

Probabilistic Graphical Models in Computer Vision (IN2329)

Csaba Domokos

Summer Semester 2015/2016

3. Introduction to Graphical models	2
Agenda for today's lecture *	3
Mixtures of Gaussians	4
Multivariate Gaussian distribution	5
Mixtures of Gaussians	6
Latent variables	7
Latent variables: responsibilities	8
Example: Mixture of three 2D Gaussians *	9
Example: Mixture of three 2D Gaussians *	10
Estimation of a mixture of Gaussians	11
Recall the EM algorithm.	12
E step *	13
M step for μ *	14
M step for μ *	15
M step for Σ *	16
M step for π *	17
The EM Algorithm for mixtures of Gaussians.	18
Example *	19

Remarks	20
Introduction to Graphical models.	21
Graphical models.	22
Bayesian networks	23
Bayesian networks	24
Markov random field	25
Markov random fields	26
Gibbs distribution	27
Examples *	28
Hammersley-Clifford theorem	29
Proof of the Hammersley-Clifford theorem (backward direction) *	30
Proof of the Hammersley-Clifford theorem (backward direction) *	31
Binomial theorem	32
Proof of the Clifford-Hammersley theorem (forward direction) *	33
Proof of the Clifford-Hammersley theorem (forward direction) *	34
Proof of the Clifford-Hammersley theorem (forward direction) *	35
Summary *	36
Literature *	37

3. Introduction to Graphical models

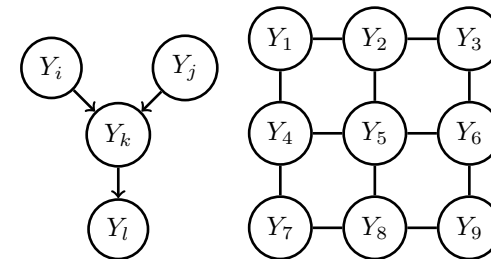
Agenda for today's lecture *

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



Multivariate Gaussian distribution

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

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

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

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

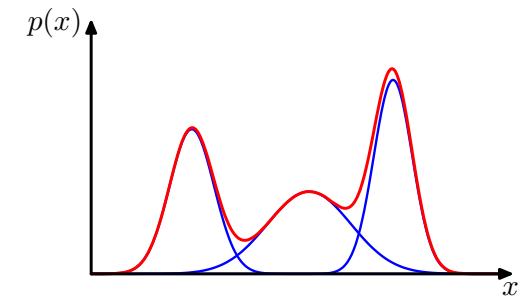
While the Gaussian distribution has some important analytical properties, it suffers from limitations when it comes to modelling real data sets. However the **linear combination of Gaussians** can give rise to very complex densities.

Let us consider a superposition of K Gaussian densities

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

which is called a **mixture of Gaussians**.

The parameters π_k are called **mixing coefficients**.



Mixture of three Gaussians

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

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

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

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

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

Latent variables: responsibilities

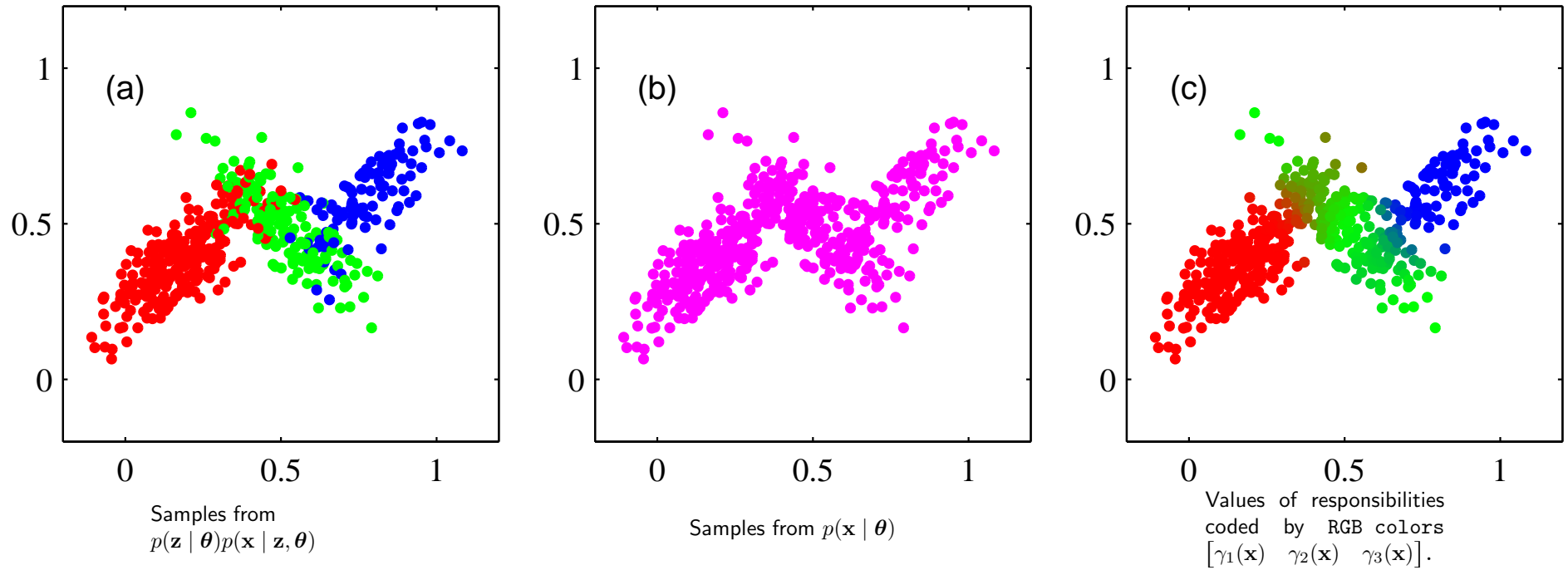
The **distribution of mixture of Gaussian**, specified by the parameter vector $\theta = (\pi, \mu, \Sigma)$, is given by

$$\begin{aligned} p(\mathbf{x}) &\triangleq p(\mathbf{x} | \theta) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z} | \theta) = \sum_{\mathbf{z}} p(\mathbf{z} | \theta) p(\mathbf{x} | \mathbf{z}, \theta) \\ &= \sum_{\mathbf{z}} \prod_{k=1}^K (\pi_k p(\mathbf{x} | \mu_k, \Sigma_k))^{z_k} = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k) . \end{aligned}$$

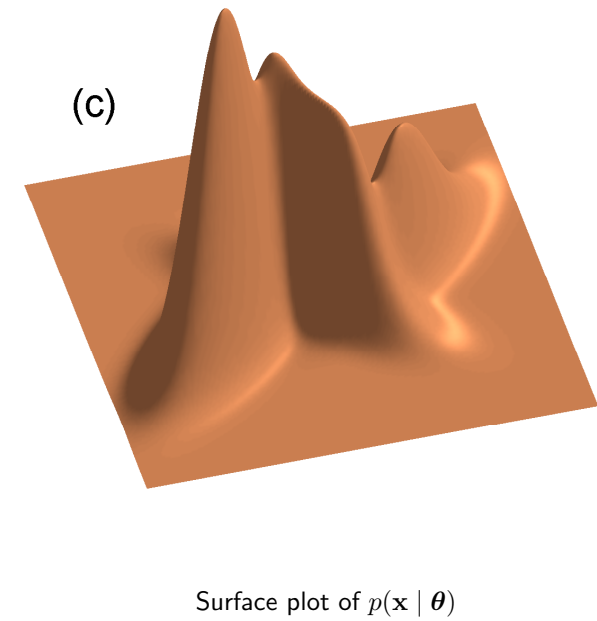
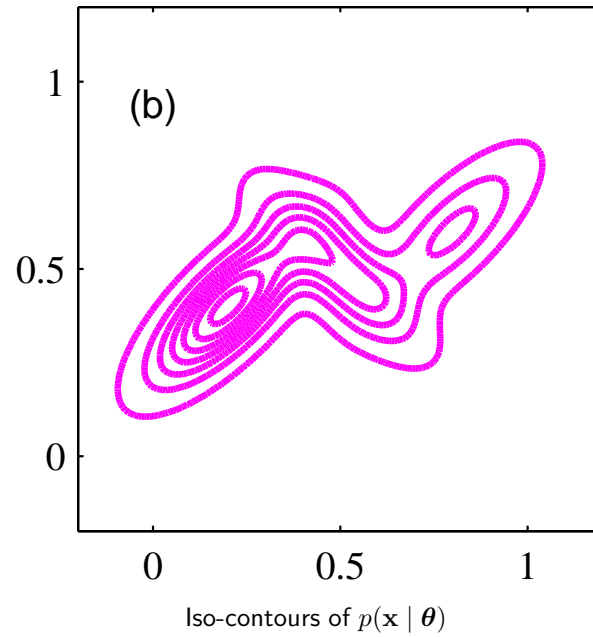
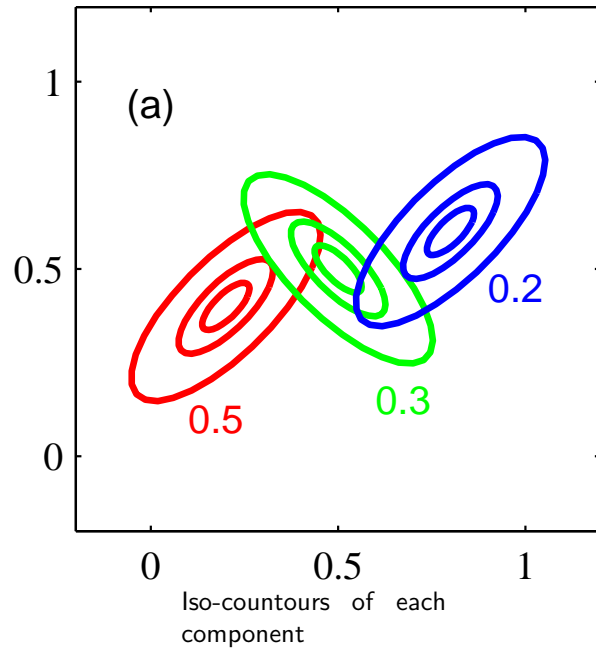
The *posterior probabilities* $p(z_k = 1 | \mathbf{x})$, denoted by $\gamma_k(\mathbf{x})$, a.k.a. **responsibilities**, show the probability that a given sample \mathbf{x} belongs to the k^{th} component.

$$\begin{aligned} \gamma_k(\mathbf{x}) &\triangleq p(z_k = 1 | \mathbf{x}) = \frac{p(\mathbf{x} | z_k = 1) p(z_k = 1)}{p(\mathbf{x})} = \frac{p(\mathbf{x} | z_k = 1) p(z_k = 1)}{\sum_{l=1}^K p(z_l, \mathbf{x})} \\ &= \frac{p(z_k = 1) p(\mathbf{x} | z_k = 1)}{\sum_{l=1}^K p(z_l = 1) p(\mathbf{x} | z_l = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k)}{\sum_{l=1}^K \pi_l \mathcal{N}(\mathbf{x} | \mu_l, \Sigma_l)} . \end{aligned}$$

Example: Mixture of three 2D Gaussians *



Example: Mixture of three 2D Gaussians *



Estimation of a mixture of Gaussians

Suppose we have a set of *i.i.d.* data samples $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ 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* $p(\mathbf{x} | \boldsymbol{\theta})$. To simplify the optimization we use the **log-likelihood function** $\mathcal{L}(\boldsymbol{\theta})$

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

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 $\theta^{(0)}$
- 2: $t \rightarrow 0$
- 3: **repeat**
- 4: $t \rightarrow t + 1$
- 5: **E step.** Evaluate $q^{(t-1)}(\mathbf{Z}) \triangleq p(\mathbf{Z} | \mathbf{X}, \theta^{(t-1)})$
- 6: **M step.** Evaluate $\theta^{(t)}$ given by

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

where

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

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

E step *

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

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

Therefore, in the **E step** we need to calculate the *responsibilities* for all data points \mathbf{x}_n .

M step for μ^*

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

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

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

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

M step for μ^*

Let us now consider the derivative of a Gaussian only

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

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

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\mu}_k} \mathcal{L}(\theta) &= \sum_{n=1}^N \frac{\gamma_k(\mathbf{x}_n)}{\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \\ & \qquad \qquad \qquad \frac{\sum_{n=1}^N \gamma_k(\mathbf{x}_n) \mathbf{x}_n}{\sum_{n=1}^N \gamma_k(\mathbf{x}_n)} = \boldsymbol{\mu}_k. \end{aligned}$$

M step for Σ *

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

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

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

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 the 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 π we use the **Lagrange multiplier method**

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

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

$$\begin{aligned} \sum_{n=1}^N \frac{\gamma_k(\mathbf{x}_n)}{\pi_k} - \lambda &= 0 \\ \sum_{n=1}^N \sum_{k=1}^K \gamma_k(\mathbf{x}_n) &= \lambda \sum_{k=1}^K \pi_k \quad \Rightarrow \quad N = \lambda \end{aligned}$$

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

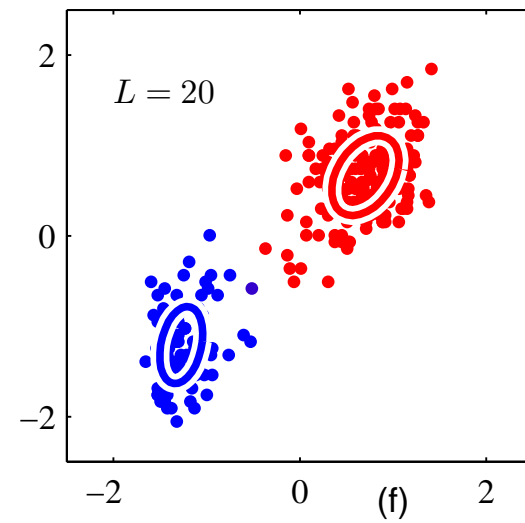
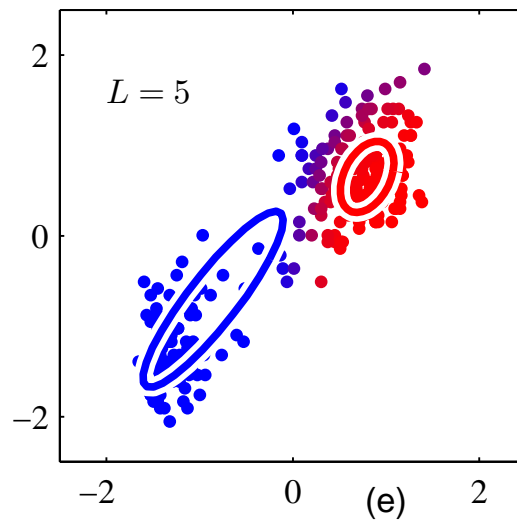
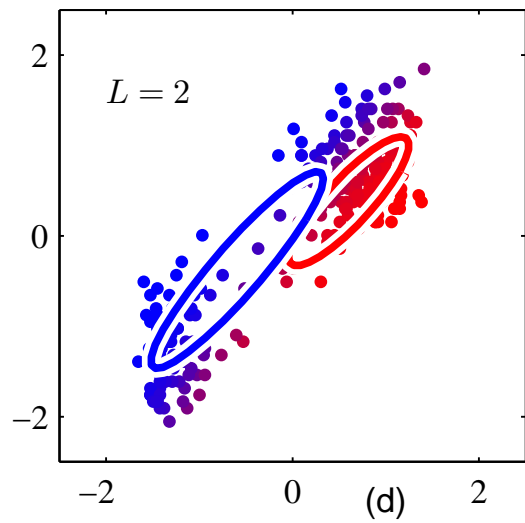
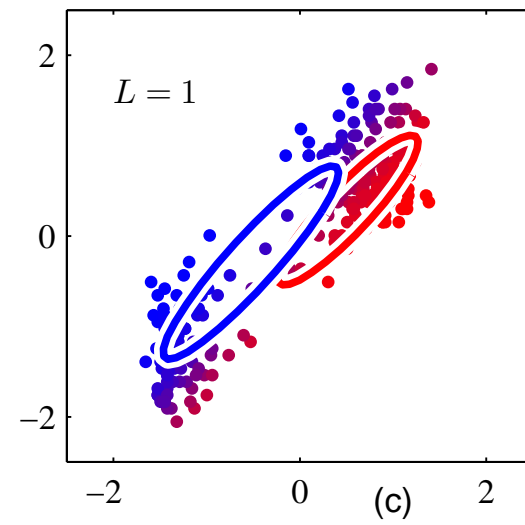
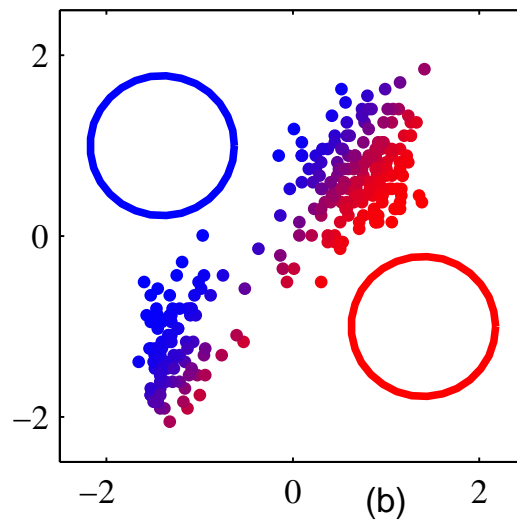
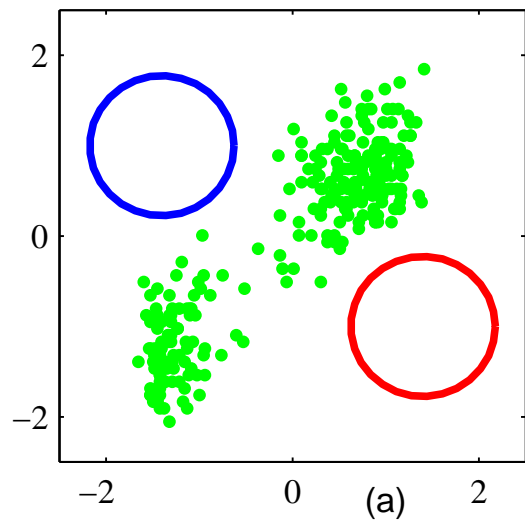
$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^K \pi_l \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \quad \text{for } 1 \leq n \leq N \text{ and } 1 \leq k \leq K .$$

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

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

- 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** necessarily 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

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

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

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 , $i \in V$.
- 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 $\mathcal{Y} = \times_{i \in V} \mathcal{Y}_i$.
- 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 $i \in V$.
- $G = (\mathcal{V}, \mathcal{E})$ is an (un)directed graph, where \mathcal{E} encodes the conditional independence assumption.

Bayesian networks

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

The factorization is given as

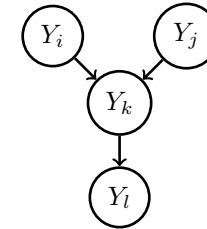
$$p(\mathbf{Y} = \mathbf{y}) = \prod_{i \in V} p(y_i \mid \mathbf{y}_{\text{pa}_G(i)}),$$

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

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 fields

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:

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

where $N(i)$ is denotes the neighbors of node i in the graph.

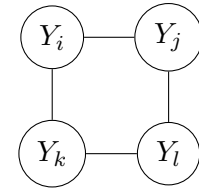
Alternatively, we use the following equivalent notation:

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

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

Example:

$$Y_i \perp\!\!\!\perp Y_l | Y_j, Y_k \Rightarrow p(Y_i, Y_l | Y_j, Y_k) = p(Y_i | Y_j, Y_k) p(Y_l | Y_j, Y_k) .$$



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 *

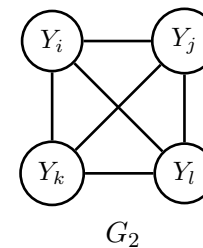
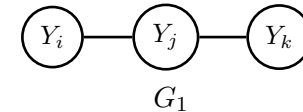
$\mathcal{C}_{G_1} = \{\{i\}, \{j\}, \{k\}, \{i, j\}, \{j, k\}\}$, hence

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

$\mathcal{C}_{G_2} = 2^{\{i, j, k, l\}}$ (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 following 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 $\text{cl}(i) = N_i \cup \{i\}$ and assume that $p(\mathbf{y})$ follows a Gibbs-distribution.

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

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

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

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

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

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 S is a subset of k distinct elements of S . If $|S| = n$, 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:

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

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

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

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

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

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 $s \cap z = \emptyset$, then it can be shown that (see exercise)

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

We have

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

where

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

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

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

- $q(\mathbf{y}_z)$ 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

$$\begin{aligned} q(\mathbf{y}_z)^1 q(\mathbf{y}_z)^{-\binom{n}{1}} q(\mathbf{y}_z)^{\binom{n}{2}} \dots q(\mathbf{y}_z)^{(-1)^n \binom{n}{n}} &= q(\mathbf{y}_z)^{\binom{n}{0} - \binom{n}{1} + \binom{n}{2} + \dots + (-1)^n \binom{n}{n}} \\ &= q(\mathbf{y}_z)^0 = 1. \end{aligned}$$

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

3. Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009
4. 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
5. J. M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. Unpublished, 1971
6. Samson Cheung. Proof of hammersley-clifford theorem. Unpublished, February 2008