Probabilistic Graphical Models in Computer Vision (IN2329)

Csaba Domokos

Summer Semester 2017

2. Graphical models ................................................................. 2
Agenda for today’s lecture * .................................................. 3
\(\sigma\)-algebra, measure, measure space * .................................. 4
Probability space * ................................................................. 5

Random variables ................................................................. 6
  Example: throwing two “fair” dice * ....................................... 7
  Preimage mapping .............................................................. 8
  Random variable ............................................................... 9
  Example: throwing two “fair” dice * .................................... 10
  Labeling via random variables ........................................... 11

Probability distributions ......................................................... 12
  Probability distribution .................................................... 13
  Joint distribution ............................................................ 14
  Marginal distributions ..................................................... 15
  Example: marginal distribution * ........................................ 16
  Conditional distribution .................................................... 17
  Summary * ........................................................................ 18
Agenda for today’s lecture *

In the previous lecture we learnt about

- Discrete probability space
- Conditional probability
- Independence, conditional independence

Today we are going to learn about

1. Random variables \((Y_1, \ldots, Y_9)\)
2. Probability distributions
   - Joint distribution \(p(y_1, \ldots, y_9)\)
   - Marginal distribution \(p(y_1)\)
   - Conditional distribution \(p(y | x)\)
3. Graphical models
\[ \sigma\text{-algebra, measure, measure space} \]

Assume an arbitrary set \( \Omega \) and \( \mathcal{A} \subseteq \mathcal{P}(\Omega) \). The set \( \mathcal{A} \) is a \textit{\( \sigma \)-algebra over} \( \Omega \) if the following conditions are satisfied:

1. \( \emptyset \in \mathcal{A} \),
2. \( A \in \mathcal{A} \Rightarrow \overline{A} \in \mathcal{A} \) (i.e. it is closed under complementation),
3. \( A_i \in \mathcal{A} \ (i \in \mathbb{N}) \Rightarrow \bigcup_{i=0}^{\infty} A_i \in \mathcal{A} \) (i.e. it is closed under countable union).

It is a consequence of this definition that \( \Omega \in \mathcal{A} \) is also satisfied. (See exercise.)

Assume an \textit{arbitrary set} \( \Omega \) and a \textit{\( \sigma \)-algebra} \( \mathcal{A} \) over \( \Omega \). A function \( P : \mathcal{A} \to [0, \infty] \) is called a \textit{measure} if the following conditions are satisfied:

1. \( P(\emptyset) = 0 \),
2. \( P \) is \( \sigma \)-additive.

Let \( \mathcal{A} \) be a \textit{\( \sigma \)-algebra over} \( \Omega \) and \( P : \mathcal{A} \to [0, \infty] \) is a \textit{measure}. \( (\Omega, \mathcal{A}, P) \) is said to be a \textit{measurable space} and the triple \( (\Omega, \mathcal{A}, P) \) is called a \textit{measure space}.
Probability space *

A **probability space** is a triple $(\Omega, \mathcal{A}, P)$, where $(\Omega, \mathcal{A})$ is a measurable space, and $P$ is a measure such that $P(\Omega) = 1$, called a **probability measure**.

To summarize:
A triple $(\Omega, \mathcal{A}, P)$ is called **probability space**, if
- the **sample space** $\Omega$ is not empty,
- $\mathcal{A}$ is a $\sigma$-algebra over $\Omega$, and
- $P : \mathcal{A} \to \mathbb{R}$ is a function with the following properties:
  1. $P(A) \geq 0$ for all $A \in \mathcal{A}$
  2. $P(\Omega) = 1$
  3. $\sigma$-additive: if $A_n \in \mathcal{A}$, $n = 1, 2, \ldots$ and $A_i \cap A_j = \emptyset$ for $i \neq j$, then

\[
P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n).
\]
Example: throwing two “fair” dice

We have the sample space \( \Omega = \{(i, j) : 1 \leq i, j \leq 6 \} \) and the (uniform) probability measure \( P(\{(i, j)\}) = \frac{1}{36} \), where \((\Omega, \mathcal{P}(\Omega), P)\) forms a (discrete) probability space.

In many cases it would be more natural to consider attributes of the outcomes. A random variable is a way of reporting an attribute of the outcome.

Let us consider the sum of the numbers showing on the dice, defined by the mapping \( X : \Omega \rightarrow \Omega', \ X(i, j) = i + j \), where \( \Omega' = \{2, 3, \ldots, 12\} \).

It can be seen that this mapping leads a probability space \((\Omega', \mathcal{P}(\Omega'), P')\), such that \( P' : \mathcal{P}(\Omega') \rightarrow [0, 1] \) is defined as

\[
P'(A') = P(\{(i, j) : X(i, j) \in A'\}) .
\]

Example: \( P'(\{11\}) = P(\{(5, 6), (6, 5)\}) = \frac{2}{36} \).
**Preimage mapping**

Let $X : \Omega \rightarrow \Omega'$ be an arbitrary mapping. The *preimage mapping* $X^{-1} : \mathcal{P}(\Omega') \rightarrow \mathcal{P}(\Omega)$ is defined as

$$X^{-1}(A') = \{ \omega \in \Omega : X(\omega) \in A' \}.$$ 

![Diagram of preimage mapping]

**Random variable**

Let $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ measurable spaces. A mapping $X : (\Omega, \mathcal{A}) \rightarrow (\Omega', \mathcal{A}')$ is called *random variable*, if

$$X^{-1}(A') = \{ \omega \in \Omega : X(\omega) \in A' \} \in \mathcal{A}.$$ 

Let $X : (\Omega, \mathcal{A}) \rightarrow (\Omega' \subseteq \mathbb{R}, \mathcal{A}')$ be a random variable and $P$ a measure over $\mathcal{A}$. Then

$$P'(A') := P_X(A') \triangleq P(X^{-1}(A'))$$

defines a measure over $\mathcal{A}'$. $P_X$ is called the *image measure* of $P$ by $X$.

Specially, if $P$ is a probability measure then $P_X$ is a probability measure over $\mathcal{A}'$. (See Exercise.)
Example: throwing two “fair” dice

We are given two sample spaces \( \Omega = \{(i, j) : 1 \leq i, j \leq 6 \} \) and \( \Omega' = \{2, 3, \ldots, 12\} \). We assume the (uniform) probability measure \( P \) over \( (\Omega, \mathcal{P}(\Omega)) \). Let us define a mapping \( X : (\Omega, \mathcal{P}(\Omega)) \rightarrow (\Omega', \mathcal{P}(\Omega')) \), where \( X(i, j) = i + j \).

Question: Is \( X \) a random variable?

\[
X^{-1}(A') = \{ \omega \in \Omega : X(\omega) \in A' \} \in \mathcal{P}(\Omega)
\]

is satisfied, since for any \( \omega' \in \Omega' \) one can find an \( \omega \in \Omega \) such that \( X(\omega) = \omega' \). Therefore \( X \) is a random variable. Moreover, \( P \) is a probability measure, hence the image measure

\[
P_X(A') = P(X^{-1}(A'))
\]

is a probability measure on \( (\Omega', \mathcal{P}(\Omega')) \).

Example: \( P_X(\{2, 4, 5\}) = P(X^{-1}(\{2, 4, 5\})) = P(\{(1, 1), (1, 3), (2, 2), (3, 1), (1, 4), (2, 3), (3, 2), (4, 1)\}) = \frac{8}{36} = \frac{2}{9} \).

Labeling via random variables

In the last lecture we defined the labeling \( L \) providing a label, taken from a label set \( \mathcal{L} \), for each pixel \( i \) on an image.

By applying a random variable

\[
X : \{(r, g, b) \in \mathbb{Z}^3 : 0 \leq r, g, b \leq 255\} \rightarrow \mathcal{L}
\]

we can model the probability of the labeling for a given pixel as

\[
P_X(\text{the given pixel has the label } l).
\]
**Probability distributions**

**Probability distribution**

Note that a random variable is a (measurable) **mapping** from a probability space to a measure space. It is **neither a variable nor random**.

Let \( X : (\Omega, \mathcal{A}, P) \rightarrow (\Omega' \subseteq \mathbb{R}, \mathcal{A}') \) be a random variable. Then the image measure \( P_X \) of \( P \) by \( X \) is called **probability distribution**.

We use the notation \( p(x) \) for \( P(X = x) \), where
\[
p(x) := P(X = x) \triangleq P(\{ \omega \in \Omega : X(\omega) = x \}).
\]

**Joint distribution**

Suppose a probability space \( (\Omega, \mathcal{A}, P) \). Let \( X : (\Omega, \mathcal{A}) \rightarrow (\Omega', \mathcal{A}') \) and \( Y : (\Omega, \mathcal{A}) \rightarrow (\Omega'', \mathcal{A}'') \) be discrete random variables, where \( x_1, x_2, \ldots \) denote the values of \( X \) and \( y_1, y_2, \ldots \) denote the values of \( Y \).

We introduce the notation
\[
p_{ij} \triangleq P(X = x_i, Y = y_j) \quad i, j = 1, 2, \ldots
\]
for the probability of the events
\[
\{ X = x_i, Y = y_j \} := \{ \omega \in \Omega : X(\omega) = x_i \text{ and } Y(\omega) = y_j \}.
\]

These probabilities \( p_{ij} \) form a **distribution**, called the **joint distribution** of \( X \) and \( Y \).

Remark that
\[
\sum_{i} \sum_{j} p_{ij} = 1.
\]
Marginal distributions

Suppose a probability space \((\Omega, \mathcal{A}, P)\). Let \(X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')\) and \(Y : (\Omega, \mathcal{A}) \to (\Omega'', \mathcal{A}'')\) be discrete random variables, where \(x_1, x_2, \ldots\) denote the values of \(X\) and \(y_1, y_2, \ldots\) denote the values of \(Y\).

The distributions defined by the probabilities

\[
p_i \triangleq P(X = x_i) \quad \text{and} \quad q_j \triangleq P(Y = y_j)
\]

are called the marginal distributions of \(X\) and of \(Y\), respectively.

Let us consider the marginal distribution of \(X\). Then

\[
p_i = P(X = x_i) = \sum_j P(X = x_i, Y = y_j) = \sum_j p_{ij}.
\]

Similarly, the marginal distribution of \(Y\) is given by

\[
q_j = P(Y = y_j) = \sum_i P(X = x_i, Y = y_j) = \sum_i p_{ij}.
\]
**Example: marginal distribution**

Consider the problem of *binary segmentation*. Let us define a pixel to be “bright”, if all its (RGB) intensities are at least 128, otherwise the given pixel is considered to be “dark”.

Assume we are given the following table with probabilities:

<table>
<thead>
<tr>
<th></th>
<th>Dark</th>
<th>Bright</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreground</td>
<td>0.163</td>
<td>0.006</td>
</tr>
<tr>
<td>Background</td>
<td>0.116</td>
<td>0.715</td>
</tr>
<tr>
<td></td>
<td>0.279</td>
<td>0.721</td>
</tr>
</tbody>
</table>

The marginal distributions of discrete random variables corresponding to the values of \{foreground, background\} and \{dark, bright\} are shown in the last column and last row, respectively.

The following also holds

\[
\sum_i p_i = \sum_i P(X = x_i) = \sum_i \sum_j P(X = x_i, Y = y_i) = \sum_i \sum_j p_{ij} = 1.
\]
Conditional distribution

Let $X$ and $Y$ be discrete random variables, where $x_1, x_2, \ldots$ denote the values of $X$ and $y_1, y_2, \ldots$ denote the values of $Y$.

The conditional distribution of $X$ given $Y$ is defined by

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{\sum_k p_{kj}} = \frac{p_{ij}}{q_j}.$$
Probabilistic graphical models encode a joint $p(x, y)$ or conditional $p(y \mid x)$ probability distribution such that given some observations $x$ 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 $\mathcal{L}$), that is $\mathcal{Y} = \times_{i \in \mathcal{V}} \mathcal{Y}_i$.
- The input domain $\mathcal{X}$ is application dependent (e.g., $\mathcal{X}$ is a set of images).
- The realization $Y = y$ means that $Y_i = y_i$ for all $i \in \mathcal{V}$.
- $G = (\mathcal{V}, \mathcal{E})$ is an (un)directed graph, which encodes the conditional independence assumption.

The factorization is given as

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

where $p(y_i \mid y_{\text{pa}_G(i)})$, assuming that $p(y_i \mid \emptyset) \equiv p(y_i)$, is a conditional probability distribution on the parents of node $i \in \mathcal{V}$, denoted by $\text{pa}_G(i)$.

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

**Example:**

- $p(y) = p(y_i \mid y_k) p(y_k \mid y_i, y_j) p(y_i) p(y_j) = p(y_i \mid y_k) p(y_i) p(y_j) = p(y_i, y_j, y_k) = p(y_i, y_j, y_k)$.
Markov random field

An undirected graphical model $G = (\mathcal{V}, \mathcal{E})$ is called a 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 can use the following equivalent notation:

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

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

**Example:**

$$Y_i \perp Y_j \mid Y_j, Y_k \implies p(y_i \mid y_j, y_k) = p(y_i \mid y_j, y_k), \text{ or}$$

$$p(y_i \mid y_j, y_k) = p(y_i \mid y_j, y_k) .$$
**Gibbs distribution**

A probability distribution $p(y)$ on an undirected graphical model $G = (V, E)$ is called **Gibbs distribution** if it can be factorized into potential functions

$$\psi_c(y_c) > 0$$

defined on cliques (i.e. fully connected subgraph) that cover all nodes and edges of $G$. That is,

$$p(y) = \frac{1}{Z} \prod_{c \in C_G} \psi_c(y_c)$$

where $C_G$ denotes the set of all (maximal) cliques in $G$ and

$$Z = \sum_{y \in Y} \prod_{c \in C_G} \psi_c(y_c)$$

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

\[ C_{G_1} = \{\{i\}, \{j\}, \{k\}, \{i, j\}, \{j, k\}\}, \text{ hence} \]
\[
P(y) = \frac{1}{Z} \psi_i(y_i) \psi_j(y_j) \psi_k(y_k) \psi_{ij}(y_i, y_j) \psi_{jk}(y_j, y_k)
\]

\[ C_{G_2} = 2^{\{i, j, k, l\}\setminus\emptyset} \text{ (i.e. all nonempty subsets of } V_2) \]
\[
P(y) = \frac{1}{Z} \prod_{c \in 2^{\{i, j, k, l\}\setminus\emptyset}} \psi_c(y_c)
\]

Hammersley-Clifford theorem

Let \( G = (V, E) \) 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 \( p(y) \) on \( G \) is a Gibbs-distribution.

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

Let \( \text{cl}(i) = N_i \cup \{i\} \) and assume that \( p(y) \) follows Gibbs-distribution.

\[
p(y_i \mid y_{N_i}) = \frac{p(y_i, y_{N_i})}{p(y_{N_i})} = \frac{\sum_{y \in \mathcal{Y}} p(y) \prod_{c \in \text{cl}(i)} \psi_c(y_c)}{\sum_{y \in \mathcal{Y}} \prod_{c \in \text{cl}(i)} \psi_c(y_c)} = \frac{\sum_{y \in \mathcal{Y}} \prod_{c \in \text{cl}(i)} \psi_c(y_c)}{\sum_{y \in \mathcal{Y}} \prod_{c \in \text{cl}(i)} \psi_c(y_c)}.
\]

Let us define two sets: \( C_i := \{c \in C_G : i \in c\} \) and \( R_i := \{c \in C_G : i \notin c\} \). Obviously, \( C_G = C_i \cup R_i \) for all \( i \in \mathcal{V} \).

Example:

\[
\begin{align*}
p(y_i \mid y_{N_i}) &= \frac{\sum_{y \in \mathcal{Y}} \prod_{c \in C_i} \psi_c(y_c) \prod_{d \in R_i} \psi_d(y_d)}{\sum_{y \in \mathcal{Y}} \prod_{c \in \text{cl}(i)} \psi_c(y_c)} = \frac{\prod_{c \in C_i} \psi_c(y_c) \cdot \sum_{y \in \mathcal{Y}} \prod_{d \in R_i} \psi_d(y_d)}{\sum_{y \in \mathcal{Y}} \prod_{c \in C_i} \psi_c(y_c)}
\end{align*}
\]

\[C_i = \{(i, j), (i, k)\}\]
\[R_i = \{(j, l), (k, l)\}\]
Proof of the Hammersley-Clifford theorem (backward direction) *

\[
p(y_i \mid y_{N_i}) = \frac{\prod_{c \in C} \psi_c(y_c)}{\sum_{y_i} \prod_{c \in C} \psi_c(y_c)} = \frac{\prod_{c \in C} \psi_c(y_c)}{\sum_{y_i} \prod_{c \in C} \psi_c(y_c)} \cdot \frac{\prod_{c \in R} \psi_c(y_c)}{\prod_{c \in R} \psi_c(y_c)} = \frac{\prod_{c \in C} \psi_c(y_c)}{\sum_{y_i} \prod_{c \in C} \psi_c(y_c)} = \frac{p(y)}{p(y_{V \setminus \{i\}})} = \frac{p(y_i, y_{V \setminus \{i\}})}{p(y_{V \setminus \{i\}})} = p(y_i \mid y_{V \setminus \{i\}}).
\]

Therefore the local Markov property holds for any node \( i \in V \).
Binomial theorem *

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

$$(x + y)^n = \sum_{k=0}^{n} \begin{pmatrix} n \\ k \end{pmatrix} x^{(n-k)} y^k,$$

where $\begin{pmatrix} n \\ k \end{pmatrix} = \frac{n!}{k!(n-k)!}$.

We will use the following identity

$$0 = (1 - 1)^n = \sum_{k=0}^{n} (-1)^k \begin{pmatrix} n \\ k \end{pmatrix}.$$

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 $\begin{pmatrix} n \\ k \end{pmatrix}$.

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

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

$$f_s(Y_s = y_s) = \prod_{z \in s} p(y_z; y^*_z)^{(-1)^{|s|-|z|}}$$

where $p(y_z; y^*_z)$ is a strictly positive distribution and $y^*_z$ means an (arbitrary but fixed) default realization of the variables $Y_z$ for the set $\bar{z} = \mathcal{V}\setminus\{z\}$. We will use the following notation:

$$q(y_z) := p(y_z; y^*_z).$$

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 $f_s(y_s) = 1$. 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

$$f_s(Y_s = y_s) = \prod_{z \in s} q(y_z)^{(-1)^{|s|-|z|}} = \prod_{w \in s \setminus \{a, b\}} \left( \frac{q(y_w) q(Y_{w\cup\{a,b\}})}{q(Y_{w\cup\{a\}}} \right)^{(-1)^a},$$

where $-1^a$ meaning either 1 or -1 is not important at all.
Proof of the Clifford-Hammersley theorem (forward direction) *

We have

\[ f_s(Y_s = y_s) = \prod_{w \subseteq s \setminus \{a, b\}} \left( \frac{q(Y_w) q(Y_{w \cup \{a, b\}}) \cdots q(Y_{w \cup \{b\}})}{q(Y_{w \cup \{a\}}) q(Y_{w \cup \{b\}})} \right)^{-1}. \]

Therefore

\[ f_s(Y_s = y_s) = \prod_{w \subseteq s \setminus \{a, b\}} 1^{-1} = 1 \quad \text{for all } s \notin C_G. \]

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

We also show that \( \prod_{s \subseteq V} f_s(y_s) = p(y) \). Consider any \( z \subseteq V \) and the corresponding factor \( q(y_z) \). Let \( n = |V| - |z| \).

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

If we multiply all those factors, we get

\[ q(y_z)^{-1} q(y_z)^{-1} \cdots q(y_z)^{-1} q(y_z)^{2} \cdots q(y_z)^{n} = q(y_z)^{-\binom{n}{1} + 1 + \cdots + \binom{n}{2}} = q(y_z)^{0} = 1. \]

So all factors cancel themselves out except of \( q(y) \), that is \( p(y) = \prod_{s \subseteq C_G} f_c(y_c) \).

\( \square \)
Factor graphs are *undirected graphical models* that **make the factorization explicit** of the probability function.

A factor graph $G = (\mathcal{V}, \mathcal{F}, \mathcal{E}')$ consists of:

- variable nodes $\mathcal{V}$ (○) and factor nodes $\mathcal{F}$ (■),
- edges $\mathcal{E}' \subseteq \mathcal{V} \times \mathcal{F}$ between variable and factor nodes,
- $N : \mathcal{F} \rightarrow 2^\mathcal{V}$ is the **scope of a factor**, defined as the set of neighboring variables, i.e. $N(F) = \{i \in \mathcal{V} : (i, F) \in \mathcal{E}\}$.

A family of distribution is defined that factorizes as:

$$p(y) = \frac{1}{Z} \prod_{F \in \mathcal{F}} \psi_F(y_{N(F)}) \quad \text{with} \quad Z = \sum_{y \in \mathcal{Y}} \prod_{F \in \mathcal{F}} \psi_F(y_{N(F)}).$$

Each factor $F \in \mathcal{F}$ connects a subset of nodes, hence we write $y_F = y_{N(F)} = (y_{v_1}, \ldots, y_{v_{\mid F\mid}})$.
Examples

An exemplar MRF

\[ p_1(y) = \frac{1}{Z_1} \psi_{ijkl}(y_i, y_j, y_k, y_l) \]

\[ p_2(y) = \frac{1}{Z_2} \psi_{ij}(y_i, y_j) \cdot \psi_{ik}(y_i, y_k) \cdot \psi_{il}(y_i, y_l) \cdot \psi_{jk}(y_j, y_k) \cdot \psi_{jl}(y_j, y_l) \cdot \psi_{kl}(y_k, y_l) \]

Factor graphs are universal, explicit about the factorization, hence it is easier to work with them.
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* $G = (V, E)$, 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 $G = (V, E)$, where the graph encodes conditional independence assumptions between the random variables.
- **Factor graphs** are universal, explicit about the factorization, hence it is easier to work with them.

In the **next lecture** we will learn about
- Conditional random field (CRF)
- Inference for graphical models
- Binary image segmentation
- EM algorithm

---

**Literature * **

**Probability theory**


**Graphical models**