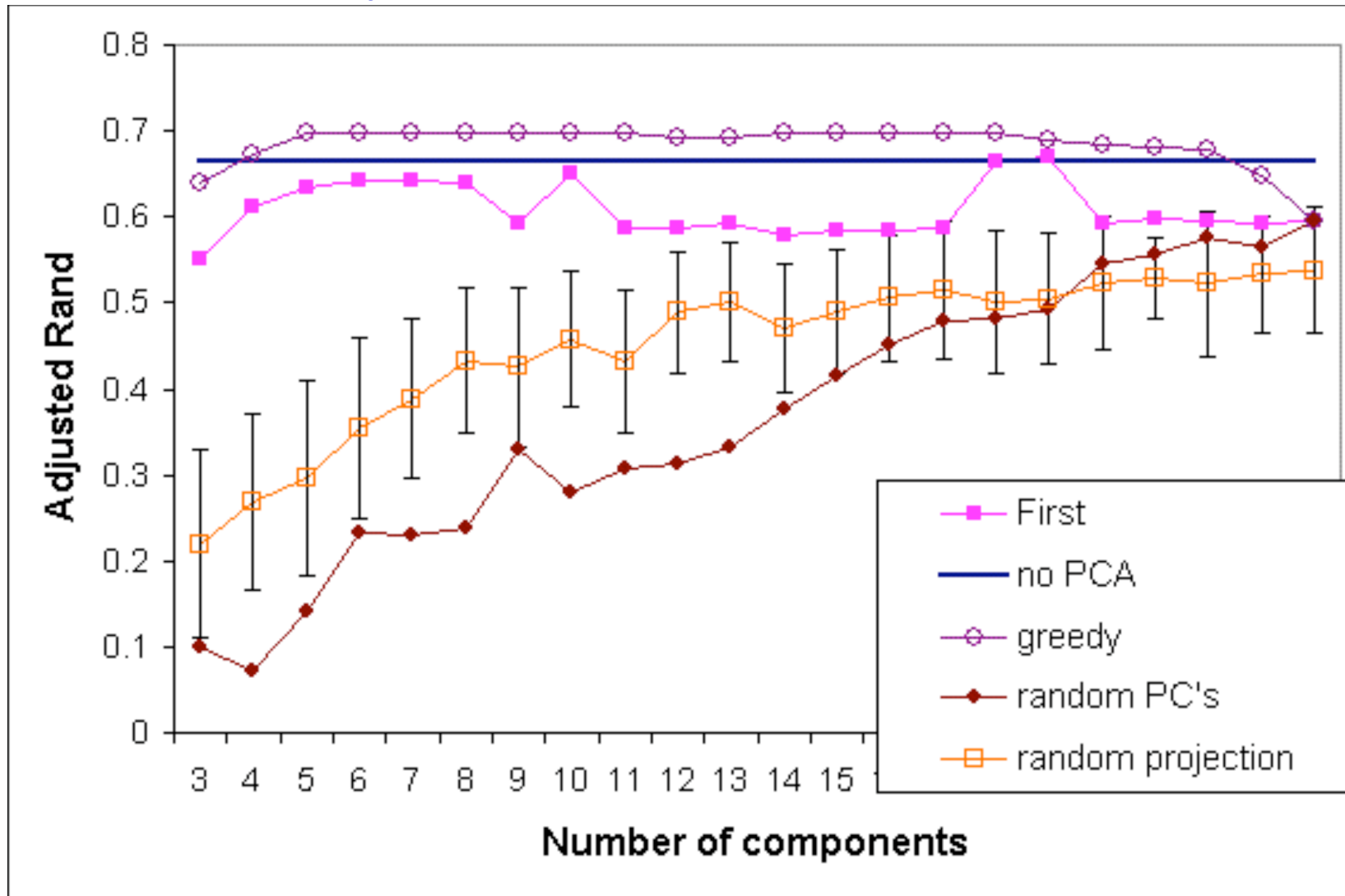
- Sharp drop of adjusted Rand index from the first 3 to first 4 PC's

Results: ovary data CAST with correlation



- Adjusted Rand index on the first components \leq without PCA
- greedy or modified greedy approach usually achieve higher adjusted Rand than without PCA

Ovary/Cast, Correlation



Real data: did PCA help?

- 2 data sets, 5 algorithms
- “+” means clustering using the first c PC's helped (for some c)

	CAST	k-Means correlation	k-Means distance	Ave-link correlation	Ave-link distance
Ovary	-	+	-	+	-
Cell cycle	-	-	-	-	-

Synth data: Did PCA help?

p-values from Wilcoxon signed rank test. (p<5% are bold)

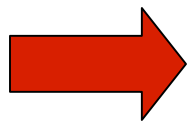
Synthetic data	Alternative hypothesis	CAST corr.	k-mean Corr	k-mean dist	ave-link corr	ave-link dist
Mixture of normal	no PCA > first	0.039	0.995	0.268	0.929	0.609
Mixture of normal	no PCA < first	0.969	0.031	0.760	0.080	0.418
Random resampled	no PCA > first	0.243	0.909	0.824	0.955	0.684
Random resampled	no PCA < first	0.781	0.103	0.200	0.049	0.337
Cyclic data	no PCA > first	0.023	NA	0.296	0.053	0.799
Cyclic data	no PCA < first	0.983	NA	0.732	0.956	0.220

Some successes

- Alter, O., Brown, P.O. and Botstein, D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, 97, 10101_10106.
- Holter, N.S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J.R. and Fedoroff, N.V. (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl Acad. Sci. USA*, 97, 8409_8414.
- Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* 2000;1(2):RESEARCH0003.

Outline of talk

- Background and motivation
- Design of our empirical study
- Results
- Summary and Conclusions



Summary & Conclusions (1)

- PCA may not improve cluster quality
 - first PC's may be worse than without PCA
 - another set of PC's may be better than first PC's
- Effect of PCA depends on clustering algorithms and similarity metrics
 - CAST with correlation: first m PC's usually worse than without PCA
 - k-means with correlation: usually PCA helps
 - k-means with Euclidean distance: worse after the first few PC's

Summary & Conclusions (2)

- No general trends in the components chosen by the greedy or modified greedy approach
 - usually the first 2 components are chosen by the exhaustive search step
- Results on the synthetic data similar to real data

Bottom Line

- Successes by other groups make it a technique worth considering, but it should not be applied blindly.

Acknowledgements

- Ka Yee Yeung
- Michèle Schummer

More Info

<http://www.cs.washington.edu/homes/{kayee,ruzzo}>



Mathematical definition of PCA

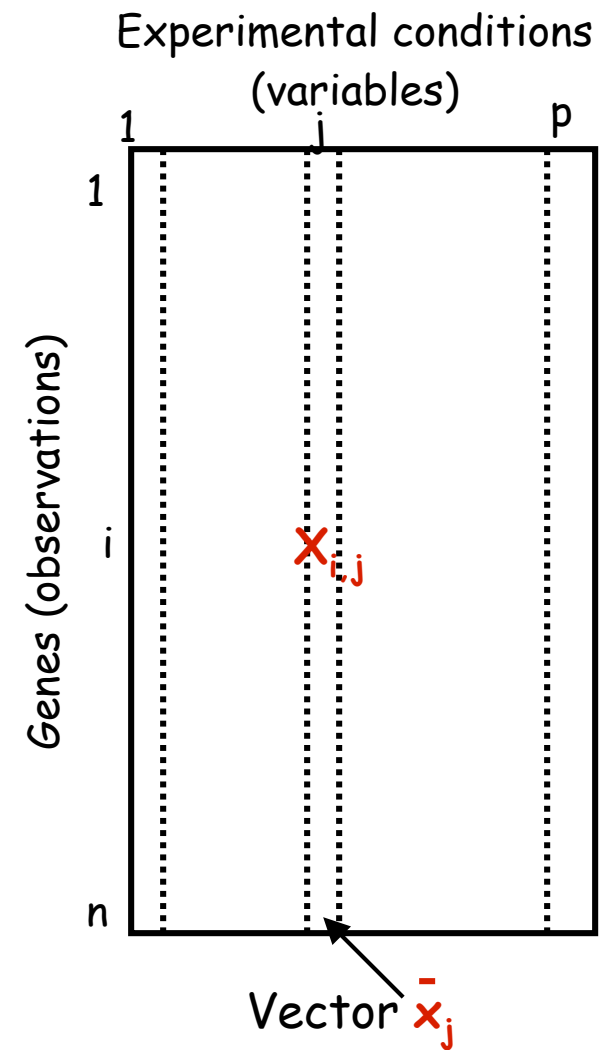
- The k-th PC:
$$z_k = \sum_{j=1}^p \alpha_{k,j} \bar{x}_j$$

- First PC : maximize

$\text{var}(z_1) = \alpha_1^T \Sigma \alpha_1$, such that $\alpha_1^T \alpha_1 = 1$, where Σ is the covariance matrix

- k-th PC: maximize

$\text{var}(z_k) = \alpha_k^T \Sigma \alpha_k$, such that $\alpha_k^T \alpha_k = 1$ and $\alpha_k^T \alpha_i = 0$, where $i < k$



More details on PCA

- It can be shown that \mathbf{v}_k is an eigenvector of \mathbf{C} corresponding to the k -th largest eigenvalue λ_k
- $\text{var}(z_k) = \lambda_k$
- Use sample covariance matrix:

$$S(j, k) = \frac{\sum_{i=1}^n (x_{i,j} - \bar{x}_j)(x_{i,k} - \bar{x}_k)}{n - 1}, \text{ where } \bar{x}_j = \frac{\sum_{i=1}^n x_{i,j}}{n}$$