

# (Gaussian) Mixture Models and the Expectation Maximization Algorithm

**Morteza Chehreghani**

Chalmers University of Technology

May 25, 2019

# Review of the Last Week

$K$ -means objective corresponds to optimizing the following problem

$$\begin{aligned} \min_{\boldsymbol{\mu}, \mathbf{Z}} R(\boldsymbol{\mu}, \mathbf{Z}; \mathbf{X}) &= \min_{\boldsymbol{\mu}, \mathbf{Z}} \sum_{n=1}^N \sum_{k=1}^K z_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|_2^2. \\ \text{s.t. } z_{nk} &\in \{0, 1\} \text{ and } \sum_{k=1}^K z_{nk} = 1 \quad \forall n. \end{aligned}$$

Where,

$$\begin{aligned} \mathbf{X} &= [\mathbf{x}_1; \cdots; \mathbf{x}_N] \in \mathbb{R}^{N \times D}, \\ \boldsymbol{\mu} &= [\boldsymbol{\mu}_1; \cdots; \boldsymbol{\mu}_K] \in \mathbb{R}^{K \times D} \text{ and} \\ \mathbf{Z} &\in \{0, 1\}^{N \times K}. \end{aligned}$$

# From Hard to Soft Clustering

- Relax the 'hard' constraint given by

$$z_{nk} \in \{0, 1\}, \sum_{k=1}^K z_{nk} = 1 ,$$

- and replace it by a 'soft' constraint:

$$z_{nk} \in [0, 1], \sum_{k=1}^K z_{nk} = 1 .$$

# From Single to Mixture Models

[Old Faithful data set](#) includes 272 measurements of eruptions of the Old Faithful geyser at Yellowstone National Park. Each measurement consists of

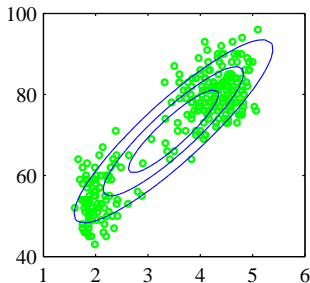
- ▶ the duration of the eruption in minutes;
- ▶ the time in minutes to the next eruption.



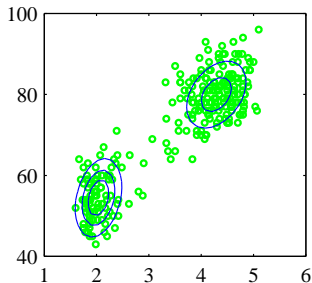
# From Single to Mixture Models

Plots of the 'old faithful' data

- ▶ Horizontal axis: the duration of the eruption in minutes.
- ▶ Vertical axis: the time in minutes to the next eruption.



(a) Modeling data with a single Gaussian distribution fitted by maximum likelihood



(b) Modeling data by a linear combination of two Gaussians fitted by maximum likelihood

# Gaussian Distribution (1-dim)

- ▶ Sample space  $\mathcal{X} = \mathbb{R}$
- ▶ Definition:

$$p(x|\mu, \sigma) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

- ▶ Statistics:

$$\mathbb{E}[X] := \mu, \text{Var}[X] := \sigma^2$$

# Gaussian Distribution (D-dim)

► Sample space  $\mathcal{X} = \mathbb{R}^D$ ,  $\mathbf{x} = (x_1, \dots, x_D)^\top$

► Definition:

$$p(\mathbf{x}|\mu, \Sigma) := \frac{1}{(\sqrt{2\pi})^D |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1}(\mathbf{x} - \mu)\right)$$

where  $\Sigma$  is the covariance matrix and  $|\Sigma|$  is its determinant

# Generative Clustering

## Generative approach

- ▶ **Goal:** explain the observed data  $\{\mathbf{x}_n\}_{n=1}^N$  by a probabilistic model  $p(\mathbf{x})$ .
- ▶ We assume the **parametric form** of the model to be chosen apriori.
- ▶ The model has **parameters** that need to be learned in order to explain the observed data well.
- ▶ Today we will focus on Gaussian Mixture Model.

## Clustering?

The model can be interpreted as assigning data points to different components/modes of a multimodal distribution.



# Introduction to Mixture Models

- ▶ Mixture of  $K$  probability densities is defined as

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k p(\mathbf{x} \mid \boldsymbol{\theta}_k).$$

Each probability distribution  $p(\mathbf{x} \mid \boldsymbol{\theta}_k)$  is a **component** of the mixture and has its own parameters  $\boldsymbol{\theta}_k$ .

- ▶ For a Gaussian component distribution the parameters  $\boldsymbol{\theta}_k$  are given by the mean  $\boldsymbol{\mu}_k$  and the covariance  $\boldsymbol{\Sigma}_k$ .

# Elements of Mixture Models

Mixture models are constructed from:

- ▶ Component distributions of the form  $p(\mathbf{x} \mid \boldsymbol{\theta}_k)$ .
- ▶ Mixing coefficients  $\pi_k$  that give the probability of each component.

In order for  $p(\mathbf{x})$  to be a proper distribution, we have to ensure that

$$\sum_{k=1}^K \pi_k = 1 \quad \text{and} \quad \pi_k \geq 0, \quad 1 \leq k \leq K.$$

Therefore, the parameters  $\pi_k, 1 \leq k \leq K$  define a categorical distribution representing the probability of each component.

# Gaussian Mixture Model

The Gaussian Mixture Model (GMM) uses Gaussians as the component distributions.

The distribution (of a particular point  $\mathbf{x}$ ) is written as

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- ▶ Given data points  $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ , the goal is to learn (estimate) the unknown parameters  $\boldsymbol{\mu}_k$ ,  $\boldsymbol{\Sigma}_k$ , and  $\pi_k$  such that we approximate the data as good as possible.
- ▶ This is equivalent to finding the parameters that **maximize the likelihood** of the given data.

# GMM: Generative Viewpoint

We assume that the the model parameters  $\Sigma, \mu, \pi$  are given.

Then, given those parameters, we sample the data  $\mathbf{x}_n$  as follows:

1. Sample a component (cluster) index  $k$  according to the probabilities  $\pi_k$ .
2. Sample a data point  $\mathbf{x}_n$  from the distribution  $p(\mathbf{x}_n \mid \mu_k, \Sigma_k)$ .

Parameter estimation based on maximizing likelihood:

Revert this process: data is given, but the parameters are unknown and should be estimated.

# Full Data Likelihood

We assume that the data points  $\mathbf{x}_n$  are independent and identically distributed (i.i.d.). The probability or likelihood of the observed data  $\mathbf{X}$ , given the parameters is then obtained by

$$p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^N p(\mathbf{x}_n) = \prod_{n=1}^N \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

# Maximum Log-Likelihood Formulation

**Goal.** find the parameters that maximize the likelihood of the data:

$$(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) \in \operatorname{argmax}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}} p(\mathbf{X} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

To simplify the calculation we take the logarithm, such that the product becomes a sum:

$$(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) \in \operatorname{argmax}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}} \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}.$$

# Maximum Log-Likelihood Estimation

- ▶ Want to solve:

$$(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) \in \operatorname{argmax}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}} \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}.$$

- ▶ Due to the presence of the summation over  $k$  inside the logarithm, the maximum likelihood solution for the parameters no longer has a closed-form analytic solution.

# Maximum Log-Likelihood Estimation

- ▶ Want to solve:

$$(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) \in \operatorname{argmax}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}} \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}.$$

- ▶ Due to the presence of the summation over  $k$  inside the logarithm, the maximum likelihood solution for the parameters no longer has a closed-form analytic solution.
- ▶ We employ an elegant powerful algorithmic technique, called **Expectation Maximization**.



# Maximum Log-Likelihood Estimation

- ▶ We want to solve:

$$(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) \in \operatorname{argmax}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}} \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}.$$

- ▶ Due to the presence of the summation over  $k$  inside the logarithm, the maximum likelihood solution for the parameters no longer has a closed-form analytic solution.
- ▶ We employ an elegant powerful algorithmic technique, called [Expectation Maximization](#).
- ▶ Intuition: if we know to which clusters the data points are assigned, then computing the maximum likelihood estimate becomes straightforward.
- ▶ Hence: we introduce a latent (or hidden) variable for the assignment of data points to clusters.

# Latent Variables

- ▶ Define  $K$ -dimensional binary random variable  $\mathbf{z}$  with a 1-of- $K$  representation.
- ▶ Only one element of  $\mathbf{z}$  is equal to 1 and all other elements are 0, i.e.,

$$z_k \in \{0, 1\}, \quad \sum_k z_k = 1.$$

# Latent Variables

- ▶ Define  $K$ -dimensional binary random variable  $\mathbf{z}$  with a 1-of- $K$  representation.
- ▶ Only one element of  $\mathbf{z}$  is equal to 1 and all other elements are 0, i.e.,

$$z_k \in \{0, 1\}, \quad \sum_k z_k = 1.$$

- ▶ The marginal distribution over  $\mathbf{z}$  is specified in terms of the mixing coefficients  $\pi_k$ , i.e.,

$$p(z_k = 1) = \pi_k.$$

# Latent Variables and Data Likelihood

- ▶  $\mathbf{z}$  uses a 1-of- $K$  representation. Thus, we write this distribution in the form of:

$$p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}.$$

- ▶ Also, the conditional distribution of  $\mathbf{x}$  given a particular instantiation (value) of  $\mathbf{z}$  is a Gaussian distribution

$$p(\mathbf{x} \mid z_k = 1) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- ▶ Therefore, we have:

$$p(\mathbf{x} \mid \mathbf{z}) = \prod_{k=1}^K \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}.$$

# Marginal Distribution with Latent Variables

The marginal distribution of  $\mathbf{x}$  can be obtained by summing the joint distribution over all possible states of  $\mathbf{z}$  to yield:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x} | \mathbf{z}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

For the full data log-likelihood we have:

$$\ln p(\mathbf{X} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}.$$

In the following, for the simplicity of presentation, we assume that the covariances  $\boldsymbol{\Sigma}$  are given (we do not need to estimate them).

# Responsibilities

- ▶  $\gamma(z_{nk})$ : probability of assigning a data point to a cluster

$$\gamma(z_{nk}) := p(z_{nk} = 1 \mid \mathbf{x}_n)$$

- ▶ Remember the generative viewpoint!
- ▶ We shall view  $\pi_k$  as the **prior** probability of  $z_{nk} = 1$ , and the quantity  $\gamma(z_{nk})$  as the corresponding **posterior** probability once we have observed  $\mathbf{x}_n$ .

# Overview of Expectation-Maximization

- ▶ We want to solve:

$$(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) \in \operatorname{argmax}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}} \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}.$$

- ▶ Due to the presence of the summation over  $k$  inside the logarithm, the maximum likelihood solution for the parameters no longer has a closed-form analytic solution.
- ▶ We employ an elegant powerful algorithmic technique, called **Expectation Maximization**.

# Overview of Expectation-Maximization

- ▶ We employ an elegant powerful algorithmic technique, called **Expectation Maximization**.
- ▶ First, we select some initial values for the means and mixing coefficients. Then, we alternate between the following two updates called the E (expectation) step and the M (maximization) step:
  1. In the expectation step, the current values for the model parameters are used to compute the posterior probabilities (responsibilities)  $\gamma(z_{nk})$ .
  2. In the maximization step, the responsibilities are used to estimate the model parameters (e.g., means and mixing coefficients).



# Expectation Step

- $\gamma(z_{nk})$ : probability of assigning  $\mathbf{x}_n$  to the  $k$ 's cluster

$$\gamma(z_{nk}) := p(z_{nk} = 1 \mid \mathbf{x}_n)$$

## Bayes' rule

The conditional probability of  $A$  given  $B$  (posterior) can be obtained by:

$$p(A|B) = \frac{p(A)p(B|A)}{p(B)}.$$

We call  $p(A)$  prior,  $p(B|A)$  likelihood and  $p(B)$  evidence.

# Expectation Step

## Bayes' rule

The conditional probability of  $A$  given  $B$  (posterior) can be obtained by:

$$p(A|B) = \frac{p(A)p(B|A)}{p(B)}.$$

We call  $p(A)$  prior,  $p(B|A)$  likelihood and  $p(B)$  evidence.

$$\gamma(z_{nk}) := p(z_{nk} = 1 \mid \mathbf{x}_n) = ?$$

We use the Bayes' rule to get

$$\begin{aligned}\gamma(z_{nk}) := p(z_{nk} = 1 \mid \mathbf{x}_n) &= \frac{p(z_{nk} = 1)p(\mathbf{x}_n \mid z_{nk} = 1)}{\sum_{j=1}^K p(z_{nj} = 1)p(\mathbf{x}_n \mid z_{nj} = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}\end{aligned}$$

# Estimating the Means

- ▶ We set the derivatives of  $\ln p(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$  with respect to the means  $\boldsymbol{\mu}_k$  to zero, and obtain:

$$0 = \sum_{n=1}^N \frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\underbrace{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}_{\gamma(z_{nk})}} \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k).$$

- ▶ Assume that  $\boldsymbol{\Sigma}_k$  is not singular. Multiplying by  $\boldsymbol{\Sigma}_k$  we obtain

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n, \quad N_k = \sum_{n=1}^N \gamma(z_{nk})$$

- ▶ The mean  $\boldsymbol{\mu}_k$  is obtained by taking a weighted mean of all the points in the data set.

# Estimating the Variances

- If we set the derivative of  $\ln p(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$  with respect to  $\boldsymbol{\Sigma}_k$  to zero we obtain

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

## Estimating the Coefficients

- ▶ Maximizing  $\ln p(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$  with respect to the mixing coefficients  $\pi_k$  and taking account of the constraint which requires the mixing coefficients to sum to one, leads to the following Lagrangian

$$\ln p(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left( \sum_{k=1}^K \pi_k - 1 \right)$$

which gives

$$\begin{aligned} 0 &= \sum_{n=1}^N \frac{\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda. \\ \Rightarrow 0 &= \sum_{n=1}^N \gamma(z_{nk}) + \pi_k \lambda = N_k + \pi_k \lambda. \end{aligned}$$

Then,  $\sum_{k=1}^K \pi_k = 1$  leads to  $\lambda = -N$ . Thus,

$$\pi_k = \frac{N_k}{N}.$$

# Description of EM

Given a GMM, the goal is to maximize the likelihood function with respect to the parameters.

1. Initialize the means  $\boldsymbol{\mu}_k$ , and mixing coefficients  $\pi_k$ . Set the  $\boldsymbol{\Sigma}_k$  to the given covariances.
2. **E-step.** Evaluate the responsibilities using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

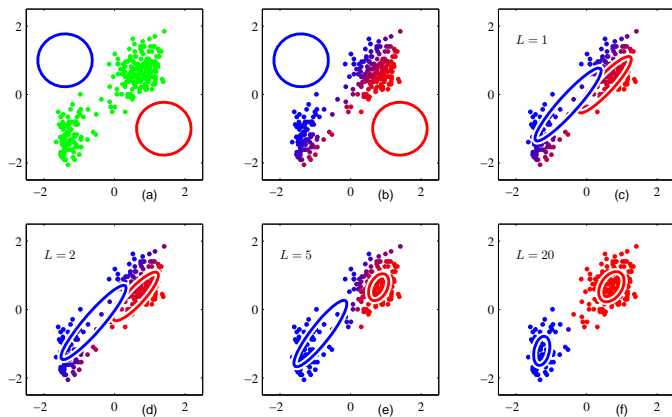
3. **M-step.** Re-estimate the parameters using the current responsibilities

$$\begin{aligned}\boldsymbol{\mu}_k &= \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n \\ \pi_k &= \frac{N_k}{N} \quad \text{where} \quad N_k = \sum_{n=1}^N \gamma(z_{nk})\end{aligned}$$

4. Compute the log-likelihood and check for the convergence of either the parameters or the log-likelihood.

# Example of EM for Gaussian Mixture Models

Illustration of the EM algorithm using the Old Faithful data set.



**Figure:** EM algorithm for mixture of two Gaussians. Note that here the covariance is also estimated (illustrated by the two ellipsoids).

# EM vs. $K$ -means Algorithm

- ▶ The  $K$ -means algorithm yields a hard assignment of data points to clusters, but the EM algorithm performs a soft assignment based on the posterior probabilities.
- ▶ The  $K$ -means algorithm does not estimate the covariances of the clusters but only the cluster means.



# $K$ -means and EM

- ▶ The EM algorithm takes many more iterations to reach convergence compared with the  $K$ -means algorithm, and each cycle requires significantly more computation.
- ▶ The  $K$ -means algorithm can be used to find a suitable initialization for a Gaussian mixture model.
- ▶ The covariance matrices can be initialized to the sample covariances of the clusters found by the  $K$ -means algorithm.
- ▶ The mixing coefficients can be set to the fractions of data points assigned to the respective clusters.
- ▶ There will generally be multiple local maxima of the log likelihood function, and EM is not guaranteed to find the largest of these maxima.

# Model Order Selection: General Principle

Trade-off between two conflicting goals:

**Data fit:** We want to predict the data accurately, e.g., maximize the likelihood. The likelihood usually improves by increasing the number of clusters.

**Complexity:** Choose a model that is not very complex which is often measured by the number of free parameters.

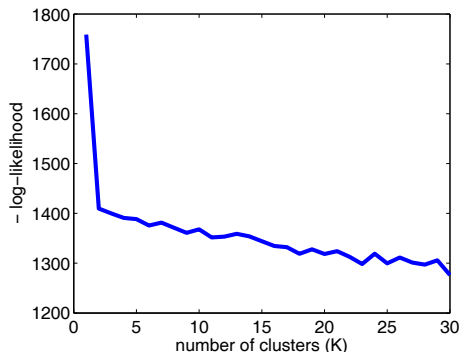
Find a trade-off between these two goals!

# Decreasing the data fit costs when increasing $K$

**Negative Log-Likelihood** of data for  $K$  mixture Gaussians:

$$-\ln p(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}.$$

The smaller the negative log-likelihood, the better the fit.



# AIC and BIC

## Trade-off

Achieve balance between data fit (measured by likelihood  $p(\mathbf{X}|\cdot)$ ) and model complexity. Complexity can be measured by the number of free parameters  $c_K$ .

## Different principles to choose $K$

- *Akaike Information Criterion* (**AIC**)

$$AIC_K = -\ln p(\mathbf{X}|\cdot) + c_K$$

# AIC and BIC

## Trade-off

Obtain a balance between data fit (measured by likelihood  $p(\mathbf{X}|\cdot)$ ) and model complexity. Complexity can be measured by the number of free (unknown) parameters  $c_K$ .

## Different principles to choose $K$

- ▶ *Akaike Information Criterion* (**AIC**)

$$AIC_K = -\ln p(\mathbf{X}|\cdot) + c_K$$

- ▶ *Bayesian Information Criterion* (**BIC**).

$$BIC_K = -\ln p(\mathbf{X}|\cdot) + \frac{1}{2}c_K \ln N$$

# AIC and BIC

Which one is more strict on the model complexity?

# AIC and BIC

Which one is more strict on the model complexity?

- Usually (on a large enough dataset), the BIC criterion penalizes complexity more than AIC.

# AIC and BIC: Remarks and Example

## Analysis

A single AIC (BIC) result is meaningless. One has to repeat the analysis for different  $K$ 's and compare the differences: the most suitable number of clusters corresponds to the smallest AIC (BIC) value.

## Example (Mixture of Gaussians with full covariance matrices)

Number of free parameters  $c_K$  is (?)



# AIC and BIC: Remarks and Example

## Analysis

A single AIC (BIC) result is meaningless. One has to repeat the analysis for different  $K$ s and compare the differences: the most suitable number of clusters corresponds to the smallest AIC (BIC) value.

## Example (Mixture of Gaussians)

Number of free parameters  $c_K$  is:

$$c_K = K \cdot D + (K - 1) + K \cdot D \cdot (D + 1)/2.$$

- ▶  $K$  is the number of estimated clusters,
- ▶  $D$  is the number of dimensions of the data

# AIC and BIC: Remarks and Example

## Example (Mixture of Gaussians)

What about if the covariance matrices are all known in advance?

# AIC and BIC: Remarks and Example

## Example (Mixture of Gaussians)

What about if the covariance matrices are all known in advance?

Number of free parameters is:

$$c_K = K \cdot D + (K - 1).$$

# AIC and BIC example: 3 clusters

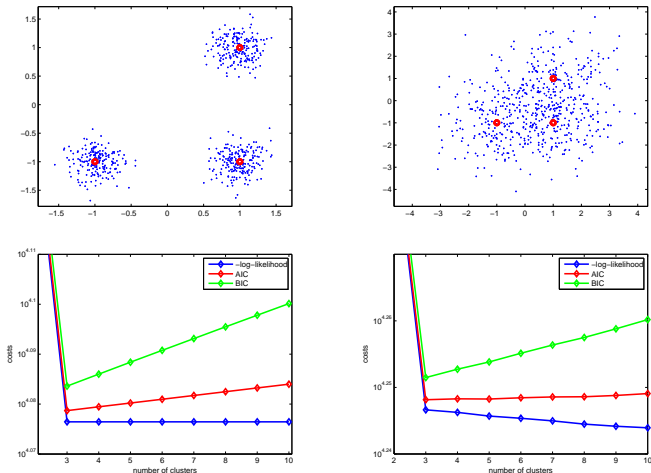


Figure: Model order selection on synthetic datasets with 3 clusters. Synthetic data has smaller variance on the left than on the right.

# AIC and BIC example: 5 clusters

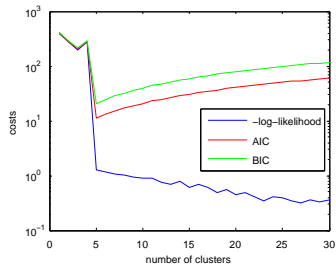
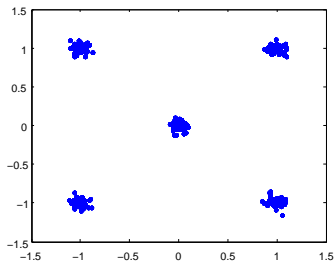


Figure: Model order selection on a synthetic dataset with 5 clusters.

## Exercise

Consider a GMM with  $\Sigma_k = \sigma_k^2 \mathbf{I}$  and one of the component means is equal to a data point:  $\mu_j = \mathbf{x}_n$ .

1. Write down the log-likelihood for this data point (i.e.  $\ln p(\mathbf{x}_n | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ )
2. Calculate  $p(\mathbf{x}_n | \mu_j, \Sigma_j)$ , the probability of  $\mathbf{x}_n$  given the  $j$ -th component.
3. In the limit  $\sigma_k \rightarrow 0$ , how does the probability in Question 2 change? Discuss the impact of this issue on the maximization of the likelihood function.
4. Can this situation occur in the case of a single Gaussian distribution (i.e., when  $K = 1$ )?
5. Propose a heuristic to avoid such a situation for a GMM.

# Reference

Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Chapter 9.