Machine Learning

Unsupervised Methods

Part 3

Sepp Hochreiter

Institute of Bioinformatics
Johannes Kepler University, Linz, Austria

Outline

1 Introduction

2 Basic Terms and Concepts

3 Principal Component Analysis

4 Independent Component Analysis

5 Factor Analysis

6 Scaling and Projection Methods

7 Clustering

8 Biclustering

9 Hidden Markov Models

10 Boltzmann Machines

7 Clustering

7.1 Mixture Models

7.2 k-Means Clustering

7.3 Hierarchical Clustering

7.4 Similarity-Based Clustering

Outline

8 Biclustering

8.1 Types of Biclusters

8.2 Overview of Biclustering Methods

8.3 FABIA Biclustering

8.4 Examples

9 Hidden Markov Models

9.1 Hidden Markov Models in Bioinformatics

9.2 Hidden Markov Model Basics

9.3 Expectation Maximization for HMM: Baum-Welch Algorithm

9.4 Viterby Algorithm

9.5 Input Output Hidden Markov Models

9.6 Factorial Hidden Markov Models

9.7 Memory Input Output Factorial Hidden Markov Models

9.8 Tricks of the Trade

9.9 Profile Hidden Markov Models

10 Boltzmann Machines

10.1 The Boltzmann Machine

10.2 Learning in the Boltzmann Machine

10.3 The Restricted Boltzmann Machine

Chapter 7

Clustering

Clustering

Clustering is one of the most popular unsupervised techniques

Clusters in the data are regions where observations group together
→ regions of high data density

clusters may correspond to a prototype from which observations are obtained via noise perturbations

Clustering extracts structures and can identify new data classes

important application of clustering: data visualization

observations are represented by prototypes: vector quantization

Mixture Models

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

Mixture models locally assign in the feature space a component which represents a cluster.

Component $\boldsymbol{j}$ out of $\boldsymbol{l}$ components has parameters like location $\boldsymbol{\mu}_j$ and width or shape $\boldsymbol{\Sigma}_j$. It has in every case a weight w_j that gives the local probability mass.

generative framework: w_j is the probability $p(j)$ of choosing component $\boldsymbol{j}$, which has density $p(\boldsymbol{x} \mid j, \boldsymbol{\theta}_j)$, where $\boldsymbol{\theta}_j$ summarizes the parameters of component $\boldsymbol{j}$

$\boldsymbol{\theta}$ summarize all parameters, which gives the generative model

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) \;=\; \sum_{j=1}^{l} p(j)\; p(\boldsymbol{x} \mid j, \boldsymbol{\theta}_j)$$

Mixture Models

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

For clustering, Bayes' formula can be used:

$$p(j \mid \boldsymbol{x}, \boldsymbol{\theta}) \; = \; \frac{p(\boldsymbol{x} \mid j, \boldsymbol{\theta}_j) \; p(j)}{p(\boldsymbol{x} \mid \boldsymbol{\theta})}$$

Observation $\boldsymbol{x}$ is assigned to the component j with largest posterior $p(j \mid \boldsymbol{x}, \boldsymbol{\theta})$

Before an observation was seen, each component or cluster has the prior probability; after observing data some clusters may be more or less probable of having produced the data, therefore the prior probability changes to the posterior.

Mixture components can be Poissons, or negative Binomials as used at our institute for analyzing sequencing data.

Optimizing Mixture Models

log-likelihood $\ln \mathcal{L} \ = \ \sum_{i=1}^{n} \ln p(\boldsymbol{x}_i \mid \boldsymbol{\theta})$

derivative with respect to parameters $\boldsymbol{\theta}_j$ of component j is

$$\frac{\partial}{\partial \boldsymbol{\theta}_j} \ln \mathcal{L} \ = \ \sum_{i=1}^{n} \frac{1}{p(\boldsymbol{x}_i \mid \boldsymbol{\theta})} \sum_{k=1}^{l} p(k) \ \frac{\partial}{\partial \boldsymbol{\theta}_j} p(\boldsymbol{x}_i \mid k, \boldsymbol{\theta}_k) \ = \ \sum_{i=1}^{n} p(j \mid \boldsymbol{x}_i, \boldsymbol{\theta}_j) \frac{\partial}{\partial \boldsymbol{\theta}_j} \ln p(\boldsymbol{x}_i \mid j, \boldsymbol{\theta}_j)$$

we used Bayes' formula

$$p(j \mid \boldsymbol{x}_i, \boldsymbol{\theta}_j) \ = \ \frac{p(\boldsymbol{x}_i \mid j, \boldsymbol{\theta}_j) \ p(j)}{p(\boldsymbol{x}_i \mid \boldsymbol{\theta})}$$

The derivative of the log-likelihood of the model with respect to the parameters of the j-th component:
posterior expectation of the derivative of the log-likelihood of component j

Mixture of Gaussians

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

We will now consider mixture of Gaussian (MoG): $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$

$$p(\boldsymbol{x}_i \mid j, \boldsymbol{\theta}_j) \ = \ p(\boldsymbol{x}_i \mid j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \ \sim \ \mathcal{N}(\boldsymbol{\mu}_j \ , \ \boldsymbol{\Sigma}_j)$$

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) \ = \ \sum_{j=1}^{l} w_j \, \mathcal{N}(\boldsymbol{\mu}_j \ , \ \boldsymbol{\Sigma}_j)$$

$$\sum_{j=1}^{l} w_j \ = \ 1 \ , \qquad w_j \ \geq \ 0$$

$$\mathcal{N}(\boldsymbol{\mu}_j \ , \ \boldsymbol{\Sigma}_j)(\boldsymbol{x}) \ = \ (2\pi)^{-m/2} \, |\boldsymbol{\Sigma}_j|^{-1/2} \exp\left(-\frac{1}{2} \ (\boldsymbol{x} \ - \ \boldsymbol{\mu}_j)^T \ \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x} \ - \ \boldsymbol{\mu}_j) \right)$$

Exponential distributions like the Gaussians are convenient because the logarithm inverts the exponential function:

$$\ln p(\boldsymbol{x} \mid j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \ = \ - \ \frac{m}{2} \ \ln(2\pi) \ - \ \frac{1}{2} \ \ln|\boldsymbol{\Sigma}_j| \ - \ \frac{1}{2} \ (\boldsymbol{x} \ - \ \boldsymbol{\mu}_j)^T \ \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x} \ - \ \boldsymbol{\mu}_j)$$

derivatives

$$\frac{\partial}{\partial \boldsymbol{\mu}_j} \ln p(\boldsymbol{x} \mid j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \ = \ \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x} \ - \ \boldsymbol{\mu}_j)$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_j} \ln p(\boldsymbol{x} \mid j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \ = \ \frac{1}{2} \ \left(\boldsymbol{\Sigma}_j^T\right)^{-1} \ + \ \frac{1}{2} \ \boldsymbol{\Sigma}_j^{-T} (\boldsymbol{x} \ - \ \boldsymbol{\mu}_j) \ (\boldsymbol{x} \ - \ \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-T}$$

Mixture of Gaussians

EM-algorithm:

E-step:

$$p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \ = \ \frac{w_j \ \mathcal{N}\left(\boldsymbol{\mu}_j \ , \ \boldsymbol{\Sigma}_j\right)\left(\boldsymbol{x}_i\right)}{\sum_{t=1}^{l} w_t \ \mathcal{N}\left(\boldsymbol{\mu}_t \ , \ \boldsymbol{\Sigma}_t\right)\left(\boldsymbol{x}_i\right)}$$

M-step:

$$\boldsymbol{w}_j^{\mathrm{new}} \ = \ \frac{1}{n} \sum_{i=1}^{n} p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

$$\boldsymbol{\mu}_j^{\mathrm{new}} \ = \ \frac{\sum_{i=1}^{n} \ p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \ \boldsymbol{x}_i}{\sum_{i=1}^{n} \ p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

$$\boldsymbol{\Sigma}_j^{\mathrm{new}} \ = \ \frac{\sum_{i=1}^{n} \ p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \ \left(\boldsymbol{x}_i \ - \ \boldsymbol{\mu}_j\right) \ \left(\boldsymbol{x}_i \ - \ \boldsymbol{\mu}_j\right)^T}{\sum_{i=1}^{n} \ p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Mixture of Gaussians: MAP

Maximum a posteriori MoG:

- proper prior for $\boldsymbol{\Sigma}_j$: Wishart density $\mathcal{W}(\boldsymbol{\Sigma}^{-1} \mid \alpha, \boldsymbol{\Psi})$
- proper prior for weights w_j: Dirichlet density $\mathcal{D}(\boldsymbol{w} \mid \gamma)$
- proper prior for means $\boldsymbol{\mu}_j$: Gaussian density $\mathcal{N}\left(\boldsymbol{\mu} \mid \boldsymbol{\nu}, \eta^{-1}\boldsymbol{\Sigma}\right)$

$$\mathcal{W}(\boldsymbol{\Sigma}^{-1} \mid \alpha, \boldsymbol{\Psi}) \; = \; c(\alpha, \boldsymbol{\Psi}) \; \left|\boldsymbol{\Sigma}^{-1}\right|^{\alpha-(m+1)/2} \; \exp\left(-\mathrm{tr}\left(\boldsymbol{\Psi} \; \boldsymbol{\Sigma}^{-1}\right)\right)$$

$$\mathcal{D}(\boldsymbol{w} \mid \gamma) \; = \; c(\gamma) \prod_{j=1}^{l} w_j^{\gamma-1}$$

$$\mathcal{N}\left(\boldsymbol{\mu} \mid \boldsymbol{\nu}, \eta^{-1}\boldsymbol{\Sigma}\right) \; = \; (2\pi)^{-m/2} \left|\eta^{-1} \; \boldsymbol{\Sigma}_j\right|^{-1/2} \exp\left(-\frac{\eta}{2} \; (\boldsymbol{\mu} \; - \; \boldsymbol{\nu})^T \; \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{\mu} \; - \; \boldsymbol{\nu})\right)$$

normalizing constants: $\alpha > (m-1)/2, c(\gamma), c(\alpha, \boldsymbol{\Psi})$

“tr” is the trace operator

Mixture of Gaussians: MAP and EM

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

expectation-maximization:

E-step:

$$p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \;=\; \frac{w_j \, \mathcal{N}\left(\boldsymbol{\mu}_j \;,\; \boldsymbol{\Sigma}_j\right)(\boldsymbol{x}_i)}{\sum_{t=1}^{l} w_t \, \mathcal{N}\left(\boldsymbol{\mu}_t \;,\; \boldsymbol{\Sigma}_t\right)(\boldsymbol{x}_i)}$$

M-step:

$$w_j^{\mathrm{new}} \;=\; \frac{\sum_{i=1}^{n} p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \;+\; \gamma \;-1}{n \;+\; l \;(\gamma \;-\; 1)}$$

$$\boldsymbol{\mu}_j^{\mathrm{new}} \;=\; \frac{\sum_{i=1}^{n} \; p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \; \boldsymbol{x}_i \;+\; \eta \; \boldsymbol{\nu}_j}{\sum_{i=1}^{n} \; p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \;+\; \eta}$$

$$\boldsymbol{\Sigma}_j^{\mathrm{new}} \;=\; \left(\sum_{i=1}^{n} p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \; (\boldsymbol{x}_i \;-\; \boldsymbol{\mu}_j) \; (\boldsymbol{x}_i \;-\; \boldsymbol{\mu}_j)^T \;+\right.$$

$$\left. \eta \; (\boldsymbol{\nu}_j \;-\; \boldsymbol{\mu}_j) \; (\boldsymbol{\nu}_j \;-\; \boldsymbol{\mu}_j)^T \;+\; 2 \; \boldsymbol{\Psi} \right)$$

$$\left(\sum_{i=1}^{n} p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \;+\; 2 \; \alpha \;-m \right)^{-1}$$

Mixture of Gaussians: Derivative of MAP EM

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

Constraints: $\sum_{j=1}^{l} w_j^{\text{new}} \ = \ 1$

Lagrangian for the constrained optimization problem for the w_j:

$$L \ = \ \sum_{i=1}^{n} \sum_{j=1}^{l} p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \ \ln w_j^{\text{new}} \ + \ \ln \mathcal{D}(\boldsymbol{w} \mid \gamma) \ - \ \lambda \left(\sum_{j=1}^{l} w_j^{\text{new}} \ - \ 1 \right)$$

Setting the derivative to zero:

$$\frac{\partial L}{\partial w_j} \ = \ \sum_{i=1}^{n} p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \ \left(w_j^{\text{new}}\right)^{-1} \ + \ (\gamma \ - \ 1) \ \left(w_j^{\text{new}}\right)^{-1} \ - \ \lambda \ = \ 0$$

$$\sum_{i=1}^{n} p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \ + \ (\gamma \ - \ 1) \ = \ \lambda \, w_j^{\text{new}}$$

Summing over j: $\sum_{i=1}^{n} \sum_{j=1}^{l} p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \ + \ l \, (\gamma \ - \ 1) \ = \ \lambda \sum_{j=1}^{l} w_j^{\text{new}}$

$$n \ + \ l \, (\gamma \ - \ 1) \ = \ \lambda$$

We obtain

$$\boxed{w_j^{\text{new}} \ = \ \frac{\sum_{i=1}^{n} p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \ + \ (\gamma \ - \ 1)}{n \ + \ l \, (\gamma \ - \ 1)}}$$

Mixture of Gaussians: Derivative of MAP EM

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

Derivatives of the posterior with respect to parameters of mixture components are set to zero:

$$\frac{\partial L}{\partial \boldsymbol{\mu}_j} = \sum_{i=1}^{n} p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \; \boldsymbol{\Sigma}_j^{-1} \; (\boldsymbol{x}_i - \boldsymbol{\mu}_j) + \eta \; \boldsymbol{\Sigma}_j^{-1} \; (\boldsymbol{\nu}_j - \boldsymbol{\mu}_j) = 0$$

gradient with respect to $\boldsymbol{\Sigma}_j^{-1}$ must be zero, too:

$$\frac{\partial L}{\partial \boldsymbol{\Sigma}_j} = \frac{\partial L}{\partial \boldsymbol{\Sigma}_j^{-1}} \frac{\partial \boldsymbol{\Sigma}_j^{-1}}{\partial \boldsymbol{\Sigma}_j} = - \boldsymbol{\Sigma}_j^{-2} \frac{\partial L}{\partial \boldsymbol{\Sigma}_j^{-1}}$$

We obtain

$$\frac{\partial L}{\partial \boldsymbol{\Sigma}_j^{-1}} = \frac{1}{2} \sum_{i=1}^{n} p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \left(\boldsymbol{\Sigma}_j - (\boldsymbol{x}_i - \boldsymbol{\mu}_j) \; (\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T \right) +$$

$$\frac{1}{2} \left(\boldsymbol{\Sigma}_j - \eta \; (\boldsymbol{\nu}_j - \boldsymbol{\mu}_j) \; (\boldsymbol{\nu}_j - \boldsymbol{\mu}_j)^T \right) + \boldsymbol{\Sigma}_j \; (\alpha - (m + 1)/2) - \boldsymbol{\Psi} = 0$$

where we used $\frac{\partial \ln |\boldsymbol{U}|}{\partial \boldsymbol{U}} = \boldsymbol{U}^{-1}$ for $\boldsymbol{U} = \boldsymbol{\Sigma}_j^{-1}$

Mixture of Gaussians: MAP Hyperparameters

Default hyperparameters:

$$\begin{aligned}
\alpha &= \frac{m}{2} \\
\boldsymbol{\Psi} &= \frac{1}{2} \boldsymbol{I} \quad \text{OR} \quad \boldsymbol{\Psi} = \frac{1}{2} \mathrm{covar}(\boldsymbol{x}) \\
\gamma &= 1 \\
\eta &= 0 \\
\boldsymbol{\nu}_j &= \mathrm{mean}(\boldsymbol{x})
\end{aligned}$$

A prior on the mean is in most cases not useful except a preferred region is known.

The posterior $p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ can be used for clustering: data belongs to the cluster for which the posterior is largest.

Mixture of Poissons

Mixture of Poissons with count data x: $p(x) \ = \ \sum_{j=1}^{l} w_j \ \mathrm{P}(x;\lambda_j)$

P is the probability mass function of the Poisson distribution:

$$\mathrm{P}(x;\lambda) \ = \ \frac{1}{x!} \ e^{-\lambda} \ \lambda^{x}$$

Bayes framework for model selection:

$$\boldsymbol{w} = (w_1, \ldots, w_l)$$
$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_l)$$

posterior: $p(\boldsymbol{w}, \boldsymbol{\lambda} \mid x) \ = \ \frac{p(x \mid \boldsymbol{w}, \boldsymbol{\lambda}) \ p(\boldsymbol{w}) \ p(\boldsymbol{\lambda})}{\int p(x \mid \boldsymbol{w}, \boldsymbol{\lambda}) \ p(\boldsymbol{w}) \ p(\boldsymbol{\lambda}) \ d\boldsymbol{w} \ d\boldsymbol{\lambda}}$

Mixture of Poissons

Dirichlet prior $\boldsymbol{w}^l \; = \; (w_2, \ldots, w_l)$ $\quad w_1 = 1 - \sum_{j=2}^{l} w_j$

$$p(\boldsymbol{w}) \; = \; \mathrm{D}(\boldsymbol{w}^l ; \boldsymbol{\gamma}) \; = \; b(\boldsymbol{\gamma}) \prod_{j=1}^{l} w_j^{\gamma_j - 1}$$

Each component w_j is distributed according to a beta distribution:

$$\mathrm{mean}(w_j) \; = \; \frac{\gamma_j}{\gamma_s}$$

$$\mathrm{mode}(w_j) \; = \; \frac{\gamma_j - 1}{\gamma_s - l}$$

$$\mathrm{var}(w_j) \; = \; \frac{\gamma_j \; (\gamma_s - \gamma_j)}{\gamma_s^2 \; (\gamma_s + 1)}$$

$$\gamma_s \; = \; \sum_{j=1}^{l} \gamma_i$$

Mixture of Poissons

Prior on λ_j is a uniform distribution on $(0, 1/t]$: $p(\lambda_j) \; = \; t$

posterior of the model parameters:

$$p(\boldsymbol{w}, \boldsymbol{\lambda} \mid x) \;=\; \frac{p(x \mid \boldsymbol{w}, \boldsymbol{\lambda}) \; p(\boldsymbol{w}) \; p(\boldsymbol{\lambda})}{\int p(x \mid \boldsymbol{w}, \boldsymbol{\lambda}) \; p(\boldsymbol{w}) \; p(\boldsymbol{\lambda}) \; d\boldsymbol{w} \; d\boldsymbol{\lambda}} \;=\; \frac{p(x \mid \boldsymbol{w}, \boldsymbol{\lambda}) \; p(\boldsymbol{w})}{\int p(x \mid \boldsymbol{w}, \lambda) \; p(\boldsymbol{w}) \; d\boldsymbol{w} \; d\boldsymbol{\lambda}}$$

$$=\; \frac{1}{c(x)} \; p(x \mid \boldsymbol{w}, \boldsymbol{\lambda}) \; p(\boldsymbol{w})$$

where $c(x)$ is independent of the model parameters

Mixture of Poissons

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

upper bound on log posterior with new variables and one data point

$$\hat{w}_j , \ \sum_{j=1}^{l} \hat{w}_j = 1$$

$$- \ \log p(\boldsymbol{w}, \boldsymbol{\lambda} \mid x) \ = \ - \ \log \left(p(x \mid \boldsymbol{w}, \boldsymbol{\lambda}) \ p(\boldsymbol{w}) \ / \ c(x) \right) \ = \ - \ \log \sum_{j=1}^{l} w_j \ \mathrm{P}(x; \lambda_j) \ - \ \log p(\boldsymbol{w}) \ + \ \log(c(x))$$

$$= \ - \ \log \sum_{j=1}^{l} \frac{\hat{w}_j}{\hat{w}_j} \ w_j \ \mathrm{P}(x; \lambda_j) \ - \ \log p(\boldsymbol{w}) \ + \ \log(c(x)) \ \leq \ - \ \sum_{j=1}^{l} \hat{w}_j \ \log \frac{w_j \ \mathrm{P}(x; \lambda_j)}{\hat{w}_j} \ - \ \log p(\boldsymbol{w}) \ + \ \log(c(x))$$

$$= \ - \ \sum_{j=1}^{l} \hat{w}_j \ \log \left(w_j \ \mathrm{P}(x; \lambda_j) \right) \ - \ \log p(\boldsymbol{w}) \ + \ \sum_{j=1}^{l} \hat{w}_j \ \log \hat{w}_j \ + \ \log(c(x))$$

where we applied Jensen's inequality and for

$$\hat{w}_j \ = \ p(j \mid x, \boldsymbol{w}, \boldsymbol{\lambda}) \ = \ \frac{w_j \ \mathrm{P}(x; \lambda_j)}{p(x \mid \boldsymbol{w}, \boldsymbol{\lambda})}$$

we obtain

$$\log \frac{w_j \ \mathrm{P}(x; \lambda_j)}{\hat{w}_j} \ = \ \log p(x \mid \boldsymbol{w}, \boldsymbol{\lambda})$$

and the inequality becomes an equality

Mixture of Poissons

Count data set: $\{x_1, \ldots, x_n\}$

Component posterior for component j and count x_i :

$$w_{ji} \ = \ p(j \mid x_i, \boldsymbol{w}, \boldsymbol{\lambda}) \ = \ \frac{p(j) \ p(x_i \mid j, \boldsymbol{w}, \boldsymbol{\lambda})}{p(x_i \mid \boldsymbol{w}, \boldsymbol{\lambda})} \ = \ \frac{w_j \ \mathrm{P}(x_i; \lambda_j)}{p(x_i \mid \boldsymbol{w}, \boldsymbol{\lambda})}$$

We introduce for each x_i variables $\hat{w}_{ji}, \ \sum_{j=1}^{l} \hat{w}_{ji} = 1$ which

approximate $p(j \mid \boldsymbol{w}, x_i, \boldsymbol{\lambda})$ and are independent of parameters

$$\hat{w}_{ji} \ = \ \frac{w_j^{\mathrm{old}} \ \mathrm{P}(x_i; \lambda_j^{\mathrm{old}})}{p(x_i; \boldsymbol{w}^{\mathrm{old}}, \boldsymbol{\lambda}^{\mathrm{old}})}$$

for the estimation the actual parameters “old” are used instead of the optimal parameters

Mixture of Poissons

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

upper bound B on the 1/n scaled negative log-posterior for all data:

$$B \;=\; - \; \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{l} \hat{w}_{ji} \; \log \left(w_j \; \mathrm{P}(x; \lambda_j) \right) \; - \; \frac{1}{n} \; \log p(\boldsymbol{w}) \; + \; \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{l} \hat{w}_{ji} \; \log \hat{w}_{ji}$$

$$+ \; \frac{1}{n} \sum_{i=1}^{n} \; \log c(x_i)$$

Mixture of Poissons

Optimizing $\boldsymbol{w}$:

$$\min_{\boldsymbol{w}} \; - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{l} \hat{w}_{ji} \log w_j \; - \; \frac{1}{n} \; \log p(\boldsymbol{w})$$

$$\text{s.t.} \quad \sum_{j=1}^{l} w_j \; = \; 1$$

Lagrangian:

$$L \; = \; - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{l} \hat{w}_{ji} \log w_j \; - \; \frac{1}{n} \; \log p(\boldsymbol{w}) \; + \; \rho \left(\sum_{j=1}^{l} w_j \; - \; 1 \right)$$

$$= \; - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{l} \hat{w}_{ji} \; \log w_j \; - \; \frac{1}{n} \sum_{j=1}^{l} (\gamma_j \; - \; 1) \; \log w_j \; + \; \rho \left(\sum_{j=1}^{l} w_j \; - \; 1 \right)$$

Derivatives zero:

$$\frac{\partial L}{\partial w_j} \; = \; - \frac{1}{n} \sum_{i=1}^{n} \hat{w}_{ji} \frac{1}{w_j} \; - \; \frac{1}{n} \; \frac{1}{w_j} \; (\gamma_j \; - \; 1) \; + \; \rho \; = \; 0$$

multiplying by w_j

$$- \frac{1}{n} \sum_{i=1}^{n} \hat{w}_{ji} \; - \; \frac{1}{n} \; (\gamma_j \; - \; 1) \; + \; \rho \, w_j \; = \; 0$$

Mixture of Poissons

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

Summation over j: $1 + \frac{1}{n}\left(\gamma_s - l\right) = \rho$

Inserting this expression:

$$-\frac{1}{n}\sum_{i=1}^{n}\hat{w}_{ji} - \frac{1}{n}\left(\gamma_j - 1\right) + \left(1 + \frac{1}{n}\left(\gamma_s - l\right)\right) w_j = 0$$

$$w_j^{\text{new}} = \frac{\hat{w}_j + \frac{1}{n}\left(\gamma_j - 1\right)}{1 + \frac{1}{n}\left(\gamma_s - l\right)} \qquad \hat{w}_j = \frac{1}{n}\sum_{i=1}^{n}\hat{w}_{ji}$$

Note that:

$$w_j = p(j) = p(j \mid \boldsymbol{w}, \boldsymbol{\lambda}) = \int p(j, x \mid \boldsymbol{w}, \boldsymbol{\lambda})\, dx = \int p(j \mid x, \boldsymbol{w}, \boldsymbol{\lambda})\, p(x \mid \boldsymbol{w}, \boldsymbol{\lambda})\, dx$$

$$= \mathrm{E}_{p(x \mid \boldsymbol{w}, \boldsymbol{\lambda})}\left(p(j \mid x, \boldsymbol{w}, \boldsymbol{\lambda})\right) \approx \frac{1}{n}\sum_{i=1}^{n} p(j \mid x_i, \boldsymbol{w}, \boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n} w_{ji} = \hat{w}_j$$

Mixture of Poissons

Optimizing λ_j: $\min_{\lambda_j} \left(- \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{l} \hat{w}_{ji} \log \mathrm{P}(x; \lambda_j) \right)$

$$\log \mathrm{P}(x_i; \lambda_j) \ = \ - \ \log(x_i!) \ - \ \lambda_j \ + \ x_i \ \log(\lambda_j)$$

the derivative of the objective with respect to λ_j is

$$- \frac{1}{n} \sum_{i=1}^{n} \left(- 1 \ + \ \frac{x_i}{\lambda_j} \right) \ \hat{w}_{ji}$$

Multiplying by λ_j and solving for it:

$$\lambda_j^{\mathrm{new}} \ = \ \frac{\sum_{i=1}^{n} x_i \ \hat{w}_{ji}}{\sum_{i=1}^{n} \hat{w}_{ji}}$$

Mixture of Poissons

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

update rules:

$$\hat{w}_{ji} = \frac{w_j^{\mathrm{old}} \; \mathrm{P}(x_i; \lambda_j^{\mathrm{old}})}{p(x_i \mid \boldsymbol{w}^{\mathrm{old}}, \boldsymbol{\lambda}^{\mathrm{old}})}$$

$$w_j^{\mathrm{new}} = \frac{\frac{1}{n} \sum_{i=1}^{n} \hat{w}_{ji} \; + \; \frac{1}{n}(\gamma_j \; - \; 1)}{1 \; + \; \frac{1}{n}(\gamma_s \; - \; l)}$$

$$\lambda_j^{\mathrm{new}} = \frac{\frac{1}{n} \sum_{i=1}^{n} \hat{w}_{ji} \; x_i}{\frac{1}{n} \sum_{i=1}^{n} \; \hat{w}_{ji}}$$

Mixture of Gaussians: Initialization

initialization of the MoG:

- “em”: first several low tolerance fast runs then a precise slow run
- “rnd”: random initializations and pick the best
- “svd”: singular value decomposition to find a good initialization

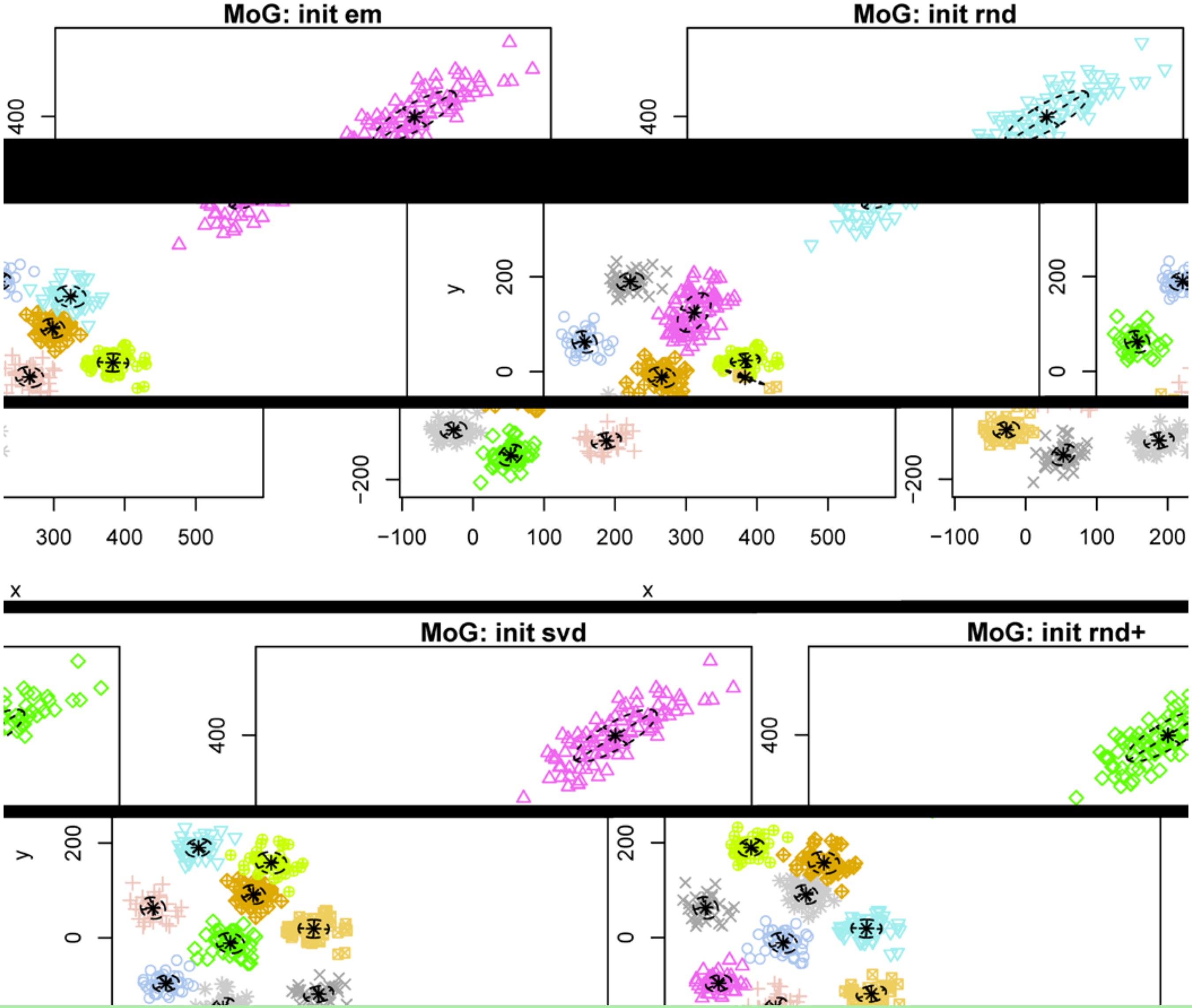
Mixture of Gaussians: Example

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine



Mixture of Gaussians: Constraints

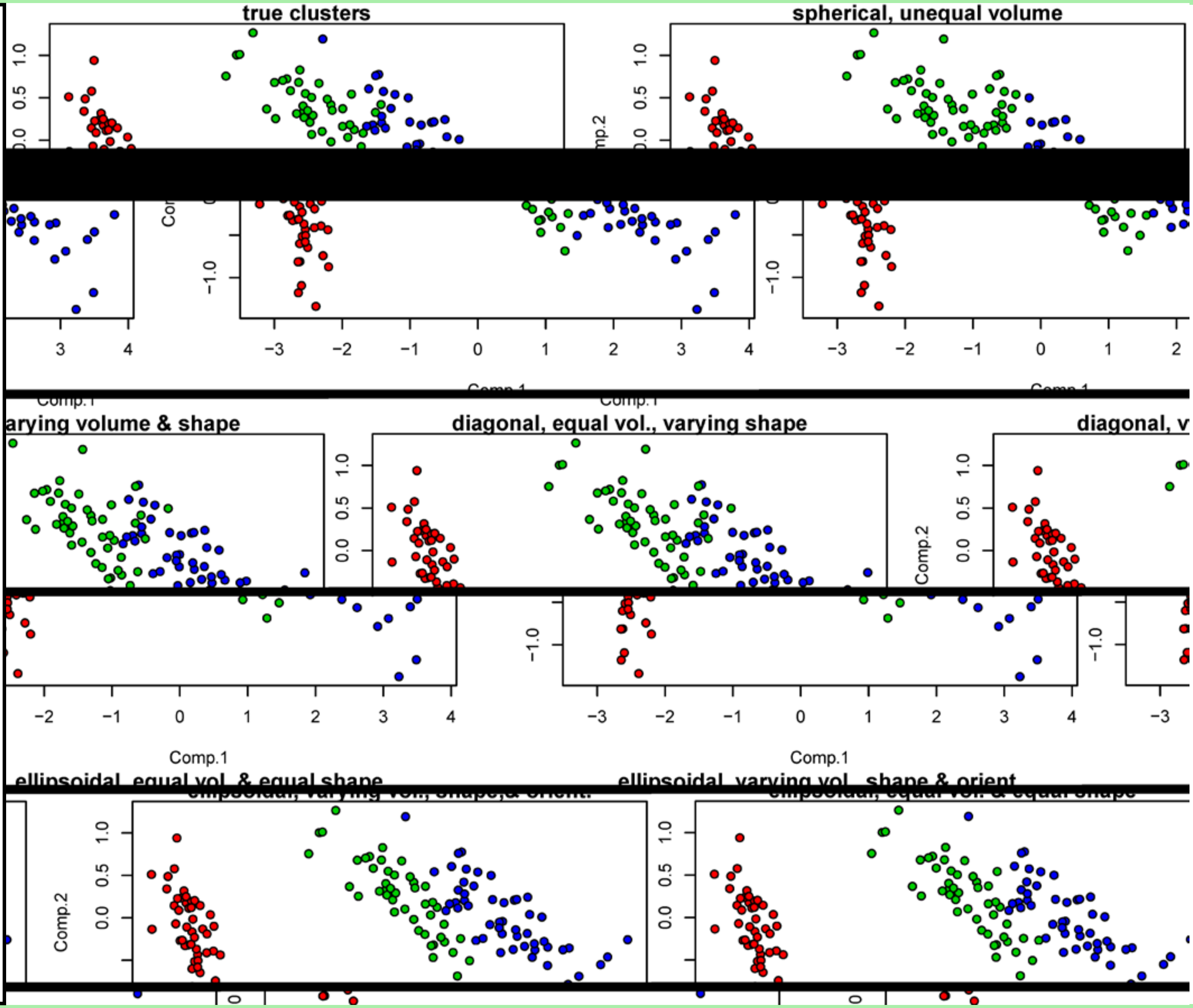
MoG for the iris data set

constraints on the parameters:

- “spherical”: covariance matrix is a multiple of the identity
- “diagonal”: diagonal covariance matrix (clusters along axis)
- “volume”: weighting factor or prior for the components

```
univariate mixture:
"E"        =          equal variance (one-dimensional)
"V"        =          variable variance (one-dimensional)
multivariate mixture:
"EII"      =          spherical, equal volume
"VII"      =          spherical, unequal volume
"EEI"      =          diagonal, equal volume and shape
"VEI"      =          diagonal, varying volume, equal shape
"EVI"      =          diagonal, equal volume, varying shape
"VVI"      =          diagonal, varying volume and shape
"EEE"      =          ellipsoidal, equal volume, shape, and orientation
"EEV"      =          ellipsoidal, equal volume and equal shape
"VEV"      =          ellipsoidal, equal shape
"VVV"      =          ellipsoidal, varying volume, shape, and orientation
single component:
"X"        =          univariate normal
"XII"      =          spherical multivariate normal
"XXI"      =          diagonal multivariate normal
"XXX"      =          elliposidal multivariate normal
```

Mixture of Gaussians: Example



MoG with 3 comp.

spherical, unequal vol.

diagonal, equal volume, varying shape

diagonal, varying volume & shape

ellipsoidal, equal vol. and equal shape

ellipsoidal, varying vol., shape, and orientation

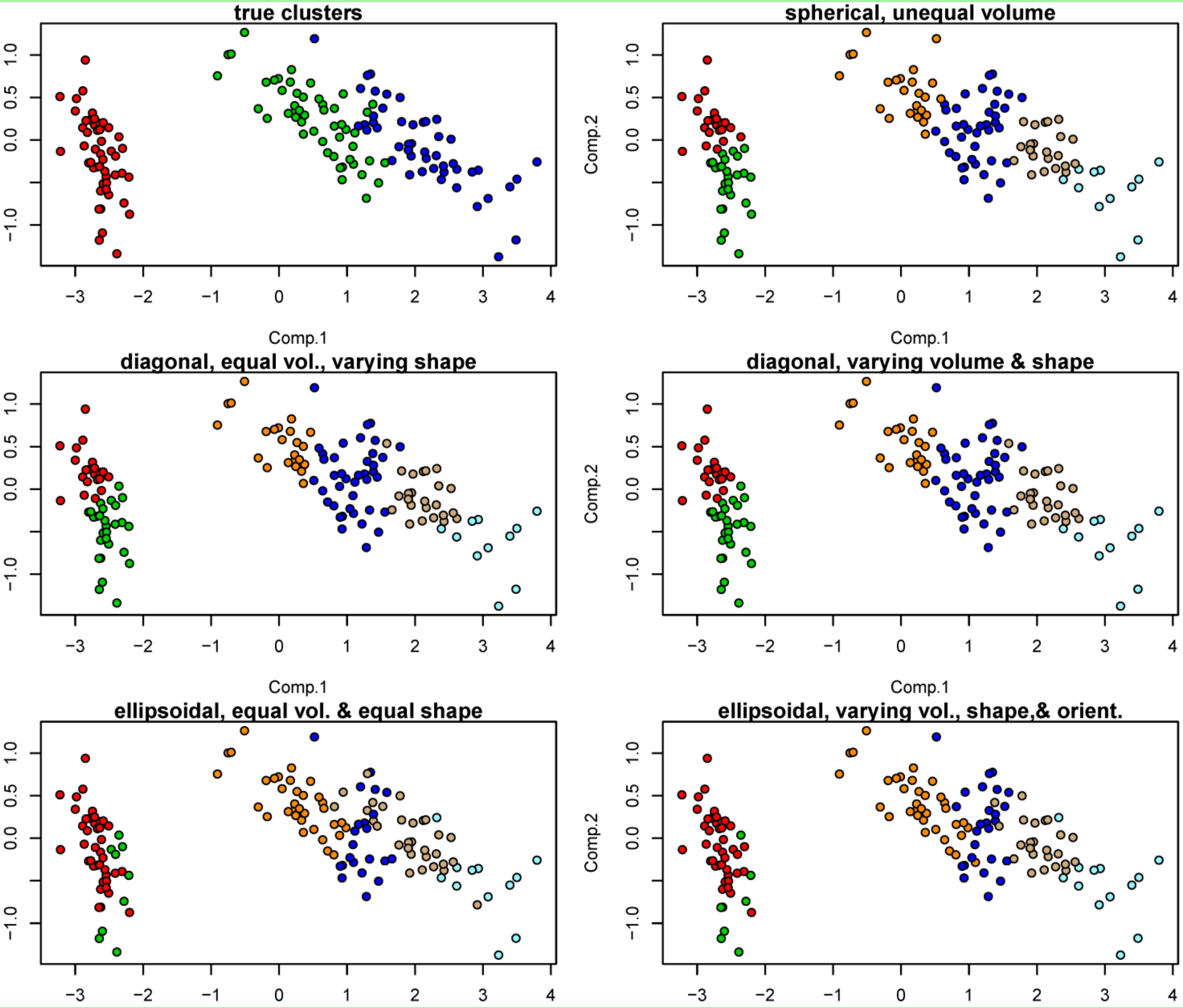
Mixture of Gaussians: Example

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine


true clusters

spherical, unequal volume

diagonal, equal vol., varying shape

diagonal, varying volume & shape

ellipsoidal, equal vol. & equal shape

ellipsoidal, varying vol., shape,& orient.

MoG with 6 comp.

spherical, unequal vol.

diagonal, equal volume, varying shape

diagonal, varying volume & shape

ellipsoidal, equal vol. and equal shape

ellipsoidal, varying vol., shape, and orientation

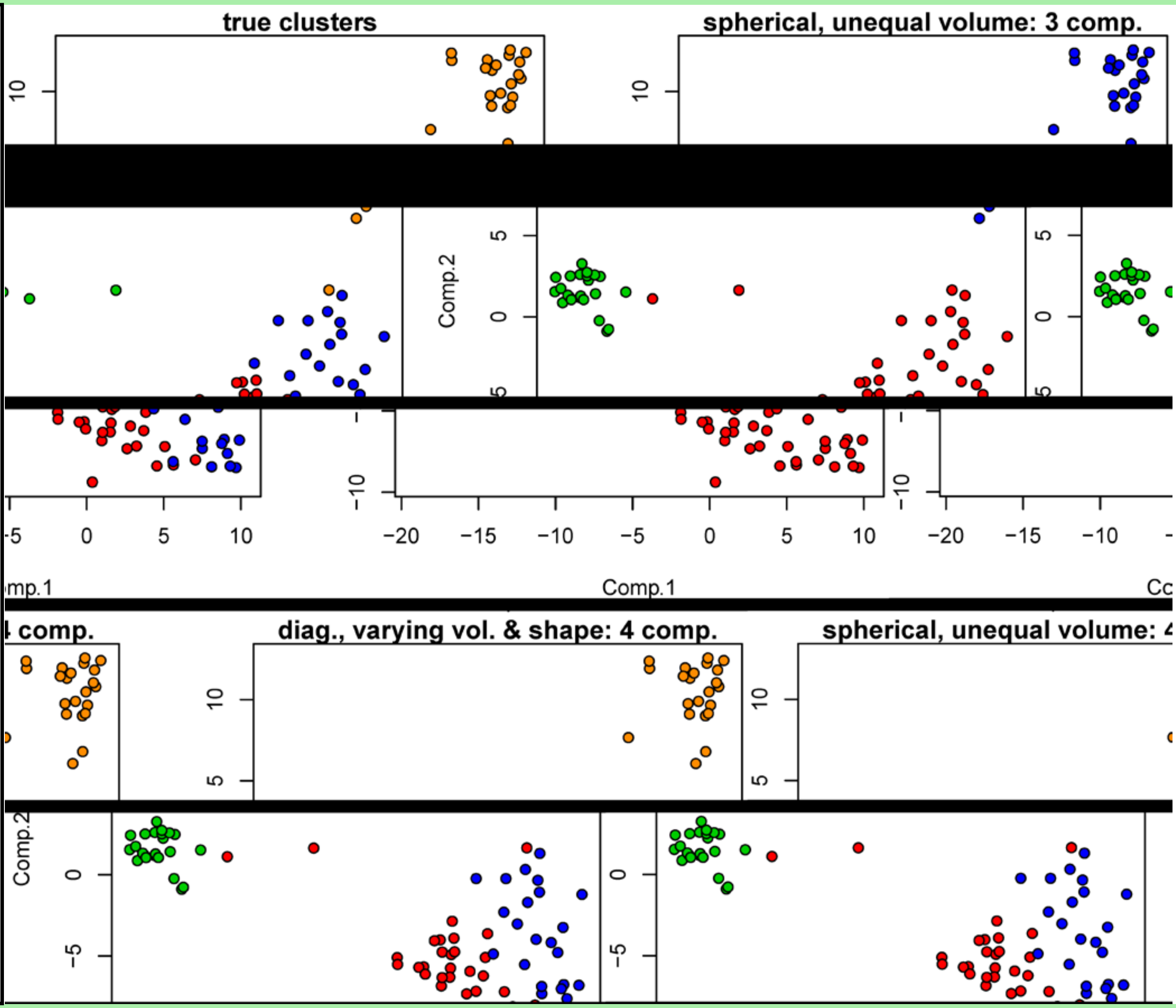
Mixture of Gaussians: Example

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine



MoG applied to multiple tissues: 76 genes with largest variance

“spherical, unequal volume”

“diagonal, equal volume, varying shape”

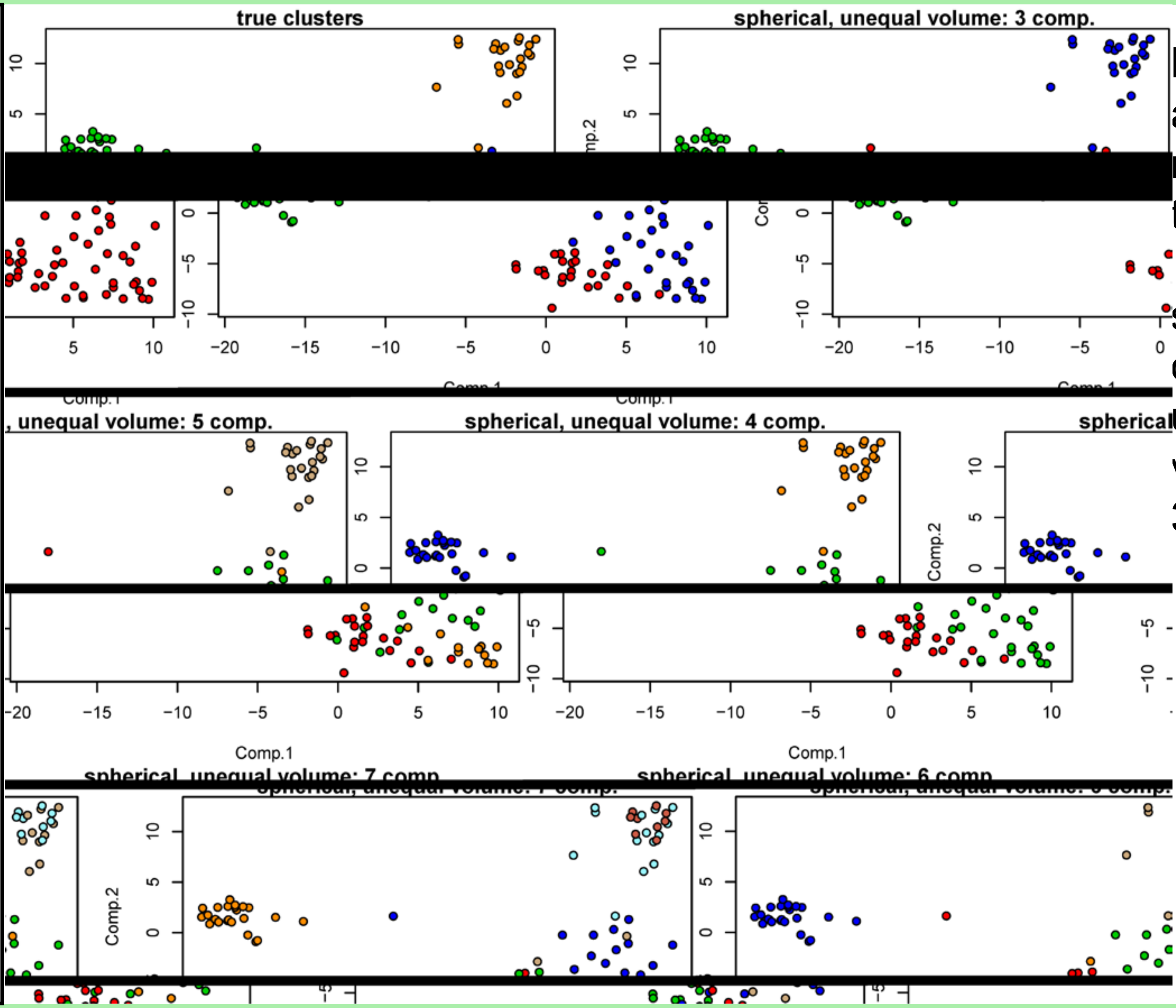
Mixture of Gaussians: Example

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine



[redacted]

MoG applied to multiple tissues

spherical compo. with unequal volume and 3 to 7 comp.

k-Means Clustering

k-means clustering is probably the best known clustering algorithm

k-means clustering is obtained from mixture clustering if the model assumptions are simplified:

- equal weight (equal volume) for each component: $w_j = \frac{1}{l}$
- spherical and equal (between components) covariance $\boldsymbol{\Sigma}_j^{-1} = \boldsymbol{I}$
- hard (discrete) cluster membership (a sample belongs to a cluster or not)

The only remaining parameters are the cluster centers.

A sample belongs to the cluster with the closest center:

$$p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j) = \begin{cases} 1 & \text{if } j = c_{\boldsymbol{x}_i} = \arg\min_k \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\| \\ 0 & \text{otherwise} \end{cases}$$

Center updates (according to the mixture model: mean of members):

$$\boldsymbol{\mu}_j^{\text{new}} = \frac{1}{n_j} \sum_{i=1,\ j=c_{\boldsymbol{x}_i}}^{n} \boldsymbol{x}_i \ , \qquad n_j = \sum_{i=1}^{n} p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \sum_{i=1,\ j=c_{\boldsymbol{x}_i}}^{n} 1$$

k-Means Clustering

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

```
 Given: data $\{\boldsymbol{x}\} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$, number of clusters $l$

BEGIN initialization
 initialize the cluster centers $\boldsymbol{\mu}_j$, $1 \leq j \leq l$
END initialization

BEGIN Iteration

 Stop=false
 while Stop=false do
   for $(i = 1 \; ; \; i \; \geq \; n \; ; \; i++)$ do
     assign $\boldsymbol{x}_i$ to the nearest $\boldsymbol{\mu}_j$
   end for
   for $(j = 1 \; ; \; j \; \geq \; l \; ; \; j++)$ do

         $\boldsymbol{\mu}_j^{\text{new}} = \frac{1}{n_j} \sum_{i=1,\; j=c_{\boldsymbol{x}_i}}^{n} \boldsymbol{x}_i$

   end for
   if stop criterion fulfilled then
     Stop=true
   end if
 end while
END Iteration
```

k-means algorithm

k-Means Clustering

k-means clustering

- fast
- robust (outliers)
- simple (advantage or disadvantage)
- prone to initialization

center near some outliers → center will stay on the outliers even if some cluster are not modeled

k-Means Clustering

membership continuous: (softmax)

$$p^b(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j) \;=\; \frac{\|\boldsymbol{x}_i \;-\; \boldsymbol{\mu}_j\|^{-2/(b-1)}}{\sum_{k=1}^{l} \|\boldsymbol{x}_i \;-\; \boldsymbol{\mu}_k\|^{-2/(b-1)}}$$

update rule

$$\boldsymbol{\mu}_j^{\mathrm{new}} \;=\; \frac{\sum_{i=1}^{n} p^b(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j)\; \boldsymbol{x}_i}{\sum_{i=1}^{n} p^b(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j)}$$

The following objective is minimized:

$$\sum_{j=1}^{l} \sum_{i=1}^{n} p^b(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j)\; \boldsymbol{x}_i \;\; \|\boldsymbol{x}_i \;-\; \boldsymbol{\mu}_j\|^2$$

This algorithm is called fuzzy k-means clustering

k-Means Clustering

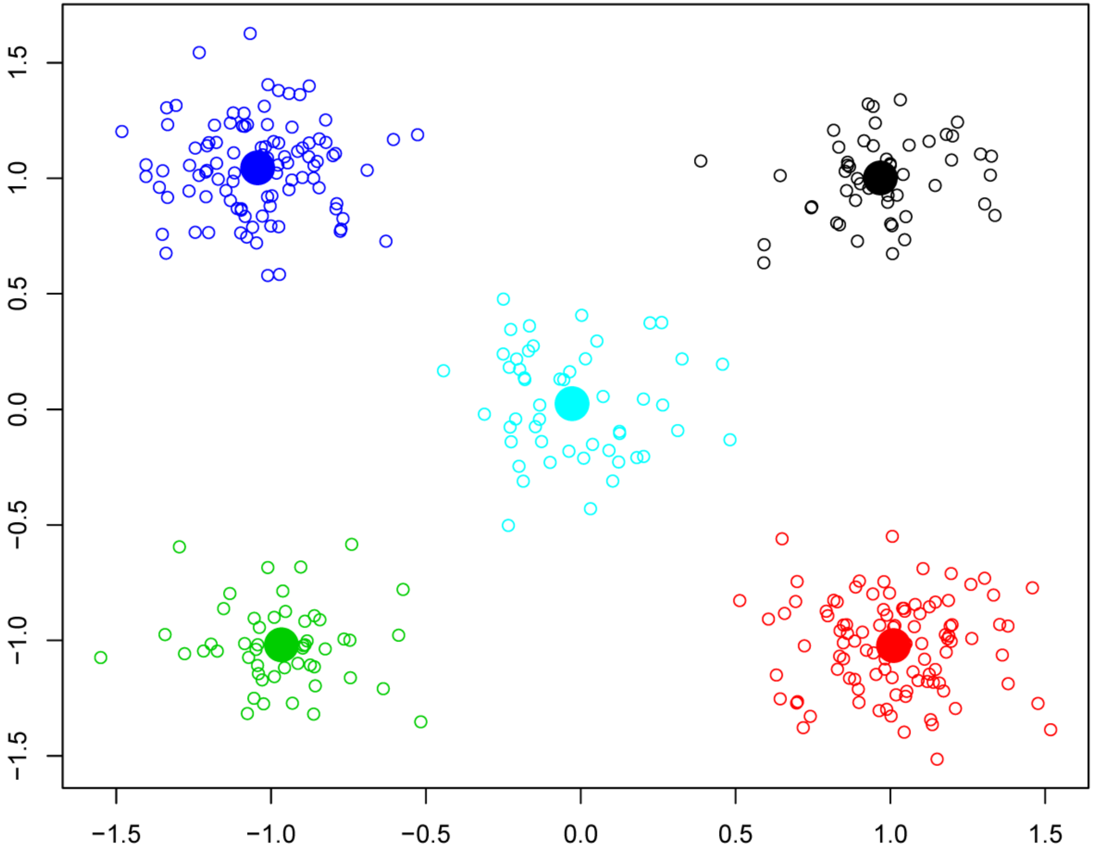
```
 Given: data $\{\boldsymbol{x}\} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$, number of clusters $l$, parameter $b$

BEGIN initialization
 initialize the cluster centers $\boldsymbol{\mu}_j$, $1 \leq j \leq l$, and $w_j(\boldsymbol{x}_i) = p(j \mid \boldsymbol{x}_i, \boldsymbol{\mu}_j)$ so
 that $\sum_{j=1}^{l} w_j(\boldsymbol{x}_i) = 1$, $w_j(\boldsymbol{x}_i) \geq 0$.
END initialization

BEGIN Iteration

 Stop=false
 while Stop=false do

     $\boldsymbol{\mu}_j^{\mathrm{new}} = \frac{\sum_{i=1}^{n} w_j(\boldsymbol{x}_i)\, \boldsymbol{x}_i}{\sum_{i=1}^{n} w_j(\boldsymbol{x}_i)}$

     $w_j(\boldsymbol{x}_i) = \frac{\| \boldsymbol{x}_i - \boldsymbol{\mu}_j \|^{-2/(b-1)}}{\sum_{k=1}^{l} \| \boldsymbol{x}_i - \boldsymbol{\mu}_k \|^{-2/(b-1)}}$

   if stop criterion fulfilled then
      Stop=true
   end if
 end while
END Iteration
```

fuzzy
k-means
algorithm

k-Means Clustering

artificial data set in two dimensions with five clusters

- color indicate cluster membership
- filled circles mark the cluster centers



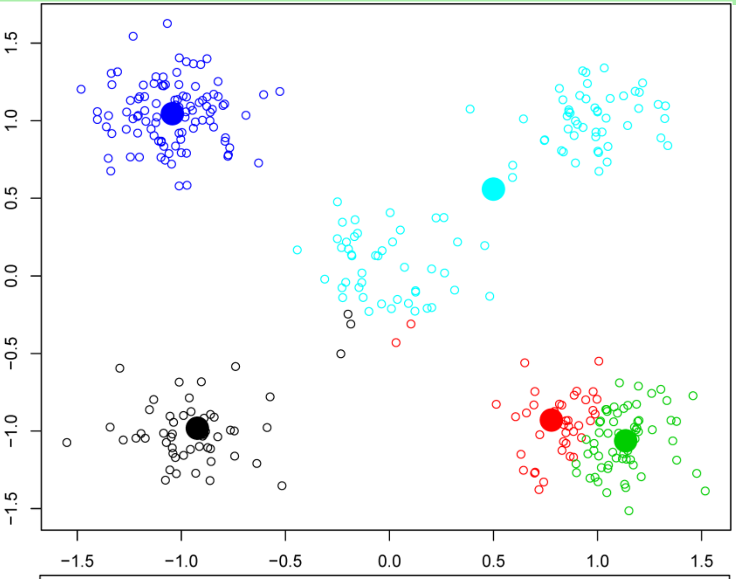
k-Means Clustering

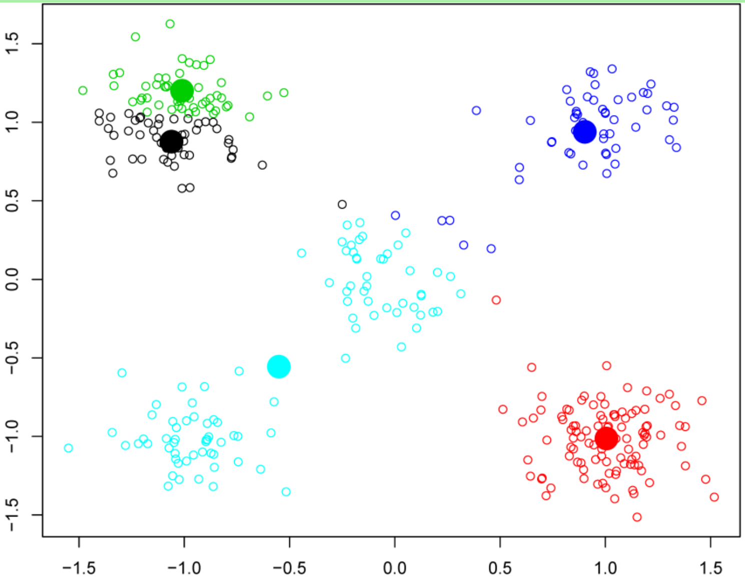
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

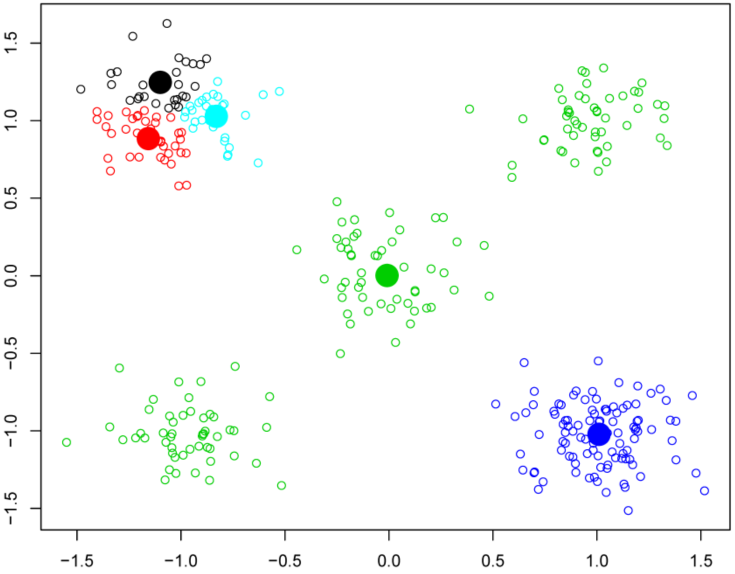
8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine







Local minima are shown:

- top row one cluster explains two true clusters while one true cluster is divided into two model clusters
- lower row: three model clusters share one true cluster

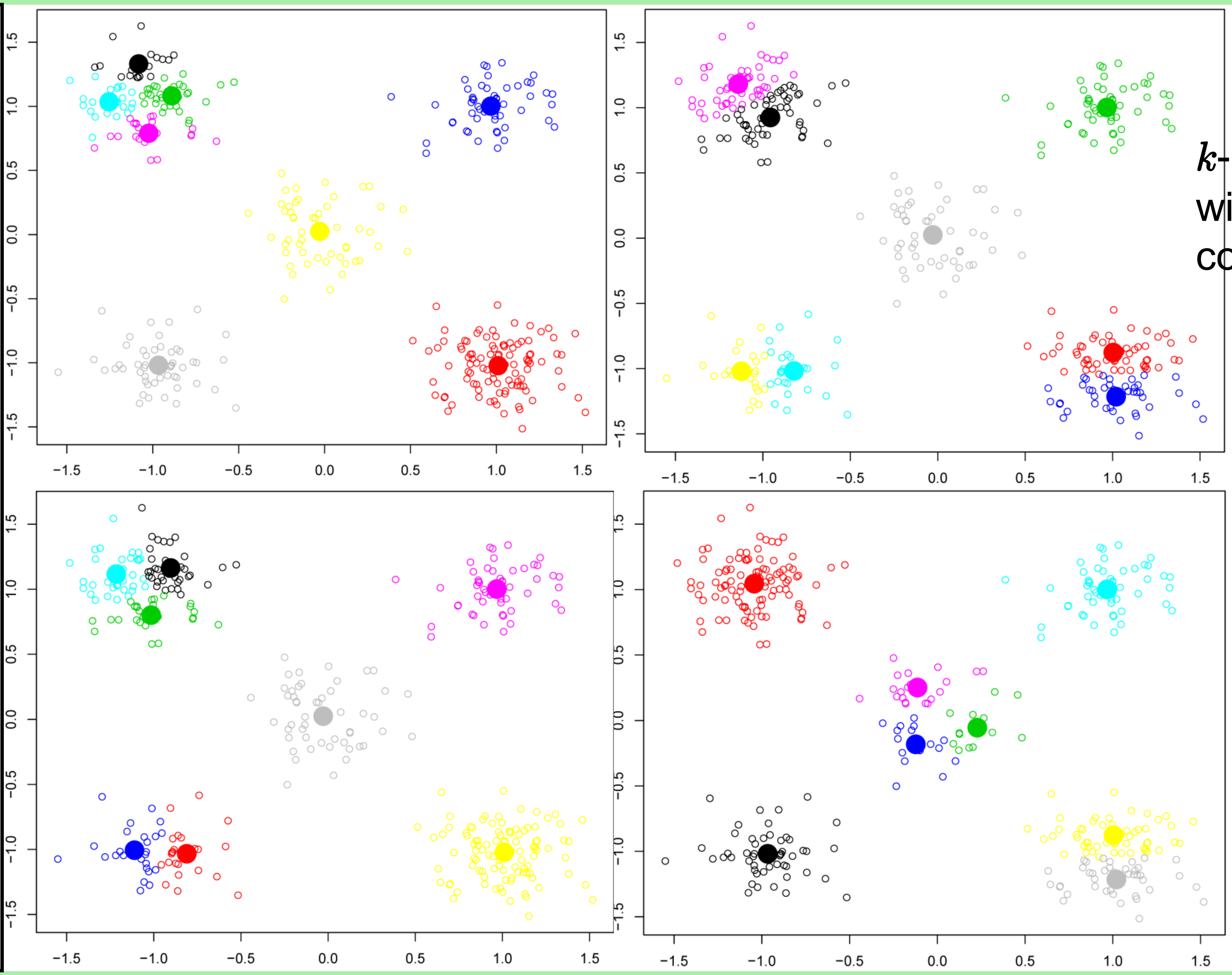
k-Means Clustering

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine



k-means with 8 comp.

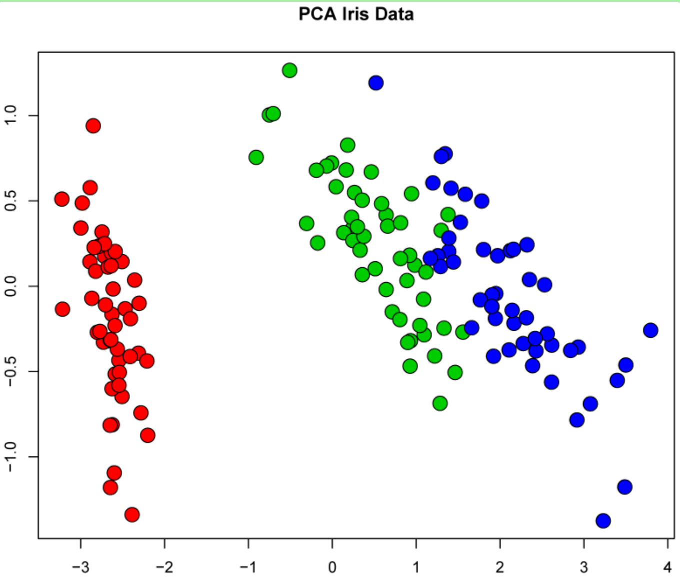
k-Means Clustering

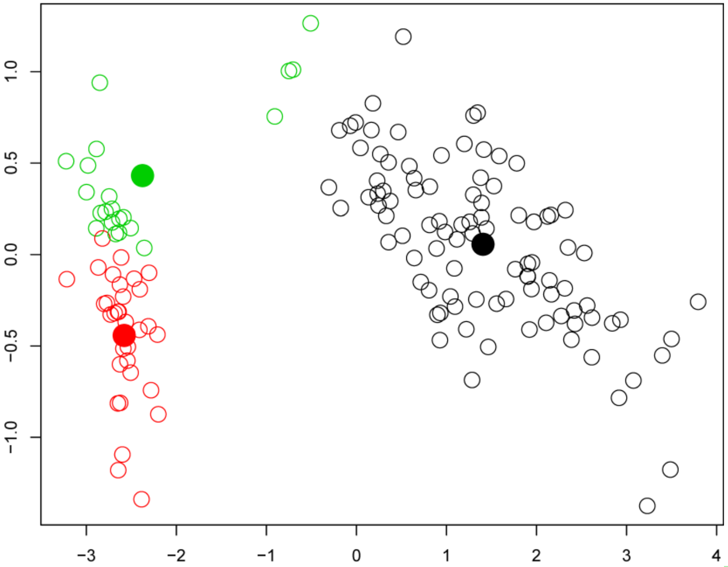
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

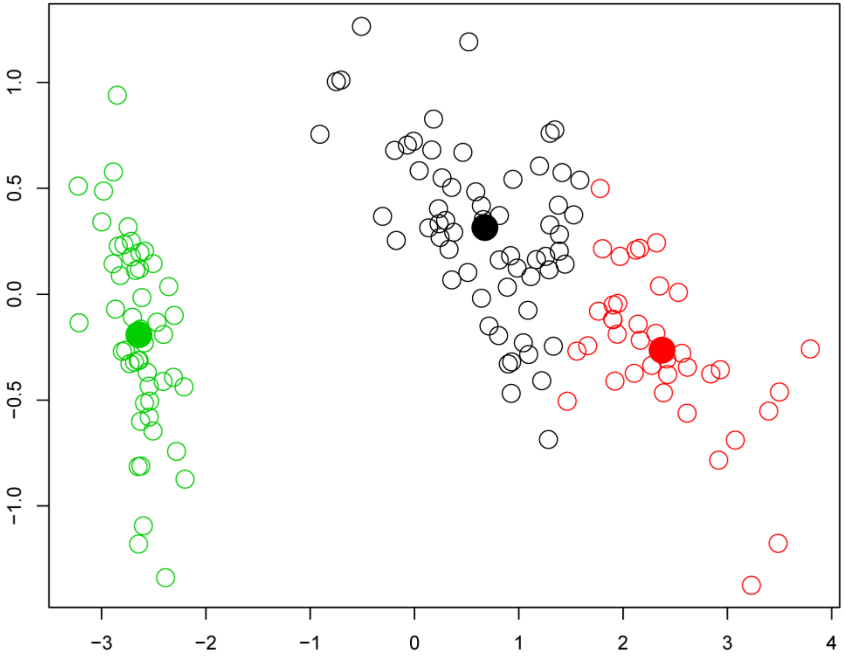
8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine







k-means applied to the Iris data

Upper right: typical quite good solution
only errors at class borders

Lower left: another typical solution
which is not as good

Can these solutions be distinguished?

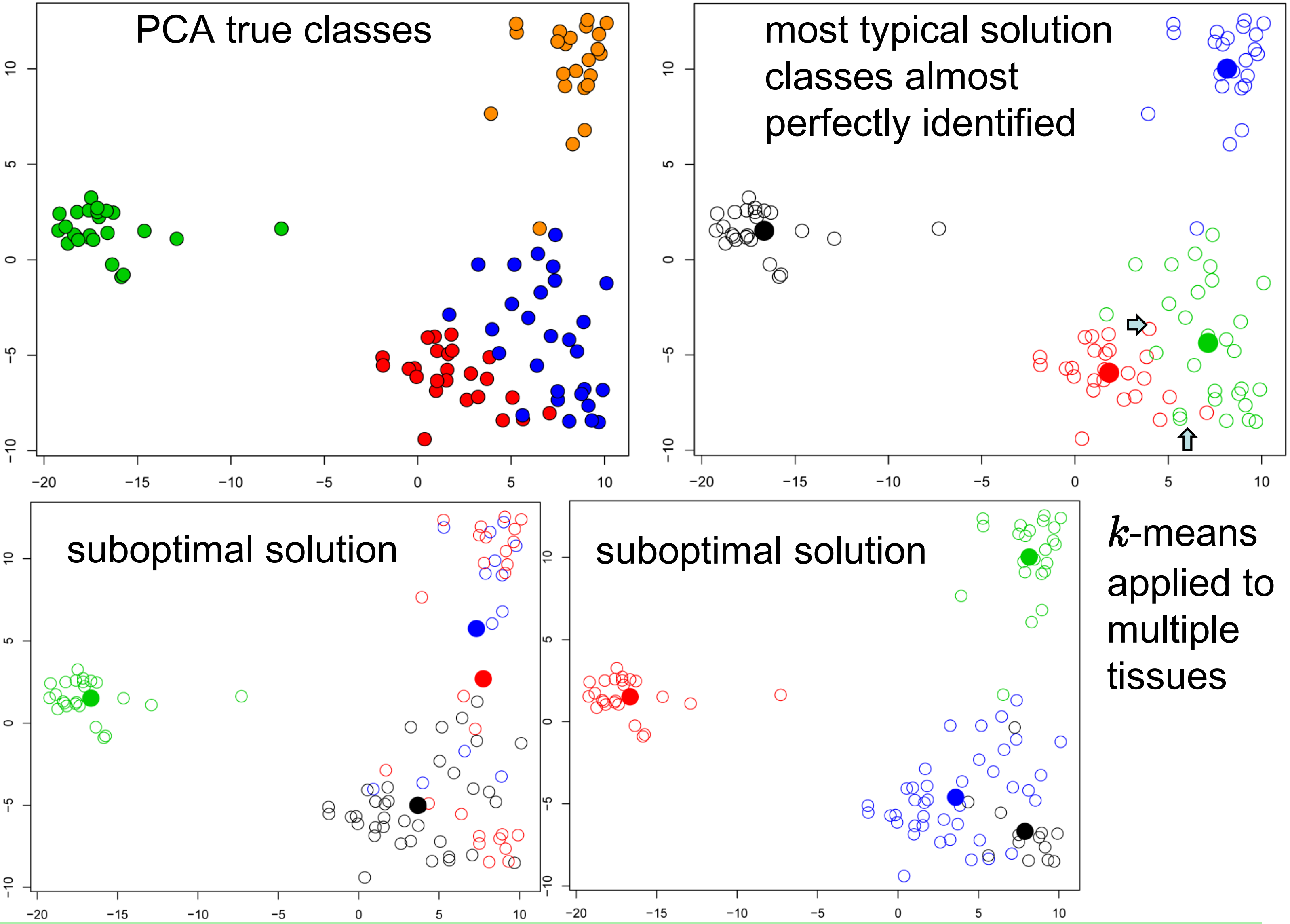
k-Means Clustering

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

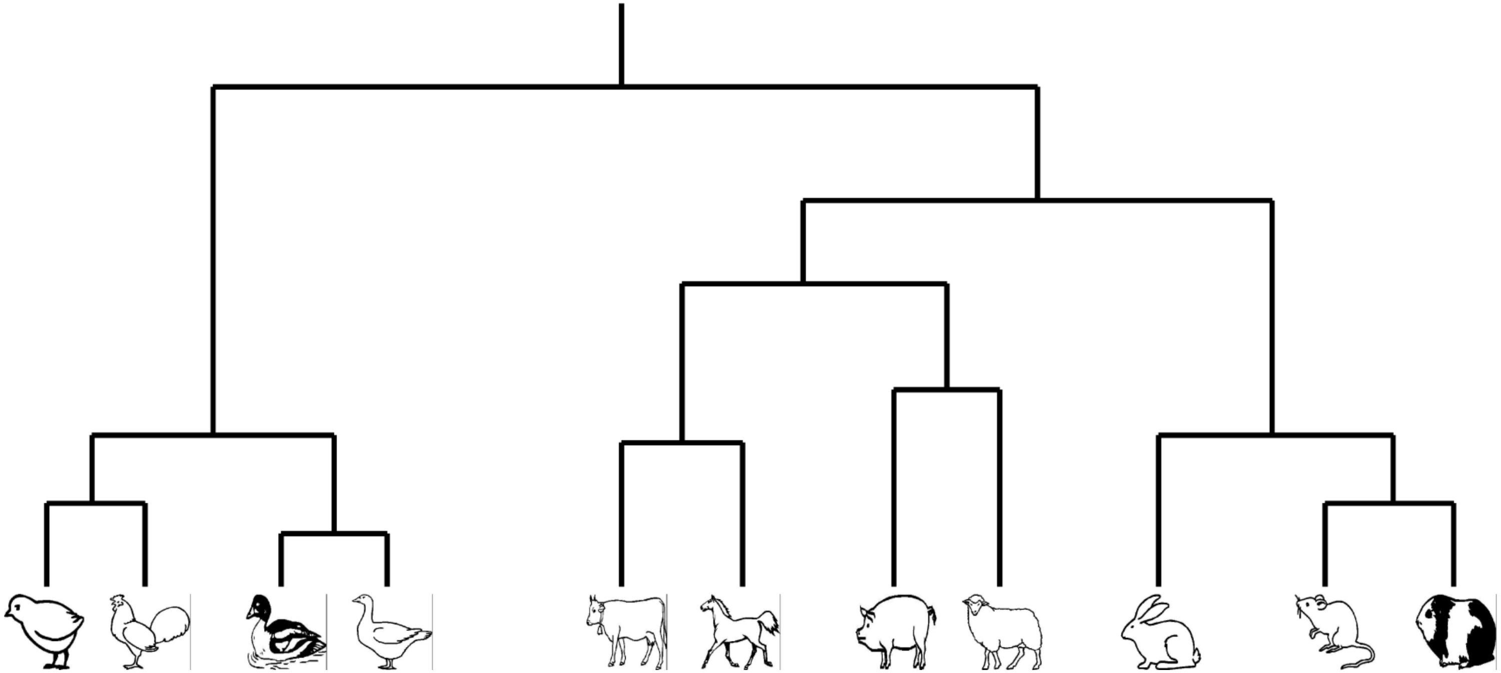
10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine



k-means applied to multiple tissues

Hierarchical Clustering

Hierarchical clustering supplies distances between clusters which are captured in a dendrogram. These distances allow to merge or cut clusters. Clustering is done agglomerative (bottom up) or divisive (top down)



Hierarchical Clustering: Cluster Distance = Linkage

agglomerative hierarchical clustering (bottom up) merges the closest clusters to new clusters.
Starts with clusters single observations and iteratively merges clusters.

different distance measures between clusters A and B are used:

$$\begin{aligned}
d_{\min}(A,B) &= \min_{\boldsymbol{a}\in A, \boldsymbol{b}\in B} \|\boldsymbol{a} - \boldsymbol{b}\| && \text{(single linkage)}\\
d_{\max}(A,B) &= \max_{\boldsymbol{a}\in A, \boldsymbol{b}\in B} \|\boldsymbol{a} - \boldsymbol{b}\| && \text{(complete linkage)}\\
d_{\mathrm{avg}}(A,B) &= \frac{1}{n_A \; n_B} \sum_{\boldsymbol{a}\in A} \sum_{\boldsymbol{b}\in B} \|\boldsymbol{a} - \boldsymbol{b}\| && \text{(average linkage)}\\
d_{\mathrm{mean}}(A,B) &= \|\bar{\boldsymbol{a}} - \bar{\boldsymbol{b}}\| && \text{(average linkage)}
\end{aligned}$$

where n_A (n_B) is the number of elements in A (B) and $\bar{\boldsymbol{a}}$ ($\bar{\boldsymbol{b}}$) is the mean of cluster A (B).
For the element distance $\|.\|$ any distance measure is possible like the Euclidean distance, the Manhattan distance, or the Mahalanobis distance.

Hierarchical Clustering: Linkage

single element clusters: distance measures are equivalent
For more elements in cluster:

- complete linkage $d_{\max}$ avoids that clusters are elongated in some direction (smallest distance between points remains small).
  → cluster may not be well separated.
- single linkage $d_{\min}$ ensures that each pair of elements from different clusters has a minimal distance. Single linkage clustering is relevant for leave-one-cluster-out cross-validation, which assumes that a whole new group of objects is unknown and left out.
- average linkage d_{avg} is “Unweighted Pair Group Method using arithmetic Averages” (UPGMA)

Divisive or top down clustering is often based on graph theoretic considerations. First the **minimal spanning tree** is built.
Then the **largest edge is removed** which gives two clusters.
Now the second largest edge can be removed and so on.
It might be more appropriate to compute the average edge length within a cluster and find the edge which is considerably larger than other edges in the cluster.

Hierarchical Clustering: Examples

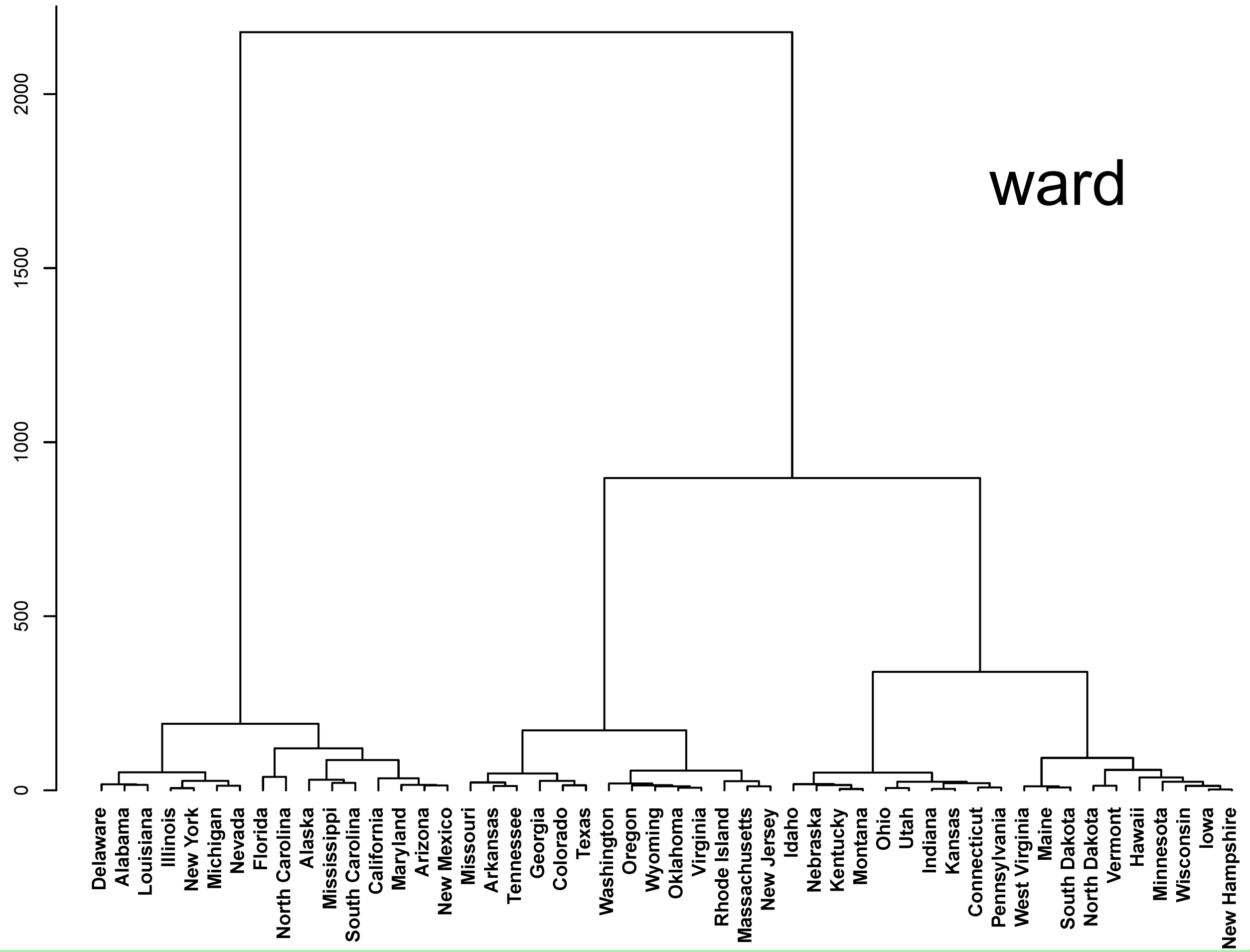
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

hierarchical clustering of the US Arrest data distance measures “ward”, “single”, “complete”, “average”, “mcquitty”, “median”, and “centroid



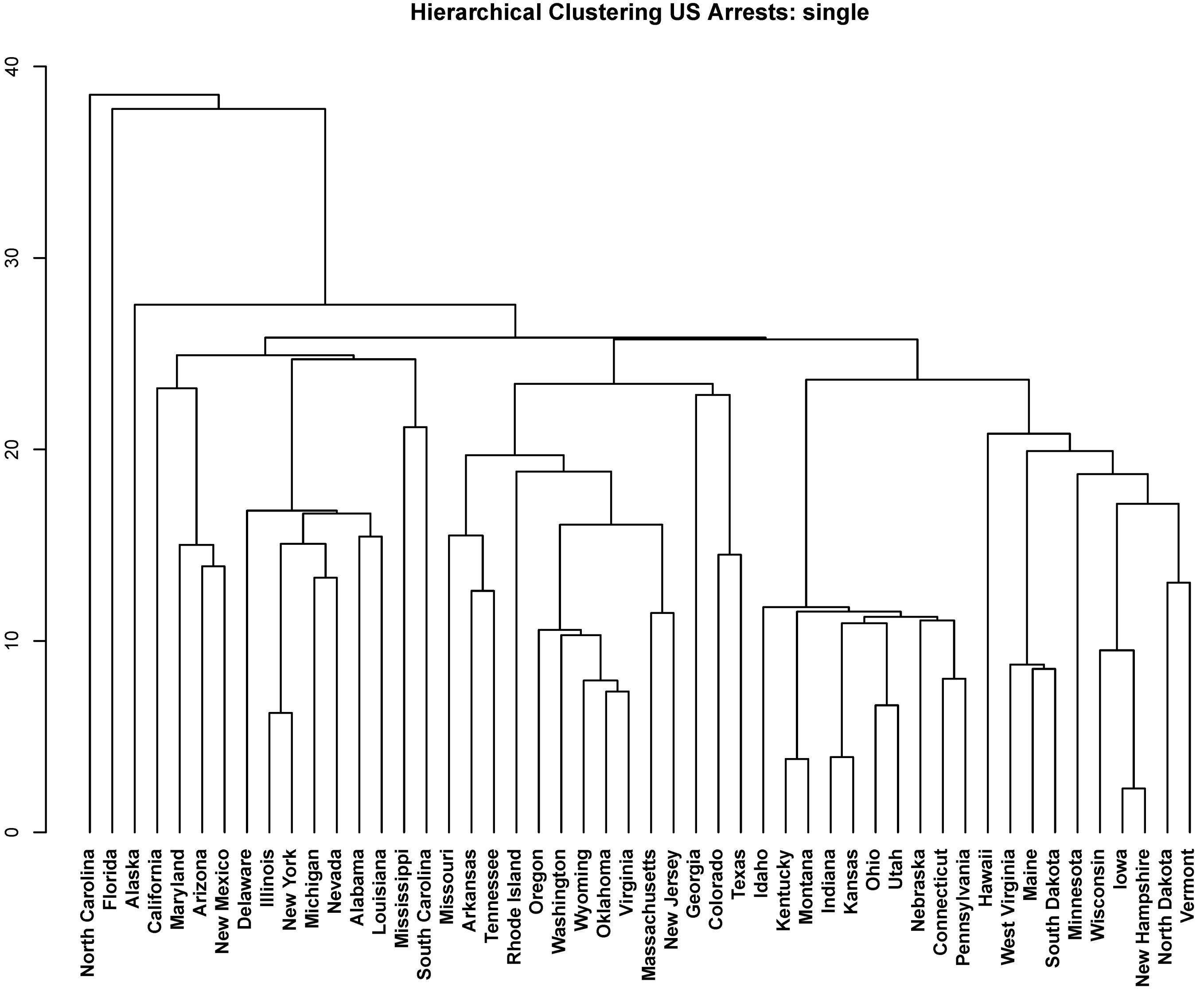
Hierarchical Clustering: Examples

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine


Hierarchical Clustering US Arrests: single

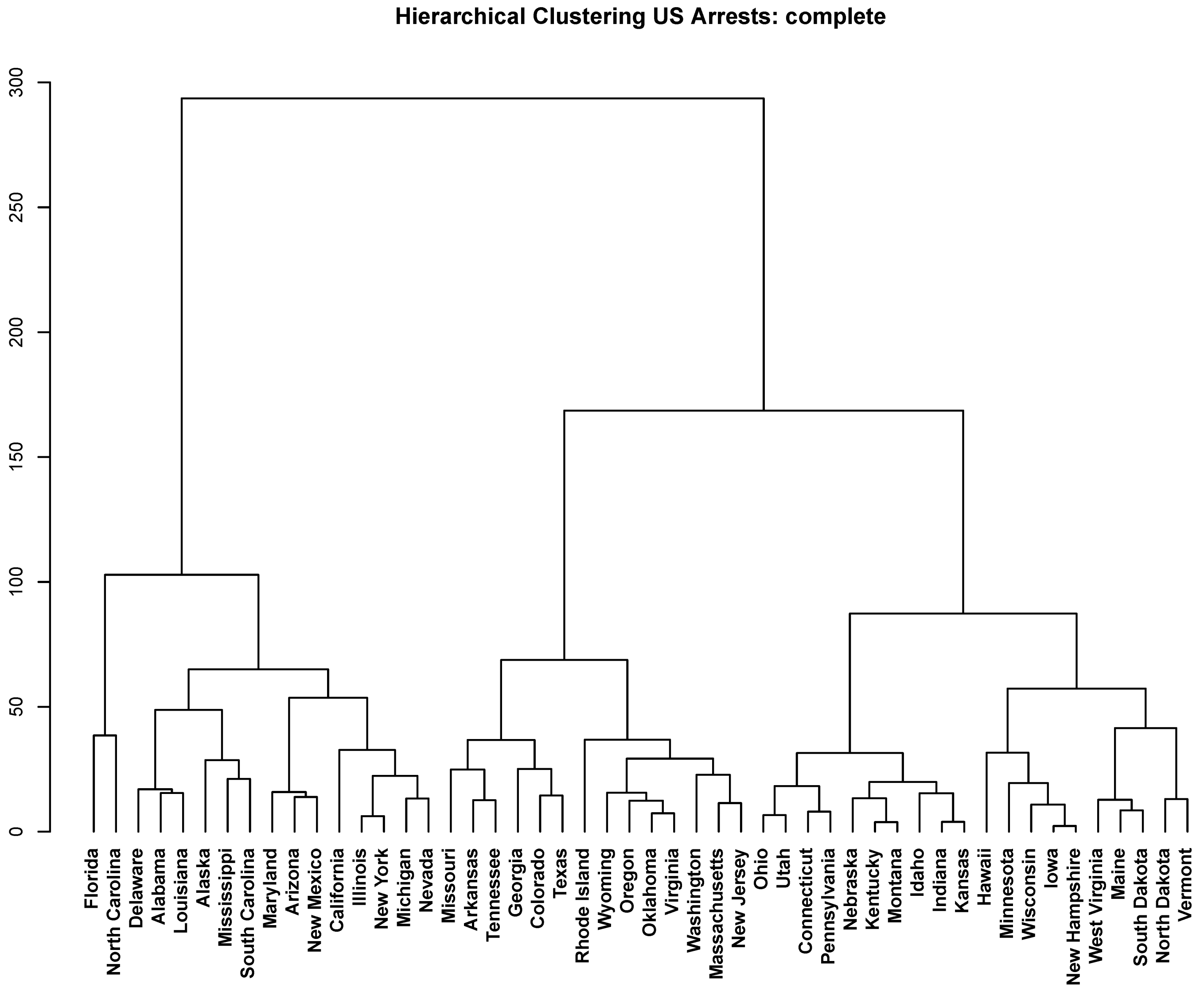
Hierarchical Clustering: Examples

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine



Hierarchical Clustering: Examples

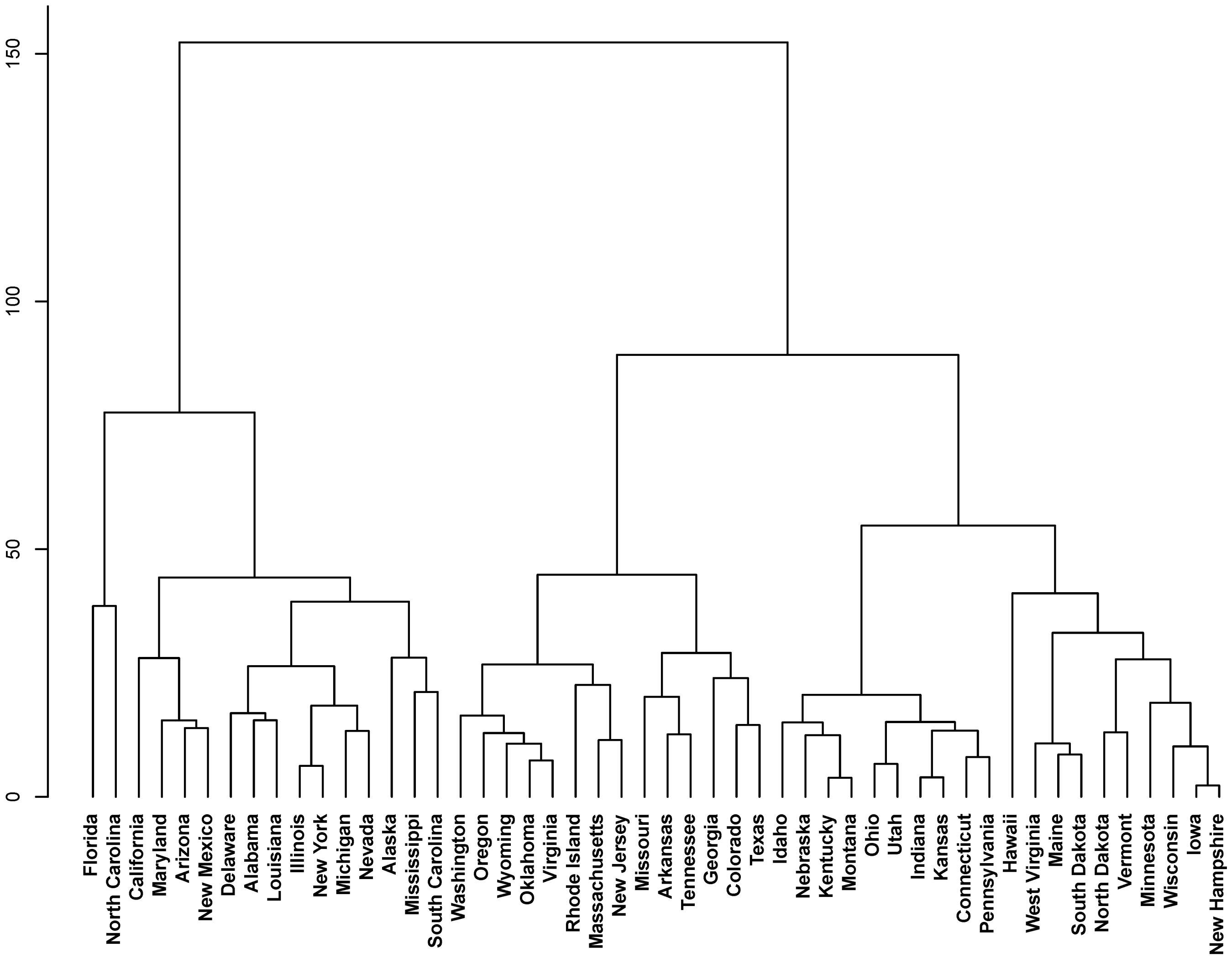
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

Hierarchical Clustering US Arrests: average



Hierarchical Clustering: Examples

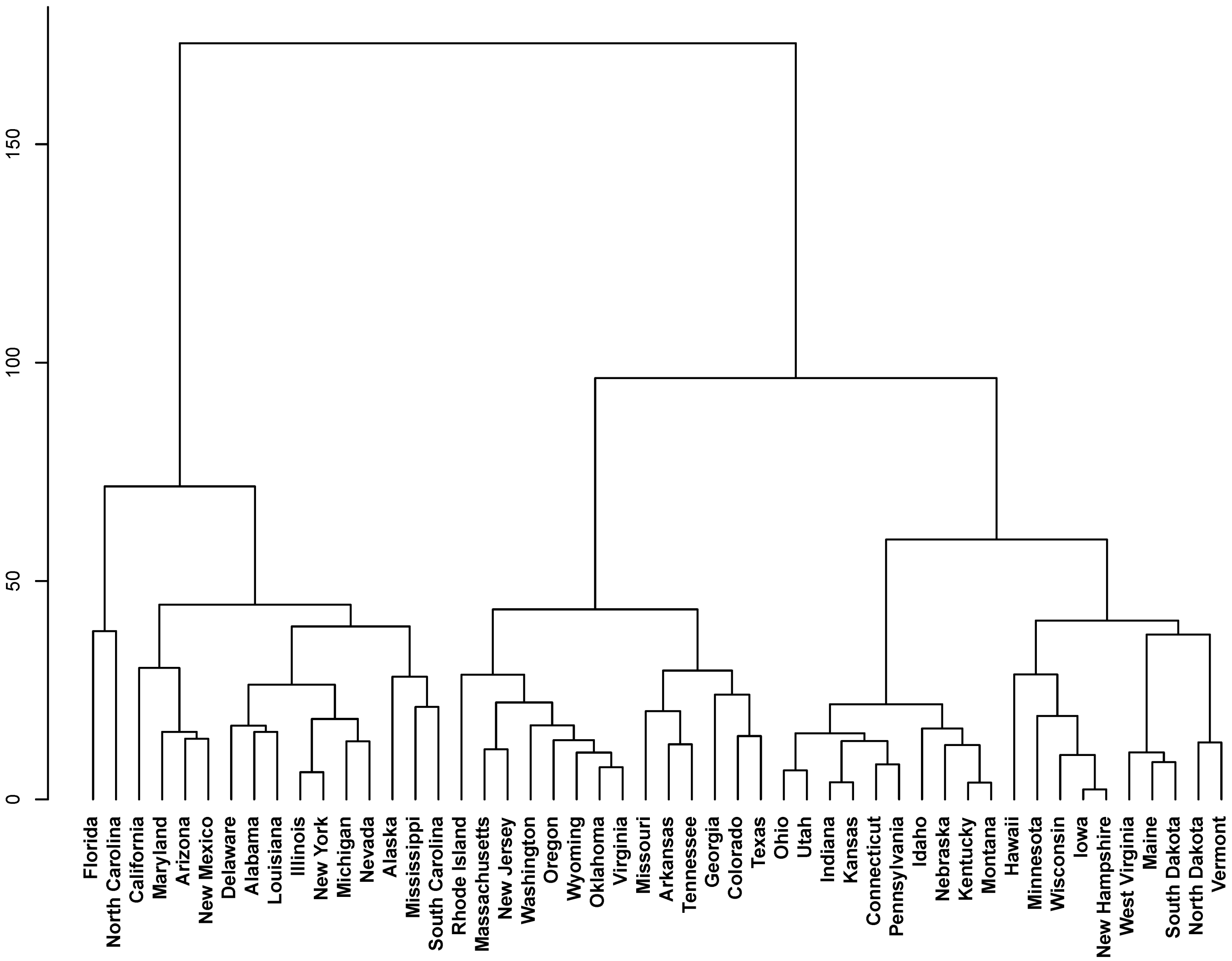
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

Hierarchical Clustering US Arrests: mcquitty



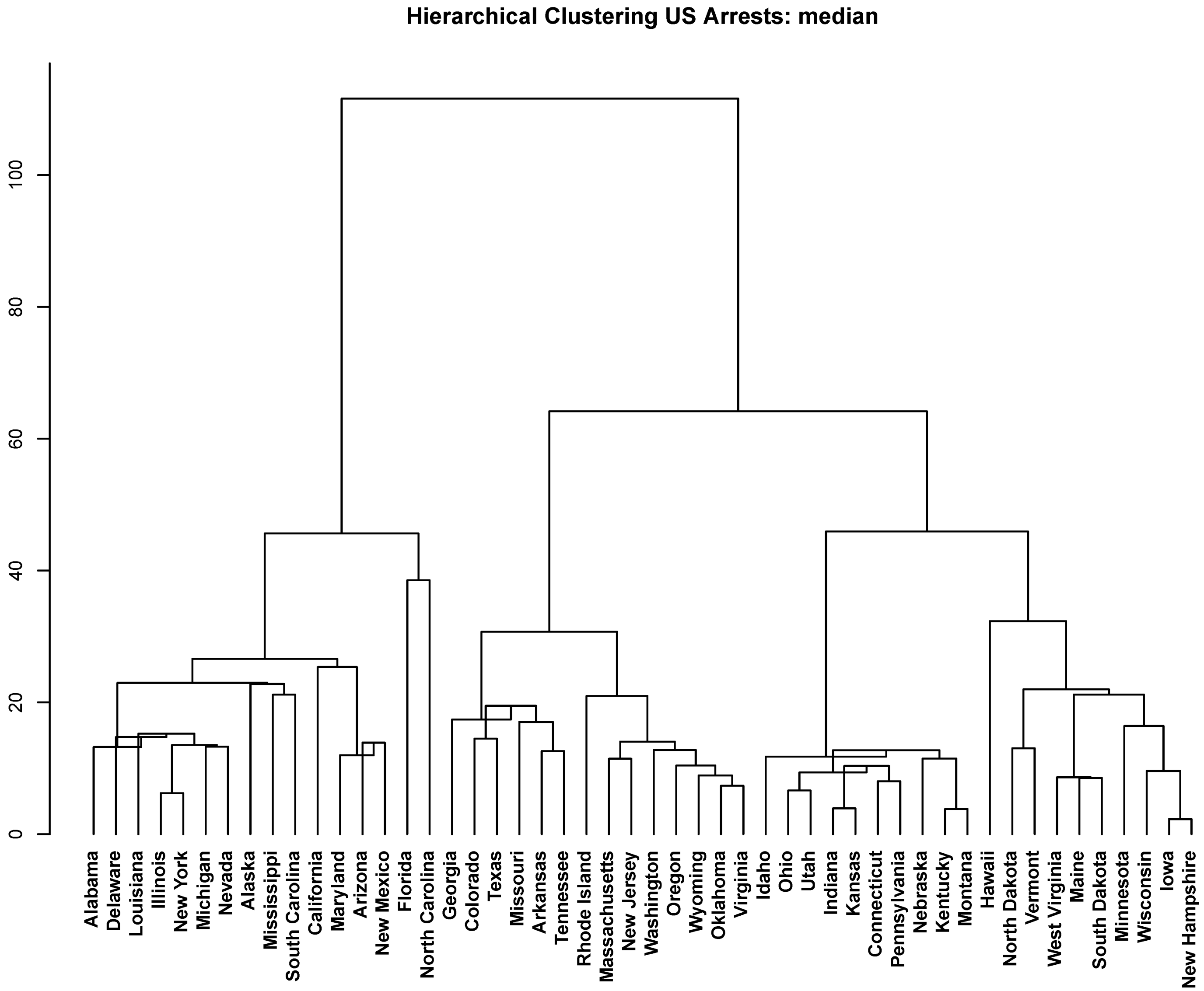
Hierarchical Clustering: Examples

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine



Hierarchical Clustering US Arrests: median

Hierarchical Clustering: Examples

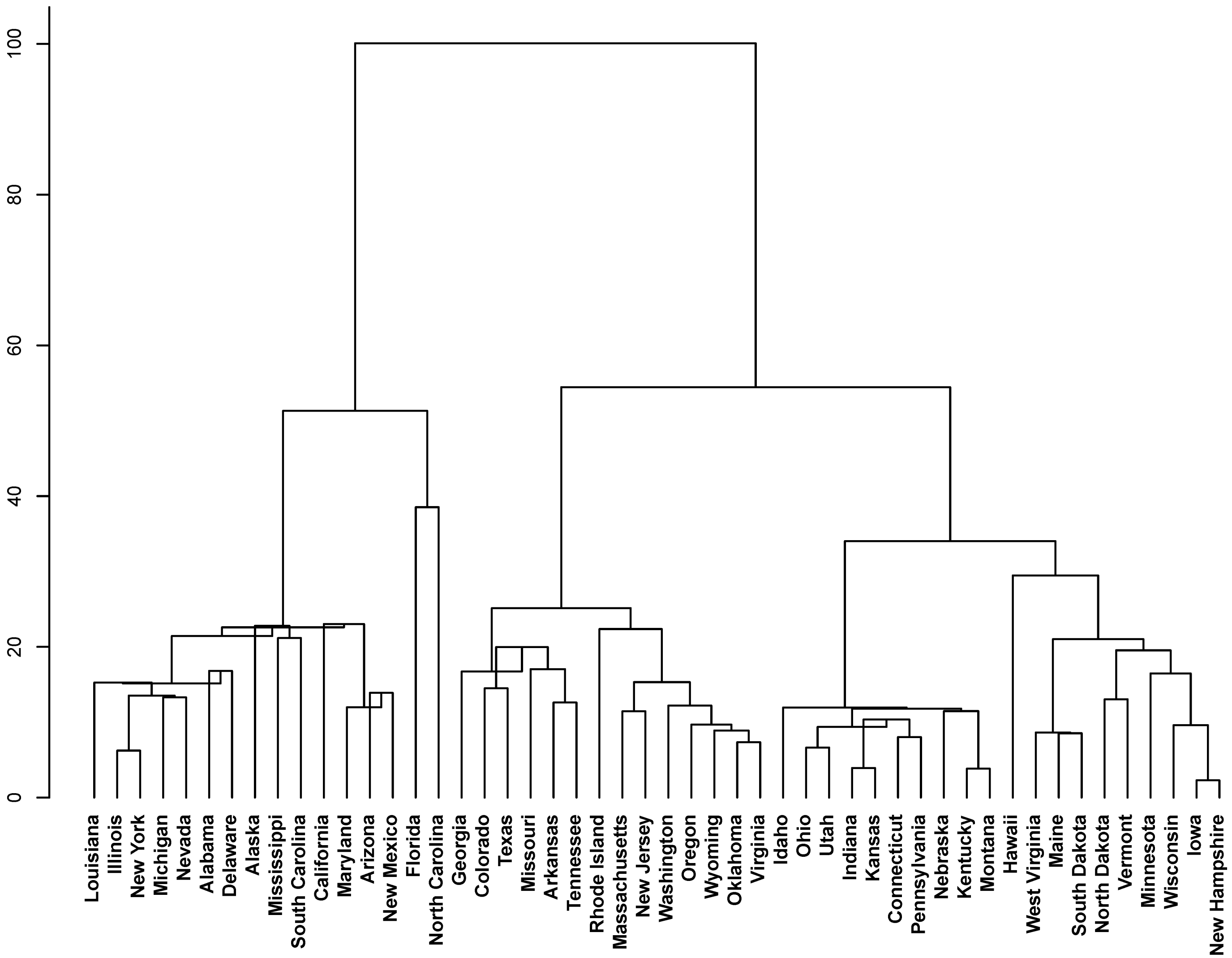
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

Hierarchical Clustering US Arrests: centroid



Hierarchical Clustering: Examples

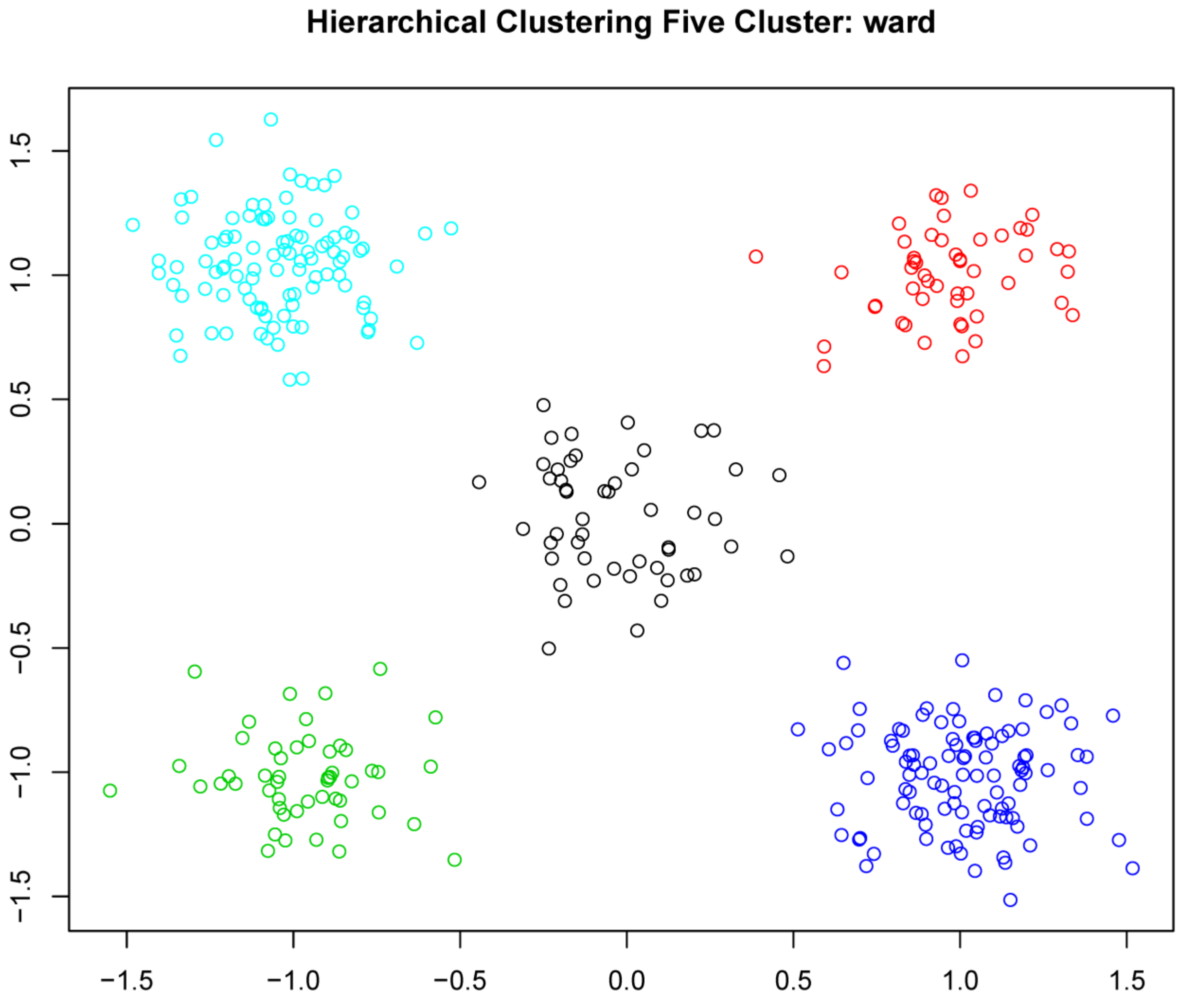
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

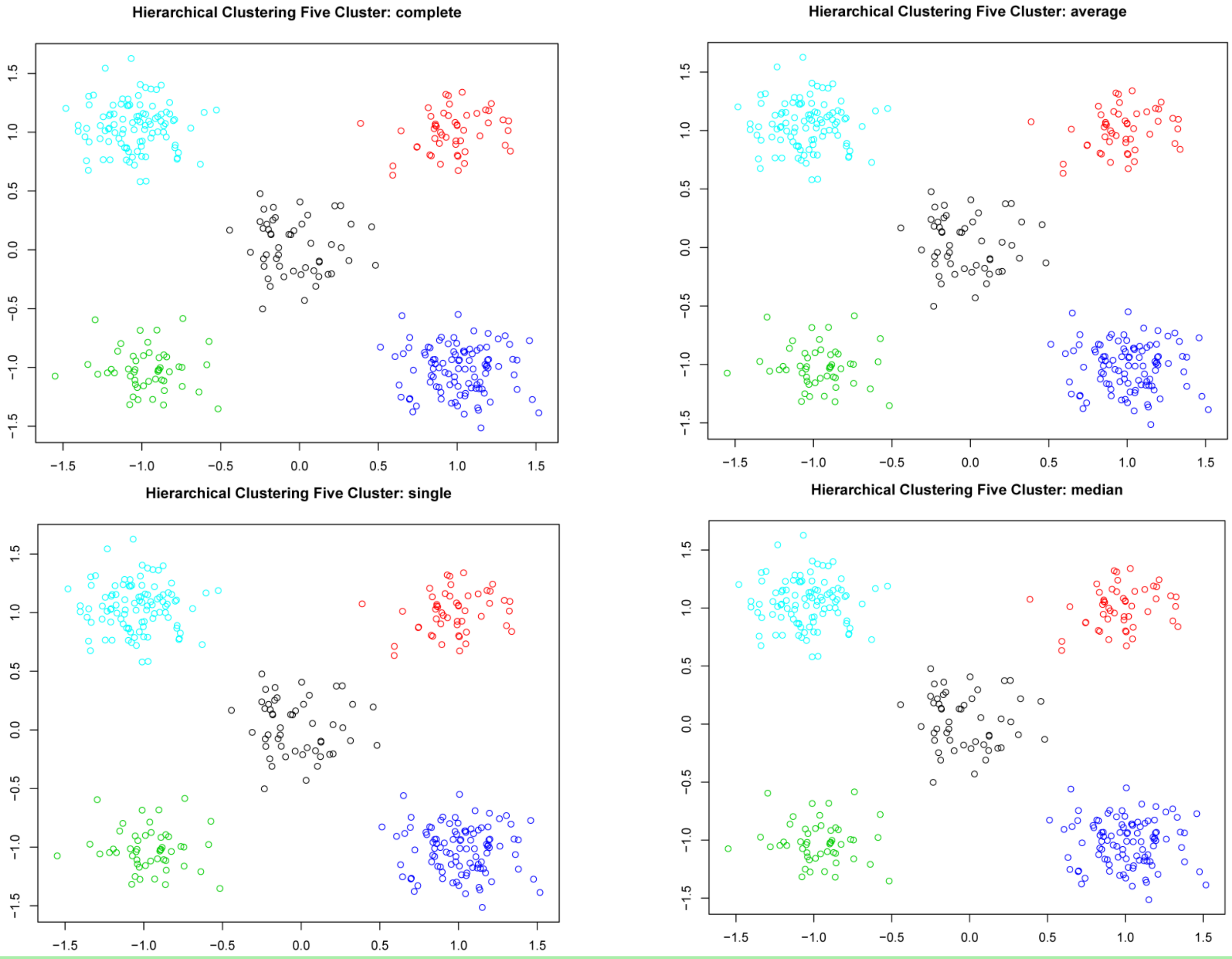
10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

hierarchical clustering of the five cluster data set: Ward's distance is perfect. To determine the clusters, the dendrogram was cut



Hierarchical Clustering: Examples

hierarchical clustering of the five cluster data set



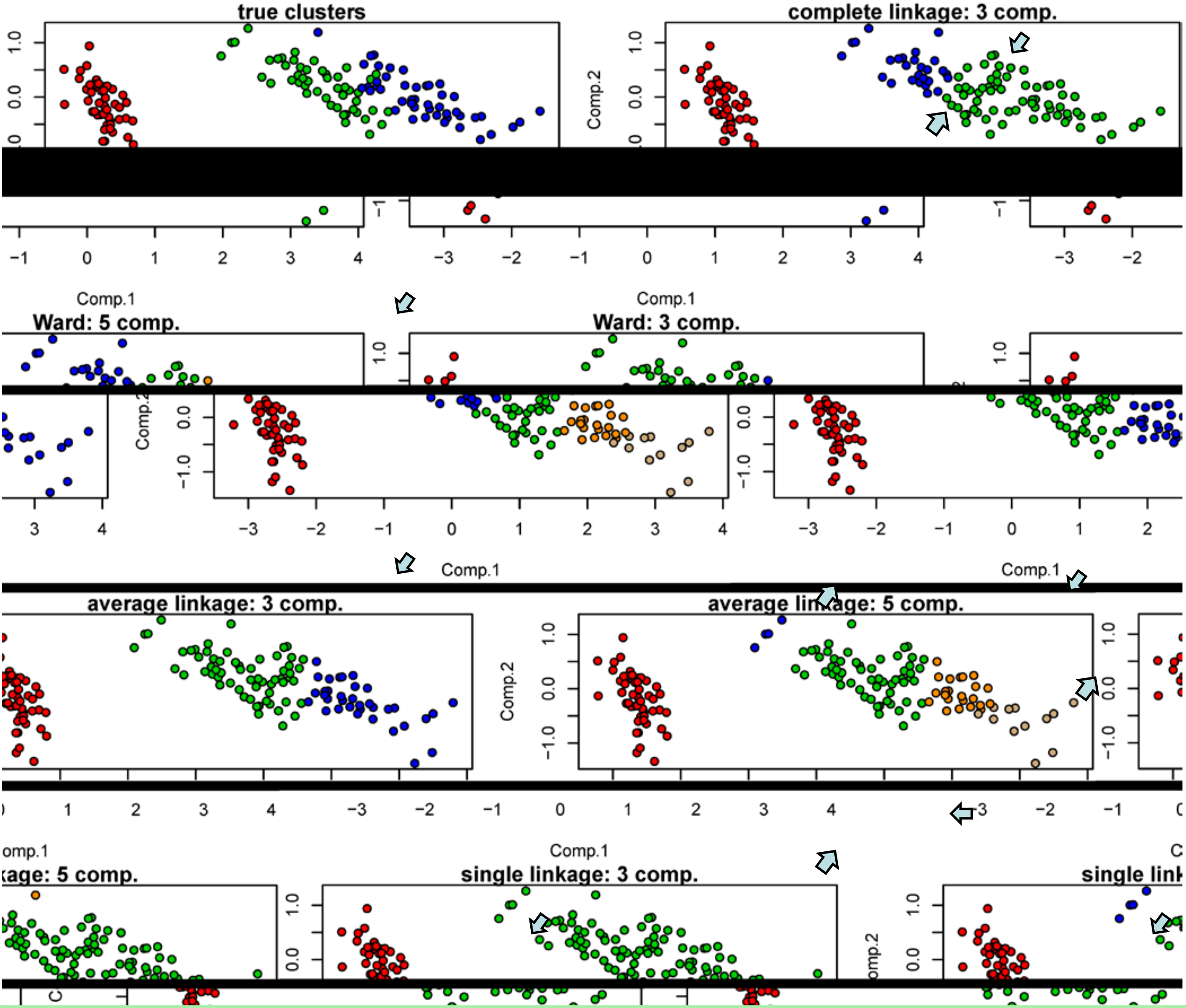
Hierarchical Clustering Five Cluster: complete

Hierarchical Clustering Five Cluster: average

Hierarchical Clustering Five Cluster: single

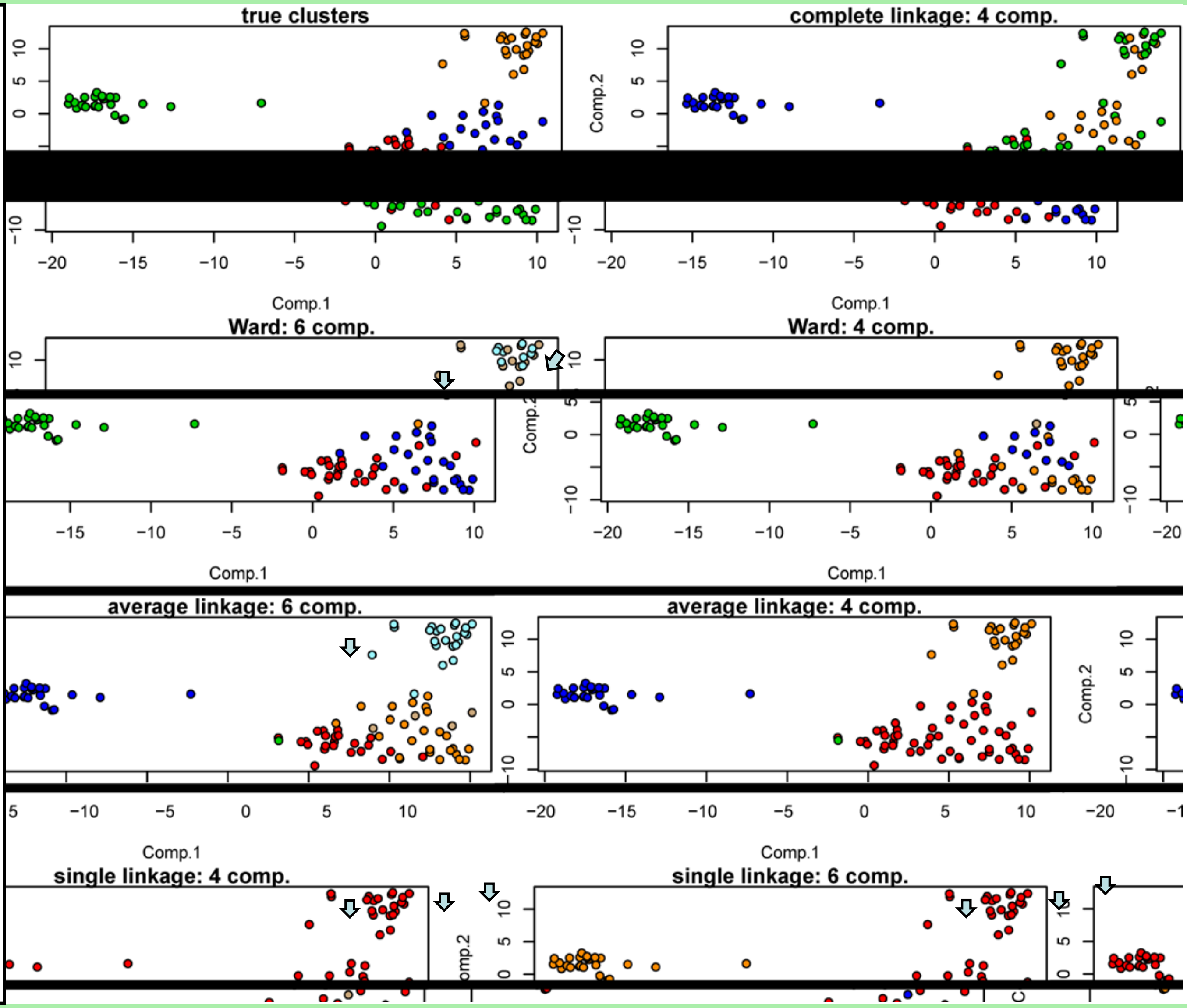
Hierarchical Clustering Five Cluster: median

Hierarchical Clustering: Examples



hierarchical clustering of the iris data set

Hierarchical Clustering: Examples



hierarchical clustering of the multiple tissue data

Similarity-Based Clustering

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

Similarity-based clustering uses similarities between objects but does not require to represent the objects via feature vectors

Similarities:

- links in the web domain
- interactions of humans (facebook)
- co-occurrences of objects (co-expression of genes or co-citations)
- spacial distances (cities on a map or atoms in a molecule)
- co-processing (compressing two documents or sorting two sets)
- alignment of two sequences
- alignment of two structures

Similarity-Based Clustering: Aspect Model

The aspect model considers discrete data of observations that are pairs (x, y), which are counted.

Example: “person x buys product y” or “person x participates in y” Applications are document-word or sample-gene relations.

the model is $$p(x,y) \ = \ \sum_z p(z) \ p(x \mid z) \ p(y \mid z)$$

class variable: $z \in \{z_1, \ldots, z_l\}$
probability of the observation: $p(x,y)$

Model assumption: x and y are independent conditioned on z

$$p(x,y) \ = \ \sum_z p(x,y,z) \ = \ \sum_z p(z) \ p(x,y \mid z) \ = \ \sum_z p(z) \ p(x \mid z) \ p(y \mid z)$$

class conditional probabilities: $p(x \mid z)$ and $p(y \mid z)$

hidden factor z has an effect on the occurrence of both x and y

Similarity-Based Clustering: Aspect Model

John Paulos' example in ABCNews.com:
“Consumption of hot chocolate is correlated with low crime rate, but both are responses to cold weather.”

- x= consumption of hot chocolate
- y= crime rate
- z= cold weather

The maximum likelihood model parameters $p(x \mid z)$ and $p(y \mid z)$ can be estimated by an EM algorithm:

E-step

$$p(z \mid x, y) \ = \ \frac{p(z)\ p(x \mid z)\ p(y \mid z)}{\sum_{z'} p(z')\ p(x \mid z')\ p(y \mid z')}$$

M-step

$$p(z) \ = \ \sum_{x,y} n(x,y)\ p(z \mid x,y)$$

$$p(y \mid z) \ = \ \frac{\sum_x n(x,y)\ p(z \mid x,y)}{p(z)}$$

$$p(x \mid z) \ = \ \frac{\sum_y n(x,y)\ p(z \mid x,y)}{p(z)}$$

$n(x,y)$is the count of observations (x,y), that is, the row of x and the column of y in the data matrix

Similarity-Based Clustering: Aspect Model

For the aspect model, clustering of the x can be based on

$$\begin{aligned} p(z \mid x) &= \frac{p(x \mid z)\ p(z)}{p(x)} \\ p(x) &= \sum_{z,y} p(z)\ p(x \mid z)\ p(y \mid z) \end{aligned}$$

z indicates the cluster, that is, each z represents one cluster

Analog formulas are obtained for clustering y or pairs (x, y)

Similarity-Based Clustering: Affinity Propagation

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

Affinity propagation is a similarity-based clustering method that is also exemplar-based clustering

exemplar-based clustering enforces cluster centers to be data points, the “prototypes” or “exemplars”

exemplar-based clustering is the k-centers clustering which starts with an initial set of randomly selected exemplars and iteratively refines this set so as to decrease the sum of squared errors.

- only for small number of clusters
- good initialization required

Affinity propagation overcomes the problems of k-centers clustering

Similarity-Based Clustering: Affinity Propagation

- similarities between object i and object k: $s(i,k)$
- preferences: $s(k,k)$ how likely object k becomes an exemplar
- responsibilities: $r(i,k)$ messages sent from object i to candidate exemplar k. responsibility reflects the evidence that k serves as an exemplar for i: how well can k represent the object i.
- availabilities: $a(i,k)$ messages sent from candidate exemplar k to object i. Availability reflects the evidence that k is indeed an exemplar.

initialization: $a(i,k) \ = \ 0$

updates:

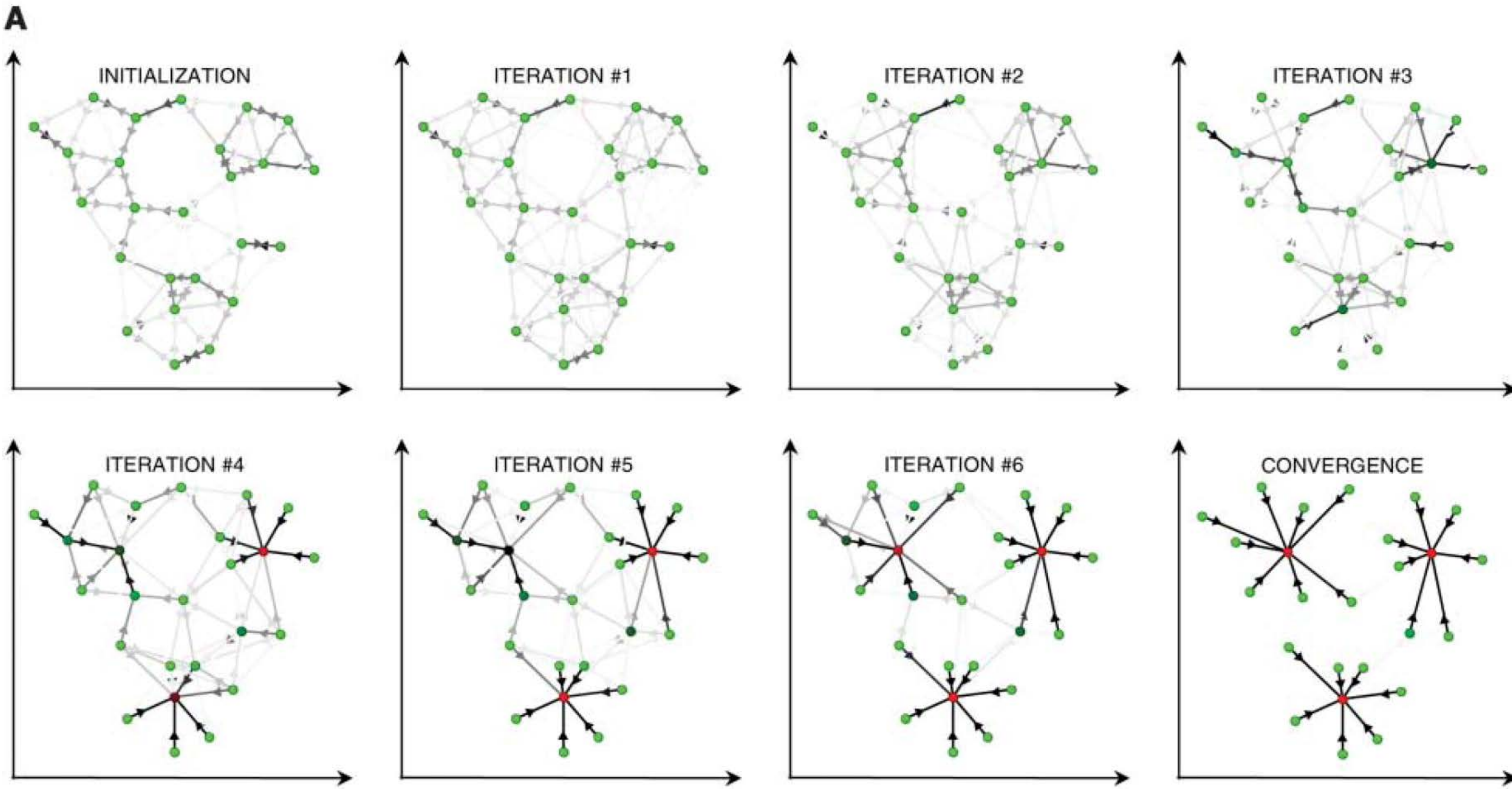
$$r(i,k) \ = \ s(i,k) \ - \ \max_{k',k' \neq k} \{a(i,k') \ + \ s(i,k')\}$$

$$a(i,k) \ = \ \min \left\{ 0 \ , \ r(k,k) \ + \ \sum_{i',i' \notin \{i,k\}} \max\{0 \ , \ r(i',k)\} \right\}$$

$$a(k,k) \ = \ \sum_{i',i' \neq k} \max\{0 \ , \ r(i',k)\}$$

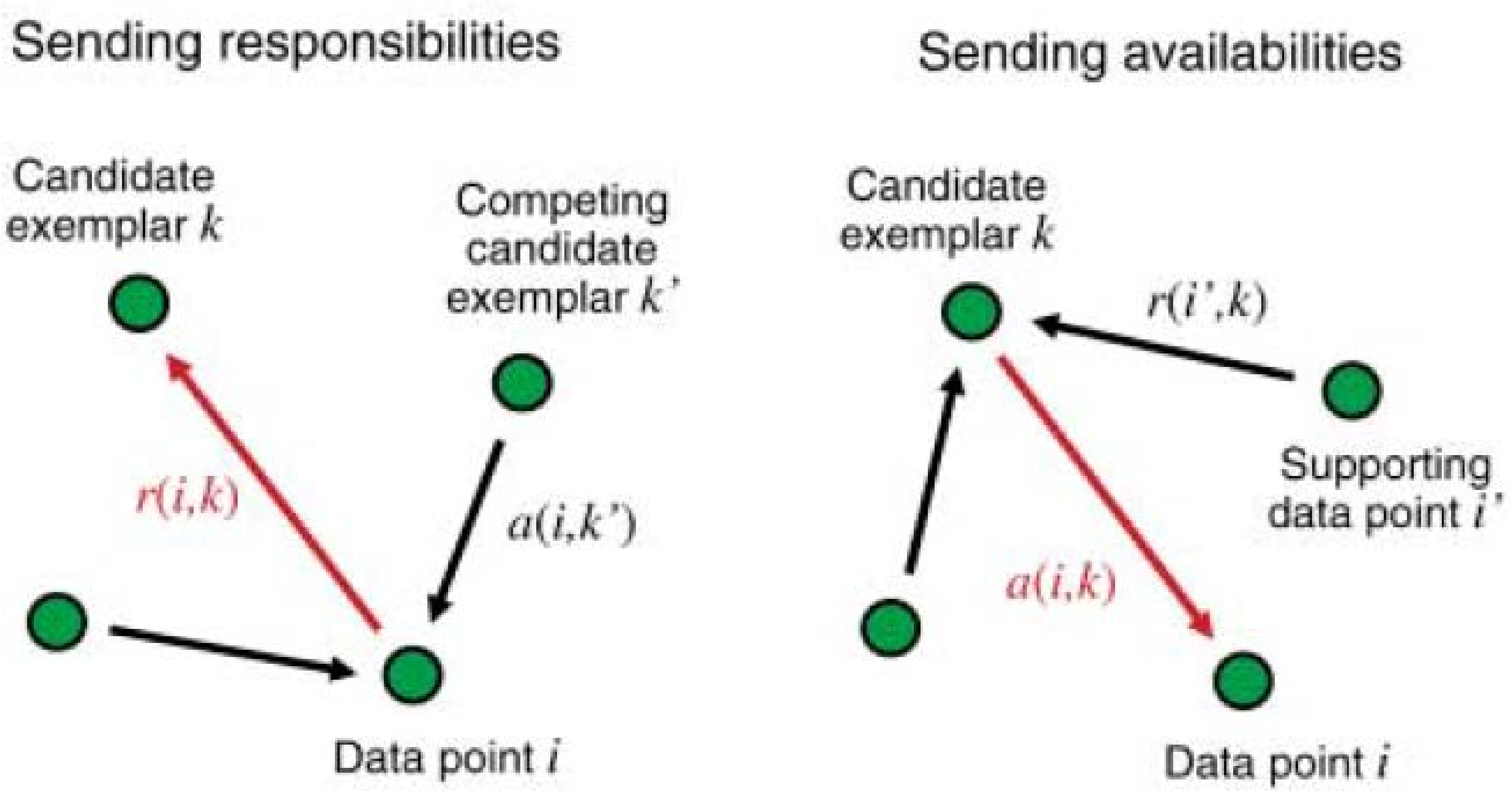
Similarity-Based Clustering: Affinity Propagation

different iterations of affinity propagation and messages sent



Similarity-Based Clustering: Affinity Propagation

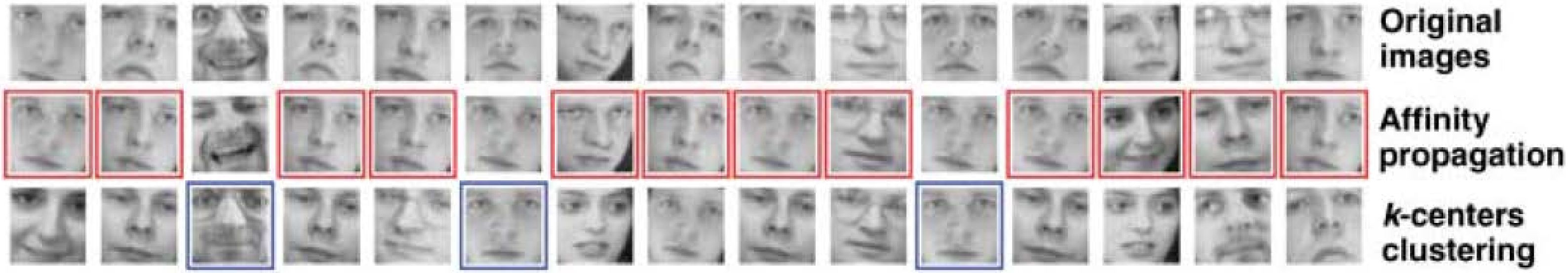
specific messages passing in the algorithm of affinity propagation



Similarity-Based Clustering: Affinity Propagation

Example of affinity propagation with images of faces

The 15 images with highest squared error under either affinity propagation or k-centers clustering are shown in the top row.
The middle and bottom rows show the exemplars assigned by the two methods, and the boxes show which of the two methods performed better for that image, in terms of squared error.
Affinity propagation found higher-quality exemplars.



Similarity-Based Clustering: Affinity Propagation

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

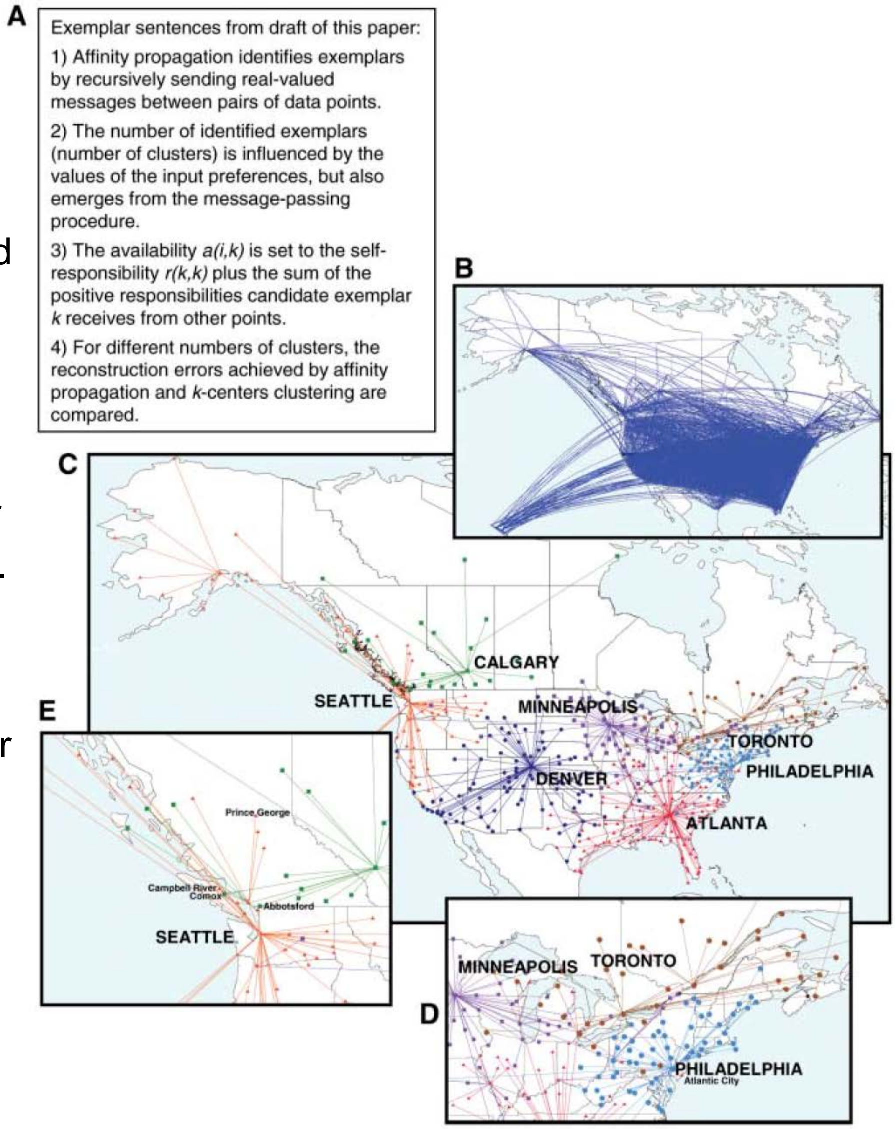
(A) Similarities between pairs of sentences in the AP manuscript were constructed by matching words. Four identified exemplars are shown.

(B) AP was applied to similarities derived from air-travel efficiency between the 456 busiest commercial airports in Canada and the United States.

(C) Seven exemplars identified by AP are color-coded, and the assignments of other cities to these exemplars is shown.

(D) The inset shows that the Canada-USA border roughly divides the Toronto and Philadelphia clusters, due to a larger availability of domestic flights vs. international flights.

(E) The west coast has extraordinarily frequent airline service between Vancouver and Seattle connects Canadian cities in the northwest to Seattle.



Exemplar sentences from draft of this paper:

1) Affinity propagation identifies exemplars by recursively sending real-valued messages between pairs of data points.

2) The number of identified exemplars (number of clusters) is influenced by the values of the input preferences, but also emerges from the message-passing procedure.

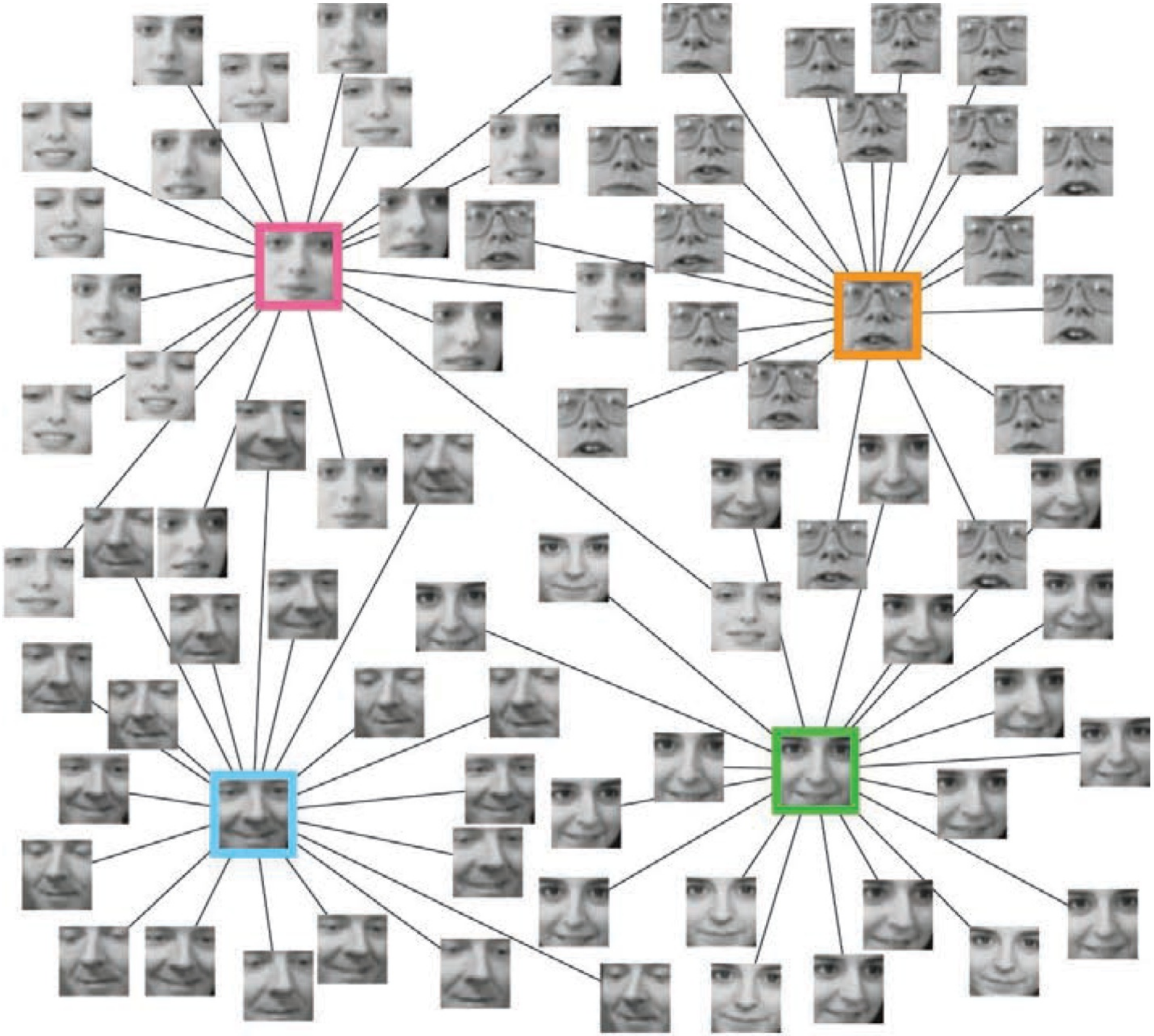
3) The availability *a(i,k)* is set to the self-responsibility *r(k,k)* plus the sum of the positive responsibilities candidate exemplar *k* receives from other points.

4) For different numbers of clusters, the reconstruction errors achieved by affinity propagation and *k*-centers clustering are compared.

Similarity-Based Clustering: Affinity Propagation

affinity propagation applied to face images

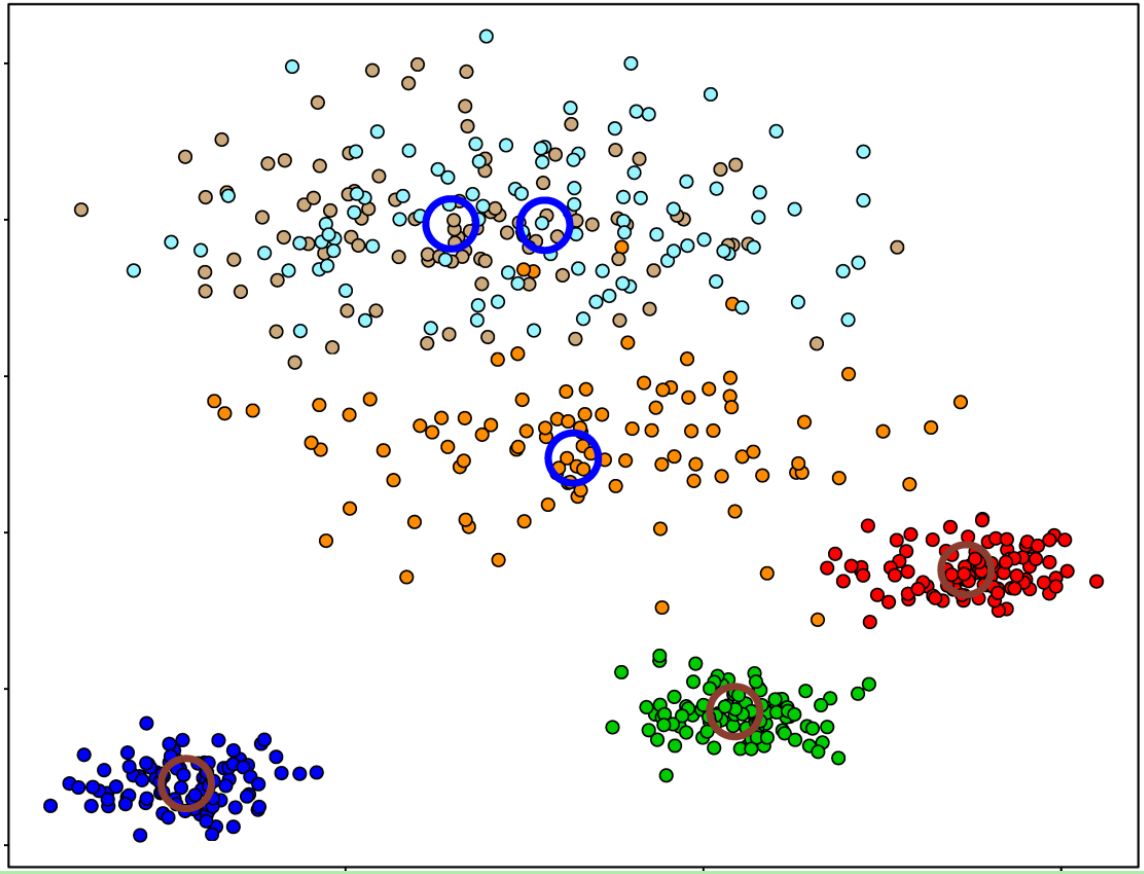
exemplars are highlighted by colored boxes



Affinity Propagation: Examples

data set with 6 clusters in a 2-D space, where three clusters are smaller (smaller variance) than the other three clusters

- Circles indicate the centers of the clusters
- Blue are the large clusters and brown the small clusters



Affinity Propagation: Examples

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

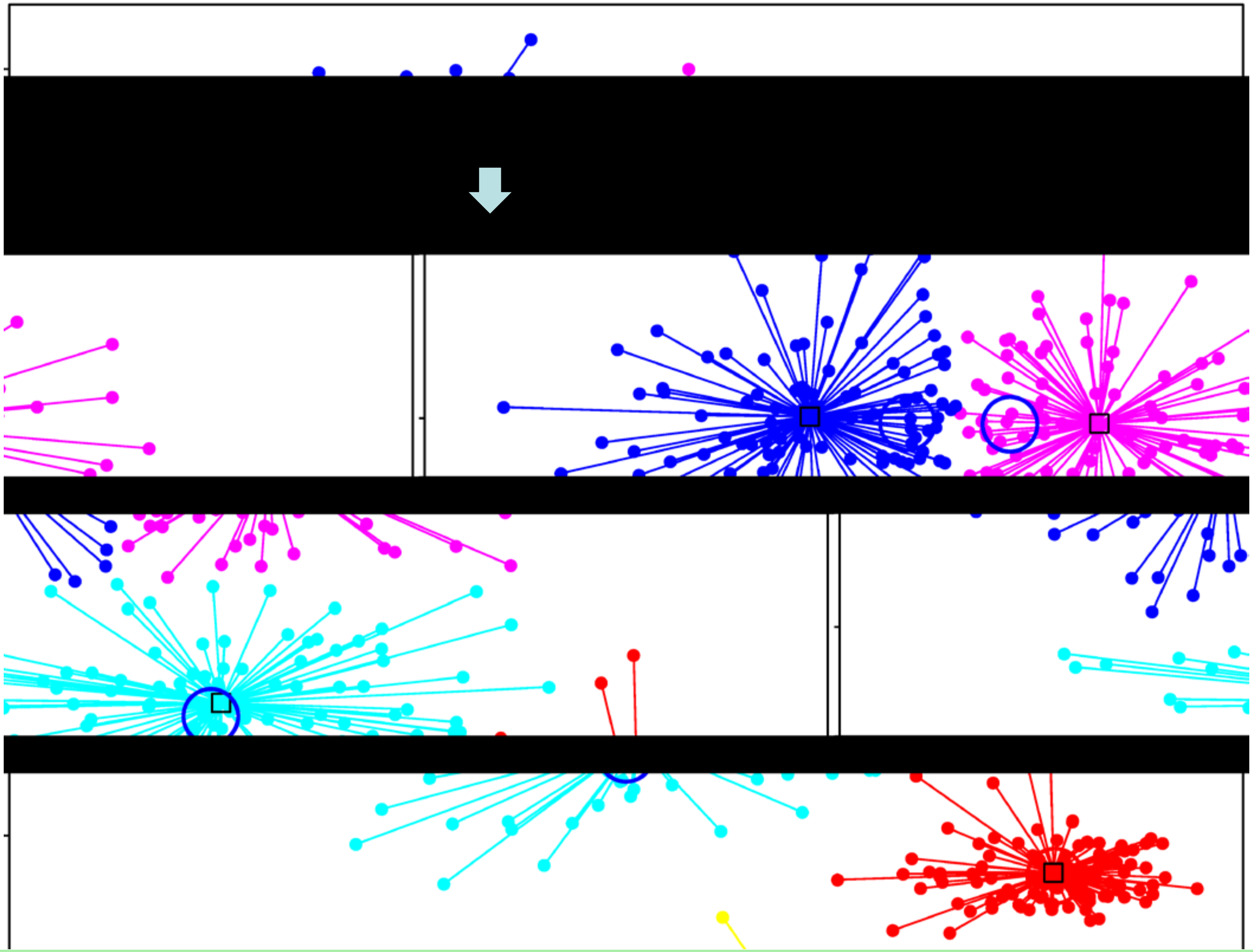
8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

- AP exemplars: black rectangles

- The two large clusters are identified
- AP prefers spherical clusters which helps here



[redacted]

[redacted]

[redacted]

Affinity Propagation: Examples

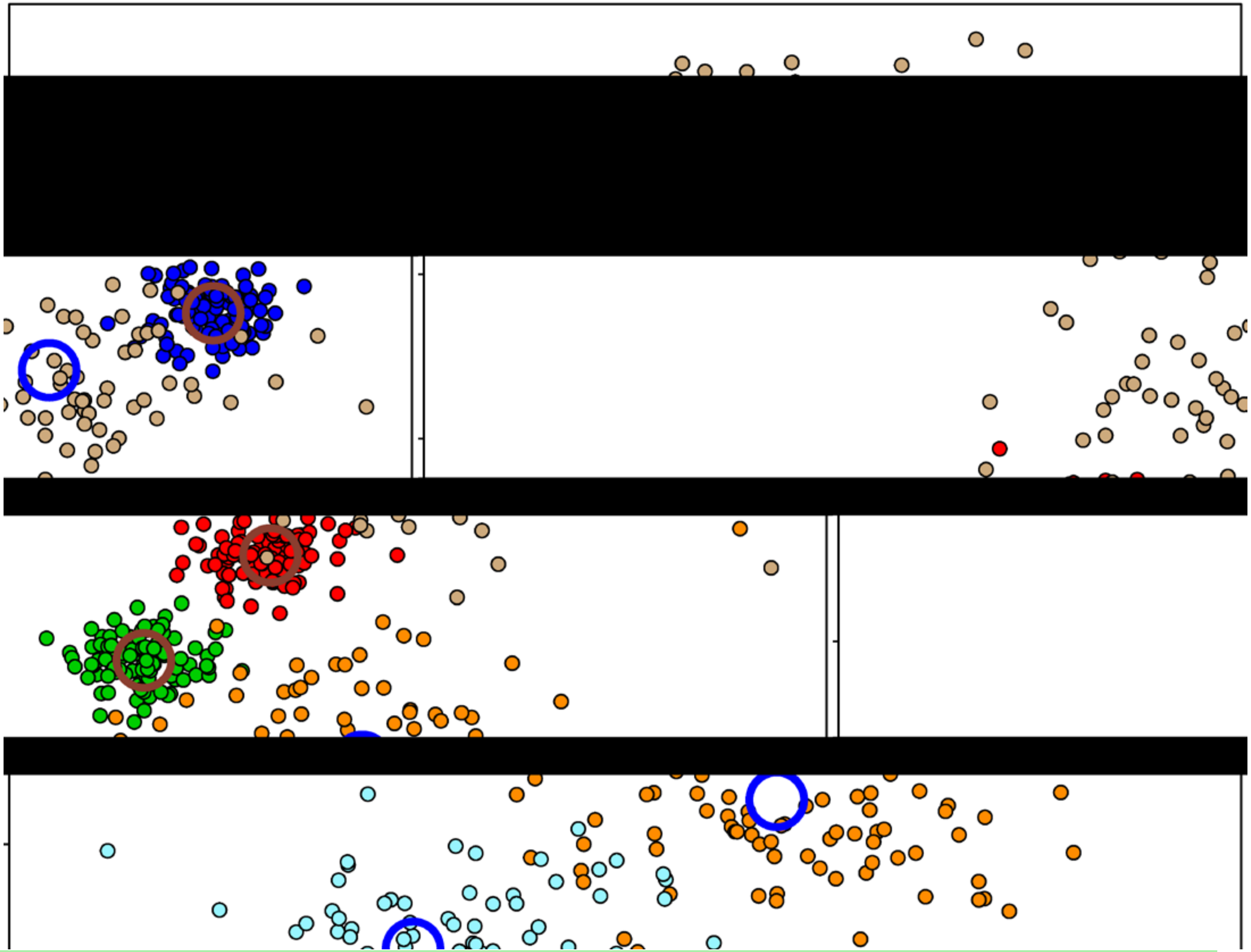
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

The x7A data set: another version of the x7 data generation



[redacted]

[redacted]

[redacted]

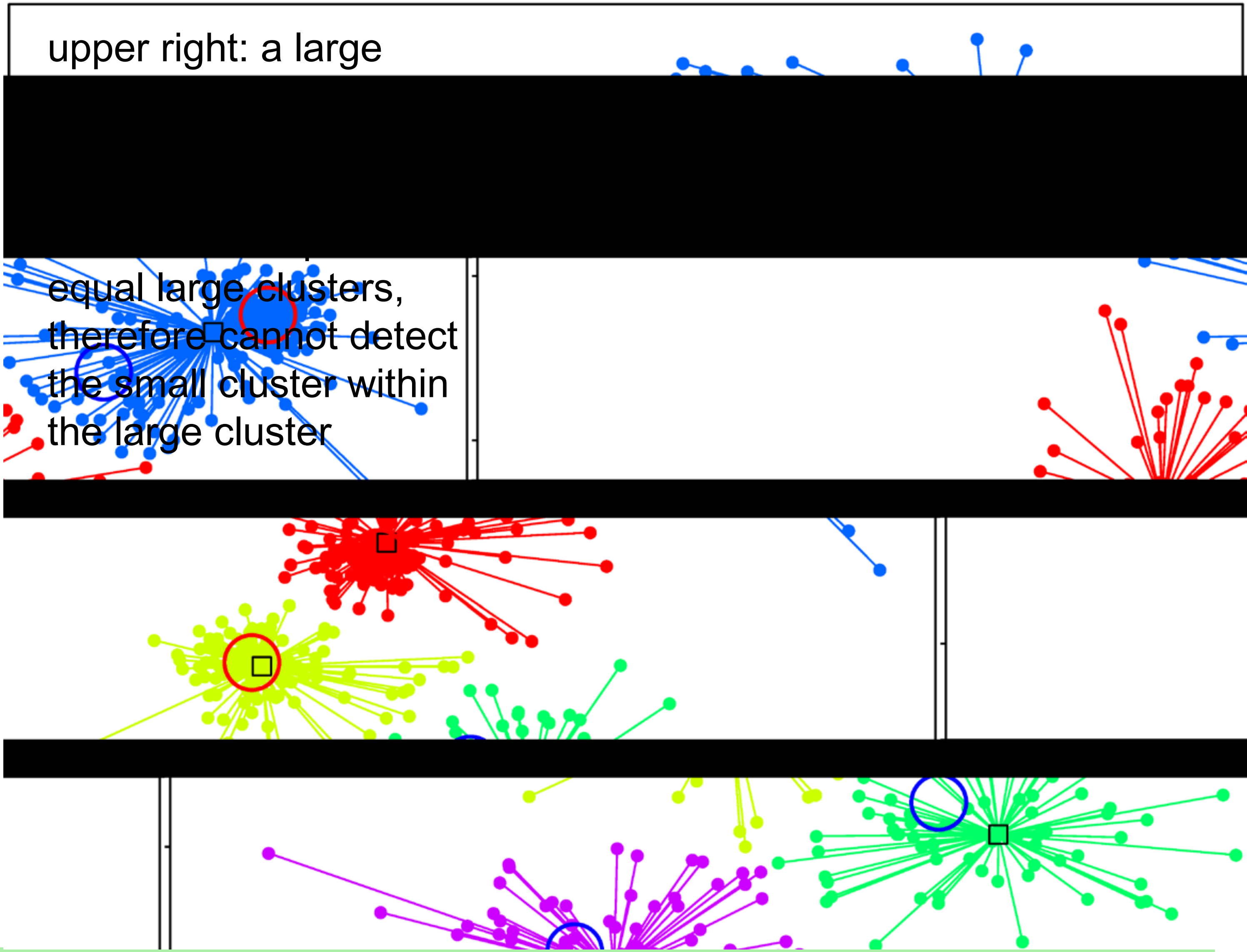
Affinity Propagation: Examples

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine



upper right: a large

[redacted]

equal large clusters, therefore cannot detect the small cluster within the large cluster

[redacted]

[redacted]

Affinity Propagation: Examples

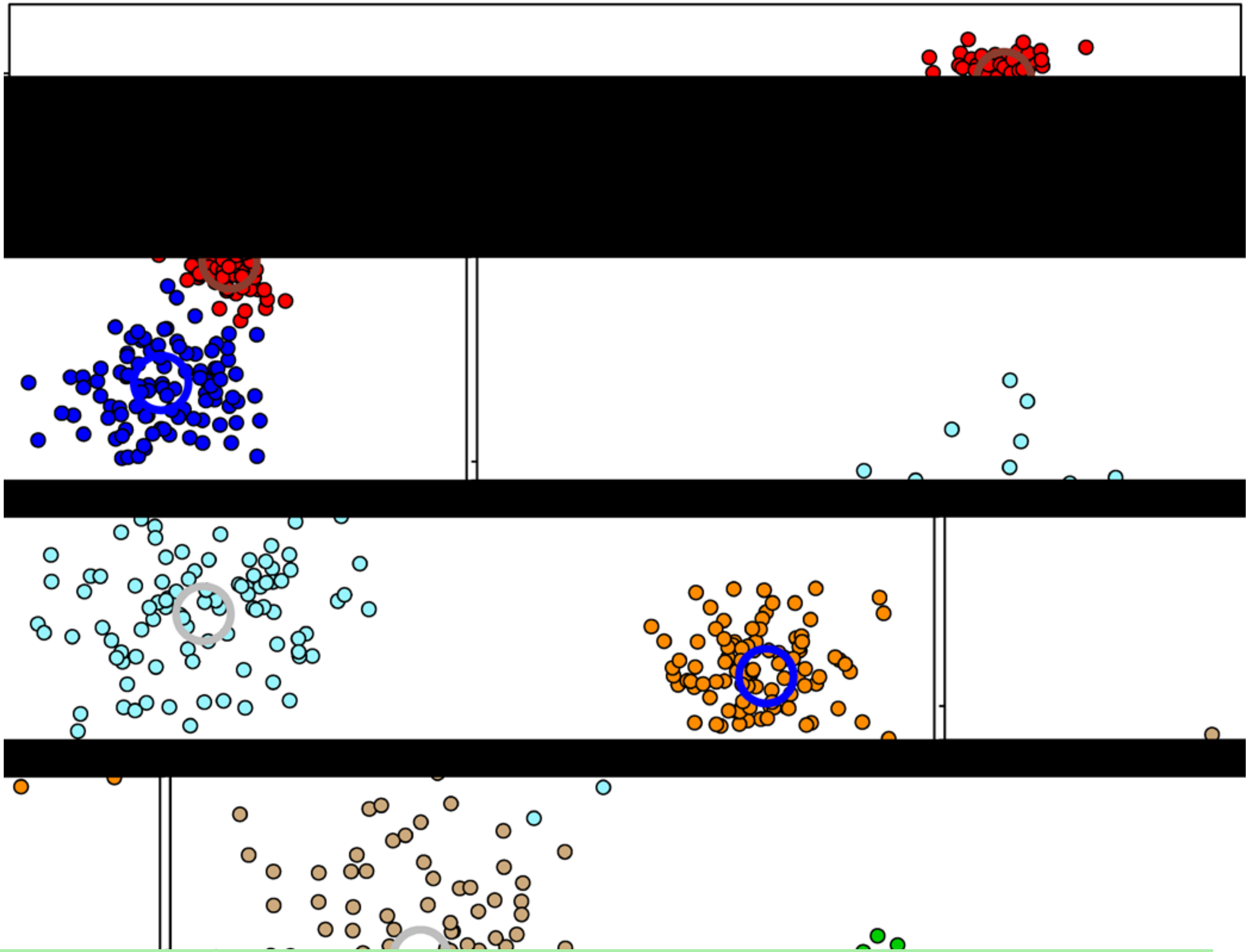
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

x9 data set: 6 clusters of which 2 are small, 2 medium sized, 2 large with respect to the variance



Affinity Propagation: Examples

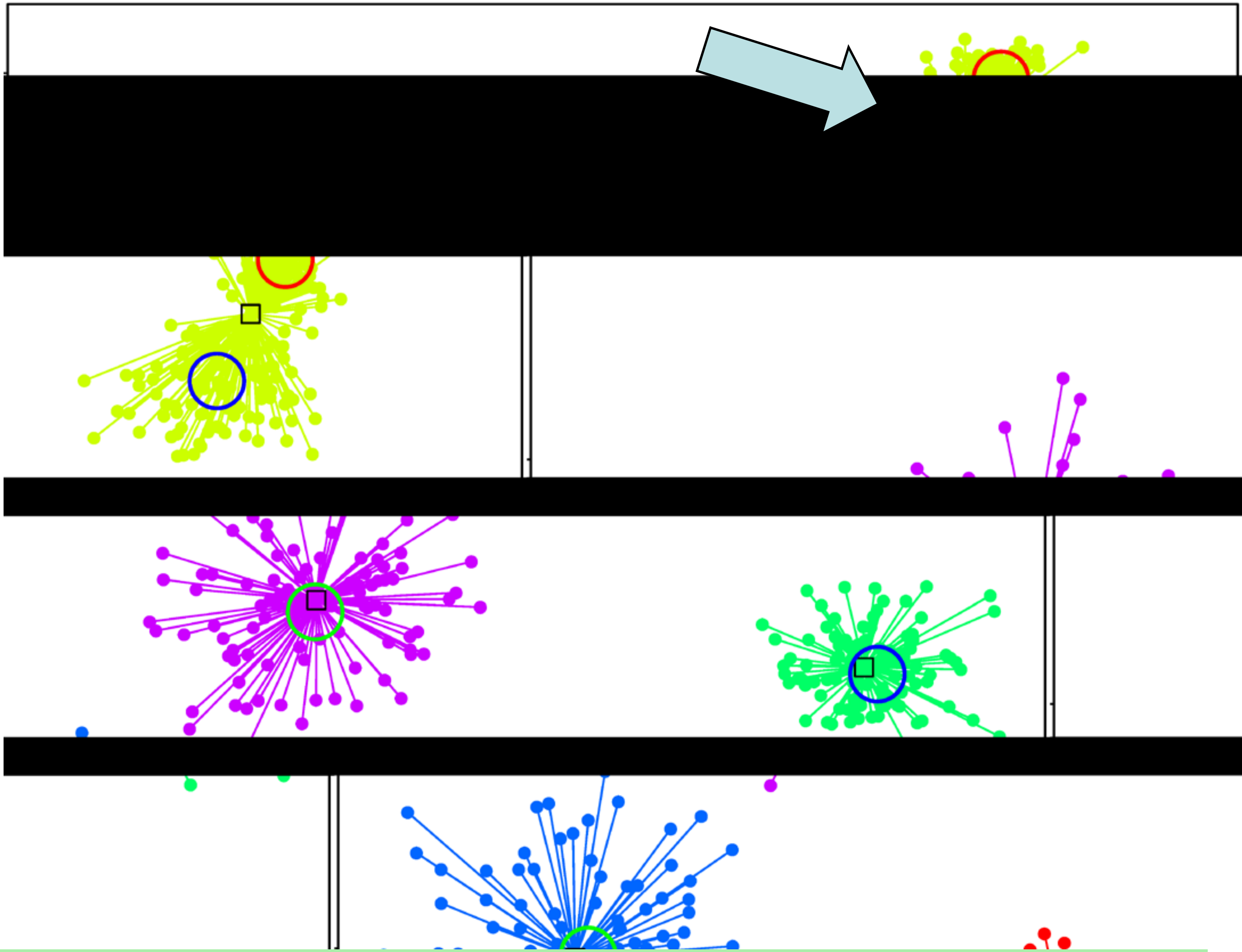
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

AP applied to x9: one cluster is not separated into the two true clusters



[redacted]

[redacted]

[redacted]

Affinity Propagation: Examples

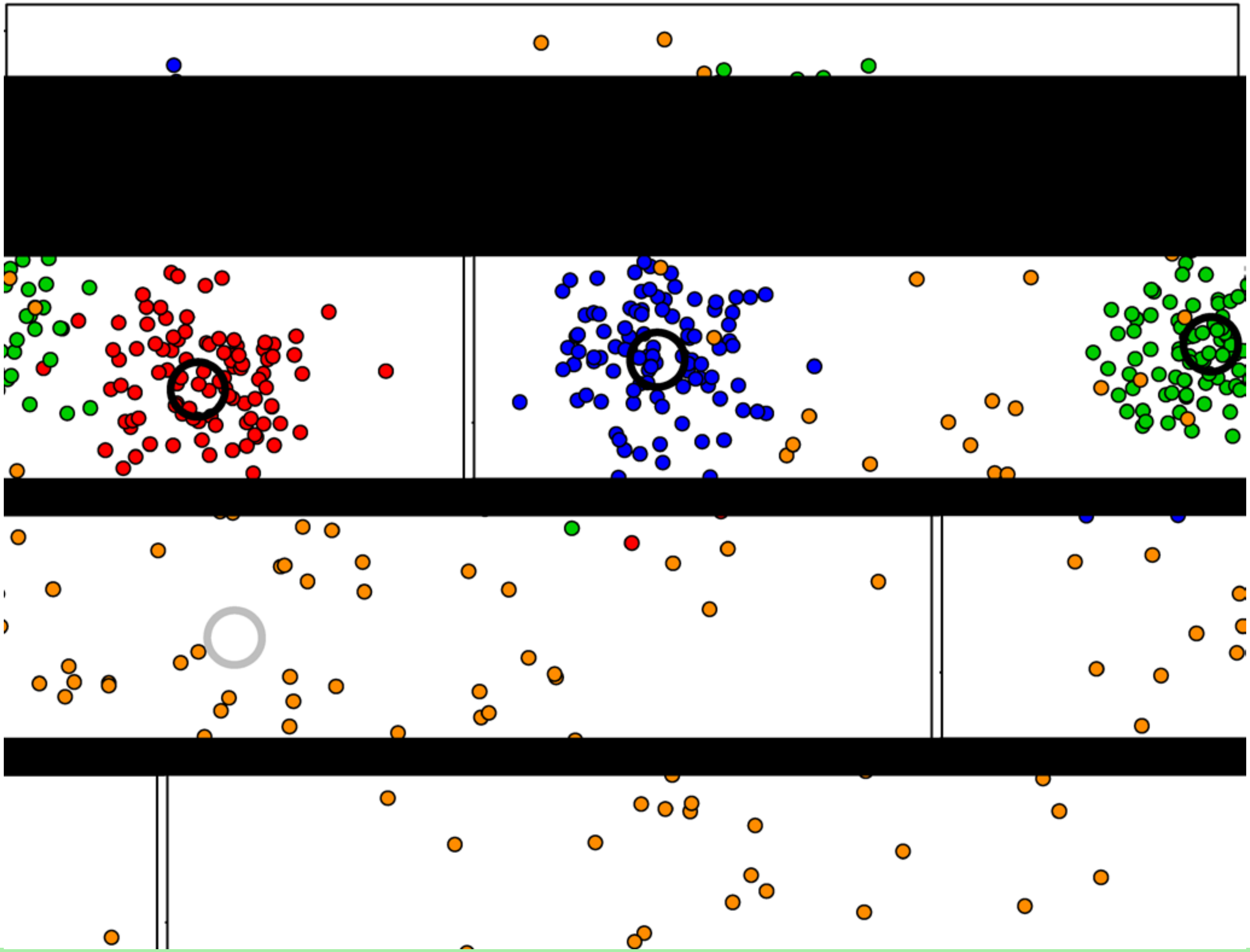
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

6-D data set with different cluster sizes (variation of their elements)
one large and three smaller clusters: 100 samples; PCA down-projection



[redacted]

[redacted]

[redacted]

Affinity Propagation: Examples

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

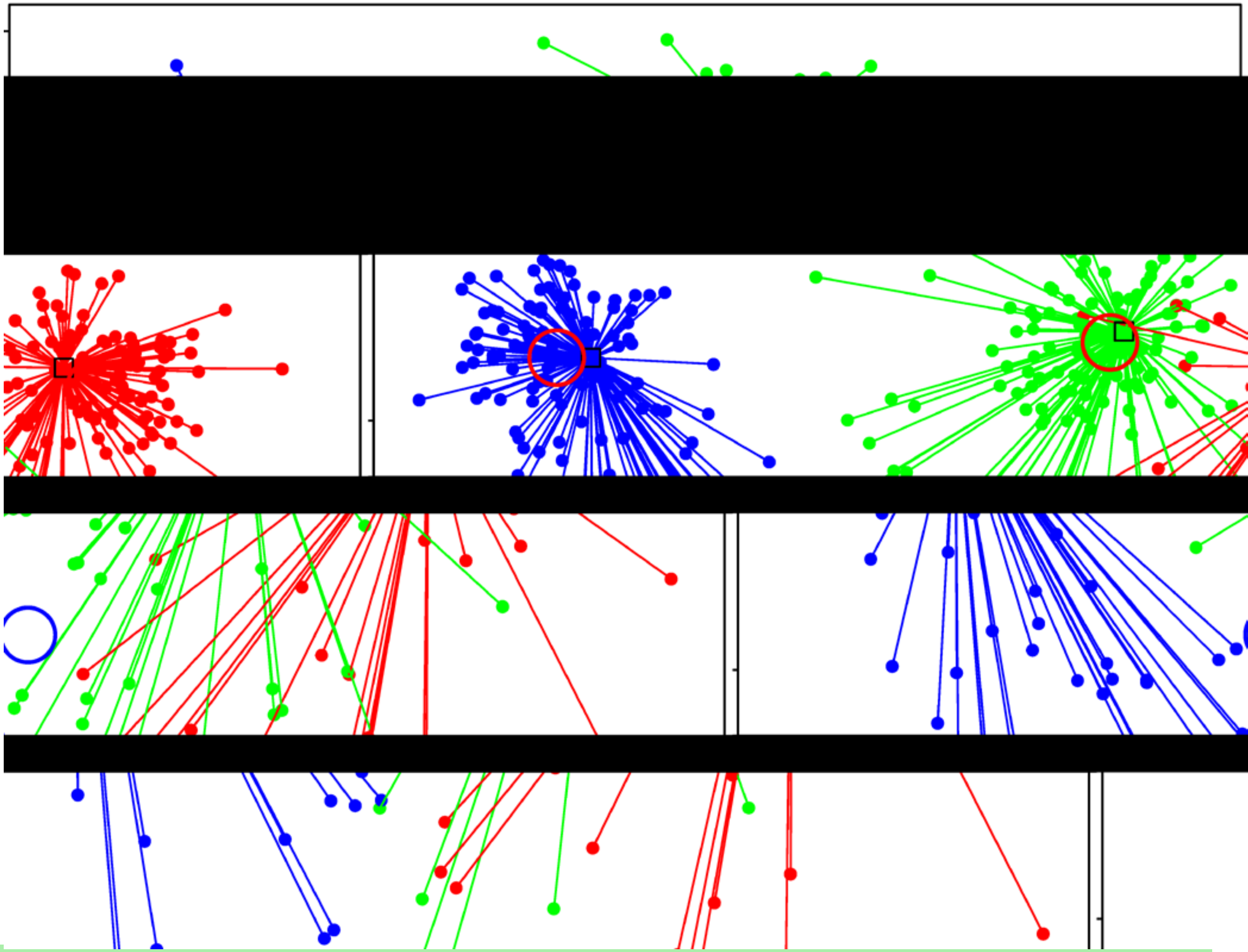
10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

affinity propagation

AP does not detect the large cluster

Elements of the large cluster are assigned to the smaller cluster

AP has problems with the different cluster sizes because it cannot adjust the variance



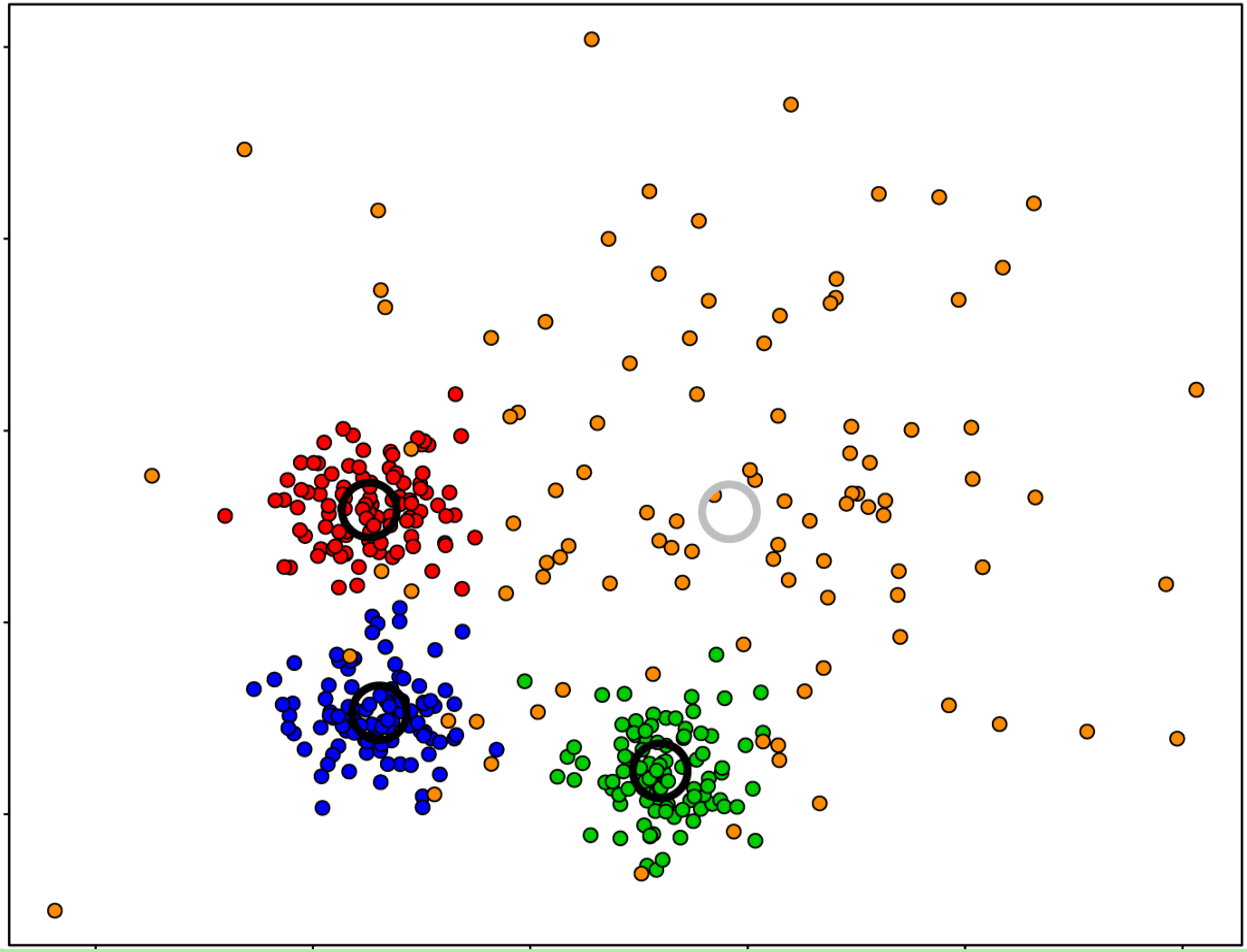
[redacted]

[redacted]

[redacted]

Affinity Propagation: Examples

Another data set of d6: d6A down-projected by PCA



Affinity Propagation: Examples

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

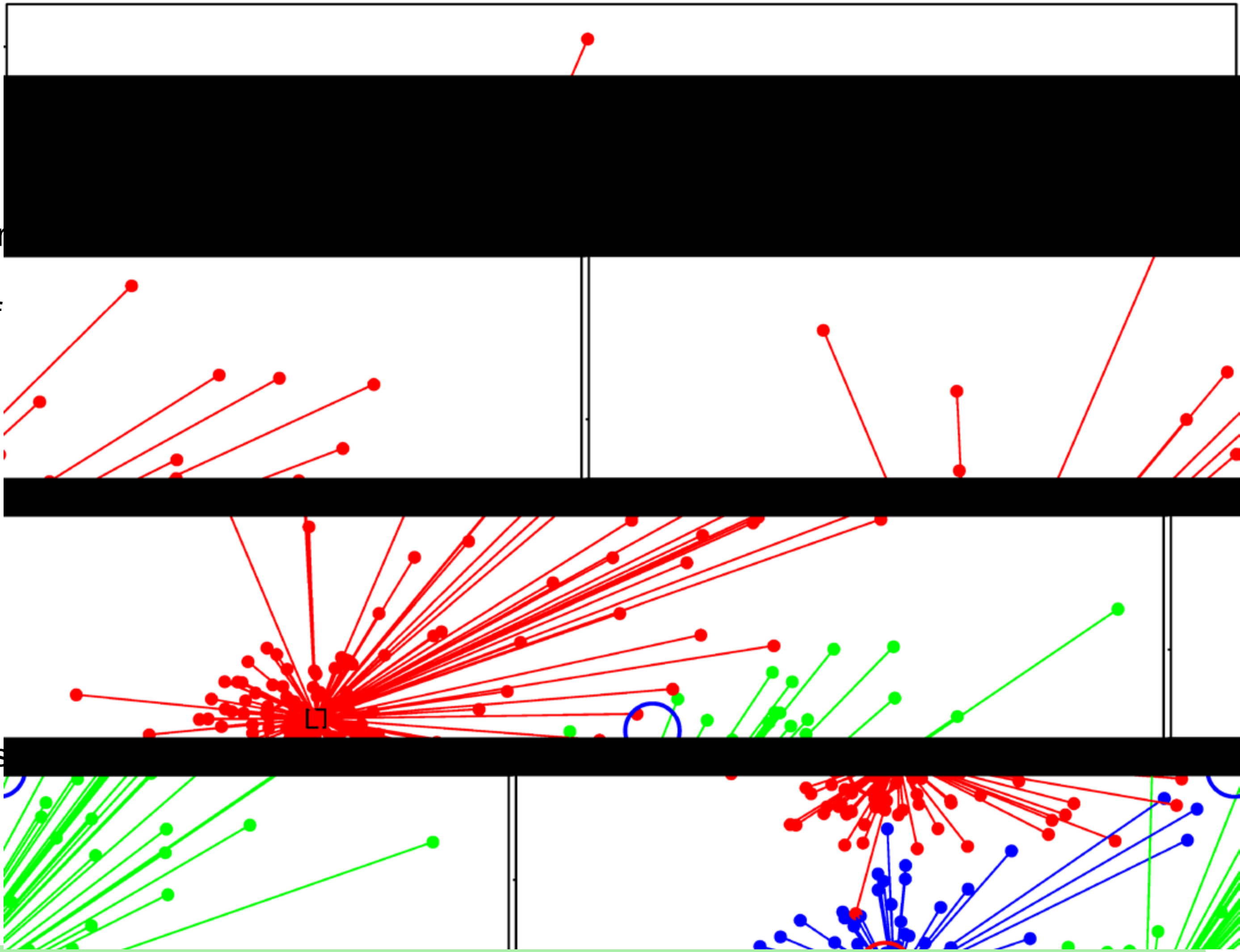
10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

affinity propagation for d6A

Again AP does not detect the large cluster

Elements of the large cluster are assigned to the smaller cluster

AP has problems with the different cluster sizes because it cannot adjust the variance



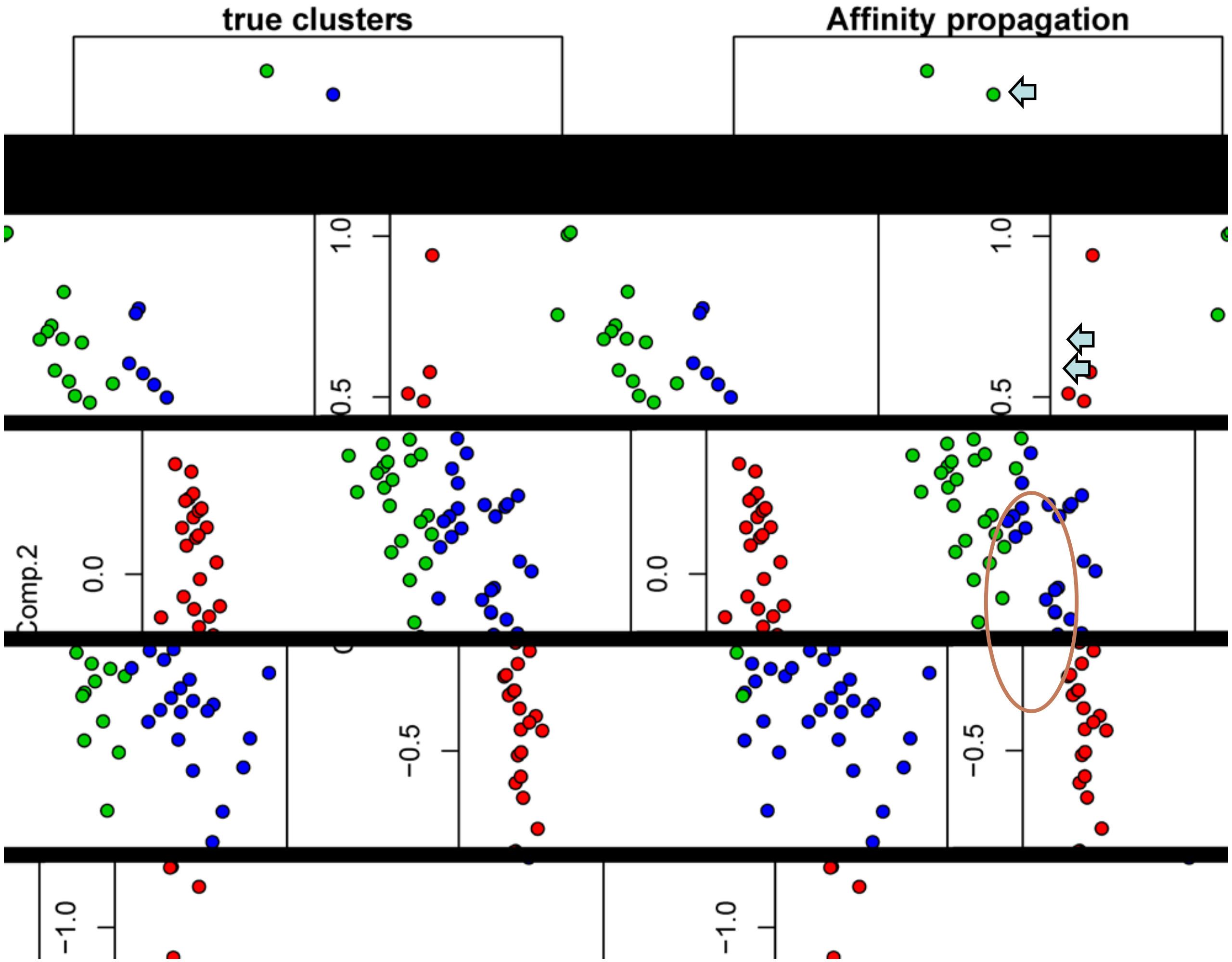
[redacted]

[redacted]

[redacted]

Affinity Propagation: Examples

affinity propagation applied to the iris data



Affinity Propagation: Examples

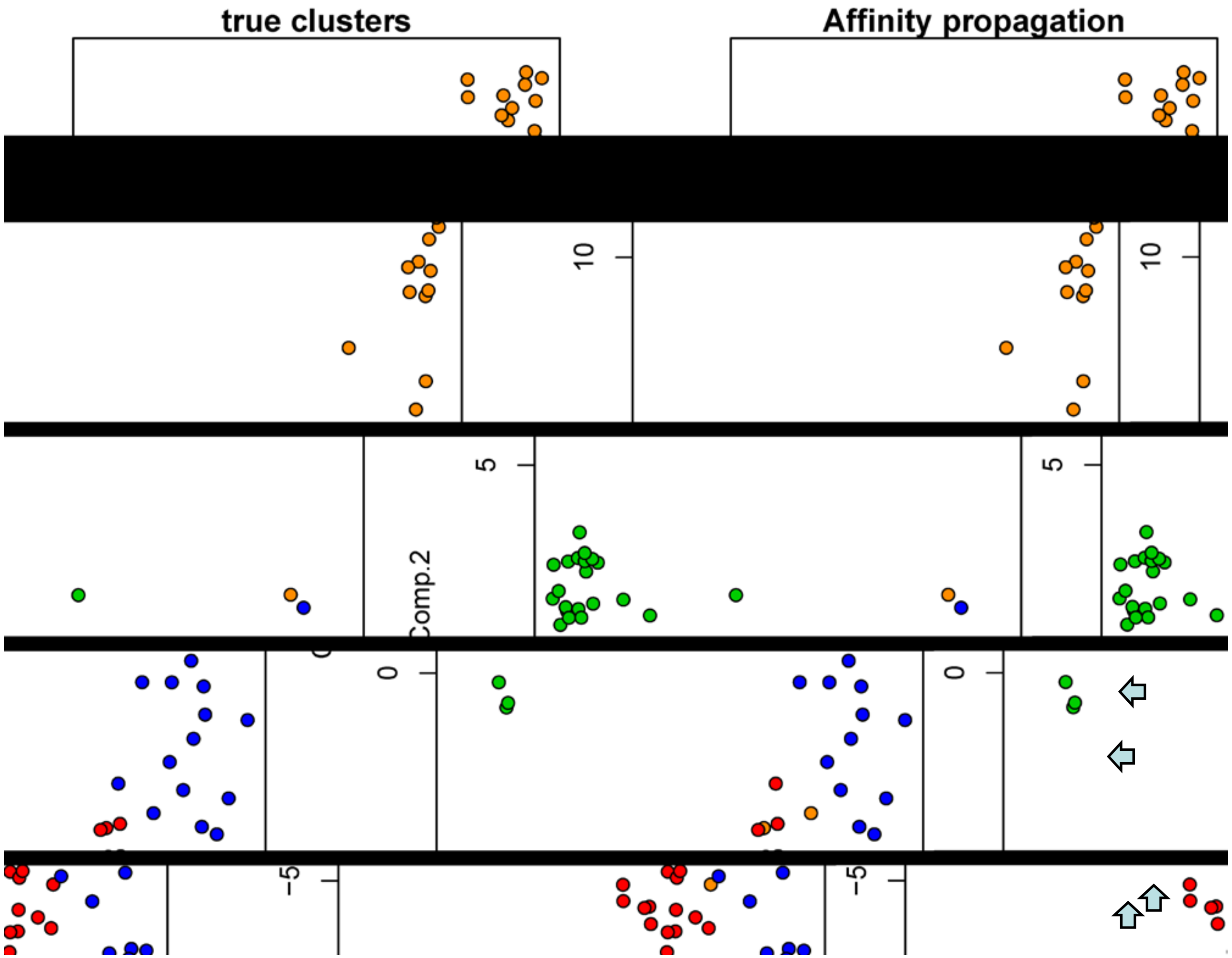
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

affinity propagation applied to the multiple tissues: p=0


[redacted]

Affinity Propagation: Examples

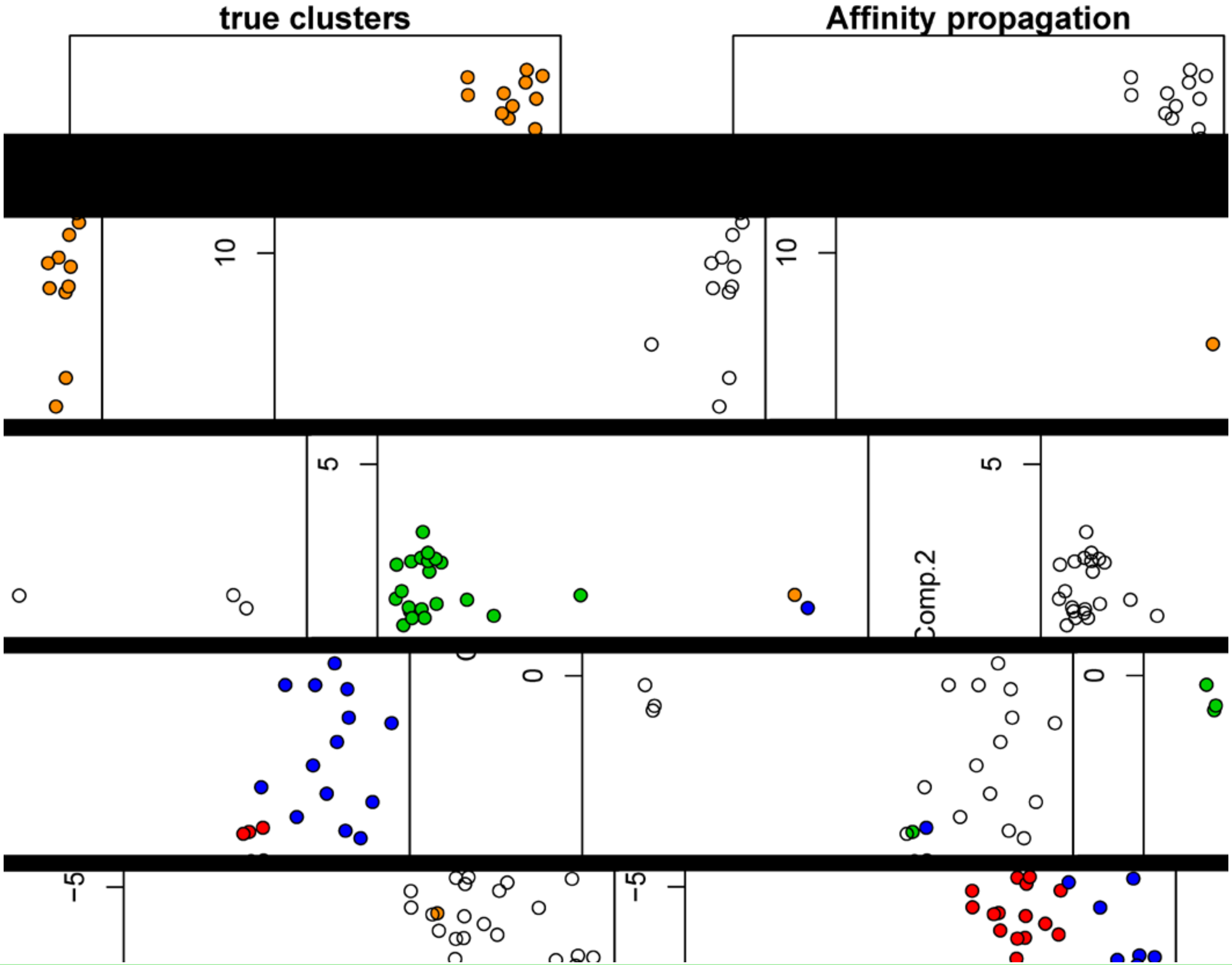
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

affinity propagation applied to the multiple tissues: p=1


[redacted]
[redacted]
[redacted]
[redacted]

Affinity Propagation: Examples

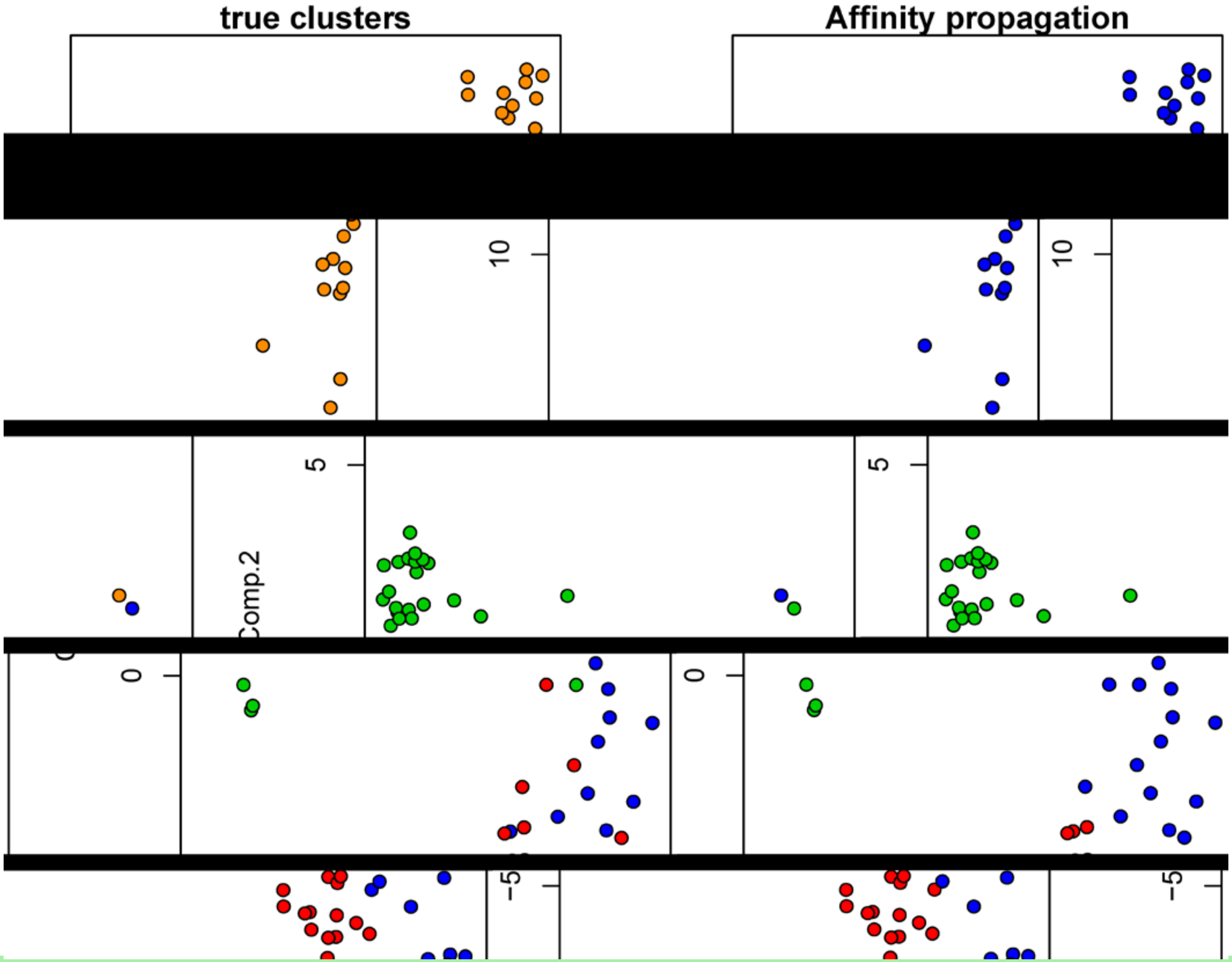
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

affinity propagation applied to the multiple tissues: p=-1(three clusters)



Similarity-Based Clustering: Mixture Models

Similarity-based mixture models are mixture models that use only similarities between objects but not feature vectors

- similarity between object $\boldsymbol{i}$ and object $\boldsymbol{j}$: $p(\boldsymbol{x}_j \mid \boldsymbol{x}_i)$ (given)
- stochastic neighbor embedding: probabilities are computed

Similarity-Based Clustering: Mixture Models

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

Similarity-based mixture of Gaussians

For $\boldsymbol{x}$ the similarities $k(\boldsymbol{x}, \boldsymbol{x}_i)$ to each element of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are given, in particular all $k(\boldsymbol{x}_j, \boldsymbol{x}_i)$ for $1 \leq i, j \leq n$ are given.

all update rules and parameters are expressed by these similarities

Gaussian mixture model $p(\boldsymbol{x}) \ = \ \sum_{i=1}^{K} p(i) \ p(\boldsymbol{x} \mid i) \ = \ \sum_{i=1}^{K} p(i) \ k(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

where $\sum_{i=1}^{K} p(i) \ = \ 1$ and $k(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is the Gaussian density with mean $\boldsymbol{\mu}_i$ and variance $\boldsymbol{\Sigma}_i$ evaluated at $\boldsymbol{x}$

We introduce a variance σ_i for component $\boldsymbol{i}$ by $\frac{1}{\sigma_i^m} k(\boldsymbol{x}; \boldsymbol{\mu}_i)^{1/\sigma_i^2}$ and obtain for the likelihood of $\boldsymbol{x}$:

$$p(\boldsymbol{x}) \ = \ \sum_{j=1}^{K} p(j) \ \frac{1}{\sigma_j^m} \ k(\boldsymbol{x}, \boldsymbol{\mu}_j)^{1/\sigma_j^2}$$

Similarity-Based Clustering: Mixture Models

We set $k(\boldsymbol{x};\boldsymbol{\mu}_i,\sigma_i^2) \;=\; \frac{1}{\sigma_i^m}\; k(\boldsymbol{x},\boldsymbol{\mu}_i)^{1/\sigma_i^2}$

The posterior gives the probability that $\boldsymbol{x}$ belongs to cluster $\boldsymbol{i}$:

$$p(i \mid \boldsymbol{x}) \;=\; \frac{p(i)\; p(\boldsymbol{x} \mid i)}{\sum_{j=1}^{K} p(j)\; p(\boldsymbol{x} \mid j)} \;=\; \frac{p(i)\; k(\boldsymbol{x};\boldsymbol{\mu}_i,\sigma_i^2)}{\sum_{j=1}^{K} p(j)\; k(\boldsymbol{x};\boldsymbol{\mu}_j,\sigma_j^2)} \;=\; \frac{p(i)\; k(\boldsymbol{x};\boldsymbol{\mu}_i,\sigma_i^2)}{p(\boldsymbol{x})}$$

With $\alpha_i = p(i)$ and $\alpha_{ik} = p(i \mid \boldsymbol{x}_k,\boldsymbol{\alpha})$ we obtain

$$\alpha_i \;=\; p(i) \;=\; p(i \mid \boldsymbol{\alpha}) \;=\; \int p(i,\boldsymbol{x} \mid \boldsymbol{\alpha})\; d\boldsymbol{x} \;=\; \int p(i \mid \boldsymbol{x},\boldsymbol{\alpha})\; p(\boldsymbol{x} \mid \boldsymbol{\alpha})\; d\boldsymbol{x} \;=\; \mathrm{E}_{p(\boldsymbol{x}|\boldsymbol{\alpha})}(p(i \mid \boldsymbol{x},\boldsymbol{\alpha}))$$

$$\approx\; \frac{1}{n}\sum_{k=1}^{n} p(i \mid \boldsymbol{x}_k,\boldsymbol{\alpha}) \;=\; \frac{1}{n}\sum_{k=1}^{n} \alpha_{ik}$$

Similarity-Based Clustering: Mixture Models

The likelihood for one data point $\boldsymbol{x}$ is

$$p(\boldsymbol{x}) \ = \ \sum_{i=1}^{K} \alpha_i \ \frac{1}{\sigma_i^m} \ k(\boldsymbol{x}_k; \boldsymbol{\mu}_i)^{1/\sigma_i^2} \ = \ \sum_{i=1}^{K} \alpha_i \ k(\boldsymbol{x}_k; \boldsymbol{\mu}_i, \sigma_i^2)$$

The objective is the negative log posterior:

$$\begin{aligned} B \ = \ & - \ \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{K} \hat{\alpha}_{ik} \ \log\left(k(\boldsymbol{x}_k; \boldsymbol{\mu}_i, \sigma_i^2)\right) \ - \ \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{K} \hat{\alpha}_{ik} \ \log\left(\alpha_i\right) \\ & - \ \frac{1}{n} \ \log p(\boldsymbol{\alpha}) \ - \ \frac{1}{n} \ \log p(\boldsymbol{\sigma}^2) \ + \ \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{K} \hat{\alpha}_{ik} \ \log \hat{\alpha}_{ik} \ + \ \frac{1}{n} \sum_{k=1}^{n} \ \log c(\boldsymbol{x}_k) \end{aligned}$$

The objective contains the Dirichlet prior $p(\boldsymbol{\alpha})$ and Wishart prior $p(\boldsymbol{\sigma}^2)$

$$\boldsymbol{\sigma}^2 \ = \ (\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2)$$

Similarity-Based Clustering: Mixture Models

Using $\hat{\alpha}_{ik} \; = \; \frac{\alpha_i^{\mathrm{old}} \; k(\boldsymbol{x}_k; \boldsymbol{\mu}_i, \sigma_i^2)}{p(\boldsymbol{x}_k)}$ and $\hat{\alpha}_i \; = \; \frac{1}{n} \sum_{k=1}^{n} \hat{\alpha}_{ik}$ and $\gamma_s = \sum_{i=1}^{n} \gamma_i$

the update rules are:

$$\alpha_i^{\mathrm{new}} \; = \; \frac{n \; \hat{\alpha}_i \; + \; \gamma_i \; - 1}{\gamma_s}$$

$$\left(\sigma_i^2\right)^{\mathrm{new}} \; = \; \frac{\sum_{k=1}^{n} \hat{\alpha}_{ik} \; \left(- \; 2 \; \log k(\boldsymbol{x}_k, \boldsymbol{\mu}_i)\right) \; + \; w \; S}{n \; \hat{\alpha}_i \; + \; w}$$

hyper-parameters γ_i come from the Dirichlet prior

hyper-parameters w and S come from the Wishart prior

Similarity-Based Clustering: Mixture Models

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

Representing the center

EM algorithm for MoG has center update:

$$\boldsymbol{\mu}_i \;=\; \frac{\sum_{k=1}^{n} \hat{\alpha}_{ik} \; \boldsymbol{x}_k}{\sum_{k=1}^{n} \hat{\alpha}_{ik}}$$

For the optimal center: $\log k(\boldsymbol{x}_k; \boldsymbol{x}_i, \boldsymbol{\sigma}_i^2) \;=\; - \, n \; \log(\sigma_i) \;-\; \sigma_i^{-2} \; (- \; \log k(\boldsymbol{x}_k, \boldsymbol{x}_i))$

$$\frac{\partial B}{\partial \boldsymbol{\mu}_i} \;=\; - \, \frac{1}{n} \sum_{k=1}^{n} \hat{\alpha}_{ik} \; \frac{1}{\sigma_i^2} \frac{\partial \log k(\boldsymbol{x}_k, \boldsymbol{\mu}_i)}{\partial \boldsymbol{\mu}_i} \;=\; 0$$

We assume:

$$\log k(\boldsymbol{x}_k, \boldsymbol{\mu}_i) \;=\; -\frac{1}{2} \; \| \boldsymbol{x}_k - \boldsymbol{\mu}_i \| \;=\; -\frac{1}{2} \; (\boldsymbol{x}_k - \boldsymbol{\mu}_i)^T (\boldsymbol{x}_k - \boldsymbol{\mu}_i)$$

$$\frac{\partial \log k(\boldsymbol{x}_k, \boldsymbol{\mu}_i)}{\partial \boldsymbol{\mu}_i} \;=\; - \; \boldsymbol{x}_k \; + \; \boldsymbol{\mu}_i$$

Similarity-Based Clustering: Mixture Models

For the update rules we only require $\|\boldsymbol{x}_l - \boldsymbol{\mu}_i\| = -2 \log k(\boldsymbol{x}_l, \boldsymbol{\mu}_i)$

We assumed for the similarities $k(\boldsymbol{x}_k, \boldsymbol{x}_j) \ = \ \frac{1}{(2\ \pi)^{m/2}} \exp\left(-\ \frac{1}{2}\ \|\boldsymbol{x}_k - \boldsymbol{x}_j\|^2\right)$

from which follows that

$$- \ 2\ \log k(\boldsymbol{x}_k, \boldsymbol{x}_j) \ = \ n \log(2\pi) \ + \ \|\boldsymbol{x}_k - \boldsymbol{x}_j\|^2 \ = \ n \log(2\pi) \ + \ \boldsymbol{x}_k^T \boldsymbol{x}_k \ - \ 2\ \boldsymbol{x}_k^T \boldsymbol{x}_j \ + \ \boldsymbol{x}_j^T \boldsymbol{x}_j$$

and, therefore, we have

$$\boldsymbol{x}_k^T \boldsymbol{x}_j \ = \ \frac{1}{2}\ n \log(2\pi) \ + \ \frac{1}{2} \boldsymbol{x}_k^T \boldsymbol{x}_k \ + \ \frac{1}{2} \boldsymbol{x}_j^T \boldsymbol{x}_j \ + \ \log k(\boldsymbol{x}_k, \boldsymbol{x}_j)$$

$$\|\boldsymbol{x}_l - \boldsymbol{\mu}_i\|^2 \ = \ (\boldsymbol{x}_l - \boldsymbol{\mu}_i)^T (\boldsymbol{x}_l - \boldsymbol{\mu}_i) \ = \ \boldsymbol{x}_l^T \boldsymbol{x}_l \ - \ 2\ \boldsymbol{x}_l^T \boldsymbol{\mu}_i \ + \ \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i \ = \ \boldsymbol{x}_l^T \boldsymbol{x}_l \ - \ 2\ \frac{\sum_{k=1}^n \hat{\alpha}_{ik}\ \boldsymbol{x}_k^T \boldsymbol{x}_l}{\sum_{k=1}^n \hat{\alpha}_{ik}} \ + \ \frac{\sum_{k,j=1}^n \hat{\alpha}_{ik}\ \hat{\alpha}_{ij}\ \boldsymbol{x}_k^T \boldsymbol{x}_j}{\left(\sum_{k=1}^n \hat{\alpha}_{ik}\right)^2}$$

$$= \ \boldsymbol{x}_l^T \boldsymbol{x}_l \ - \ \frac{\sum_{k=1}^n \hat{\alpha}_{ik}\ \left(\boldsymbol{x}_k^T \boldsymbol{x}_k \ + \ 2 \log k(\boldsymbol{x}_k, \boldsymbol{x}_l)\right)}{\sum_{k=1}^n \hat{\alpha}_{ik}} \ - \ n \log\ (2\pi) \ - \ \boldsymbol{x}_l^T \boldsymbol{x}_l$$

$$+ \ \frac{\sum_{k,j=1}^n \hat{\alpha}_{ik}\ \hat{\alpha}_{ij}\ \left(1/2\ n \log(2\pi) \ + \ 1/2\ \boldsymbol{x}_k^T \boldsymbol{x}_k \ + \ 1/2\ \boldsymbol{x}_j^T \boldsymbol{x}_j \ + \ \log k(\boldsymbol{x}_k, \boldsymbol{x}_j)\right)}{\left(\sum_{k=1}^n \hat{\alpha}_{ik}\right)^2}$$

$$= - \ \frac{\sum_{k=1}^n \hat{\alpha}_{ik}\ \boldsymbol{x}_k^T \boldsymbol{x}_k}{\sum_{k=1}^n \hat{\alpha}_{ik}} \ - \ 2\ \frac{\sum_{k=1}^n \hat{\alpha}_{ik}\ \log k(\boldsymbol{x}_k, \boldsymbol{x}_l)}{\sum_{k=1}^n \hat{\alpha}_{ik}} \ - \ n \log(2\pi)$$

$$+ \ 1/2\ n \log(2\pi) \ + \ \frac{1}{2}\ \frac{\sum_{k=1}^n \hat{\alpha}_{ik}\ \boldsymbol{x}_k^T \boldsymbol{x}_k}{\sum_{k=1}^n \hat{\alpha}_{ik}} \ + \ \frac{1}{2}\ \frac{\sum_{j=1}^n \hat{\alpha}_{ij}\ \boldsymbol{x}_j^T \boldsymbol{x}_j}{\sum_{k=1}^n \hat{\alpha}_{ik}} \ + \ \frac{\sum_{k,j=1}^n \hat{\alpha}_{ik}\ \hat{\alpha}_{ij}\ \log k(\boldsymbol{x}_k, \boldsymbol{x}_j)}{\left(\sum_{k=1}^n \hat{\alpha}_{ik}\right)^2}$$

$$= \ - \ 1/2\ n \log(2\pi) \ - \ 2\ \frac{\sum_{k=1}^n \hat{\alpha}_{ik}\ \log k(\boldsymbol{x}_k, \boldsymbol{x}_l)}{\sum_{k=1}^n \hat{\alpha}_{ik}} \ + \ \frac{\sum_{k,j=1}^n \hat{\alpha}_{ik}\ \hat{\alpha}_{ij}\ \log k(\boldsymbol{x}_k, \boldsymbol{x}_j)}{\left(\sum_{k=1}^n \hat{\alpha}_{ik}\right)^2}$$

Similarity-Based Clustering: Mixture Models

we can compute $k(\boldsymbol{x}_k, \boldsymbol{\mu}_i) \ = \ \frac{1}{(2\ \pi)^{m/2}} \exp\left(-\ \frac{1}{2}\ \|\boldsymbol{x}_k - \boldsymbol{\mu}_i\|^2\right)$

If we ignore $1/(2\pi)^{m/2}$ and define $k(\boldsymbol{x}, \boldsymbol{x}_i) \ = \ \exp\left(-\ \frac{1}{2}\ \|\boldsymbol{x} - \boldsymbol{x}_i\|^2\right)$

then we would obtain

$$\|\boldsymbol{x}_l - \boldsymbol{\mu}_i\|^2 \ = \ -\ 2\ \frac{\sum_{k=1}^{n} \hat{\alpha}_{ik}\ \log k(\boldsymbol{x}_k, \boldsymbol{x}_l)}{\sum_{k=1}^{n} \hat{\alpha}_{ik}} \ + \ \frac{\sum_{k,j=1}^{n} \hat{\alpha}_{ik}\ \hat{\alpha}_{ij}\ \log k(\boldsymbol{x}_k, \boldsymbol{x}_j)}{\left(\sum_{k=1}^{n} \hat{\alpha}_{ik}\right)^2}$$

and

$$k(\boldsymbol{x}_k, \boldsymbol{\mu}_i) \ = \ \exp\left(-\ \frac{1}{2}\ \|\boldsymbol{x}_k - \boldsymbol{\mu}_i\|^2\right)$$

Scaling of $\boldsymbol{k}$ does not change the updates.

Similarity-Based Clustering: Mixture Models

Update Rules

$$
\begin{aligned}
k(\boldsymbol{x}_k; \boldsymbol{\mu}_i, \sigma_i^2) &= \frac{1}{\sigma_i^m} \; k(\boldsymbol{x}_k, \boldsymbol{\mu}_i)^{1/\sigma_i^2} \\
p(\boldsymbol{x}_k) &= \sum_{i=1}^{l} \alpha_i^{\text{old}} \; k(\boldsymbol{x}_k; \boldsymbol{\mu}_i, \sigma_i^2) \\
\hat{\alpha}_{ik} &= \frac{\alpha_i^{\text{old}} \; k(\boldsymbol{x}_k; \boldsymbol{\mu}_i, \sigma_i^2)}{p(\boldsymbol{x}_k)} \\
\hat{\alpha}_i &= \frac{1}{n} \sum_{k=1}^{n} \hat{\alpha}_{ik} \\
\alpha_i^{\text{new}} &= \frac{n \; \hat{\alpha}_i \; + \; \gamma_i \; - 1}{\gamma_s} \\
(\sigma_i^2)^{\text{new}} &= \frac{\sum_{k=1}^{n} \hat{\alpha}_{ik} \; (- \; 2 \; \log k(\boldsymbol{x}_k, \boldsymbol{\mu}_i)) \; + \; w \; S}{n \; \hat{\alpha}_i \; + \; w} \\
- \; 2 \; \log k(\boldsymbol{x}_l, \boldsymbol{\mu}_i) &= \|\boldsymbol{x}_l - \boldsymbol{\mu}_i\|^2 \; = \; \frac{\sum_{k=1}^{n} \hat{\alpha}_{ik} \; (- \; 2 \; \log k(\boldsymbol{x}_k, \boldsymbol{x}_l))}{n \; \hat{\alpha}_i} \; - \; \frac{1}{2} \; \frac{\sum_{k,j=1}^{n} \hat{\alpha}_{ik} \; \hat{\alpha}_{ij} \; (- \; 2 \; \log k(\boldsymbol{x}_k, \boldsymbol{x}_j))}{n^2 \; \hat{\alpha}_i^2} \\
k(\boldsymbol{x}_l, \boldsymbol{\mu}_i)^{\text{new}} &= \exp\Big(- \; \frac{1}{2} \; \|\boldsymbol{x}_l - \boldsymbol{\mu}_i\|^2 \Big)
\end{aligned}
$$

Similarity-Based Clustering: Mixture Models

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

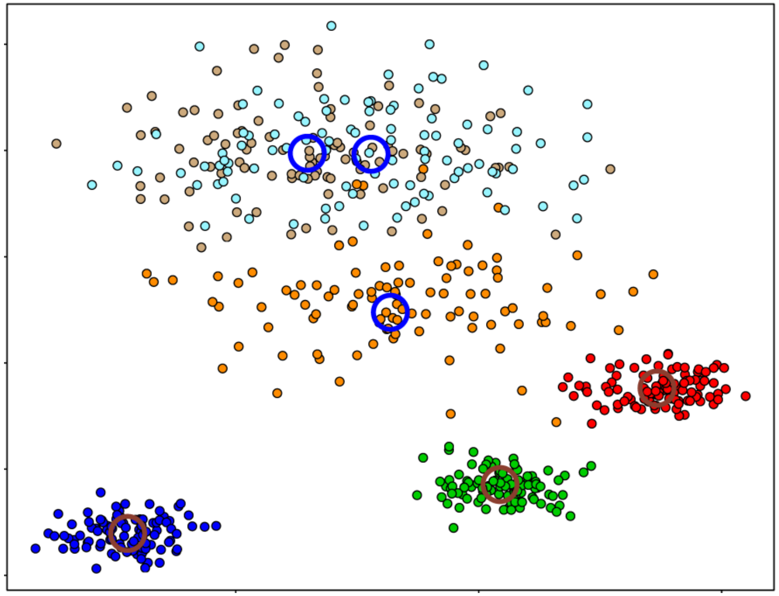
The covariance matrix can also be represented by the similarities. However, we would require a matrix inversion where the dimension of this matrix is the number of observations.

This is computationally very expensive and the algorithm is no longer stable.

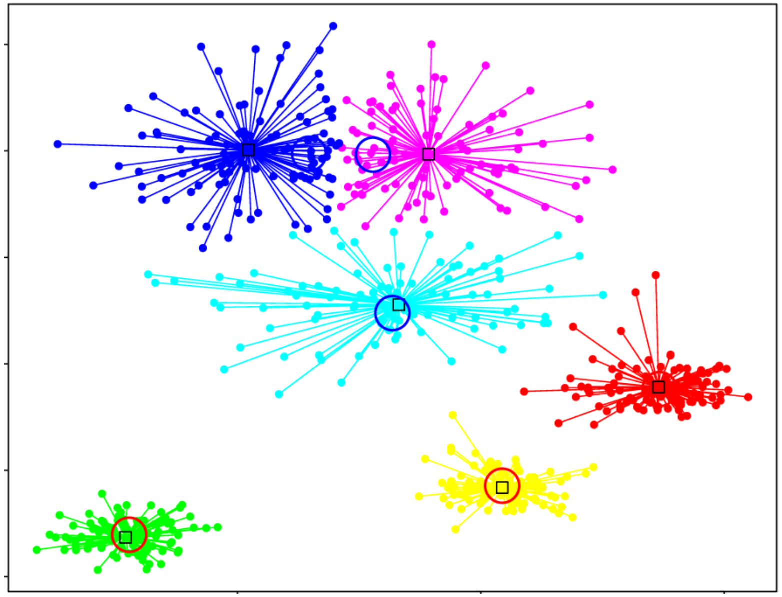
Similarity-Based Mixture Models: Examples

x7 data set clustered by similarity-based mixture clustering

x7 data



result affinity propagation



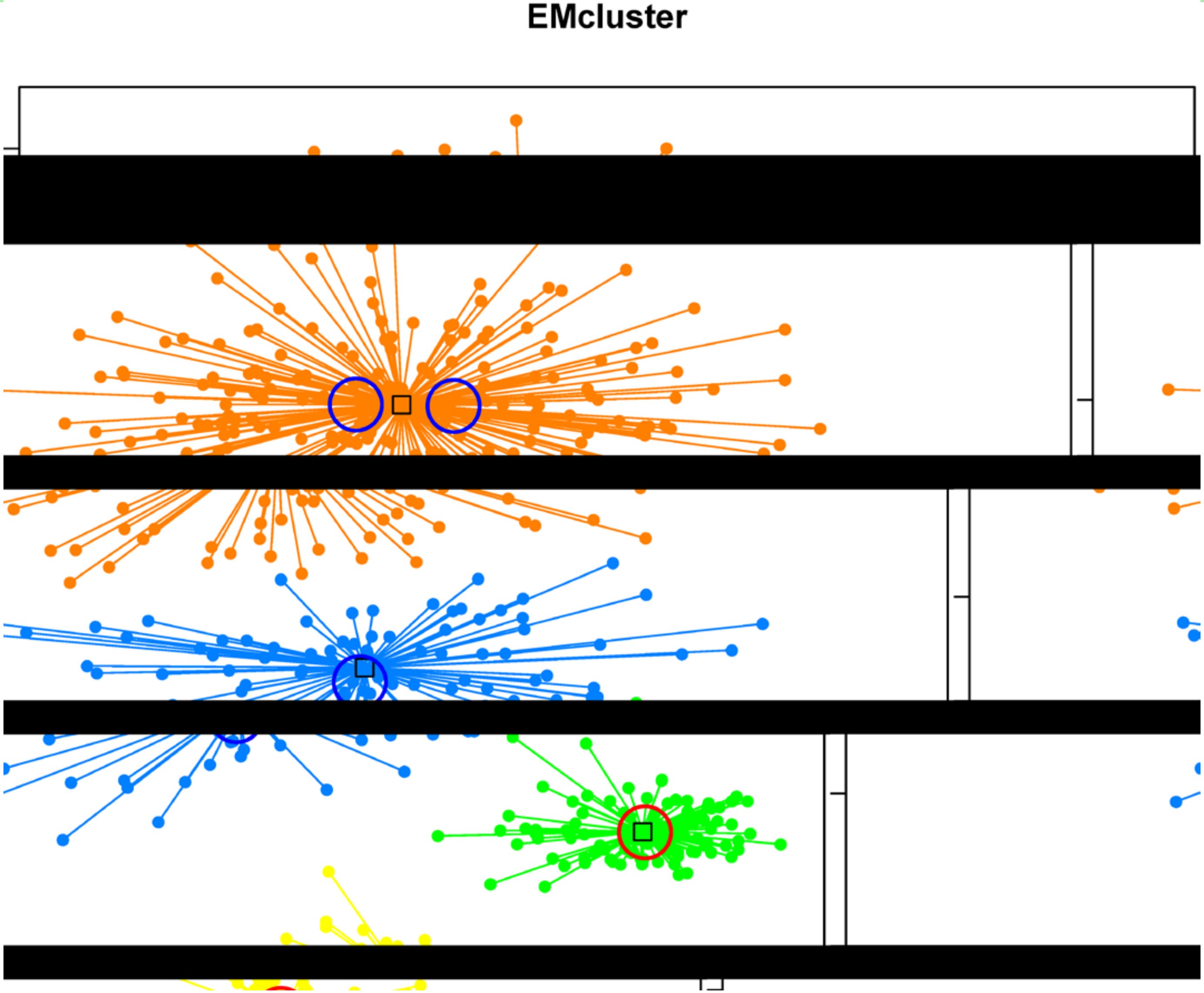
Similarity-Based Mixture Models: Examples

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine



[redacted]

[redacted]

[redacted]

[redacted]

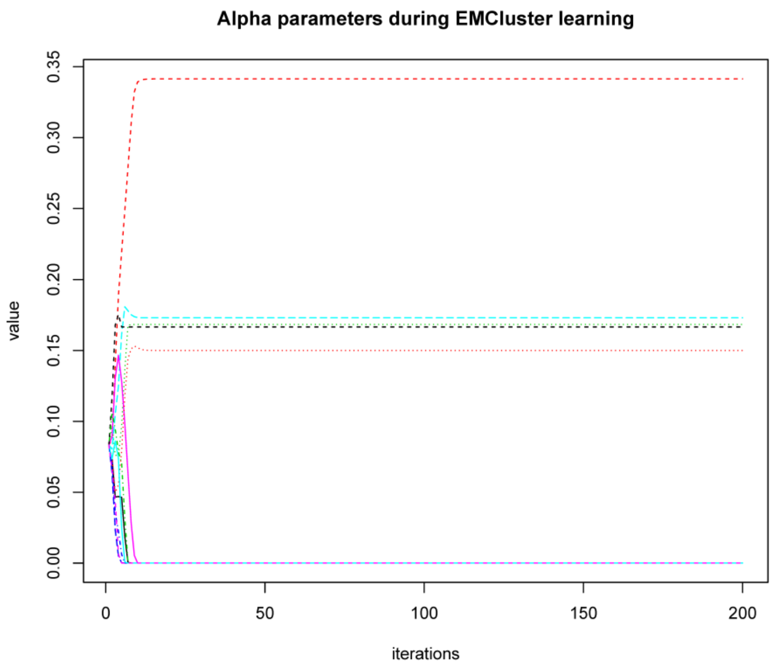
Similarity-Based Mixture Models: Examples

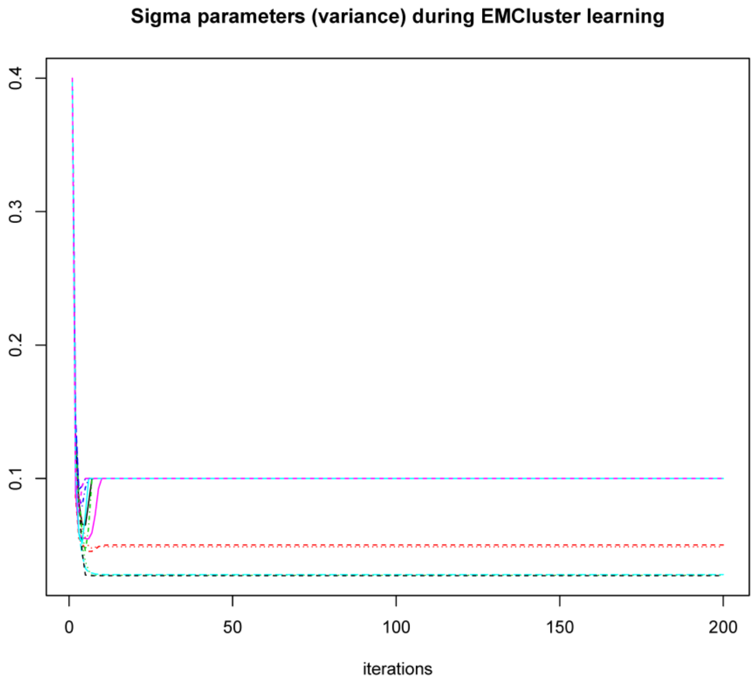
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine





Similarity-Based Mixture Models: Examples

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

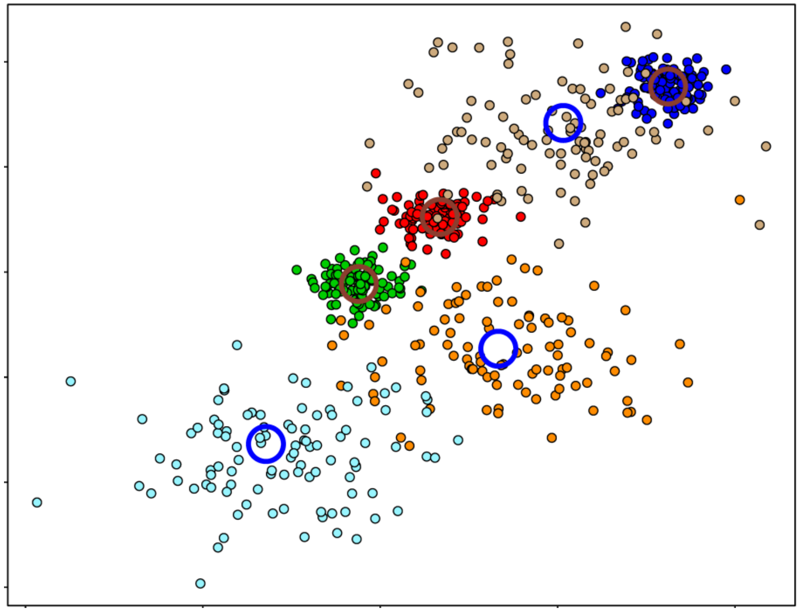
8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

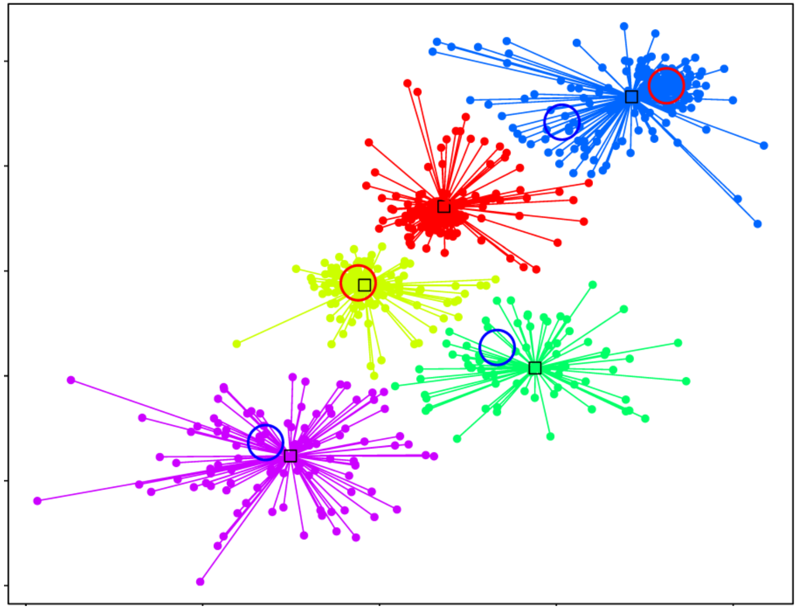
10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

x7A

x7A data



result affinity propagation



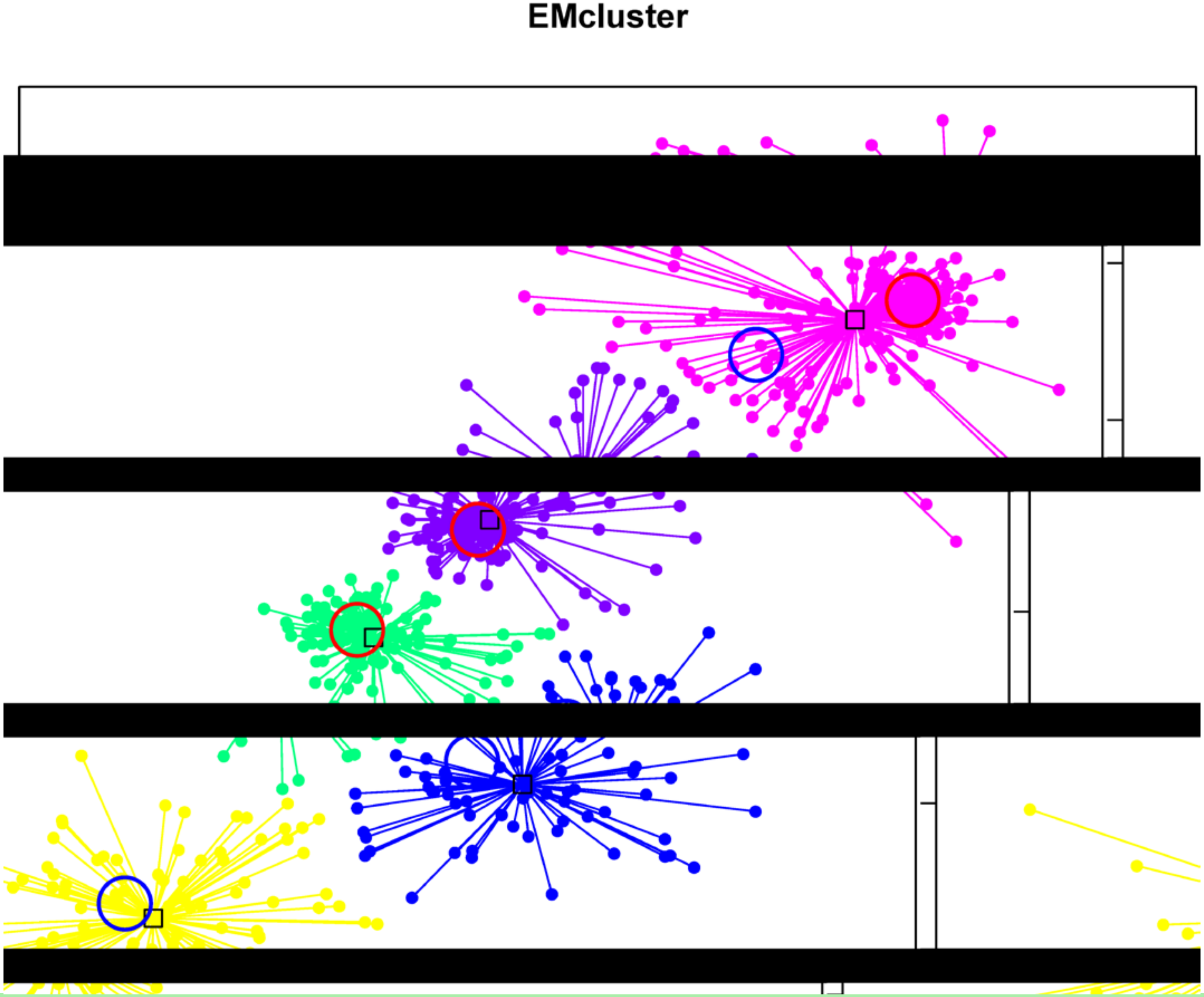
Similarity-Based Mixture Models: Examples

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

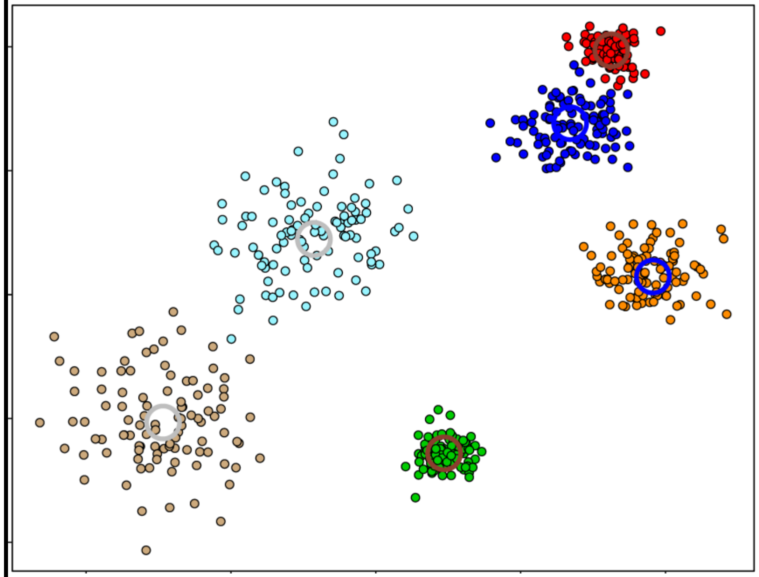
10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine



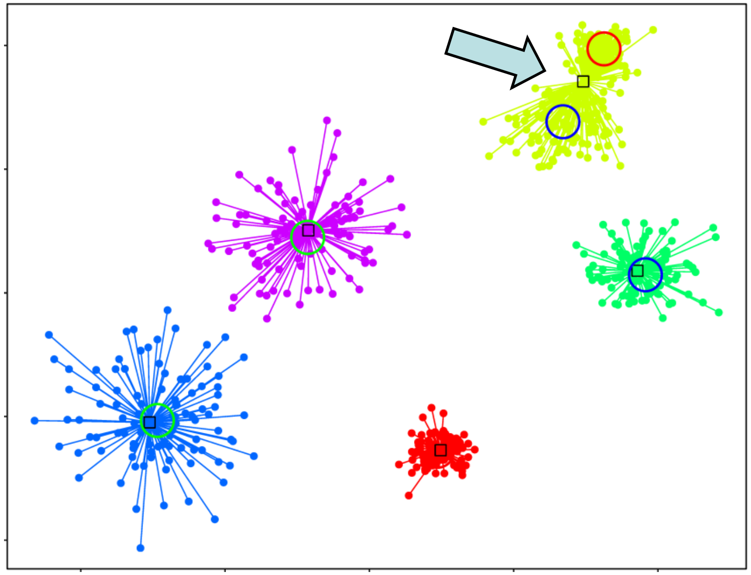
Similarity-Based Mixture Models: Examples

x9

x9 data



result affinity propagation



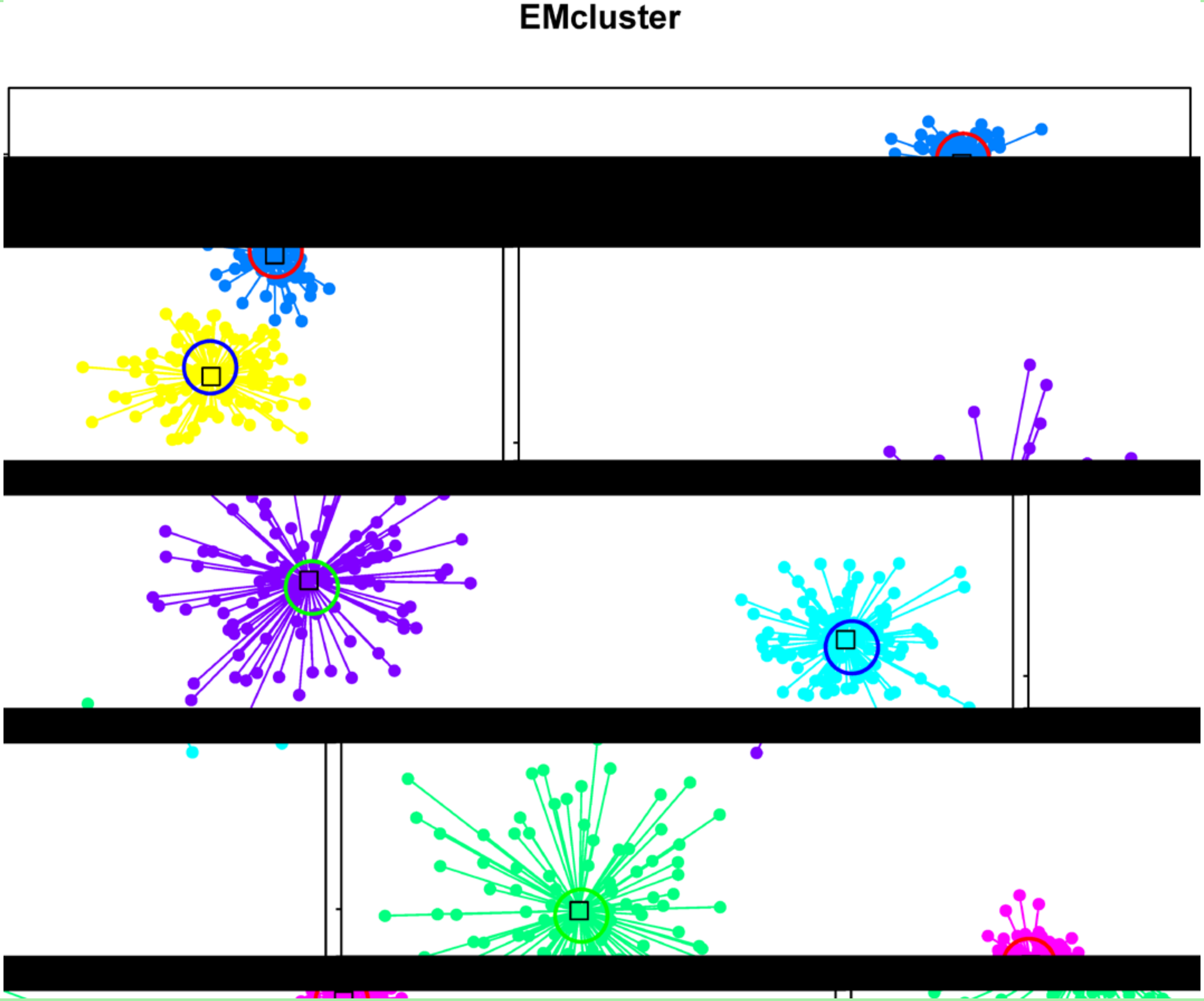
Similarity-Based Mixture Models: Examples

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine



[redacted]

[redacted]

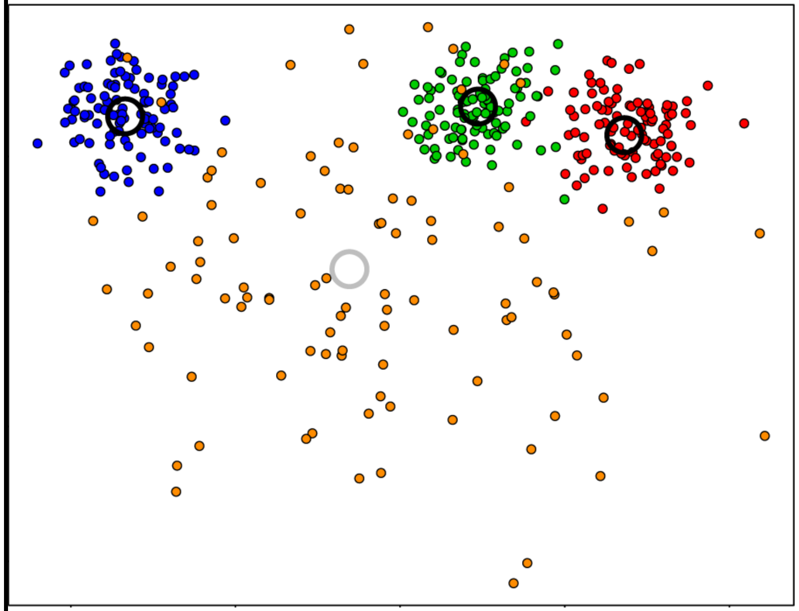
[redacted]

[redacted]

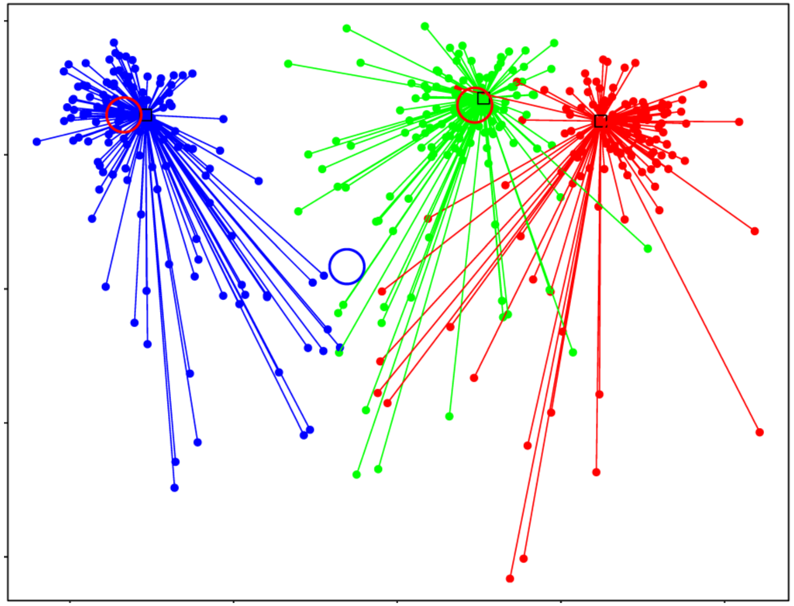
Similarity-Based Mixture Models: Examples

d6

d6 data



result affinity propagation



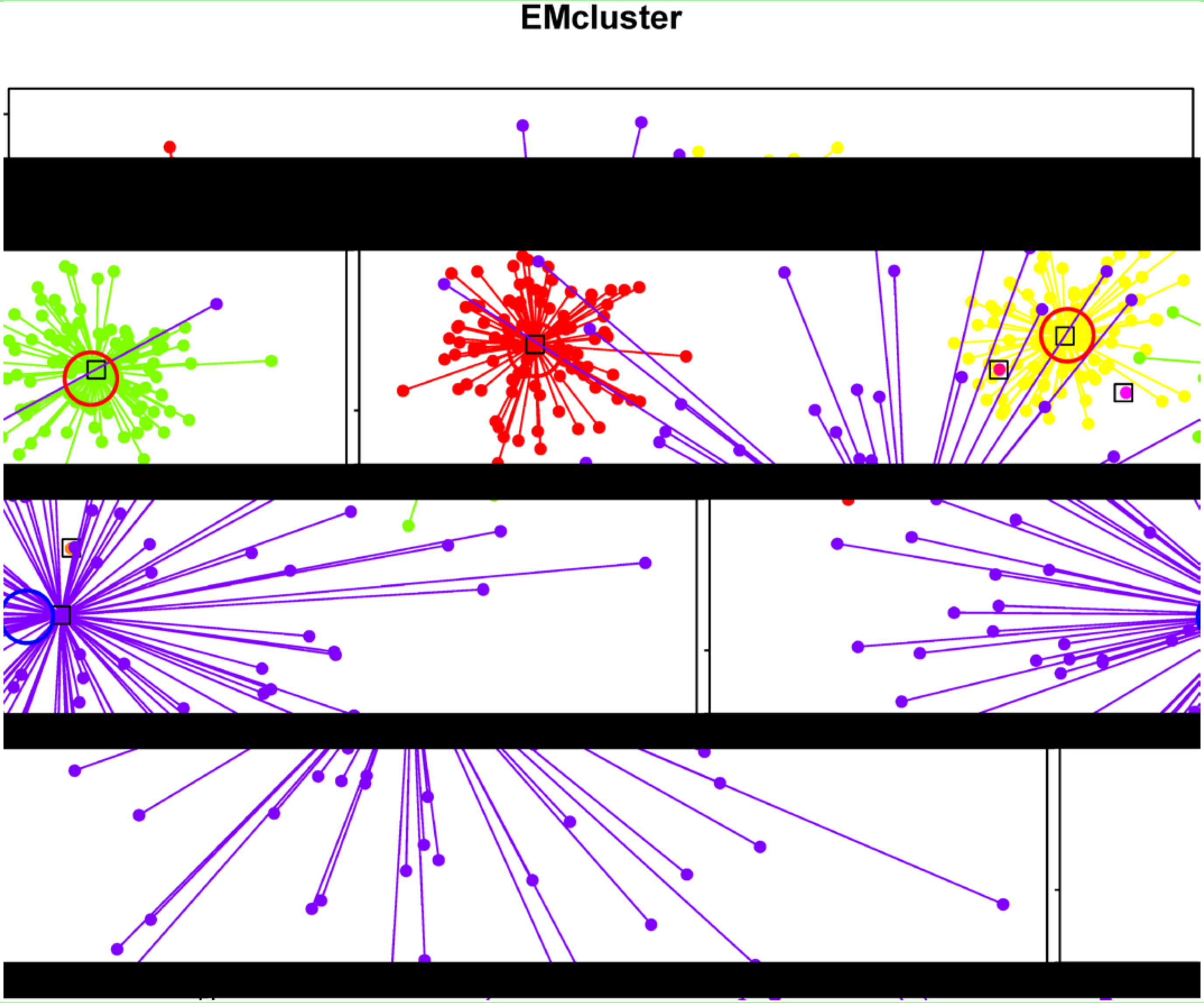
Similarity-Based Mixture Models: Examples

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine



[redacted]

[redacted]

[redacted]

[redacted]

Similarity-Based Mixture Models: Examples

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

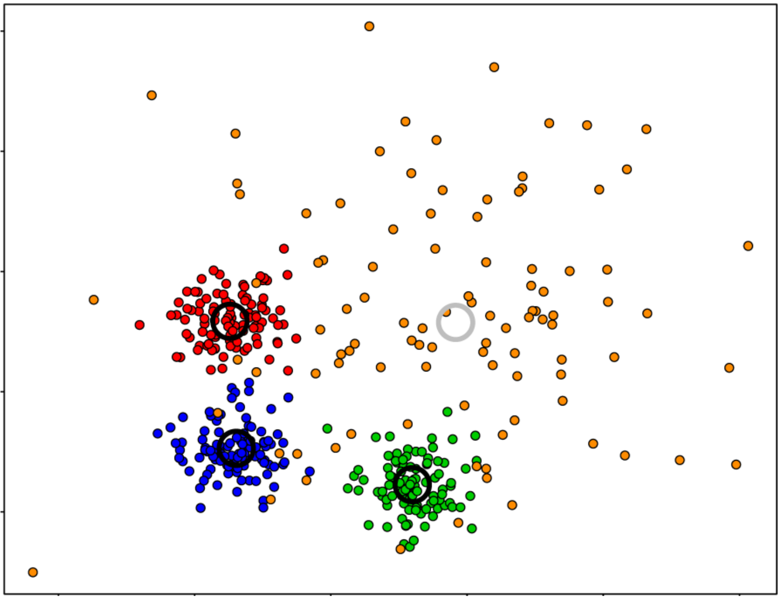
8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

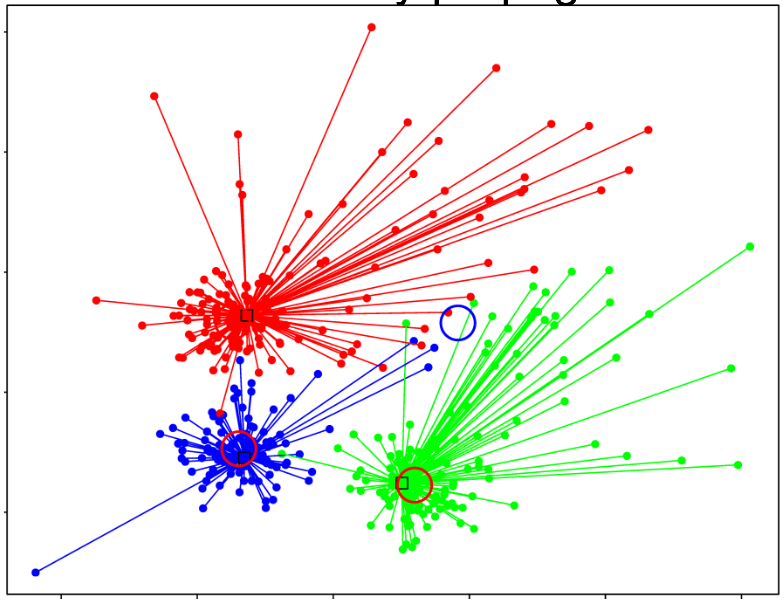
10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

d6A

d6A data



result affinity propagation



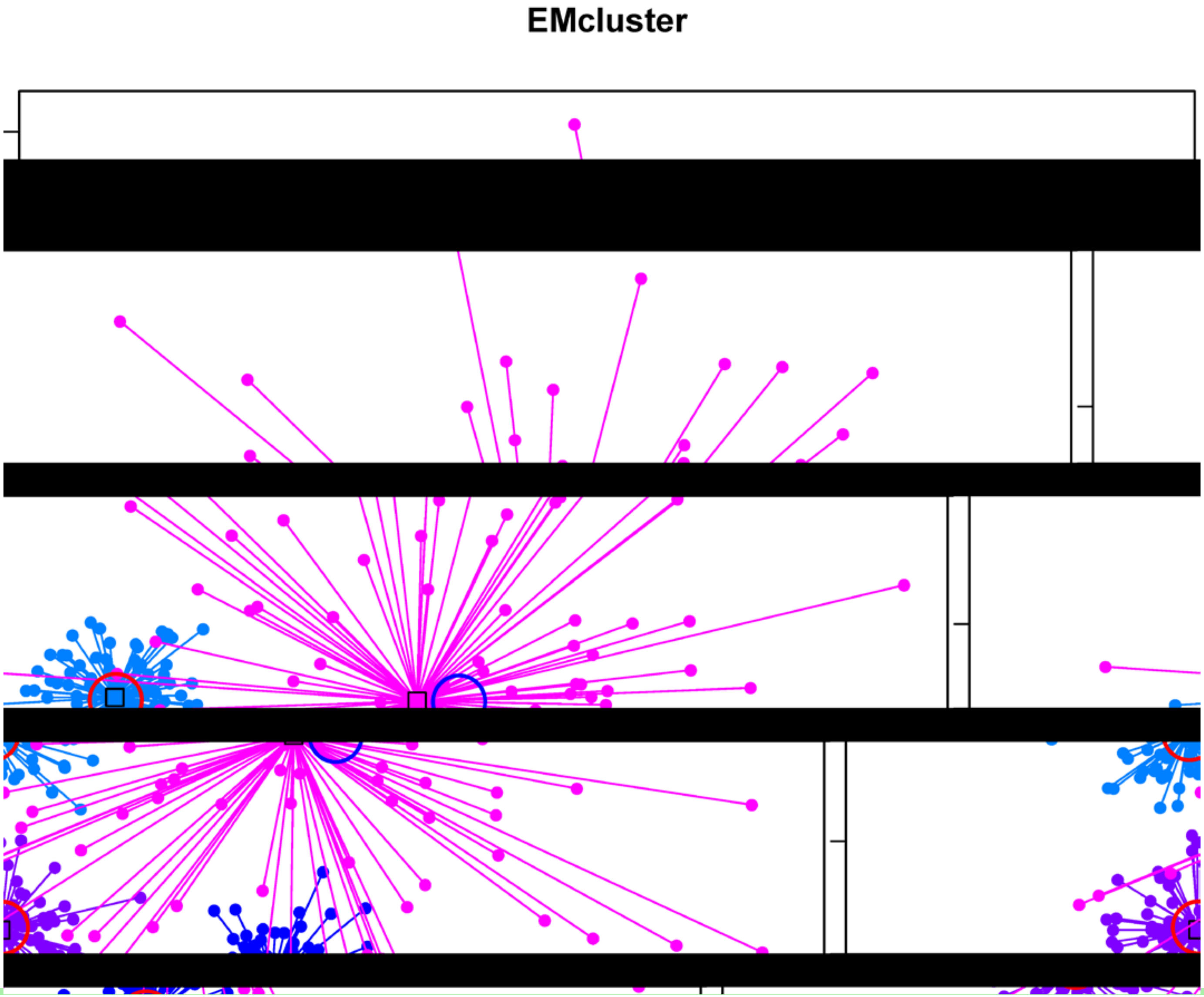
Similarity-Based Mixture Models: Examples

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine



EMcluster

Similarity-Based Mixture Models: Examples

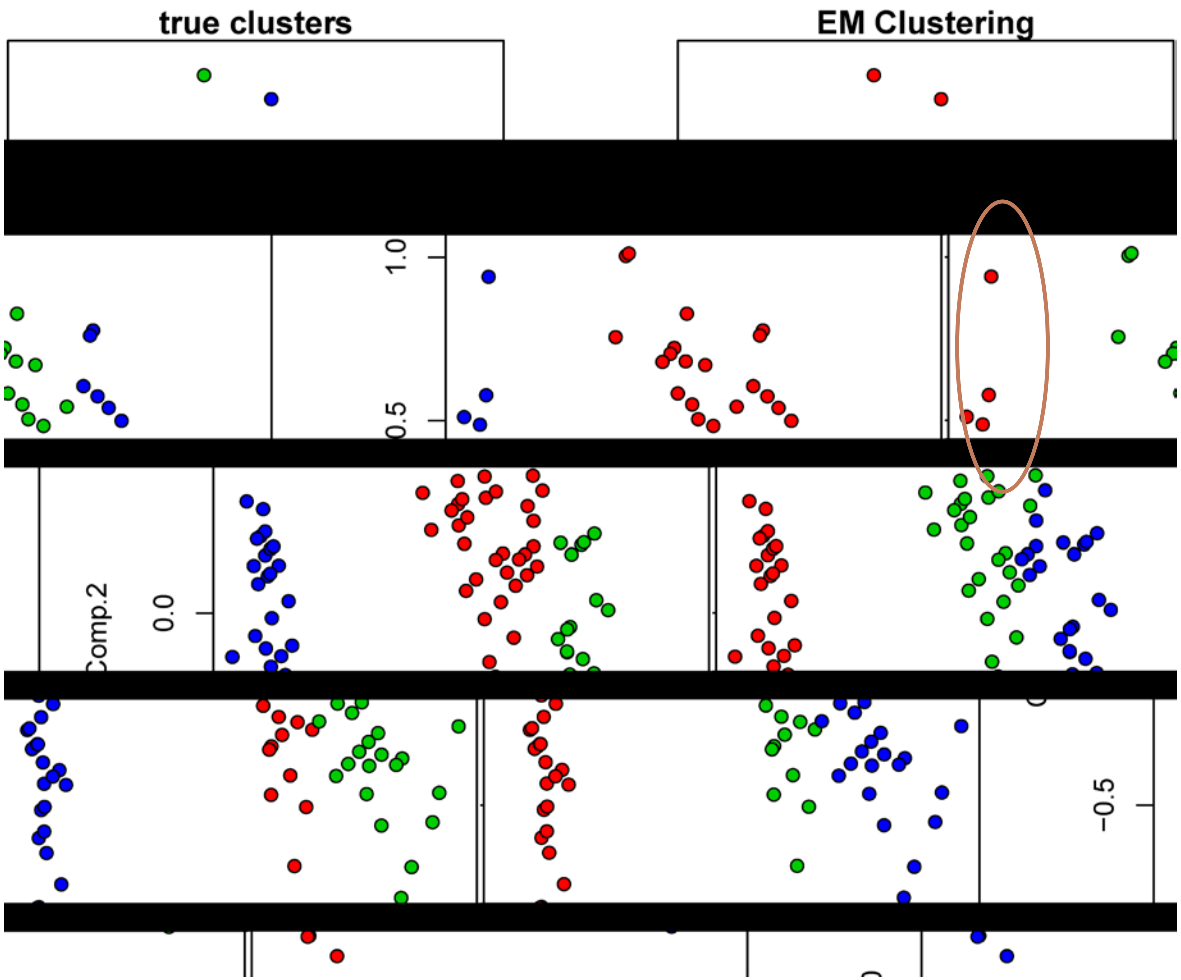
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

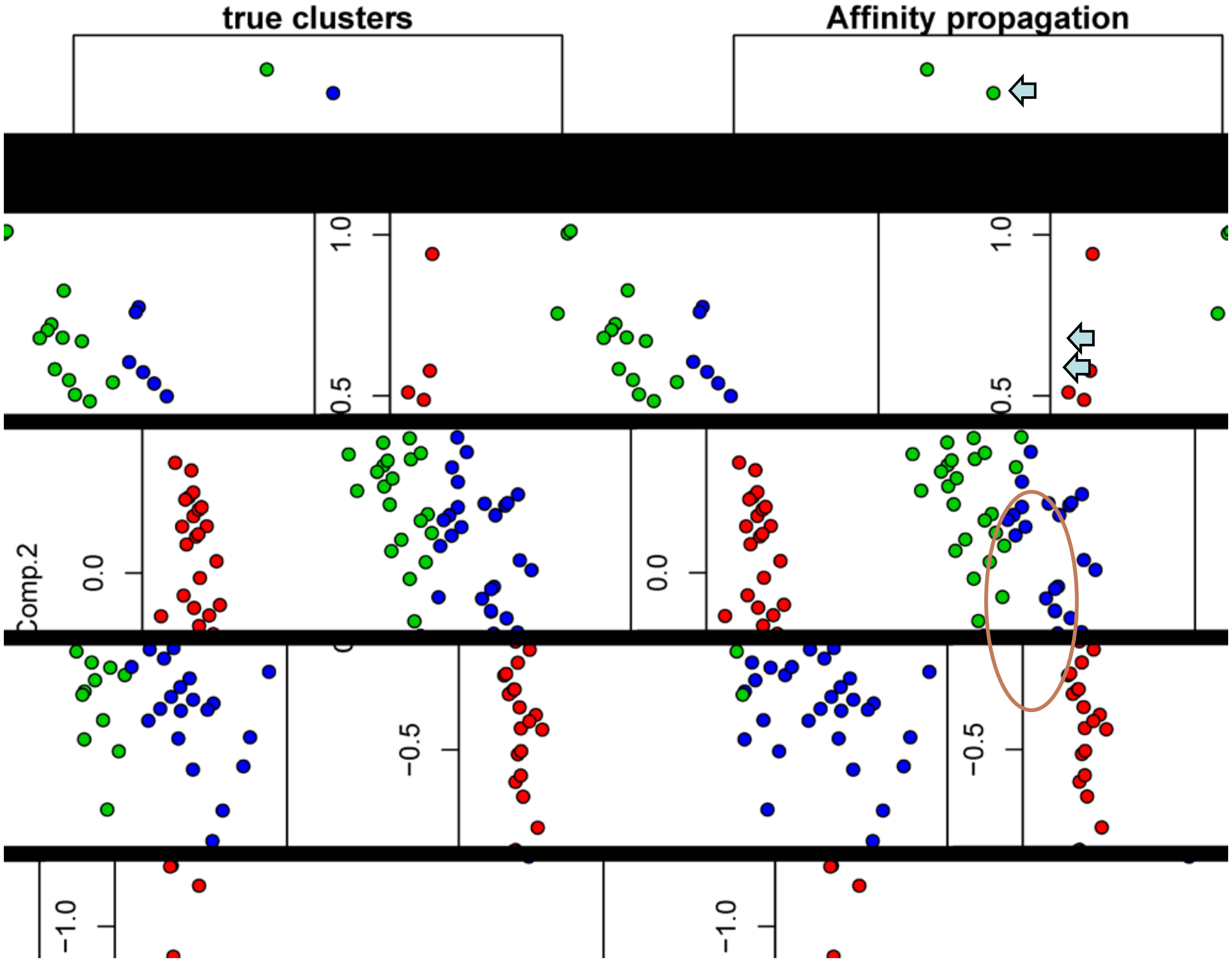
9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

similarity-based mixture clustering of the iris data

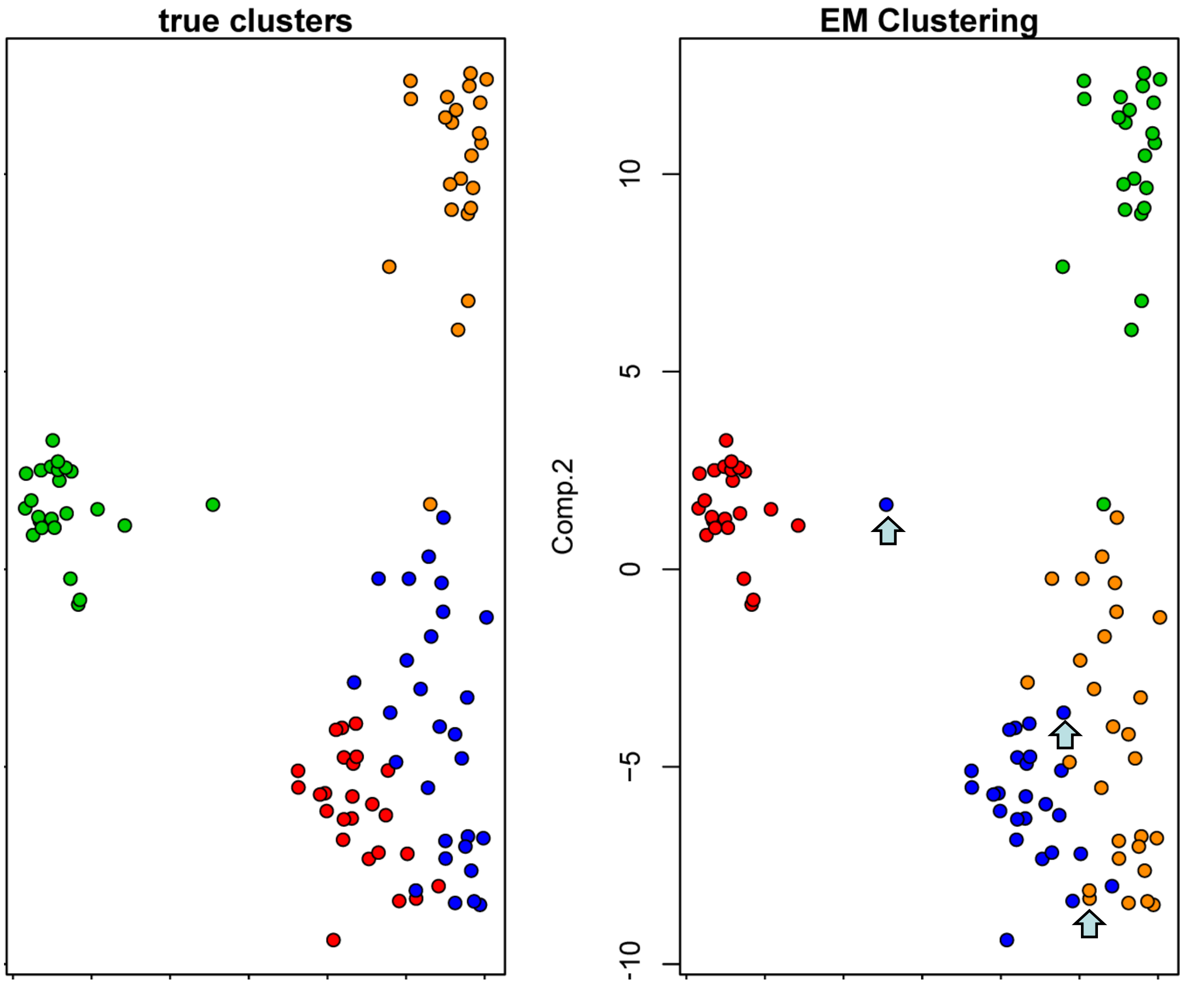


Similarity-Based Mixture Models: Examples

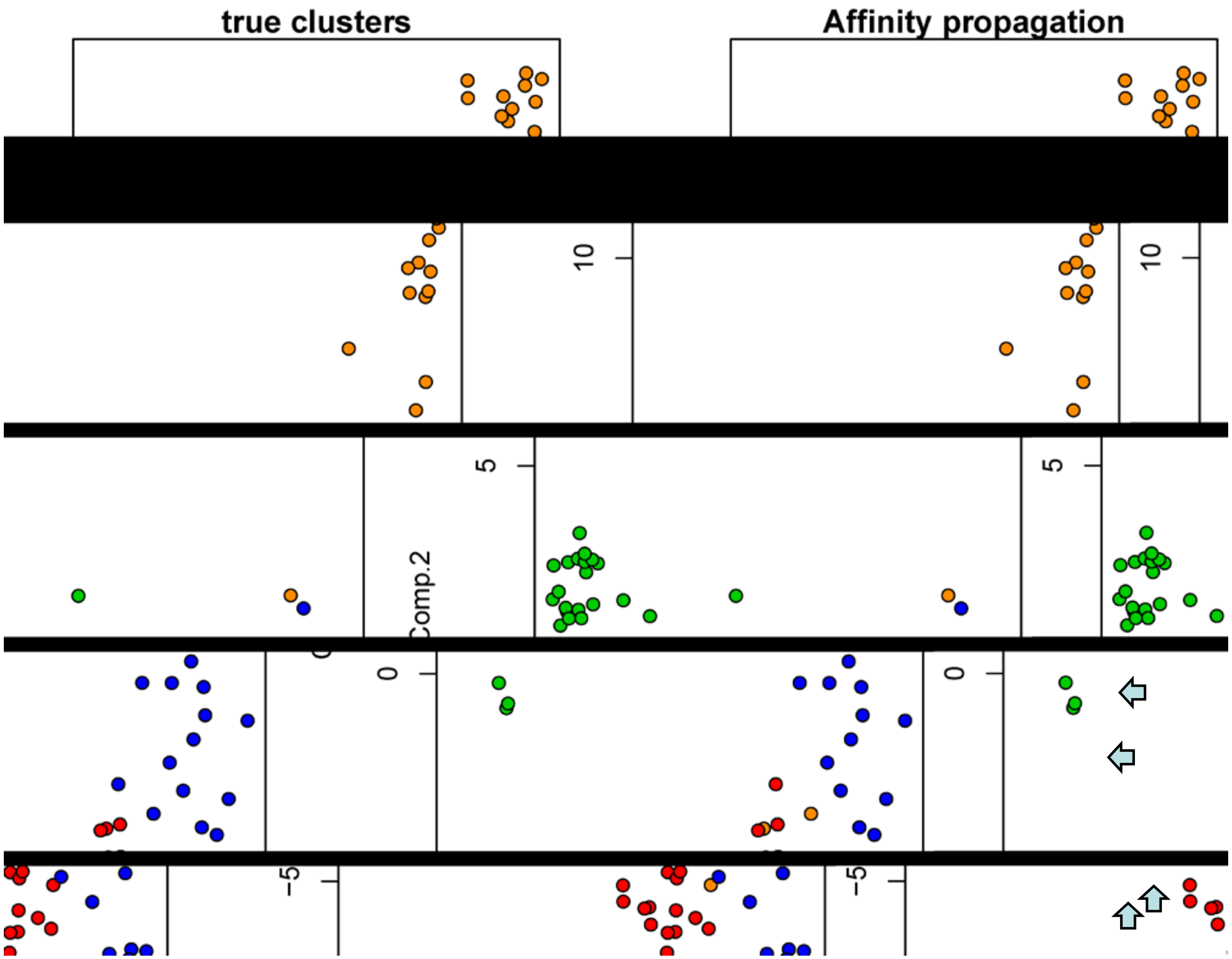


Similarity-Based Mixture Models: Examples

similarity-based mixture clustering of the multiple tissues



Similarity-Based Mixture Models: Examples



Biclustering

Biclustering simultaneously clusters the rows and the columns of a matrix

Bicluster:

- pair of a set of row elements and a set of column elements
- column elements are similar to each other on row elements and vice versa

row and column elements can belong to multiple or to no bicluster

contrast to standard clustering

- row elements clustered only on a subgroup of column elements
- multiple or not cluster membership

Biclustering

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

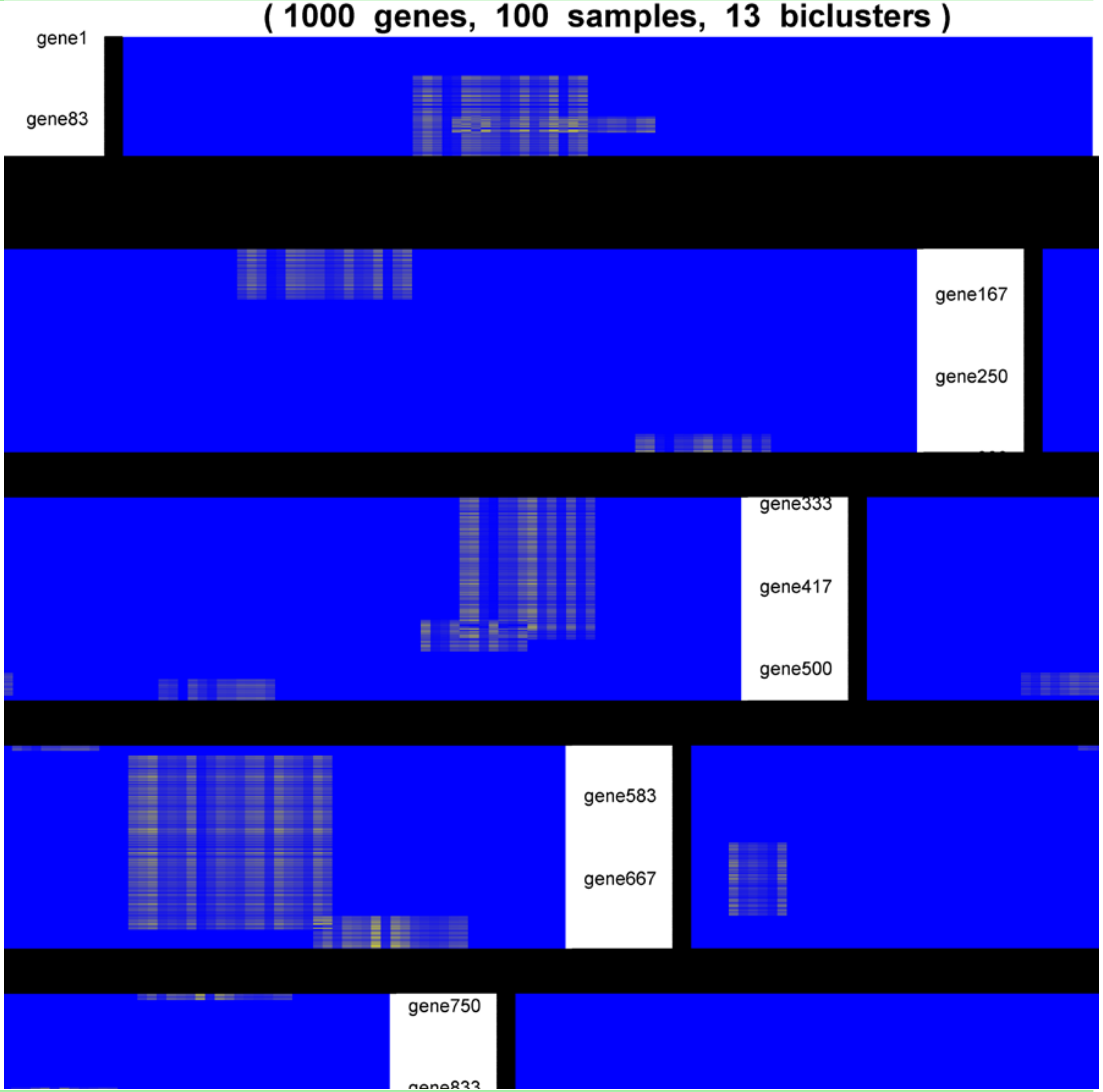
8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

examples of biclusters

for visualization the biclusters consist of adjacent row and column elements



Biclustering: Types of Biclusters

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

Types of bicluster:

- constant values
- constant rows values
- constant column values
- additive coherent values
- multiplicative coherent values
- general coherent values

1.0	4.0	5.0	0.0	1.5
4.0	7.0	8.0	3.0	4.5
3.0	6.0	7.0	2.0	3.5
5.0	8.0	9.0	4.0	5.5
2.0	5.0	6.0	1.0	2.5

coherent values: additive

1.0	0.5	2.0	0.2	0.8
2.0	1.0	4.0	0.4	1.6
3.0	1.5	6.0	0.6	2.4
4.0	2.0	8.0	0.8	3.2
5.0	2.5	10.0	1.0	4.0

coherent values: multiplicative

2.0	2.0	2.0	2.0	2.0
2.0	2.0	2.0	2.0	2.0
2.0	2.0	2.0	2.0	2.0
2.0	2.0	2.0	2.0	2.0
2.0	2.0	2.0	2.0	2.0

constant bicluster

1.0	1.0	1.0	1.0	1.0
2.0	2.0	2.0	2.0	2.0
3.0	3.0	3.0	3.0	3.0
4.0	4.0	4.0	4.0	4.0
4.0	4.0	4.0	4.0	4.0

constant row bicluster

1.0	2.0	3.0	4.0	5.0
1.0	2.0	3.0	4.0	5.0
1.0	2.0	3.0	4.0	5.0
1.0	2.0	3.0	4.0	5.0
1.0	2.0	3.0	4.0	5.0

constant column bicluster

Biclustering: Methods

Biclustering methods:

- variance minimization methods
- two-way clustering methods
- motif and pattern recognition methods
- probabilistic and generative approaches

variance minimization methods define clusters as blocks in the matrix with minimal deviation of their elements: Cheng-Church δ-biclusters; δ-cluster methods search for blocks of elements having a deviation below δ; δ-pClusters search sub-matrices with pairwise edge differences less than δ

Two-way clustering methods apply conventional clustering to the columns and rows and (iteratively) combine the results: Coupled Two-Way Clustering (CTWC), Interrelated Two-Way Clustering (ITWC), Double Conjugated Clustering (DCC)

Biclustering: Methods

Motif and pattern recognition methods define a bicluster as samples sharing a common pattern or motif: xMOTIF, Order-Preserving Sub-Matrices (OPSM), Spectral clustering (SPEC), Iterative Signature Algorithm (ISA)

Probabilistic and generative methods use model-based techniques to define biclusters: Statistical-Algorithmic Method for Bicluster Analysis (SAMBA), Probabilistic Relational Models (PRMs), ProBic, cMonkey, plaid models, Bayesian BiClustering model (BBC), Factor Analysis for Bicluster Acquisition (FABIA)

BBC and FABIA are generative models:

1. well-understood model selection techniques (maximum likelihood)
2. hyperparameter selection within the Bayesian framework
3. signal-to-noise ratios
4. model comparisons via the likelihood or posterior
5. tests like the likelihood ratio test
6. global model to explain all data

Biclustering: FABIA

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

FABIA biclustering defines a bicluster as an outer product $\boldsymbol{u}\ \boldsymbol{y}^T$. The vector $\boldsymbol{u}$ corresponds to a **prototype column vector** that contains zeros for features not participating in the bicluster. The vector $\boldsymbol{y}$ is a vector of **factors** with which the prototype column vector is scaled for each sample and contains zeros for samples not participating in the bicluster.

Vectors containing many zeros or values close to zero are called **sparse vectors**.

$$\overbrace{\begin{pmatrix}0\\0\\0\\0\\1\\2\\3\\4\\0\\0\\0\\0\\0\\0\\0\end{pmatrix}}^{\boldsymbol{u}} * \overbrace{\begin{pmatrix}0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\end{pmatrix}}^{\boldsymbol{y}^T} = \overbrace{\begin{pmatrix}
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\\
0&0&0&0&0&0&0&2&4&6&8&10&0&0&0\\
0&0&0&0&0&0&0&3&6&9&12&15&0&0&0\\
0&0&0&0&0&0&0&4&8&12&16&20&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{pmatrix}}^{\boldsymbol{u} * \boldsymbol{y}^T}$$

Biclustering: FABIA

model for $\boldsymbol{l}$ biclusters and additive noise:

$$\boldsymbol{X} \ = \ \sum_{j=1}^{l} \boldsymbol{u}_j \ \boldsymbol{y}_j^T \ + \ \boldsymbol{\Upsilon} \ = \ \boldsymbol{Y} \ \boldsymbol{U}^T \ + \ \boldsymbol{\Upsilon}$$

data matrix: $\boldsymbol{X} \in \mathbb{R}^{n \times m}$
additive noise: $\boldsymbol{\Upsilon} \in \mathbb{R}^{n \times m}$
sparse prototype vector: $\boldsymbol{u}_i \in \mathbb{R}^m$
sparse vector of factors: $\boldsymbol{y}_j \in \mathbb{R}^n$ ($\boldsymbol{y}_j^T = (y_{1j}, \ldots, y_{nj})$ values for samples)
sparse prototype matrix: $\boldsymbol{U} \in \mathbb{R}^{m \times l}$
sparse factor matrix: $\boldsymbol{Y} \in \mathbb{R}^{n \times l}$

$$\boldsymbol{x}_i \ = \ \sum_{j=1}^{l} \boldsymbol{u}_j \ y_{ij} \ + \ \boldsymbol{\epsilon}_i \ = \ \boldsymbol{U} \ \tilde{\boldsymbol{y}}_i \ + \ \boldsymbol{\epsilon}_i$$

column of the noise matrix: $\boldsymbol{\epsilon}_i$
$\boldsymbol{i}$-th row of factor matrix: $\tilde{\boldsymbol{y}}_i = (y_{i1}, \ldots, y_{il})^T$ (one value per bicluster)

Biclustering: FABIA

factor analysis model with $\boldsymbol{l}$ factors:

$$\boldsymbol{x} \ = \ \sum_{j=1}^{l} \boldsymbol{u}_j \ \tilde{y}_j \ + \ \boldsymbol{\epsilon} \ = \ \boldsymbol{U} \ \tilde{\boldsymbol{y}} \ + \ \boldsymbol{\epsilon}$$

observations: $\boldsymbol{x}$

loading matrix: $\boldsymbol{U}$

$\boldsymbol{j}$-th factor: $\tilde{y}_j$

vector of factors: $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \ldots, \tilde{y}_l)^T$

additive noise: $\boldsymbol{\epsilon} \in \mathbb{R}^m$

Assumptions:

noise is independent of factors

$\tilde{\boldsymbol{y}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$

$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi})$

$\boldsymbol{\Psi} \in \mathbb{R}^{l \times l}$ is diagonal (independent Gaussian noise)

parameter $\boldsymbol{U}$ explains the dependent (common) and $\boldsymbol{\Psi}$ the independent variance in the observations

Biclustering: FABIA

FABIA formulation requires sparse factors and loadings to represent biclusters. Sparseness is obtained by a component-wise independent Laplace distribution:

$$p(\tilde{\boldsymbol{y}}) \;=\; \left(\frac{1}{\sqrt{2}}\right)^{l} \prod_{j=1}^{l} e^{-\;\sqrt{2}\;|\tilde{y}_j|}$$

$$p(\boldsymbol{u}_j) \;=\; \left(\frac{1}{\sqrt{2}}\right)^{m} \prod_{k=1}^{m} e^{-\;\sqrt{2}\;|u_{kj}|}$$

Unfortunately, for FABIA, the likelihood is analytically intractable because of the Laplacian priors.

Expectation maximization (EM) has been extended to variational expectation maximization for model selection and finding the optimal parameters. The variational EM maximizes the posterior.

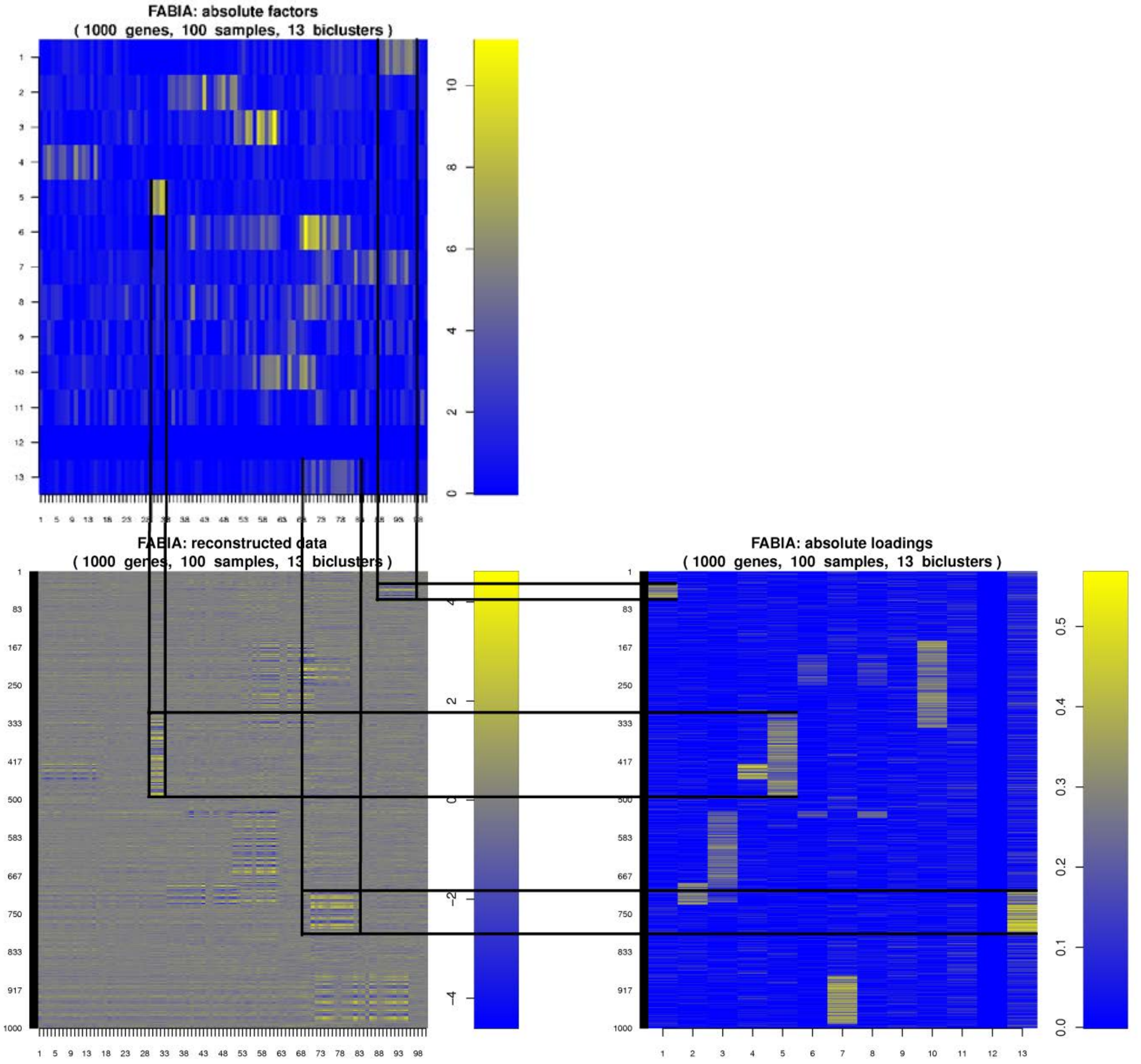
Biclustering: FABIA

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

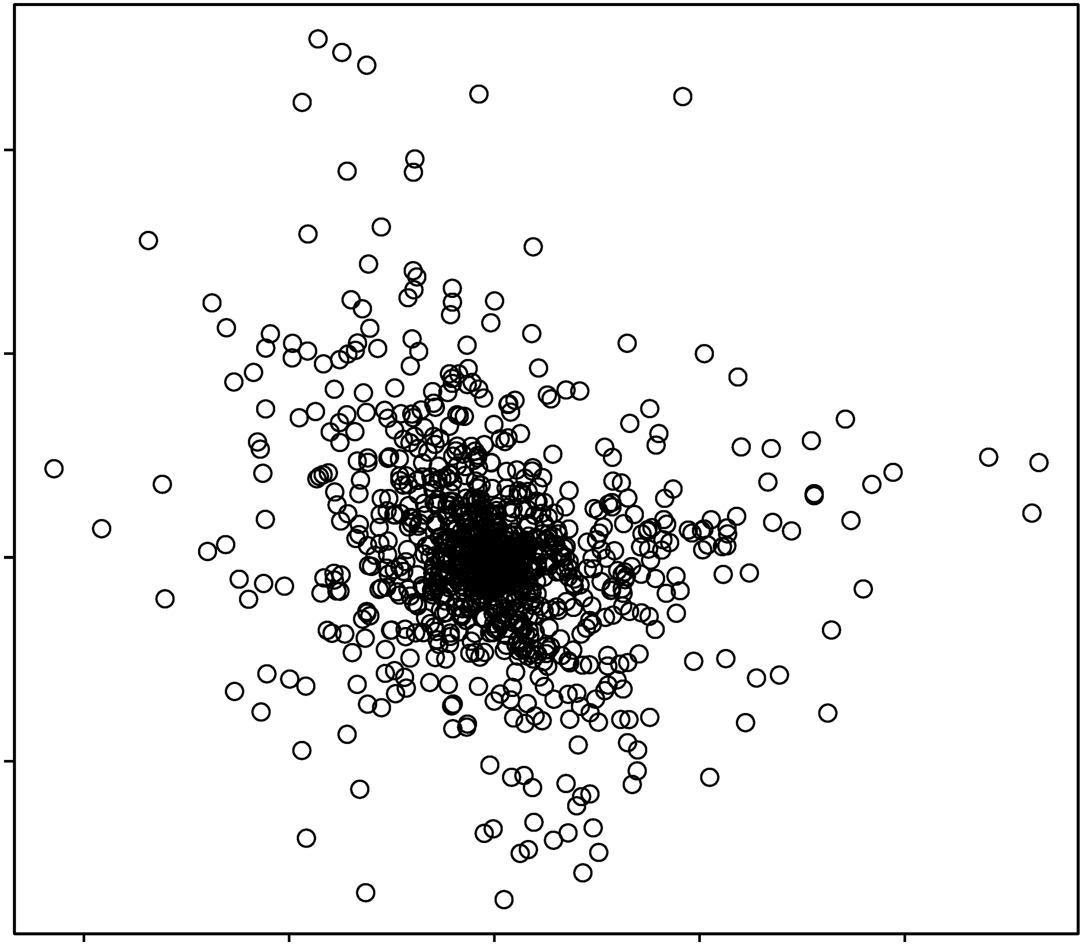
9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine



Biclustering: Examples

We test FABIA on a 50-dimensional data set with linearly mixed super-Gaussians



Biclustering: Examples

FABIA biclustering of the iris data

The loadings of FABIA are

```
             bicluster2 bicluster3 bicluster1
Sepal.Length  0.6490253 -0.2155847          0
Sepal.Width  -0.1828042 -0.3511878          0
Petal.Length  1.7465614  0.0000000          0
Petal.Width   0.7285857 -0.0453688          0
```

Only two biclusters have been found.

- first bicluster “bicluster2” focuses on “Petal.Length” which is correlated to “Petal.Width” and “Sepal.Length”. This bicluster is related to petal but includes also sepal length.
- second bicluster “bicluster3” removed “Petal.Length” and “Petal.Width” due to sparseness and focused on “Sepal.Width” including also “Sepal.Length”. This bicluster is related to sepal.

Biclustering: Examples

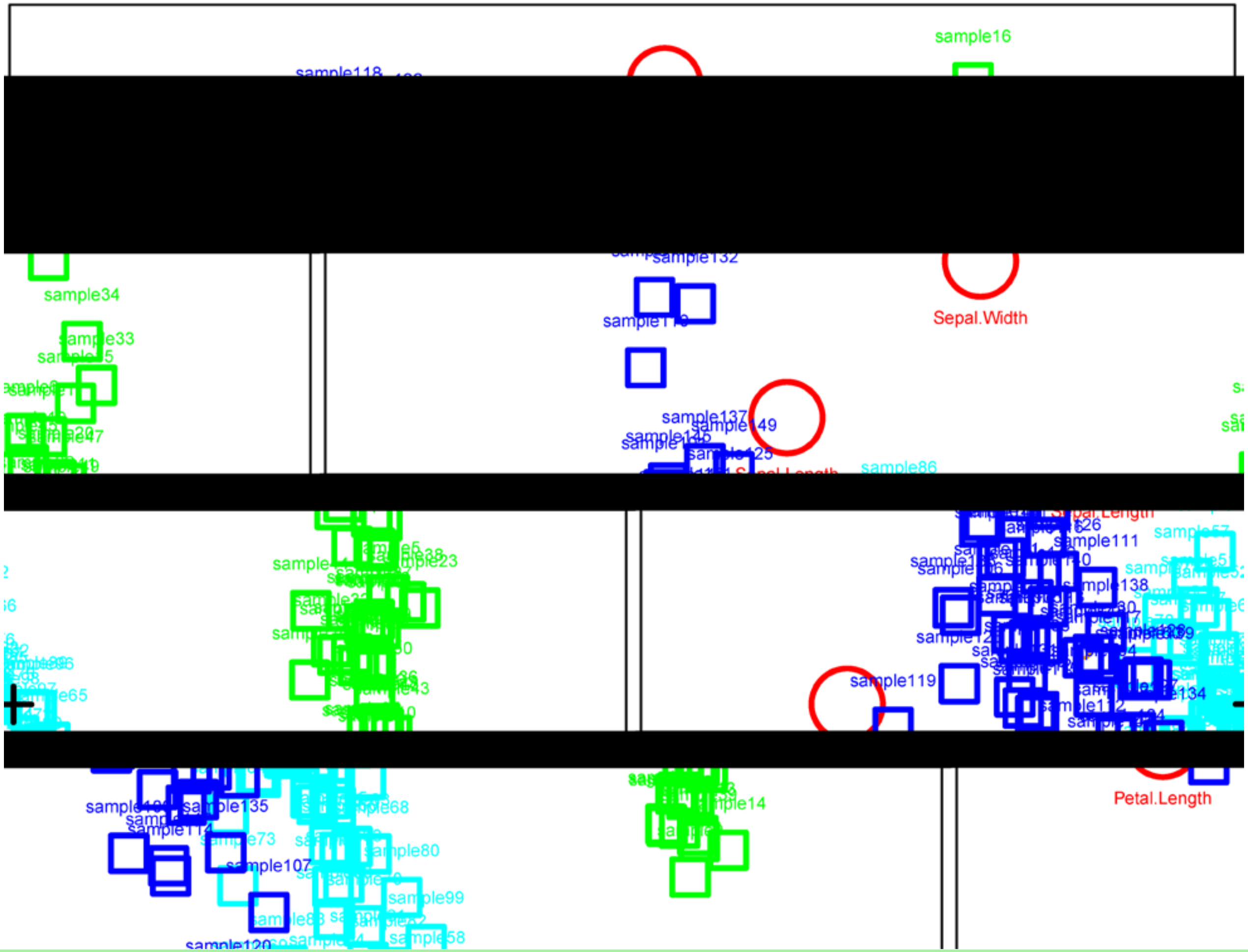
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

biplot that plots both the features and the samples into one plot


[redacted]
[redacted]
[redacted]

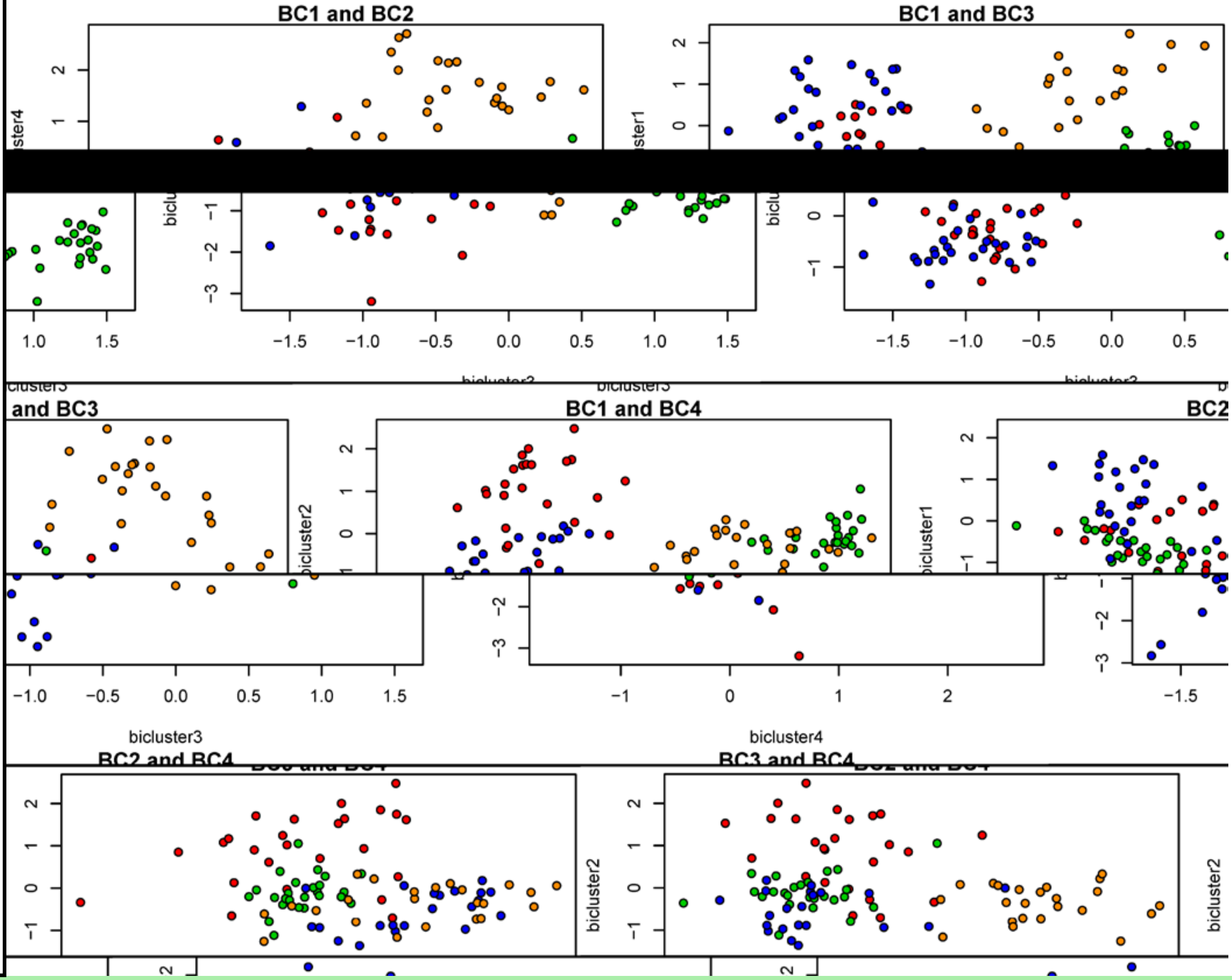
First bicluster on the x-axis is driven by Petal.Length and separates the species

Second bicluster is driven by Sepal.Width

Sepal.Length belongs to both clusters

Biclustering: Examples

FABIA applied to the multiple tissue data: 200 iterations



[redacted]

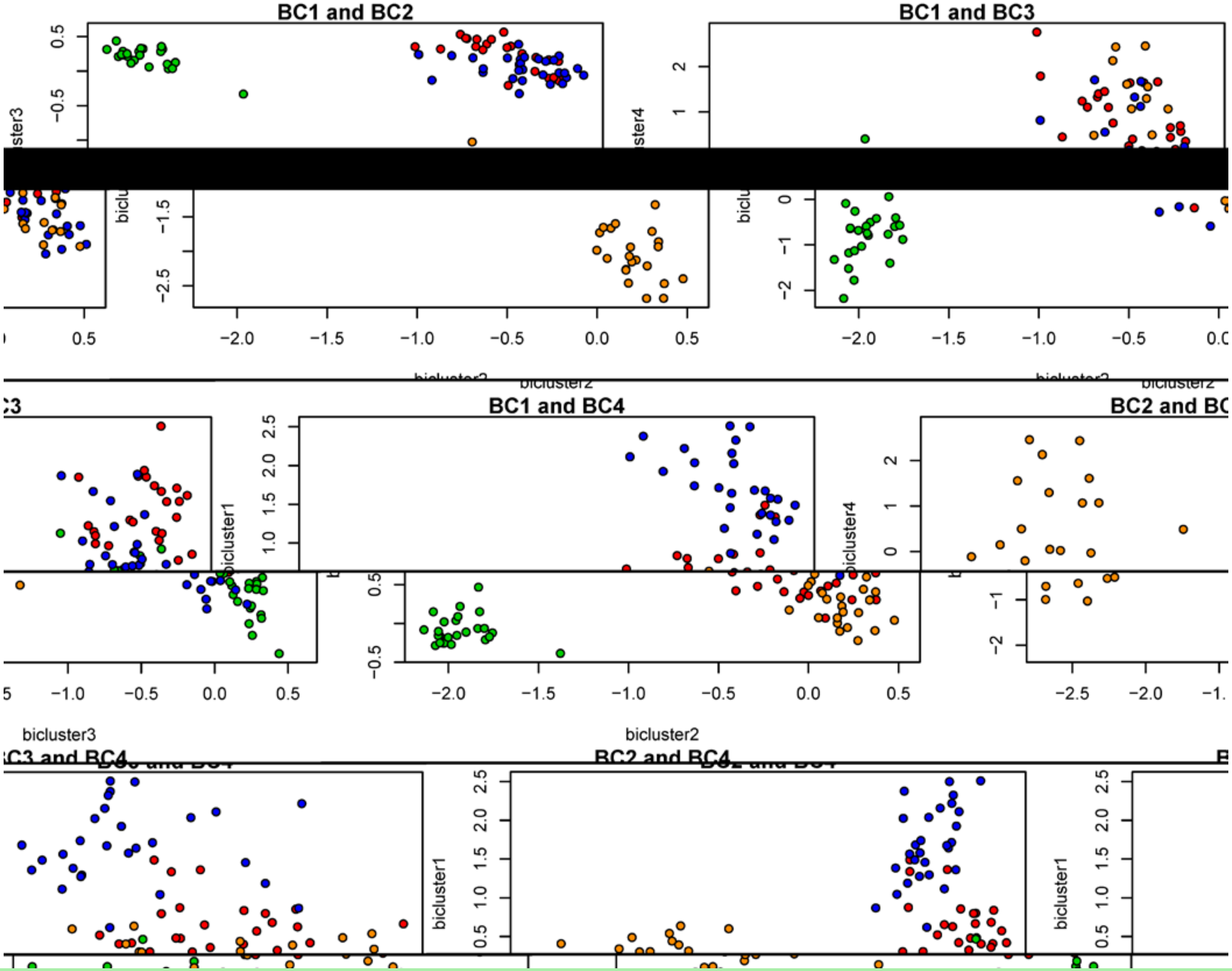
first bicluster (“bicluster3”) separates prostate (green) and colon (orange)

second bicluster (“bicluster4”) separates colon (orange)

fourth bicluster (“bicluster2”) separates breast (red), though not perfectly.

Biclustering: Examples

FABIA applied to the multiple tissue data: 500 iterations



[redacted]

first bicluster ("bicluster2") separates prostate (green)

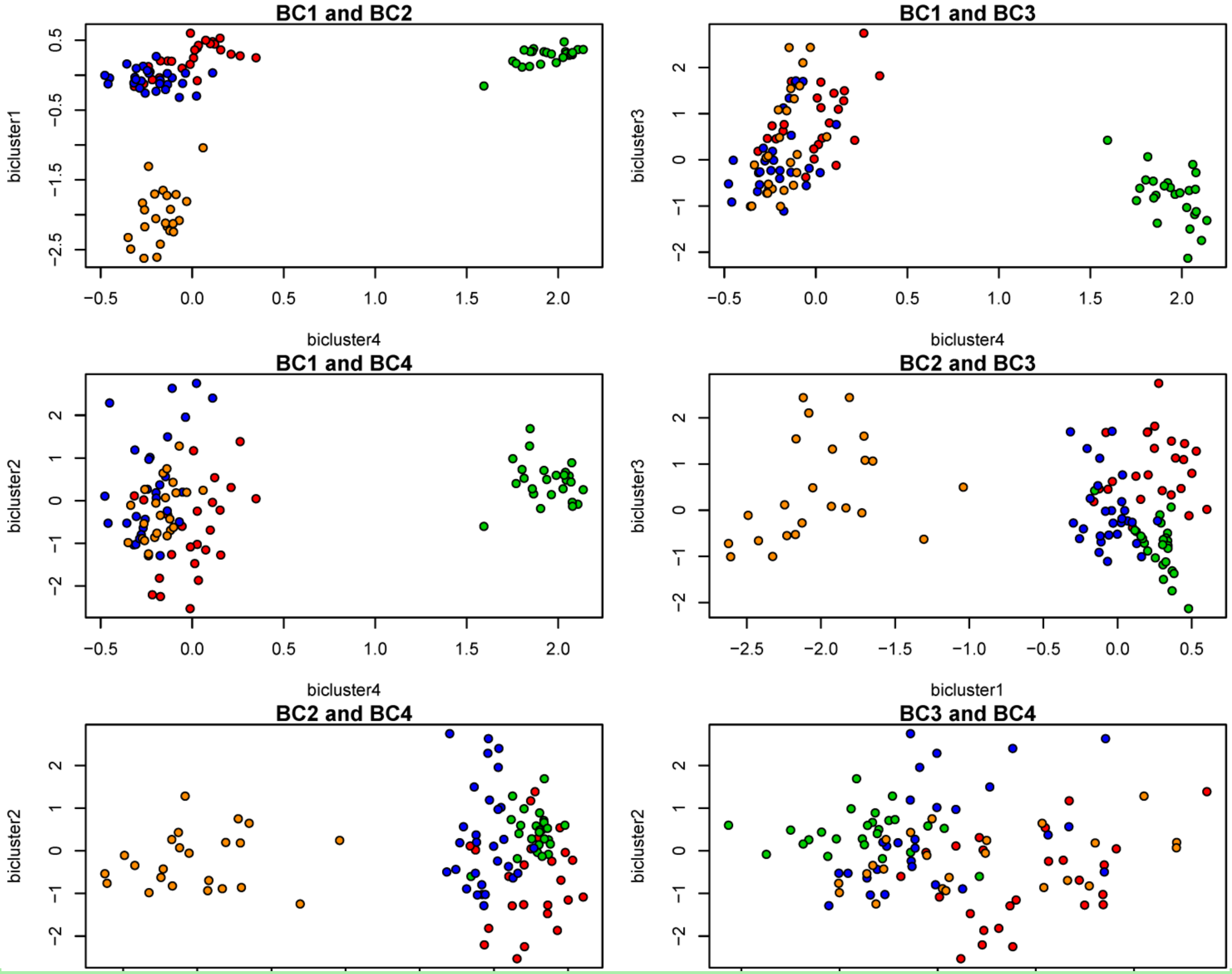
second bicluster ("bicluster3") separates colon (orange)

fourth bicluster ("bicluster2") separates lung (blue) from the rest

Separation is better than with 200 iterations

Biclustering: Examples

FABIA applied to the multiple tissue data: 2,000 iterations



first bicluster ("bicluster4") clearly separates prostate (green

The second bicluster ("bicluster1") separates colon (orange

third and fourth bicluster ("bicluster3/2") help to separate lung (blue) and breast (red)

more iterations: the solutions become more sparse and more clear separation

Biclustering: Examples

Since the FABIA solution is sparse, it allows for an interpretation.
most relevant genes of the first bicluster:

```
"KLK3"  "ACPP"  "KLK2"  "CUTL2" "RNF41"
```

- KLK3 is known as the prostate specific antigen. “Serum level of this protein, called PSA in the clinical setting, is useful in the diagnosis and monitoring of prostatic carcinoma.”
- ACPP “is synthesized under androgen regulation and is secreted by the epithelial cells of the prostate gland.”
- KLK2 “is primarily expressed in prostatic tissue and is responsible for cleaving pro-prostate-specific antigen into its enzymatically active form.”

This means that the most relevant genes of the first bicluster are all strongly associated with prostate tissue.

Biclustering: Examples

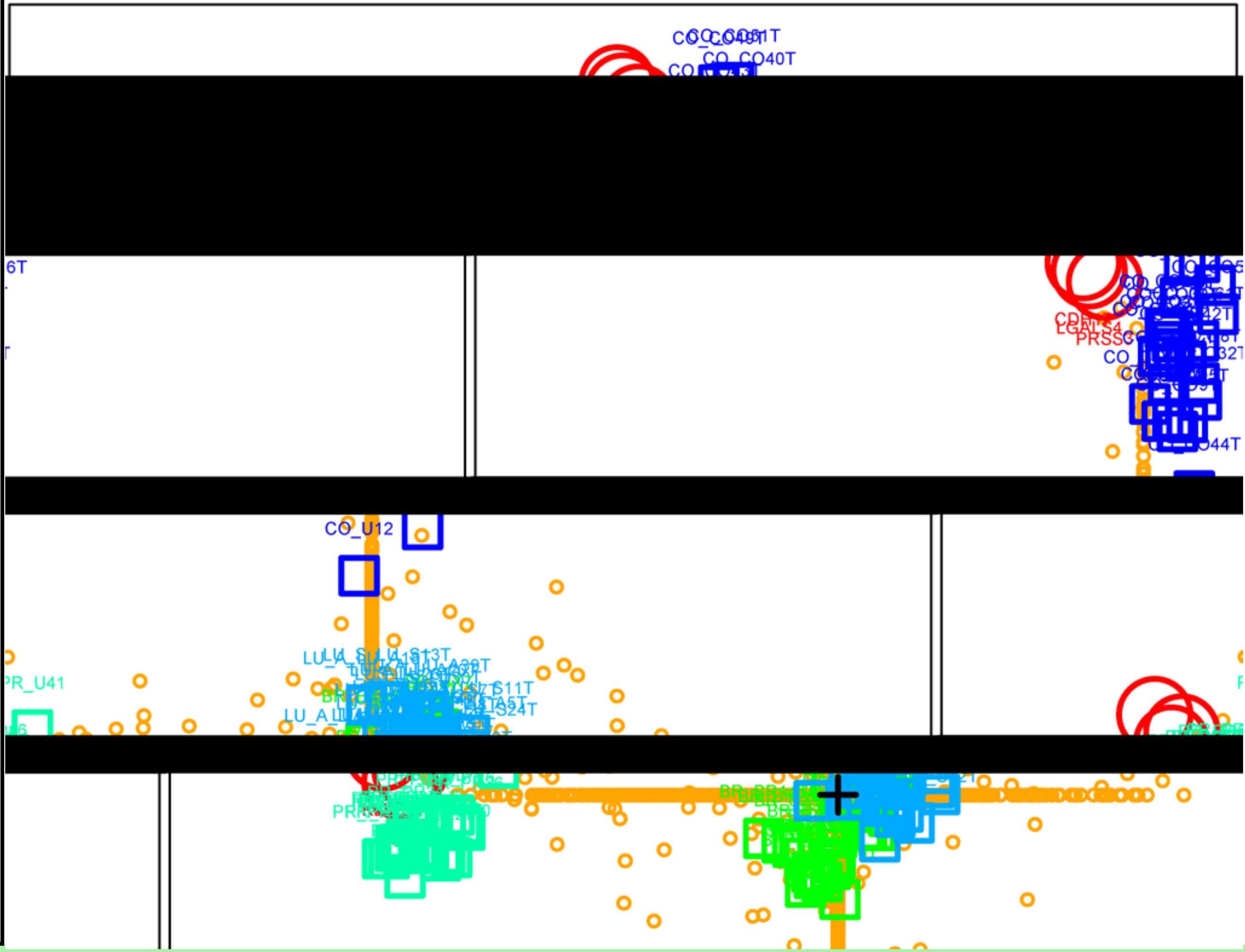
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

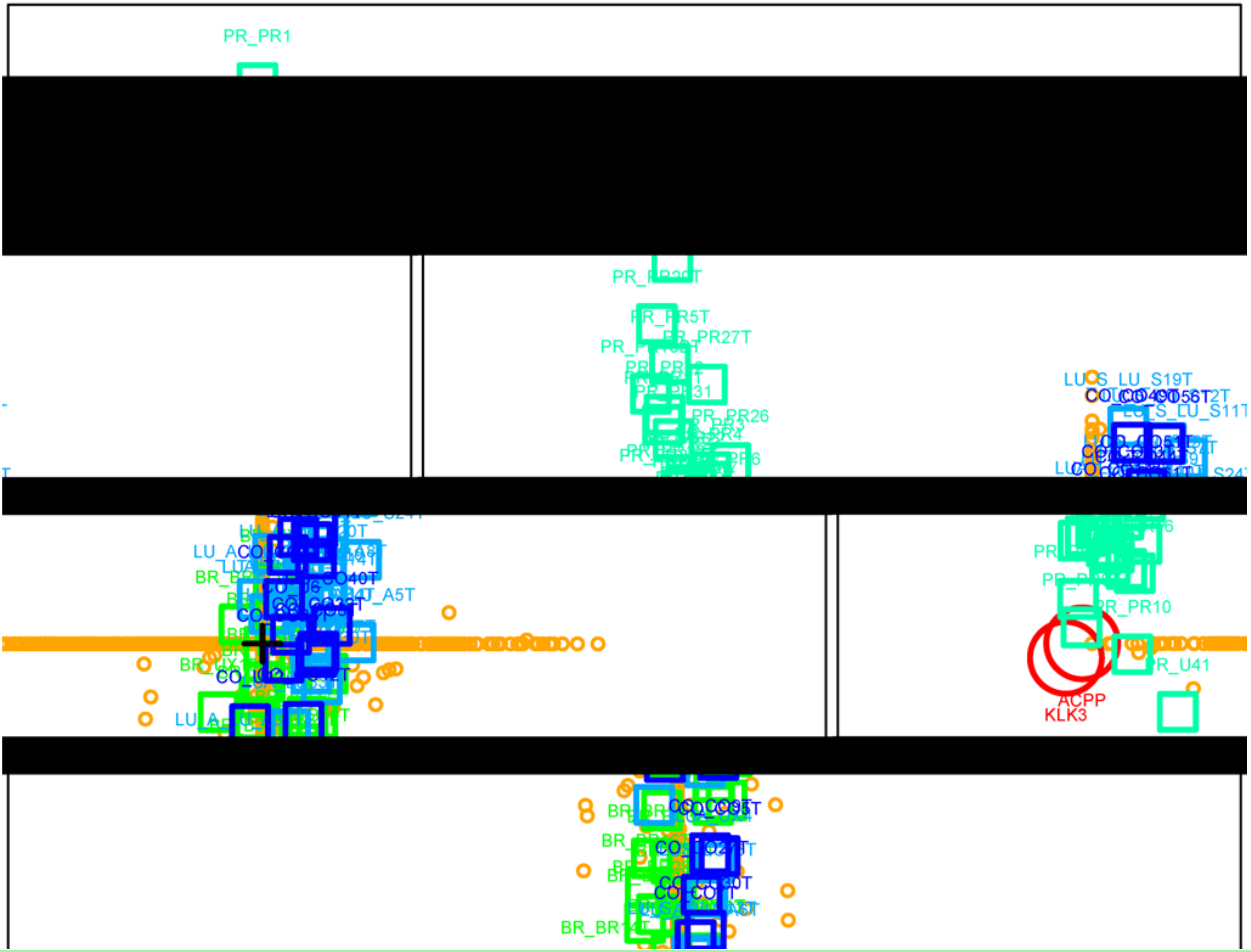
biplot of the first and second bicluster of FABIA (2000 iterations)


[redacted]
[redacted]
[redacted]

The large red circles are the features (genes) driving the bicluster while the golden small circles are features with minor influence. Sparseness pushes most features to zero in one direction (the axes are golden due to the features pushed to zero).
The turquoise prostate samples at the x-axis are separated from the other samples.

Biclustering: Examples

Biplot for the first and third bicluster: 2,000 iterations



[redacted]

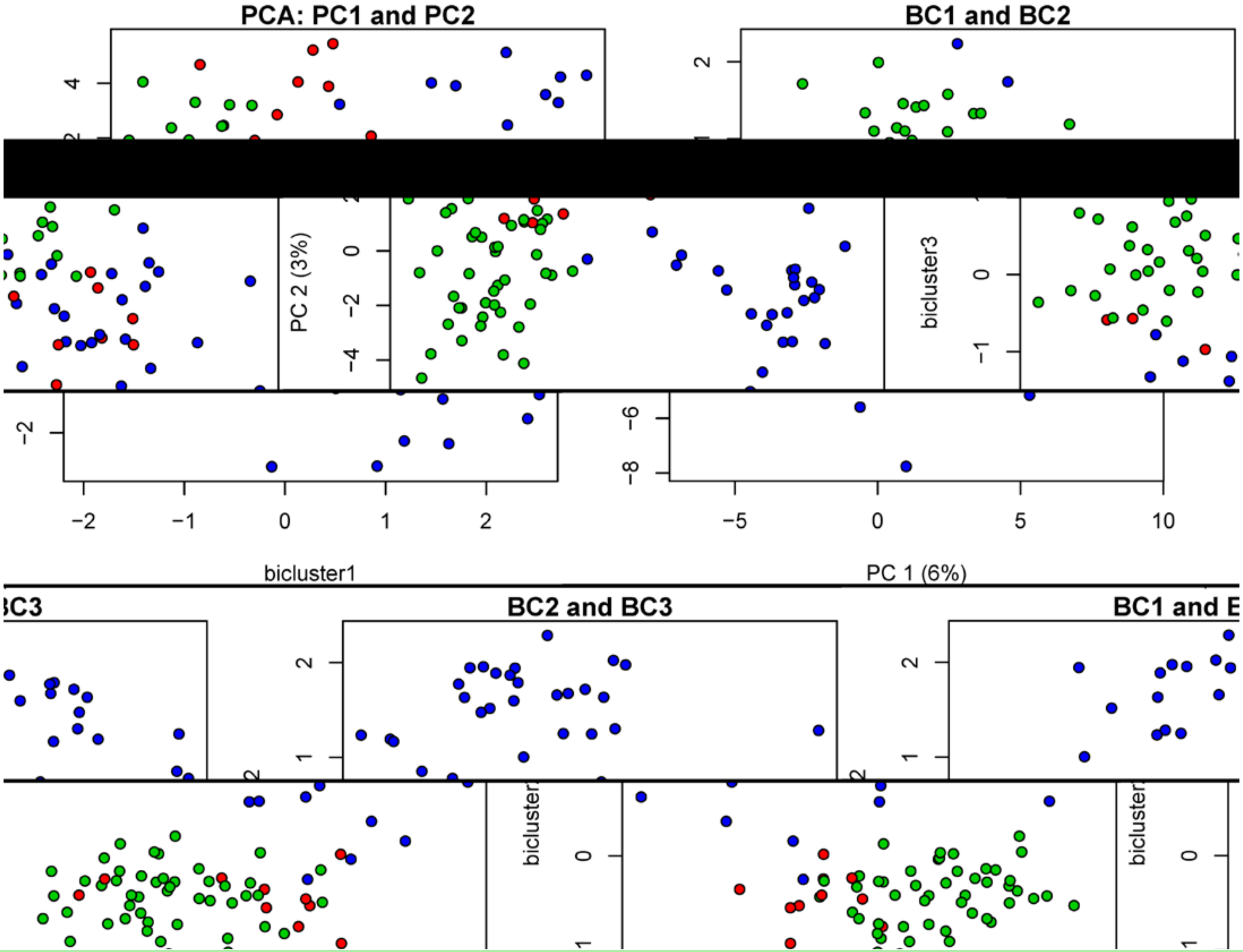
[redacted]

[redacted]

For the first bicluster at the x-axis the genes KLK3 (prostate specific antigen) and ACPP (secreted by the epithelial cells of the prostate gland) can be recognized. These genes separate the turquoise prostate samples at the x-axis from the other samples.

Biclustering: Examples

FABIA applied to the breast cancer data:



[redacted]

Both PCA and biclustering separate the blue class but have problems to separate the red class.

Biclustering: Examples

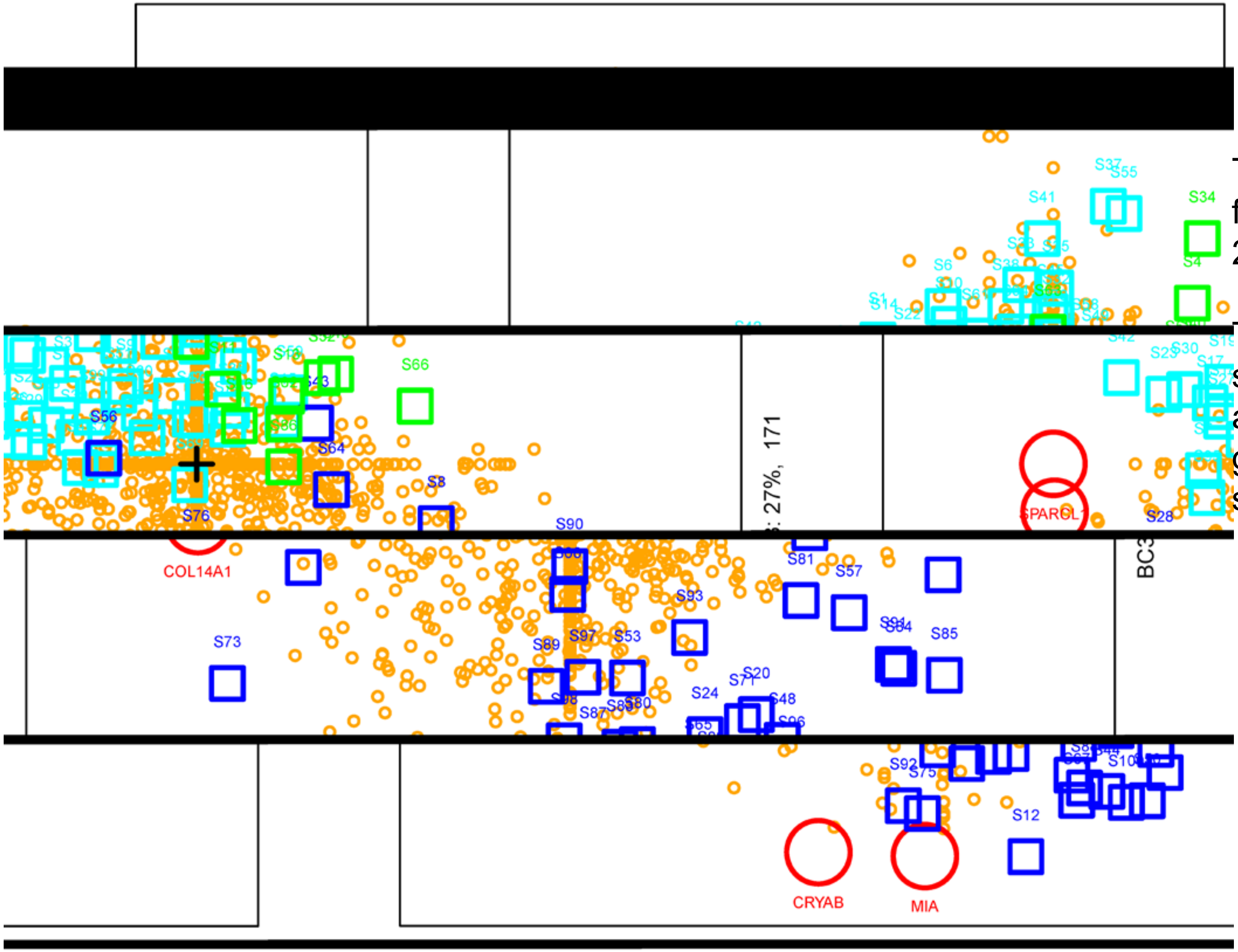
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

Biplot breast cancer

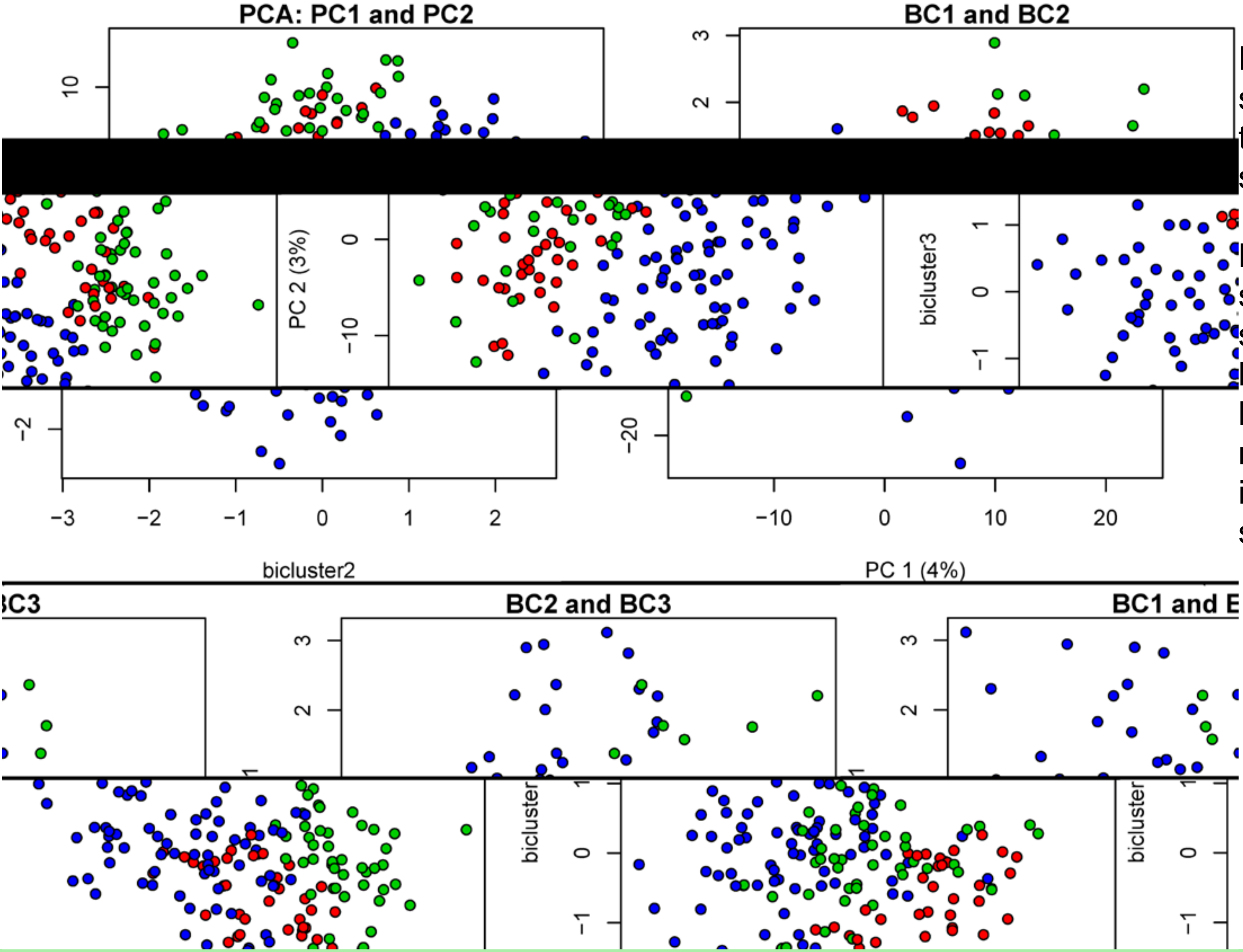


The biplot for bicluster 2 and 3.

The subclasses are quite good separated.

Biclustering: Examples

FABIA applied to the DLBCL data:

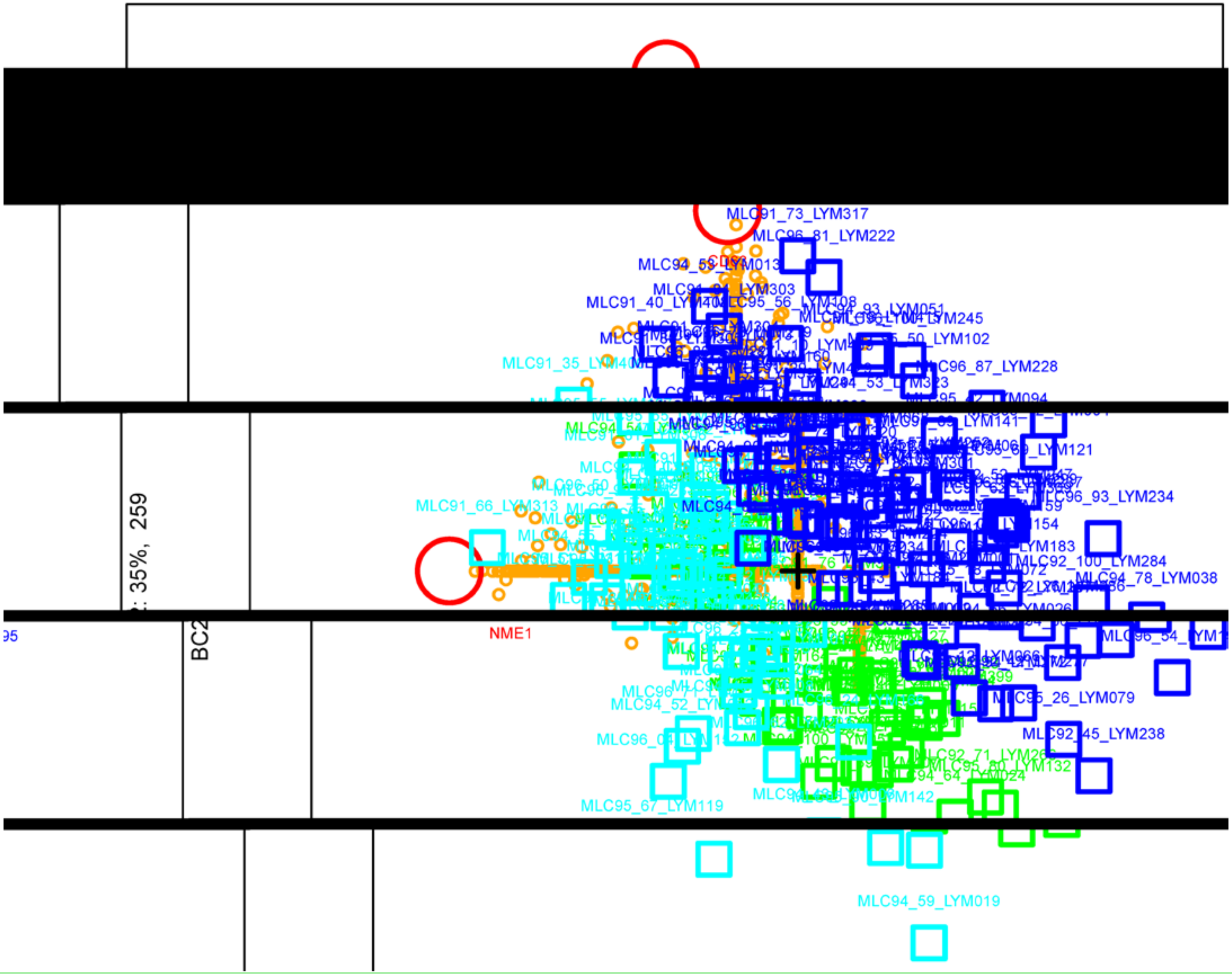


[redacted]

PCA separates the blue subclass.

Biclustering separates subclasses better because red cluster is better separated.

Biclustering: Examples


[redacted]

FABIA applied to DLBCL data.

The biplot for bicluster 1 and 2.

The subclasses are quite good separated.

Chapter 9

Hidden Markov Models

Hidden Markov Models

first chapter devoted to unsupervised learning

→ generative models

most prominent generative model in bioinformatics
hidden Markov model

well suited for analyzing protein or DNA sequences: discrete

Hidden Markov Models in Bioinformatics

generative approach: generating output sequences

New sequences: similar to the model-building sequences

Similar: → certain patterns in a sequences

Hidden Markov Models in Bioinformatics

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

HMMs for gene prediction

exons and introns are identified e.g. GENSCAN

base-pair specificity: 50% - 80%

Profile HMMs

store a multiple alignment in a hidden Markov model

new sequences: evaluated by likelihood

HMMs are build from unaligned sequences: local maxima

avoid: → deterministic annealing (HMMER)

software packages: HMMER
SAM

- cannot discover long-range dependencies
- cannot deal with real-valued inputs
- cannot detect higher order correlations
- cannot use a negative set

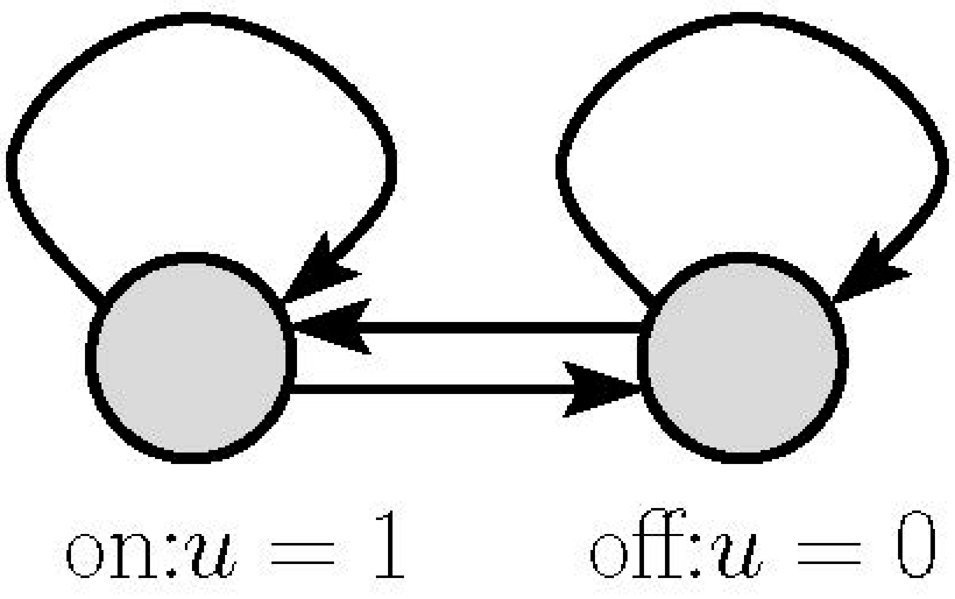
Other HMMs Applications

- remote homology detection
- scoring
- combined with trees
- data base is build on HMMs: PFAM (protein family database) \

Hidden Markov Model Basics

graph of connected hidden states $u \in \{1, \ldots, S\}$

each state produces a probabilistic output

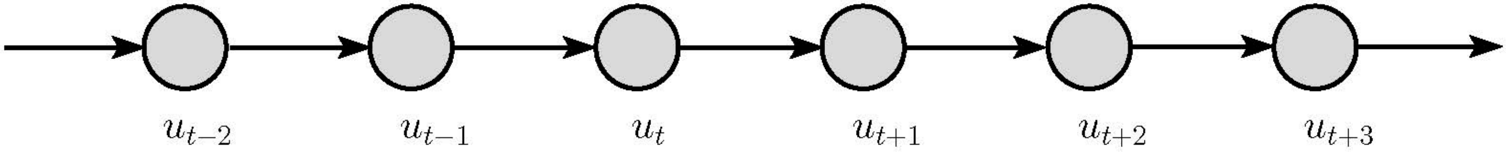


Hidden Markov Model Basics

model evolves over time t (sequence position)

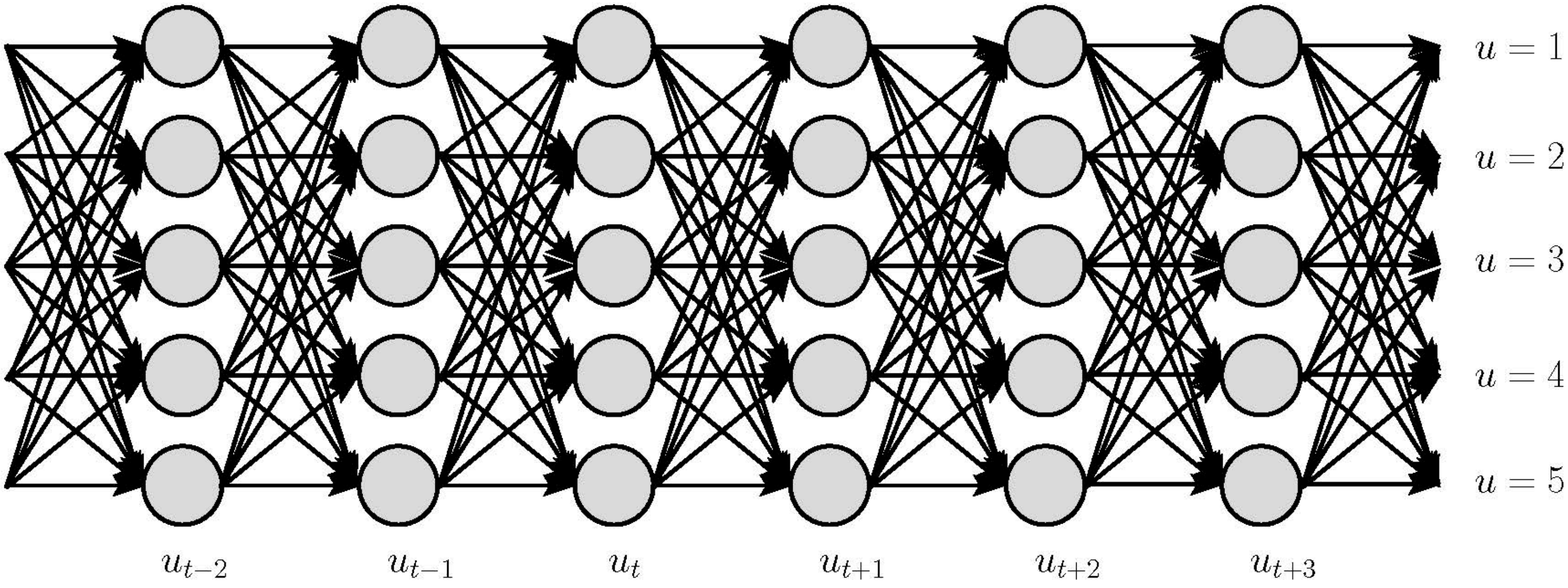
At each step it jumps into another state or remains in current

At t: $u_t \in \{1, \ldots, S\}$



Hidden Markov Model Basics

all possible sequences of values of the hidden state:



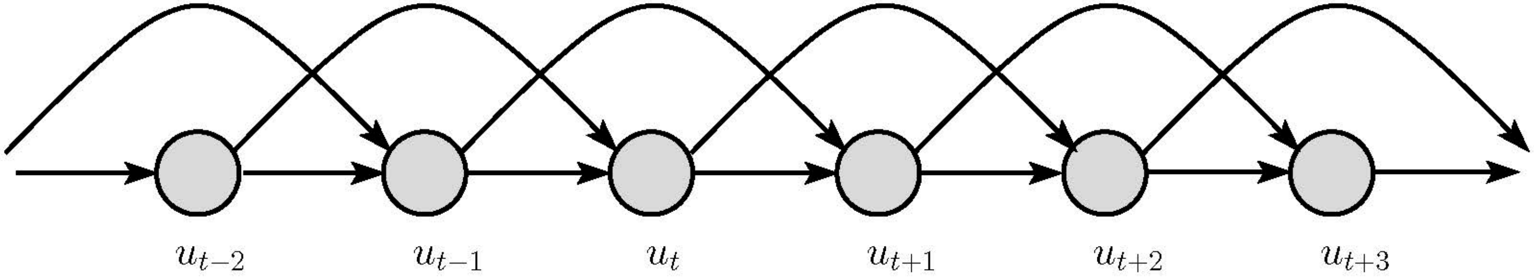
Each path from left to right is a possible sequence

Hidden Markov Model Basics: Transition

transition probabilities: $p(a \mid b)$ $a, b \in \{1, \ldots, S\}$
b is the current state, a next state

Markov assumption: next state only depends on current state

Higher order hidden Markov models: eg second order $p(a \mid b, c)$



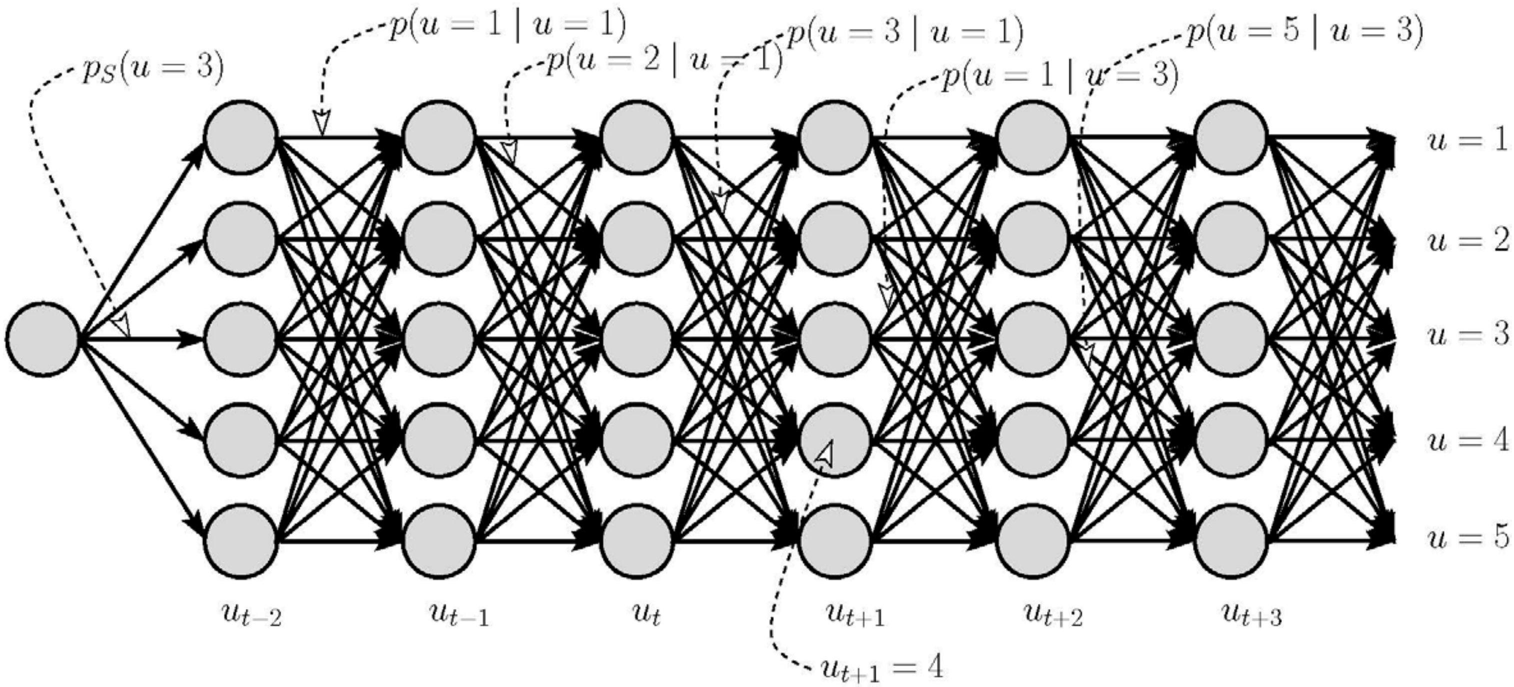
Hidden Markov Model Basics: Transition

transition probability: $p(u_t \mid u_{t-1})$

start state probability: $p_S(u_1)$

probability of observing sequence $u^T = (u_1, u_2, u_3, \ldots, u_T)$ length T

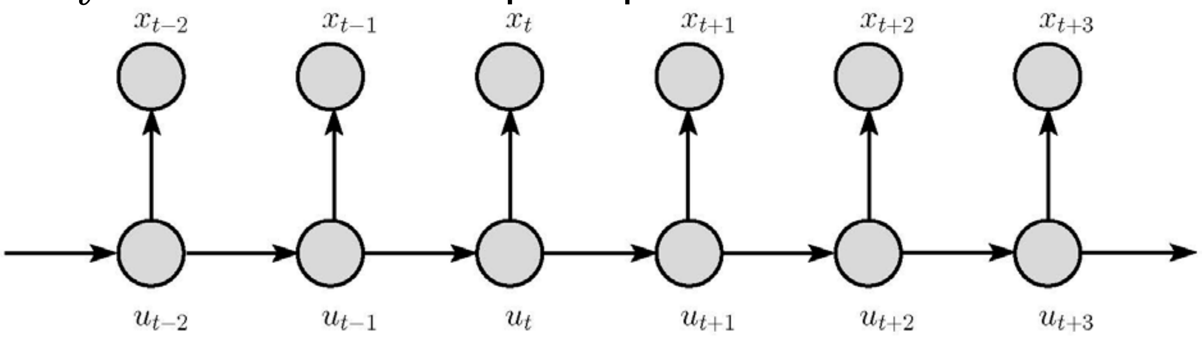
$$p(u^T) \ = \ p_S(u_1) \prod_{t=2}^{T} p(u_t \mid u_{t-1})$$



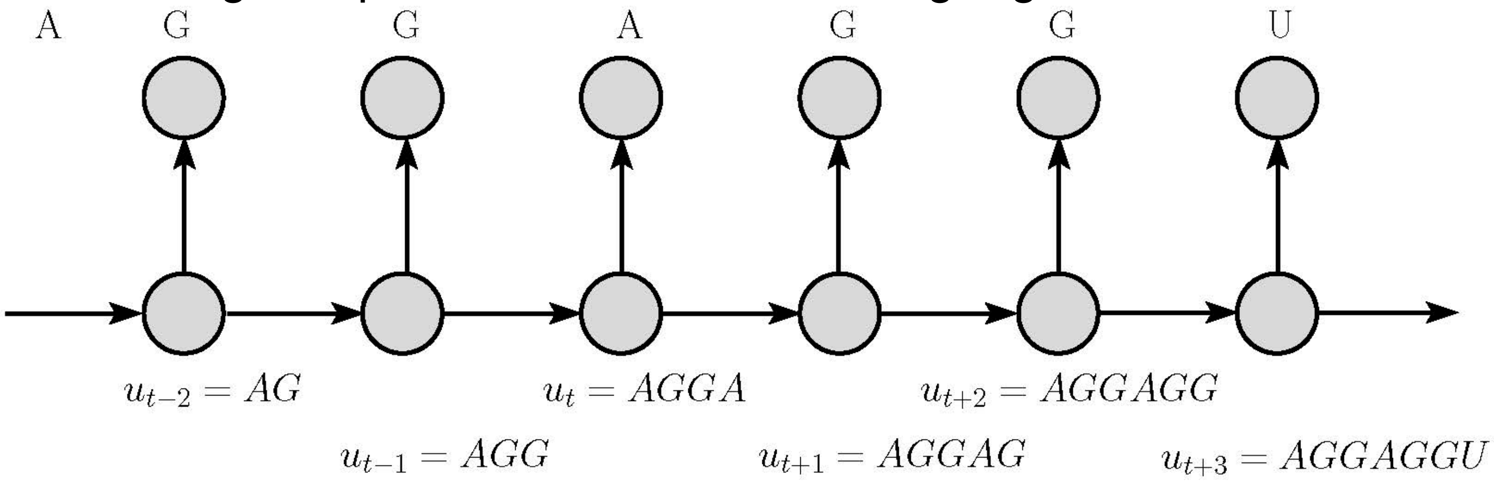
Hidden Markov Model Basics: Emission

emission probability $p_E(x_t \mid u_t)$

x_t: element of the output alphabet of size $\boldsymbol{P}$ $x_t \in \{A, T, C, G\}$



Shine-Dalgarno pattern for ribosome binding regions



Hidden Markov Model Basics: Probability

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

joint probability of the output sequence $x^T = (x_1, x_2, x_3, \ldots, x_T)$ of length $\boldsymbol{T}$ and the hidden state sequence $u^T = (u_1, u_2, u_3, \ldots, u_T)$

$$p(u^T, x^T) \ = \ p_S(u_1) \ \prod_{t=2}^{T} p(u_t \mid u_{t-1}) \prod_{t=1}^{T} p_E(x_t \mid u_t)$$

probability of the output sequence → marginalization:

$$p(x^T) \ = \ \sum_{u^T} p(u^T, x^T) \ = \ \sum_{u^T} p_S(u_1) \ \prod_{t=2}^{T} p(u_t \mid u_{t-1}) \prod_{t=1}^{T} p_E(x_t \mid u_t)$$

$\sum_{u^T}$ sum over all possible sequences of length $\boldsymbol{T}$ of values $\{1, \ldots, S\}$

(first order) Markov assumption → recursively compute this sum

Hidden Markov Model Basics: Forward Pass

$x^t = (x_1, x_2, x_3, \ldots, x_t)$ prefix sequence of x^T

$p(x^t, u_t)$ probability of the model producing x^t and being in u_t at end

$$\begin{aligned} p(x^t, u_t) \ &= \ p(x_t \mid x^{t-1}, u_t) \ p(x^{t-1}, u_t) \ = \\ p_E(x_t \mid u_t) \ &\sum_{u_{t-1}} p(x^{t-1}, u_t, u_{t-1}) \ = \\ p_E(x_t \mid u_t) \ &\sum_{u_{t-1}} p(u_t \mid x^{t-1}, u_{t-1}) \ p(x^{t-1}, u_{t-1}) \ = \\ p_E(x_t \mid u_t) \ &\sum_{u_{t-1}} p(u_t \mid u_{t-1}) \ p(x^{t-1}, u_{t-1}) \end{aligned}$$

Markov assumptions $\begin{aligned} p(x_t \mid x^{t-1}, u_t) \ &= \ p_E(x_t \mid u_t) \\ p(u_t \mid x^{t-1}, u_{t-1}) \ &= \ p(u_t \mid u_{t-1}) \end{aligned}$

marginalization $p(x^{t-1}, u_t) \ = \ \sum_{u_{t-1}} p(x^{t-1}, u_t, u_{t-1})$

Cond. prob. $\quad p(x^{t-1}, u_t, u_{t-1}) \ = \ p(u_t \mid x^{t-1}, u_{t-1}) \ p(x^{t-1}, u_{t-1})$

Hidden Markov Model Basics: Forward Pass

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

each recursion step: sum over all S u_{t-1}

for each value u_t

for each time step t

Complexity: $O(T\ S^2)$

recursion starts with $p(x^1, u_1) \ = \ p_S(u_1)\ p_E(x_1 \mid u_1)$

and the probability is $p(x^T) \ = \ \sum_{u_T} p(x^T, u_T)$

“forward pass” or “forward phase” to compute the probability of x^T

Hidden Markov Model Basics: Forward Pass

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

HMM Forward Pass

Given: sequence $x^T = (x_1, x_2, x_3, \ldots, x_T)$, state values $u \in \{1, \ldots, S\}$, start probabilities $p_S(u_1)$, transition probabilities $p(u_t \mid u_{t-1})$, and emission probabilities $p_E(x_t \mid u_t)$; Output: likelihood $p(x^T)$ and $p(x^t, u_t)$

```
BEGIN initialization

        $p(x^1, u_1) \ = \ p_S(u_1) \, p_E(x_1 \mid u_1)$

END initialization
BEGIN Recursion
  for $(t = 2 \ ; \ t \ \leq \ T \ ; \ t++)$ do
    for $(a = 1 \ ; \ a \ \leq \ S \ ; \ a++)$ do

      $p(x^t, u_t = a) \ = \ p_E(x_t \mid u_t = a) \sum_{u_{t-1}=1}^{S} p(u_t = a \mid u_{t-1}) \, p(x^{t-1}, u_{t-1})$

    end for
  end for
END Recursion
BEGIN Compute Likelihood

        $p(x^T) \ = \ \sum_{a=1}^{S} p(x^T, u_T = a)$

END Compute Likelihood
```

Expectation Maximization for HMM: Baum-Welch Algorithm

Learning / parameter selection

Parameters: $\boldsymbol{S}$ start probabilities $p_S(u_1)$
$\boldsymbol{S}^2$ transitions probabilities $p(u_t \mid u_{t-1})$
$\boldsymbol{SP}$ emission probabilities $p_E(x_t \mid u_t)$

training sequences: $\{\boldsymbol{x}^i\}, 1 \leq i \leq l$ -- not subsequence!!

parameter optimization: maximize likelihood

Expectation Maximization algorithm

EM algorithm defines a lower bound on the likelihood:

$$\mathcal{F}(Q, \boldsymbol{w}) \ = \ \int_U Q(\boldsymbol{u} \mid \boldsymbol{x}) \ \ln p(\boldsymbol{x}, \boldsymbol{u}; \boldsymbol{w}) \ d\boldsymbol{u} \ - \int_U Q(\boldsymbol{u} \mid \boldsymbol{x}) \ \ln Q(\boldsymbol{u} \mid \boldsymbol{x}) \ d\boldsymbol{u}$$

Expectation Maximization for HMM: Baum-Welch Algorithm

$Q(\boldsymbol{u} \mid \boldsymbol{x})$: estimation for $p(\boldsymbol{u} \mid \boldsymbol{x}; \boldsymbol{w})$

discrete HMMs: $\boldsymbol{u}$ is the sequence of hidden states
$\boldsymbol{x}$ the sequence of output states
$\boldsymbol{w}$ all probabilities (parameters) in the model

discrete: integral → sum

estimation for the state sequence:

$$Q(\boldsymbol{u} \mid \boldsymbol{x}) \;=\; p(u_1 = a_1, u_2 = a_2, \ldots, u_T = a_T \mid x^T; \boldsymbol{w}^{\mathrm{old}})$$

Expectation Maximization for HMM: Baum-Welch Algorithm

$$\begin{aligned}
\mathcal{F}(Q, \boldsymbol{w}) \ &= \ \sum_{a_1=1}^{S} \ldots \sum_{a_T=1}^{S} \\
& p(u_1 = a_1, u_2 = a_2, \ldots, u_T = a_T \mid x^T; \boldsymbol{w}^{\mathrm{old}}) \ \ln p(x^T, u^T; \boldsymbol{w}) \ - \\
& \sum_{a_1=1}^{S} \ldots \sum_{a_T=1}^{S} p(u_1 = a_1, u_2 = a_2, \ldots, u_T = a_T \mid x^T; \boldsymbol{w}^{\mathrm{old}}) \\
& \ln p(u_1 = a_1, u_2 = a_2, \ldots, u_T = a_T \mid x^T; \boldsymbol{w}^{\mathrm{old}}) \ = \\
& \sum_{a_1=1}^{S} \ldots \sum_{a_T=1}^{S} p(u_1 = a_1, u_2 = a_2, \ldots, u_T = a_T \mid x^T; \boldsymbol{w}^{\mathrm{old}}) \\
& \ln p(x^T, u^T; \boldsymbol{w}) - c
\end{aligned}$$

c independent of $\boldsymbol{w}$

Expectation Maximization for HMM: Baum-Welch Algorithm

$$\ln p(u^T, x^T; \boldsymbol{w}) =$$

$$\ln p_S(u_1) + \sum_{t=2}^{T} \ln p(u_t \mid u_{t-1}) + \sum_{t=1}^{T} \ln p_E(x_t \mid u_t)$$

a_t summed out

$$\mathcal{F}(Q, \boldsymbol{w}) = \sum_{a=1}^{S} p(u_1 = a \mid x^T; \boldsymbol{w}^{\text{old}}) \ \ln p_S(u_1 = a) +$$

$$\sum_{t=1}^{T} \sum_{a=1}^{S} p(u_t = a \mid x^T; \boldsymbol{w}^{\text{old}}) \ \ln p_E(x_t \mid u_t = a) +$$

$$\sum_{t=2}^{T} \sum_{a=1}^{S} \sum_{b=1}^{S} p(u_t = a, u_{t-1} = b \mid x^T; \boldsymbol{w}^{\text{old}}) \ \ln p(u_t = a \mid u_{t-1} = b) - c$$

Expectation Maximization for HMM: Baum-Welch Algorithm

parameters $\boldsymbol{w}$:
start probabilities $p_S(a)$
emission probabilities $p_E(x \mid a)$
transition probabilities $p(a \mid b)$

Constraints:

$$\sum_a p_S(a) \ = \ 1$$

$$\sum_x p_E(x \mid a) \ = \ 1$$

$$\sum_a p(a \mid b) \ = \ 1$$

$$\text{Now:} \quad \boldsymbol{w} = \boldsymbol{w}^{\text{old}}$$

Expectation Maximization for HMM: Baum-Welch Algorithm

M-step optimization problem

$$\min_{\boldsymbol{w}} \quad \sum_t \sum_a c_{ta} \ \ln w_a$$

$$\text{s.t.} \quad \sum_a w_a \ = \ 1$$

Lagrangian

$$L \ = \ \sum_t \sum_a c_{ta} \ \ln w_a \ - \ \lambda \left(\sum_a w_a \ - \ 1 \right)$$

$$\frac{\partial L}{\partial w_a} \ = \ \sum_t c_{ta} \ \frac{1}{w_a} \ - \ \lambda \ = \ 0 \ \Rightarrow \ \sum_t c_{ta} \ - \ \lambda \ w_a \ = \ 0$$

summing over a gives

$$\sum_a \sum_t c_{ta} \ = \ \lambda$$

Expectation Maximization for HMM: Baum-Welch Algorithm

We obtain $w_a \; = \; \frac{\sum_t c_{ta}}{\sum_a \sum_t c_{ta}}$

(constraint) maximization step (M-step) for the different probabilities:

$$p_S(a) \; = \; \frac{p(u_1 = a \mid x^T ; \boldsymbol{w})}{\sum_{a'} p(u_1 = a' \mid x^T ; \boldsymbol{w})}$$

$$p_E(x \mid a) \; = \; \frac{\sum_{t=1}^{T} \delta_{x_t = x} \; p(u_t = a \mid x^T ; \boldsymbol{w})}{\sum_y \sum_{t=1}^{T} \delta_{x_t = y} \; p(u_t = a \mid x^T ; \boldsymbol{w})}$$

$$p(a \mid b) \; = \; \frac{\sum_{t=2}^{T} p(u_t = a , u_{t-1} = b \mid x^T ; \boldsymbol{w})}{\sum_{a'} \sum_{t=2}^{T} p(u_t = a' , u_{t-1} = b \mid x^T ; \boldsymbol{w})}$$

Expectation Maximization for HMM: Baum-Welch Algorithm

$$p_S(a) \;=\; p(u_1 = a \mid x^T ; \boldsymbol{w})$$

$$p_E(x \mid a) \;=\; \frac{\sum_{t=1}^{T} \delta_{x_t = x} \; p(u_t = a \mid x^T ; \boldsymbol{w})}{\sum_{t=1}^{T} p(u_t = a \mid x^T ; \boldsymbol{w})}$$

$$p(a \mid b) \;=\; \frac{\sum_{t=2}^{T} p(u_t = a, u_{t-1} = b \mid x^T ; \boldsymbol{w})}{\sum_{t=2}^{T} p(u_{t-1} = b \mid x^T ; \boldsymbol{w})}$$

estimation step (E-step):

Estimate $p(u_t = a \mid x^T ; \boldsymbol{w})$ and $p(u_t = a, u_{t-1} = b \mid x^T ; \boldsymbol{w})$

suffix sequence $x^{t \leftarrow T} = (x_t, x_{t+1}, \ldots, x_T)$ of length $T - t + 1$

probability $p(x^{t+1 \leftarrow T} \mid u_t = a)$ of suffix sequence

$x^{t+1 \leftarrow T} = (x_{t+1}, \ldots, x_T)$ if starting from $u_t = a$

Expectation Maximization for HMM: Baum-Welch Algorithm

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

$$p(u_t = a \mid x^T; \boldsymbol{w}) \;=\; \frac{p(u_t = a, x^T; \boldsymbol{w})}{p(x^T)} \;=$$

$$\frac{p(x^t, u_t = a; \boldsymbol{w})\; p(x^{t+1 \leftarrow T} \mid u_t = a)}{p(x^T)}$$

$$p(u_t = a, u_{t-1} = b \mid x^T; \boldsymbol{w}) \;=\; \frac{p(u_t = a, u_{t-1} = b, x^T; \boldsymbol{w})}{p(x^T)} \;=$$

$$p(x^{t-1}, u_{t-1} = b; \boldsymbol{w})\; p(u_t = a \mid u_{t-1} = b)\; p_E(x_t \mid u_t = a)$$
$$p(x^{t+1 \leftarrow T} \mid u_t = a) \;/\; p(x^T)$$

$$p(u_t = a \mid x^T; \boldsymbol{w}) \;=\; \sum_{b=1}^{S} p(u_t = a, u_{t-1} = b \mid x^T; \boldsymbol{w})$$

Expectation Maximization for HMM: Baum-Welch Algorithm

backward recursion for computing $p(x^{t+1 \leftarrow T} \mid u_t = a)$

$$p(x^{t+1 \leftarrow T} \mid u_t = a) \ = \ \sum_{b=1}^{S} p_E(x_{t+1} \mid u_{t+1} = b) \ p(u_{t+1} = b \mid u_t = a) \ p(x^{t+2 \leftarrow T} \mid u_{t+1} = b)$$

starting conditions

$$p(x^{T \leftarrow T} \mid u_{T-1} = a) \ = \ \sum_{b=1}^{S} p_E(x_T \mid u_T = b) \ p(u_T = b \mid u_{T-1} = a)$$

or, alternatively

$$\forall_a : \quad p(x^{T+1 \leftarrow T} \mid u_T = a) \ = \ 1$$

Expectation Maximization for HMM: Baum-Welch Algorithm

HMM Backward Pass

Given: sequence $x^T = (x_1, x_2, x_3, \ldots, x_T)$, state values $u \in \{1, \ldots, S\}$, start probabilities $p_S(u_1)$, transition probabilities $p(u_t \mid u_{t-1})$, and emission probabilities $p_E(x_t \mid u_t)$; Output: likelihood $p(x^T)$ and $p(x^{t+1 \leftarrow T} \mid u_t = a)$

```
BEGIN initialization
    $\forall_a : \; p(x^{T+1 \leftarrow T} \mid u_T = a) \; = \; 1$
END initialization
BEGIN Recursion
  for $(t = T-1 \; ; \; t \; \geq \; 1 \; ; \; t--)$ do
    for $(a = 1 \; ; \; a \; \leq \; S \; ; \; a++)$ do
        $p(x^{t+1 \leftarrow T} \mid u_t = a) \; = \sum_{b=1}^{S} p_E(x_{t+1} \mid u_{t+1} = b) \; p(u_{t+1} = b \mid u_t = a) \; p(x^{t+2 \leftarrow T} \mid u_{t+1} = b) \; .$
    end for
  end for
END Recursion
BEGIN Compute Likelihood
    $p(x^T) \; = \; \sum_{a=1}^{S} p_S(u_1 = a) \; (p(x^{1 \leftarrow T} \mid u_1 = a)$
END Compute Likelihood
```

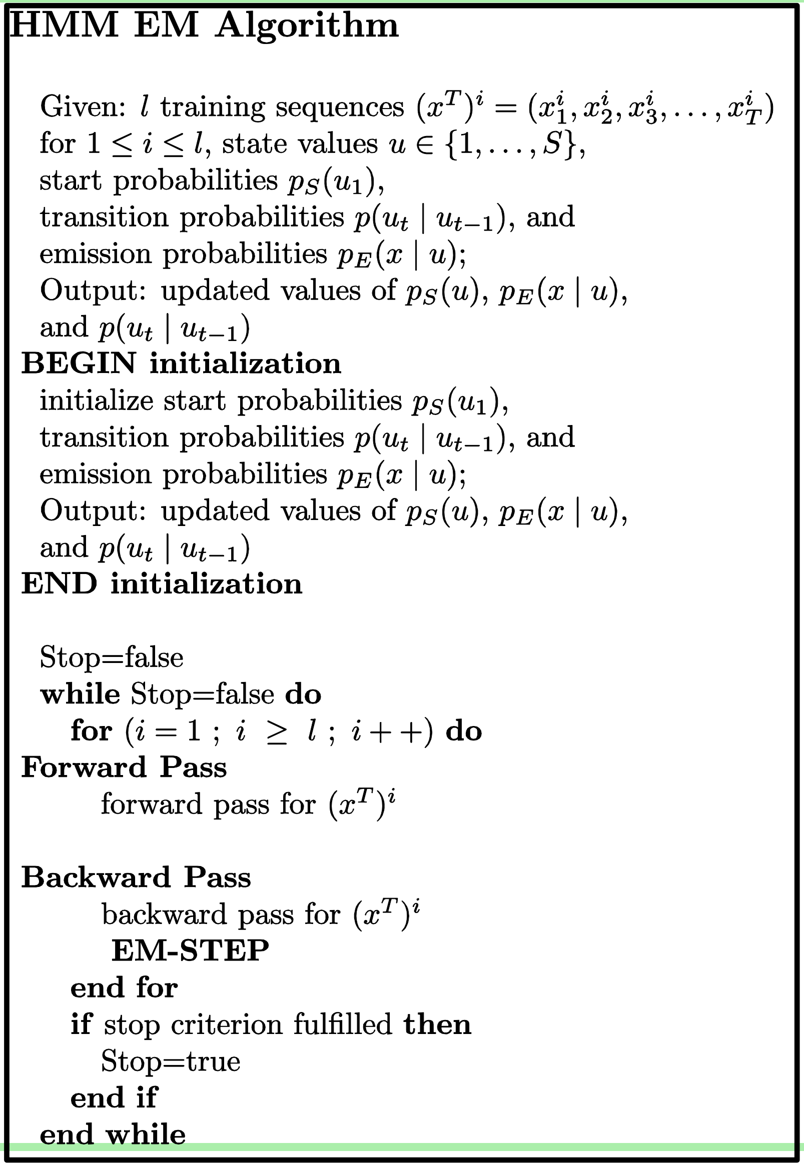
Expectation Maximization for HMM: Baum-Welch Algorithm

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine



HMM EM Algorithm

```
  Given: $l$ training sequences $(x^T)^i = (x_1^i, x_2^i, x_3^i, \ldots, x_T^i)$
  for $1 \leq i \leq l$, state values $u \in \{1, \ldots, S\}$,
  start probabilities $p_S(u_1)$,
  transition probabilities $p(u_t \mid u_{t-1})$, and
  emission probabilities $p_E(x \mid u)$;
  Output: updated values of $p_S(u)$, $p_E(x \mid u)$,
  and $p(u_t \mid u_{t-1})$
BEGIN initialization
  initialize start probabilities $p_S(u_1)$,
  transition probabilities $p(u_t \mid u_{t-1})$, and
  emission probabilities $p_E(x \mid u)$;
  Output: updated values of $p_S(u)$, $p_E(x \mid u)$,
  and $p(u_t \mid u_{t-1})$
END initialization

  Stop=false
  while Stop=false do
    for $(i = 1 \; ; \; i \; \geq \; l \; ; \; i++)$ do
Forward Pass
        forward pass for $(x^T)^i$

Backward Pass
        backward pass for $(x^T)^i$
         EM-STEP
    end for
    if stop criterion fulfilled then
        Stop=true
    end if
  end while
```

Expectation Maximization for HMM: Baum-Welch Algorithm

HMM EM Algorithm

E-Step

```
for $(a = 1 \; ; \; a \; \leq \; S \; ; \; a++)$ do
  for $(b = 1 \; ; \; b \; \leq \; S \; ; \; b++)$ do
      $p(u_t = a, u_{t-1} = b \mid (x^T)^i; \boldsymbol{w}) \; = $
      $p((x^{t-1})^i, u_{t-1} = b; \boldsymbol{w}) \; p(u_t = a \mid u_{t-1} = b)$
      $p_E(x_t^i \mid u_t = a)$
      $p((x^{t+1 \leftarrow T})^i \mid u_t = a) \; / \; p((x^T)^i)$
  end for
end for
for $(a = 1 \; ; \; a \; \leq \; S \; ; \; a++)$ do
    $p(u_t = a \mid (x^T)^i; \boldsymbol{w}) \; = \; \sum_{b=1}^{S} p(u_t = a, u_{t-1} = b \mid (x^T)^i; \boldsymbol{w})$
end for
```

HMM EM Algorithm

M-Step

```
for $(a = 1 \; ; \; a \; \leq \; S \; ; \; a++)$ do
    $p_S(a) \; = \; p(u_1 = a \mid (x^T)^i; \boldsymbol{w})$
end for
for $(a = 1 \; ; \; a \; \leq \; S \; ; \; a++)$ do
  for $(x = 1 \; ; \; x \; \leq \; P \; ; \; x++)$ do
      $p_E(x \mid a) \; = \; \frac{\sum_{t=1}^{T} \delta_{x_t^i = x} \; p(u_t = a \mid (x^T)^i; \boldsymbol{w})}{\sum_{t=1}^{T} p(u_t = a \mid (x^T)^i; \boldsymbol{w})}$
  end for
end for
for $(a = 1 \; ; \; a \; \leq \; S \; ; \; a++)$ do
  for $(b = 1 \; ; \; b \; \leq \; S \; ; \; b++)$ do
      $p(a \mid b) \; = \; \frac{\sum_{t=2}^{T} p(u_t = a, u_{t-1} = b \mid (x^T)^i; \boldsymbol{w})}{\sum_{t=2}^{T} p(u_{t-1} = b \mid (x^T)^i; \boldsymbol{w})}$
  end for
end for
```

Viterby Algorithm

likelihood of $x^T$is an integral - more exactly a sum - over all probabilities of possible sequences of hidden states multiplied by the probability that the hidden sequence emits x^T

Often a specific hidden sequence dominates sum:

$$(u^T)^* \;=\; \arg\max_{u^T} p(u^T \mid x^T) \;=\; \arg\max_{u^T} p(u^T, x^T)$$

semantic meaning: hidden states are interpretable what is $(u^T)^*$

Bioinformatics: $(u^T)^*$ needed for alignment sequence ←→ HMM

Alignment → obtained through dynamic programming
path in from left to right in figure

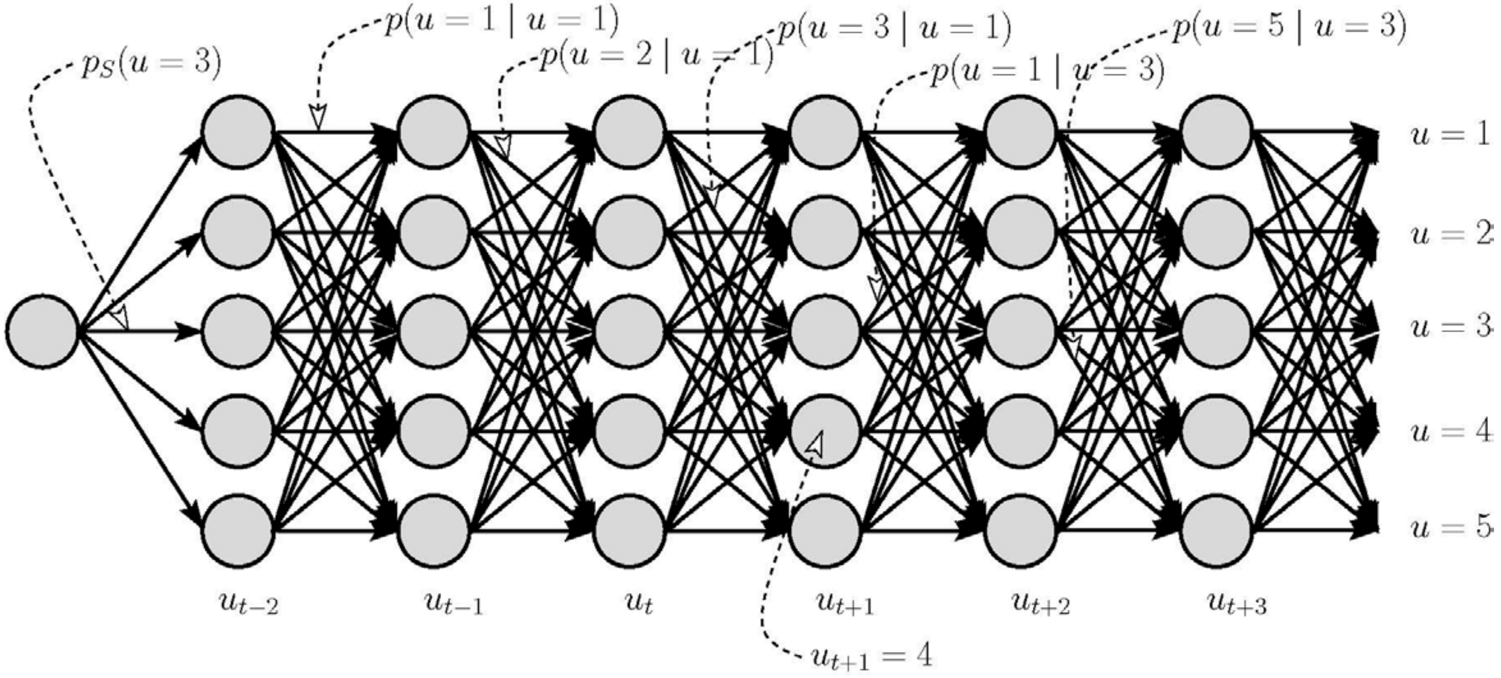
Viterby Algorithm

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine



circles: represented by a matrix $\boldsymbol{V}$

$V_{t,a}$: maximal probability of a sequence of length t ending in state a

$$V_{t,a} = \max_{u^{t-1}} p(x^t, u^{t-1}, u_t = a)$$

Viterby Algorithm

Markov conditions → recursive formulae

$$V_{t,a} \; = \; p_E(x_t \mid u_t = a) \; \max_b p(u_t = a \mid u_{t-1} = b) \; V_{t-1,b}$$

Initialization

$$V_{1,a} \; = \; p_S(a) p_E(x_1 \mid u_1 = a)$$

Result

$$\max_{u^T} p(u^T, x^T) \; = \; \max_a V_{T,a} \; .$$

sequence by backtracing

$$b(t,a) \; = \; \arg\max_b p(u_t = a \mid u_{t-1} = b) \; V_{t-1,b}$$

complexity of Viterby algorithm: $O(T \; S^2)$

(maximum over S terms)

Viterby Algorithm

HMM Viterby

Given: sequence $x^T = (x_1, x_2, x_3, \ldots, x_T)$, state values $u \in \{1, \ldots, S\}$, start probabilities $p_S(u_1)$, transition probabilities $p(u_t \mid u_{t-1})$, and emission probabilities $p_E(x_t \mid u_t)$; Output: most likely sequence of hidden state values $(u^T)^*$ and its probability $p\left(x^T, (u^T)^*\right)$

```
BEGIN initialization
    $V_{1,a} \ = \ p_S(a) p_E(x_1 \mid u_1 = a)$
END initialization
BEGIN Recursion
  for $(t = 2 \ ; \ t \ \leq \ T \ ; \ t++)$ do
    for $(a = 1 \ ; \ a \ \leq \ S \ ; \ a++)$ do
        $V_{t,a} \ = \ p_E(x_t \mid u_t = a) \ \max_b p(u_t = a \mid u_{t-1} = b) \ V_{t-1,b}$
        $b(t,a) \ = \ \arg\max_b p(u_t = a \mid u_{t-1} = b) \ V_{t-1,b}$
    end for
  end for
END Recursion
BEGIN Compute Probability
    $p\left(x^T, (u^T)^*\right) \ = \ \max_{a=1}^{S} V(T,a)$
END Compute Probability
```

Viterby Algorithm

HMM Viterby-Backtracing

```
BEGIN Backtracing
        $s \; = \; \arg\max_{a=1}^{S} V(T, a)$
        print $s$
  for $(t = T \; ; \; t \; \geq \; 2 \; ; \; t--)$ do
        $s \; = \; b(t, s)$
        print $s$
  end for
END Backtracing
```

Viterby Algorithm

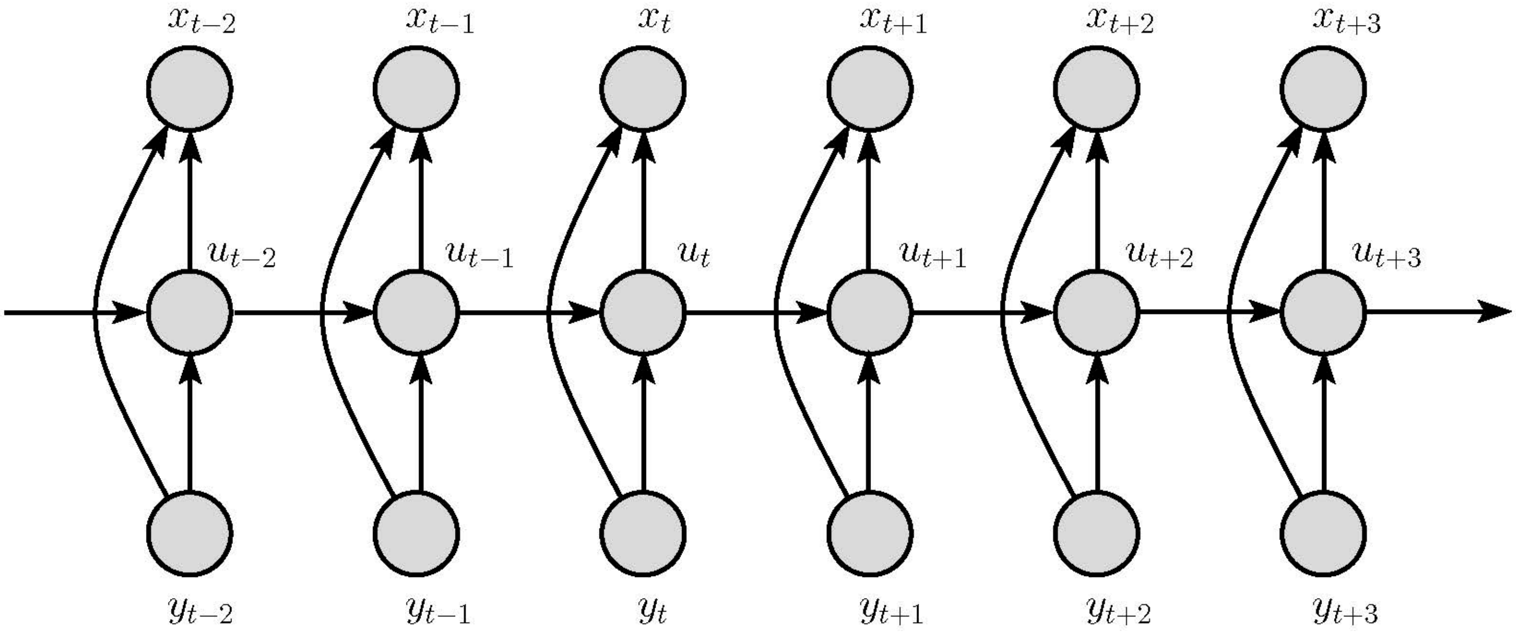
Viterby algorithm to improve a multiple alignment:

1. initialize HMM
2. align all sequences to the HMM via the Viterby algorithm
3. make frequency counts per column and compute the transition probabilities to update the HMM
4. if not converged go to step 2

Input Output Hidden Markov Models

Input Output Hidden Markov Models (IOHMMs)

Output sequence $x^T = (x_1, x_2, x_3, \ldots, x_T)$ conditioned on input sequence $y^T = (y_1, y_2, y_3, \ldots, y_T)$



Input Output Hidden Markov Models

probabilities are conditioned on the input:

$$p_S(u_1 \mid y_1) \qquad p(u_t \mid y_t, u_{t-1}) \qquad p_E(x_t \mid y_t, u_t)$$

IOMM: classification including negative examples

$$y_T = 1 \qquad y_T = -1$$

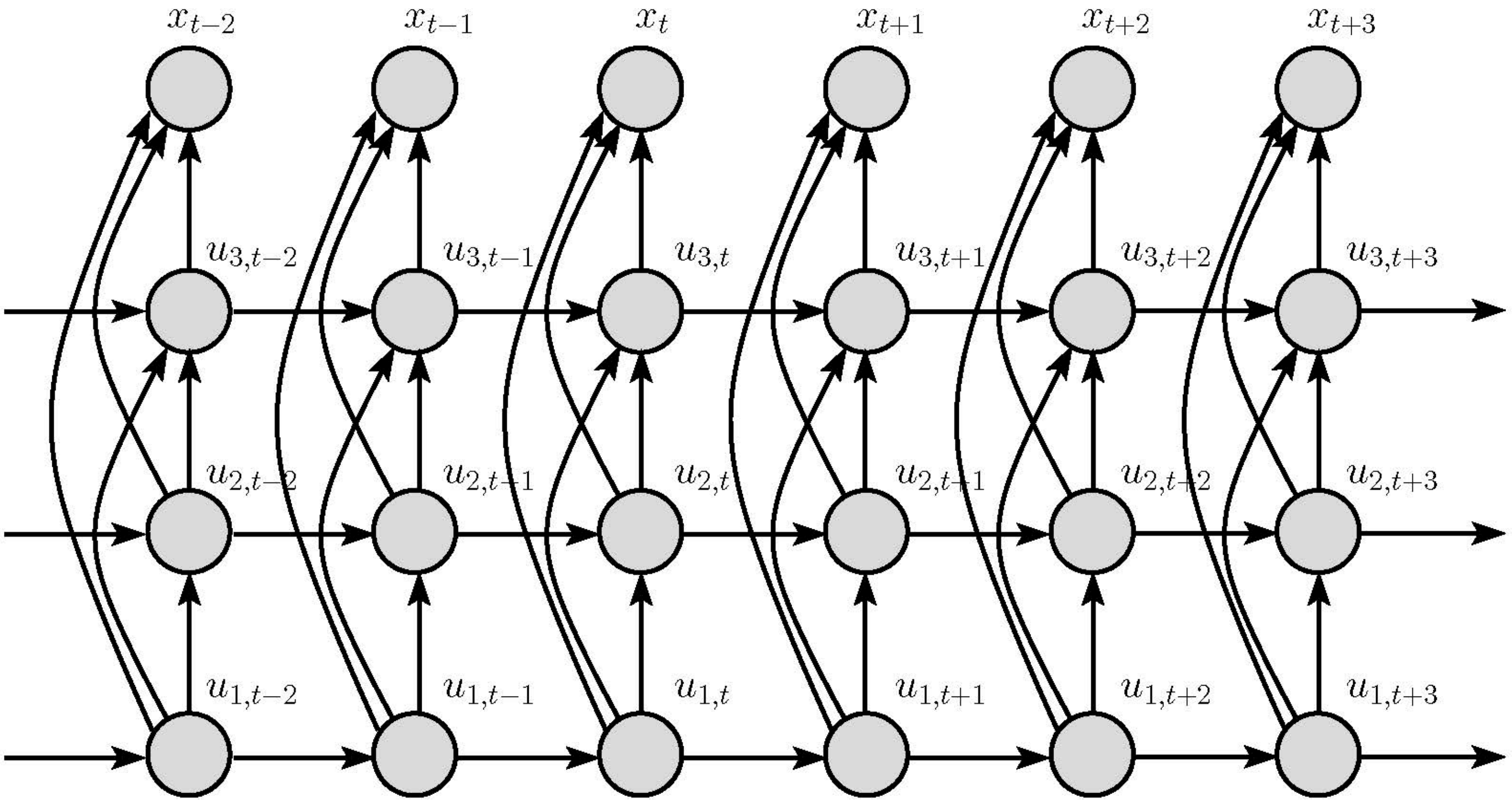
Subclassses of the negative class which is very similar to the positive class can be better discriminated

parameters increase proportional to input symbols

Learning analog to original method

Factorial Hidden Markov Models

hidden state is divided into more components



Factorial Hidden Markov Models

The transition probability of hidden state variable u_i is conditioned on all u_k with $k \leq i$

emission probability depends on all hidden states

Idea: u_1 evolves very slowly, u_2 evolves faster, etc
state variable depends on all slower ones

factorial HMMs is computational expensive to learn

approximative methods have been developed to speed up learning

Memory Input Output Factorial Hidden Markov Models

Sean Eddy
"HMMs are reasonable models of linear sequence problems, but they don't deal well with correlations between residues or states, especially long-range correlations."

"The state path of an HMM has no way of remembering
what a distant state generated when a second state generates its residue."

only way an HMM can store information:

→go into a certain state value and don't change it

BUT: event which led the HMM enter the fixed state is memorized

Memory Input Output Factorial Hidden Markov Models

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

$$p(u_t = a \mid u_{t-1} = a) \ = \ 1$$

if the state takes on the value a then the state is fixed forever

memory is enforced by fixing $p(u_t = a \mid u_{t-1} = a) \ = \ 1$

However: after the storage process systems dynamics nor other events to memorize can be dealt with

→ factorial HMM

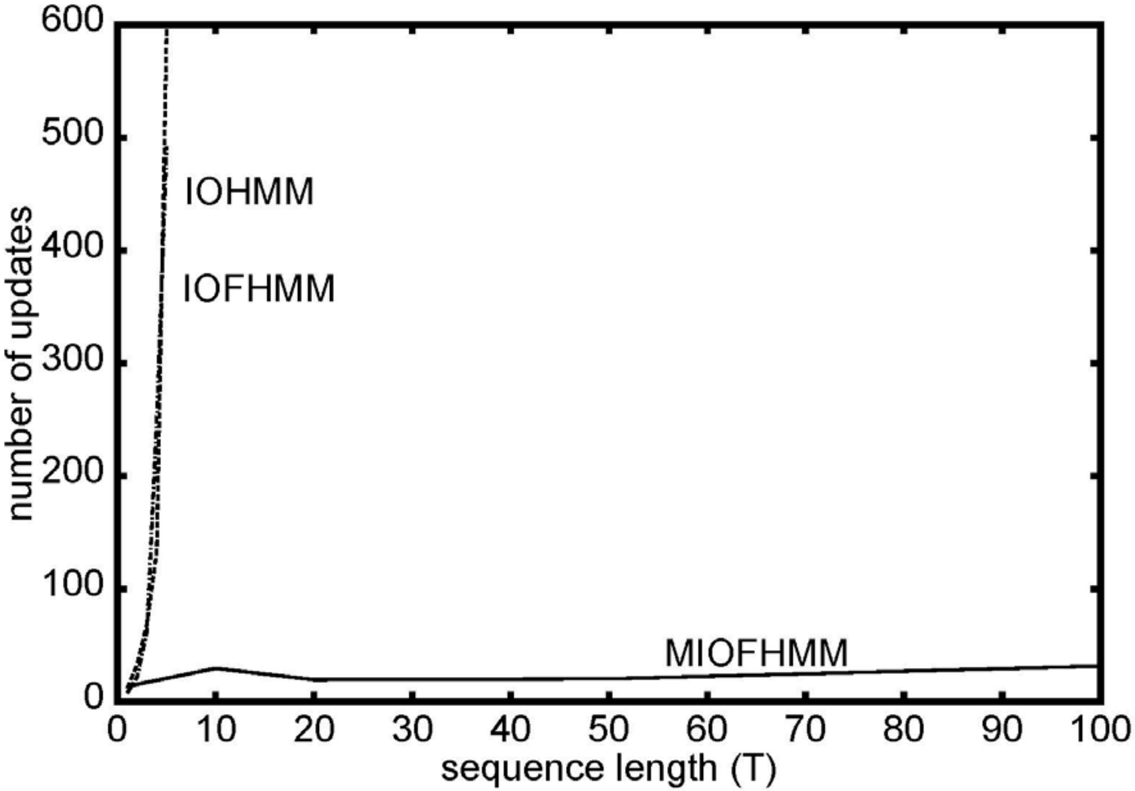
Storing events → input output HMMs

Memory-based Input-Output Factorial HMM

Initially: all state variables have “uncommitted” values
then various inputs can trigger the memory state

Memory Input Output Factorial Hidden Markov Models

number of updates required to train three models:
input output HMM (IOHMM),
input output factorial HMM (IOFHMM),
and “Memory-based Input-Output Factorial HMM” (MIOFHMM)



Tricks of the Trade

1. Bioinformatics: delete states which do not emit symbols

2. variable length of the sequences
   carefully compare likelihoods → length dependent

3. small likelihood → log-space

4. kept probabilities above a threshold ε to allow all solutions

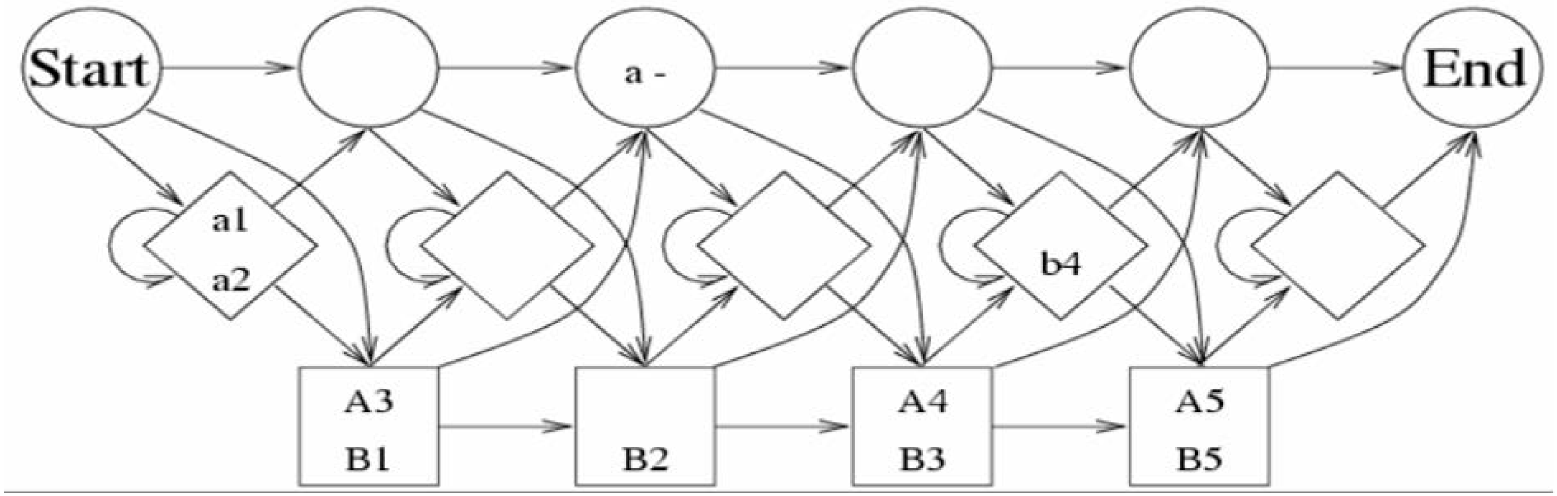
5. EM-algorithm: probabilities > 0 → set to zero after learning
   → helps to generalize

6. prone to local minima → global optimization strategies
   (deterministic annealing)

Profile Hidden Markov Models

Profile Hidden Markov Models code a multiple sequence alignment

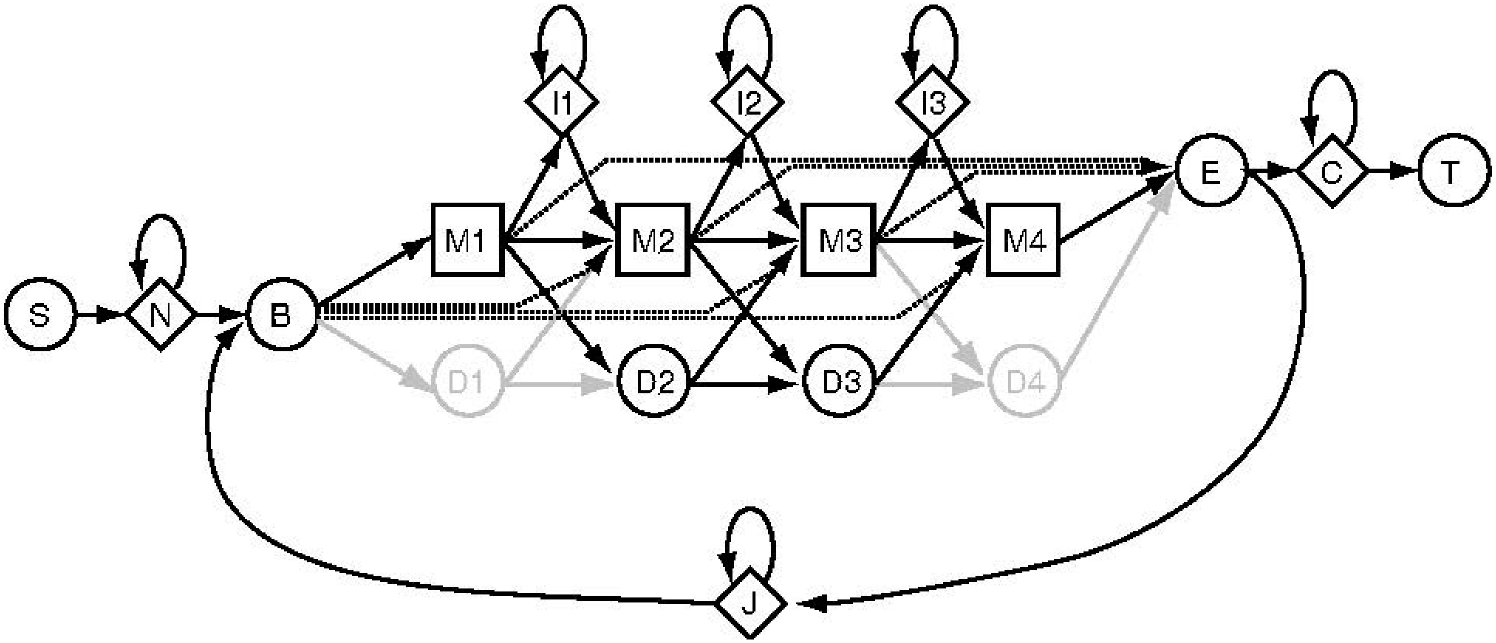
position-specific scoring system → search databases for homologous



1. The top row with states indicated with circles are a pattern.
2. The diamond states are inserted strings.
3. The bottom row with states indicated as squares are deletions, where a letter from the pattern is skipped.

Profile Hidden Markov Models

HMMER package by Sean Eddy



1. squares “Mx”: consensus string
2. circles “Dx”: are deletion states (non-emitting states)
3. diamonds “Ix”: insertion states

Profile Hidden Markov Models

package SAM
Sequence Alignment and Modeling system (SAM -
http://www.cse.ucsc.edu/research/compbio/sam.html
creating, refining, and using HMMs

SAM models: multiple alignment as HMMER

Databases: Protein FAMily database (Pfam)

67% proteins contain at least one Pfam profile HMM

Profile Hidden Markov Models

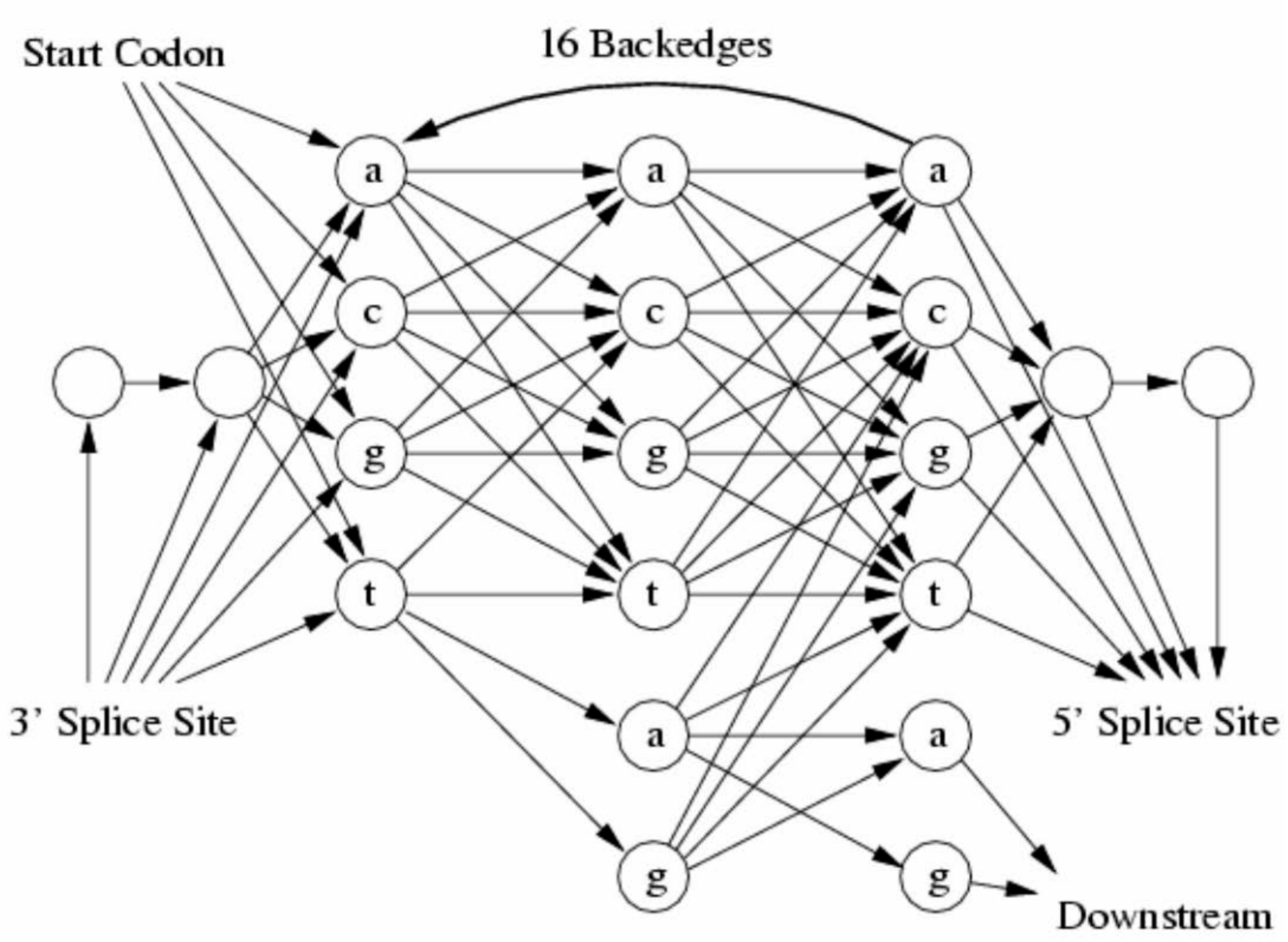
7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

HMM for splice site detection



Chapter 10

The Boltzmann Machine

Boltzmann Machines

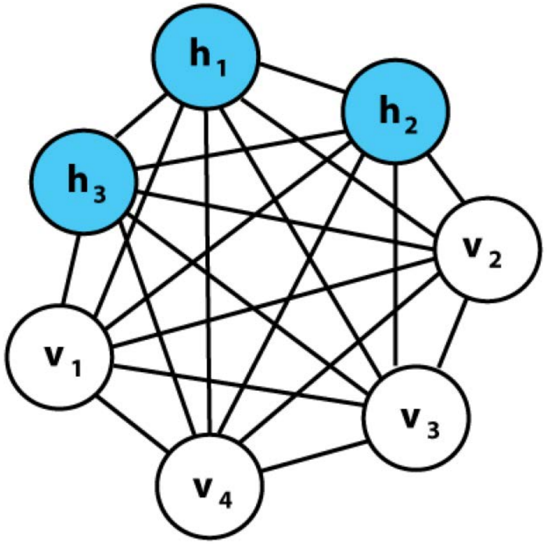
Boltzmann machine: stochastic recurrent neural network proposed by Geoffrey Hinton and Terry Sejnowski in 1985.

Boltzmann machines: not useful for practical problems

restricted Boltzmann machines: gained high popularity in the context of deep learning

name from statistical mechanics (normalizing distribution and used for sampling)

A graphical representation of an example Boltzmann machine. Each undirected edge represents dependency. In this example there are 3 hidden units and 4 visible units



Boltzmann Machines

Boltzmann machine: network of units with an “energy” of its current state; binary stochastic units, global energy

$$E \;=\; - \left(\sum_{i,j;i<j} w_{ij}\, s_i\, s_j + \sum_{i} \theta_i\, s_i \right)$$

- w_{ij} is the connection strength between unit j and unit i
- s_i is the state of unit i which is either 1 or 0
- θ_i is the bias or the activation threshold of unit i
- $w_{ii} = 0$ for all i, that is, units do not have self-connections
- $w_{ij} = w_{ji}$ for all i,j, that is, connections are symmetric

Weights are written by a symmetric matrix $\boldsymbol{W}$ with zero diagonal

Boltzmann Machines

7 Clustering
7.1 Mixture Models
7.2 k-Means
7.3 Hierarchical
7.4 Similarity-Based

8 Biclustering
8.1 Types of Biclusters
8.2 Overview Methods
8.3 FABIA
8.4 Examples

9 Hidden Markov Models
9.1 HMMs in Bioinf.
9.2 HMM Basics
9.3 EM / Baum-Welch
9.4 Viterby
9.5 Input Output HMM
9.6 Factorial HMM
9.7 Memory IOHMM
9.8 Tricks of the Trade
9.9 Profile HMM

10 Boltzmann Machines
10.1 The Boltzmann Machine
10.2 Learning
10.3 Restricted Boltzmann Machine

energy difference of a single unit: $\Delta E_i \ = \ \sum_j w_{ij}\, s_j \ + \ \theta_i$

Boltzmann distribution: energy of a state is proportional to the negative log probability of that state.

Boltzmann Factor: energy difference if one unit is flipped

$$\Delta E_i \ = \ -k_B\, T \ln(p_{\mathrm{i=off}}) \ - \ (-k_B\, T \ln(p_{\mathrm{i=on}}))$$

k_B : Boltzmann's constant is absorbed into the temperature T

energy difference at temperature T:

$$\begin{aligned}
& \frac{\Delta E_i}{T} \ = \ \ln(p_{\mathrm{i=on}}) \ - \ \ln(p_{\mathrm{i=off}}) \\
\Leftrightarrow \quad & \frac{\Delta E_i}{T} \ = \ \ln(p_{\mathrm{i=on}}) \ - \ \ln(1 - p_{\mathrm{i=on}}) \\
\Leftrightarrow \quad & \frac{\Delta E_i}{T} \ = \ \ln\left(\frac{p_{\mathrm{i=on}}}{1 - p_{\mathrm{i=on}}}\right) \\
\Leftrightarrow \quad & -\frac{\Delta E_i}{T} \ = \ \ln\left(\frac{1 - p_{\mathrm{i=on}}}{p_{\mathrm{i=on}}}\right) \\
\Leftrightarrow \quad & -\frac{\Delta E_i}{T} \ = \ \ln\left(\frac{1}{p_{\mathrm{i=on}}} \ - 1\right) \\
\Leftrightarrow \quad & \exp\left(-\frac{\Delta E_i}{T}\right) \ = \ \frac{1}{p_{\mathrm{i=on}}} \ - \ 1
\end{aligned}$$

Boltzmann Machines

➔ probability that the i-th unit is on:

$$p_{\mathrm{i=on}} = \frac{1}{1 + \exp(-\frac{\Delta E_i}{T})}$$

→ logistic function is used as activation probability

The network runs iteratively choosing a unit and setting its state according to its probability from above.

When the machine is “at thermal equilibrium” the probability distribution of global states has converged.

starts with a high temperature which gradually decreases

Boltzmann Machines: Learning

Learning in the Boltzmann Machine

Training: weights so that the global states with the highest probabilities will get the lowest energies

- “**visible**” units V: receive binary training input from the “environment”
- “**hidden**” units H

distribution over the training set: $P^{+}(V)$
distribution over converged global states and after marginalization over the hidden units: $P^{-}(V)$

Boltzmann Machine learning: $P^{-}(V) \approx P^{+}(V)$

Objective is the Kullback-Leibler divergence:

$$D_{\mathrm{KL}}(P^{+} \parallel P^{-}) \ = \ \sum_{v} P^{+}(v) \ \ln \left(\frac{P^{+}(v)}{P^{-}(v)} \right)$$

sum is over all the possible states of V

Boltzmann Machines: Learning

Analog to the EM algorithm, Boltzmann machine training has two phases that are performed alternating:

- “positive‘” phase: visible units' states are clamped to the training data
- ”negative” phase the network runs freely

gradient of the objective with respect to a given weight:

$$\frac{\partial D}{\partial w_{ij}} = - \frac{1}{\eta} \left(p_{ij}^{+} - p_{ij}^{-}\right)$$

- p_{ij}^{+} is the probability of units i and j both being on when the machine is at equilibrium on the positive phase
- p_{ij}^{-} is the probability of units i and j both being on when the machine is at equilibrium on the negative phase
- η is the learning rate

Learning by using only “local information” → biological plausible

Boltzmann Machines: Learning

training the bias weight: $$\frac{\partial D}{\partial \theta_i} = -\frac{1}{\eta}\left(p_i^+ - p_i^-\right)$$

Problems:

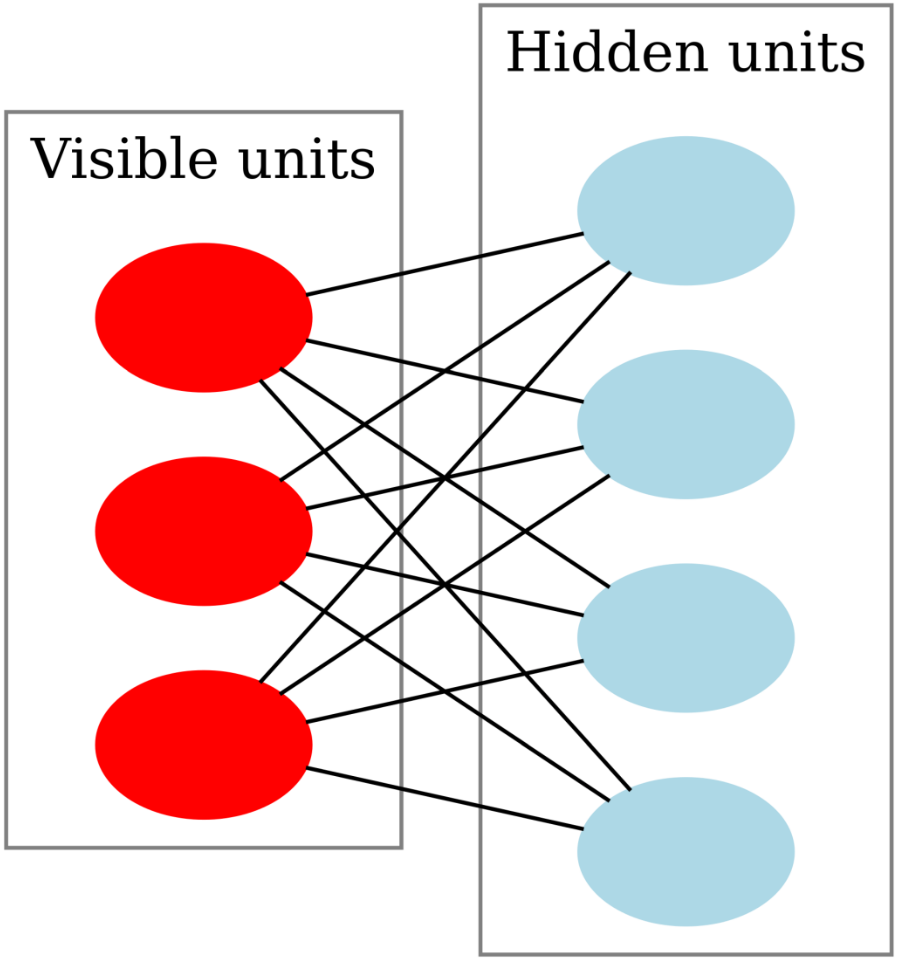
- the time to equilibrium grows exponentially with the size and with the magnitude of the connection strengths

- probabilities between zero and one → more plastic, variance trap
net effect → noise causes the connection strengths to random walk until the activities saturate

Boltzmann Machines: RBM

Restricted Boltzmann Machine (RBM)

- no intralayer connections between hidden or visible units
- activities of its hidden units are a representation of the visible units
- **deep learning** idea: hidden units of one RBM are the visible units of a higher-level RBM → stacked RBMs



Graphical representation of a restricted Boltzmann machine. The four blue units represent hidden units, and the three red units represent visible states. In restricted Boltzmann machines there are only connections (dependencies) between hidden and visible units, and none between units of the same type.

Boltzmann Machines: RBM

$w_{i,j}$: connection between hidden unit h_j and visible unit v_i
bias weight visible unit: a_i
bias weight visible unit: b_j
energy:

$$E(\boldsymbol{v},\boldsymbol{h}) \;=\; -\sum_i a_i\, v_i \;-\; \sum_j b_j\, h_j \;-\; \sum_{i,j} h_j\, v_i\, w_{i,j}$$

$$E(\boldsymbol{v},\boldsymbol{h}) \;=\; -\,\boldsymbol{a}^T\boldsymbol{v} \;-\; \boldsymbol{b}^T\boldsymbol{h} \;-\; \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}$$

Probability distributions over hidden and visible vectors:

$$p(\boldsymbol{v},\boldsymbol{h}) \;=\; \frac{1}{Z}\, e^{-E(\boldsymbol{v},\boldsymbol{h})}$$

, where Z is a normalizing constant or the partition function defined as the sum of $e^{-E(\boldsymbol{v},\boldsymbol{h})}$ over all possible configurations.

Marginalizing → probability of a visible (input) vector:

$$p(\boldsymbol{v}) \;=\; \frac{1}{Z} \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h})}$$

Boltzmann Machines: RBM

Mutual conditional independence:

$$p(\boldsymbol{v} \mid \boldsymbol{h}) \;=\; \prod_{i=1}^{m} p(v_i \mid \boldsymbol{h}) \qquad p(h_j = 1 \mid \boldsymbol{v}) \;=\; \sigma(b_j + \sum_{i=1}^{m} w_{i,j}\; v_i)$$

$$p(\boldsymbol{h} \mid \boldsymbol{v}) \;=\; \prod_{j=1}^{n} p(h_j \mid \boldsymbol{v}) \qquad p(v_i = 1 \mid \boldsymbol{h}) \;=\; \sigma(a_i + \sum_{j=1}^{n} w_{i,j}\; h_j)$$

where σ is the sigmoid function

Training $\quad \arg\max_{\boldsymbol{W}} \prod_{\boldsymbol{v} \in V} p(\boldsymbol{v}) \qquad \arg\max_{\boldsymbol{W}} \mathrm{E}\left(\sum_{\boldsymbol{v} \in V} \log p(\boldsymbol{v})\right)$

Boltzmann Machines: RBM

train RBMs by contrastive divergence (CD):

1. Take a training sample $\boldsymbol{v}$, compute the probabilities of the hidden units and sample a hidden activation vector $\boldsymbol{h}$ from this probability distribution

2. Compute the outer product of $\boldsymbol{v}$ and $\boldsymbol{h}$ and call this the “positive gradient”

3. From $\boldsymbol{h}$, sample a reconstruction $\boldsymbol{v}'$ of the visible units, then resample the hidden activations $\boldsymbol{h}'$ from this

4. Compute the outer product of $\boldsymbol{v}'$ and $\boldsymbol{h}'$ and call this the “negative gradient”

5. Let the weight update to $w_{i,j}$ be the “positive gradient” minus the “negative gradient”, times some learning rate η