Model-based clustering and data transformations of gene expression data

Walter L. Ruzzo

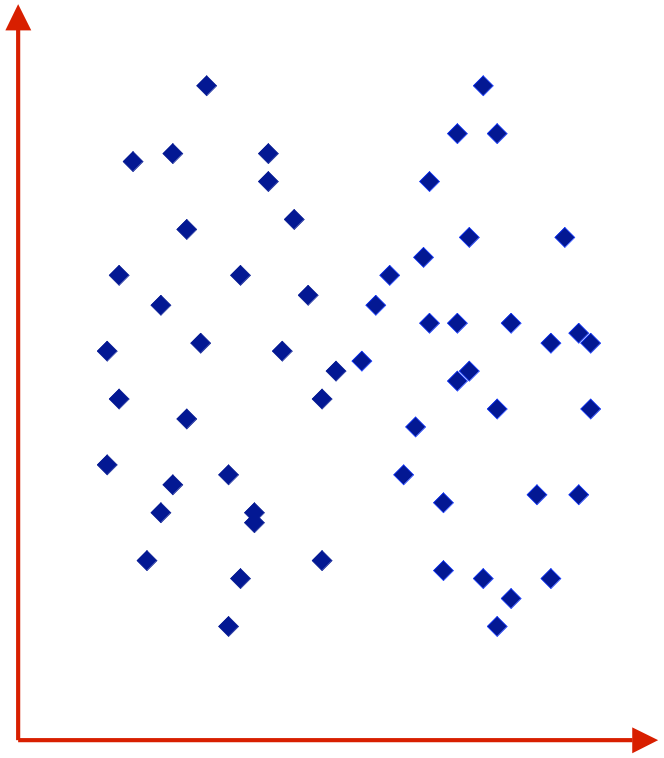
University of Washington

UW CSE Computational Biology Group

Overview

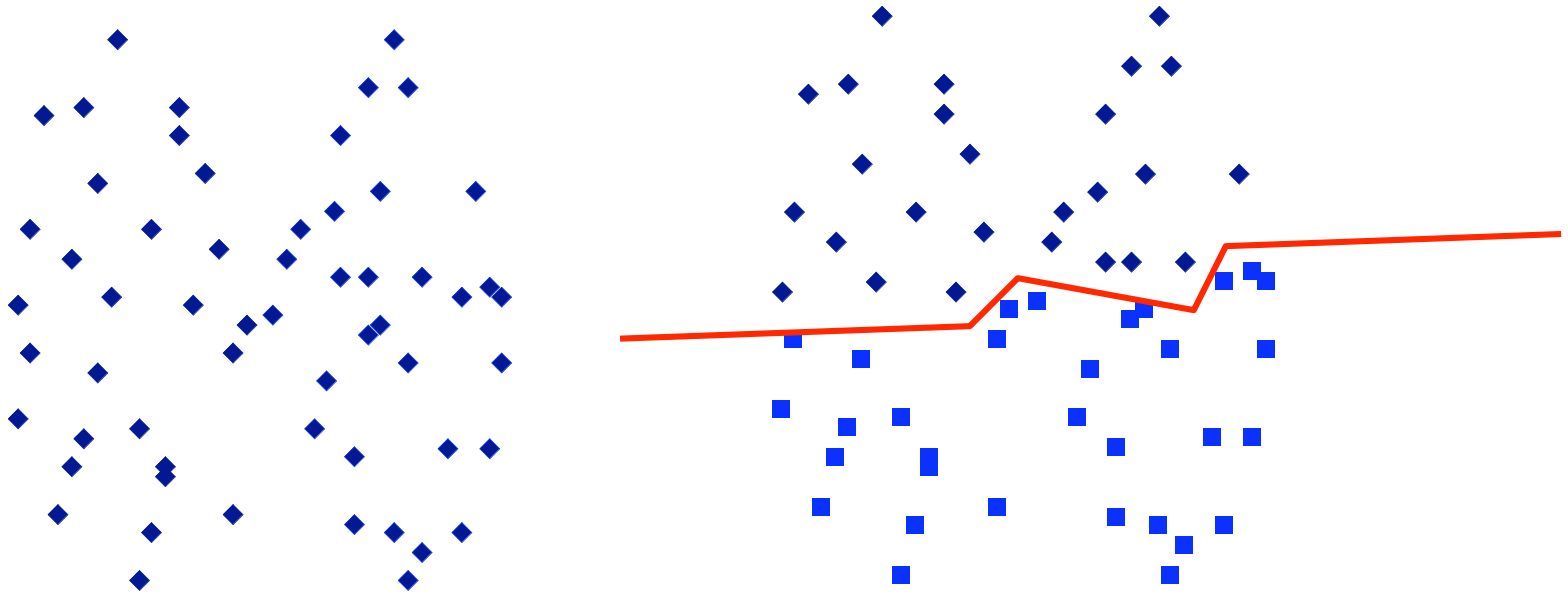
- Motivation
- Model-based clustering
- Validation
- Summary and Conclusions

Toy 2-d Clustering Example

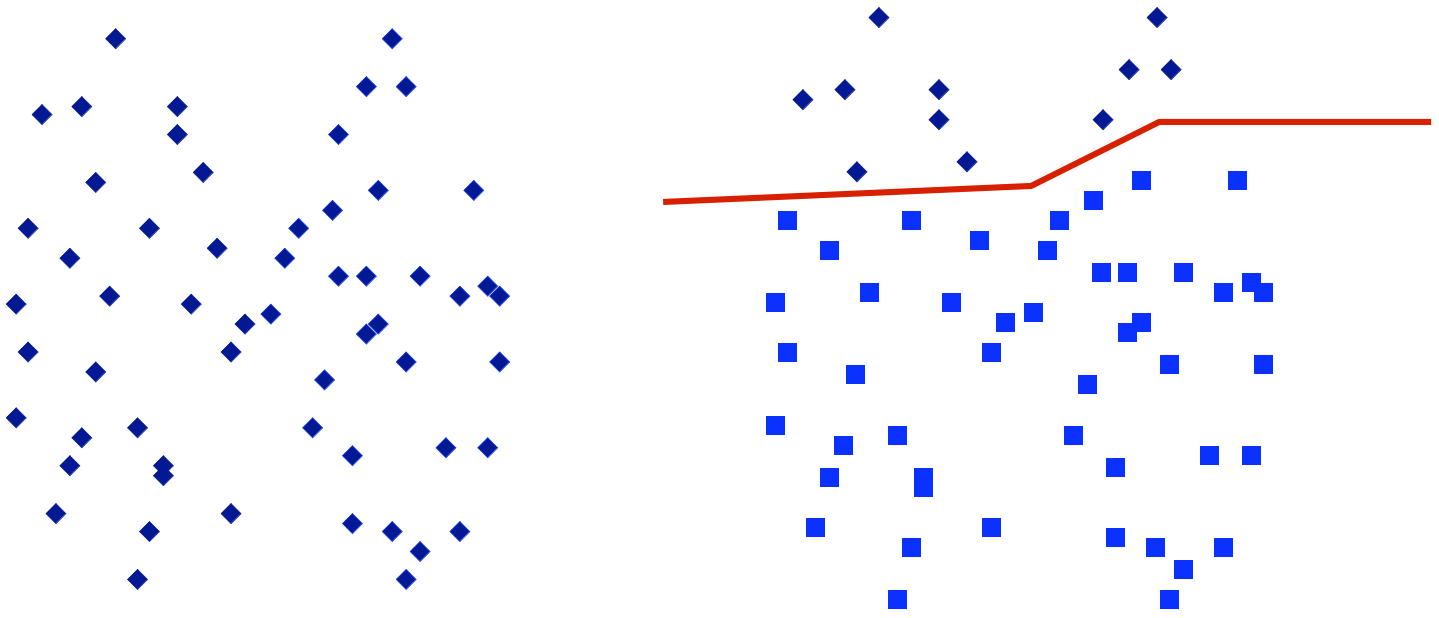


?

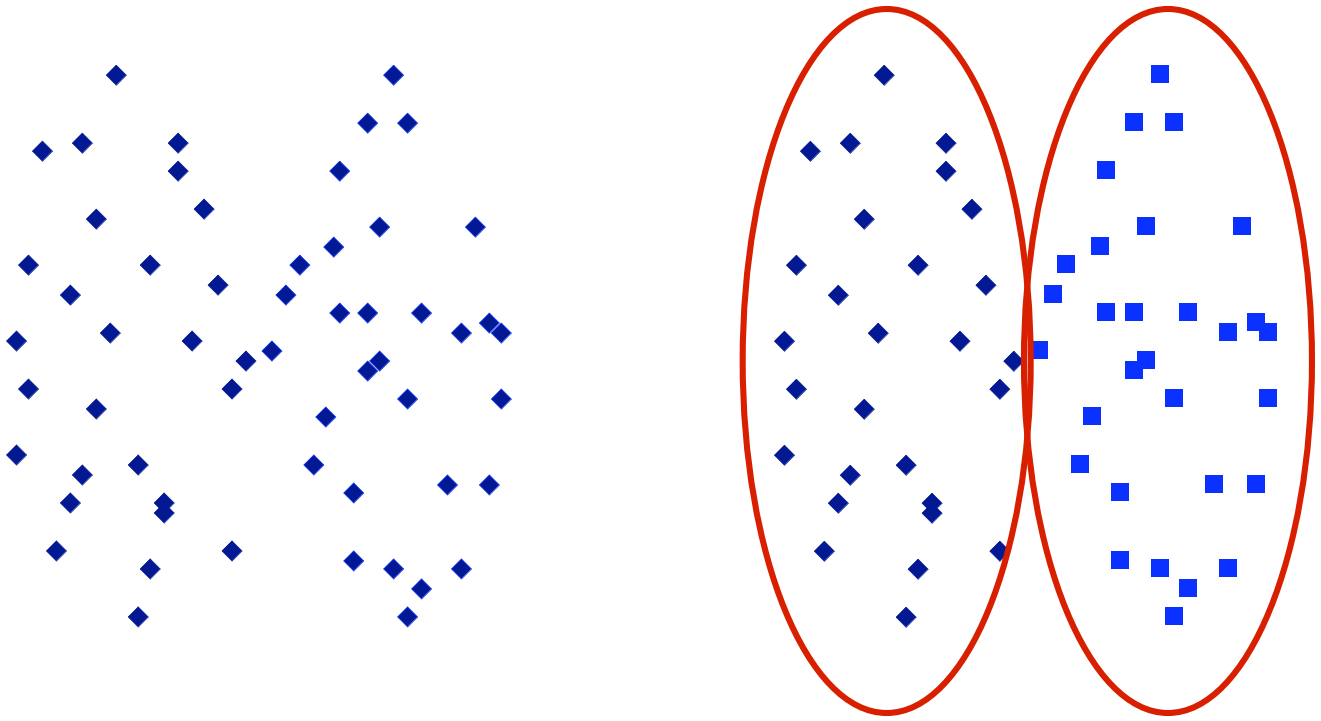
K-Means



Hierarchical Average Link



Model-Based (If You Want)

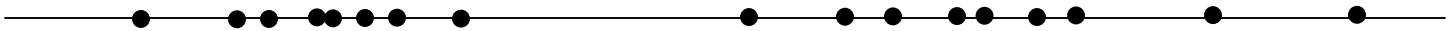


Overview

- Motivation
- Model-based clustering
- Validation
- Summary and Conclusions

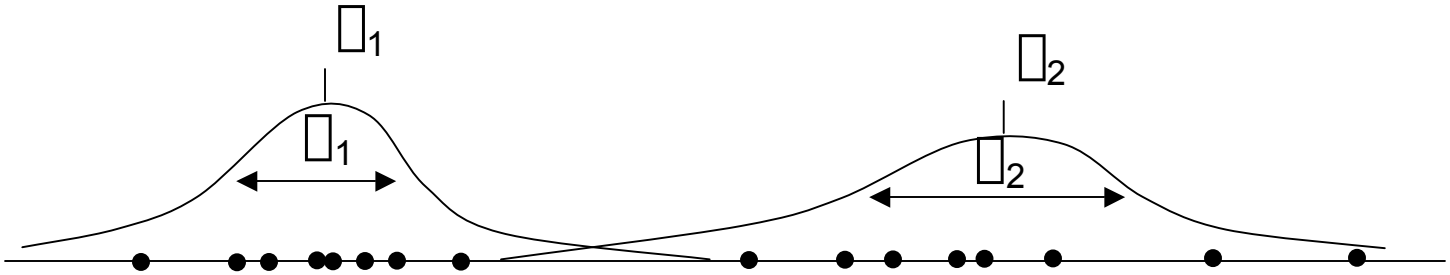
Model-based clustering

- Gaussian mixture model:
  - Assume each cluster is generated by a multivariate normal distribution
  - Cluster k has parameters :
    - Mean vector: μ_k
    - Covariance matrix: Σ_k



Model-based clustering

- Gaussian mixture model:
  - Assume each cluster is generated by a multivariate normal distribution
  - Cluster k has parameters :
    - Mean vector: μ_k
    - Covariance matrix: Σ_k



Variance & Covariance

- Variance

$$\mathrm{var}(x) = E((x - \bar{x})^2)$$

- Covariance

$$\mathrm{cov}(x, y) = E((x - \bar{x})(y - \bar{y}))$$

- Correlation

$$\mathrm{cor}(x, y) = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

Gaussian Distributions

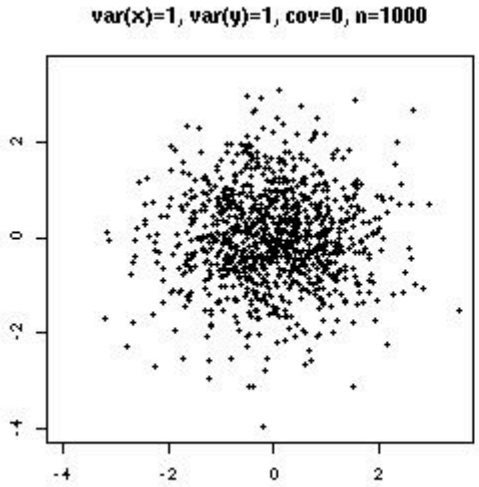
- Univariate $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(x-\bar{x})^2/\sigma^2}$

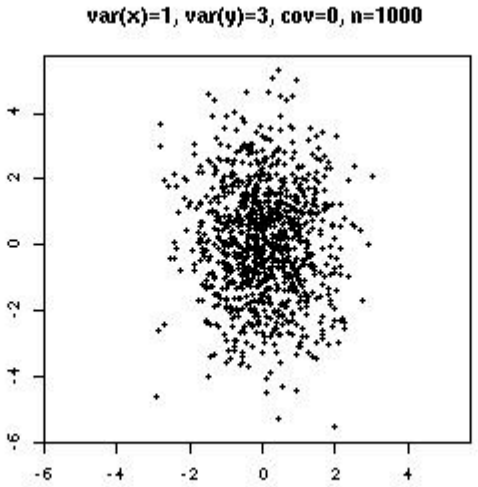
- Multivariate $\frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-\bar{x})^T(\Sigma^{-1})(x-\bar{x})}$

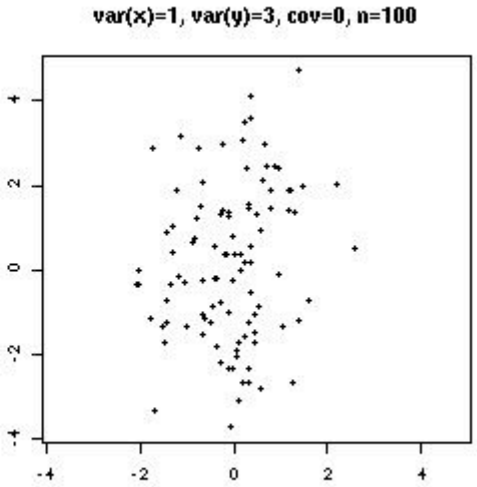
  where Σ is the variance/covariance matrix:

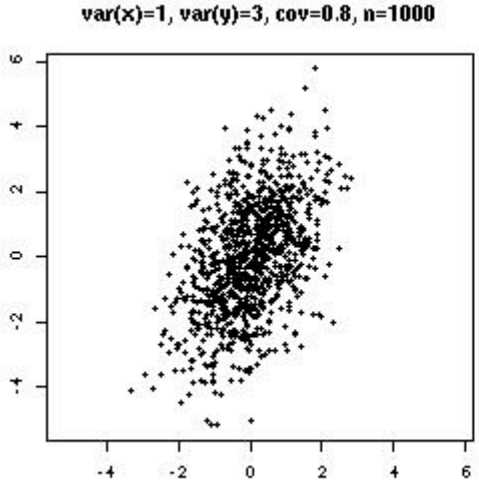
$$\Sigma_{i,j} = E((x_i - \bar{x}_i)(x_j - \bar{x}_j))$$

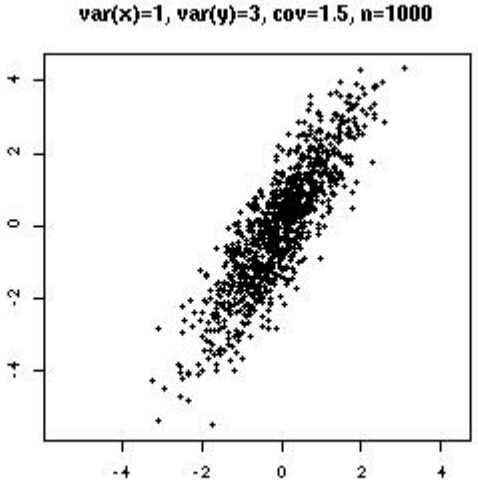
Variance/Covariance

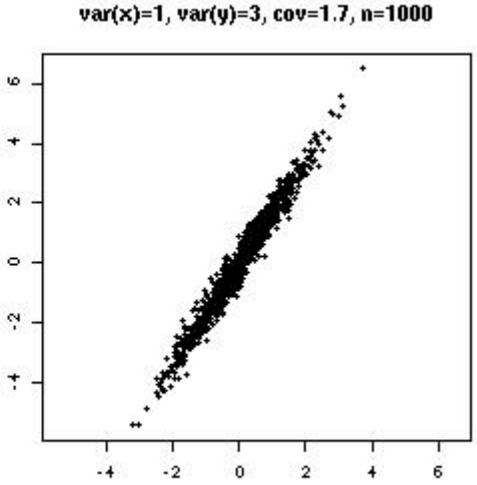












Covariance models

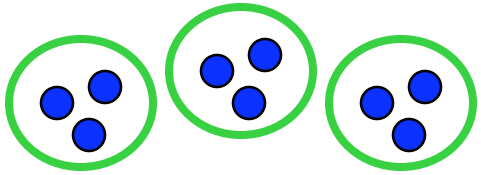
(Banfield & Raftery 1993)

$$\Sigma_k = \lambda_k D_k A_k D_k^T$$

volume shape orientation

- Equal volume spherical model (EI): ~ kmeans

  $\Sigma_k = \lambda I$



Covariance models

(Banfield & Raftery 1993)

$$\Sigma_k = \lambda_k D_k A_k D_k^T$$

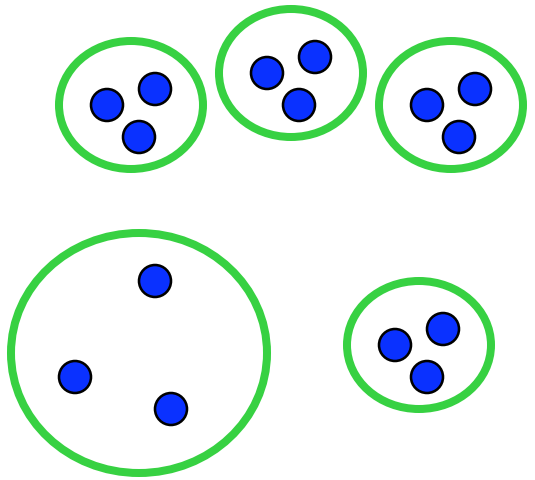
volume shape orientation

- Equal volume spherical model (EI): ~ kmeans

  $\Sigma_k = \lambda I$

- Unequal volume spherical (VI):

  $\Sigma_k = \lambda_k I$



Covariance models

(Banfield & Raftery 1993)

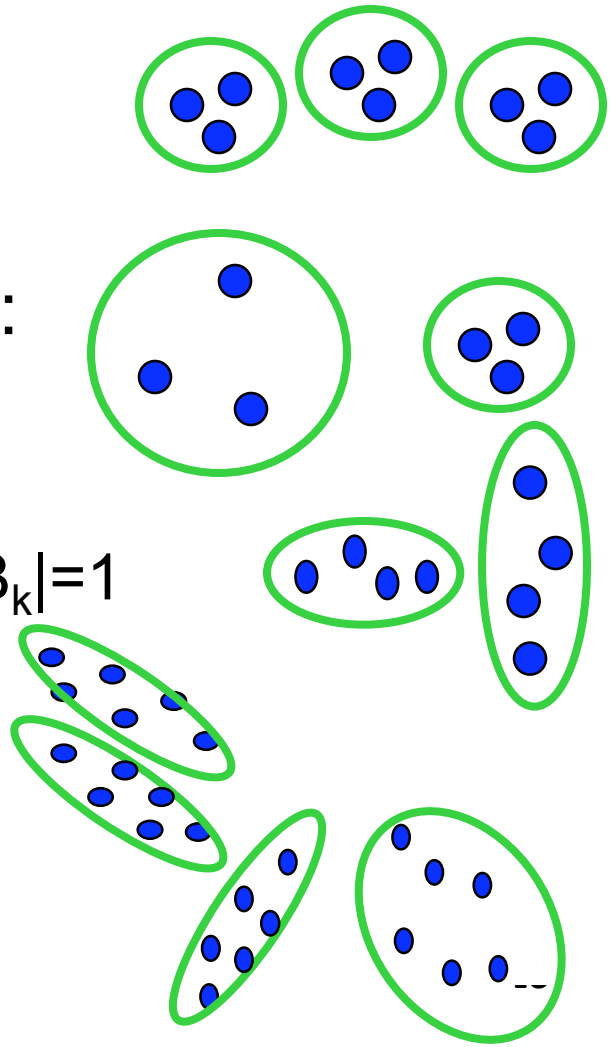
$$\Sigma_k = \lambda_k D_k A_k D_k^T$$

volume shape orientation

- Equal volume spherical model (EI): ~ kmeans
  $\Sigma_k = \lambda I$
- Unequal volume spherical (VI):
  $\Sigma_k = \lambda_k I$
- Diagonal model:
  $\Sigma_k = \lambda_k B_k$, where B_k is diagonal, $|B_k|=1$
- EEE elliptical model:
  $\Sigma_k = \lambda D A D^T$
- Unconstrained model (VVV):
  $\Sigma_k = \lambda_k D_k A_k D_k^T$

More flexible

But more parameters



EM algorithm

- General approach to maximum likelihood
- Iterate between E and M steps:
  - E step: compute the probability of each observation belonging to each cluster using the current parameter estimates
  - M-step: estimate model parameters using the current group membership probabilities

Advantages of model-based clustering

- Higher quality clusters
- Flexible models
- Model selection – **A principled way to choose right model and right # of clusters**
  - Bayesian Information Criterion (BIC):
    - Approximate Bayes factor: posterior odds for one model against another model
    - Roughly: data likelihood, penalized for number of parameters
  - A large BIC score indicates strong evidence for the corresponding model.

Definition of the BIC score

$$2\log p(D \mid M_k) \approx 2\log p(D \mid \hat{\theta}_k, M_k) - \nu_k \log(n) = BIC_k$$

- The integrated likelihood $p(D|M_k)$ is hard to evaluate,

  where D is the data, M_k is the model.
- BIC is an approximation to log $p(D|M_k)$
- ν_k: number of parameters to be estimated in model M_k

Overview

- Motivation
- Model-based clustering
- Validation
  - Methodology
  - Data Sets
  - Results
- Summary and Conclusions

Validation Methodology

- Compare on data sets with *external criteria*
  (BIC scores do **not** require the external criteria)
- To compare clusters with external criterion:
  - Adjusted Rand index (Hubert and Arabie 1985)
  - Adjusted Rand index = 1 → perfect agreement
  - 2 random partitions have an expected index of 0
- Compare quality of clusters to those from:
  - a leading heuristic-based algorithm: CAST (Ben-Dor & Yakhini 1999)
  - k-Means (EI).

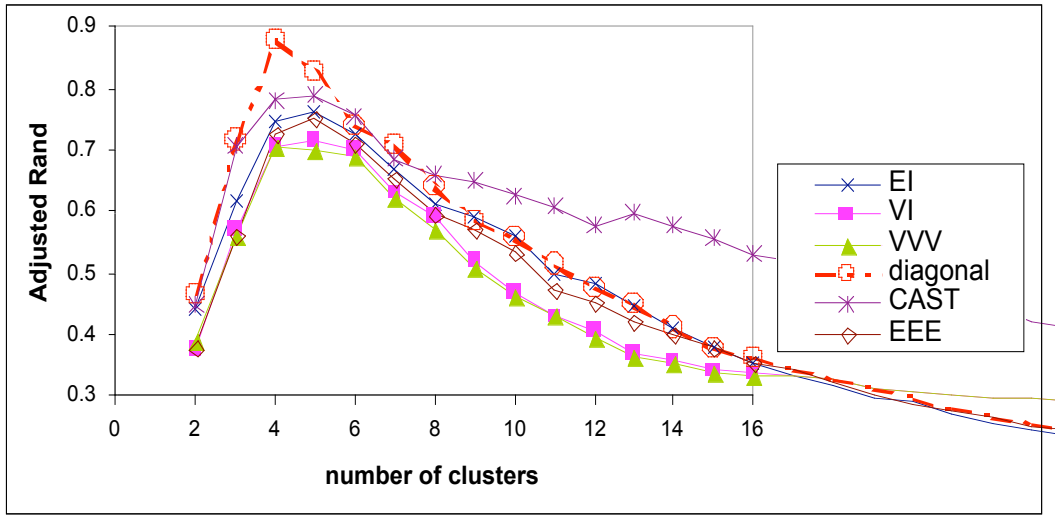
Gene expression data sets

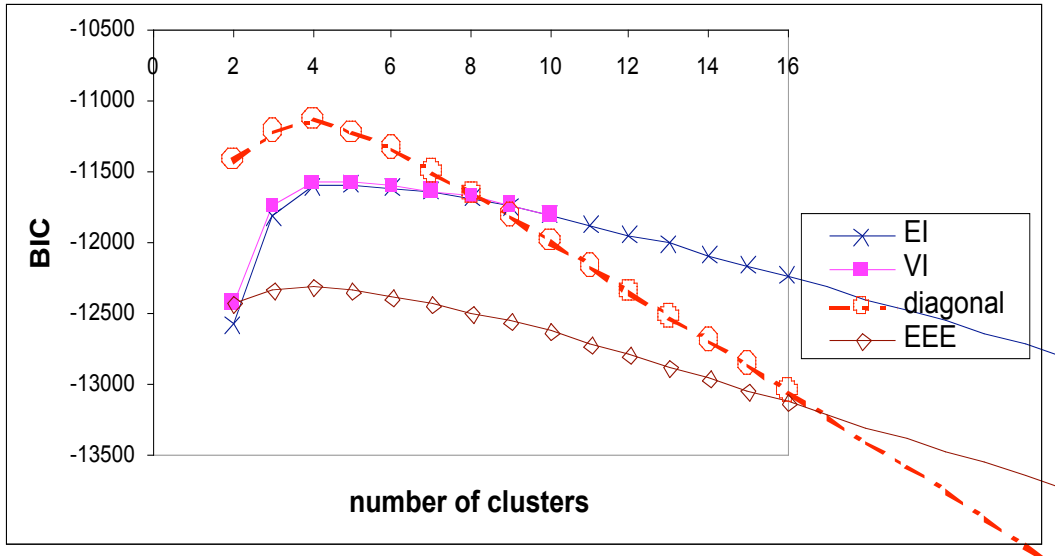
- Ovarian cancer data set
  (Michel Schummer, Institute of Systems Biology)
  - Subset of data: 235 clones
    24 experiments (cancer/normal tissue samples)
  - 235 clones correspond to 4 genes
- Yeast cell cycle data (Cho *et al* 1998)
  - 17 time points
  - Subset of 384 genes associated with 5 phases of cell cycle

Synthetic data sets

Both based on ovary data

- Randomly resampled ovary data
  - For each class, randomly sample the expression levels in each experiment, independently
  - Near diagonal covariance matrix
- Gaussian mixture
  - Generate multivariate normal distributions with the sample covariance matrix and mean vector of each class in the ovary data

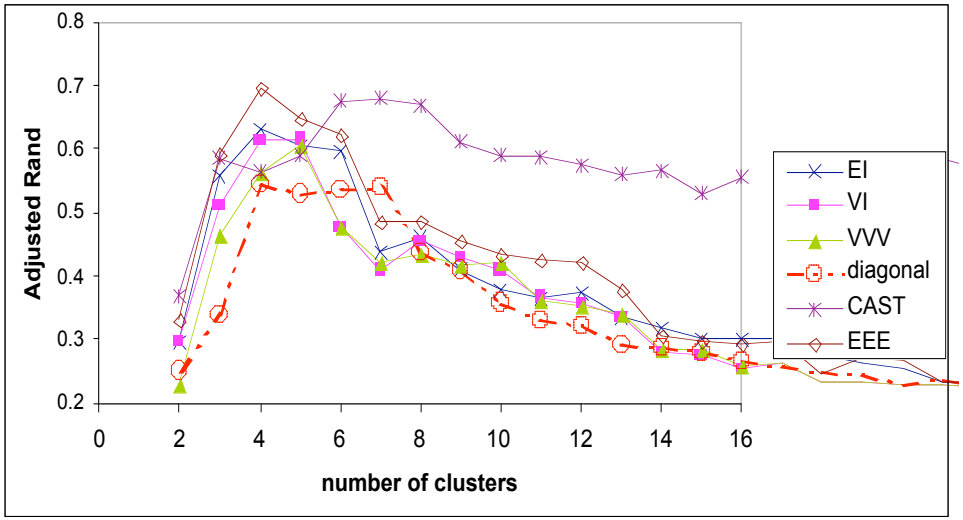


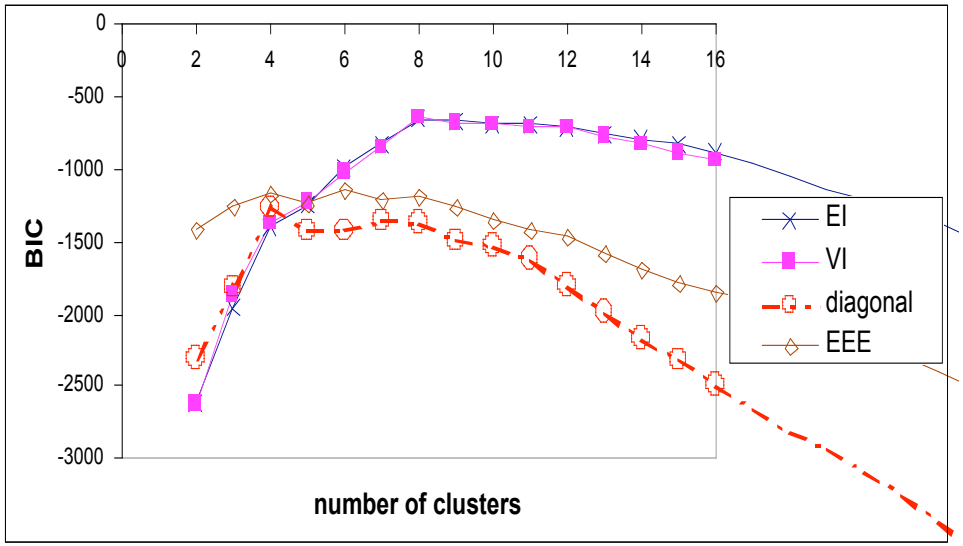


Results: randomly resampled ovary data

- Diagonal model achieves max BIC score (~expected)
- max BIC at 4 clusters (~expected)
- max adjusted Rand
- beats CAST

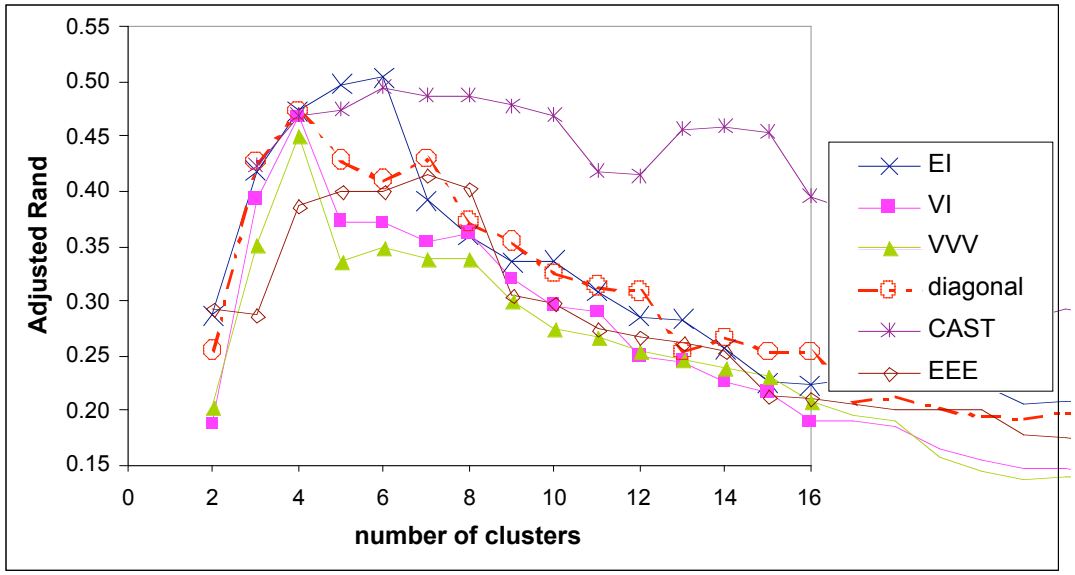
Results: square root ovary data

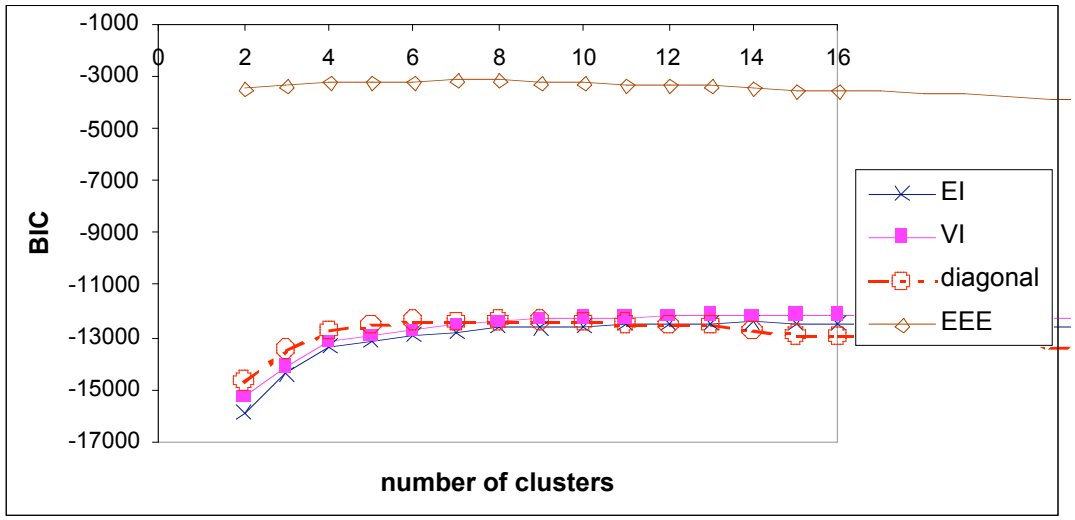




- Adjusted Rand: max at EEE 4 clusters (> CAST)
- BIC analysis:
  - EEE and diagonal models → local max at 4 clusters
  - Global max → VI at 8 clusters (8 ≈ split of 4).

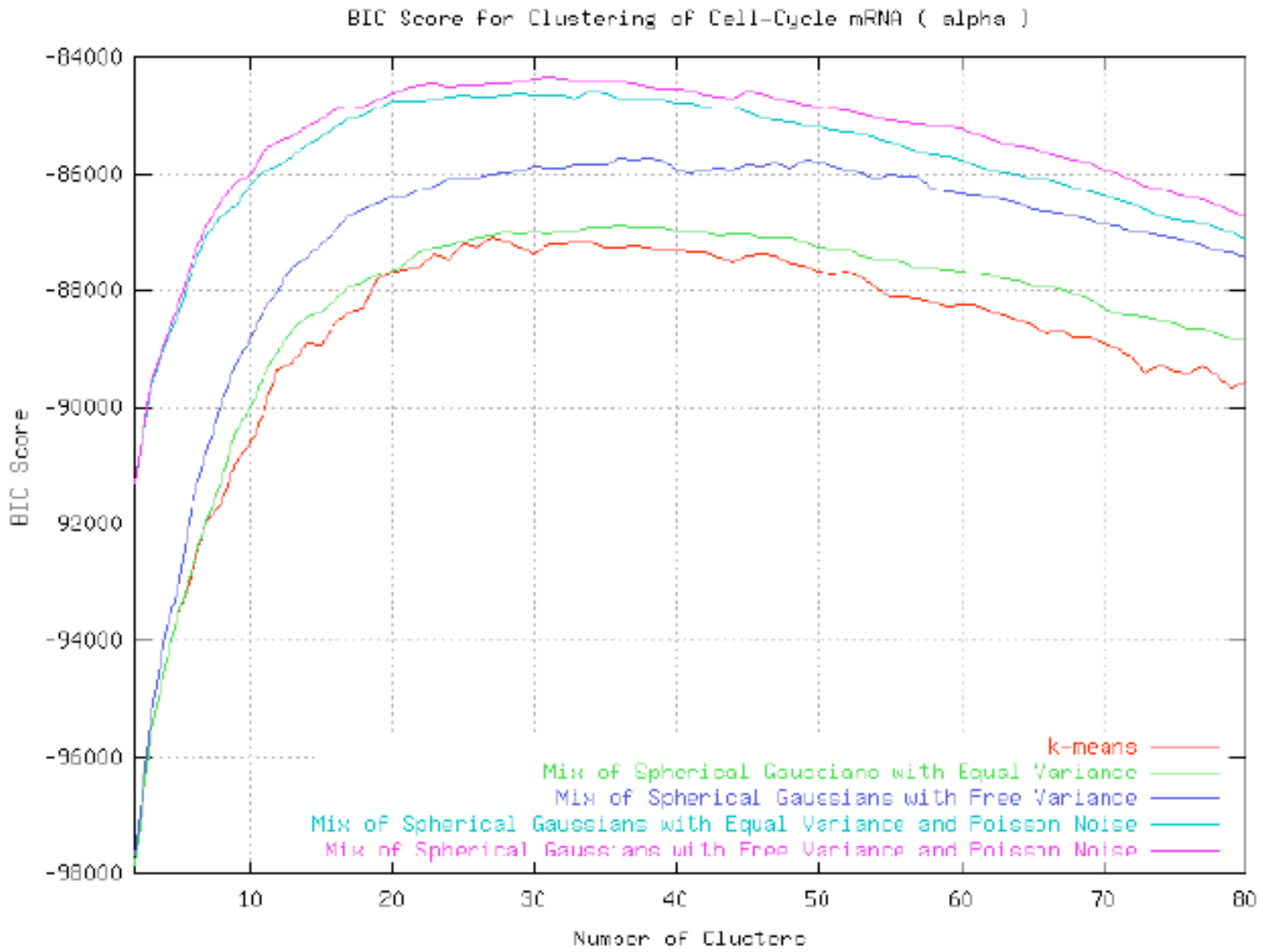
Results: standardized yeast cell cycle data



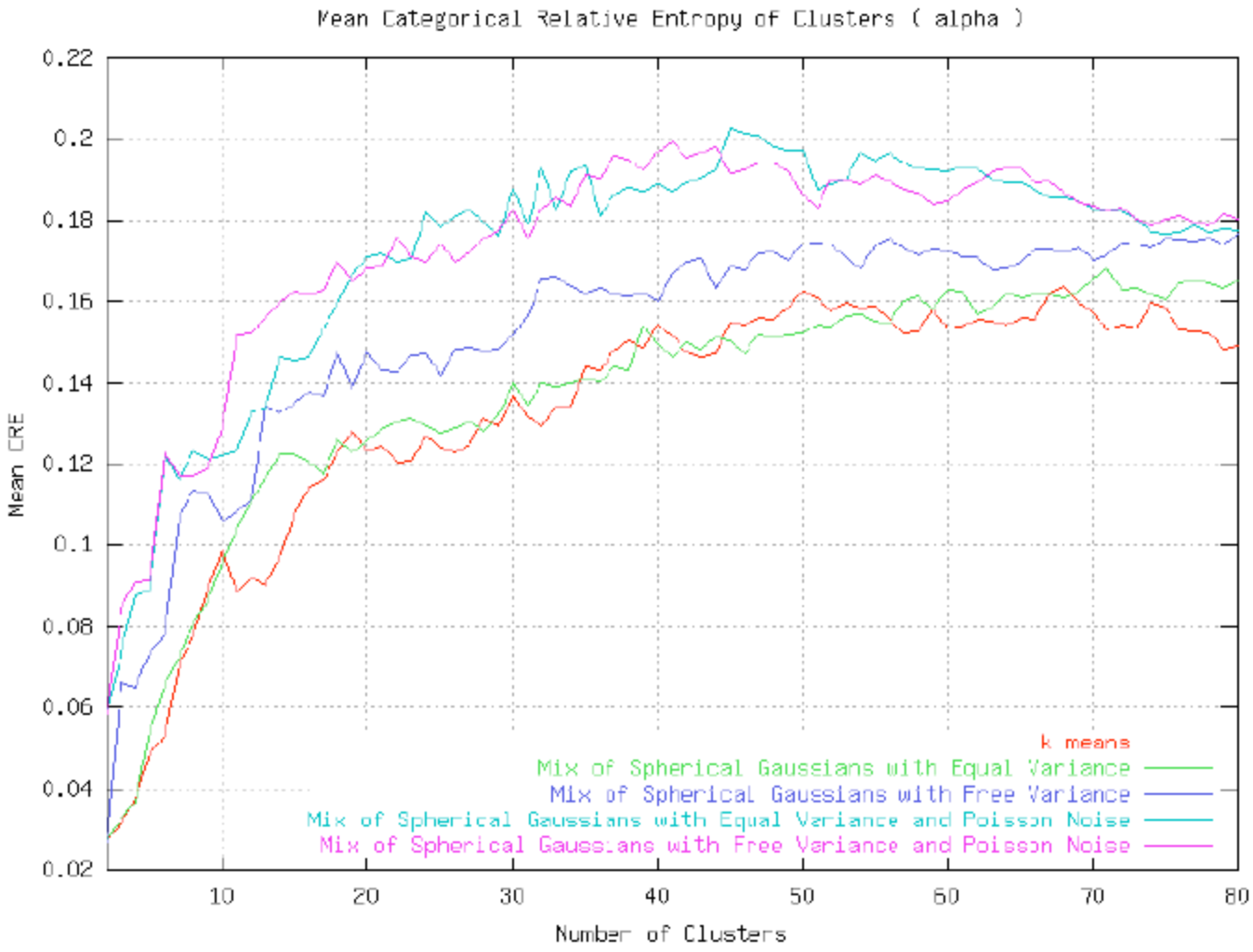


- Adjusted Rand: EI slightly > CAST at 5 clusters.
- BIC: selects EEE at 5 clusters.

BIC Scores for Clustering of Alpha–Factor Data with Noise Mixture Models



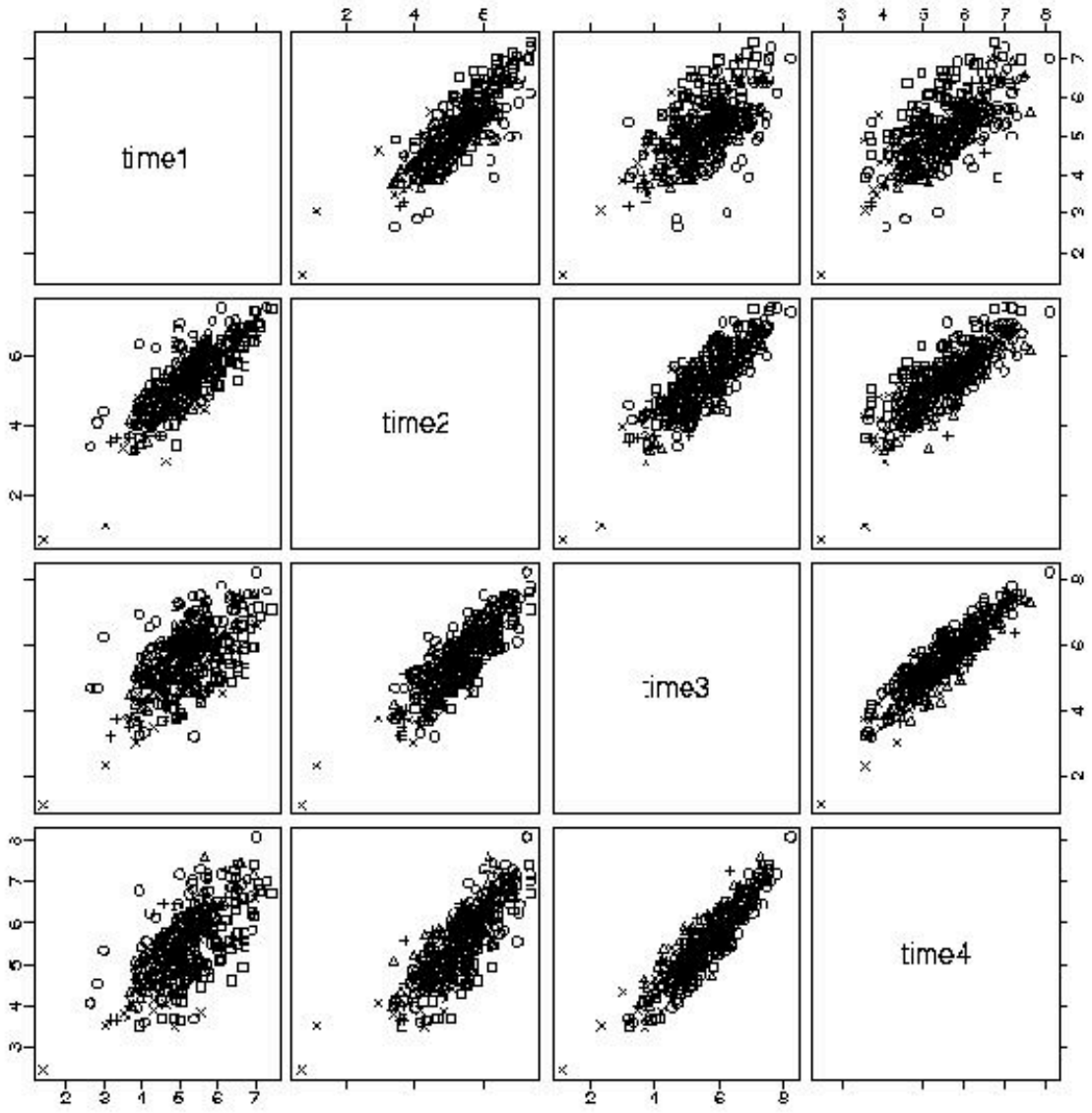
CRE Scores for Clustering of Alpha–Factor Data with Noise Mixture Models



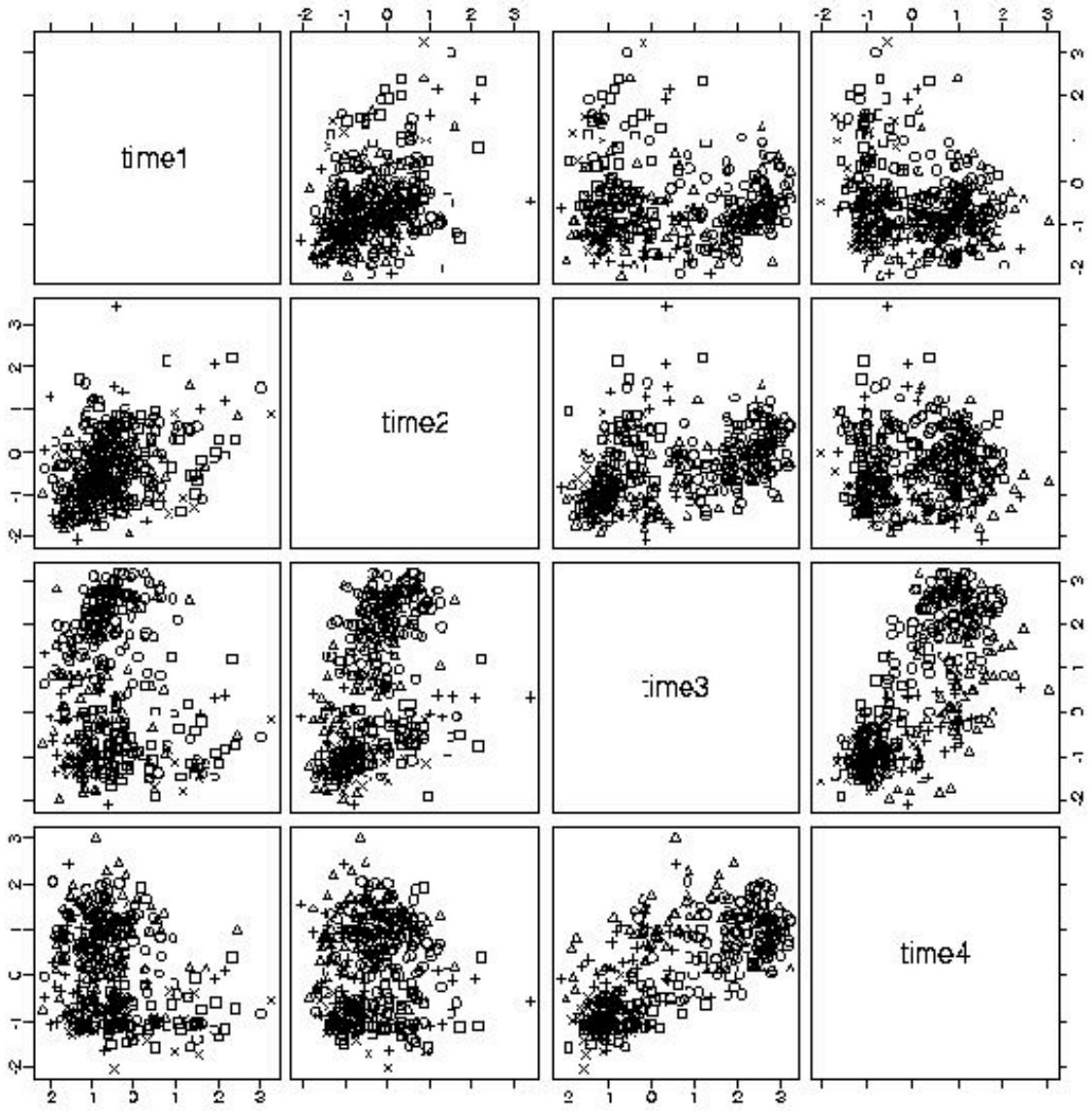
Overview

- Motivation
- Model-based clustering
- Validation
- Importance of Data Transformation
- Summary and Conclusions

log yeast cell cycle data



Standardized yeast cell cycle data



Overview

- Motivation
- Model-based clustering
- Validation
- Summary and Conclusions

Summary and Conclusions

- Synthetic data sets:
  - With the correct model, model-based clustering better than a leading heuristic clustering algorithm
  - BIC selects the right model & right number of clusters
- Real expression data sets:
  - Comparable adjusted Rand indices to CAST
  - BIC gives a good hint as to the number of clusters
- Appropriate data transformations increase normality & cluster quality (See paper & web.)

Acknowledgements

- Ka Yee Yeung[1], Chris Fraley[2,4], Alejandro Murua[4], Adrian E. Raftery[2]
- Michèl Schummer[5] – the ovary data
- Jeremy Tantrum[2] – help with MBC software (diagonal model)
- Chris Saunders[3] – CRE & noise model

[1]Computer Science & Engineering
[2]Statistics
[3]Genome Sciences
[4]Insightful Corporation
[5]Institute of Systems Biology

More Info

http://www.cs.washington.edu/homes/ruzzo

Adjusted Rand Example

	c#1(4)	c#2(5)	c#3(7)	c#4(4)
class#1(2)	2	0	0	0
class#2(3)	0	0	0	3
class#3(5)	1	4	0	0
class#4(10)	1	1	7	1

$$a = \binom{2}{2} + \binom{3}{2} + \binom{4}{2} + \binom{7}{2} = 31$$

$$b = \binom{4}{2} + \binom{5}{2} + \binom{7}{2} + \binom{4}{2} - a = 43 - 31 = 12$$

$$c = \binom{2}{2} + \binom{3}{2} + \binom{5}{2} + \binom{10}{2} - a = 59 - 31 = 28$$

$$d = \binom{20}{2} - a - b - c = 119$$

$$Rand, R = \frac{a+d}{a+d+c+d} = 0.789$$

$$\text{Adjusted Rand} = \frac{R - E(R)}{1 - E(R)} = 0.469$$