

Clustering

Rui Kuang

**Department of Computer Science and Engineering
University of Minnesota**



How to Compress a Image?



Original transparent PNG
File size: **57 KB**

VS

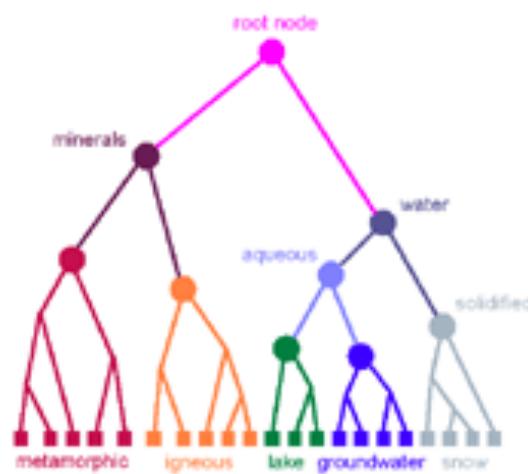
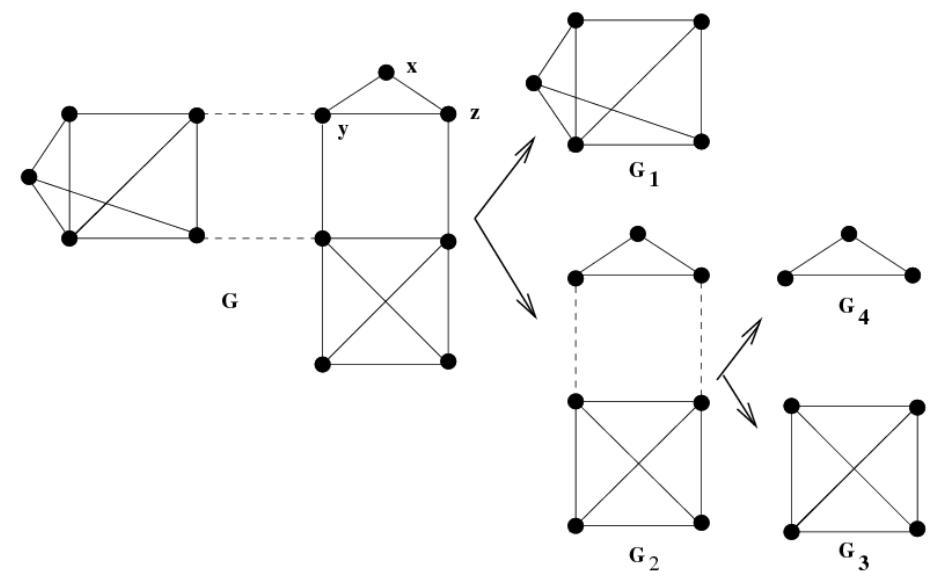
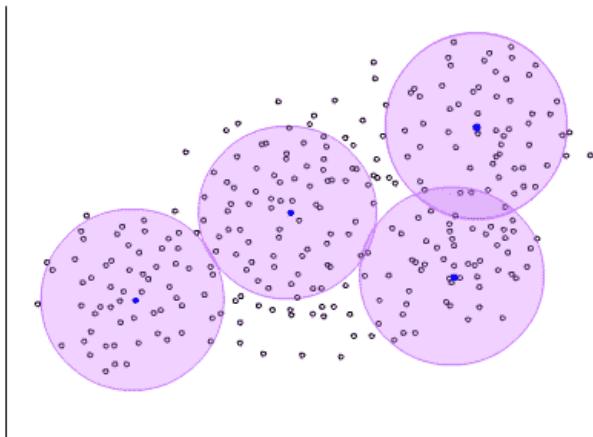


Shrunk transparent PNG
File size: **16 KB**

RGB Colour Codes

| | | | | | | | |
|---------|---------|---------|---------|---------|---------|---------|---------|
| #000000 | #330000 | #660000 | #990000 | #CC0000 | #FF0000 | #110000 | #001100 |
| #003300 | #333300 | #663300 | #993300 | #CC3300 | #FF3300 | #220000 | #002200 |
| #006600 | #336600 | #666600 | #996600 | #CC6600 | #FF6600 | #330000 | #003300 |
| #009900 | #339900 | #669900 | #999900 | #CC9900 | #FF9900 | #440000 | #004400 |
| #00CC00 | #33CC00 | #66CC00 | #99CC00 | #CCCC00 | #FFCC00 | #550000 | #005500 |
| #00FF00 | #33FF00 | #66FF00 | #99FF00 | #CCFF00 | #FFFF00 | #660000 | #006600 |
| #000033 | #330033 | #660033 | #990033 | #CC0033 | #FF0033 | #770000 | #007700 |
| #003333 | #333333 | #663333 | #993333 | #CC3333 | #FF3333 | #880000 | #008800 |
| #006633 | #336633 | #666633 | #996633 | #CC6633 | #FF6633 | #990000 | #009900 |
| #009933 | #339933 | #669933 | #999933 | #CC9933 | #FF9933 | #AA0000 | #00AA00 |
| #00CC33 | #33CC33 | #66CC33 | #99CC33 | #CCCC33 | #FFCC33 | #BB0000 | #00BB00 |
| #00FF33 | #33FF33 | #66FF33 | #99FF33 | #CCFF33 | #FFFF33 | #CC0000 | #00CC00 |
| #000066 | #330066 | #660066 | #990066 | #CC0066 | #FF0066 | #DD0000 | #00DD00 |
| #003366 | #333366 | #663366 | #993366 | #CC3366 | #FF3366 | #EE0000 | #00EE00 |
| #006666 | #336666 | #666666 | #996666 | #CC6666 | #FF6666 | #FF0000 | #00FF00 |
| #009966 | #339966 | #669966 | #999966 | #CC9966 | #FF9966 | #000011 | #110011 |
| #00CC66 | #33CC66 | #66CC66 | #99CC66 | #CCCC66 | #FFCC66 | #000022 | #220033 |
| #00FF66 | #33FF66 | #66FF66 | #99FF66 | #CCFF66 | #FFFF66 | #000033 | #330033 |
| #000099 | #330099 | #660099 | #990099 | #CC0099 | #FF0099 | #000044 | #440044 |
| #003399 | #333399 | #663399 | #993399 | #CC3399 | #FF3399 | #000055 | #550055 |
| #006699 | #336699 | #666699 | #996699 | #CC6699 | #FF6699 | #000066 | #660066 |
| #009999 | #339999 | #669999 | #999999 | #CC9999 | #FF9999 | #000077 | #770077 |

How to Do Clustering?



k -Means Clustering

- Find k reference vectors
(prototypes/codebook vectors/codewords)
which best represent data $\|\mathbf{x}^t - \mathbf{m}_i\| = \min_j \|\mathbf{x}^t - \mathbf{m}_j\|$
- Reference vectors, $\mathbf{m}_j, j = 1, \dots, k$
- Reconstruction error:

$$E\left(\left\{\mathbf{m}_i\right\}_{i=1}^k | \mathcal{X}\right) = \sum_t \sum_i b_i^t \|\mathbf{x}^t - \mathbf{m}_i\|^2$$

$$b_i^t = \begin{cases} 1 & \text{if } \|\mathbf{x}^t - \mathbf{m}_i\| = \min_j \|\mathbf{x}^t - \mathbf{m}_j\| \\ 0 & \text{otherwise} \end{cases}$$

k -Means Clustering

- Minimizing reconstruction error:

$$E\left(\left\{\mathbf{m}_i\right\}_{i=1}^k | \mathcal{X}\right) = \sum_t \sum_i b_i^t \|\mathbf{x}^t - \mathbf{m}_i\|^2$$

$$b_i^t = \begin{cases} 1 & \text{if } \|\mathbf{x}^t - \mathbf{m}_i\| = \min_j \|\mathbf{x}^t - \mathbf{m}_j\| \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial \sum_t \sum_i b_i^t \|\mathbf{x}^t - \mathbf{m}_i\|^2}{\partial m_i} = 2 \sum_t b_i^t (\mathbf{x}^t - \mathbf{m}_i) = 0$$

$$\mathbf{m}_i = \frac{\sum_t b_i^t \mathbf{x}^t}{\sum_t b_i^t}$$

k-means Clustering

Initialize $\mathbf{m}_i, i = 1, \dots, k$, for example, to k random \mathbf{x}^t

Repeat

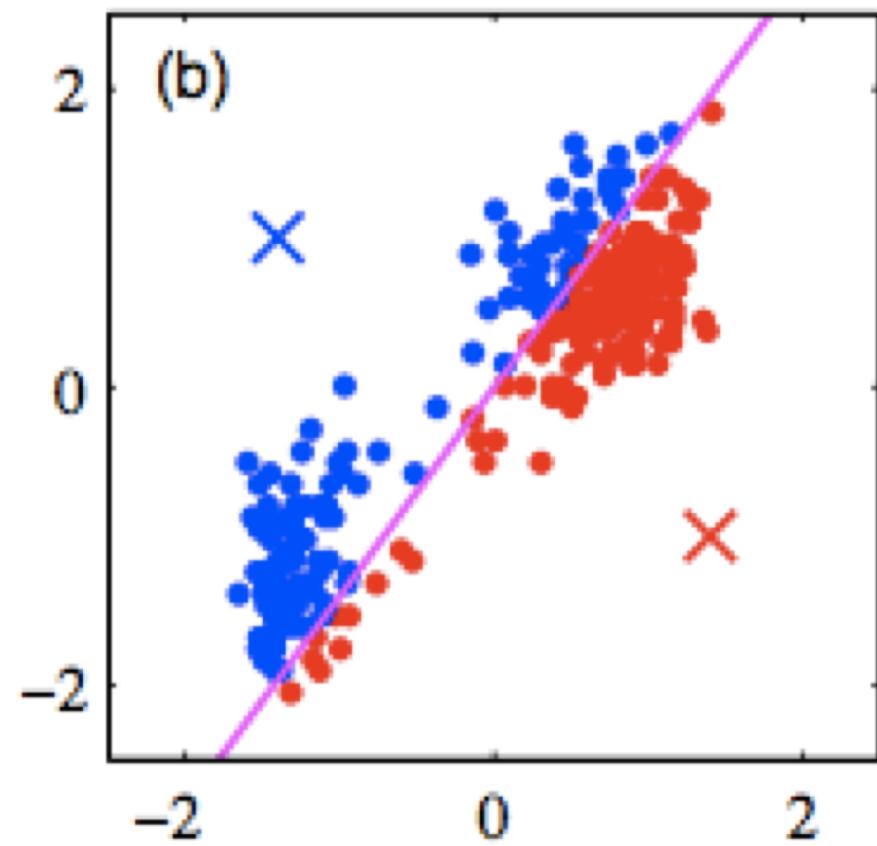
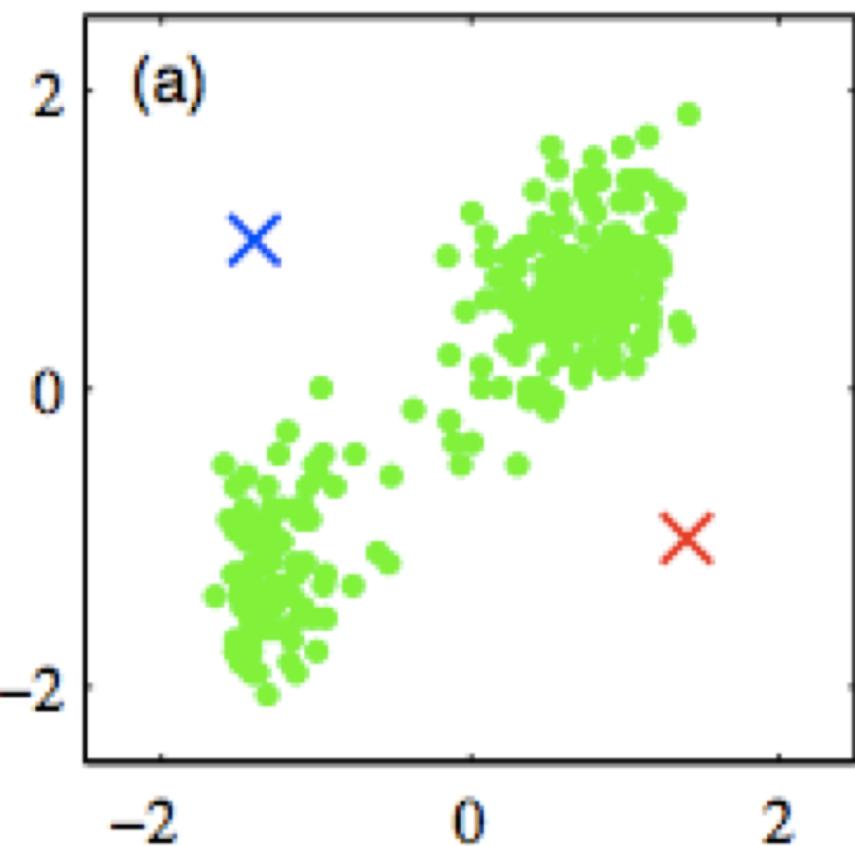
For all $\mathbf{x}^t \in \mathcal{X}$

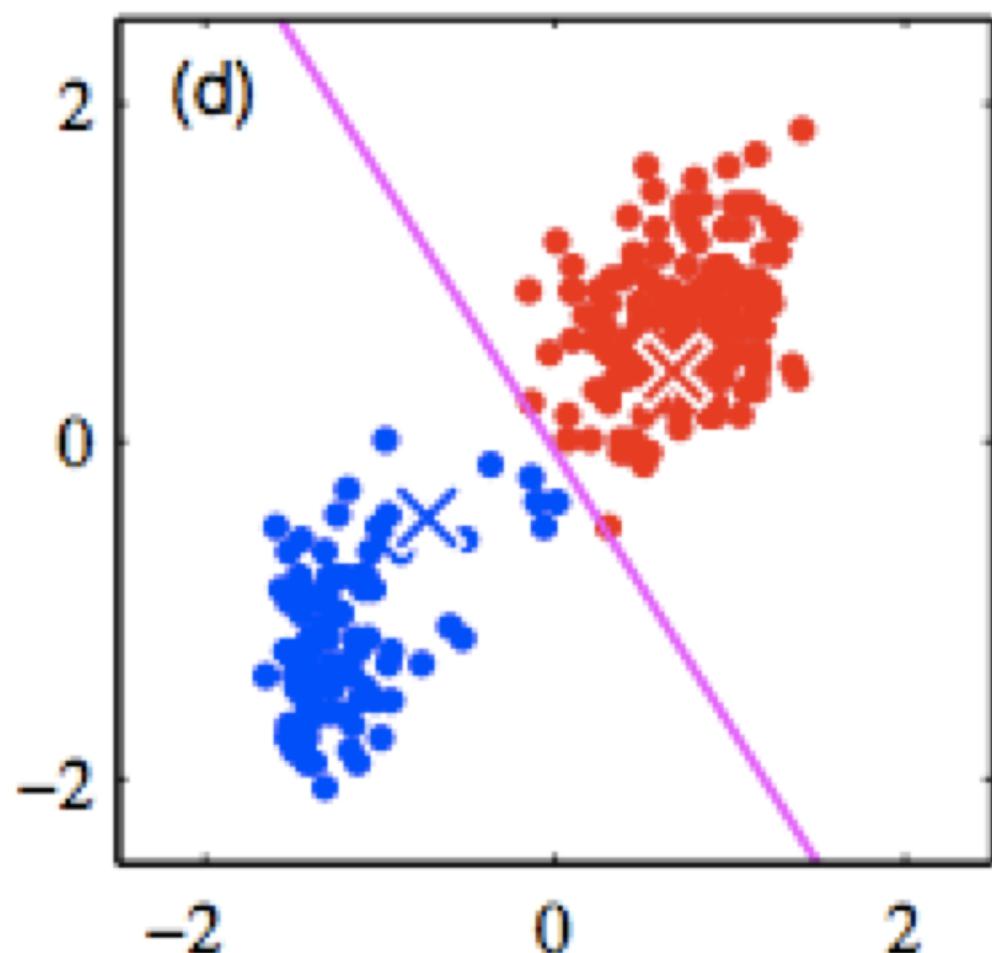
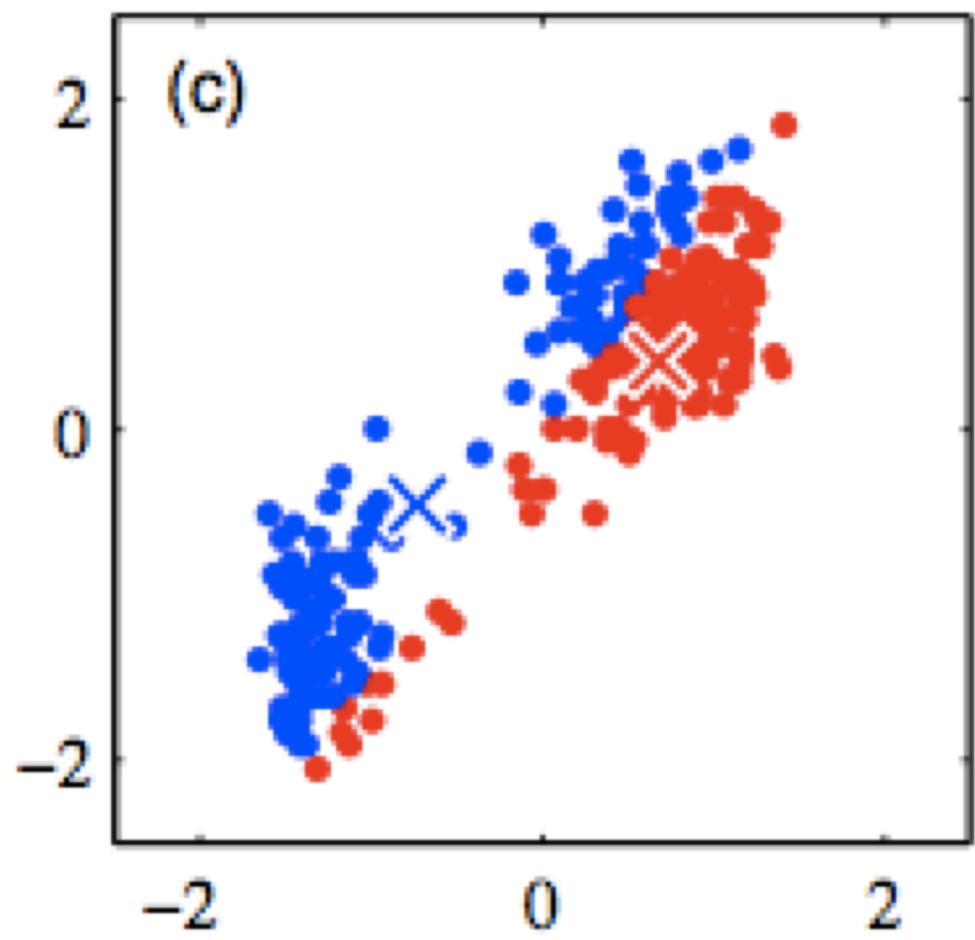
$$b_i^t \leftarrow \begin{cases} 1 & \text{if } \|\mathbf{x}^t - \mathbf{m}_i\| = \min_j \|\mathbf{x}^t - \mathbf{m}_j\| \\ 0 & \text{otherwise} \end{cases}$$

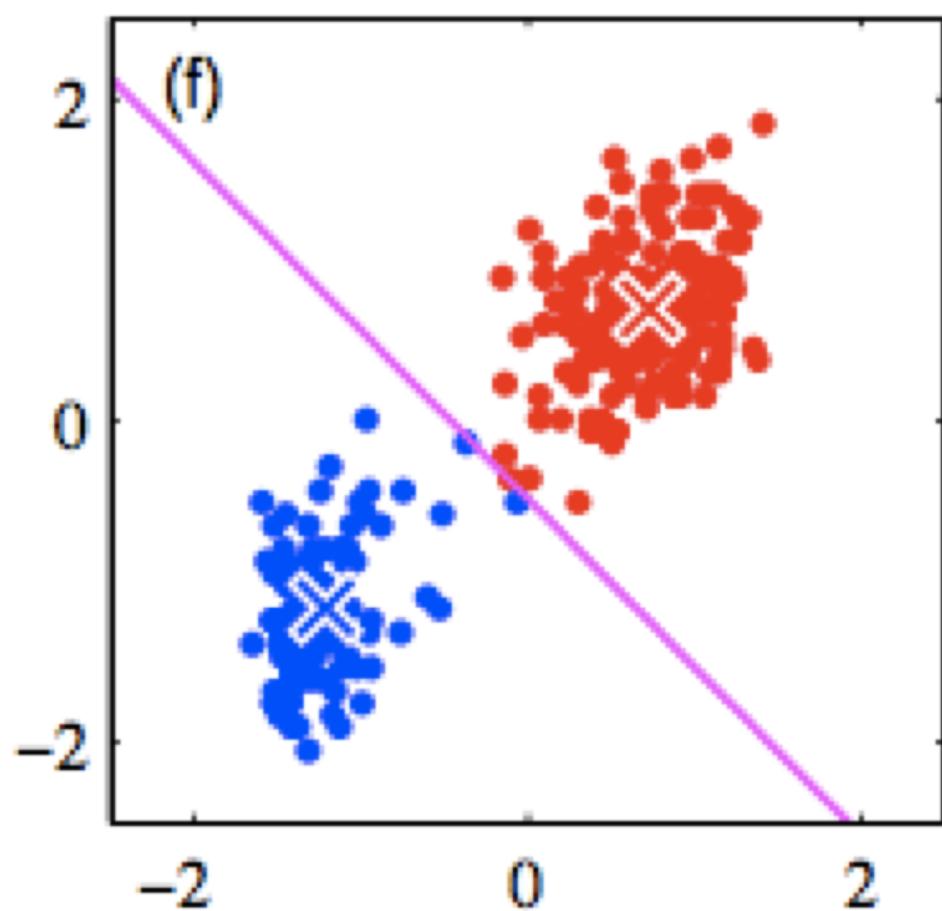
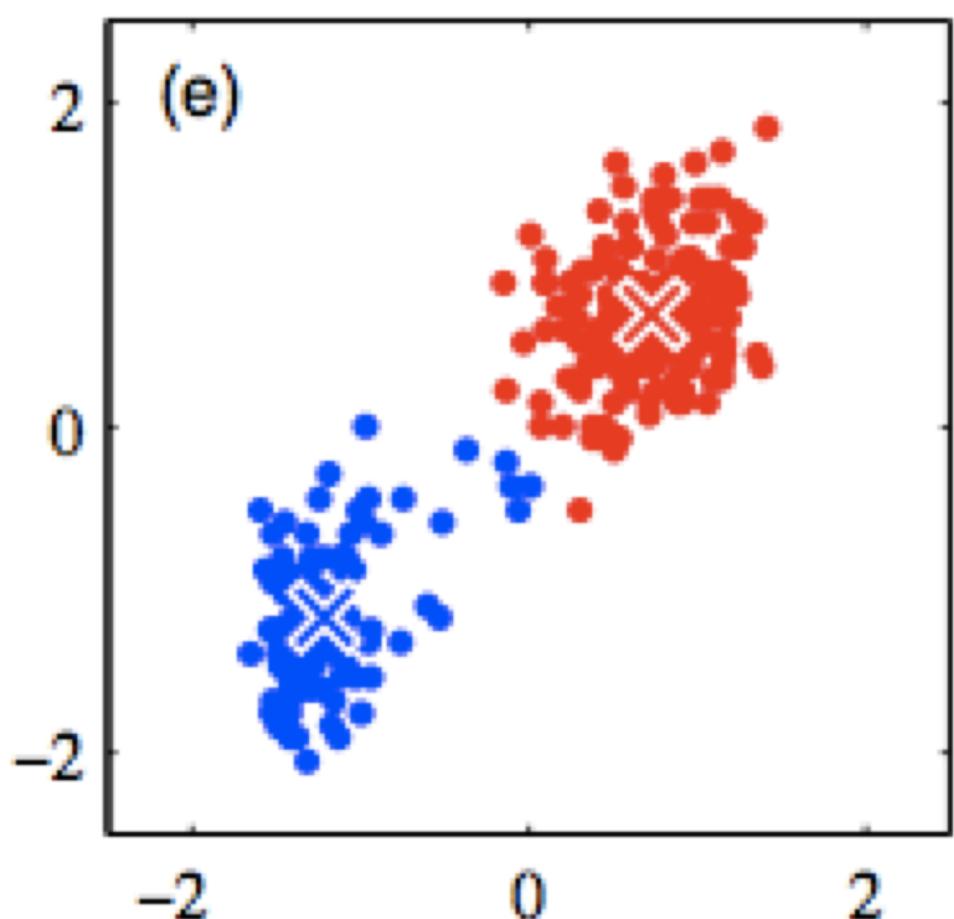
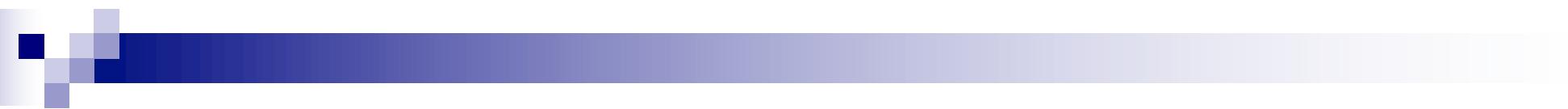
For all $\mathbf{m}_i, i = 1, \dots, k$

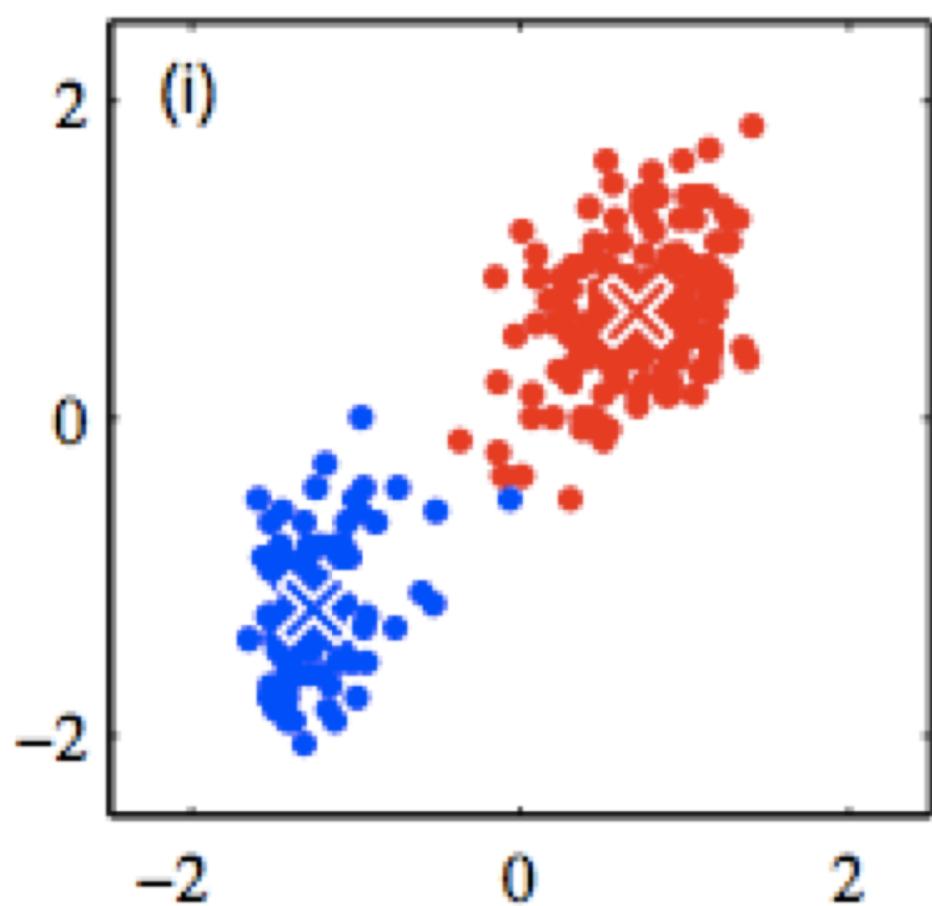
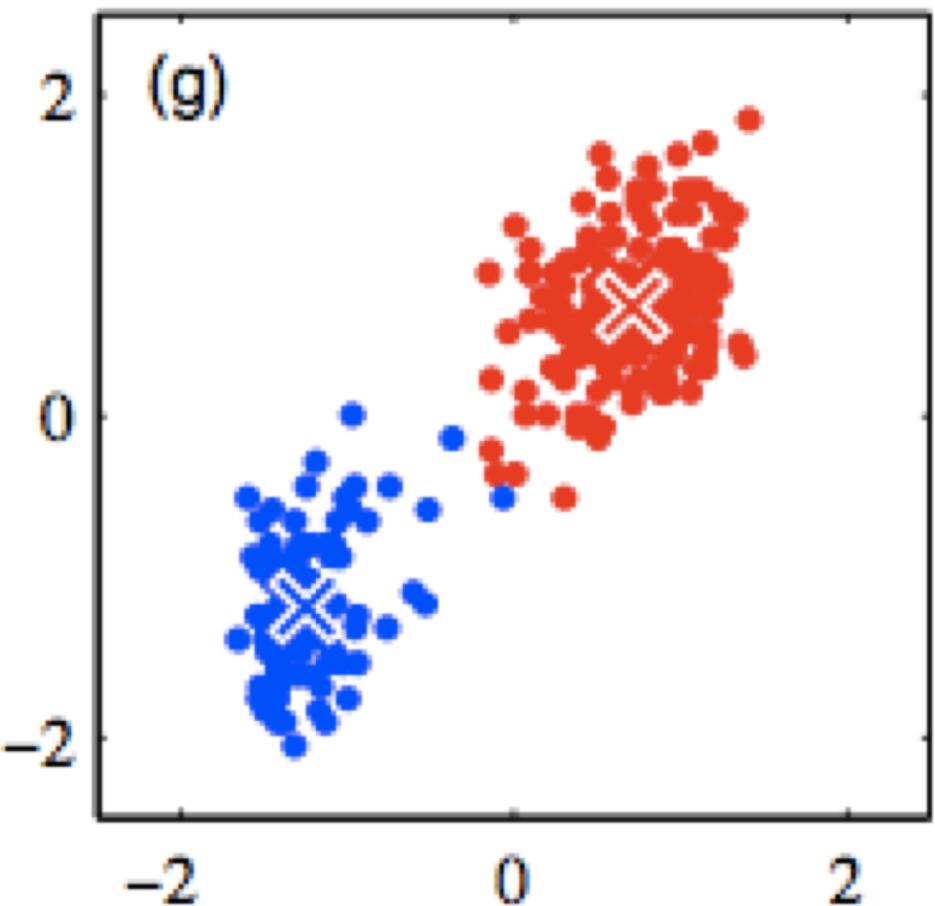
$$\mathbf{m}_i \leftarrow \sum_t b_i^t \mathbf{x}^t / \sum_t b_i^t$$

Until \mathbf{m}_i converge

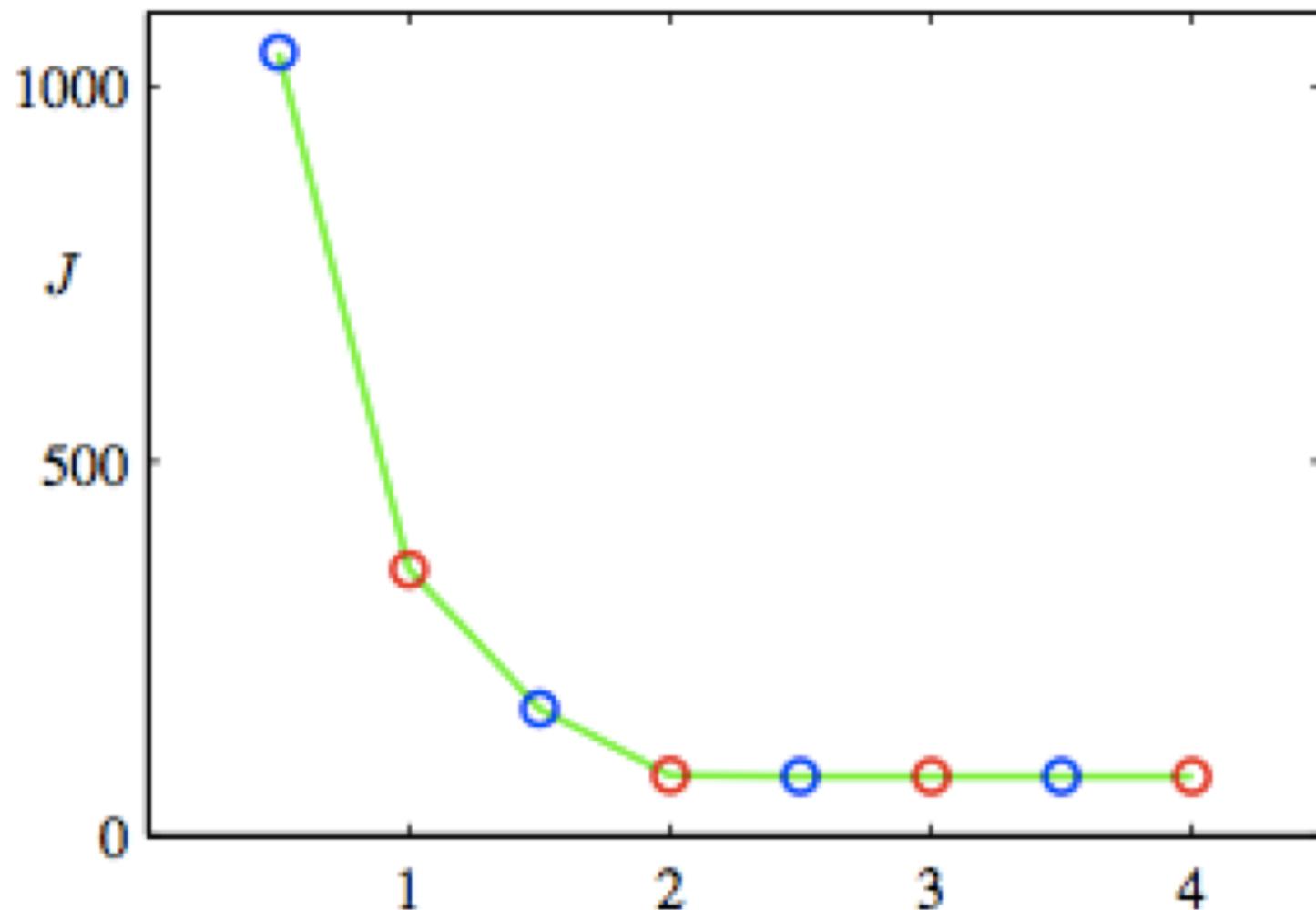








Reconstruction Error

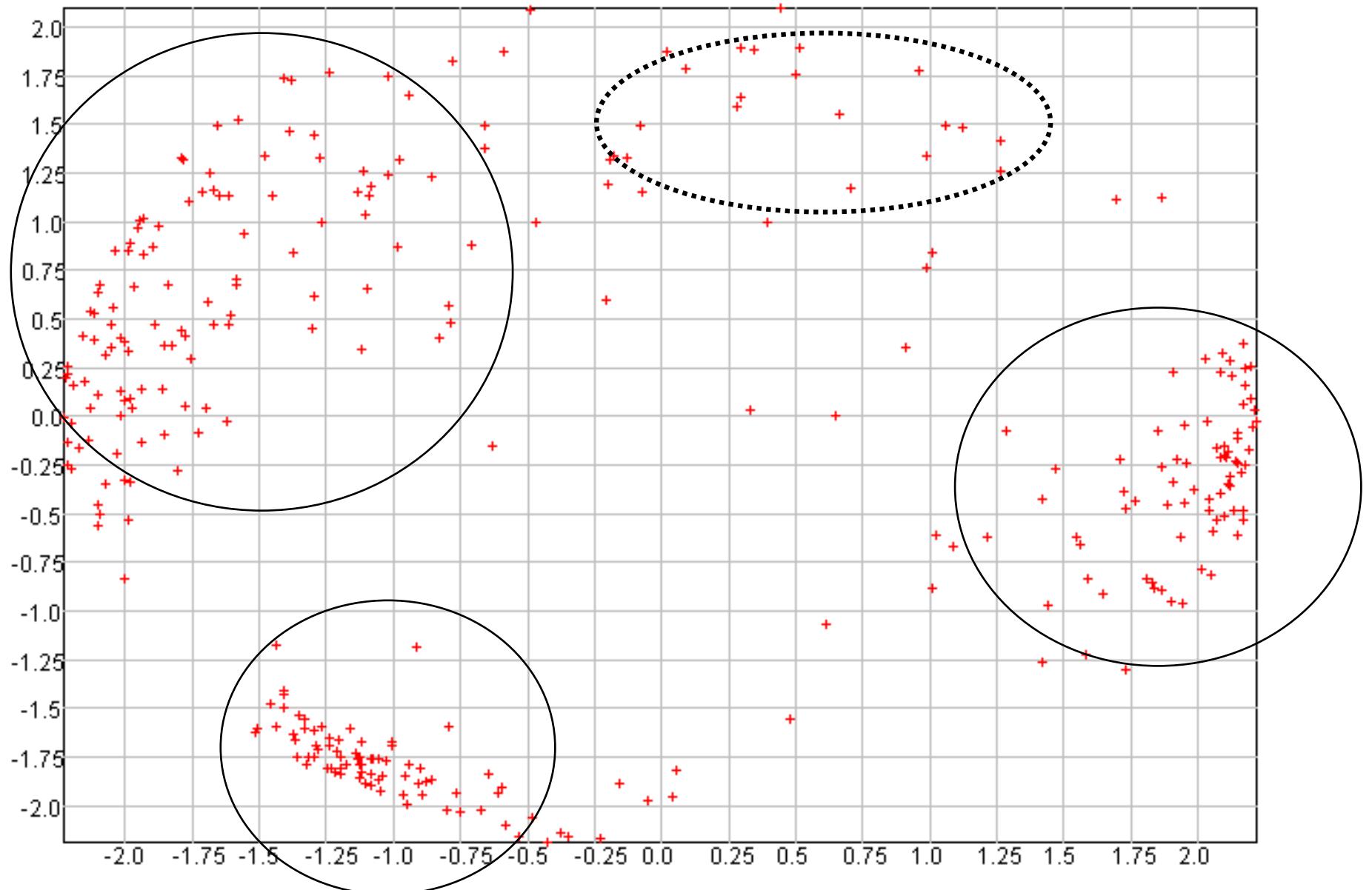


Choosing k

- Defined by the application, e.g., image quantization
- Incremental (leader-cluster) algorithm: Add one at a time until “elbow” (reconstruction error/log likelihood/intergroup distances)
- Plot data (after PCA) and check for clusters

PCA – Example

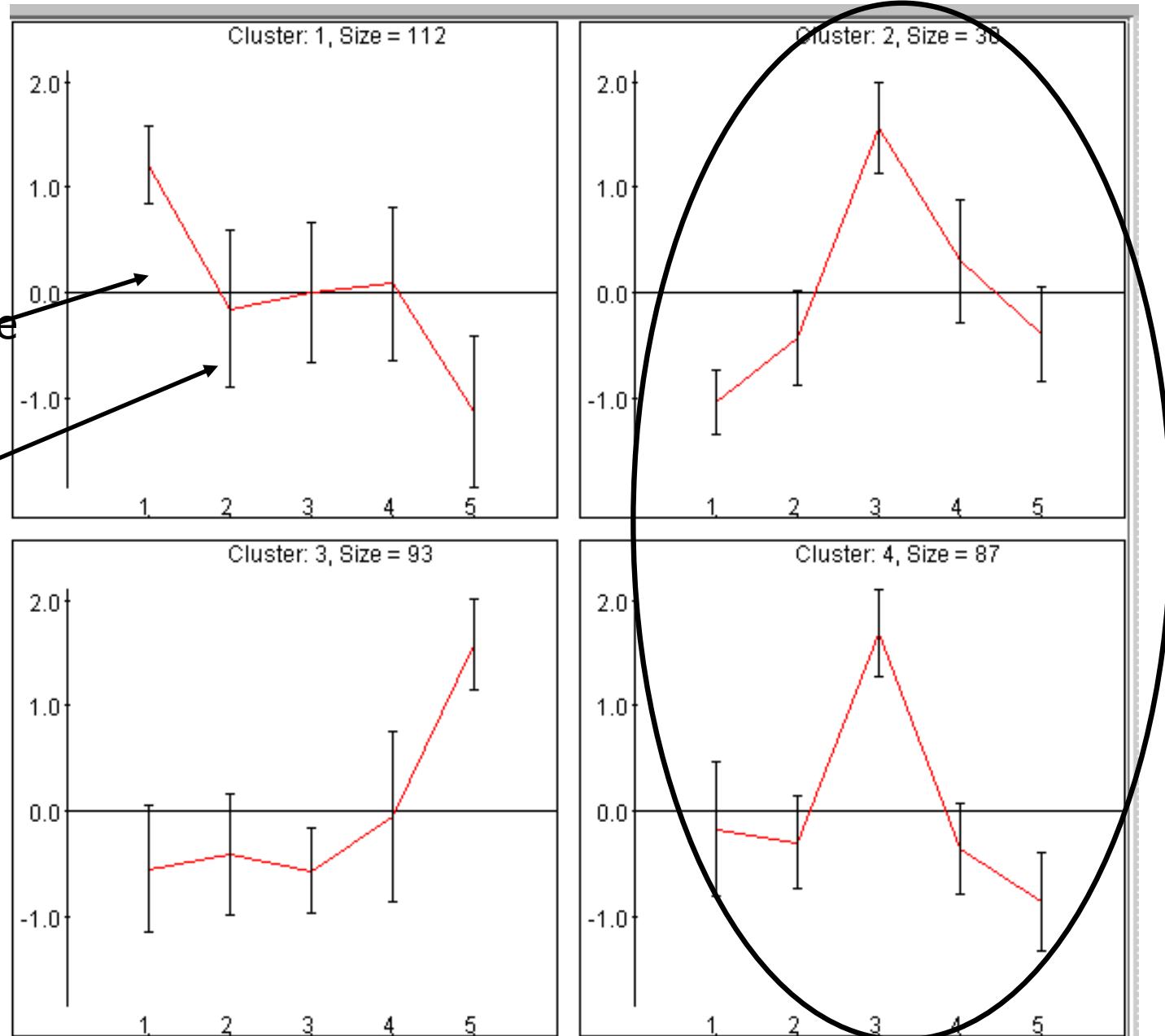
Visual estimation of the number of clusters in the data

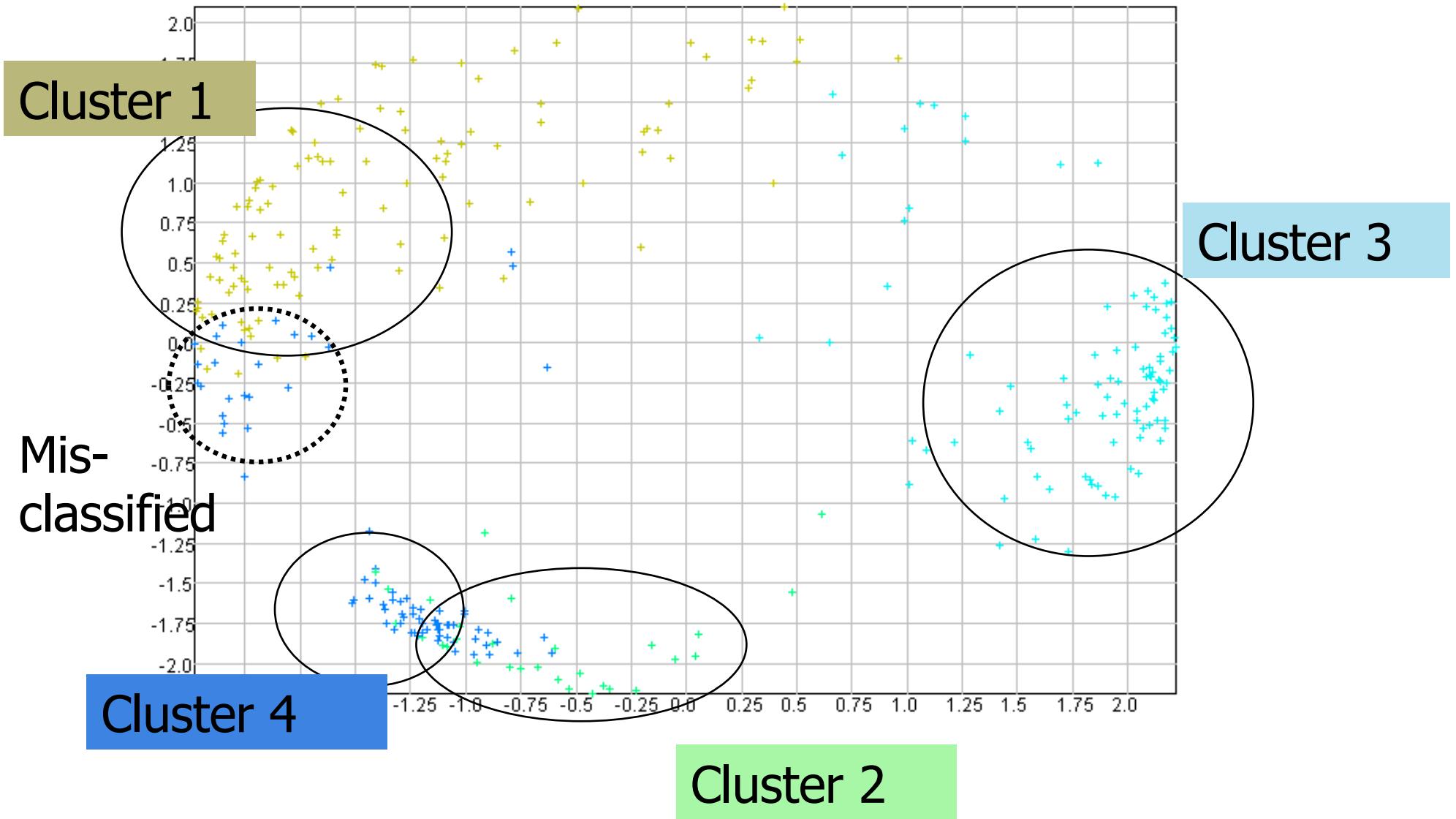
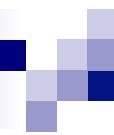


K-MEANS example: 4 clusters

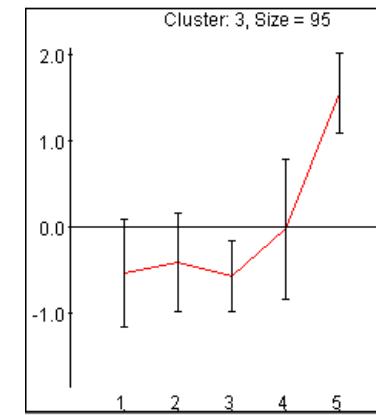
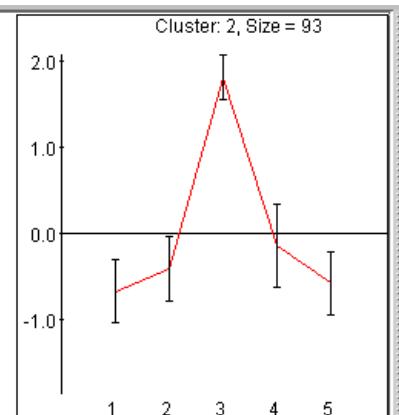
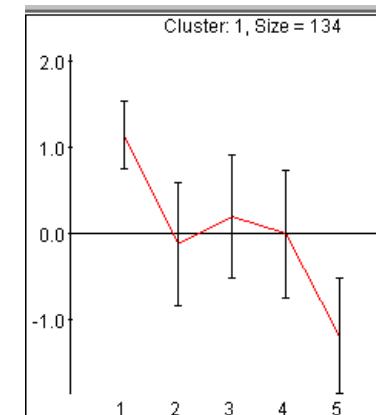
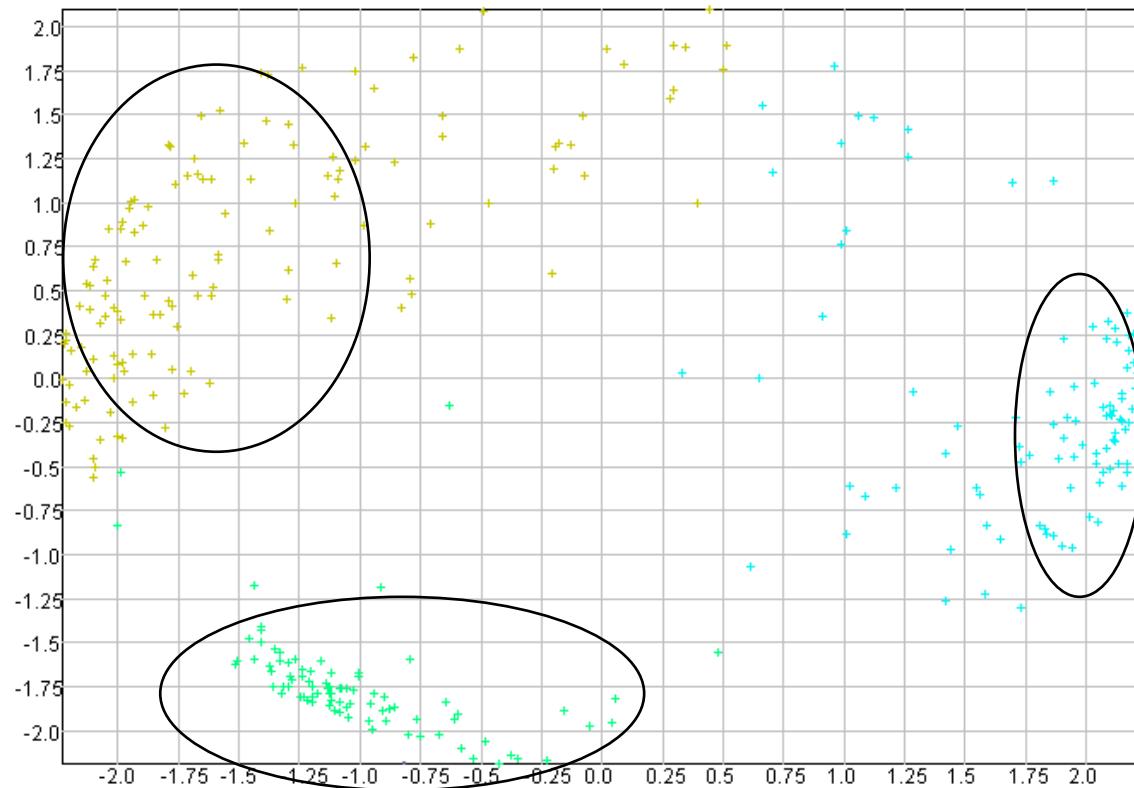
Mean profile

Std in each dimension

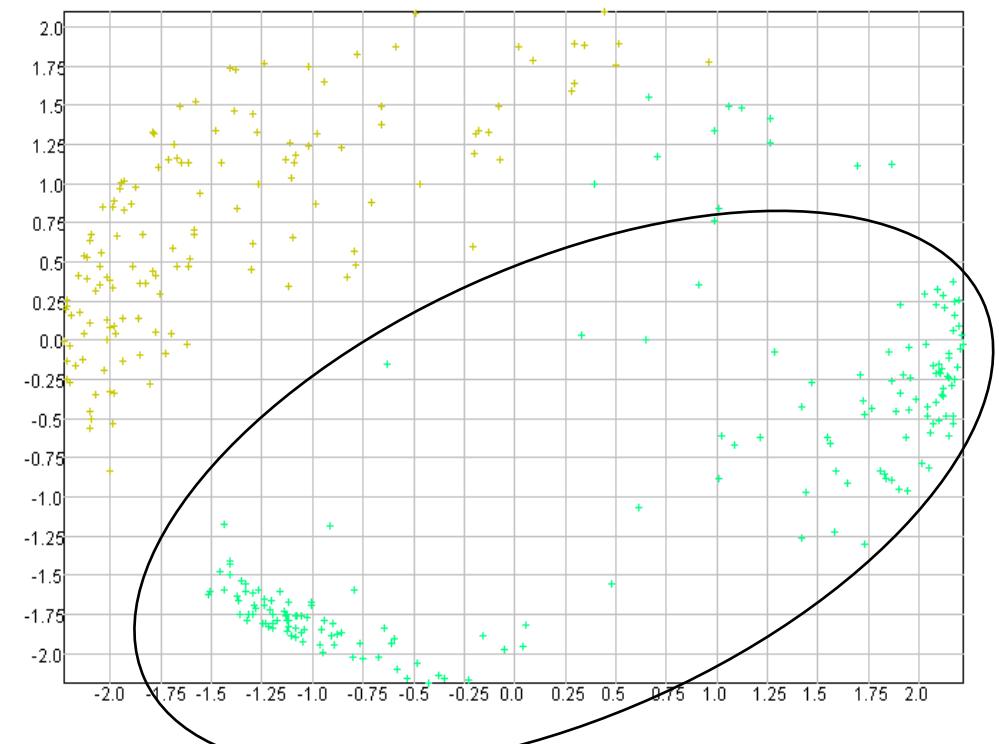
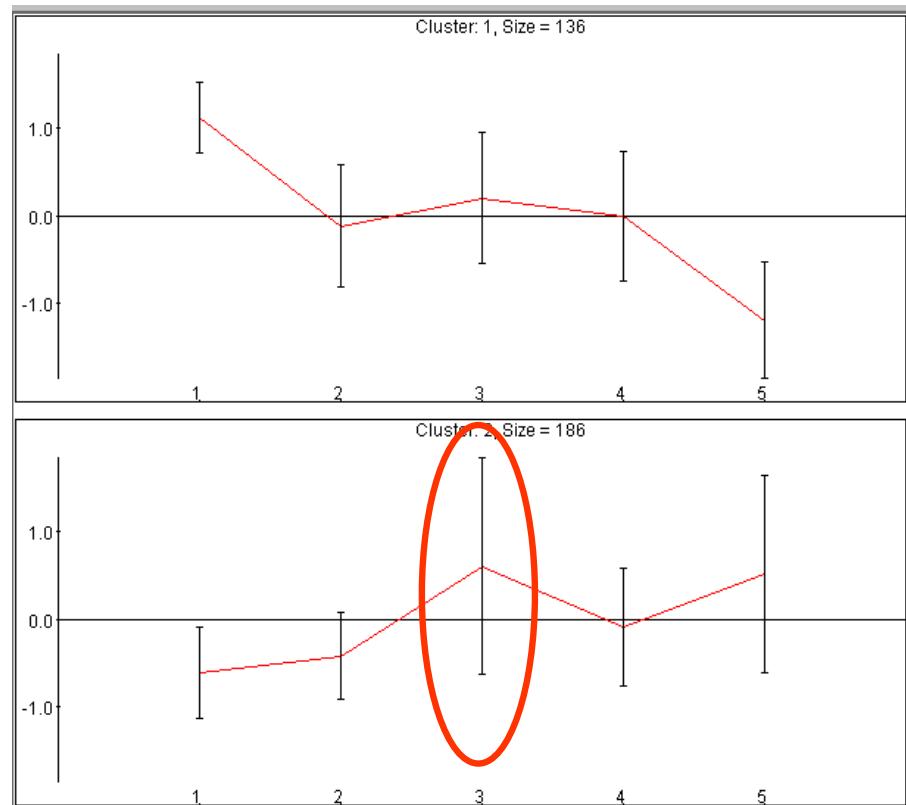




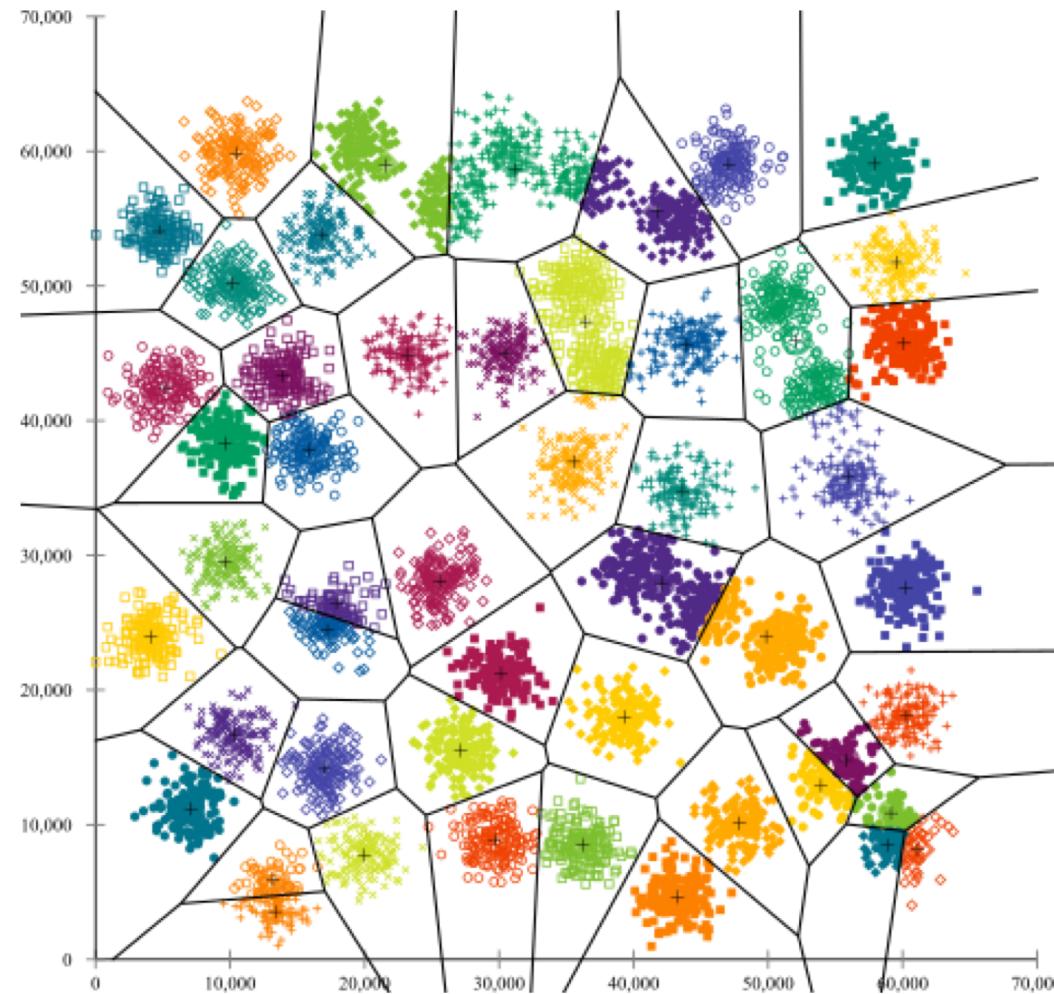
K-means example: 3 clusters



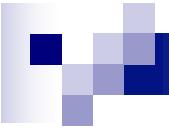
Too few clusters: K=2



K -means Local Minimum



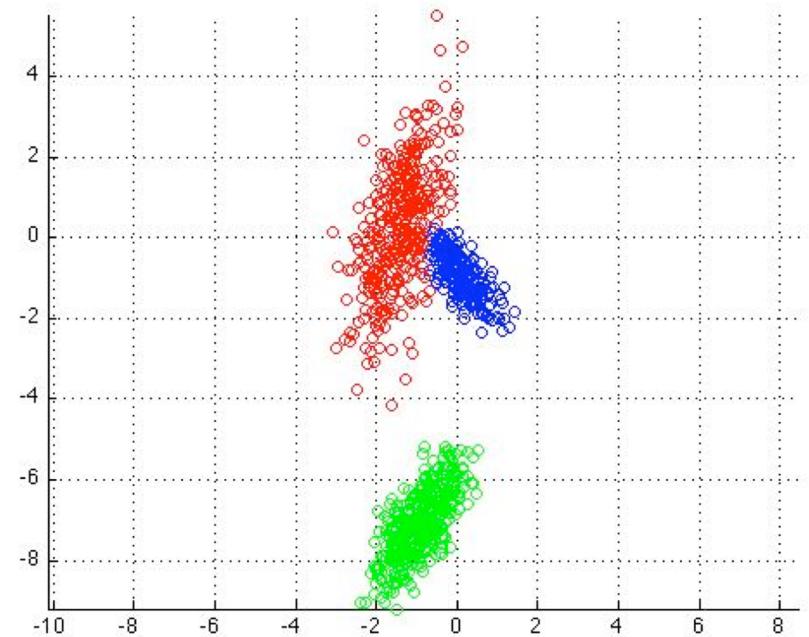
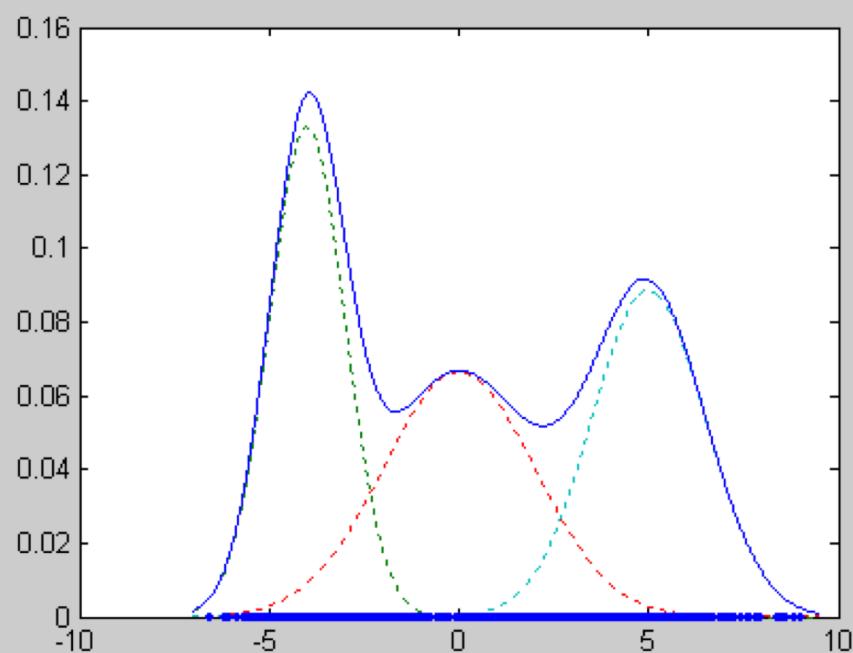
<https://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means>



K-means Tips

- Local minimum: run multiple times and choose the one with the lowest reconstruction error.
- Collapsing on one data point and empty clusters.
- SSE (sum of squared errors) assumes spherical variances.

Mixture of Gaussians



$$p(\mathbf{x} | \Phi) = \sum_{i=1}^K \pi_i \mathcal{N}(\mathbf{x} | \mu_i, \Sigma_i),$$

with $0 \leq \pi_i \leq 1$ and $\sum_{i=1}^K \pi_i = 1$.

Difficult ML Problem

■ Maximum Likelihood

$$\mathcal{L}(\Phi|\mathcal{X}) = \sum_t \log \sum_{i=1}^k p(\mathbf{x}^t | G_i) P(G_i) = \sum_t \log \sum_{i=1}^k \pi_i \mathcal{N}(\mathbf{x}^t | \mu_i, \Sigma_i)$$

■ Take derivative wrt μ_i

$$0 = - \sum_{t=1}^N \frac{\pi_i \mathcal{N}(x^t | \mu_i, \Sigma_i)}{\sum_{j=1}^K \pi_j \mathcal{N}(x^t | \mu_j, \Sigma_j)} \Sigma_i^{-1} (x^t - \mu_i)$$

$\gamma(z_i^t)$

$$\mu_i = \frac{1}{N_i} \sum_{t=1}^N \gamma(z_i^t) \mathbf{x}^t, \quad N_i = \sum_{t=1}^N \gamma(z_i^t)$$

Mixture of Gaussians

■ Maximum Likelihood

$$\mathcal{L}(\Phi | \mathcal{X}) = \sum_t \log \sum_{i=1}^k p(\mathbf{x}^t | G_i) P(G_i) = \sum_t \log \sum_{i=1}^k \pi_i \mathcal{N}(\mathbf{x} | \mu_i, \Sigma_i)$$

■ Take derivative wrt Σ_i

$$\Sigma_i = \frac{1}{N_i} \sum_{t=1}^N \gamma(z_i^t)(x^t - \mu_i)(x^t - \mu_i)^T$$

■ Take derivative wrt π_i

$$\frac{\partial \sum_t \log \sum_{i=1}^k \pi_i \mathcal{N}(\mathbf{x} | \mu_i, \Sigma_i) + \alpha (\sum_{i=1}^K \pi_i - 1)}{\partial \pi_i} = 0 \Leftrightarrow \pi_i = \frac{N_i}{N}$$

EM on Gaussian Mixtures

■ E-step:

$$\gamma(z_i) = \frac{\pi_i \mathcal{N}(\mathbf{x} | \mu_i, \Sigma_i)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x} | \mu_j, \Sigma_j)}$$

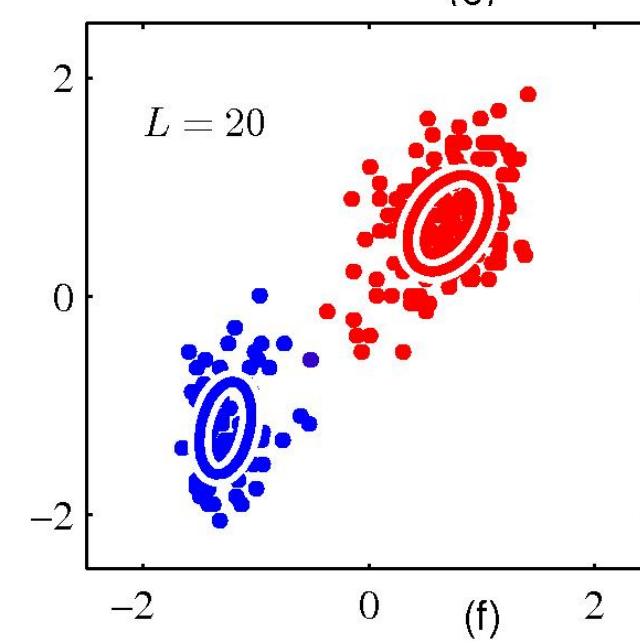
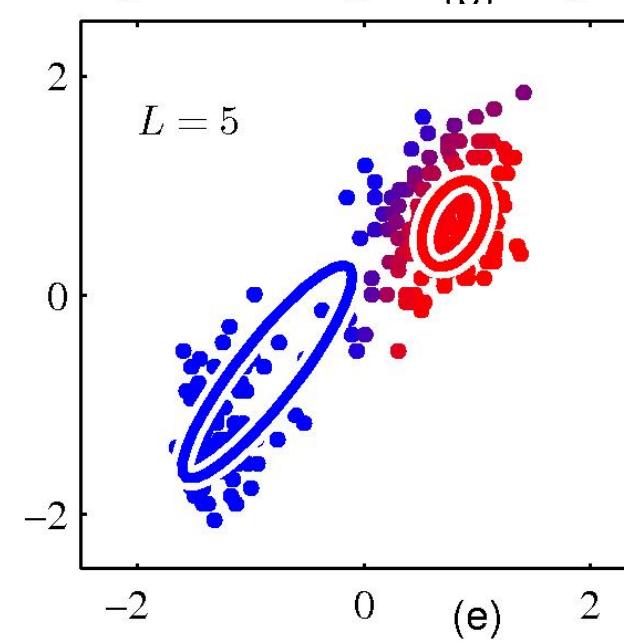
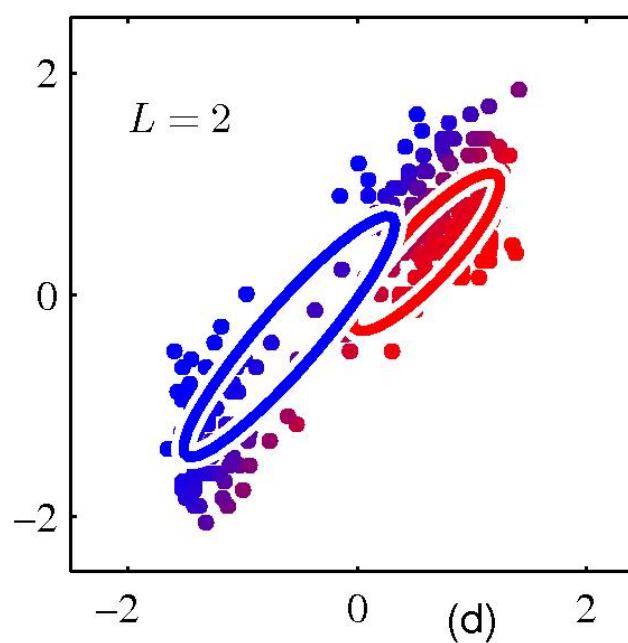
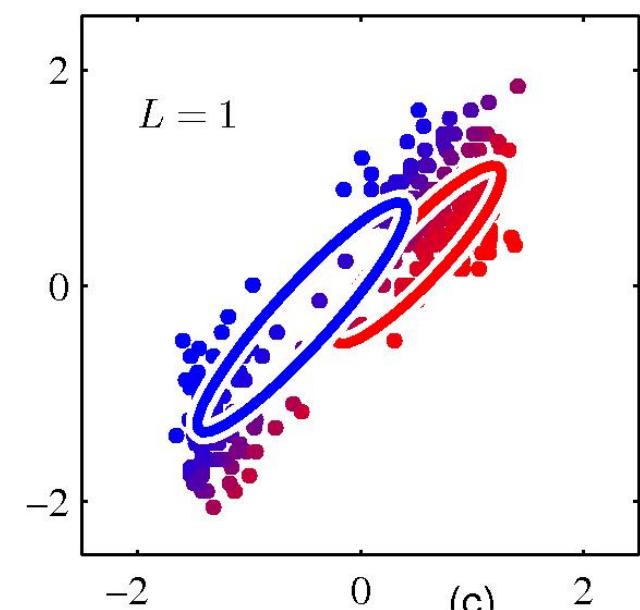
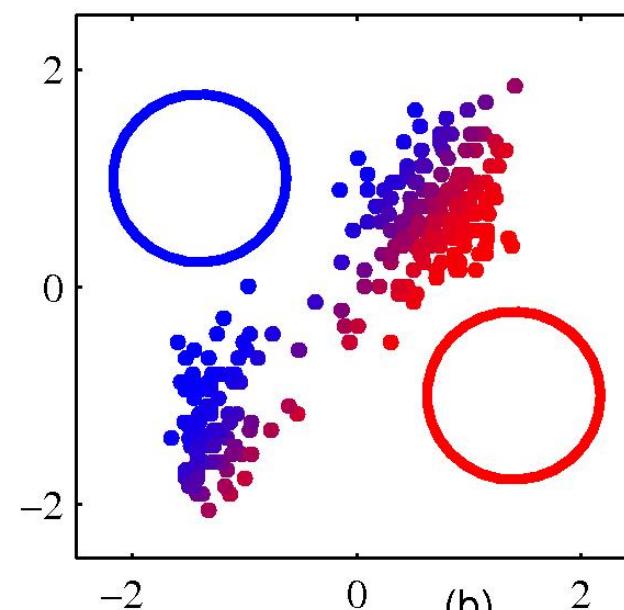
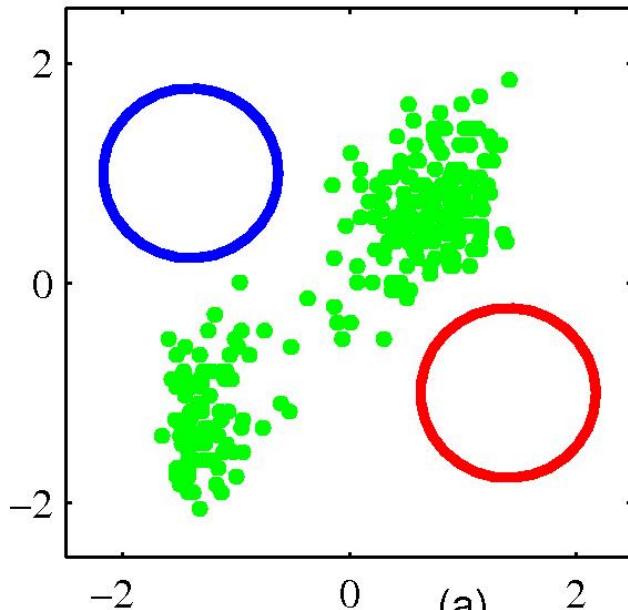
■ M-step:

$$\mu_i = \frac{1}{N_i} \sum_{t=1}^N \gamma(z_i^t) x_i, \quad N_i = \sum_{t=1}^N \gamma(z_i^t)$$

$$\pi_i = \frac{N_i}{N}$$

$$\Sigma_i = \frac{1}{N_i} \sum_{t=1}^N \gamma(z_i^t) (x^t - \mu_i)(x^t - \mu_i)^T$$

EM on Gaussian Mixtures Example



The Role of Hidden Variables

- Consider coupling a multinomial variable with Gaussian. $z_i = 1$ if \mathbf{x} belongs to G_i , 0 otherwise;

$$p(z) = \prod_{i=1}^K \pi_i^{z_i} \quad p(x|z) = \prod_{i=1:K} \mathcal{N}(\mathbf{x} | \mu_i, \Sigma_i)^{z_i}$$

- Joint distribution

$$p(x, z) = p(z) * p(x | z) = \prod_{i=1}^K (\pi_i \mathcal{N}(\mathbf{x} | \mu_i, \Sigma_i))^{z_i}$$

- Work with complete likelihood on $p(x, z)$.

Complete Likelihood

- Complete likelihood, $\mathcal{L}_c(\Phi | X, Z)$, in terms of x and z

$$\begin{aligned}\mathcal{L}_c(\Phi | \mathcal{X}) &= \log \prod_t p(x^t, z^t | \Phi) = \sum_t \log p(x^t, z^t | \Phi) \\ &= \sum_t \log p(z^t | \Phi) + \log p(x^t | z^t, \Phi) \\ &= \sum_t \sum_i z_i^t [\log \pi_i + \log p(x^t | \Phi_i)]\end{aligned}$$

- Work with the expectation of $\mathcal{L}_c(\Phi | X)$

$$\begin{aligned}\mathcal{Q}(\Phi | \Phi^l) &= E[\mathcal{L}_c(\Phi | X, Z) | \chi, \Phi^l] \\ &= \sum_t \sum_i E[z_i^t | \chi, \Phi^l] [\log \pi_i + \log p(x^t | \Phi_i^l)]\end{aligned}$$

Expectation (EM)

■ E-Step: Expectation

$$\begin{aligned} Q(\Phi | \Phi^l) &= E\left[\mathcal{L}_c(\Phi | \mathcal{X}, \mathcal{Z}) | \chi, \Phi^l\right] \\ &= \sum_t \sum_i E\left[z_i^t | \chi, \Phi^l\right] [\log \pi_i + \log p(x^t | \Phi_i^l)] \end{aligned}$$

$$\begin{aligned} E\left[z_i^t | \chi, \Phi^l\right] &= E\left[z_i^t | x^t, \Phi^l\right] = p(z_i^t = 1 | x^t, \Phi^l) \\ &= \frac{p(z_i^t = 1 | \Phi^l) p(x^t | z_i^t = 1, \Phi^l)}{p(x^t | \Phi^l)} \\ &= \frac{\pi_i p(x^t | \Phi_i^l)}{\sum_j \pi_j p(x^t | \Phi_j^l)} = P(z_i^t = 1 | x^t, \Phi^l) \equiv \gamma(z_i^t) \end{aligned}$$

Maximization (EM)

- M-Step: Maximization

$$\Phi^{l+1} = \operatorname{argmax}_{\Phi} Q(\Phi | \Phi^l)$$

$$Q(\Phi | \Phi^l) = \sum_t \sum_i \gamma(z_i^t) [\log \pi_i + \log p(x^t | \Phi_i^l)]$$

- Maximum likelihood learning, e.g. gaussian

$$\pi_i = \frac{N_i}{N} \quad \mu_i = \frac{1}{N_i} \sum_{t=1}^N \gamma(z_i^t) x_i, \quad N_i = \sum_{t=1}^N \gamma(z_i^t)$$

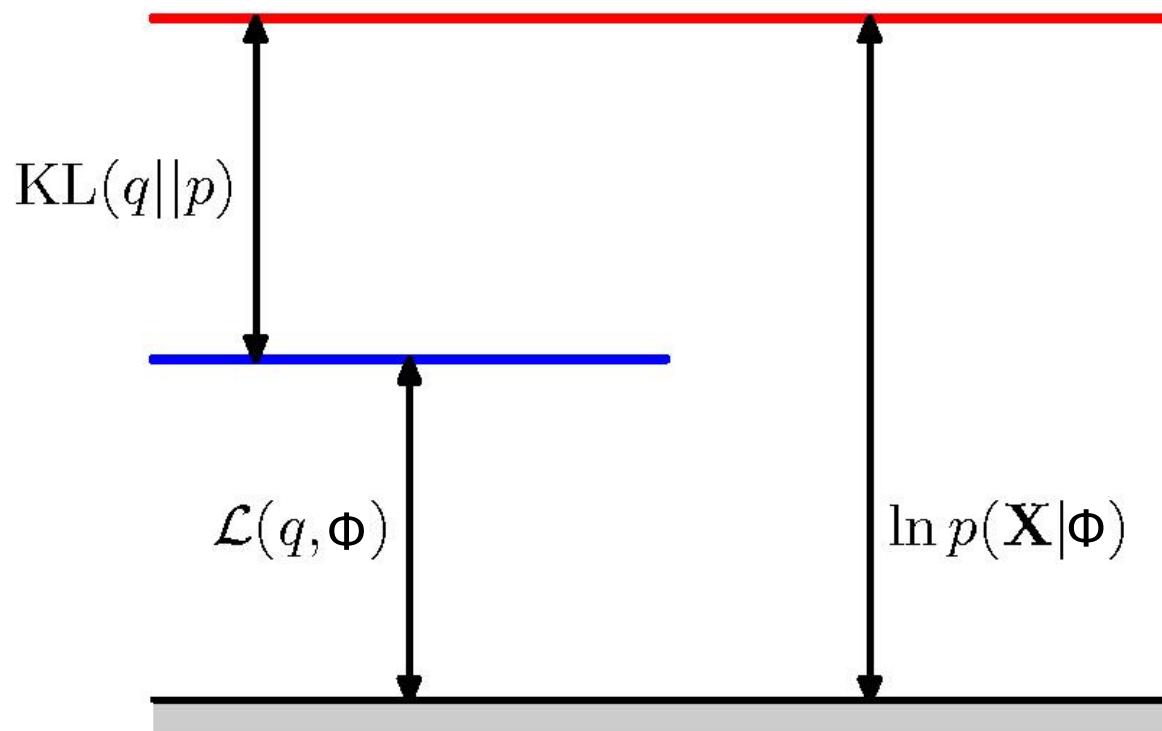
$$\Sigma_i = \frac{1}{N_i} \sum_{t=1}^N \gamma(z_i^t) (x^t - \mu_i)(x^t - \mu_i)^T$$

EM summary

- The likelihood function of mixture of Gaussians is intractable (log sum problem).
- The complete likelihood function is a product of Gaussians selected by z_i (log product).
- EM aims to maximize the expectation (over z) of the complete likelihood wrt the parameters of mixture of Gaussians.
- The expectation of the complete likelihood relies on the expectation of each z_i , which is re-estimated in each iteration.

Why EM Converges?

$$\begin{aligned}\ln p(X | \Phi) &= \sum_Z q(Z) \ln p(X | \Phi) = \sum_Z q(Z) \ln \frac{p(X, Z | \Phi)}{p(Z | X, \Phi)} \\ &= \sum_Z q(Z) \ln \frac{p(X, Z | \Phi)}{q(Z)} - \sum_Z q(Z) \ln \frac{p(Z | X, \Phi)}{q(Z)} \\ &= \mathcal{L}(q, \Phi) + KL(q \| p)\end{aligned}$$

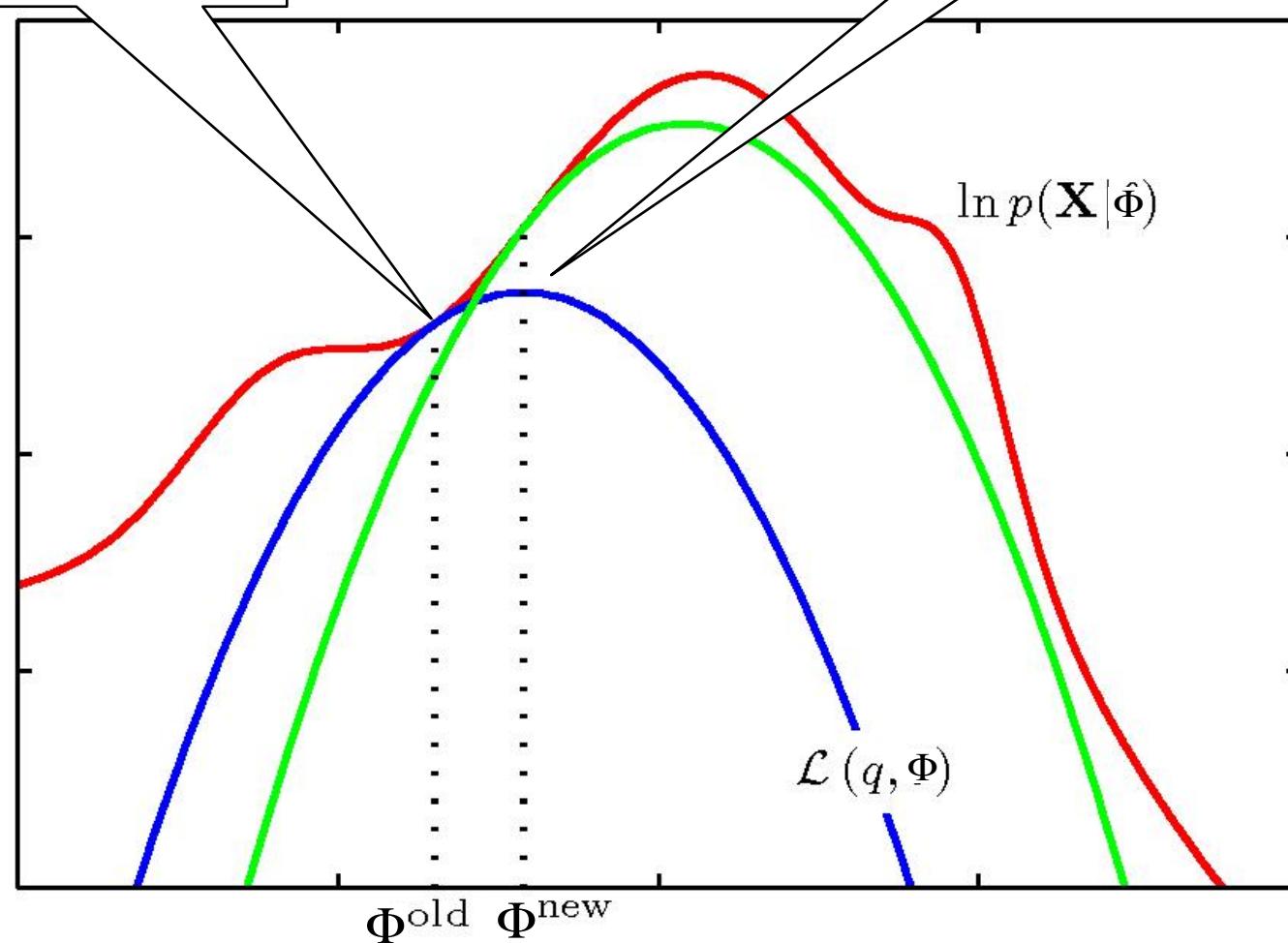


Since $KL(q \| p) \geq 0$,
 $\ln p(X | \Phi) \geq \mathcal{L}(q, \Phi)$.

Analysis of EM

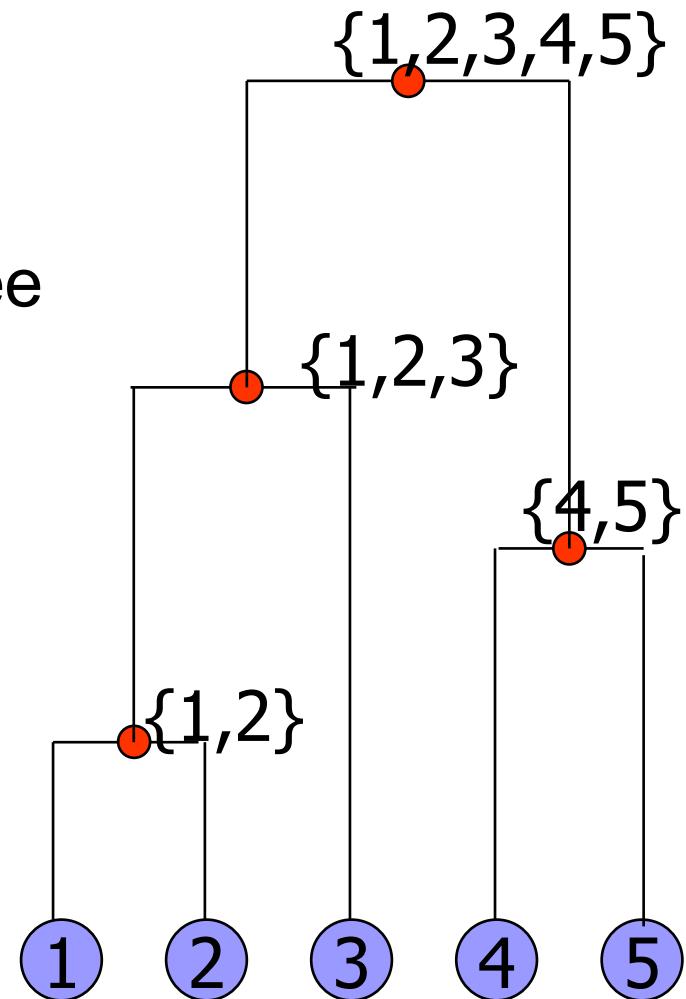
When $KL(p//q)=0$, \mathcal{L} touches $\ln p(\mathbf{X}|\Phi)$.

Each new Φ maximize \mathcal{L} .



Hierarchical Clustering

- Organize the samples in a structure of a hierarchical tree
- Agglomerative (bottom up)
- Divisive (top down)



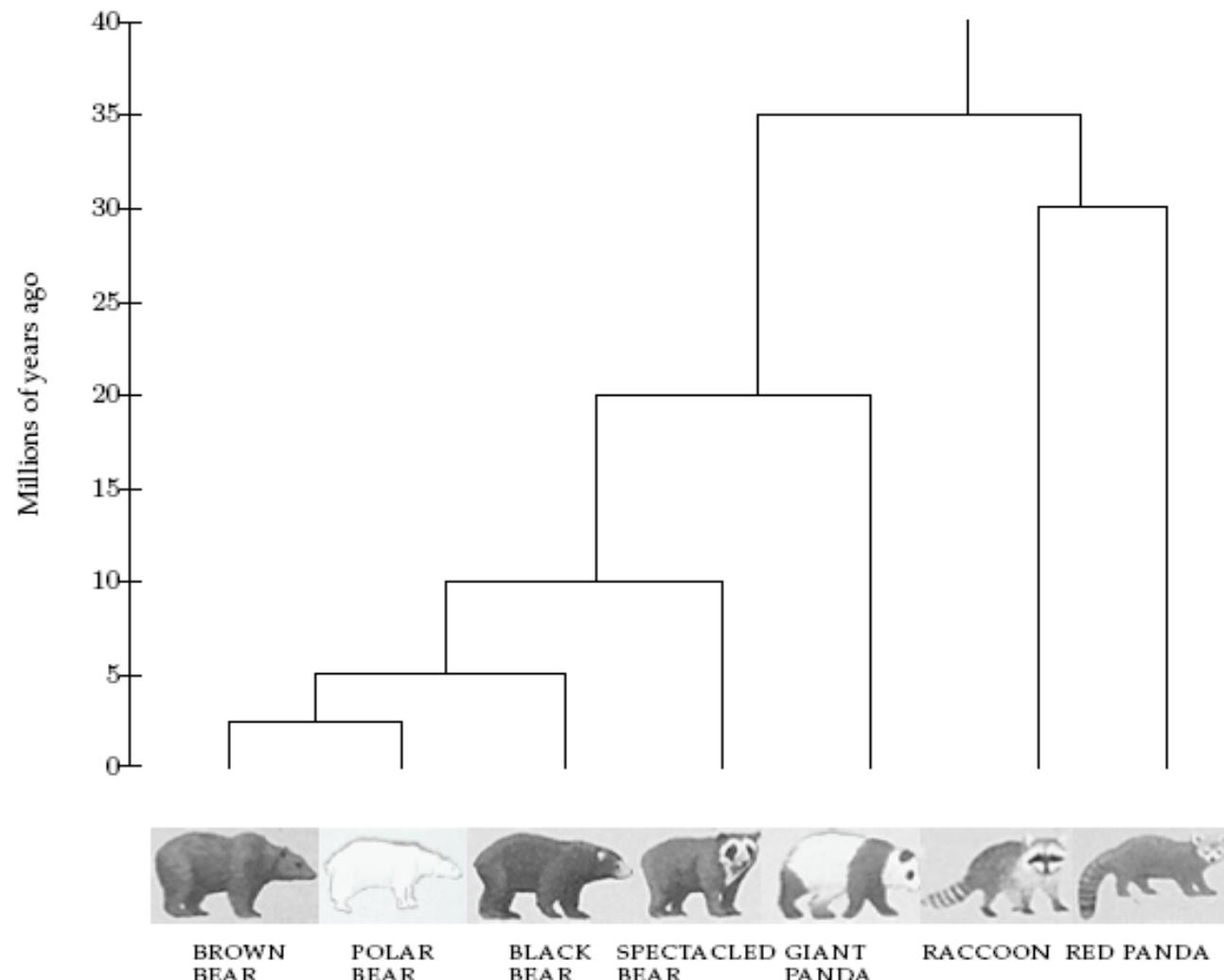
Evolution and DNA Analysis: the Giant Panda Riddle

- For roughly 100 years scientists were unable to figure out which family the giant panda belongs to



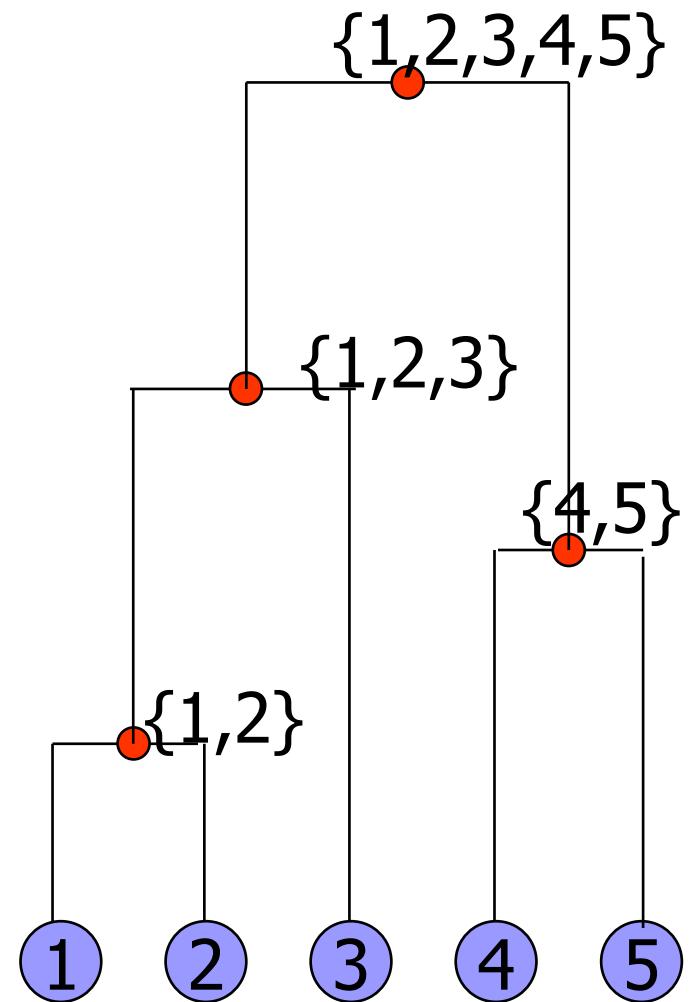
- Giant pandas look like bears but have features that are unusual for bears and typical for raccoons, e.g., they do not hibernate
- In 1985, Steven O'Brien and colleagues solved the giant panda classification problem using DNA sequences and algorithms

Phylogenetic tree of bears and raccoons



Hierarchical Clustering: Agglomerative

- Initial step: each sample is regarded as a cluster with one item
- Find the 2 most similar clusters and merge them into a common node
- The length of the branch is proportional to the distance
- Iterate on merging nodes until all samples are contained in one cluster- the root of the tree.



Hierarchical Clustering

- Cluster based on similarities/distances
- Distance measure between instances \mathbf{x}^r and \mathbf{x}^s

Minkowski (L_p) (Euclidean for $p = 2$)

$$d_m(\mathbf{x}^r, \mathbf{x}^s) = \left[\sum_{j=1}^d (x_j^r - x_j^s)^p \right]^{1/p}$$

City-block distance

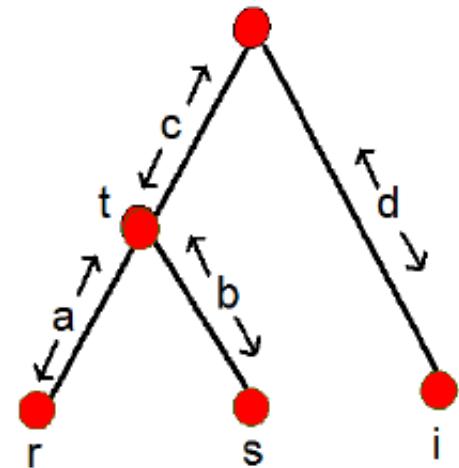
$$d_{cb}(\mathbf{x}^r, \mathbf{x}^s) = \sum_{j=1}^d |x_j^r - x_j^s|$$

Algorithm

- Input: The distance matrix D
- Find the pair r, s with the least distance
- Merge clusters r, s .
- Delete elements r, s and add a new element t with

$$D_{it} = D_{ti} = \frac{D_{ir} + D_{is} - D_{rs}}{2}$$

- Repeat until one element is left

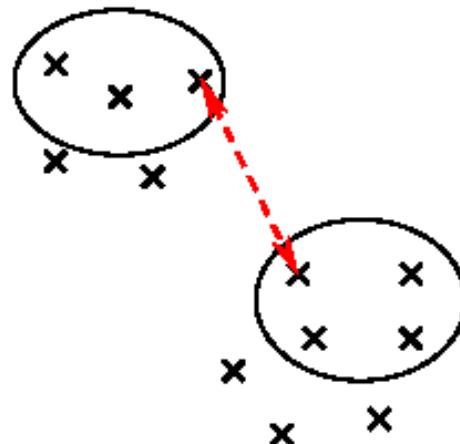


$$D_{it} = c + d = \frac{(c + d + a) + (c + d + b) - (a + b)}{2} = \frac{D_{ir} + D_{is} - D_{rs}}{2}$$

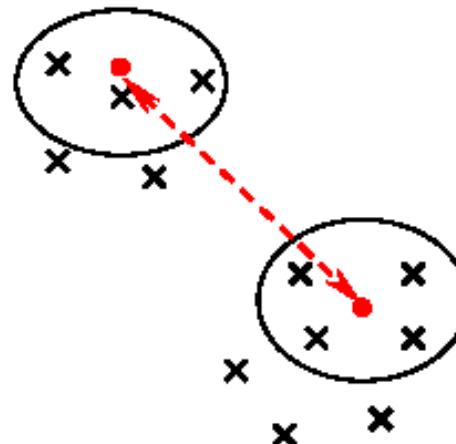
Algorithm

- Average linkage: distance between the centroids
- Single-link: minimum distance (long and thin)
- Complete-link: maximum distance (tight)

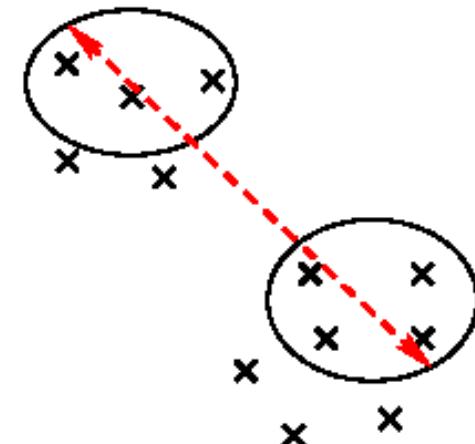
Single-linkage



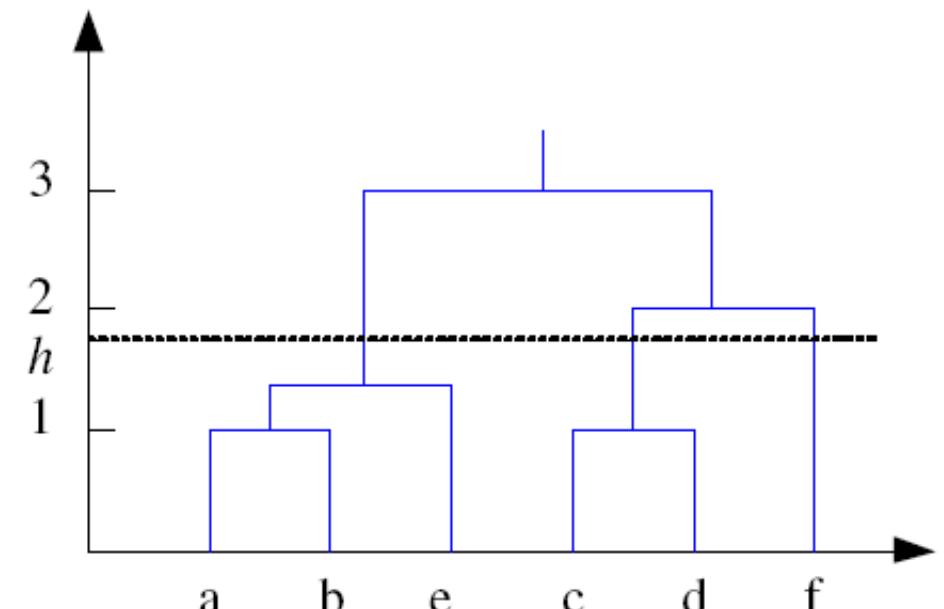
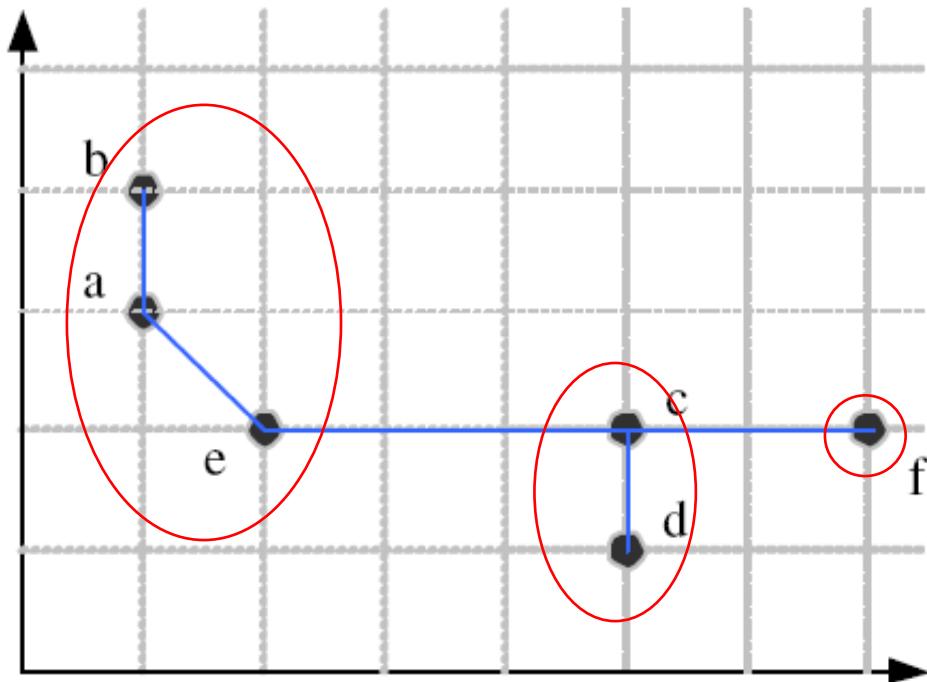
Average-linkage



Complete-linkage



Example: Single-Link Clustering



Dendrogram