Probabilistic Machine Learning
Lecture 16
Graphical Models

Philipp Hennig

15 June 2021

Eberhard Karls Universität Tübingen

Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning

#	date	content	Ex	#	date	content	Ex
1	20.04.	Introduction	1	14	09.06.	Generalized Linear Models	
2	21.04.	Reasoning under Uncertainty		15	15.06.	Exponential Families	8
3	27.04.	Continuous Variables	2	16	16.06.	Graphical Models	
4	28.04.	Monte Carlo		17	22.06.	Factor Graphs	9
5	04.05.	Markov Chain Monte Carlo	3	18	23.06.	The Sum-Product Algorithm	
6	05.05.	Gaussian Distributions		19	29.06.	Example: Topic Models	10
7	11.05.	Parametric Regression	4	20	30.06.	Mixture Models	
8	12.05.	Learning Representations		21	06.07.	EM	11
9	18.05.	Gaussian Processes	5	22	07.07.	Variational Inference	
10	19.05.	Understanding Kernels		23	13.07.	Topics	
11	26.05.	Gauss-Markov Models		25	14.07.	Example: Inferring Topics	
12	25.05.	An Example for GP Regression	6	24	20.07.	Example: Kernel Topic Models	
13	08.06.	GP Classification	7	26	21.07.	Revision	

Recap from Lecture 1

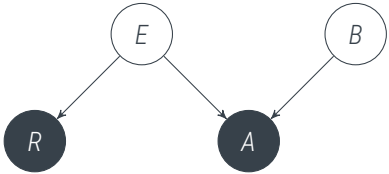
Conditional Independence Affects Computational Complexity

Joint probability distribution has
$2^4 - 1 = 15 = 8 + 4 + 2 + 1$ parameters

$$p(A, E, B, R) = p(A \mid R, E, B) \cdot p(R \mid E, B) \cdot p(E \mid B) \cdot p(B)$$

Removing irrelevant conditions (domain knowledge!) reduces to $8 = 4 + 2 + 1 + 1$ parameters:

$$p(A, E, B, R) = p(A \mid E, B) \cdot p(R \mid E) \cdot p(E) \cdot p(B)$$



Procedural construction of **directed graphical model**

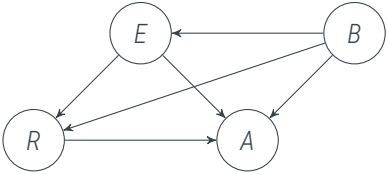
1. For each variable in the joint distribution, draw a circle
2. For each term $p(x_1, \dots \mid y_1, \dots)$ in the factorized joint distribution, draw an arrow *from* every **parent** (right side) node y_i *to* every **child** (left side) node x_i.
3. fill in all **observed** variables (variables on which we want to *condition*).

Every Probability Distribution is a DAG

It's just not always a helpful concept

By the Product Rule, every joint can be factorized into a (dense) DAG.

$$p(A, E, B, R) = p(A \mid E, B, R) \cdot p(R \mid E, B) \cdot p(E \mid B) \cdot p(B)$$



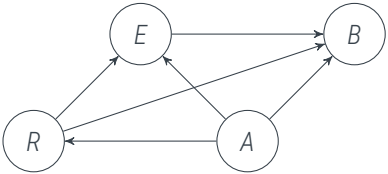
A = the alarm was triggered
E = there was an earthquake
B = there was a break-in
R = an announcement is made on the radio

Every Probability Distribution is a DAG

It's just not always a helpful concept

The direction of the arrows is **not** a causal statement.

$$p(A, E, B, R) = p(B \mid A, E, R) \cdot p(E \mid A, R) \cdot p(R \mid A) \cdot p(A)$$



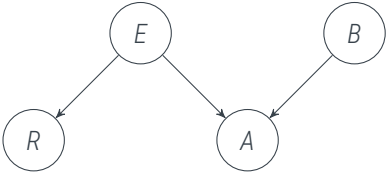
A = the alarm was triggered
E = there was an earthquake
B = there was a break-in
R = an announcement is made on the radio

Every Probability Distribution is a DAG

It's just not always a helpful concept

But the representation is particularly interesting when it reveals **independence**.

$$p(A, E, B, R) = p(A \mid E, B) \cdot p(R \mid E) \cdot p(E) \cdot p(B)$$



$A =$ the alarm was triggered
$E =$ there was an earthquake
$B =$ there was a break-in
$R =$ an announcement is made on the radio

Directed Graphs are an Imperfect Representation

The Graph for Two Coins and a Bell

example by Stefan Harmeling

$$P(A=1)=0.5 \qquad P(C=1 \mid A=1, B=1)=1 \qquad P(C=1 \mid A=1, B=0)=0$$
$$P(B=1)=0.5 \qquad P(C=1 \mid A=0, B=1)=0 \qquad P(C=1 \mid A=0, B=0)=1$$

These CPTs imply $P(A|B)=P(A)$, $P(B|C)=P(B)$ and $P(C|A)=P(C)$ and $P(C \mid B)=P(C)$.

Directed Graphs are an Imperfect Representation

The Graph for Two Coins and a Bell

example by Stefan Harmeling

$$P(A=1)=0.5 \qquad P(C=1 \mid A=1, B=1)=1 \qquad P(C=1 \mid A=1, B=0)=0$$
$$P(B=1)=0.5 \qquad P(C=1 \mid A=0, B=1)=0 \qquad P(C=1 \mid A=0, B=0)=1$$

These CPTs imply $P(A|B)=P(A)$, $P(B|C)=P(B)$ and $P(C|A)=P(C)$ and $P(C \mid B)=P(C)$.

We thus have three factorizations:

1. $P(A,B,C)=P(C|A,B)\cdot P(A|B)\cdot P(B)=P(C|A,B)\cdot P(A)\cdot P(B)$
2. $P(A,B,C)=P(A|B,C)\cdot P(B|C)\cdot P(C)=P(A|B,C)\cdot P(B)\cdot P(C)$
3. $P(A,B,C)=P(B|C,A)\cdot P(C|A)\cdot P(A)=P(B|C,A)\cdot P(C)\cdot P(A)$

Directed Graphs are an Imperfect Representation

The Graph for Two Coins and a Bell

example by Stefan Harmeling

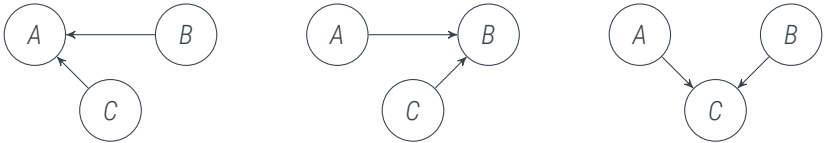
$$P(A=1)=0.5 \qquad P(C=1 \mid A=1, B=1)=1 \qquad P(C=1 \mid A=1, B=0)=0$$
$$P(B=1)=0.5 \qquad P(C=1 \mid A=0, B=1)=0 \qquad P(C=1 \mid A=0, B=0)=1$$

These CPTs imply $P(A|B)=P(A)$, $P(B|C)=P(B)$ and $P(C|A)=P(C)$ and $P(C \mid B)=P(C)$.

We thus have three factorizations:

1. $P(A,B,C)=P(C|A,B)\cdot P(A|B)\cdot P(B)=P(C|A,B)\cdot P(A)\cdot P(B)$
2. $P(A,B,C)=P(A|B,C)\cdot P(B|C)\cdot P(C)=P(A|B,C)\cdot P(B)\cdot P(C)$
3. $P(A,B,C)=P(B|C,A)\cdot P(C|A)\cdot P(A)=P(B|C,A)\cdot P(C)\cdot P(A)$

Each corresponds to a graph. Note that each can only express some of the independencies:



Today: More about graphs

- extended syntax for directed graphical models
- constructing conditional independence from directed graphs
- an alternative framework, in which conditional independence is easy, but the joint is hard
- some theory on its representational power

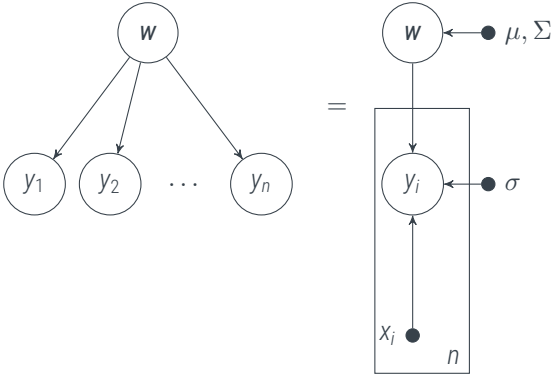
Overarching Goal: Representing probability distributions in a graphical way, to guide and simplify the design of advanced probabilistic models

Plates and Hyperparameters

some syntactic sugar for practical uses

$$p(\boldsymbol{y}, \boldsymbol{w}) = \prod_{i=1}^{n} \mathcal{N}(y_i; \phi(x_i)^\intercal \boldsymbol{w}, \sigma^2) \mathcal{N}(\boldsymbol{w}; \mu, \Sigma)$$

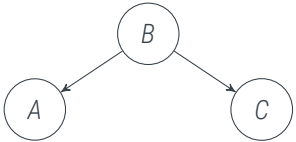
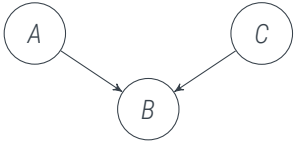
- A box with sharp edges, drawn around a set of nodes and labeled with a number n is called a **plate** and denotes n copies of the content of the box.
- a small filled circle denots a (hyper-) parameter that is set or optimized, and which is not part of the generative model.



Atomic Independence Structures

DAGs imply conditional independence, but not dependence!

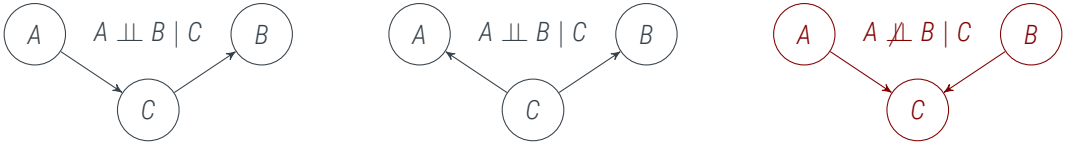
For uni- and bi-variate graphs, conditional independence is trivial.
For tri-variate sub-graphs, there are three possible structures:

	graph	factorization	implications
(i)		$p(A,B,C) = p(C \mid B) \cdot p(B \mid A) \cdot p(A)$	$A \perp\!\!\!\perp C \mid B$ but not, i.e., $A \not\perp\!\!\!\perp C$
(ii)		$p(A,B,C) = p(A \mid B) \cdot p(C \mid B) \cdot p(B)$	$A \perp\!\!\!\perp C \mid B$ but not, i.e., $A \not\perp\!\!\!\perp C$
(iii)		$p(A,B,C) = p(B \mid A,C) \cdot p(C) \cdot p(A)$	$A \perp\!\!\!\perp C$ but not, i.e., $A \not\perp\!\!\!\perp C \mid B$

d-separation

A Generalization of the Atomic Structures above

[J. Pearl, *Probabilistic Reasoning in Intelligent Systems* 1988]



Theorem (d-separation, Pearl, 1988. Formulation taken from Bishop, 2006)

Consider a general directed acyclic graph, in which A, B, C are nonintersecting sets of nodes whose union may be smaller than the complete graph. To ascertain whether $A \perp\!\!\!\perp B \mid C$, consider all possible paths (connections along lines in the graph, regardless of the direction) from any node in A to any node in B. Any such path is considered blocked if it includes a node such that either

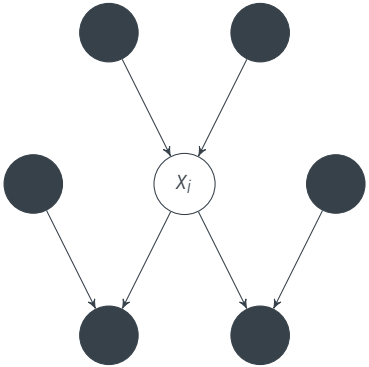
- *the arrows on the path meet either head-to-tail or tail-to-tail at the node, and the node is in C, or*
- *the arrows meet head-to-head at the node, and neither the node, nor any of its descendants is in C.*

*If all paths are blocked, then A is said to be d-**separated** from B by C, and $A \perp\!\!\!\perp B \mid C$.*

Markov Blankets

Thus, all further considerations about computations on the graph can be made in a local fashion.

$$
\begin{aligned}
& p(x_i \mid \mathbf{x}_{j \neq i}) \\
&= \frac{p(x_1, \dots, x_d)}{\int p(x_1, \dots, x_d)\, dx_i} = \frac{\prod_k p(x_k \mid \text{parents}_k)}{\int \prod_k p(x_k \mid \text{parents}_k)\, dx_i} \\
&= \frac{\prod_{k' \notin \text{blanket}} p(x_{k'} \mid \text{parents}_{k'}) \prod_{k \in \text{blanket}} p(x_k \mid \text{parents}_k)}{\prod_{k' \notin \text{blanket}} p(x_{k'} \mid \text{parents}_{k'}) \int \prod_{k \in \text{blanket}} p(x_k \mid \text{parents}_k)\, dx_i} \\
&= \frac{\prod_{k \in \text{blanket}} p(x_k \mid \text{parents}_k)}{\int \prod_{k \in \text{blanket}} p(x_k \mid \text{parents}_k)\, dx_i}
\end{aligned}
$$

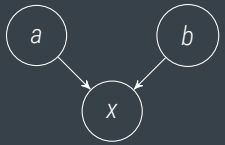


Definition (Markov Blanket — for directed graphs)

The **Markov Blanket** of node x_i is the set of all *parents, children, and co-parents* of x_i. Conditioned on the blanket, x_i is independent of the rest of the graph.

Directed Graphical Models

- The directed nature of connections in Bayesian belief networks reflects the fact that a conditional probability has a left- and right-hand side



$$= p(x \mid a, b) \cdot p(a) \cdot p(b)$$

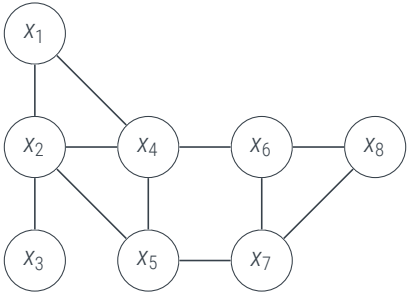
- This is convenient since it allows writing down the graph directly from the factorization.
- But conditional independence statements (*d*-separation) is tricky. Blocking a path requires notions of parents and co-parents, and different rules depending on whether arrows meet head-to-head or head-to-tail.
- There are joint distributions whose set of conditional independences can not be represented by a single directed graph.

Is there another notation, in which conditional independence can be more simply stated as "two nodes are independent if all paths connecting them are blocked?"

Undirected Graphical Models

aka. Markov Random Fields

[example from Bishop, PRML, 2006]



Definition (Markov Random Field)

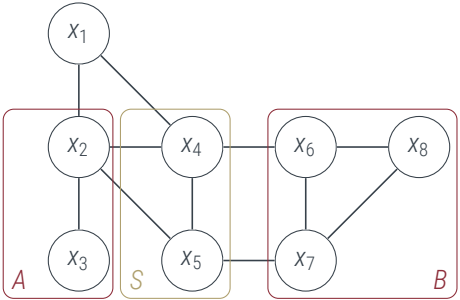
An *undirected Graph* $G = (V, E)$ is a set V of nodes and edges E. An undirected graph G and a set of random variables $X = \{X_v\}_{v \in V}$ is a **Markov Random Field** if, for any subsets $A, B \subset V$ and a *separating set* S (i.e. a set such that every path from A to B passes through S), $X_A \perp\!\!\!\perp X_B \mid X_S$.

The above definition is known as the *global Markov property*. It implies the weaker *pairwise Markov property*: Any two nodes u, v that do not share an edge are conditionally independent given all other variables: $X_u \perp\!\!\!\perp X_v \mid X_{V \setminus \{u,v\}}$.

Undirected Graphical Models

aka. Markov Random Fields

[example from Bishop, PRML, 2006]



Definition (Markov Random Field)

An *undirected Graph* $G = (V, E)$ is a set V of nodes and edges E. An undirected graph G and a set of random variables $X = \{X_v\}_{v \in V}$ is a **Markov Random Field** if, for any subsets $A, B \subset V$ and a *separating set* S (i.e. a set such that every path from A to B passes through S), $X_A \perp\!\!\!\perp X_B \mid X_S$.

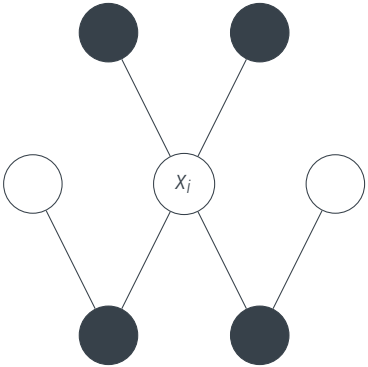
The above definition is known as the *global Markov property*. It implies the weaker *pairwise Markov property*: Any two nodes u, v that do not share an edge are conditionally independent given all other variables: $X_u \perp\!\!\!\perp X_v \mid X_{V \setminus \{u,v\}}$.

Markov Blankets, again

The strength of undirected graphs

Definition (Markov Blanket — for undirected graphs)

For a Markov Random Field, the **Markov Blanket** of node x_i is the set of all direct *neighbors* of x_i (the set of all nodes that share an edge with x_i). Conditioned on the blanket, x_i is independent of the rest of the graph.



Essentially by definition,
MRFs allow a more compact definition of conditional independence than directed graphs.
But what is the associated *joint* probability distribution?

Cliques

[example and definition from Bishop, PRML, 2006]

By the pairwise Markov property, any two nodes not connected by an edge have to be conditionally independent given the rest of the graph. Thus, the joint has to factorize as

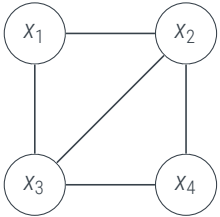
$$p(x_i, x_j \mid \mathbf{x}_{\setminus\{i,j\}}) = p(x_i \mid \mathbf{x}_{\setminus\{i,j\}}) \cdot p(x_j \mid \mathbf{x}_{\setminus\{i,j\}})$$

Hence, for the factorization to hold, nodes that do not share an edge must not be in the same factor. What kind of factors does this leave us with?

Definition (Cliques)

Given a graph $G = (V, E)$, a **clique** is a subset $c \subset V$ such that there exists an edge between all pairs of nodes in c. A **maximal clique** is a clique such that it is impossible to include any other nodes from V without it ceasing to be a clique.

In the following slides, the set of all maximal cliques of a graph will be denoted C.



Cliques

[example and definition from Bishop, PRML, 2006]

By the pairwise Markov property, any two nodes not connected by an edge have to be conditionally independent given the rest of the graph. Thus, the joint has to factorize as

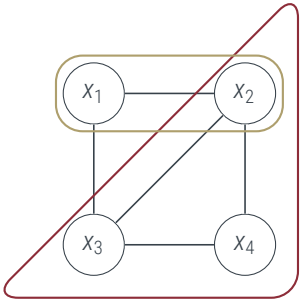
$$p(x_i, x_j \mid \mathbf{x}_{\setminus\{i,j\}}) = p(x_i \mid \mathbf{x}_{\setminus\{i,j\}}) \cdot p(x_j \mid \mathbf{x}_{\setminus\{i,j\}})$$

Hence, for the factorization to hold, nodes that do not share an edge must not be in the same factor. What kind of factors does this leave us with?

Definition (Cliques)

Given a graph $G = (V, E)$, a **clique** is a subset $c \subset V$ such that there exists an edge between all pairs of nodes in c. A **maximal clique** is a clique such that it is impossible to include any other nodes from V without it ceasing to be a clique.

In the following slides, the set of all maximal cliques of a graph will be denoted C.



Potentials

the price of dropping direction from edges

[derivation adapted from Bishop, PRML, 2016]

By the above, any distribution $p(\boldsymbol{x})$ that satisfies the conditional independence structures of the graph G can be written as a factorization over all cliques, and thus also just over all *maximal* cliques (since any clique is part of at least one maximal clique).

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{c \in C} \psi_c(\boldsymbol{x}_c) \qquad (\star)$$

- in directed graphs, each factor $p(x_{\text{ch}} \mid x_{\text{pa}})$ had to be a probability distribution of the children (but not of the parents!). But in MRFs there is no distinction between parents and children. So we only know that each **potential function** $\psi_c(x_c) \geq 0$. For simplicity, we will restrict $\psi_c(x_c) > 0$.
- The normalization constant Z is the **partition function**

$$Z := \sum_{\boldsymbol{x}} \prod_{c \in C} \psi_c(\boldsymbol{x}_c).$$

Because of the loss of structure from directed to undirected graphs, we have to explicitly compute Z. This can be NP-hard, and is the primary downside of MRFs. (e.g. consider n discrete variables with k states each, then computing Z may require summing k^n terms).

The Boltzmann distribution

Markov Random Fields with Positive Potentials are Exponential Families (but not necessarily of the helpful kind)

Because $\psi_c(\boldsymbol{x}_c) > 0$, we can write

$$\psi_c(\boldsymbol{x}_c) > 0 = \exp(-E_c(\boldsymbol{x}_c))$$

and introduce scaling factors w_c to get

$$p(\boldsymbol{x}) = \exp\left(-\sum_{c \in C} w_c E_c(\boldsymbol{x}_c) - \log Z\right)$$

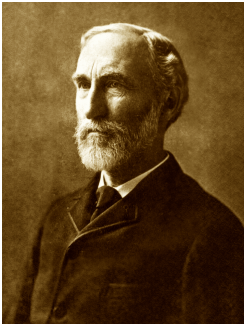
Definition (Boltzmann distribution / Gibbs measure)

A probability distribution with pdf of the form

$$p(\boldsymbol{x}) = e^{-E(\boldsymbol{x})}$$

is called a **Boltzmann** or **Gibbs** distribution. $E(\boldsymbol{x})$ is known as the **energy function**.

Ludwig E. Boltzmann
(1844–1906)



Josiah W. Gibbs
(1839–1903)

The Boltzmann distribution

Markov Random Fields with Positive Potentials are Exponential Families (but not necessarily of the helpful kind)

Because $\psi_c(\boldsymbol{x}_c) > 0$, we can write

$$\psi_c(\boldsymbol{x}_c) > 0 = \exp(-E_c(\boldsymbol{x}_c))$$

and introduce scaling factors w_c to get

$$p(\boldsymbol{x}) = \exp\left(-\sum_{c \in C} w_c E_c(\boldsymbol{x}_c) - \log Z\right)$$

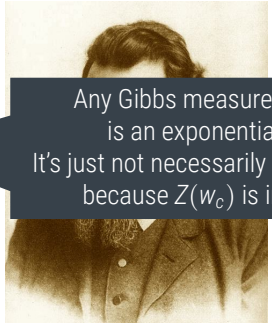
Definition (Boltzmann distribution / Gibbs measure)

A probability distribution with pdf of the form

$$p(\boldsymbol{x}) = e^{-E(\boldsymbol{x})}$$

is called a **Boltzmann** or **Gibbs** distribution. $E(\boldsymbol{x})$ is known as the **energy function**.

Any Gibbs measure (any MRF!)
is an exponential family!
It's just not necessarily the helpful kind
because $Z(w_c)$ is intractable!



Ludwig E. Boltzmann
(1844–1906)

Josiah W. Gibbs
(1839–1903)

The Hammersley-Clifford Theorem

formal statement of the rough derivation to here

Theorem (Hammersley-Clifford (unpublished, 1971. Clifford, 1990))

Consider the set of all possible strictly positive distributions $p(X_V)$ defined over a set V of variables corresponding to the nodes in the undirected graph $G = (V, E)$. Let $\mathcal{U}_I$ be the subset of such distributions that are consistent with the conditional independences that can be read off from G using graph separation. And let $\mathcal{U}_F$ be the set of such distributions that can be expressed as a Gibbs measure with the factorization $(\star)$. Then $\mathcal{U}_I = \mathcal{U}_F$.

Informally: "Any strictly positive MRF is a Gibbs measure, and every Gibbs measure is an MRF."

The Gaussian Case

For Gaussians, the MRF can be read off directly from the precision matrix

recap from Lecture 3

Consider a set of variables $\boldsymbol{x}$ that are jointly Gaussian distributed:

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$$

If the *inverse* covariance (aka. precision) matrix contains a zero at element $[\Sigma^{-1}]_{ij}$, then $x_i \perp\!\!\!\perp x_j \mid \boldsymbol{x}_{\setminus i,j}$.

Thus, for joint Gaussian models, the MRF can be constructed directly from the inverse covariance matrix:

1. draw a variable x_i for every element of $\boldsymbol{x}$
2. if $[\Sigma^{-1}]_{ij} \neq 0$, draw an edge between x_i and x_j.

Directed Graphical Models / Bayesian Networks

- directly encode a facorization of the joint (it can be read off by parsing the graph from the children to the parents)
- however, reading off conditional independence structure is tricky (it requires considering *d*-separation)
- directed graphs are for encoding **generative knowledge** (think: scientific modelling)

Undirected Graphical Models / Markov Random Fields (MRFs)

- directly encode conditional independence structure (by definition)
- however, reading off the joint from the graph is tricky (it requires finding all maximal cliques, normalization constant is intractable)
- MRFs are for encoding **computational constraints** (think: computer vision)