

Annotated  
Version

Machine Learning Course - CS-433

# Gaussian Mixture Models

Nov 7, 2019

changes by Martin Jaggi 2019, changes by Rüdiger Urbanke 2018, changes by Martin Jaggi  
2016, 2017 ©Mohammad Emtiyaz Khan 2015

Last updated on: November 7, 2019

**EPFL**

# Motivation

K-means forces the clusters to be spherical, but sometimes it is desirable to have *elliptical* clusters. Another issue is that, in K-means, each example can only belong to one cluster, but this may not always be a good choice, e.g. for data points that are near the “border”. Both of these problems are solved by using Gaussian Mixture Models.

## Clustering with Gaussians

The first issue is resolved by using full covariance matrices  $\Sigma_k$  instead of *isotropic* covariances.

$$p(\mathbf{X} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{z}) = \prod_{n=1}^N \prod_{k=1}^K [\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_{nk}}$$

gives ellipses instead of spheres

## Soft-clustering

The second issue is resolved by defining  $z_n$  to be a random variable. Specifically, define  $z_n \in \{1, 2, \dots, K\}$  that follows a multinomial distribution.

k-means:

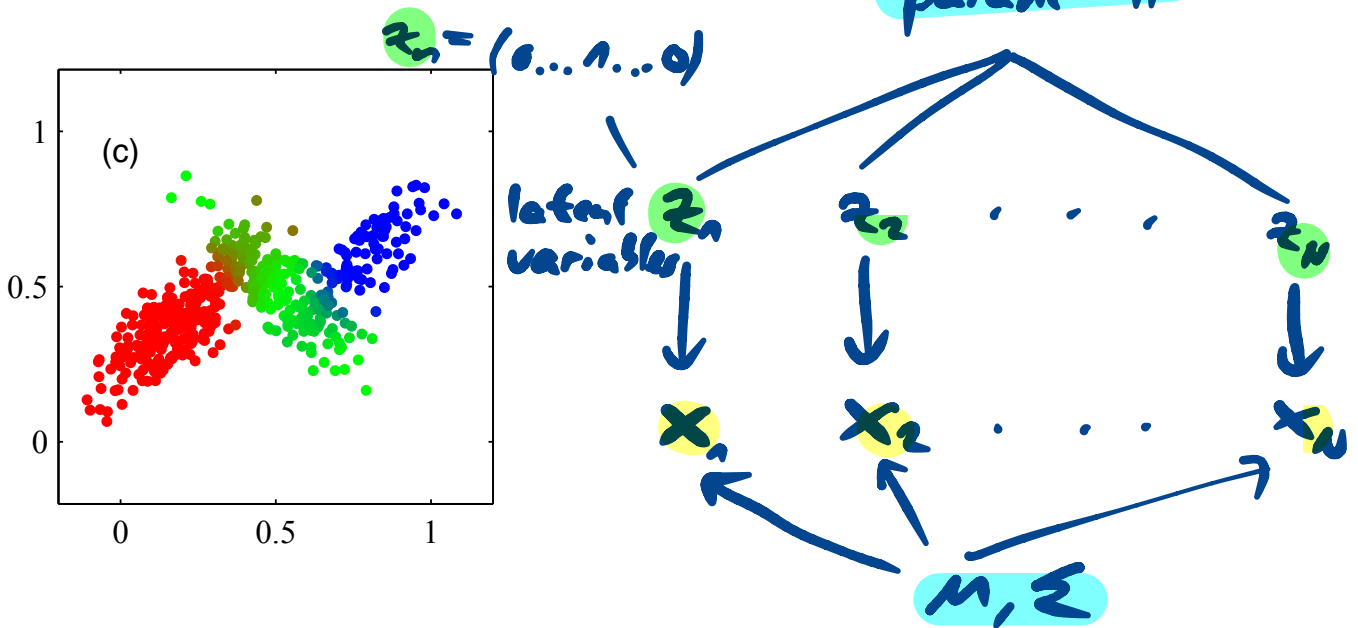
$$\Sigma_k = \mathbf{I}$$

parameters	
$\boldsymbol{\mu}$	$\mathbb{R}^{D \cdot K}$
$\boldsymbol{\Sigma}$	$\mathbb{R}^{D \cdot D \cdot K}$
$\boldsymbol{\pi}$	$\mathbb{R}^K$

importance weight of group k

$$p(z_n = k) = \pi_k \text{ where } \pi_k > 0, \forall k \text{ and } \sum_{k=1}^K \pi_k = 1$$

This leads to **soft-clustering** as opposed to having “hard” assignments.



## Gaussian mixture model

Together, the **likelihood** and the **prior** define the **joint** distribution of Gaussian mixture model (GMM).

$$p(\mathbf{X}, \mathbf{z} | \mu, \Sigma, \pi) = \prod_{n=1}^N p(\mathbf{x}_n, z_n | \mu, \Sigma, \pi)$$

Bayes Rule:  $p(a, b) = p(a|b) \cdot p(b)$

$$= \prod_{n=1}^N p(\mathbf{x}_n | z_n, \mu, \Sigma) p(z_n | \pi)$$

likelihood prior

$$= \prod_{n=1}^N \prod_{k=1}^K [\mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)]^{z_{nk}} \prod_{k=1}^K [\pi_k]^{z_{nk}}$$

$z_n = (0, 0, \dots, 1, \dots, 0)$

$z_{nk}$

Here,  $\mathbf{x}_n$  are observed data vectors,  $z_n$  are *latent* unobserved variables, and the unknown **parameters** are given by  $\theta := \{\mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K, \pi\}$ .

# Marginal likelihood

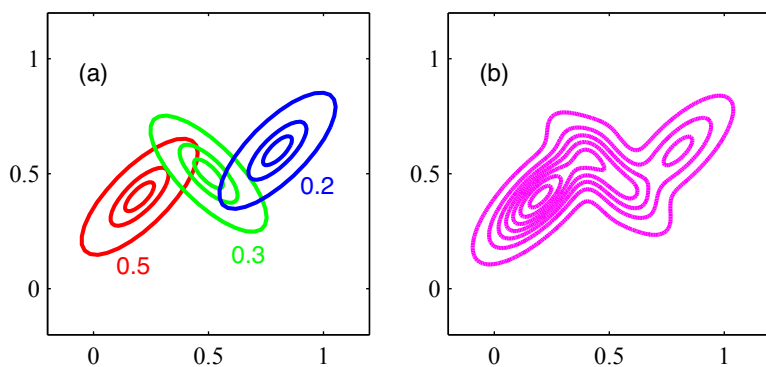
GMM is a **latent variable model** with  $z_n$  being the unobserved (latent) variables. An advantage of treating  $z_n$  as latent variables instead of *parameters* is that we can *marginalize* them out to get a cost function that does not depend on  $z_n$ , i.e. as if  $z_n$  never existed.

**likelihood:  $p(\mathbf{x}, \mathbf{z} | \theta)$**

Specifically, we get the following **marginal likelihood** by marginalizing  $z_n$  out from the likelihood:

**marginal likelihood**

$$p(\mathbf{x}_n | \theta) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)$$



Deriving cost functions this way, is good for *statistical efficiency*. Without a latent variable model, the number of parameters grow at rate  $O(N)$ . After marginalization, the growth is reduced to  $O(D^2 K)$  (assuming  $D, K \ll N$ ).

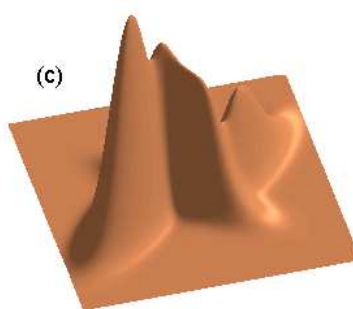
$$z_n = \begin{cases} (1, 0, \dots, 0) & \text{if } k=1 \\ \vdots & \\ (0, 0, \dots, 1) & \text{if } k=K \end{cases}$$

**joint**

$$p(\mathbf{x}_n, z_n)$$

**marginal**

$$\begin{aligned} p(\mathbf{x}_n) &:= \sum_{k=1}^K p(\mathbf{x}_n, z_n = k) \\ &= \sum_{k=1}^K p(\mathbf{x} | z) \cdot p(z = k) \\ &\quad \parallel \pi_k \end{aligned}$$



~~$$z : N$$~~

$$\begin{aligned} \theta = \mu_k &: K \cdot D \\ \Sigma_k &: K \cdot D^2 \\ \pi_k &: K \end{aligned}$$

# Maximum likelihood

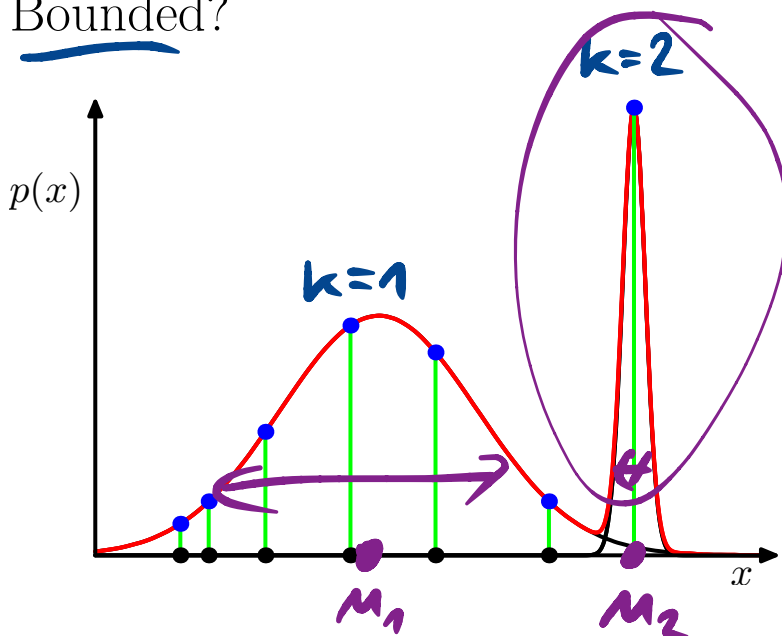
To get a maximum (marginal) likelihood estimate of  $\theta$ , we maximize the following:

$$\max_{\theta} \sum_{n=1}^N \log \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)$$

$$\log / P(\mathbf{x}_n | \theta) = \prod_{n=1}^N p(\mathbf{x}_n | \theta) = \sum_{k=1}^K \pi_k \mathcal{N}(\dots)$$

Is this cost convex? Identifiable? Bounded?

$\mathcal{L}(\theta)$



① non-convex

② non-unique optima

permutation of  $[k]$

$k \rightarrow k'$

$\pi_k \leftrightarrow \pi_{k'}$

$\mu_k \leftrightarrow \mu_{k'}$

$\Sigma_k \leftrightarrow \Sigma_{k'}$

③ non-bounded

$\mathcal{L} \rightarrow \infty$

if  $\Sigma_k = \sigma \cdot \mathbf{I}$  width

in the limit  $\sigma \rightarrow 0$