MSA101/MVE187 2022 Lecture 14
Graphical models

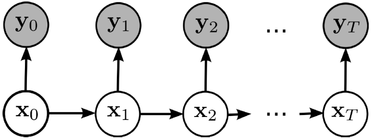
Petter Mostad

Chalmers University

October 12, 2022

From simple to complex models

- We have looked at Bayesian inference for small models where you may work with the entire posterior distribution, using, e.g., MCMC.
- For larger models, one needs to specify and systematically use conditional independencies between variables.
- Example: Algorithms developed for State Space Models (or Hidden Markov Models (HMM)):



- The ideas there may be generalized to general networks of variables.

Overview

- Graphical models: A way to specify stochastic models.
- Bayesian networks for modelling and model visualization.
- Using the graph to infer conditional independencies.
- Markov networks.
- Example: Gaussian Markov Random Fields.
- Using the graph for posterior inference.

Graphical representations of conditional independencies

- In complex models with many variables, it is crucial to model how variables depend on each other.
- Idea: Represent dependencies in a graph.
  - Helpful for visualization.
  - May use graph theory in connection with computations.
- We will look at two examples of graphical models:
  - Bayesian networks: Represent the probability density as a product of conditional densities:

$$\pi(x, y, z, v, w) = \pi(x \mid y, z) \cdot \pi(y \mid z) \cdot \pi(z \mid v, w) \cdot \pi(v) \cdot \pi(w)$$

  - Markov networks: Represent the probability density as a product of factors:

$$\pi(x, y, z, v, w) = C \cdot f_1(x, y, z) \cdot f_2(y, z) \cdot f_3(z, v, w) \cdot f_4(v) \cdot f_5(w)$$

Bayesian networks

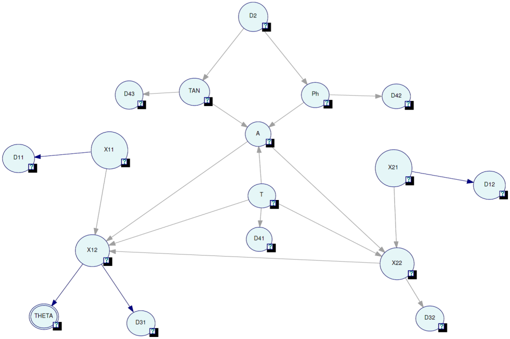
- Any joint density can always be written as a product over conditional densities:

$$\pi(x_1, \dots, x_n) = \pi(x_1)\pi(x_2 \mid x_1)\pi(x_3 \mid x_1, x_2) \dots \pi(x_n \mid x_1, \dots, x_{n-1})$$

- Given a specific model, we might be able to drop the conditioning on some of the variables in some factors. The representation then conveys the structure of the model.
- Re-ordering the variables will often give a different representation!
- The graph with an arrow $x \to y$ for each of the conditionings $\pi(y \mid \dots x \dots)$ in the representation above is the Bayesian Network representation. x is "parent", y is "child".
- Note that, following the arrows, you can never get a cycle. Thus the graph is a *directed acyclic graph* (DAG).
- Conversely, given any DAG and conditional densities for each child given its parents, the product of these gives a joint probability density.

Bayesian networks for visualization

- To the right: An example of a specific graphical network.
- Hierarchical models are, by definition, specified as a series of conditional distributions. The graph represents essential model information.
- Visualizations may use "plates" to represent repeated components.
- Note: Get a sample from the unconditional joint density by "propagating" simulation through network.



Conditional independence

- If x and y become independent when we fix the value of z we say that x and y are conditionally independent given z. We write $x \coprod y \mid z$.
- Equivalent formulations:
  - $\pi(x, y \mid z) = \pi(x \mid z)\pi(y \mid z)$
  - $\pi(x \mid y, z) = \pi(x \mid z)$
  - $\pi(y \mid x, z) = \pi(y \mid z)$
- We use the same definitions and notation when X, Y and Z are *disjoint groups of variables*.
- Example: When the data x_1, x_2, x_3 is *iid* given the parameter θ, we get for example $\{x_1, x_2\} \coprod x_3 \mid \theta$.

Reading off conditional independencies from a Bayesian network

- Some conditional independence statements can be "read off" the DAG of a Bayesian network.
- Is there a general way to prove that two sets of variables are conditionally independent given a third set based only on the Bayesian network graph?
- Preliminary observation: Two children with a single common parent are conditionally independent given the parent.
- Preliminary observation: Two parents with a single common child are generally NOT conditionally independent given the child.
- Definition: A "v-structure" is a part of a network consisting of a child with two parents.

d-separation

- A "trail" in a DAG is an *undirected path* in the graph.
- Assume X, Y, Z are sets of variables. An "active trail" from X to Y given Z is one where, for every v-structure $x_{i-1} \to x_i \leftarrow x_{i+1}$ in the trail, x_i or a decendant is in Z, and no other node in the trail is in Z.
- We say X and Y are *d-separated* given Z if there is no active trail between any $x \in X$ and $y \in Y$ given Z.
- Theorem: If X and Y are d-separated given Z in a Bayesian network representation of a stochastic model, then $X \coprod Y \mid Z$.
- Theorem: If X and Y are *not* d-separated given Z in a DAG, then there exists a stochastic model where X and Y are not conditionally independent given Z that has the DAG as a Bayesian network.
- See Koller & Friedman: "Probabilistic Graphical Models" for more details.

A way to check d-separation

Let X, Y, Z be disjoint sets of nodes in a Bayesian Network. Perform the following steps:

1. Remove all links from Z to their children.
2. Repeatedly, remove all childless nodes not in X, Y, or Z.

Then X and Y are d-separated given Z in the original network if and only if there is no trail from X to Y in the reduced network.
To prove this, prove following statements:

- Step 1 above does not change the d-separation.
- Step 2 above does not change the d-separation.
- After steps,
  - All nodes not in X, Y, Z have a descendant in X, Y, or Z.
  - Nodes in Z have no descendants.
- In a network fulfilling conditions above, any trail $X \rightarrow Y$ is active.

Markov networks

- For many models, the probability (density) function may be written as a product of positive factors where each involves only a subset of the variables. Example:

$$\pi(x, y, z, v, w) = C \cdot f_1(x, y, z) \cdot f_2(y, z) \cdot f_3(z, v, w) \cdot f_4(v) \cdot f_5(w)$$

- Note: The f_i functions are *not* necessarily densities (i.e., do not necessarily integrate to 1).
- Assume the representation is maximally reduced, i.e., for any pair of variables x, y occuring in a factor, the factor cannot be written as a product of two factors where the first does not contain x and the second does not contain y.
- The corresponding Markov network contains an *undirected* edge between x and y for all nodes x and y occurring together in a factor.
- A Bayesian network may generally be converted into a Markov network using a process called *moralization*.

Conditional independence in Markov networks

Given a Markov network and a set X of variables.

- A *Markov blanket* Z is a set of variables such that $X \coprod Y \mid Z$ where Y is any collection of variables not in X or Z.
- A *Markov boundary* is a minimal Markov blanket.
- The Markov boundary consists of all variables directly linked to X in the Markov network.
- Given a probability density on a set of variables, it can be specified as the set of conditional distributions of each variable given its Markov boundary.
- However, specifying a conditional distribution for each variable given its neighbours in a graph does not always result in a probability density for all variables.

Simulation in Markov networks using Gibbs sampling

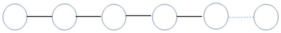
- With a Markov network representation of a posterior, we can set up a Gibbs sampling from the posterior by iteratively simulating from the conditional distribution of each node given its Markov boundary.
- Explicitly: Write down the joint density of all variables, and for each variable θ_i in sequence:
  - Regard all other variables as constants, throw away all factors not depending on θ_i.
  - Interpret the remaining function of θ_i as a standard density, or use it in some more advanced simulation method.
- Note: You need to check that the joint density is *proper*.
- We may simulate from a posterior represented as a Bayesian network by converting it to a Markov network (using moralization) and then simulate as above.
- Widely used programs like BUGS (WinBugs, OpenBugs), Jags (Just Another Gibbs Sampler), and **Stan** offer "black box" implementations of Gibbs sampling on wide classes of Bayesian Networks.

Gaussian Markov random fields (GMRF)

- A density $\pi(x_1, \ldots, x_n)$ can be considered a GMRF if it can be written as
$$\pi(x_1, \ldots, x_n) = \exp\left(-f(x_1, \ldots, x_n)\right)$$
where $f(x_1, \ldots, x_n)$ is a quadratic polynomial.
- We can then always re-write the density on $x = (x_1, \ldots, x_n)$ so that
$$\pi(x) = \exp\left(-\frac{1}{2}(x-\mu)^t P(x-\mu) + C\right).$$
where μ is a vector, P is a symmetric matrix, and C is a constant.
- The density is *proper* if and only if P is *positive definite*. In this case we can re-write the density as
$$\pi(x) = \frac{1}{|2\pi P^{-1}|}\exp\left(-\frac{1}{2}(x-\mu)^t P(x-\mu)\right),$$
so that $x \sim \text{Normal}(\mu, P^{-1})$.
- In many cases it may be useful to consider the Markov network for the GMRF.

GMRF and precision matrices

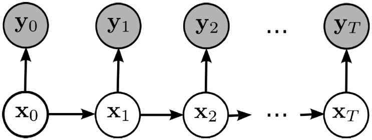
- For a GMRF and two variables x_i and x_j, the following are equivalent:
  1. There is no line between x_i and x_j in the Markov network.
  2. In the term $a_{ij}x_ix_j$ in the quadratic polynomial f defining the density, we have $a_{ij} = 0$.
  3. In the precision matrix P, the ij-th entry p_{ij} is zero.
- Thus, we can read off the Markov network directly from the precision matrix: Its non-zero terms correspond to edges in the Markov network.
- Example: If P is zero everywhere except along the main diagonal and the diagonals closest to it (i.e., $p_{ij} = 0$ unless $|i - j| \leq 1$) then the Markov network looks like the graph below (with number of nodes corresponding to number of variables).



Inference for graphical models (BNs or Markov networks)

- Two types of inference:
  - Given a network, and given observed values for some variables, how can we make predictions for (or simulate from) some remaining variables using the conditional distribution?
  - Given observations for some variables, how do we find a graphical model for these variables from the data?
- The second goal above, learning networks from data, can be extremely difficult. Active area of research.
- For the first question, several options exist, for example:
  - Doing Metropolis Hastings on the joint density of the variables (if not too many).
  - Using the network structure and simulate from the posterior using Gibbs sampling.
  - Using the network structure for exact or approximate inference with algorithms similar to those used with State Space Models / Hidden Markov Models.

Revisiting SSM/HMM



- We may prove that $\{y_{i+1}, \ldots, y_T\} \coprod \{y_0, \ldots, y_i\} \mid x_i$ using d-separation.
- It follows that $\pi(y_{i+1}, \ldots, y_T \mid x_i, y_0, \ldots, y_i) = \pi(y_{i+1}, \ldots, y_T \mid x_i)$ and thus Bayes formula gives

$$\pi(x_i \mid y_0, \ldots, y_T) \propto_{x_i} \pi(y_{i+1}, \ldots, y_T \mid x_i)\pi(x_i \mid y_0, \ldots, y_i)$$

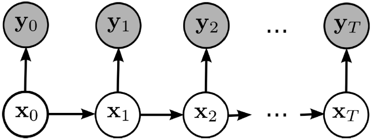
- We can use "Forward" and "Backward" algorithms to recursively compute, respectively,

$$\pi(x_i \mid y_0, \ldots, y_i)$$

and

$$\pi(y_{i+1}, \ldots, y_T \mid x_i).$$

Revisiting SSM/HMM



- Forward: For $i = 0, \dots, T$ compute $\pi(x_i \mid y_0, \dots, y_i)$ using

$$
\begin{aligned}
& \pi(x_i \mid y_0, \dots, y_i) \\
\propto_{x_i} \quad & \pi(y_i \mid x_i)\pi(x_i \mid y_0, \dots, y_{i-1}) \\
= \quad & \pi(y_i \mid x_i) \int \pi(x_i \mid x_{i-1})\pi(x_{i-1} \mid y_0, \dots, y_{i-1})\, dx_{i-1}
\end{aligned}
$$

- Backward: For $i = T-1, \dots, 0$ compute $\pi(y_{i+1}, \dots, y_T \mid x_i)$ using

$$
\begin{aligned}
& \pi(y_{i+1}, \dots, y_T \mid x_i) \\
= \quad & \int \pi(y_{i+2}, \dots, y_T \mid x_{i+1})\pi(y_{i+1} \mid x_{i+1})\pi(x_{i+1} \mid x_i)\, dx_{i+1}
\end{aligned}
$$

The message-passing algorithm

To generalize the ideas above to a general Markov network:

- Represent groups of variables with new variables such that the resulting Markov network becomes a tree.
- Propagate "messages" (i.e., densities) through the tree with algorithms similar to the Forward and Backward algorithms.
- This makes it possible to find the marginal distribution at each node of the tree, and thus for each variable.
- May be called the sum-product algorithm when the variables have a finite number of possible values.

Summary: Posterior inference for graphical models

- We want to fix some variables (called *data*) and compute the posterior distribution of *some* other variables of interest.
- For a Markov network, fixing some variables produces directly another similar Markov network.
- A Bayesian Network may first be converted to a Markov network, using moralization.
- Run a version of a message passing algorithm: The details vary with the type of variables and conditional distributions:
  - When all variables have a finite number of possible values, computations can be done exactly.
  - Exact computations can also be done when all conditional distributions are multivariate normal.
  - In most other cases, one must use approximations. Example: Particle filters.