Undirected Graphical Models I

Definition and Basic Properties

Chris Williams
(based on slides by Michael U. Gutmann)

Probabilistic Modelling and Reasoning (INFR11134)
School of Informatics, The University of Edinburgh

Spring Semester 2023

Recap

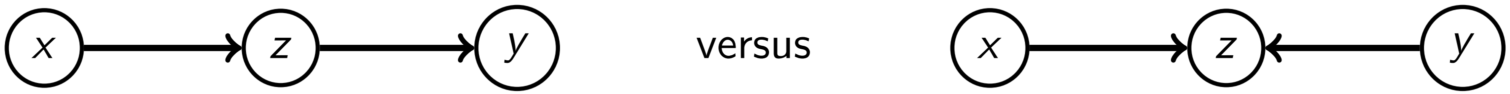
- The number of free parameters in probabilistic models increases with the number of random variables involved.
- Making statistical independence assumptions reduces the number of free parameters that need to be specified.
- Starting with the chain rule and an ordering of the random variables, we used statistical independencies to simplify the representation.
- We thus obtained a factorisation in terms of a product of conditional pdfs that we visualised as a DAG.
- In turn, we used DAGs to define sets of distributions ("directed graphical models").
- We discussed independence properties satisfied by the distributions, d-separation, and the equivalence to the factorisation.

The directionality in directed graphical models

- So far we mainly exploited the property

$$\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z} \Longleftrightarrow p(\mathbf{y}|\mathbf{x}, \mathbf{z}) = p(\mathbf{y}|\mathbf{z})$$

- But when working with $p(\mathbf{y}|\mathbf{x}, \mathbf{z})$ we impose an ordering or directionality from $\mathbf{x}$ and $\mathbf{z}$ to $\mathbf{y}$.
- Directionality matters in directed graphical models



- In some cases, directionality is natural but in others we do not want to choose one direction over another.
- We now discuss how to visualise and represent probability distributions and independencies in a symmetric manner without assuming a directionality or ordering of the variables.

Program

1. Visualising factorisations with undirected graphs
2. Undirected graphical models

Program

1. Visualising factorisations with undirected graphs
   - Undirected characterisation of statistical independence
   - Gibbs distributions
   - Visualising Gibbs distributions with undirected graphs

2. Undirected graphical models

Further characterisation of statistical independence

- From exercises: For non-negative functions $a(\mathbf{x}, \mathbf{z}), b(\mathbf{y}, \mathbf{z})$:

$$\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z} \Longleftrightarrow p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = a(\mathbf{x}, \mathbf{z})b(\mathbf{y}, \mathbf{z})$$

- Equivalent to $p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z})p(\mathbf{z})$ but does not assume that the factors are (conditional) pdfs/pmfs.
- No directionality or ordering of the variables is imposed.
- Unconditional version: For non-negative functions $a(\mathbf{x}), b(\mathbf{y})$:

$$\mathbf{x} \perp\!\!\!\perp \mathbf{y} \Longleftrightarrow p(\mathbf{x}, \mathbf{y}) = a(\mathbf{x})b(\mathbf{y})$$

- The important point is the factorisation of $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ into two non-negative factors:
  - if the factors share a variable $\mathbf{z}$, then we have conditional independence,
  - if not, we have unconditional independence.

Further characterisation of statistical independence

- Since $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ must sum (integrate) to one, we must have

$$\sum_{\mathbf{x},\mathbf{y},\mathbf{z}} a(\mathbf{x}, \mathbf{z}) b(\mathbf{y}, \mathbf{z}) = 1$$

- Normalisation condition often ensured by re-defining $a(\mathbf{x}, \mathbf{z}) b(\mathbf{y}, \mathbf{z})$:

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \frac{1}{Z} \phi_A(\mathbf{x}, \mathbf{z}) \phi_B(\mathbf{y}, \mathbf{z}) \qquad Z = \sum_{\mathbf{x},\mathbf{y},\mathbf{z}} \phi_A(\mathbf{x}, \mathbf{z}) \phi_B(\mathbf{y}, \mathbf{z})$$

- Z: normalisation constant (related to partition function, see later)
- ϕ_i: factors (also called potential functions).
  Do generally not correspond to (conditional) pdfs/pmfs.

What does it mean?

$$\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z} \Longleftrightarrow p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \frac{1}{Z}\phi_A(\mathbf{x}, \mathbf{z})\phi_B(\mathbf{y}, \mathbf{z})$$

"$\Rightarrow$" If we want our model to satisfy $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$ we should write the pdf (pmf) as

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}) \propto \phi_A(\mathbf{x}, \mathbf{z})\phi_B(\mathbf{y}, \mathbf{z})$$

"$\Leftarrow$" If the pdf (pmf) can be written as $p(\mathbf{x}, \mathbf{y}, \mathbf{z}) \propto \phi_A(\mathbf{x}, \mathbf{z})\phi_B(\mathbf{y}, \mathbf{z})$ then we have $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$

equivalent for unconditional version

Example

Consider $p(x_1, x_2, x_3, x_4) \propto \phi_1(x_1, x_2)\phi_2(x_2, x_3)\phi_3(x_4)$

What independencies does p satisfy?

- We can write

$$p(x_1, x_2, x_3, x_4) \propto \underbrace{[\phi_1(x_1, x_2)\phi_2(x_2, x_3)]}_{\tilde{\phi}_1(x_1,x_2,x_3)}[\phi_3(x_4)]$$
$$\propto \tilde{\phi}_1(x_1, x_2, x_3)\phi_3(x_4)$$

  so that $x_4 \perp\!\!\!\perp x_1, x_2, x_3$.

- Integrating out x_4 gives

$$p(x_1, x_2, x_3) = \int p(x_1, x_2, x_3, x_4)\mathrm{d}x_4 \propto \phi_1(x_1, x_2)\phi_2(x_2, x_3)$$

  so that $x_1 \perp\!\!\!\perp x_3 \mid x_2$

Gibbs distributions

- Example is a special case of a class of pdfs/pmfs that factorise as
$$p(x_1, \ldots, x_d) = \frac{1}{Z} \prod_c \phi_c(\mathcal{X}_c)$$
  - $\mathcal{X}_c \subseteq \{x_1, \ldots, x_d\}$
  - ϕ_c are non-negative factors (potential functions)
    Do generally not correspond to (conditional) pdfs/pmfs.
    They measure "compatibility", "agreement", or "affinity"
  - Z is a normalising constant so that $p(x_1, \ldots, x_d)$ integrates (sums) to one.
- Known as Gibbs (or Boltzmann) distributions
- $\tilde{p}(x_1, \ldots, x_d) = \prod_c \phi_c(\mathcal{X}_c)$ is said to be an unnormalised model: $\tilde{p} \geq 0$ but does not necessarily integrate (sum) to one.

Energy-based model

- With $\phi_c(\mathcal{X}_c) = \exp\left(-E_c(\mathcal{X}_c)\right)$, we have equivalently

$$p(x_1, \ldots, x_d) = \frac{1}{Z} \exp\left[-\sum_c E_c(\mathcal{X}_c)\right]$$

- $\sum_c E_c(\mathcal{X}_c)$ is the energy of the configuration $(x_1, \ldots, x_d)$.
  low energy $\iff$ high probability

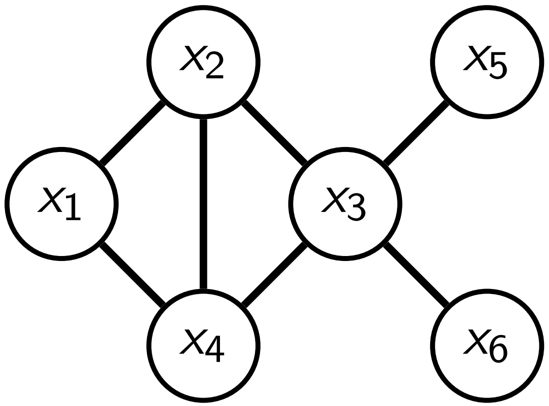
Visualising Gibbs distributions with undirected graphs

$p(x_1, \ldots, x_d) \propto \prod_c \phi_c(\mathcal{X}_c)$

- Node for each x_i
- For all factors ϕ_c: draw an undirected edge between all x_i and x_j that belong to $\mathcal{X}_c$
- Results in a fully-connected subgraph for all x_i that are part of the same factor (this subgraph is called a clique).

Example:

Graph for $p(x_1, \ldots, x_6) \propto \phi_1(x_1, x_2, x_4)\phi_2(x_2, x_3, x_4)\phi_3(x_3, x_5)\phi_4(x_3, x_6)$



Program

1. Visualising factorisations with undirected graphs
   - Undirected characterisation of statistical independence
   - Gibbs distributions
   - Visualising Gibbs distributions with undirected graphs

2. Undirected graphical models

Program

1. Visualising factorisations with undirected graphs

2. Undirected graphical models
   - Definition
   - Examples
   - Conditionals, marginals, and change of measure

Undirected graphical models (UGMs)

- We started with a factorised pdf/pmf and associated a undirected graph with it. We now go the other way around and start with an undirected graph.
- *Definition* An undirected graphical model based on an undirected graph H with d nodes and associated random variables x_i is the set of pdfs/pmfs that factorise as

$$p(x_1, \ldots, x_d) = \frac{1}{Z} \prod_c \phi_c(\mathcal{X}_c)$$

  where Z is the normalisation constant, $\phi_c(\mathcal{X}_c) \geq 0$, and the $\mathcal{X}_c$ correspond to the maximal cliques in the graph.
- Remark: a pdf/pmf $p(x_1, \ldots, x_d)$ that can be written as above is said to "factorise over the graph H". We also say that it has property $F(H)$ ("F" for factorisation).

Remarks

- An undirected graph defines the pdfs/pmfs in terms of Gibbs distributions.
- The undirected graphical model corresponds to a set of probability distributions. This is because we did not specify any numerical values for the factors $\phi_c(\mathcal{X}_c)$. We only specified which variables the factors take as input.
- Individual pdfs/pmf in the set are typically also called a undirected graphical model.
- Other names for an undirected graphical model: Markov network (MN), Markov random field (MRF)
- The $\mathcal{X}_c$ form maximal cliques in the graph.

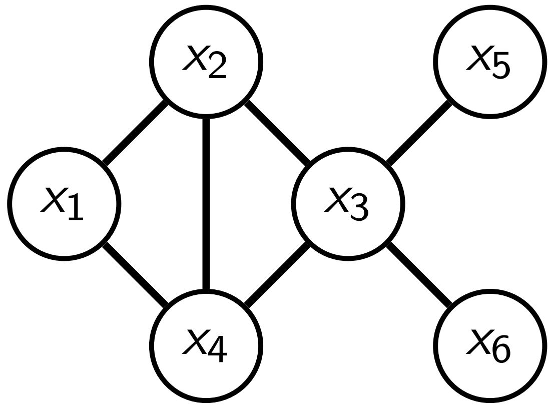
  Maximal clique: a set of fully connected nodes (clique) that is not contained in another clique.

Why maximal cliques?

- The mapping from Gibbs distribution to graph is many to one. We may obtain the same graph for different Gibbs distributions, e.g.

$$p(\mathbf{x}) \propto \phi_1(x_1, x_2, x_4)\phi_2(x_2, x_3, x_4)\phi_3(x_3, x_5)\phi_4(x_3, x_6)$$

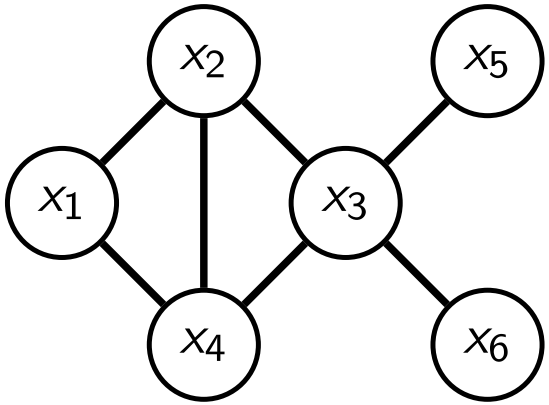
$$p(\mathbf{x}) \propto \tilde{\phi}_1(x_1, x_2)\tilde{\phi}_2(x_1, x_4)\tilde{\phi}_3(x_2, x_4)\tilde{\phi}_4(x_2, x_3)\tilde{\phi}_5(x_3, x_4)\tilde{\phi}_6(x_3, x_5)\tilde{\phi}_7(x_3, x_6)$$



- By using maximal cliques, we take a conservative approach and do not make additional assumptions on the factorisation.

Example

Undirected graph:



Random variables: $\mathbf{x} = (x_1, \ldots, x_6)$

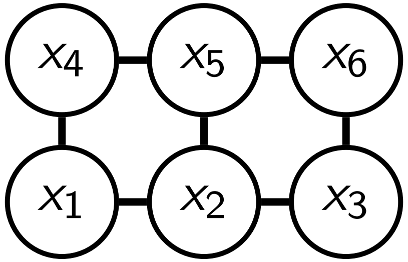
Maximal cliques: $\{x_1, x_2, x_4\}, \quad \{x_2, x_3, x_4\}, \quad \{x_3, x_5\}, \quad \{x_3, x_6\}$

Undirected graphical model: set of pdfs/pmfs $p(\mathbf{x})$ that factorise as:

$$\begin{aligned} p(\mathbf{x}) &= \frac{1}{Z}\phi_1(x_1, x_2, x_4)\phi_2(x_2, x_3, x_4)\phi_3(x_3, x_5)\phi_4(x_3, x_6) \\ &\propto \phi_1(x_1, x_2, x_4)\phi_2(x_2, x_3, x_4)\phi_3(x_3, x_5)\phi_4(x_3, x_6) \end{aligned}$$

Example (pairwise Markov network)

Graph:



Random variables: $\mathbf{x} = (x_1, \ldots, x_6)$

Maximal cliques: all neighbours

$\{x_1, x_2\}$ $\{x_2, x_3\}$ $\{x_4, x_5\}$ $\{x_5, x_6\}$ $\{x_1, x_4\}$ $\{x_2, x_5\}$ $\{x_3, x_6\}$

Undirected graphical model: set of pdfs/pmfs $p(\mathbf{x})$ that factorise as:

$$p(\mathbf{x}) \propto \phi_1(x_1, x_2)\phi_2(x_2, x_3)\phi_3(x_4, x_5)\phi_4(x_5, x_6)\phi_5(x_1, x_4)\phi_6(x_2, x_5)\phi_7(x_3, x_6)$$

Example of a pairwise Markov network.

Example: Ising model

- Variables x_i taken on values in $\{-1, +1\}$ ("spins")
- Laid out on a grid (pairwise Markov network)
- $E(x_i, x_j) = -Jx_ix_j$ if i and j are neighbours, 0 otherwise
- If $J > 0$ then we get low energy (high probability) when $x_i = x_j$, and higher energy when $x_i \neq x_j$
- This is "ferromagnetic" behaviour in physics (spins align)
- Lots of theory in statistical physics, e.g. on phase transitions

Example: Graphical Gaussian models

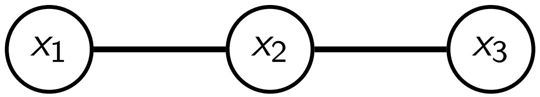
- Gaussian pdf $N(x; \boldsymbol{\mu}, \Sigma)$:

$$p(\mathbf{x}) \propto \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

- Set $\Lambda = \Sigma^{-1}$, the *precision matrix*, then

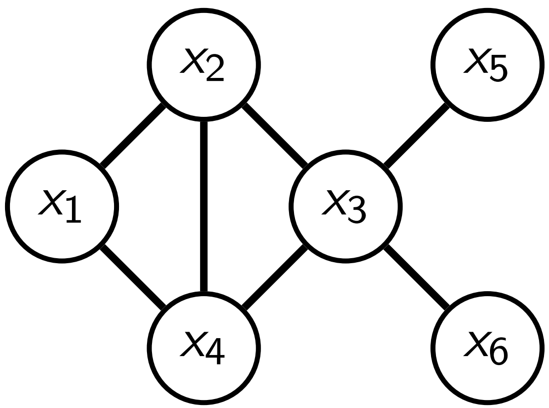
$$p(\mathbf{x}) \propto \exp\left(-\frac{1}{2}\mathbf{x}^T\Lambda\mathbf{x} + \mathbf{h}^T\mathbf{x}\right)$$

  with $\mathbf{h} = \Lambda\boldsymbol{\mu}$
- If $\Lambda_{ij} = \Lambda_{ji} = 0$, then there is no edge between i and j in the graph
- Zeros in Λ define a Graphical Gaussian model, e.g.



Conditionals

- For DGMs, the factors $k(x_i|\mathrm{pa}_i)$ defining $p(\mathbf{x})$ are the conditional pdfs/pmfs of x_i given pa_i under $p(\mathbf{x})$, i.e. $p(x_i|\mathrm{pa}_i)$. We do not have such a correspondence for UGMs.
- But conditioning on random variables corresponds to a simple graph operation: removing their nodes from the graph.
- Example: For $p(x_1, \ldots, x_6)$ specified by the graph below, what is $p(x_1, x_2, x_4, x_5, x_6|x_3 = \alpha)$?



Conditionals

- The graph specifies the factorisation

$$p(x_1, \dots, x_6) \propto \phi_1(x_1, x_2, x_4)\phi_2(x_2, x_3, x_4)\phi_3(x_3, x_5)\phi_4(x_3, x_6)$$

- By definition: $p(x_1, x_2, x_4, x_5, x_6 | x_3 = \alpha)$

$$= \frac{p(x_1, x_2, x_3 = \alpha, x_4, x_5, x_6)}{\int p(x_1, x_2, x_3 = \alpha, x_4, x_5, x_6) \mathrm{d}x_1 \mathrm{d}x_2 \mathrm{d}x_4 \mathrm{d}x_5 \mathrm{d}x_6}$$
$$= \frac{\phi_1(x_1, x_2, x_4)\phi_2(x_2, \alpha, x_4)\phi_3(\alpha, x_5)\phi_4(\alpha, x_6)}{\int \phi_1(x_1, x_2, x_4)\phi_2(x_2, \alpha, x_4)\phi_3(\alpha, x_5)\phi_4(\alpha, x_6) \mathrm{d}x_1 \mathrm{d}x_2 \mathrm{d}x_4 \mathrm{d}x_5 \mathrm{d}x_6}$$
$$= \frac{1}{Z(\alpha)} \phi_1(x_1, x_2, x_4)\phi_2^\alpha(x_2, x_4)\phi_3^\alpha(x_5)\phi_4^\alpha(x_6)$$

- Gibbs distribution with derived factors ϕ_i^α of reduced domain and new normalisation “constant” $Z(\alpha)$
- Note that $Z(\alpha)$ depends on the conditioning value α.

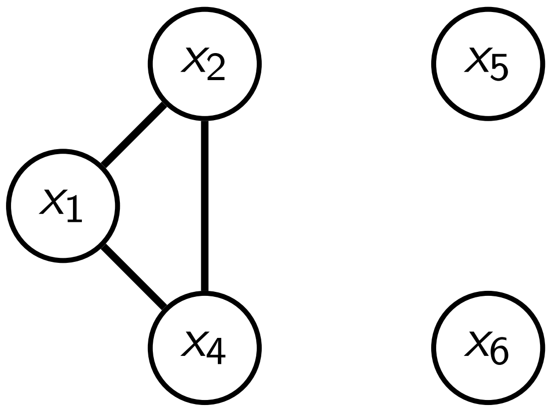
Conditionals

Let $p(x_1, \ldots, x_6) \propto \phi_1(x_1, x_2, x_4)\phi_2(x_2, x_3, x_4)\phi_3(x_3, x_5)\phi_4(x_3, x_6)$.

- Conditional $p(x_1, x_2, x_4, x_5, x_6 | x_3 = \alpha)$ is

$$\frac{1}{Z(\alpha)}\phi_1(x_1, x_2, x_4)\phi_2^\alpha(x_2, x_4)\phi_3^\alpha(x_5)\phi_4^\alpha(x_6)$$

- Conditioning on variables removes the corresponding nodes and connecting edges from the undirected graph



Marginals

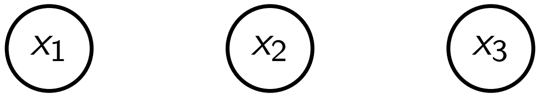
- For DGMs, the product of the first j terms in the factorisation, $\prod_{i=1}^{j} k(x_i|\mathrm{pa}_i)$, equaled the marginal $p(x_1, \ldots, x_j)$.
- UGMs do not have such a general property. But we can exploit the factorisation when computing the marginals.
- Will be the discussed in the "inference part" of the course.

Change of measure

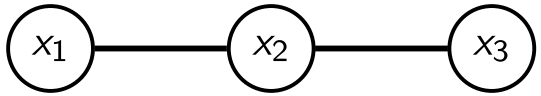
- A way to create new pdf/pmfs is to reweight existing ones, which is a special instance of a "change of measure".
- For example, assume $q(x_1, x_2, x_3) = \prod_i q_i(x_i)$ to be a given pmf. We want to generate a new pmf that assigns higher probabilities to $(x_1, x_2) \in A$, and to $(x_2, x_3) \in B$, for some sets A and B.
- We can thus define the Gibbs distribution

$$p(\mathbf{x}) = \frac{1}{Z}\phi_A(x_1, x_2)\phi_B(x_2, x_3)\prod_{i=1}^{3} q_i(x_i)$$

where $\phi_A(x_1, x_2) = 1$ for $(x_1, x_2) \notin A$, $\phi_A(x_1, x_2) > 1$ for $(x_1, x_2) \in A$, and equivalently for ϕ_B.



graph for $q(\mathbf{x})$



graph for $p(\mathbf{x})$

Change of measure

- Similarly, we can think that an undirected graph defines how a base distribution, e.g. $q(\mathbf{x}) = \prod_i q_i(x_i)$, should be reweighted by factors $\phi_c(\mathcal{X}_c)$, thus defining a change of measure.
- Two different ways of defining models: Reweighting for UGMs vs data generation for DGMs.
- Reweighting is clear when computing expectations, e.g.

$$\begin{aligned}
\mathbb{E}_p[h] &= \sum_{\mathbf{x}} h(\mathbf{x})p(\mathbf{x}) \\
&= \frac{1}{Z} \sum_{x_1,x_2,x_3} h(x_1,x_2,x_3)\phi_A(x_1,x_2)\phi_B(x_2,x_3) \prod_i q_i(x_i) \\
&= \frac{1}{Z} \mathbb{E}_q[h\phi_A\phi_B]
\end{aligned}$$

- Since $Z = \sum_{x_1,x_2,x_3} \phi_A(x_1,x_2)\phi_B(x_2,x_3) \prod_i q_i(x_i) = \mathbb{E}_q[\phi_A\phi_B]$

$$\mathbb{E}_p[h] = \frac{\mathbb{E}_q[h\phi_A\phi_B]}{\mathbb{E}_q[\phi_A\phi_B]}$$

Change of measure

Program recap

1. Visualising factorisations with undirected graphs
   - Undirected characterisation of statistical independence
   - Gibbs distributions
   - Visualising Gibbs distributions with undirected graphs

2. Undirected graphical models
   - Definition
   - Examples
   - Conditionals, marginals, and change of measure

Credits

These slides are modified from ones produced by Michael Gutmann, made available under Creative Commons licence CC BY 4.0.

©Michael Gutmann and Chris Williams, The University of Edinburgh 2018-2023 CC BY 4.0 (cc)(i).