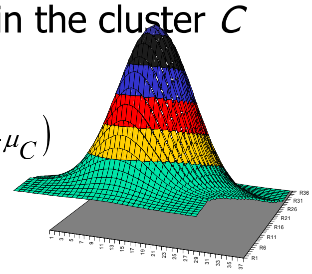
Expectation Maximization (EM)

Basic Notions [Dempster, Laird & Rubin 1977]

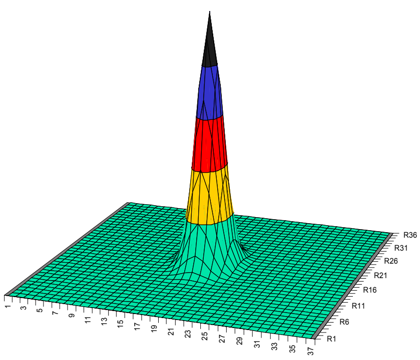
- Consider points $p = (x_1, \ldots, x_d)$ from a d-dimensional Euclidean vector space
- Clusters are represented by probability distributions
- Typically: mixture of Gaussian distributions
- Representation of a cluster C
  - Center point μ_C of all points in the cluster
  - d x d Covariance matrix Σ_C for the points in the cluster C
- Density function for cluster C:

$$P(x \mid C) = \frac{1}{\sqrt{(2\pi)^d \left|\sum C\right|}} \cdot e^{\frac{1}{2}\cdot(x-\mu_C)^T \cdot (\Sigma_C)^{-1} \cdot (x-\mu_C)}$$

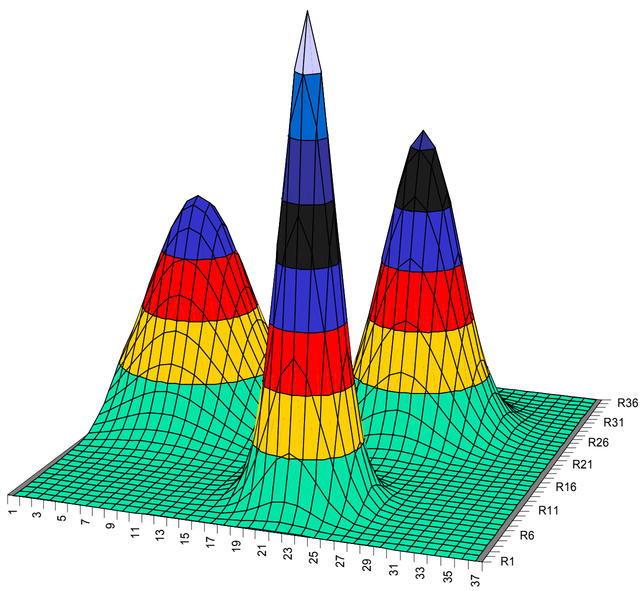


EM: Gaussian Mixture – 2D examples

A single Gaussian density function



A Gaussian Mixture Model, $k = 3$



EM – Basic Notions

- Density function for clustering $M = \{C_1, \ldots, C_k\}$
  - Let W_i denote the fraction of cluster C_i in the entire data set D

$$P(x) = \sum_{i=1}^{k} W_i \cdot P(x \mid C_i)$$

- Assignment of points to clusters
  - A point may belong to several clusters with different probabilities

$$P(C_i \mid x) = W_i \cdot \frac{P(x \mid C_i)}{P(x)}$$

- Maximize *E(M),* a measure for the quality of a clustering *M*
  - *E(M)* indicates the probability that the data D have been generated by following the distribution model M

$$E(M) = \sum_{x \in D} \log(P(x))$$

EM – Algorithm

ClusteringbyExpectationMaximization (point set *D*, int *k*)

Generate an initial model M' = (C_1', …, C_k')

repeat

// (re-) assign points to clusters

Compute P(x|C_i), P(x) and P(C_i|x) for each object x from *D* and each cluster (= Gaussian) C_i

// (re-) compute the models

Compute a new model M = {C_1, …, C_k} by recomputing W_i, μ_C and Σ_C for each Cluster C_i

Replace M' by M

until |E(M) – E(M')| < ε

return M

EM – Recomputation of Parameters

- Weight W_i of cluster C_i $$W_i = \frac{1}{n} \sum_{x \in D} P(C_i \mid x)$$
- Center μ_i of cluster C_i $$\mu_i = \frac{\sum_{x \in D} x \cdot P(C_i \mid x)}{\sum_{x \in D} P(C_i \mid x)}$$
- Shape matrix Σ_i of cluster C_i $$\Sigma_i = \frac{\sum_{x \in D} P(C_i \mid x)(x - \mu_i)^2}{\sum_{x \in D} P(C_i \mid x)}$$

EM – Discussion

- Convergence to (possibly local) minimum
- Computational effort:
  - O(n * |M| * #iterations)
  - #iterations is quite high in many cases
- Both result and runtime strongly depend on …
  - … the initial assignment
  - … a proper choice of parameter k (= desired number of clusters)
- Modification to obtain a really *partitioning* variant
  - Objects may belong to several clusters
  - Assign each object to the cluster to which it belongs with the highest probability

Clustering (II) – Summary

- Advanced Clustering Topics
  - Incremental Clustering, Generalized DBSCAN, Outlier Detection
- Scaling Up Clustering Algorithms
  - BIRCH, Data Bubbles, Index-based Clustering, GRID Clustering
- Sets of Similarity Queries (Similarity Join)
- Subspace Clustering
- Expectation Maximization: a statistical approach