Chapter 3
Density Estimation
Parametric Density Estimation

Statistical Methods and Learning in Computer Vision
WS 2012/2013

Claudia Nieuwenhuis
Lehrstuhl für Computer Vision and Pattern Recognition
Fakultät für Informatik
Technische Universität München

Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

Overview

1. Parametric Density Estimation
2. The EM-Algorithm
3. Problems and Solutions

Overview

1. **Parametric Density Estimation**
2. The EM-Algorithm
3. Problems and Solutions

Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

Parametric Densities

In case of high dimensional sample data non-parametric methods often fail. In these cases a parametric density estimator can help. These estimators predefine the basic class of density and only estimate specific parameters from the sample data. This makes the estimation problem simpler. However, the result is restricted to the chosen family of distributions.

Parametric Density Estimation

Parametric density estimators estimate probability density functions for which the basic class is already defined and only the parameters θ are to be estimated, e.g. the multivariate normal distribution class, which depends on the mean vector μ and the covariance matrix Σ.

$$\begin{aligned} \mathcal{N}(x|\theta) &= \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-0.5(x-\mu)^T\Sigma^{-1}(x-\mu)\right), \\ \theta &= \mu, \Sigma \end{aligned}$$

Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

The EM-Algorithm for computing Mixtures of Gaussians

Generative Model

Assume the following generative model, which generates samples $x_n, n = 1, .., N$.
There are K components. Each component generates data from a Gaussian with given parameters $\mathcal{N}(\mu_k, \Sigma_k)$.
Each data sample is generated in the following way:

- Randomly select a cluster from which to draw the sample x with probability $P(C = k) = \pi_k$.
- Draw the data sample from the normal distribution with density $f(x|C = k) = \mathcal{N}(x|\mu_k, \Sigma_k)$

So we arrive at the generative model described by the probability density function

$$f(x) = \sum_{k=1}^{K} f(C = k) f(x|C = k)$$

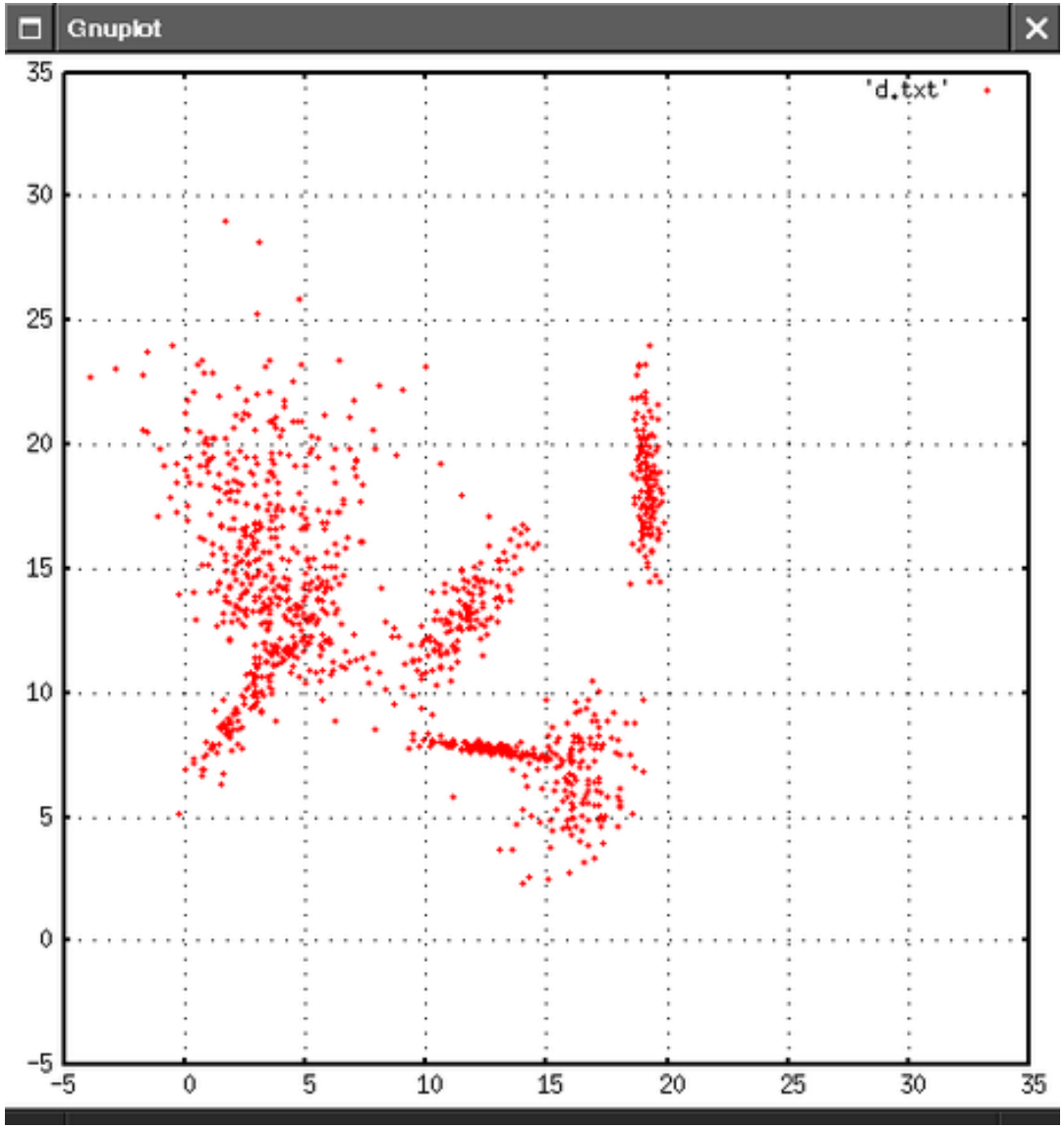
$$f(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k), \quad 0 \leq \pi_k \leq 1, \sum_{k=1}^{K} \pi_k = 1$$

Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

Dataset for EM-algorithm

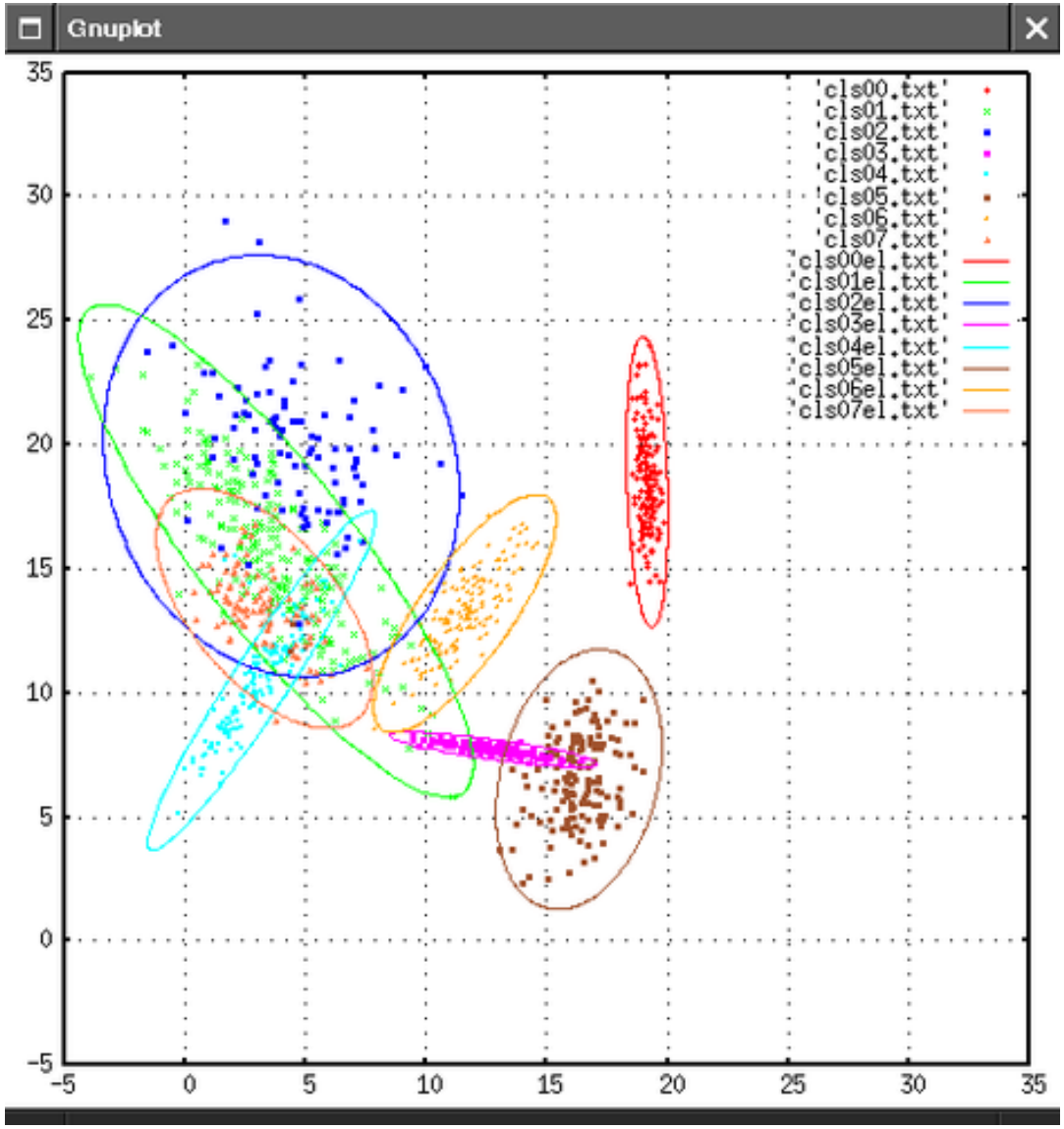


Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

Result of the EM-algorithm

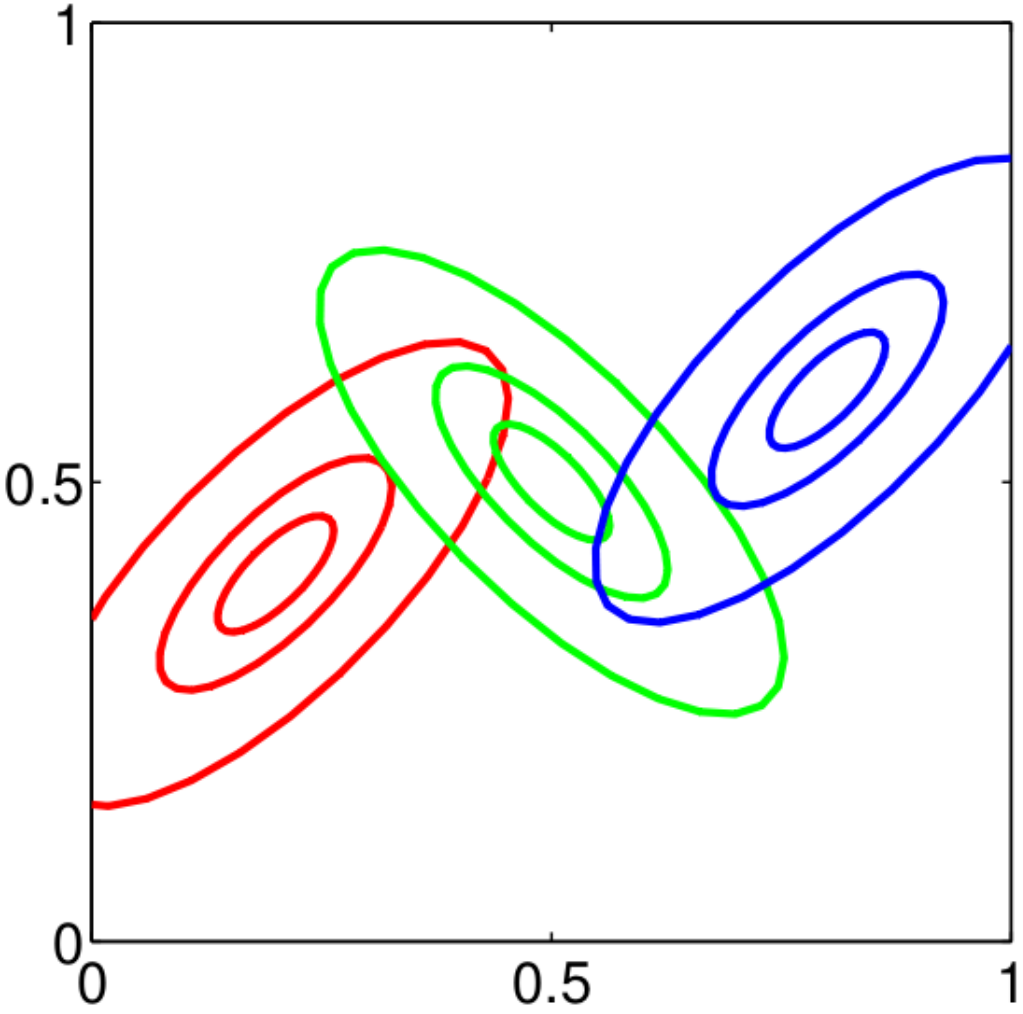


Parametric Density Estimation

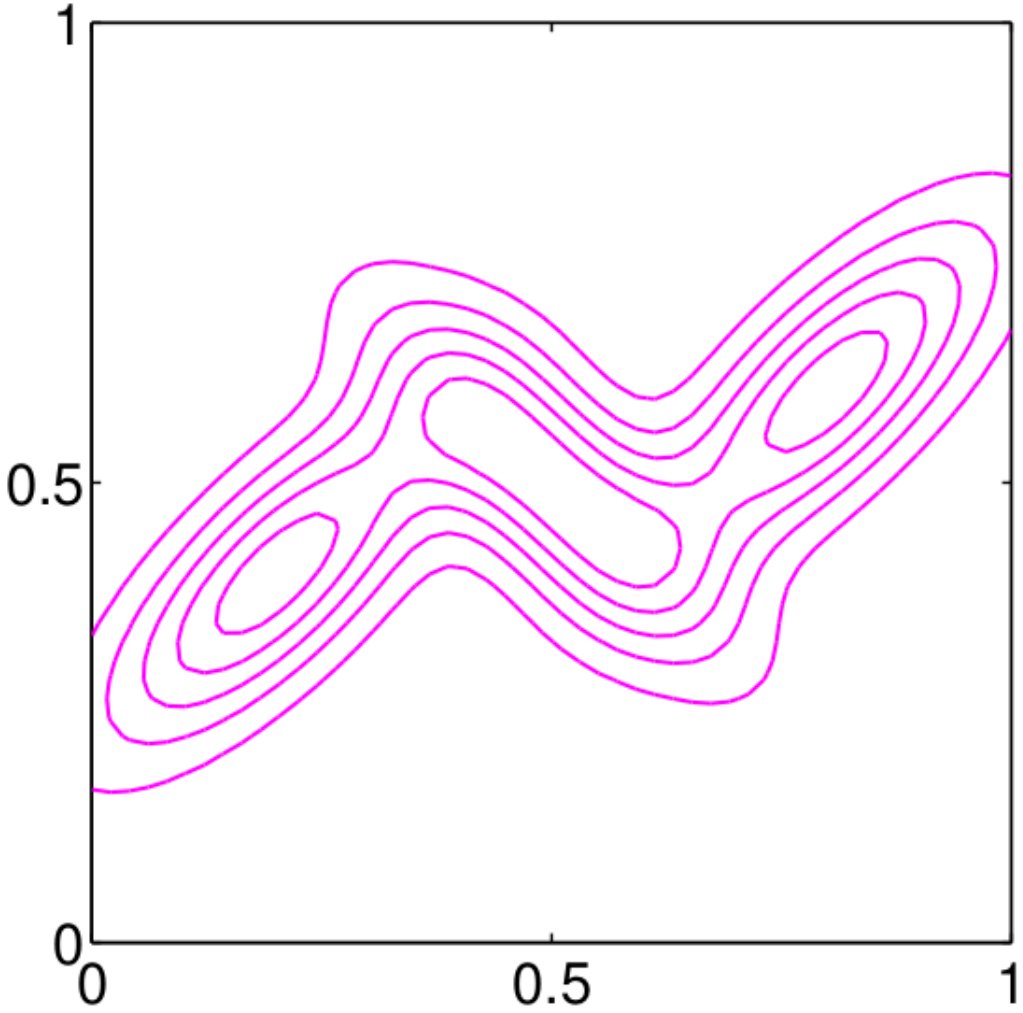
The EM-Algorithm

Problems and Solutions

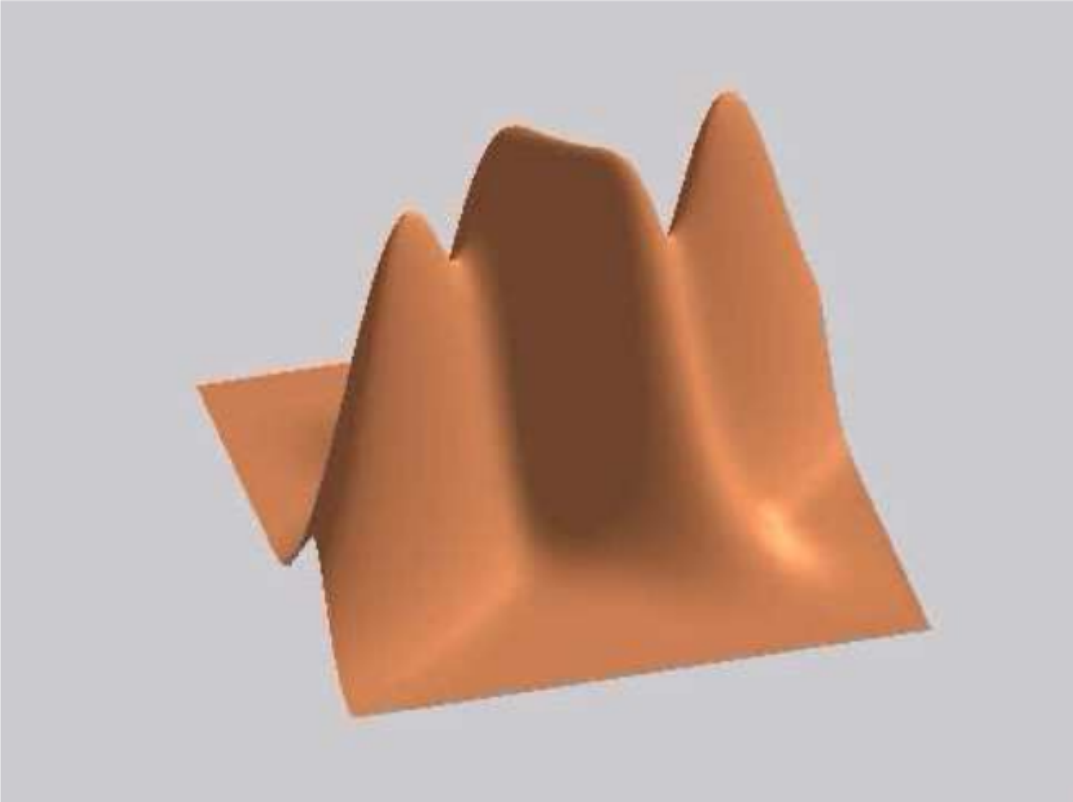
Gaussian Mixtures



Gaussian Mixtures



Gaussian Mixtures



Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

Maximum Likelihood Approach

Our goal is to find the parameter vector $\theta = (\mu, \Sigma)$ specifying the model from which the samples x_n have most likely been drawn. The events are independent, so we want to find the parameters which maximize the likelihood function

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\theta_k).$$

To simplify optimization we use the log-likelihood function

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\theta_k).$$

Note: there is no closed-form maximum-likelihood solution for this model.

The Maximum Likelihood Estimate for μ

For d-dimensional data points we have...

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k), \quad \pi_k \geq 0, \sum_{k=1}^{K} \pi_k = 1$$

$$\mathcal{N}(x|\theta) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Setting the derivative with respect to μ_i to 0 we obtain

$$\frac{\partial \mathcal{N}(x|\theta)}{\partial \mu_i} = \sum_{n=1}^{N} \underbrace{\frac{\pi_i \mathcal{N}(x_n|\mu_i, \Sigma_i)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}}_{\gamma_{ni}} \Sigma_i^{-1}(x_n - \mu_i) = 0$$

$$0 = \sum_{n=1}^{N} \gamma_{ni}(x_n - \mu_i)$$

$$\mu_i = \frac{\sum_{n=1}^{N} \gamma_{ni} x_n}{\sum_{n=1}^{N} \gamma_{ni}}$$

Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

The Maximum Likelihood Estimate for Σ

For d-dimensional data points we have...

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k), \quad \pi_k \geq 0, \sum_{k=1}^{K} \pi_k = 1$$

$$\mathcal{N}(x|\theta) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

Setting the derivative with respect to Σ_i to 0 we obtain

$$\Sigma_i = \frac{\sum_{n=1}^{N} \gamma_{ni} (x_n - \mu_i)(x_n - \mu_i)^T}{\sum_{n=1}^{N} \gamma_{ni}}$$

Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

The Maximum Likelihood Estimate for π

For d-dimensional data points we have...

To integrate the conditions on π we use Lagrange Multipliers

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) + \lambda(1 - \sum_{k=1}^{K} \pi_k)$$

$$\mathcal{N}(x|\theta) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

Setting the derivative with respect to π_i to 0 we obtain

$$\sum_{n=1}^{N} \frac{\mathcal{N}(x_n|\mu_i, \Sigma_i)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)} - \lambda = 0$$

$$\sum_{n=1}^{N} \frac{\sum_{i=1}^{K} \pi_i \mathcal{N}(x_n|\mu_i, \Sigma_i)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)} = \sum_{i=1}^{K} \pi_i \lambda$$

$$\lambda = N$$

$$\pi_i = \frac{\sum_{n=1}^{N} \gamma_{ni}}{N}$$

Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

The Expectation Maximization Algorithm

1. Initialize μ_i, Σ_i and the mixing coefficients π_i.
2. **E-Step:** Compute the responsibilities γ_{ni}

$$\gamma_{ni} = \frac{\pi_i \mathcal{N}(x_n|\mu_i, \Sigma_i)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}$$

3. **M-Step:** Reestimate the parameters

$$\begin{aligned}
\mu_i^{new} &= \frac{\sum_{n=1}^{N} \gamma_{ni} x_n}{\sum_{n=1}^{N} \gamma_{ni}} \\
\Sigma_i^{new} &= \frac{\sum_{n=1}^{N} \gamma_{ni} (x_n - \mu_i^{new})(x_n - \mu_i^{new})^T}{\sum_{n=1}^{N} \gamma_{ni}} \\
\pi_i^{new} &= \frac{\sum_{n=1}^{N} \gamma_{ni}}{N}
\end{aligned}$$

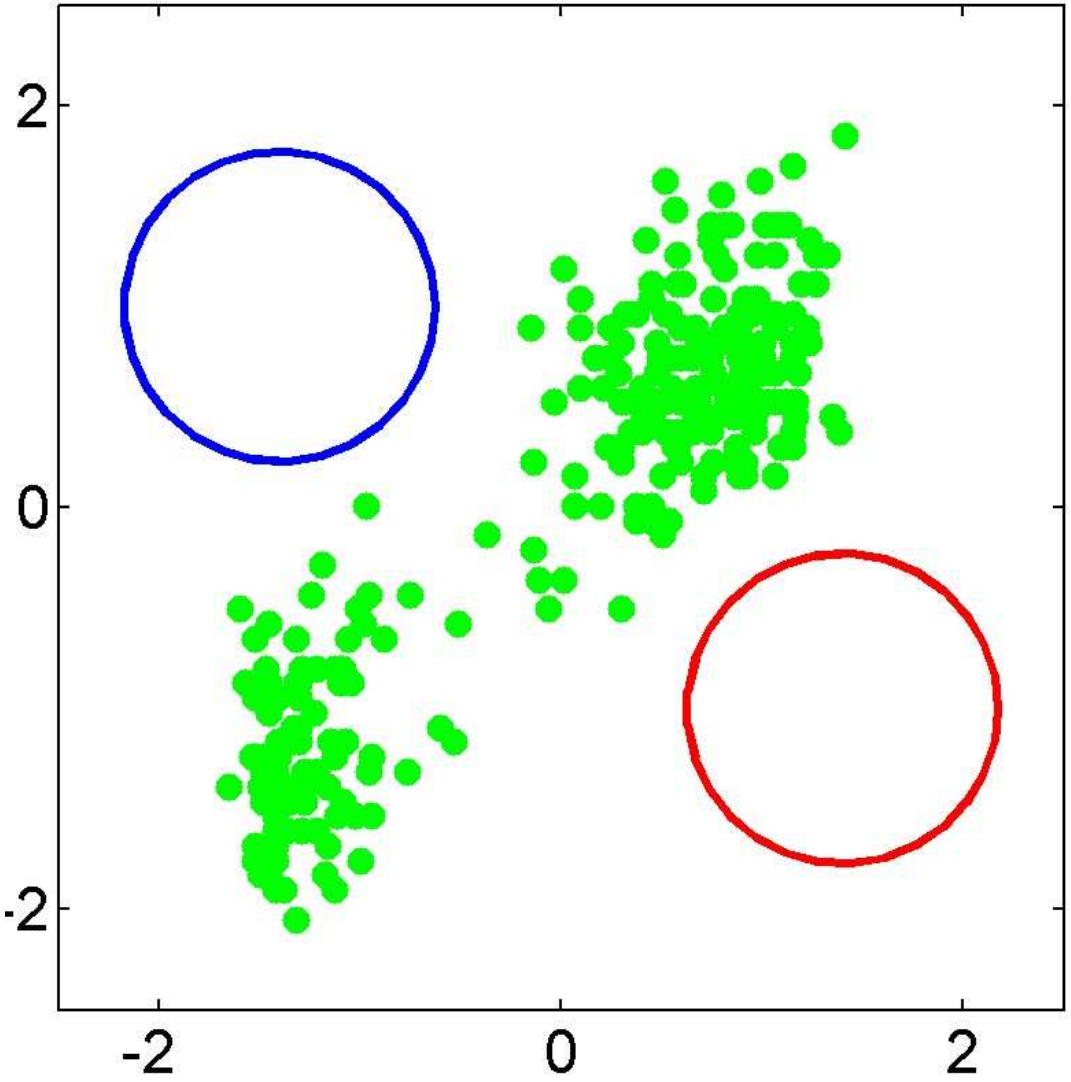
4. Check for convergence of the log-likelihood or the parameters. If not converged return to 2.

Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

The EM-Algorithm

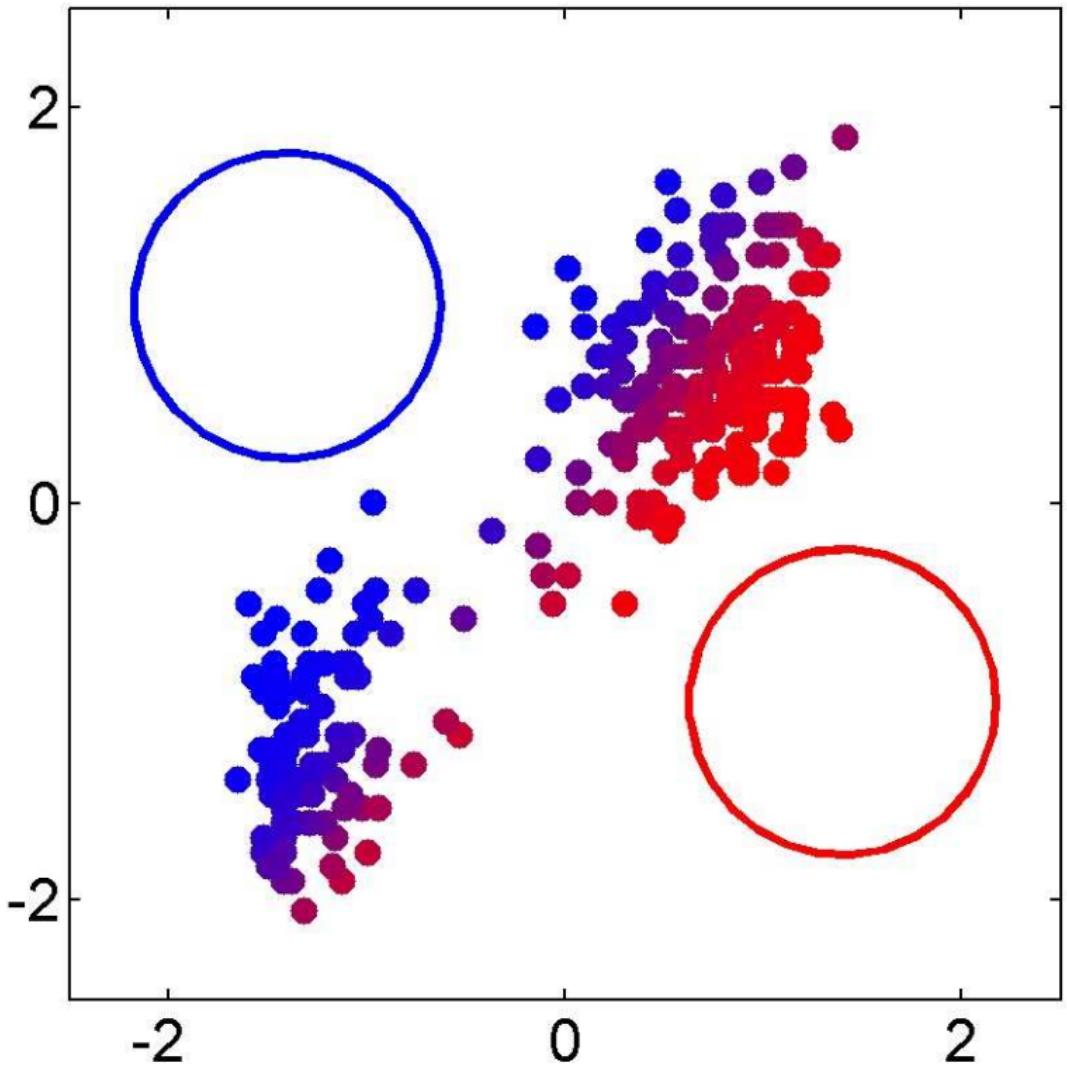


Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

The EM-Algorithm

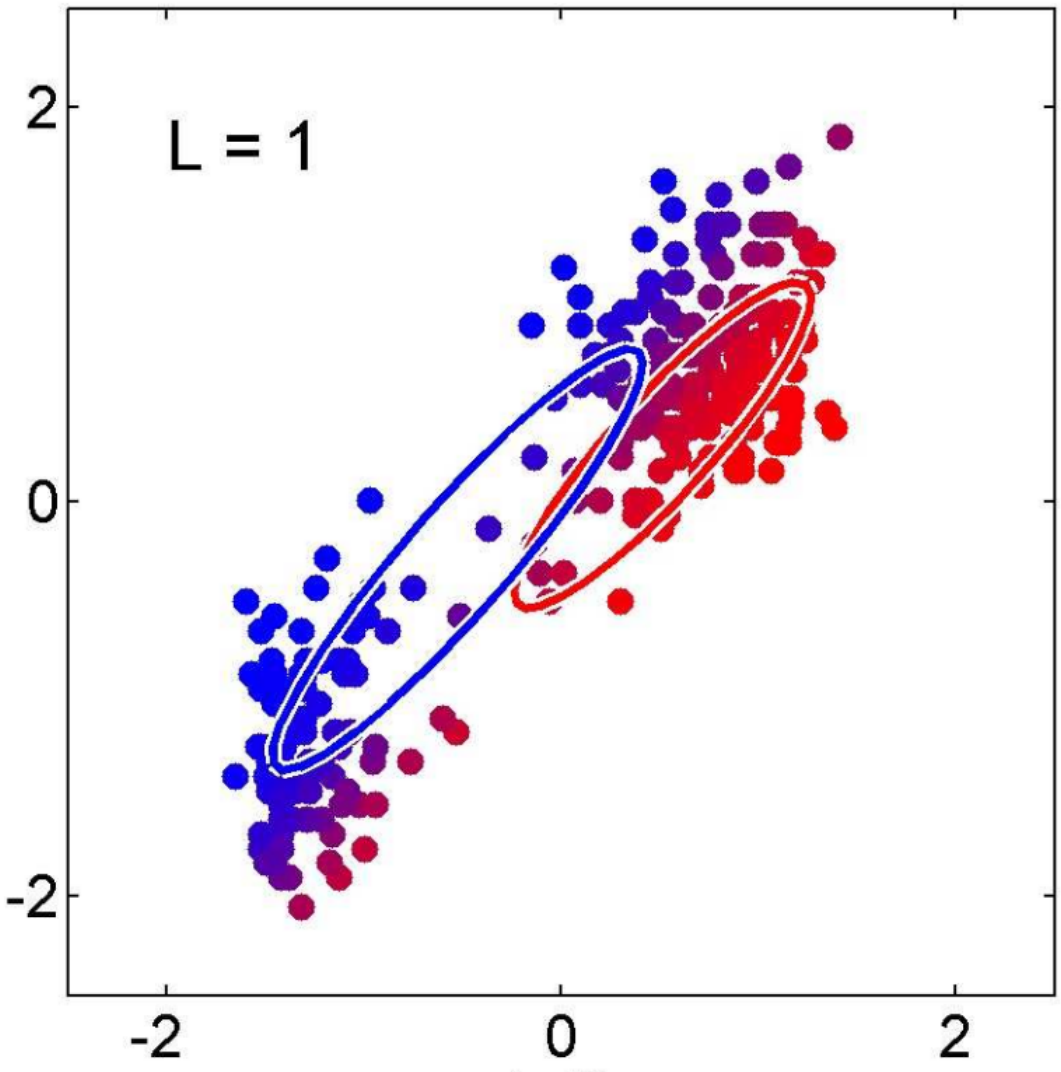


Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

The EM-Algorithm

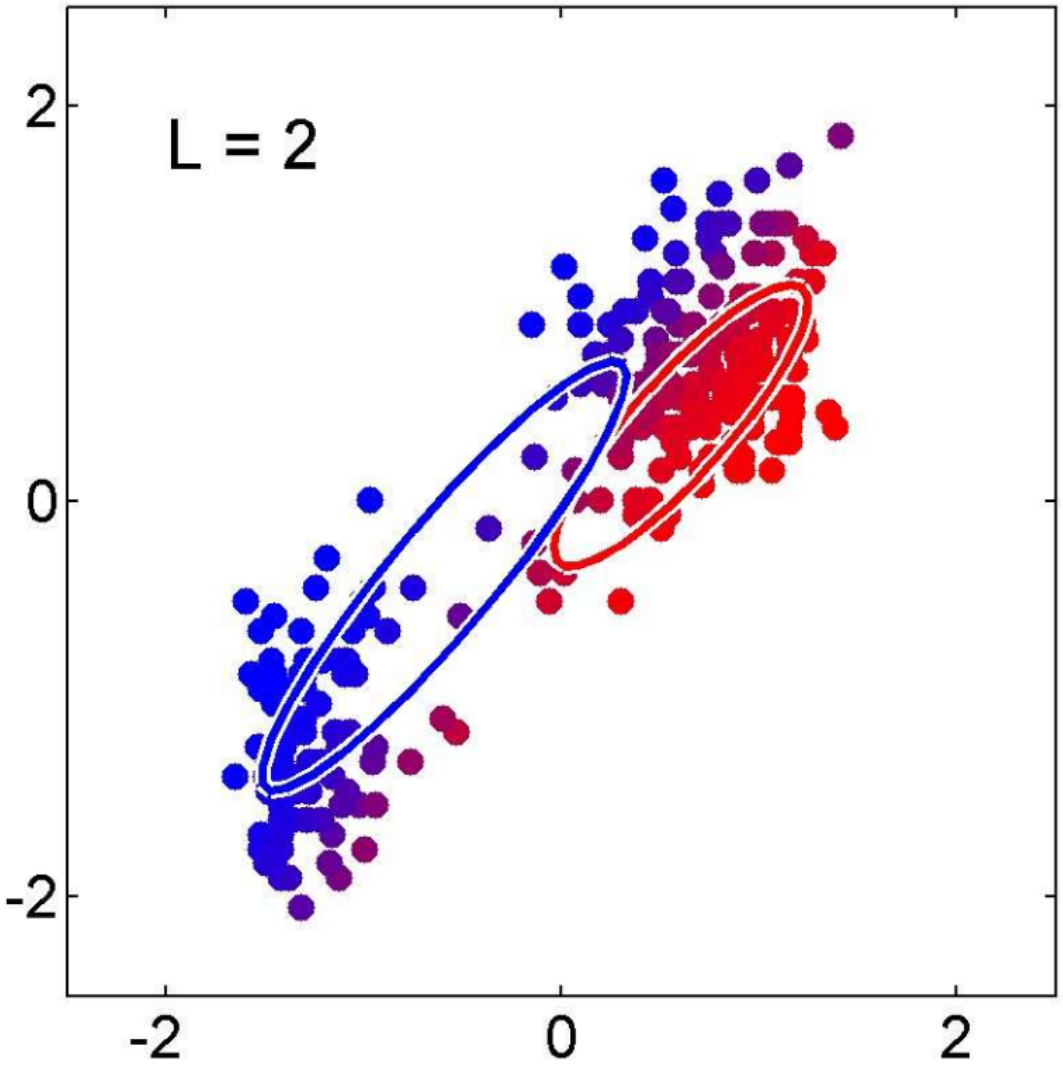


Parametric Density Estimation

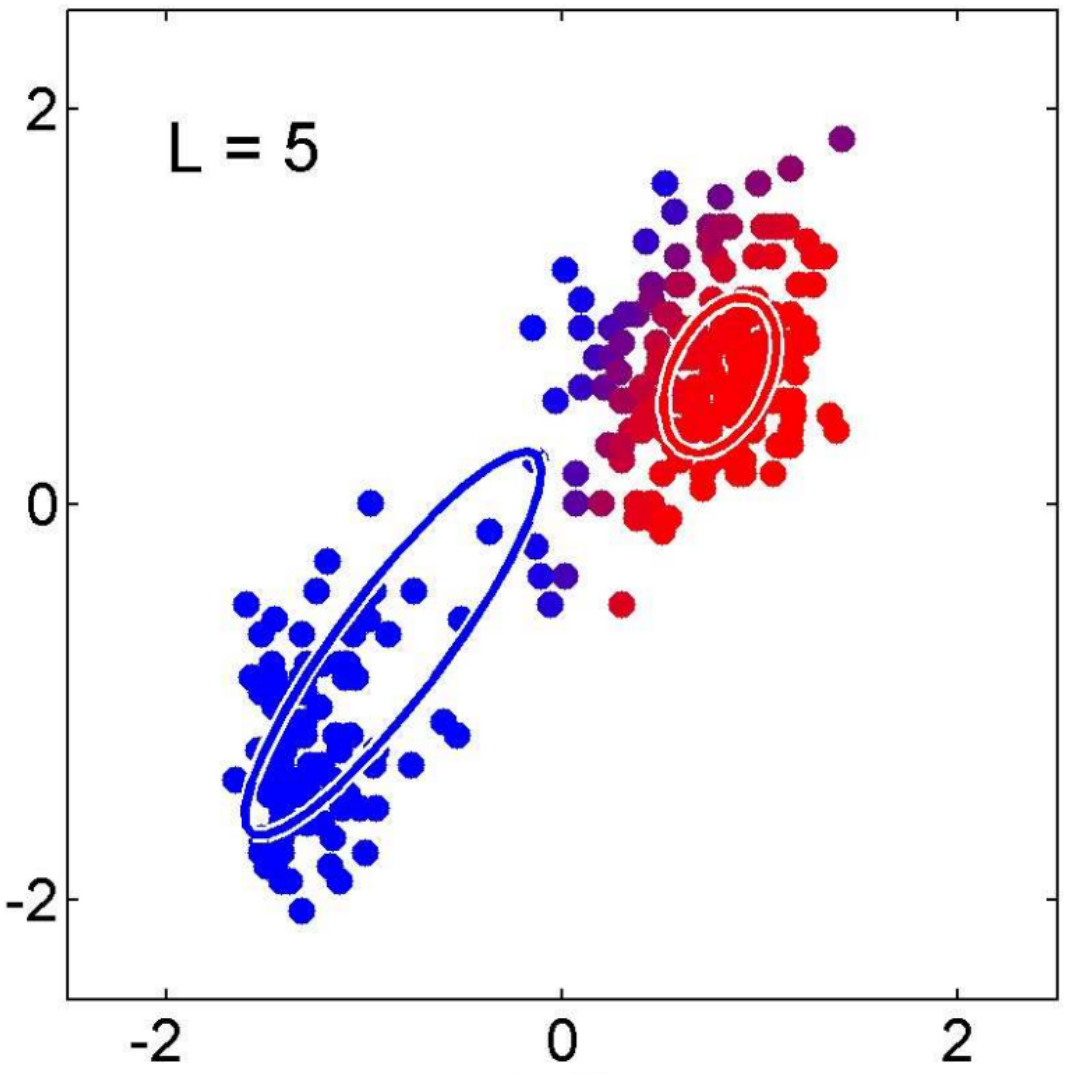
The EM-Algorithm

Problems and Solutions

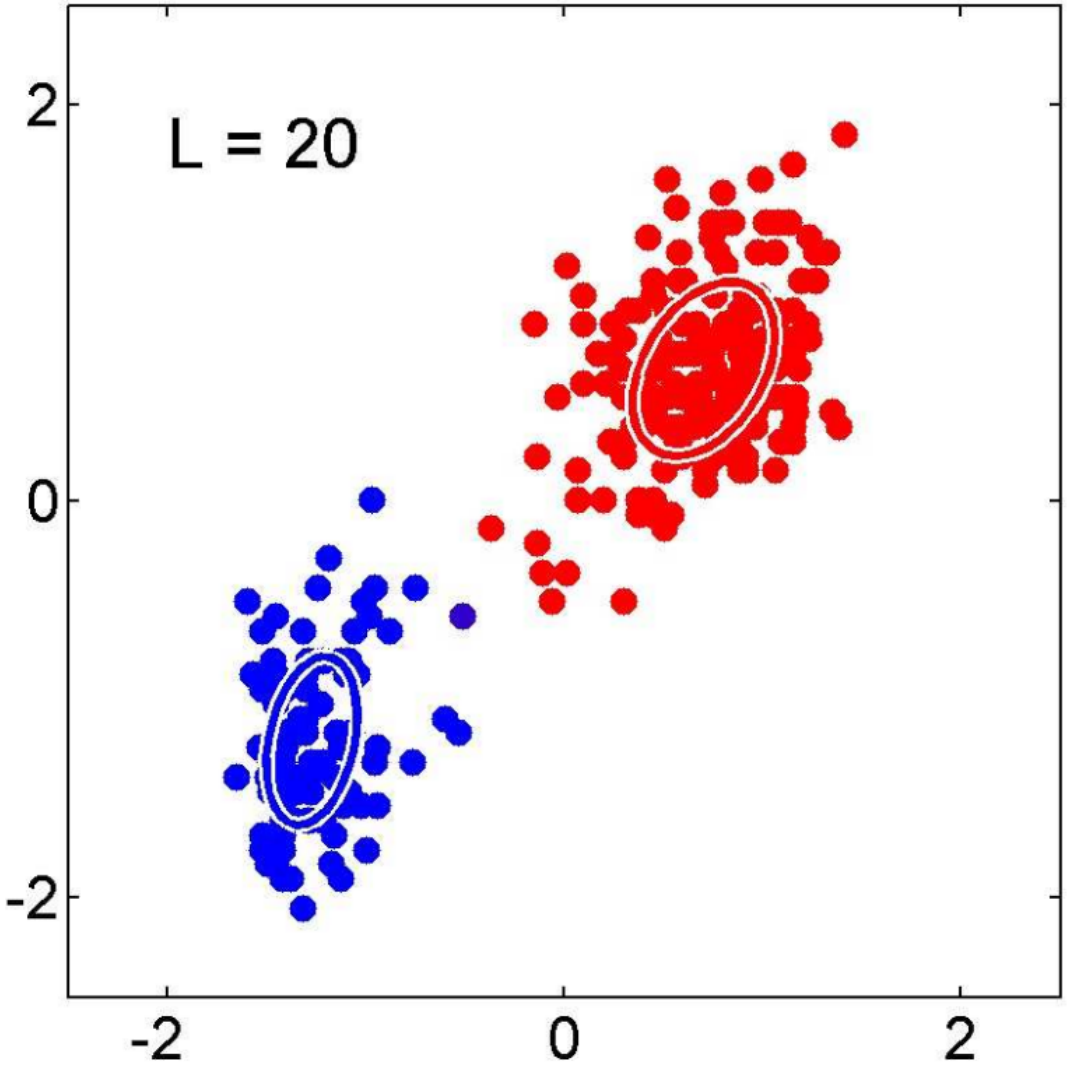
The EM-Algorithm



The EM-Algorithm



The EM-Algorithm



Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

Note:
The EM algorithm is not limited to Mixtures of Gaussians but can also be applied to other probability density functions.

Problems:

- The runtime can be large
- The algorithm does not necessarily yield global maxima
- Covariance matrices can become singular in case of data lying in lower dimensional subspaces
- How is the number of components K chosen?

Initialization: K-Means Clustering

Goal: identify clusters of data points in multidimensional space. Each point is assigned to the closest cluster indicated by μ_k. The optimization function is the distortion measure

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2,$$

where r_{nk} is a binary variable indicating if x_n is closest to cluster k

$$r_{nk} = \begin{cases} 1, & \text{if } k = \text{argmin}_j \|x_n - \mu_j\| \\ 0, & \text{otherwise} \end{cases}$$

Setting the derivative with respect to μ_k to 0 we obtain

$$\begin{aligned} \frac{\partial J}{\partial \mu_k} &= \sum_{n=1}^{N} r_{nk} 2(x_n - \mu_k) = 0 \\ \mu_k &= \frac{\sum_{n=1}^{N} r_{nk} x_n}{\sum_{n=1}^{N} r_{nk}} \end{aligned}$$

Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

K-Means Clustering

1. Initialize the distinct cluster positions μ_k
2. Compute the cluster assignment

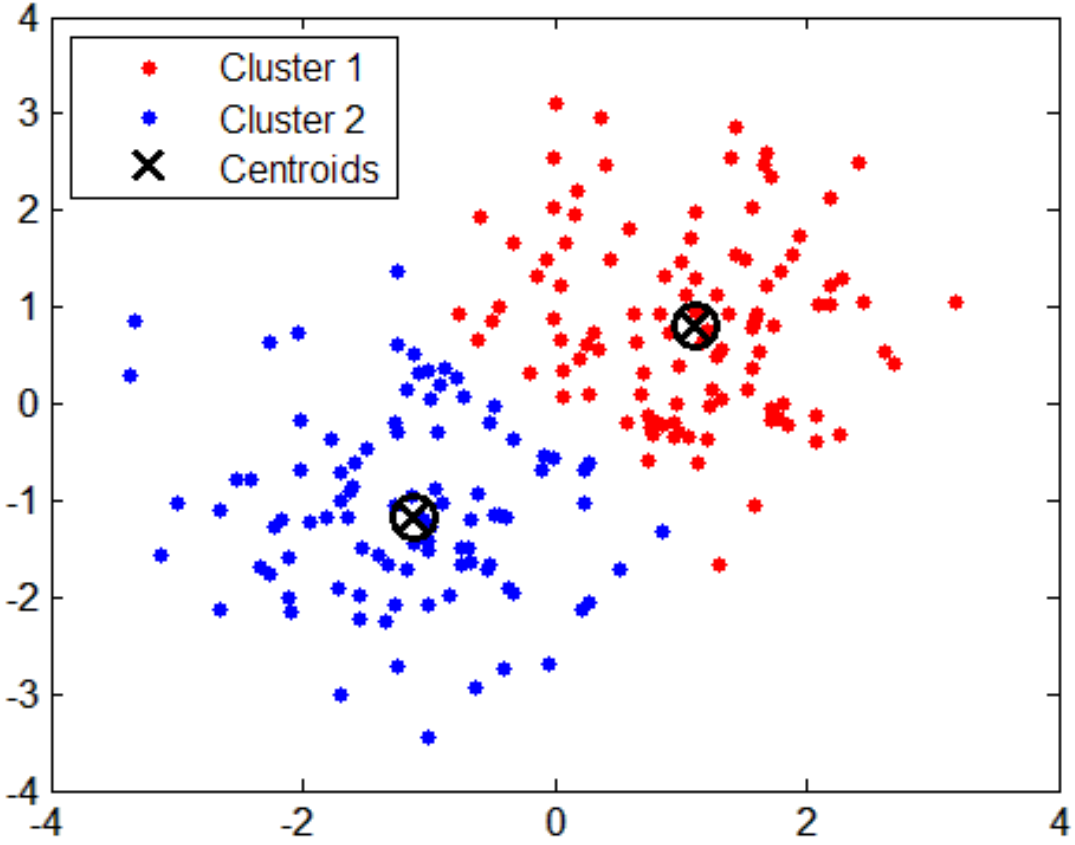
$$r_{nk} = \begin{cases} 1 & \text{if } k = \text{argmin}_j \|x_n - \mu_j\| \\ 0 & \text{otherwise} \end{cases}$$

3. Update the cluster positions

$$\mu_k^{new} = \frac{\sum_{n=1}^{N} r_{nk} x_n}{\sum_{n=1}^{N} r_{nk}}$$

4. Check for convergence of J or μ_k. If the algorithm has not converged go to 2.

The K-Means-Algorithm



Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

EM and K-Means Clustering

K-Means Clustering is a special case of the EM-Algorithm with binary responsiblities γ_{nk}. This corresponds to covariance matrices $\Sigma_k = \epsilon\, I$ for $\epsilon \to 0$.

$$\mathcal{N}(x_n|\mu_k, \Sigma_k) = \frac{1}{(2\pi\epsilon)^{\frac{d}{2}}} \exp\left\{-\frac{1}{2\epsilon}\|x_n - \mu_k\|^2\right\}$$

$$\begin{aligned}\gamma_{nk} &= \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} \\ &= \frac{\pi_k \exp\left\{-\frac{1}{2\epsilon}\|x_n - \mu_k\|^2\right\}}{\sum_{j=1}^{K} \pi_j \exp\left\{-\frac{1}{2\epsilon}\|x_n - \mu_j\|^2\right\}}\end{aligned}$$

For a given data point x_n and $\epsilon \to 0$ all terms go to zero. However, the term for which $\|x_n - \mu_j\|$ is smallest will go to zero most slowly. Hence, all responsibilities will go to zero except for γ_{nj}, which goes to 1. We, thus, obtain the binary cluster assignment r_{nj} of the K-Means algorithm.

Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

Parametric Density Estimation

The EM-Algorithm

Problems and Solutions

Runtime

Initializing the means of the EM-Algorithm with the cluster centers found by the K-Means algorithm the EM-algorithm converges more quickly.

Singular Covariance Matrices

The estimated covariance matrices can become singular if the data points lie on a lower dimensional subspace.

1. **Remedy 1:** in each step a constant matrix ϵI can be added to the covariance matrix.
2. **Remedy 2:** assume prior distributions for the parameters π, Σ and μ, which leads to MAP solutions instead of ML solutions for the parameters, which avoid singularities. This is called Bayesian Mixture of Gaussians.

Avoiding local maxima

The algorithm is restarted with different initializations and the result with the highest log likelihood after convergence is chosen.

Chosing the number of components K

Open problem in optimization.
A heuristic is: Start with only one Gaussian and subsequently add further components as long as the Bayesian Information Criterion (BIC) grows.
When estimating model parameters using maximum likelihood estimation, it is possible to increase the likelihood by adding components, which may result in overfitting. The BIC resolves this problem by introducing a penalty term for the number of parameters in the model.

$$\begin{aligned} BIC &= 2L(\theta) - 6K\log(N) \\ &= \prod_{n=1}^{N}\sum_{k=1}^{K}\pi_k\mathcal{N}(x_n|\theta_k) - 6K\log(N) \end{aligned}$$

Parametric Density Estimation

The EM-Algorithm

Problems and Solutions