MSA101/MVE187 2017 Lecture 11

Petter Mostad

Chalmers University

October 3, 2017

Graphical representation of conditional independencies

- Any stochastic model can be described as a set of variables and a joint probability distribution for them.
- The key to describing, understanding, and computing with such models is to describe conditional independencies between the variables.
- Various graphical model types use various definitions to represent conditional independencies. The main advantages are:
  - Visualization and communciation of models.
  - The possibiliy to apply graph-theoretic algorithms.
- Several ways of employing graphs to represent independencies exist. We will consider the two most important:
  - Bayesian Networks, using directed acyclic graphs (DAGs)
  - Markov Networks, using undirected graphs.

Bayesian Networks

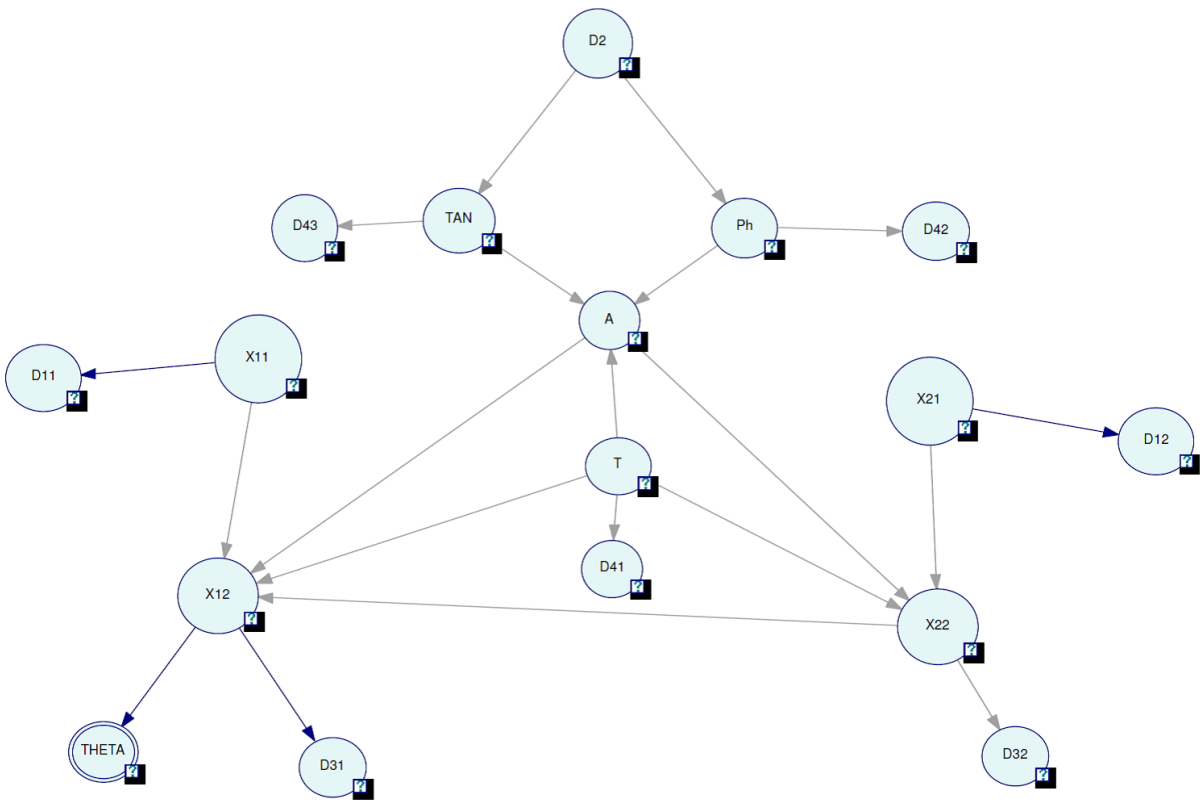
- A Bayesian Network consists of
  - A *directed acyclic graph* (DAG), and
  - For each node i in the network, a variable x_i and a conditioinal probability distribution

    $$\pi(x_i \mid X_{\mathrm{Pa}(i)})$$

    where $X_{\mathrm{Pa}(i)}$ is the set of variables corresponding to the *parent nodes* of i in the network.
- The corresponding stochastic model is the product over the conditional probability distributions.
- Note:
  - The same stochastic model can be represented with many different graphs.
  - The graphs do not necessarily represents *causality* (we look at this later).
  - Statements about conditional independencies can be computed from the graph.

Examples

- Markov chains, hidden Markov models, ...
- Hierarchical models.
- Models specified directly as Bayesian Networks: An example:



Conditional independencies in Bayesian Networks

- Nodes that do not share *decendants* are conditionally independent given their shared *ancestors*.
- Note: Fixing the value of a common *decendant* of two variables can make them *dependent*.

A taste of graph theory for Bayesian Networks

- A *trail* between nodes a and b in a DAG is a path disregarding the arrow directions.
- A trail from a to b is *blocked* by a set of nodes S if it contains a node γ such that
  - either the trail meets in a "V" at γ and γ and its decendants are not in S,
  - or the trail does *not* meet in a "V" at γ, and $\gamma \in S$.
- If *all* trails between a and b are blocked by S for all $a \in A$ and $b \in B$, then A and B are said to be *d-separated* by S.
- If A and B are d-separated by S in a BN, then $A \perp\!\!\!\perp B \mid S$, meaning that A and B are conditionally independent given S.
- If A and B are *not* d-separated by S in a DAG, then there exists a BN with this DAG as its graph where we do *not* have $A \perp\!\!\!\perp B \mid S$.

Markov networks

- A Markov Network consists of
  - an undirected graph,
  - for each node in the graph, a variable, and
  - a set of functions (called *factors*) on subsets of variables, such that all nodes corresponding to the variables in the subset are connected by edges.
- The product of all the factors is the joint density of the model IF the product has a finite integral (or sum) over the variables, so that it can be scaled to 1.
- Conditional independencies can be deduced from the graph.
- All Bayesian Networks can be transformed into Markov networks.
- There are joint distributions whose set of conditional independencies can be represented with a Markov network, but not with a Bayesian network.
- An important issue when using these networks is to ensure that the joint distribution is proper.

Example: Gaussian Markov random fields

- A Gaussian Markov random field (GMRF) is one where all the conditional distributions are normal, as follows:

$$Z \mid Z_1, \ldots, Z_k \sim \text{Normal}(\mu_Z + \beta_1 Z_1 + \cdots + \beta_k Z_k, \Sigma_Z)$$

- The joint distribution of all variables in the network becomes multivariate normal, IF it is proper.
- For example, one may set up a spatial model identifying a set of *neighbours* for each variable, and define its conditional distribution to depend only on the neighbours, specifically to be a normal distribution with expectation equal to the mean of the neighbours. This model is initially not proper; one may for example add the condition that the mean of all variables should be zero to make it proper.

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* (inverse covariance 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

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

GMRFs vs. precision matrices

For a Gaussian Markov random field, the following holds:

- If a corresponding Markov network does not contain an edge between two variables, the corresponding entry of the precision matrix for the joint distribution is zero.
- Given a precision matrix for a joint distribution, we can construct a factorization with a Markov network which only contains edges where the precision matrix is non-zero.
- Thus, as Gaussian Markov random fields tend to be defined with simple Markov networks, they have *sparse* precision matrices, which is taken advantage of in computational methods.

Computations for Graphical models

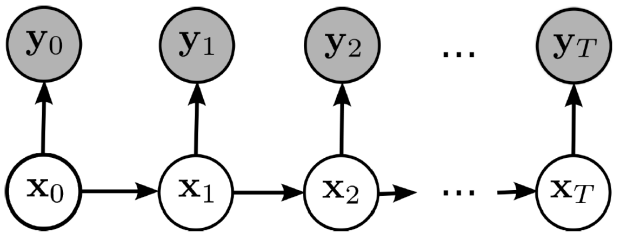
- Given a graphical model, we might want to
  - Follow the Bayesian paradigm and find the conditional distribution of some nodes (variables) given fixed values (data) for some other nodes. Two approaches: *Simulation* or *exact inference*.
  - Given fixed values (data) for some nodes, find the maximum aposteriori (MAP) for some other set of nodes, i.e., their values maximizing the posterior density. *As an example of this, one might estimate the ML values of parameters in a network specified with unknown parameters*.
- Given data and prior knowlege, one may want to *learn* the structure of a suitable BN or Markov model.

Exact posterior inference for graphical models

- Given a model represented as a Bayesian Network (or Markov network), the goal of inference (as in any Bayesian computation) is to compute the marginal distribution of some variables of interest, conditionally on fixing some other variables, called *data*.
- For a Markov network, fixing some variables produces directly another similar Markov network.
- A Bayesian Network may first be converted to a Markov network.
- A direct way to obtain a marginal distribution in a Markov network is *variable elimination*:
  - Integrating (or summing) out variables in factors.
  - Multiplying together factors.
- Any inference algorithm depends on the basic operations above, but they can be "scheduled" in smart ways, using e.g. "message passing" algorithms.

Example: The Forward-Backward algorithm

Message passing applied to the following Bayesian Network: A *Hidden Markov Model*



Objective: Compute the marginal posterior distribution of every x_i given data $y_0, \ldots, y_T$: Use $\pi(x_i \mid y_0 \ldots, y_T) \propto \pi(y_{i+1}, \ldots, y_T \mid x_i)\pi(x_i \mid y_0, \ldots, y_i)$ and

1. Forward: For $i = 0, \ldots, T$ compute $\pi(x_i \mid y_0, \ldots, y_i)$ using

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

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

$$\pi(y_{i+1}, \ldots, y_T \mid x_i) = \int \pi(y_{i+2}, \ldots, y_T \mid x_{i+1})\pi(y_{i+1} \mid x_{i+1})\pi(x_{i+1} \mid x_i)\, dx_{i+1}$$