Probabilistic Graphical Models in Computer Vision (IN2329)

Csaba Domokos

Summer Semester 2015/2016

3. Introduction to Graphical models

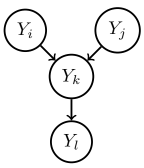
Agenda for today's lecture *

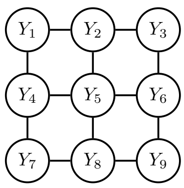
In the **previous lecture** we learnt about

- Expectation-maximization algorithm, which is an iterative method for parameter estimation, where the model also depends on *latent variables*

Today we are going to learn about

1. Expectation-maximization algorithm for mixture of Gaussians
2. Introduction to Graphical models
   - *Directed* graphical models: Bayesian network
   - *Undirected* graphical models: Markov random field





Mixtures of Gaussians

Multivariate Gaussian distribution

Assume a *D-dimensional random vector* $\mathbf{X} = (X_1, \ldots, X_D)$, i.e. a vector whose components are random variables, with the joint density function

$$p(x_1, \ldots, x_D) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) .$$

$\mathbf{X}$ is said to have **multivariate Gaussian (or Normal) distribution** with parameters $\boldsymbol{\mu} \in \mathbb{R}^D$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$ assuming that $\boldsymbol{\Sigma}$ is *positive definite*.

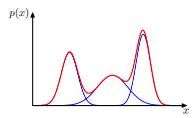
$\boldsymbol{\mu}$ is called the **mean vector** and $\boldsymbol{\Sigma}$ is called the **covariance matrix**. We often use the notation $\mathbf{X} \sim \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denoting $\mathbf{X}$ has Normal distribution.

Reminder: A symmetric $\mathbf{A} \in \mathbb{R}^{n\times n}$ matrix is said to be **positive definite**, if $\mathbf{u}^T\mathbf{A}\mathbf{u} > 0$ for all non-zero $\mathbf{u} \in \mathbb{R}^n$.

Mixtures of Gaussians

While the Gaussian distribution has some important analytical properties, it suffers from limitations when it comes to modelling real data sets. However the **linear combination of Gaussians** can give rise to very complex densities.
Let us consider a superposition of K Gaussian densities

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



Mixture of three Gaussians

which is called a **mixture of Gaussians**.
The parameters π_k are called **mixing coefficients**.

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

All the density functions are non-negative, hence $\pi_k \geqslant 0$ for $1 \leqslant k \leqslant K$, therefore

$$0 \leqslant \pi_k \leqslant 1 \quad \text{for all} \quad k = 1, \ldots, K .$$

Latent variables

We introduce a K-dimensional **binary random variable** z having a *1-of-K representation*, i.e. $z_k = 1$ and all other elements are equal to 0. Let us define the *marginal distribution over z* as

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

which is considered as the *prior probability* of picking the k^{th} component of a mixture of Gaussians. This distribution can be also written in the form

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

Moreover, the conditional distribution of $\mathbf{x}$ given a particular value for $\mathbf{z}$, i.e. *the likelihood*, can be written as

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

Latent variables: responsibilities

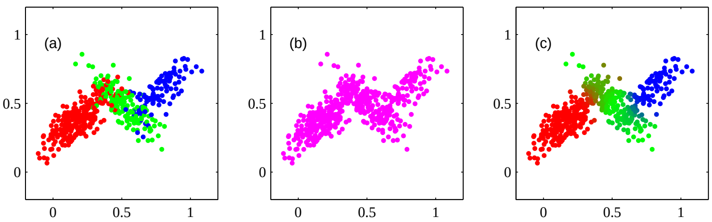
The **distribution of mixture of Gaussian**, specified by the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, is given by

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

The *posterior probabilities* $p(z_k = 1 \mid \mathbf{x})$, denoted by $\gamma_k(\mathbf{x})$, a.k.a. **responsibilities**, show the probability that a given sample $\mathbf{x}$ belongs to the k^{th} component.

$$\gamma_k(\mathbf{x}) \triangleq p(z_k = 1 \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid z_k = 1)p(z_k = 1)}{p(\mathbf{x})} = \frac{p(\mathbf{x} \mid z_k = 1)p(z_k = 1)}{\sum_{l=1}^{K} p(z_l, \mathbf{x})}$$
$$= \frac{p(z_k = 1)p(\mathbf{x} \mid z_k = 1)}{\sum_{l=1}^{K} p(z_l = 1)p(\mathbf{x} \mid z_l = 1)} = \frac{\pi_k \, \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} \pi_l \, \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} .$$

Example: Mixture of three 2D Gaussians *

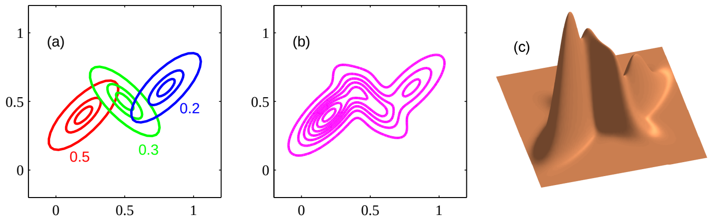


Samples from $p(\mathbf{z} \mid \boldsymbol{\theta})p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\theta})$

Samples from $p(\mathbf{x} \mid \boldsymbol{\theta})$

Values of responsibilities coded by `RGB colors` $[\gamma_1(\mathbf{x}) \quad \gamma_2(\mathbf{x}) \quad \gamma_3(\mathbf{x})]$.

Example: Mixture of three 2D Gaussians *



Iso-countours of each component

Iso-contours of $p(\mathbf{x} \mid \boldsymbol{\theta})$

Surface plot of $p(\mathbf{x} \mid \boldsymbol{\theta})$

Estimation of a mixture of Gaussians

Suppose we have a set of *i.i.d.* data samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ drawn from a mixture of Gaussians. The data set is represented by $\mathbf{X} \in \mathbb{R}^{N \times D}$.

The goal is to find the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, specifying the model from which the samples $\mathbf{x}_n$ have most likely been drawn. We may find the parameters which maximize the *likelihood function* $p(\mathbf{x} \mid \boldsymbol{\theta})$. To simplify the optimization we use the **log-likelihood function** $\mathcal{L}(\boldsymbol{\theta})$

$$\begin{aligned}\hat{\boldsymbol{\theta}} \in \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) &= \operatorname*{argmax}_{\boldsymbol{\theta}} \ \ln p(\mathbf{X} \mid \boldsymbol{\theta}) \overset{i.i.d.}{=} \operatorname*{argmax}_{\boldsymbol{\theta}} \ \ln \prod_{n=1}^{N} p(\mathbf{x}_n \mid \boldsymbol{\theta}) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \ \ln \prod_{n=1}^{N} \prod_{k=1}^{K} \left(\pi_k \ \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{z_{nk}} \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left(\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) .\end{aligned}$$

Note that there is no closed-form solution for this model $\Rightarrow$ iterative solution.

Recall the EM algorithm

1: Choose an initial setting for the parameters $\boldsymbol{\theta}^{(0)}$
2: $t \to 0$
3: **repeat**
4: $t \to t + 1$
5: **E step**. Evaluate $q^{(t-1)}(\mathbf{Z}) \triangleq p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t-1)})$
6: **M step**. Evaluate $\boldsymbol{\theta}^{(t)}$ given by

$$\boldsymbol{\theta}^{(t)} = \operatorname*{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) ,$$

where

$$\begin{aligned}Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) &\triangleq \mathbb{E}[\ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) \mid \mathbf{X}, \boldsymbol{\theta}^{(t)}] \\ &= \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t-1)}) \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})\end{aligned}$$

7: **until** convergence of either the parameters $\boldsymbol{\theta}$ or the log likelihood $\mathcal{L}(\boldsymbol{\theta}; \mathbf{X})$

E step *

We need to calculate $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}})$. It is calculated based on $p(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\theta}^{\text{old}})$ for all $n = 1, \ldots, N$

$$\begin{aligned}p(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\theta}^{\text{old}}) &= \frac{p(\mathbf{x}_n \mid \mathbf{z}_n, \boldsymbol{\theta}^{\text{old}}) \ p(\mathbf{z}_n \mid \boldsymbol{\theta}^{\text{old}})}{p(\mathbf{x}_n \mid \boldsymbol{\theta}^{\text{old}})} \\ &= \frac{\prod_{k=1}^{K} \left(\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{z_{nk}} \ \pi_k^{z_{nk}}}{\sum_{l=1}^{K} \pi_l \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \\ &= \frac{\pi_k \ \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} \pi_l \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \triangleq \gamma_k(\mathbf{x}_n) .\end{aligned}$$

Therefore, in the **E step** we need to calculate the *responsibilities* $\gamma_k(\mathbf{x}_n)$ for all data points $\mathbf{x}_n$ and components $k = 1, \ldots, K$.

M step for μ *

We have already known that $z_{nk} = \gamma_k(\mathbf{x}_n)$. Therefore, we may consider

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k(\mathbf{x}_n) \left(\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \ \text{s.t.} \ \pi_k > 0 , \sum_{k=1}^{K} \pi_k = 1 .$$

We calculate the derivative of $\mathcal{L}(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\mu}_k$

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \gamma_k(\mathbf{x}_n) \frac{1}{\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \frac{\partial}{\partial \boldsymbol{\mu}_k} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) .$$

M step for μ *

Let us now consider the derivative of a Gaussian only

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu}_k} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) &= \frac{1}{\sqrt{|2\pi \boldsymbol{\Sigma}_k|}} \frac{\partial}{\partial \boldsymbol{\mu}_k} \exp\left(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\right) \\ =& \frac{1}{\sqrt{|2\pi \boldsymbol{\Sigma}_k|}} \exp\left(\frac{-1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\right) \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \\ =& \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) .\end{aligned}$$

By substituting back and setting the derivative of $\mathcal{L}(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\mu}_k$ to 0, we get

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \frac{\gamma_k(\mathbf{x}_n)}{\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) = 0$$

$$\frac{\sum_{n=1}^{N} \gamma_k(\mathbf{x}_n) \ \mathbf{x}_n}{\sum_{m=1}^{N} \gamma_k(\mathbf{x}_m)} = \boldsymbol{\mu}_k .$$

M step for Σ *

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k(\mathbf{x}_n) \left(\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \ \text{s.t.} \ \pi_k > 0 , \sum_{k=1}^{K} \pi_k = 1 .$$

Setting the derivative of $\mathcal{L}(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\Sigma}_k$ to 0, one can obtain (see exercise)

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma_k(\mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{m=1}^{N} \gamma_k(\mathbf{x}_m)} .$$

Remark: A $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$ matrix, calculated as

$$\boldsymbol{\Sigma} = \frac{1}{N-1} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^T ,$$

is called **sample covariance matrix** of data points $\{\mathbf{x}_n \in \mathbb{R}^D\}_{n=1}^{N}$, where $\boldsymbol{\mu}$ is the **sample mean**.

M step for π *

To integrate the conditions on $\boldsymbol{\pi}$ we use the **Lagrange multiplier method**

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k(\mathbf{x}_n) \big(\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \big) + \lambda (1 - \sum_{k=1}^{K} \pi_k) \ .$$

Setting the derivative w.r.t. π_k to 0, we obtain

$$\sum_{n=1}^{N} \frac{\gamma_k(\mathbf{x}_n)}{\pi_k} - \lambda = 0$$

$$\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k(\mathbf{x}_n) = \lambda \sum_{k=1}^{K} \pi_k \quad \Rightarrow \quad N = \lambda$$

therefore

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma_k(\mathbf{x}_n)}{N} \ .$$

The EM Algorithm for mixtures of Gaussians

1: Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k for all $k = 1, \dots, K$

2: **repeat**

3: **E step**. Evaluate the responsibilities using the current parameter values

$$\gamma_k(\mathbf{x}_n) = \frac{\pi_k \, \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} \pi_l \, \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \quad \text{for } 1 \leqslant n \leqslant N \text{ and } 1 \leqslant k \leqslant K \ .$$

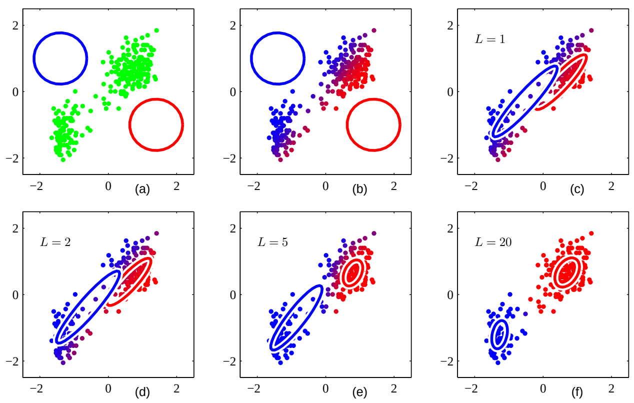
4: **M step**. Re-estimate the parameters $(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ for all $k = 1, \dots, K$

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{\sum_{n=1}^{N} \gamma_k(\mathbf{x}_n) \mathbf{x}_n}{\sum_{m=1}^{N} \gamma_k(\mathbf{x}_m)} \ , \quad \boldsymbol{\Sigma}_k^{\text{new}} = \frac{\sum_{n=1}^{N} \gamma_k(\mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})^T}{\sum_{m=1}^{N} \gamma_k(\mathbf{x}_m)}$$

$$\pi_k^{\text{new}} = \frac{\sum_{n=1}^{N} \gamma_k(\mathbf{x}_n)}{N}$$

5: **until** convergence of either the parameters $\boldsymbol{\theta}$ or the log likelihood $\mathcal{L}(\boldsymbol{\theta})$

Example *



Remarks

- The EM algorithm is **not limited** to mixtures of Gaussians, but it can also be applied to *other probability distributions*.
- The algorithm does **not** necessary yield global maxima. In practice, it is restarted with *different initializations* and the result with the highest log-likelihood after convergence is chosen.
- One can think the EM algorithm as an **alternating minimization** procedure. Considering $f(\boldsymbol{\theta}, q)$ as the objective function, one iteration of the EM algorithm can be reformulated as

$$\text{E-step:} \quad q^{(t+1)} \in \operatorname*{argmax}_{q} f(\boldsymbol{\theta}^{(t)}, q)$$

$$\text{M-step:} \quad \boldsymbol{\theta}^{(t+1)} \in \operatorname*{argmax}_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, q^{(t)})$$

Introduction to Graphical models

Graphical models

Probabilistic graphical models encode a joint $p(\mathbf{x}, \mathbf{y})$ or conditional $p(\mathbf{y} \mid \mathbf{x})$ probability distribution such that given some observations we are provided with a full probability distribution over all feasible solutions.

The graphical models allow us to encode relationships between a set of random variables using a concise language, by means of a graph.

We will use the following notations

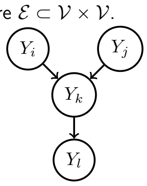
- $\mathcal{V}$ denotes a **set of output variables** (e.g., for pixels) and the corresponding random variables are denoted by Y_i for all $i \in \mathcal{V}$.
- The **output domain** $\mathcal{Y}$ is given by the product of individual variable domains $\mathcal{Y}_i$ (e.g., a single label set $\mathcal{L}$), so that $\mathcal{Y} = \times_{i \in \mathcal{V}} \mathcal{Y}_i$.
- The **input domain** $\mathcal{X}$ is application dependent (e.g., $\mathcal{X}$ is a set of images).
- The **realization** $\mathbf{Y} = \mathbf{y}$ means that $Y_i = y_i$ for all $i \in \mathcal{V}$.
- $G = (\mathcal{V}, \mathcal{E})$ is an (un)directed graph, where $\mathcal{E}$ encodes the conditional independence assumption.

Bayesian networks

Assume a **directed, acyclic** graphical model $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$.

The factorization is given as

$$p(\mathbf{Y} = \mathbf{y}) = \prod_{i \in \mathcal{V}} p(y_i \mid \mathbf{y}_{\text{pa}_G(i)}) \ ,$$



where $p(y_i \mid \mathbf{y}_{\text{pa}_G(i)})$ is a conditional probability distribution on the parents of node $i \in \mathcal{V}$.

The *conditional independence assumption* is encoded by G that is a variable is conditionally independent of its non-descendants given its parents.

For example:

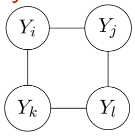
$$\begin{aligned} p(\mathbf{y}) &= p(y_l \mid y_k) \ p(y_k \mid y_i, y_j) \ p(y_i) \ p(y_j) \\ &= p(y_l \mid y_k) \ p(y_k \mid y_i, y_j) \ p(y_i, y_j) = p(y_l \mid y_k) \ p(y_i, y_j, y_k) \\ &= p(y_l \mid y_i, y_j, y_k) \ p(y_i, y_j, y_k) = p(y_i, y_j, y_k, y_l) \ . \end{aligned}$$

Markov random field

Markov random field

An *undirected graphical model* $G = (\mathcal{V}, \mathcal{E})$ is called **Markov Random Field** (MRF) if two nodes are conditionally independent whenever they are not connected. In other words, for any node i in the graph, the **local Markov property** holds:

$$p(Y_i \mid Y_{\mathcal{V}\setminus\{i\}}) = p(Y_i \mid Y_{N(i)}) ,$$



where $N(i)$ is denotes the neighbors of node i in the graph. Alternatively, we use the following equivalent notation:

$$Y_i \perp\!\!\!\perp Y_{\mathcal{V}\setminus \mathrm{cl}(i)} \mid Y_{N(i)} ,$$

where $\mathrm{cl}(i) = N(i) \cup \{i\}$ is the *closed neighborhood* of i.

Example:

$$Y_i \perp\!\!\!\perp Y_l \mid Y_j, Y_k \quad \Rightarrow \quad \begin{aligned} p(y_i \mid y_j, y_k, y_l) &= p(y_i \mid y_j, y_k) , \\ p(y_l \mid y_i, y_j, y_k) &= p(y_l \mid y_j, y_k) . \end{aligned}$$

Gibbs distribution

A *probability distribution* $p(\mathbf{y})$ on an *undirected graphical model* $G = (\mathcal{V}, \mathcal{E})$ is called **Gibbs distribution** if it can be factorized into potential functions $\psi_c(\mathbf{y}_c) > 0$ defined on cliques (i.e. fully connected subgraph) that cover all nodes and edges of G. That is,

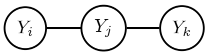
$$p(\mathbf{y}) = \frac{1}{Z} \prod_{c \in \mathcal{C}_G} \psi_c(\mathbf{y}_c) ,$$

where $\mathcal{C}_G$ denotes the set of all (maximal) cliques in G and

$$Z = \sum_{\mathbf{y} \in \mathcal{Y}} \prod_{c \in \mathcal{C}_G} \psi_c(\mathbf{y}_c) .$$

is the normalization constant. Z is also known as **partition function**.

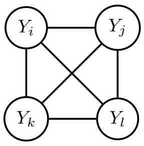
Examples *



G_1

$\mathcal{C}_{G_1} = \{\{i\}, \{j\}, \{k\}, \{i,j\}, \{j,k\}\}$, hence

$$p(\mathbf{y}) = \frac{1}{Z} \psi_i(y_i) \psi_j(y_j) \psi_k(y_k) \psi_{ij}(y_i, y_j) \psi_{jk}(y_j, y_k)$$



G_2

$\mathcal{C}_{G_2} = 2^{\{i,j,k,l\}}$ (i.e. all subsets of $\mathcal{V}_2$)

$$p(\mathbf{y}) = \frac{1}{Z} \prod_{c \in 2^{\{i,j,k,l\}}} \psi_c(\mathbf{y}_c)$$

$$\begin{aligned} 2^{\{i,j,k,l\}} = &\{\{i\}, \{j\}, \{k\}, \{l\}, \\ &\{i,j\}, \{i,k\}, \{i,l\}, \{j,k\}, \{j,l\}, \\ &\{i,j,k\}, \{i,j,l\}, \{i,k,l\}, \{j,k,l\}, \\ &\{i,j,k,l\}\} \end{aligned}$$

Hammersley-Clifford theorem

Let $G = (\mathcal{V}, \mathcal{E})$ be an *undirected graphical model*. The Hammersley-Clifford theorem tells us that the followings are equivalent:

- G is an MRF model.
- The joint probability distribution $p(\mathbf{y})$ on G is a Gibbs-distribution.

An MRF defines a family of **joint probability distributions** by means of an undirected graph $G = (\mathcal{V}, \mathcal{E})$, $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ (there are no self-edges), where the graph encodes *conditional independence assumptions* between the random variables corresponding to $\mathcal{V}$.

Proof of the Hammersley-Clifford theorem (backward direction) *

Let $\mathrm{cl}(i) = N_i \cup \{i\}$ and assume that $p(\mathbf{y})$ follows a *Gibbs-distribution*.

$$p(y_i \mid \mathbf{y}_{N_i}) = \frac{p(y_i, \mathbf{y}_{N_i})}{p(\mathbf{y}_{N_i})} = \frac{\sum_{\mathcal{V}\setminus \mathrm{cl}(i)} p(\mathbf{y})}{\sum_{y_i} \sum_{\mathcal{V}\setminus \mathrm{cl}(i)} p(\mathbf{y})} = \frac{\sum_{\mathcal{V}\setminus \mathrm{cl}(i)} \frac{1}{Z} \prod_{c \in \mathcal{C}_G} \psi_c(\mathbf{y}_c)}{\sum_{x_i} \sum_{\mathcal{V}\setminus \mathrm{cl}(i)} \frac{1}{Z} \prod_{c \in \mathcal{C}_G} \psi_c(\mathbf{y}_c)}$$

Let us define two sets: $\mathcal{C}_i := \{c \in \mathcal{C}_G : i \in c\}$ and $\mathcal{R}_i := \{c \in \mathcal{C}_G : i \notin c\}$.

$$\begin{aligned} &= \frac{\sum_{\mathcal{V}\setminus \mathrm{cl}(i)} \prod_{c \in \mathcal{C}_i} \psi_c(\mathbf{y}_c) \prod_{d \in \mathcal{R}_i} \psi_d(\mathbf{y}_d)}{\sum_{y_i} \sum_{\mathcal{V}\setminus \mathrm{cl}(i)} \prod_{c \in \mathcal{C}_i} \psi_c(\mathbf{y}_c) \prod_{d \in \mathcal{R}_i} \psi_d(\mathbf{y}_d)} \\ &= \frac{\prod_{c \in \mathcal{C}_i} \psi_c(\mathbf{y}_c) \sum_{\mathcal{V}\setminus \mathrm{cl}(i)} \prod_{d \in \mathcal{R}_i} \psi_d(\mathbf{y}_d)}{\sum_{y_i} \prod_{c \in \mathcal{C}_i} \psi_c(\mathbf{y}_c) \sum_{\mathcal{V}\setminus \mathrm{cl}(i)} \prod_{d \in \mathcal{R}_i} \psi_d(\mathbf{y}_d)} \\ &= \frac{\prod_{c \in \mathcal{C}_i} \psi_c(\mathbf{y}_c)}{\sum_{y_i} \prod_{c \in \mathcal{C}_i} \psi_c(\mathbf{y}_c)} \end{aligned}$$

Proof of the Hammersley-Clifford theorem (backward direction) *

$$\begin{aligned} p(y_i \mid \mathbf{y}_{N_i}) &= \frac{\prod_{c \in \mathcal{C}_i} \psi_c(\mathbf{y}_c)}{\sum_{y_i} \prod_{c \in \mathcal{C}_i} \psi_c(\mathbf{y}_c)} \\ &= \frac{\prod_{c \in \mathcal{C}_i} \psi_c(\mathbf{y}_c)}{\sum_{y_i} \prod_{c \in \mathcal{C}_i} \psi_c(\mathbf{y}_c)} \cdot \frac{\prod_{c \in \mathcal{R}_i} \psi_c(\mathbf{y}_c)}{\prod_{c \in \mathcal{R}_i} \psi_c(\mathbf{y}_c)} \\ &= \frac{\prod_{c \in \mathcal{C}_G} \psi_c(\mathbf{y}_c)}{\sum_{y_i} \prod_{c \in \mathcal{C}_G} \psi_c(\mathbf{y}_c)} \\ &= \frac{p(\mathbf{y})}{p(\mathbf{y}_{\mathcal{V}\setminus\{i\}})} = \frac{p(\mathbf{y}_{\mathcal{V}\setminus\{i\}}, y_i)}{p(\mathbf{y}_{\mathcal{V}\setminus\{i\}})} \\ &= p(y_i \mid \mathbf{y}_{\mathcal{V}\setminus\{i\}}) . \end{aligned}$$

Therefore the *local Markov property* holds for any node $i \in \mathcal{V}$.

Binomial theorem *

Reminder: Let $x, y \in \mathbb{R}$ and $n \in \mathbb{N}$, then

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} x^{(n-k)} y^k ,$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

We will use the following identity

$$0 = (1-1)^n = \sum_{k=0}^{n} (-1)^k \binom{n}{k} .$$

Reminder: A k-combination of a set $\mathcal{S}$ is a subset of k distinct elements of $\mathcal{S}$. If $|\mathcal{S}| = n$, then number of k-combinations is equal to $\binom{n}{k}$.

Proof of the Clifford-Hammersley theorem (forward direction) *

We define a *candidate* potential function for any subset $s \subseteq \mathcal{V}$ as follows:

$$f_s(\mathbf{Y}_s = \mathbf{y}_s) = \prod_{z \subseteq s} p(\mathbf{y}_z, \mathbf{y}^*_{\bar{z}})^{(-1^{|s|-|z|})}$$

where $p(\mathbf{y}_z, \mathbf{y}^*_{\bar{z}})$ is a strictly positive distribution. We will use the following notation:

$$q(\mathbf{y}_z) := p(\mathbf{y}_z, \mathbf{y}^*_{\bar{z}}) .$$

Assume that the *local Markov property* holds for any node $i \in \mathcal{V}$.
First, we show that, if s is not a clique, then $f_s(\mathbf{y}_s) = 1$. For this sake, let us assume that s is **not** a clique, therefore there exist $a, b \in s$ that are not connected to each other. Hence

$$f_s(\mathbf{Y}_s = \mathbf{y}_s) = \prod_{z \subseteq s} q(\mathbf{y}_z)^{(-1^{|s|-|z|})} = \prod_{z \subseteq s \setminus \{a,b\}} \left(\frac{q(\mathbf{y}_w) \; q(\mathbf{y}_{w \cup \{a,b\}})}{q(\mathbf{y}_{w \cup \{a\}}) \; q(\mathbf{y}_{w \cup \{b\}})} \right)^{(-1^*)} ,$$

where -1^* meaning either 1 or -1 is not important at all.

Proof of the Clifford-Hammersley theorem (forward direction) *

Let us consider $s \subseteq \mathcal{V}$ such that $s \cap z = \varnothing$, then it can be shown that (see exercise)

$$q(\mathbf{y}_z \mid \mathbf{y}_s) \triangleq p(\mathbf{y}_z \mid \mathbf{y}_s, \mathbf{y}^*_{\mathcal{V}\setminus(z\cup s)}) = \frac{q(\mathbf{y}_z, \mathbf{y}_s)}{q(\mathbf{y}_s)} .$$

We have

$$f_s(\mathbf{Y}_s = \mathbf{y}_s) = \prod_{w \subseteq s\setminus\{a,b\}} \left(\frac{q(\mathbf{y}_w)\; q(\mathbf{y}_{w\cup\{a,b\}})}{q(\mathbf{y}_{w\cup\{a\}})\; q(\mathbf{y}_{w\cup\{b\}})} \right)^{(-1^*)} ,$$

where

$$\frac{q(\mathbf{y}_w)\; q(\mathbf{y}_w, y_a, y_b)}{q(\mathbf{y}_w, y_a)\; q(\mathbf{y}_w, y_b)} = \frac{q(y_a \mid y_b, \mathbf{y}_w)}{q(y_a \mid \mathbf{y}_w)} \overset{a \perp\!\!\!\perp b}{=} \frac{q(y_a \mid \mathbf{y}_w)}{q(y_a \mid \mathbf{y}_w)} = 1 .$$

Proof of the Clifford-Hammersley theorem (forward direction) *

We also show that $\prod_{s\subseteq\mathcal{V}} f_s(\mathbf{y}_s) = p(\mathbf{y})$. Consider any $z \subset \mathcal{V}$ and the corresponding factor $q(\mathbf{y}_z)$. Let $n := |\mathcal{V}| - |z|$.

- $q(\mathbf{y}_z)$ occurs in $f_z(\mathbf{y}_z)$ as $q(\mathbf{y}_z)^{(-1^0)} = q(\mathbf{y}_z)$.
- $q(\mathbf{y}_z)$ also occurs in the functions $f_s(\mathbf{y}_s)$ for $s \subseteq \mathcal{V}$, where $|s| = |z| + 1$. The number of such factors is $\binom{n}{1}$. The exponent of those factors is $-1^{|s|-|z|} = -1^1 = -1$.
- $q(\mathbf{y}_z)$ occurs in the functions $f_s(\mathbf{y}_s)$ for $s \subseteq \mathcal{V}$, where $|s| = |z| + 2$. The number of such factors is $\binom{n}{2}$ and their exponent is $-1^{|s|-|z|} = 1$.

If we multiply **all** those factors, we get

$$q(\mathbf{y}_z)^1\; q(\mathbf{y}_z)^{-\binom{n}{1}}\; q(\mathbf{y}_z)^{\binom{n}{2}}\; \ldots\; q(\mathbf{y}_z)^{(-1^n)\binom{n}{n}} = q(\mathbf{y}_z)^{\binom{n}{0}-\binom{n}{1}+\binom{n}{2}+\cdots+(-1)^n\binom{n}{n}}$$
$$= q(\mathbf{y}_z)^0 = 1 .$$

So all factors cancel themselves out except of $q(\mathbf{y}) = p(\mathbf{y})$. □

Summary *

- A **graphical models** allow us to *encode relationships between a set of random variables* using a concise language, by means of a graph.
- A **Bayesian network** is a *directed acyclic* graphical model $G = (\mathcal{V}, \mathcal{E})$, where conditional independence assumption is encoded by G that is a variable is conditionally independent of its non-descendants given its parents.
- An **MRF** defines a family of **joint probability distributions** by means of an undirected graph $G = (\mathcal{V}, \mathcal{E})$, where the graph encodes conditional independence assumptions between the random variables.

In the **next lecture** we will learn about

- Conditional random fields (CRF)
- Binary image segmentation

Literature *

The EM algorithm for Mixtures of Gaussians

1. Frank Dellaert. The expectation maximization algorithm. Technical Report GIT-GVU-02-20, Georgia Institute of Technology, Atlanta, GA, USA, 2002
2. Shane M. Haas. The expectation-maximization and alternating minimization algorithms. Unpublished, 2002
3. Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006

Graphical models

4. Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009
5. Sebastian Nowozin and Christoph H. Lampert. Structured prediction and learning in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 6(3–4), 2010
6. J. M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. Unpublished, 1971
7. Samson Cheung. Proof of hammersley-clifford theorem. Unpublished, February 2008