MSA101/MVE187 2018 Lecture 11

Petter Mostad

Chalmers University

October 9, 2018

Overview of today

- Graphical models. Bayesian networks and Markov networks.
- Conditional independencies and graphical models.
- Gibbs sampling for graphical models.
- Learning networks.
- Causal networks.

Graphical representations of conditional independencies

- In complex models with many variables, it is crucial to model and keep track of 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 (posterior) 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 random fields: Represent the (posterior) 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

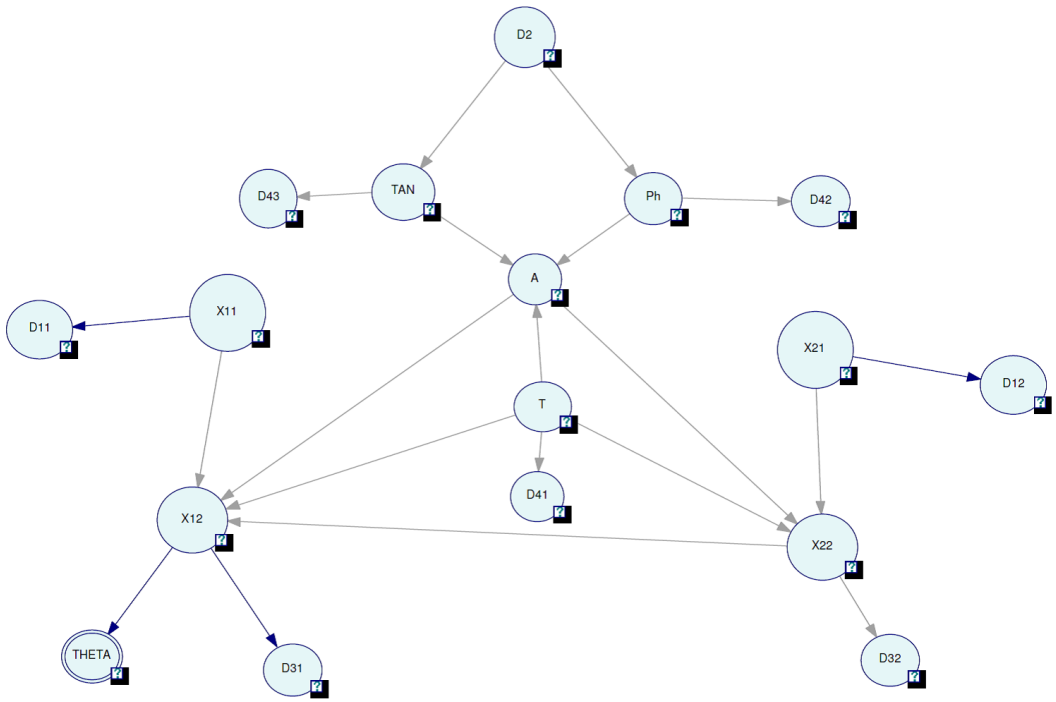
- Any joint distribution can be written as a product over conditional distributions:

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

- 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 $\pi(y \mid \ldots x \ldots)$ 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 distributions for each child given its parents, the product of these gives a joint probability distribution. (Show this).

Bayesian networks for visualization

- Examples.
- Hierarchical models.
- Using repeated graph components.



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:
  - $\pi(x, y \mid z) = \pi(x \mid z)\pi(y \mid z)$
  - $\pi(x \mid y, z) = \pi(x \mid y)$
  - $\pi(z \mid y, z) = \pi(z \mid y)$
- 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

- Conditional independence statements can be "read off" the DAG of a Bayesian network. Examples...
- Note: Conditioning on children may create dependencies.
- 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. (An undirected path in the DAG is a "trail"; it is "active" given Z if, for any "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).
- Theorem: If X and Y are d-separated given Z in a Bayesian network representation of a stochastic model, then $X \perp\!\!\!\perp Y \mid Z$.
- 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 details.

Markov networks

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

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

- 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.
- Examples.
- Note: A Bayesian network may generally be converted into a Markov network using *moralization*.

Conditional independence in Markov networks

- For a variable x, its *Markov blanket* Z is the set of variables directly connected to x in the Markov network representation.
- We then have $x \coprod Y \mid Z$ for any set Y of variables not containing x or Z.
- We define in the same way the Markov blanket of a set of variables X; the same holds conclusion about conditional independence holds.
- Examples
- A way to specify a stochastic model on a set of variables is to construct a graph connecting the variables and specify the conditional distribution of each variable given values of the variables it is connected to. NOTE: This does not necessarily result in a *proper* distribution!

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 blanket.
- Examples.
- Note: In order to simulate from the posterior, we need to know it is *proper*. This is not always the case for Markov networks.
- 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
$$\pi(x) = \exp\left(-\frac{1}{2}(x-\mu)^t Q(x-\mu) + C\right).$$
where μ is a vector, Q is a symmetric matrix, and C is a constant.
- The density is *proper* if and only if Q is *positive definite*. In this case we can re-write the density as
$$\pi(x) = \frac{1}{|2\pi P^{-1}|} \exp\left(-\frac{1}{2}(x-\mu)^t P(x-\mu)\right),$$
where P is a scalar multiple of Q, 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, if X, Y, and Z are groups of variables and we write

$$P = \begin{bmatrix} P_{XX} & P_{XY} & P_{XZ} \\ P_{YX} & P_{YY} & P_{YZ} \\ P_{ZX} & P_{ZY} & P_{ZZ} \end{bmatrix}$$

  for the precision matrix of their joint distribution, we have $X \coprod Y \mid Z$ if and only if $P_{XY} = 0$.
- Examples.

A note about multivariate normal distributions

Assume the joint distribution for variables $Z_1, \ldots, Z_k$ is multivariate normal. Then

- If we integrate out Z_i, the convariance matrix for the remaining variables is equal to the submatrix corresponding to these variables of the covariance matrix for the joint distribution. (We knew this).
- If we fix Z_i, the *precision matrix* for the remaining variables is equal to the submatrix corresponding to these variables of the *precision matrix* for the joint distribution. See below:

Given the joint normal distribution $[\theta_1, \theta_2] \sim \text{Normal}\left([\mu_1, \mu_2], \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}^{-1}\right)$, we get the conditional distribution $\theta_1 \mid \theta_2 \sim \text{Normal}\left(\mu_1 - P_{11}^{-1}P_{12}(\theta_2 - \mu_2), P_{11}^{-1}\right)$. Proof: Use the identity

$$\begin{aligned} & \left(\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)^t \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \left(\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right) \\ = \quad & \left(\theta_1 - \mu_1 + P_{11}^{-1}P_{12}(\theta_2 - \mu_2)\right)^t P_{11} \left(\theta_1 - \mu_1 + P_{11}^{-1}P_{12}(\theta_2 - \mu_2)\right) \\ & + (\theta_2 - \mu_2)^t (P_{22} - P_{21}P_{11}^{-1}P_{12})(\theta_2 - \mu_2). \end{aligned}$$