Learning with hidden variables

Generative and Deep Learning (GDL)

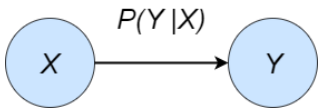
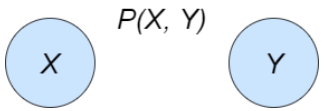
Davide Bacciu (davide.bacciu@unipi.it)

UNIVERSITÀ DI PISA

Lecture Outline

- Learning with non observed variables
- Latent/hidden variable models
- Maximum likelihood learning with latent variable
- Expectation-Maximization algorithm
- Exact maximum likelihood learning in mixture models

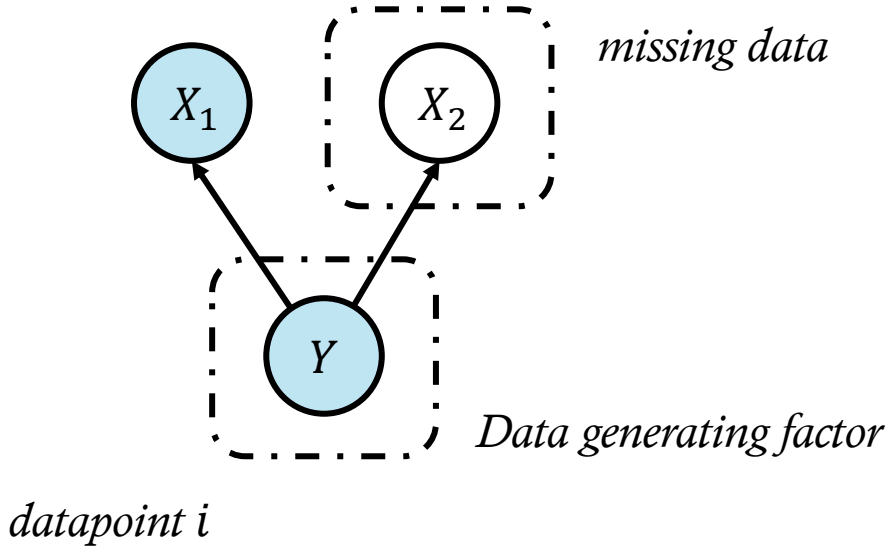
Learning with graphical models

		Structure	
		Fixed Structure 	Fixed Variables 
Data	Complete	Naive Bayes Calculate Frequencies (ML)	Discover dependencies from the data Structure Search Independence tests
	Incomplete	**Latent variables** **EM Algorithm (ML)** **MCMC, VBEM (Bayesian)**	Difficult Problem Structural EM

Latent/Hidden Variable Models

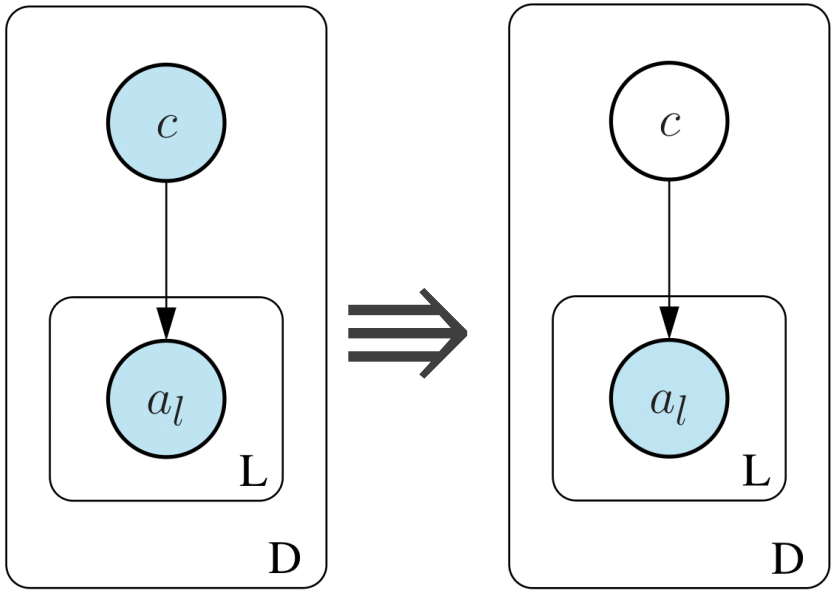
Dealing with Incomplete Data

- Incomplete data
  - Missing observations (imputation)
  - Unobserved random variables
- Key idea
  - Complete the data making hypothesis on the unobserved variables
  - Review the hypothesis
  - Iterate



What if...

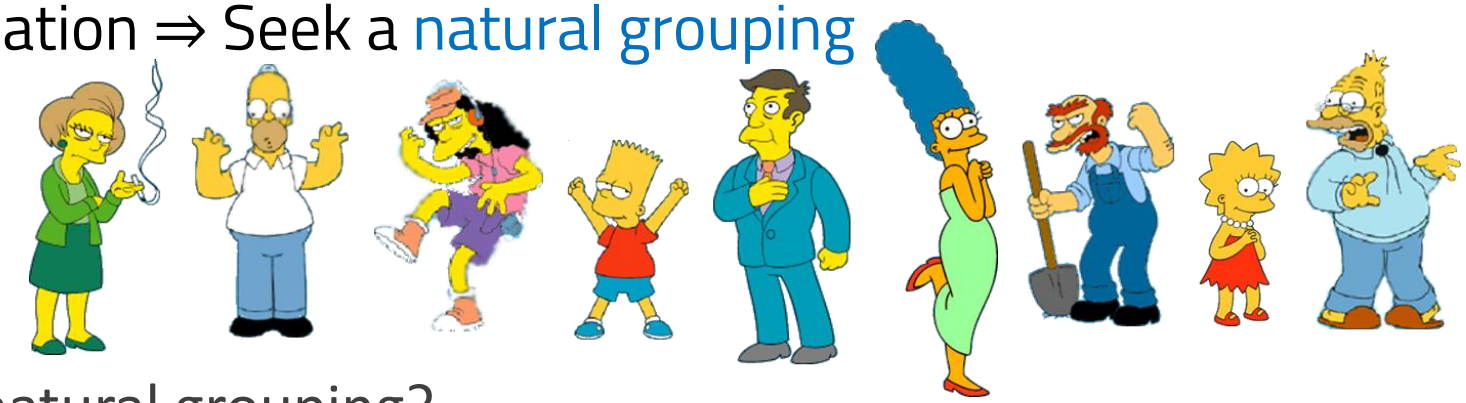
Sample classification is no longer observable



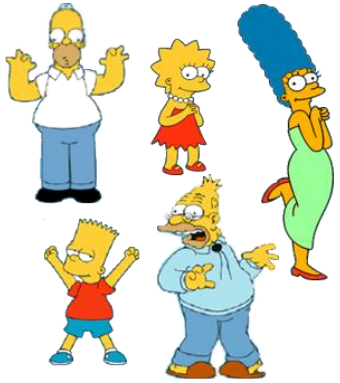
c is now a **Hidden/Latent Variable** within a **Mixture Model**

Mixture Model

No classification ⇒ Seek a natural grouping



What is a natural grouping?



Family

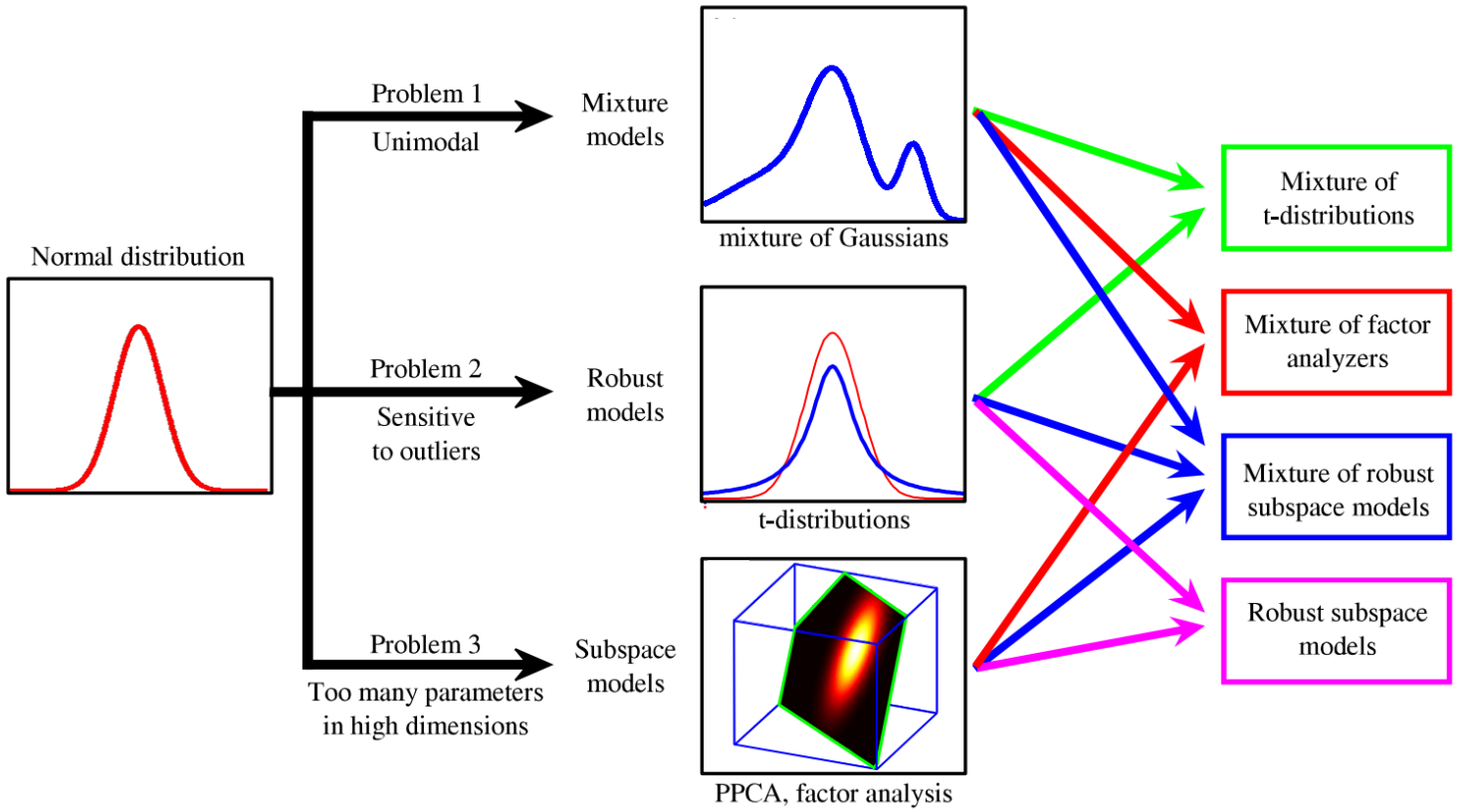
School

Females

Males

Credit goes to Eamonn Keogh @ UCR

Mixture Models



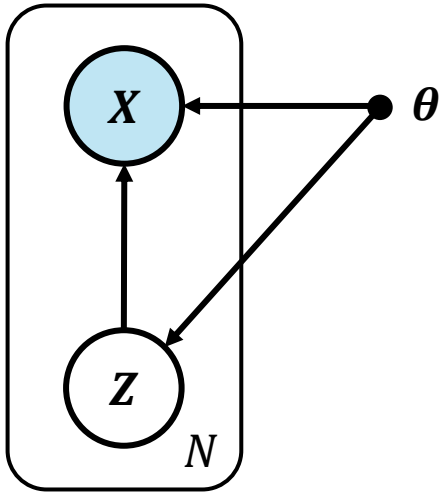
Latent Variable Models

We want to model a probability distribution $P(\boldsymbol{X} = \boldsymbol{x}|\theta)$, parametrized by $\boldsymbol{\theta}$, with $\boldsymbol{X}$ being a multivariate RV.

high-dimensional data $\boldsymbol{x}$

digit, rotation, scale, etc..

latent factors $\boldsymbol{z}$ *(low-dimensional)*



N datapoints

These models contains both

- Observed random variables **X** (i.e. for which we have training data)
- Unobserved (hidden/latent) variables **Z** (e.g. data clusters)

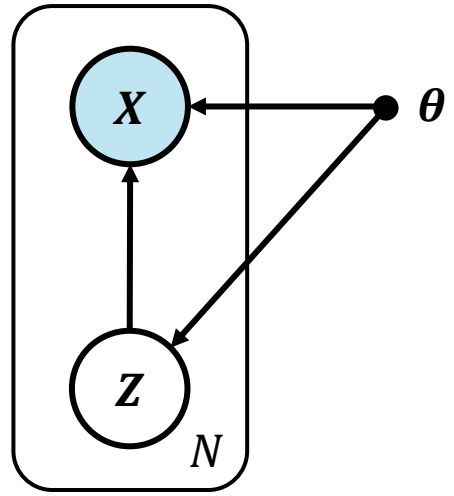
Joint distribution with latents

We want to model a probability distribution $P(\boldsymbol{X} = \boldsymbol{x}|\theta)$, with $\boldsymbol{X}$ being a multivariate RV

$\boldsymbol{\theta}$ is a generic term to denote params, but these distributions have **different** params

Joint distribution **defined** as

$$P(\boldsymbol{x}, \boldsymbol{z}|\theta) = P(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\theta})P(\boldsymbol{z}|\boldsymbol{\theta})$$



Marginal likelihood (marginalize out the multivariate RV $\boldsymbol{Z}$)

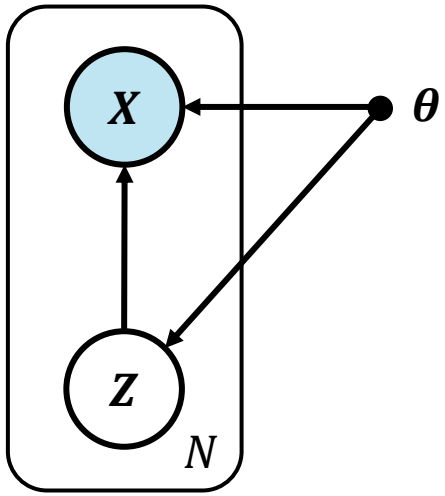
continuous $\quad P(\boldsymbol{x}|\theta) = \int P(\boldsymbol{x}, \boldsymbol{z}|\theta)d\boldsymbol{z} = \int P(\boldsymbol{x}|\boldsymbol{z}, \theta)P(\boldsymbol{z}|\theta)d\boldsymbol{z}$

discrete $\quad P(\boldsymbol{x}|\theta) = \sum_{\boldsymbol{z}} P(\boldsymbol{x}, \boldsymbol{z}|\theta) = \sum_{\boldsymbol{z}} P(\boldsymbol{x}|\boldsymbol{z}, \theta)P(\boldsymbol{z}|\theta)$

in both cases $\quad P(\boldsymbol{x}|\theta) = \mathbb{E}_{P(\boldsymbol{z}|\theta)}[P(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\theta})]$

Learning with latent variables

Learning in latent variable models



Learning: Obtain the value (ML/MAP) or a probability distribution (Bayesian) for the parameters $\boldsymbol{\theta}$ of all the distributions involved
Here: $P(\boldsymbol{X}|\boldsymbol{Z},\theta)$ and $P(\boldsymbol{Z}|\theta)$

We focus on Maximum Likelihood estimation

$$\theta_{ML} = \arg\max_{\theta} P(\boldsymbol{X}|\theta)$$

but we need to circumvent the issue that we don't know how to compute $P(\boldsymbol{X}|\theta)$ because it does not contain the $\boldsymbol{Z}$ needed by the model

Completing the likelihood

- Given the likelihood maximization

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} \mathcal{L}(\theta|\mathbf{X}) = \arg\max_{\theta} \mathrm{P}(\mathbf{X}|\theta)$$

- $\mathcal{L}(\theta|\mathbf{X})$ is the incomplete likelihood and $\mathbf{X}$ is the incomplete data containing only observed random variables
- Assume complete data $\mathbf{d} = (\mathrm{X}, \mathrm{Z})$ exists where $\mathrm{z} \in \mathbf{Z}$ are hidden or latent variables
- Obtain the complete likelihood

$$\mathcal{L}_{\mathrm{c}}(\theta|\mathbf{d}) = \mathrm{P}(\mathbf{d}|\theta) = \mathrm{P}(\mathbf{X}, \mathbf{Z}|\theta) = \mathrm{P}(\mathbf{Z}|\mathbf{X}, \theta)\mathrm{P}(\mathbf{X}|\theta)$$

Key Idea

Maximize the complete likelihood $\mathcal{L}_c$ instead of $\mathcal{L}$

Expectation Maximization

Algorithm solving an optimization problem involving maximization of complete log-likelihood $\mathcal{L}_c(\theta)$ w.r.t. model parameters θ

$$\theta^{(k+1)} = \arg\max_{\theta} \sum_{\boldsymbol{z}} P\left(\boldsymbol{Z} = \boldsymbol{z} \middle| \boldsymbol{X}, \theta^{(k)}\right) \log \mathrm{P}(\mathbf{X}, \mathbf{Z} = \mathbf{z}|\theta)$$

Requires that posterior computation is feasible

$$\mathbb{E}_{P(\boldsymbol{Z}|\boldsymbol{X},\theta^{(k)})}[\cdot]$$

A 2-step iterative algorithm

E-Step: Given the current estimate of the model parameters $\theta^{(k)}$, compute

$$Q^{(k+1)}\left(\theta \middle| \theta^{(k)}\right) = \mathbb{E}_{\boldsymbol{z} \sim P(\boldsymbol{Z}|\boldsymbol{X},\theta^{(k)})}\left[\log P\left(\boldsymbol{X}, \boldsymbol{Z} = \boldsymbol{z} \middle| \theta^{(k)}\right)\right]$$

M-Step: Find the new estimate of the model parameters

$$\theta^{(k+1)} = \arg\max_{\theta} Q^{(k+1)}(\theta|\theta^{(k)})$$

Iterate until

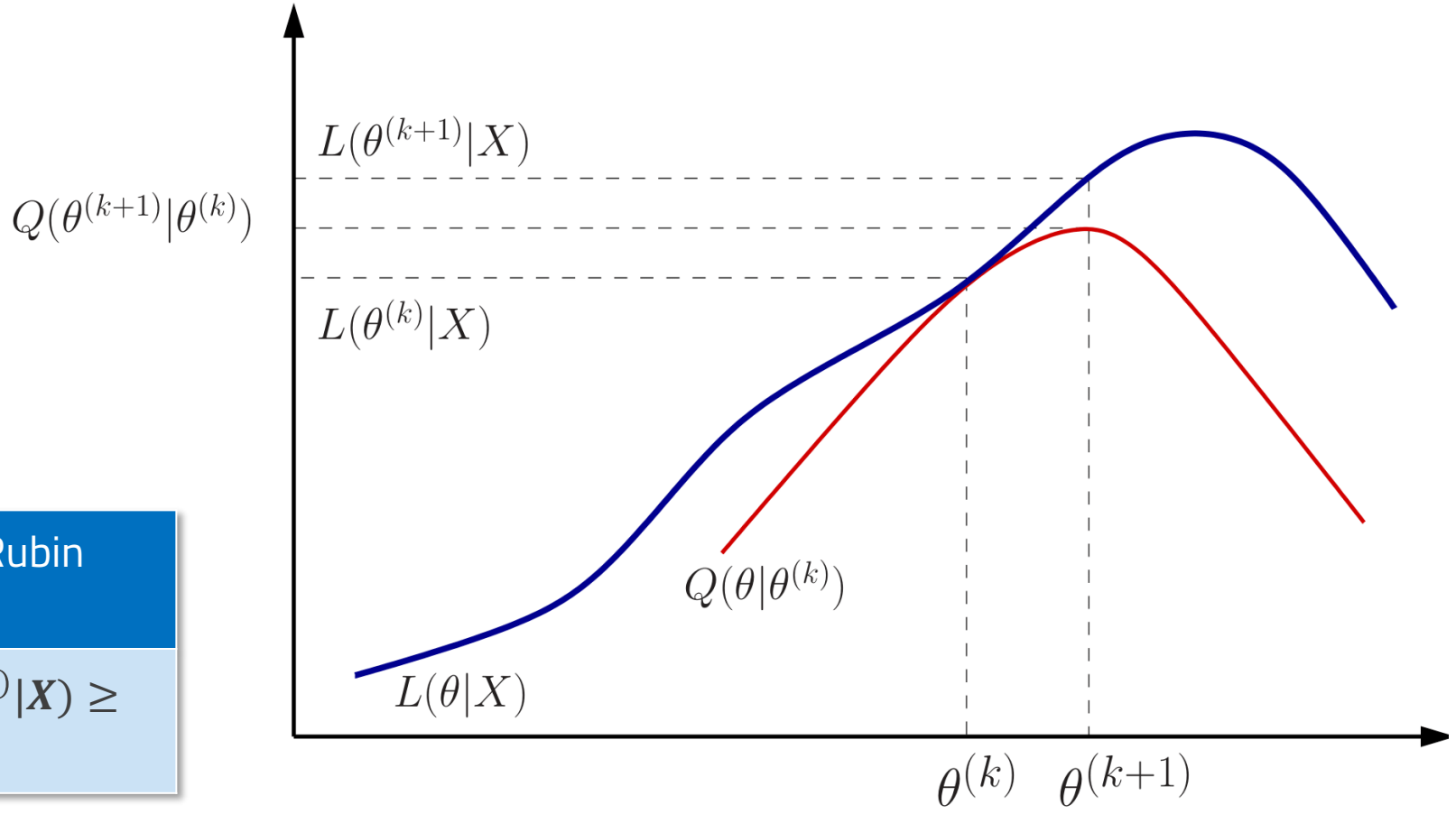
$$\left|\mathcal{L}_c(\theta)^{k+1} - \mathcal{L}_c(\theta)^{k}\right| < \epsilon$$

(or stop if maximum number of iterations is reached)

EM Graphically

Theorem (Dempster, Laird and Rubin (1997))

An EM process ensures $\mathcal{L}(\theta^{(k+1)}|\boldsymbol{X}) \geq \mathcal{L}(\theta^{(k)}|\boldsymbol{X})$



What if posterior is intractable?

Sampling methods

- Monte Carlo
  - Basic integration
  - Importance Sampling
  - Rejection Sampling
  - …
- Markov Chain Monte Carlo (MCMC)
  - Metropolis Algorithm
  - Gibbs Sampling
  - Hamiltonian Monte Carlo (HMC)

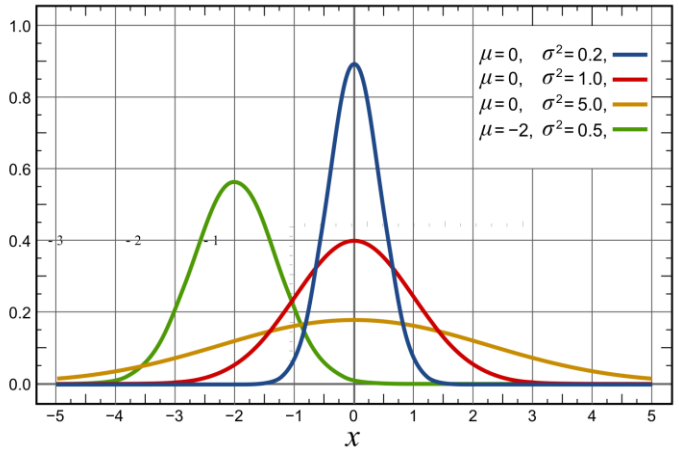
Variational methods

- Mean Field Variational Bayes
- Variational Expectation Maximization (VEM)
- Stochastic Variational Inference (SVI)
- Amortized Variational Inference
  - Variational Autoencoders

We shall see (some of) these in the next lectures!

Gaussian Mixture Models

Univariate Gaussian



Random Variable

Unidimensional $x \in \mathbb{R}$

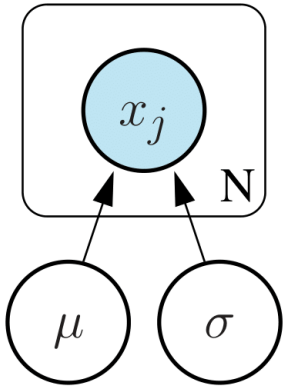
Parameterized Family

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

Model Parameters $\theta = (\mu,\sigma)$

Graphical model

Fully observable with single RV $x_j \in \mathbb{R}$ and two parameters μ and σ

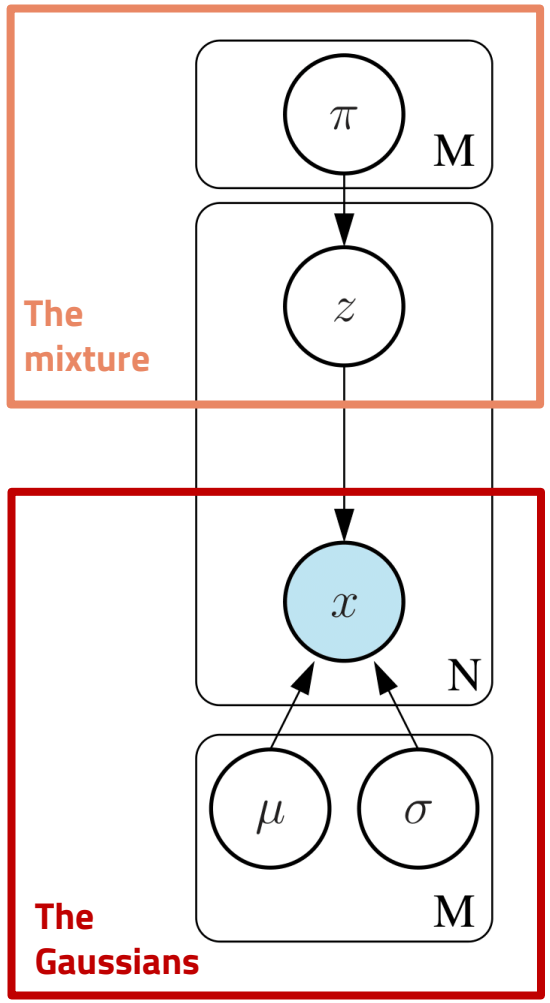


Can be learned by maximum likelihood optimization

$$<\mu_{ML},\sigma_{ML}> = \arg\max_{\mu,\sigma}\log P(\boldsymbol{d}|\mu,\sigma) = \arg\max_{\mu,\sigma}\sum_{j}^{N}\log P(x_j|\mu,\sigma)$$

$$= \arg\max_{\mu,\sigma}\sum_{j}^{N}\left(\log\sqrt{2\pi}\sigma - \left(\frac{(x_j-\mu)^2}{2\sigma^2}\right)\right)$$

Gaussian Mixture Model (GMM)



- A model for continuous observable data $\boldsymbol{x} \in \mathbb{R}^d$ involving M Gaussians that are mixed by latent/hidden variables $z \in \{1, .., M\}$
  - Each observation $\boldsymbol{x}$ is generated by a Gaussian with probability

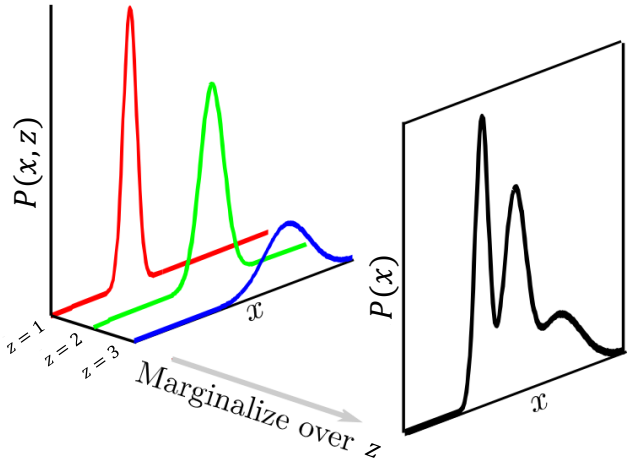
$$P(\boldsymbol{x}|z = m) \sim \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m)$$

  - z is the hidden mixture selection variable with prior

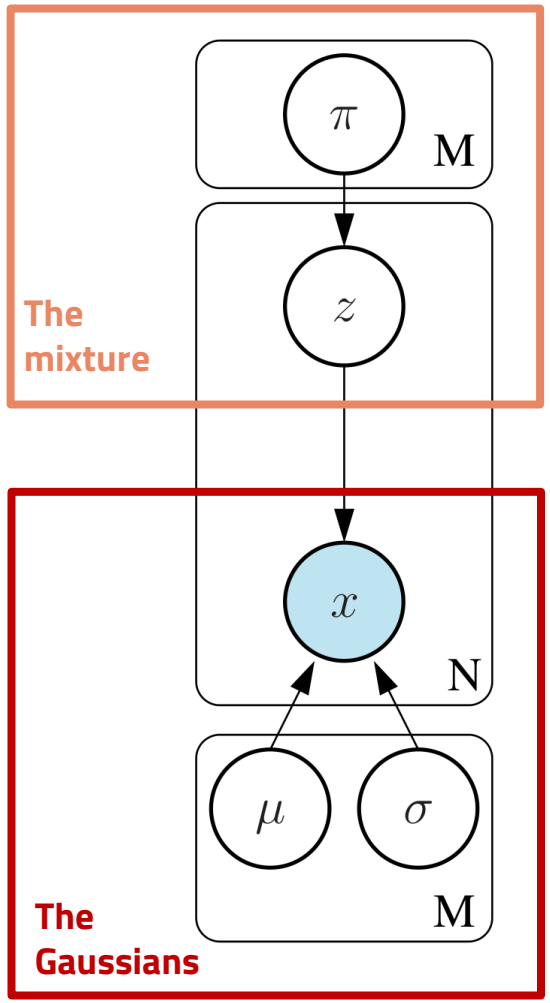
$$P(z = m) = \pi_m$$

- Model parameters $\theta = (\pi, \mu, \sigma)$ estimated by the EM algorithm

The mixture density is created by marginalizing the hidden z in the joint $P(x, z)$



Gaussian Mixture Model (GMM)



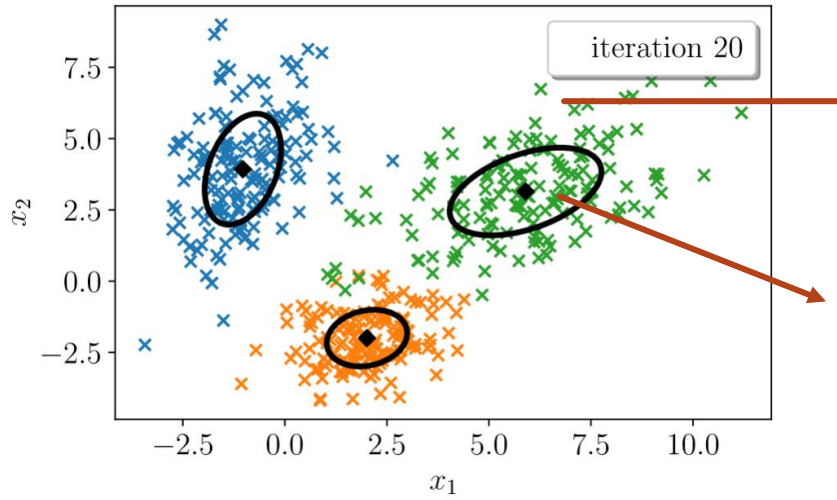
- A model for continuous observable data $\boldsymbol{x} \in \mathbb{R}^d$ involving M Gaussians that are mixed by latent/hidden variables $z \in \{1, .., M\}$
  - Each observation $\boldsymbol{x}$ is generated by a Gaussian with probability

$$P(\boldsymbol{x}|z = m) \sim \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m)$$

  - z is the hidden mixture selection variable with prior

$$P(z = m) = \pi_m$$

- Model parameters $\theta = (\pi, \mu, \sigma)$ estimated by the EM algorithm



Can be used to cluster data into groups

The "group" corresponds to the value of the latent variable

Inference (feasible!)

$$P_{\boldsymbol{\theta}}(z = m|\boldsymbol{x}) = \frac{P_{\boldsymbol{\theta}}(\boldsymbol{x}|z = m)P_{\theta}(z = m)}{P_{\boldsymbol{\theta}}(\boldsymbol{x})}$$

GMM Likelihood

univariate to ease notation

A set of i.i.d observations $\boldsymbol{X} = \{x_1, \ldots, x_n\}$ has incomplete likelihood

$$\mathcal{L}(\theta|\boldsymbol{X}) = P(\boldsymbol{X}|\theta) = \prod_{j=1}^{N} P(x_j|\theta)$$

Use marginalization to introduce the hidden variable $z_j = m$ (for all m)

$$\mathcal{L}(\theta|\boldsymbol{X}) = \prod_{j=1}^{N} \sum_{m=1}^{M} P\big(x_j, z_j = m\big|\theta\big)$$

Following the independence relationships in the graphical model this rewrites

$$\mathcal{L}(\theta|\boldsymbol{X}) = \prod_{j=1}^{N} \sum_{m=1}^{M} P(z_j = m|\pi) P(x_j|z_j = m, \mu, \sigma)$$

Categorical

univariate Gaussian from slide 18

A technicality towards the complete likelihood

- We define the complete log-likelihood using auxiliary indicator variables

$$\bar{z}_{jm}(z_j) = \begin{cases} 1 & \text{if } z_j = m \\ 0 & \text{otherwise} \end{cases}$$

- In practice: it is like assuming we know the hidden assignment of each data point x_j to the mixture m for all j, m

Rationale: contribution of $P(\boldsymbol{x}_j | z_j = m, \theta) \cdot P(z_j = m | \theta)$ factor to the likelihood should be 0 if we know that sample j belongs to mixture $m' \neq m$

Indicator variables are binary RVs for which, by definition, we have $P(\bar{z}_{jm} = 1) = P(z_j = m | x_j)$

GMM Complete Likelihood

Assume we know the hidden assignment $z_{jm} = 1$ for each sample j, we have the following complete likelihood

$$\mathcal{L}_c(\theta|\boldsymbol{X}, \boldsymbol{Z}) = \prod_{j=1}^{N} \sum_{m=1}^{M} z_{jm} \pi_m \, P\left(x_j \middle| \mu_m, \sigma_m\right)$$

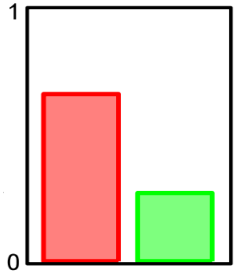
GMM Complete Likelihood

Assume we know the hidden assignment $z_{jm} = 1$ for each sample j, we have the following complete likelihood

$$\mathcal{L}_c(\theta|\boldsymbol{X},\boldsymbol{Z}) = \prod_{j=1}^{N}\sum_{m=1}^{M} z_{jm}\pi_m P(x_j|\mu_m,\sigma_m) = \prod_{j=1}^{N}\prod_{m=1}^{M}\left(\pi_m P(x_j|\mu_m,\sigma_m)\right)^{z_{jm}}$$

Can be better handled by taking the log

$$\log\mathcal{L}_c(\theta|\boldsymbol{X},\boldsymbol{Z}) = \sum_{j=1}^{N}\sum_{m=1}^{M} z_{jm}\log\left(\pi_m P(x_j|\mu_m,\sigma_m)\right)$$



Now we can compute

$$Q(\theta|\theta^{(k)}) = E_{P(\boldsymbol{Z}|\boldsymbol{X},\theta^{(k)})}[\log\mathcal{L}_c(\theta|\boldsymbol{X},\boldsymbol{Z})]$$

$$= \sum_{j=1}^{N}\sum_{m=1}^{M} P(z_j = m|x_j,\theta^{(k)})\log\left(\pi_m^{(k)} P\left(x_j\middle|\mu_m^{(k)},\sigma_m^{(k)}\right)\right)$$

$$P(\bar{z}_{jm} = 1) = P(z_j = m|x_j)$$

Expectation Step

To compute the auxiliary function

$$Q(\theta|\theta^{(k)}) = \sum_{j=1}^{N}\sum_{m=1}^{M} P\left(z_j = m \middle| x_j, \theta^{(k)}\right) \log\left(\pi_m^{(k)} P\left(x_j \middle| \mu_m^{(k)}, \sigma_m^{(k)}\right)\right)$$

we need to estimate the posterior $P\left(z_j = m \middle| x_j, \theta^{(k)}\right)$ based on the current $\theta^{(k)}$ values

$$P\left(z_j = m \middle| x_j, \theta^{(k)}\right) = \frac{\pi_m^{(k)} P\left(x_j \middle| \mu_m^{(k)}, \sigma_m^{(k)}\right)}{\sum_{m'} \pi_{m'}^{(k)} P\left(x_j \middle| \mu_{m'}^{(k)}, \sigma_{m'}^{(k)}\right)}$$

Maximization Step

Maximization of the auxiliary function with respect to the parameters $\theta = (\pi, \mu, \sigma)$

$$Q\left(\theta \middle| \theta^{(k)}\right) = \sum_{j=1}^{N} \sum_{m=1}^{M} P\left(z_j = m \middle| x_j, \theta^{(k)}\right) \log \pi_m$$

$$+ \sum_{j=1}^{N} \sum_{m=1}^{M} P\left(z_j = m \middle| x_j, \theta^{(k)}\right) \log P\left(x_j \middle| \mu_m, \sigma_m\right)$$

As usual this is performed by solving

$$\frac{\partial Q\left(\theta \middle| \theta^{(k)}\right)}{\partial \theta} = 0$$

while taking into account the sum-to-one constraint $\sum_{m=1}^{M} \pi_m = 1$

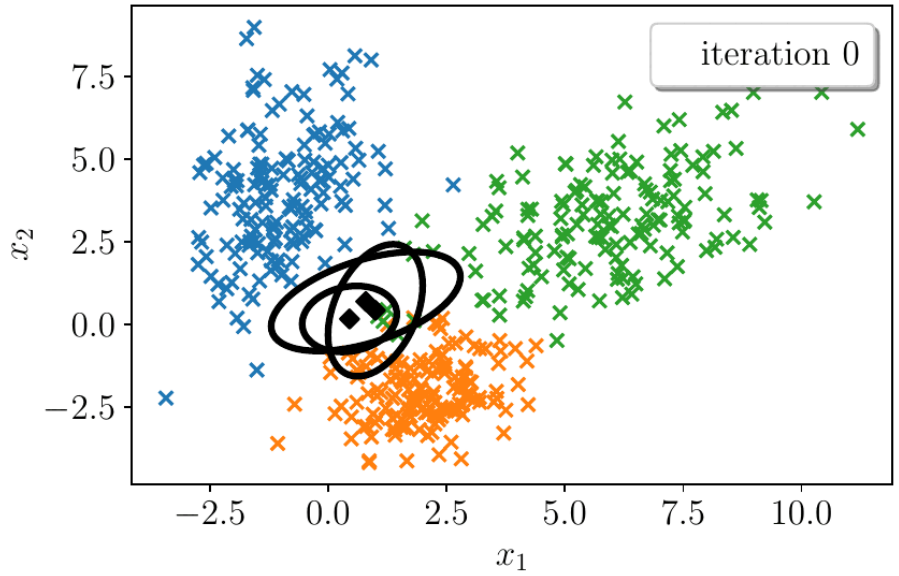
M-Step Learning Equations

Solving the independent maximization problems with respect to π_m, μ_m and σ_m yields

$$\pi_m^{(k+1)} = \frac{\sum_{j=1}^{N} P\left(z_j = m \middle| x_j, \theta^{(k)}\right)}{N}$$

$$\mu_m^{(k+1)} = \frac{\sum_{j=1}^{N} x_j P\left(z_j = m \middle| x_j, \theta^{(k)}\right)}{\sum_{j=1}^{N} P\left(z_j = m \middle| x_j, \theta^{(k)}\right)}$$

$$\sigma_m^{(k+1)} = \sqrt{\frac{\sum_{j=1}^{N} P\left(z_j = m \middle| x_j, \theta^{(k)}\right)\left(x_j - \mu^{(k)}\right)^2}{\sum_{j=1}^{N} P\left(z_j = m \middle| x_j, \theta^{(k)}\right)}}$$



Wrap-Up

The EM Algorithm in a Nut-Shell

The EM algorithm can be summarized by the optimization problem

$$\theta^{(k+1)} = \arg\max_{\theta} \sum_{\boldsymbol{z}} P\left(\boldsymbol{Z} = \boldsymbol{z} \middle| \boldsymbol{X}, \theta^{(k)}\right) \log \mathcal{L}_c(\theta | \boldsymbol{X}, \boldsymbol{Z} = \boldsymbol{z})$$

The **posterior** $\boldsymbol{P\left(Z = z \middle| X, \theta^{(k)}\right)}$ provides an **expected count** for the **event** $\boldsymbol{z}$ that is the EM counterpart of basic ML counting

Take Home Messages

- Latent variable models allow generation of visible information (data) to be steered by hidden knowledge about non-observable variables
  - Assuming conditional independence relationships are known
- Expectation-Maximization
  - Use it for dealing with ML learning with unobserved variables
  - Complete observations with hidden variables
  - Use posterior to estimate unobserved counts (at least for the countable case)
  - Compute observed terms using such pseudo-counts
- All of the above hinges on the fact that we can compute the exact posterior
  - At some point we will have to deal with how to approximate the posterior

Next Lecture(s)

- A probabilistic model for sequences: Hidden Markov Models (HMMs)
- Exact inference on a chain with observed and unobserved variables
- The Expectation-Maximization algorithm for HMMs
- Graphical models with varying structure: Dynamic Bayesian Networks