STAT 535 Lecture 3
Graphical representations of conditional independence
Part I Markov Random Fields

©Marina Meilă
mmp@stat.washington.edu

Reading KF 2.2 (graphs), 4.2, 4.3, (4.4. additional topics)

1 Representing independence in graphs

Graphical model = graphical representation of (conditional) independence relationships in a joint distribution
= the distribution itself

graphical model - **structure** (a graph)
- **parametrization** (depends on the graph, parameters are "local")

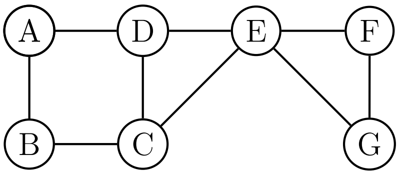
A graph is defined as $G = (V, \mathcal{E})$ where

- V is the set of graph **vertices** (or **nodes**); each node represents a variable
- $\mathcal{E}$ is the set of graph **edges**; edges encode the dependencies.

More precisely: a missing edge encodes an independence relationship.

Idea: Independence in the joint distribution $\longleftrightarrow$ **Separation** in graph

This mapping is **not unique** and **not perfect**. But even so, graphical representations are useful. Allowing for efficient computations is the main reason. Helping scientists understand a problem and communicate about it is another reason.



Examples: $F, G \perp A, B, C, D \mid E$
$A \perp C \mid B, D$
$A \perp C \mid B, D, E, F$

Figure 1: An undirected graph and some independencies encoded by it.

There are multiple "languages" for representing independence relations in graphs. The most popular ones are **Markov random fields** and **Bayesian Networks**. Later on we will also study **factor graphs, decomposable models** and **junction tree** representations. These graphical representations are tools for understanding and designing inference algorithms.

2 Markov Random Fields (Markov networks)

2.1 Encoding independencies in undirected graphs

An arbitrary undirected graph can be seen as encoding a set of independencies. The following rules states when two sets of variables $U_1, U_2 \subseteq V$, $U_1 \cap U_2 = \emptyset$ are **separated** in an undirected graph. We denote separation by $\perp$ and take it to mean "independence in the joint distribution over V".

$$\boxed{U_1 \perp U_2 \mid U_3 \iff \text{all paths between sets } U_1 \text{ and } U_2 \text{ pass through set } U_3}$$

We say that U_3 **blocks** the paths between X and Y; think of it as "blocking the flow of information".

$n(A)$ = the **neighbors** of variable A

A consequence of this rule is the following **Markov property for MRFs** also called the **local Markov property**:

$$\boxed{A \perp \text{everything else} \mid n(A)}$$

A set of variables that separates node A from the rest of the graph is called a **Markov blanket** for A. The set $n(A)$ is thus a Markov blanket, and it is the minimal Markov blanket of A. Adding a node to a Markov blanket preserves the Markov blanket property.

2.2 I-maps and perfect maps

Let us consider an undirected graph $G = (V, \mathcal{E})$ and a probability distribution P over the set of variables V. If every separation relationship "U_1 separated from U_2 by U_3" in G corresponds to a conditional independence $U_1 \perp U_2 \mid U_3$ under P, then we say that the G is an **I-map** (independence map) of P.

One can think of a graph G as a representant of the family of all probability distributions on V for which G is an I-map. Some of these distributions may have additional independencies, which do not appear in G. For example, any G is an I-map for the distribution over V in which all the variables are mutually independent.

If G is an I-map of and P has no independencies except for those represented by G, the we say that G is a **perfect map** for P.

Any undirected graph G has a perfect map (Geiger and Pearl,88), but not any P has a perfect map as an undirected graph. The distributions representable by graphs are a subclass of all distributions. An example of limitation imposed by the graph representation is that, in any graph (see for instance the graph in Figure 1) $A \perp G \mid E$ and $B \perp G \mid E$ is equivalent to $\{A, B\} \perp G \mid E$. However, this is not always true in a distribution (example: the parity function).

2.3 Factorization

Now we will characterize the set of distributions for which a graph G is an I-map. For this, we need a new definition. A **clique** of a graph G is a set of nodes $C \subseteq V$ which are **fully connected** in G (i.e all possible edges between nodes in C appear in $\mathcal{E}$). A **maximal** clique is a clique which is not contained in any other clique of the graph.

For example, in figure 1, all the nodes are cliques of size one (but not maximal), all the edges are cliques of size two, and the triangles CDE, EFG are cliques of size three. The maximal cliques are AB, BC, AD, CDE, EFG.

Theorem 1 *Let G be a graph and assume P can be factored in the following way*

$$P = \prod_{C \text{maximal clique}} \phi_C(x_C) \tag{1}$$

where ϕ_C is a non-negative function depending only on the variables in C. Then, G is an I-map of P.

We will illustrate this theorem by an example shortly. The converse is a more powerful result, and is known as the Hammersley-Clifford theorem.

Theorem 2 (Hammersley-Clifford) *If $P > 0$ and G is an I-map of P, then P can be written as a product of functions defined over the cliques of G as in (1)*[1].

Exercise The theorem doesn't always hold if $P(x) = 0$ for some x. Can you construct such a counterexample? (Hint: give P lots of zeros.)

If a distribution P can be written in the form (1) for some graph G we say that P **factors according to graph** G.

[1] The original Hammersley-Clifford theorem is stronger; it only assumes that P obeys the local Markov property according to G

Example Factorization for the undirected G in figure 1

$$P_{ABCDEFG} = \phi_{AB}(a,b)\phi_{AD}(a,d)\phi_{BC}(b,c)\phi_{CDE}(c,d,e)\phi_{EFG}(e,f,g)$$

The functions ϕ are called **clique potentials**. They are required to be non-negative (positive if $P > 0$). Clique potentials are not uniquely defined. One can obtain equivalent factorizations by dividing/multiplying with functions of variables that are common between cliques. For instance, we can rewrite the above joint distribution as

$$\begin{aligned}
&P_{ABCDEFG} = \\
&= (2\phi_{AB}(a,b))(\phi_{AD}(a,d)/2)\phi_{BC}(b,c)\phi_{CDE}(c,d,e)\phi_{EFG}(e,f,g) \\
&= (h(a)\phi_{AB}(a,b))(\phi_{AD}(a,d)/h(a))\phi_{BC}(b,c)\phi_{CDE}(c,d,e)\phi_{EFG}(e,f,g) \quad \text{for any } h(a) > 0 \\
&= \Phi_{AB}(a,b)\phi_{AD}(a,d)\phi_{BC}(b,c)(\phi'_{CDE}(c,d,e)h(c,d))\phi_{EFG}(e,f,g)
\end{aligned}$$

The last example shows why we only need to consider maximal cliques in the factorization of P. Because of the non-unicity of the ϕ's, the parameters of the clique potentials are hard to interpret. The potentials do not, in general, represent probability tables. However, there are some important special cases when the ϕ's have probabilistic interpretations – these will be the decomposable models we will study later. The Hidden Markov model you have already encountered is one of them.

2.4 The clique potentials are not marginals in Markov Random Fields - an example

The following simple example shows that *potential* $\neq$ *marginal* even if the potential is normalized.

Let $V = \{A, B, C\}$, $\mathcal{E} = \{AB, BC, CA\}$ and

$$\phi_{AB} = \phi_{BC} = \phi_{AC} = \begin{bmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{bmatrix}$$

Note that this is not exactly a Markov field, as the potentials are given on the edges, not on the maximal clique ABC. However, we shall use this example for simplicity (otherwise we'd need 4 or more nodes to prove our point). Namely, we will show that $P_{AB} \not\propto \phi_{AB}$.

$$\begin{aligned}
P_{AB}(0,0) &\propto \phi_{AB}(0,0)\sum_{C}\phi_{AC}(0,c)\phi_{BC}(0,c) = \frac{1}{3}\frac{1}{3}\frac{1}{3}+\frac{1}{3}\frac{1}{6}\frac{1}{6} = \frac{5}{3^3\cdot 4}\\
P_{AB}(0,1) &\propto \phi_{AB}(0,1)\sum_{C}\phi_{AC}(0,c)\phi_{BC}(1,c) = \frac{1}{6}\frac{1}{6}\frac{1}{3}+\frac{1}{6}\frac{1}{3}\frac{1}{6} = \frac{2}{3^3\cdot 4}
\end{aligned}$$

By symmetry, $P_{AB}(1,1) = P_{AB}(0,0)$ and $P_{AB}(0,1) = P_{AB}(1,0)$. Hence

$$Z = 2(P_{AB}(0,0)+P_{AB}(0,1)) = \frac{7}{54}$$

and

$$\frac{P_{AB}(0,0)}{P_{AB}(0,1)} = \frac{5}{2} \neq \frac{\phi_{AB}(0,0)}{\phi_{AB}(0,1)} = \frac{2}{1}$$

2.5 Where do the ϕ potentials come from?

Sometimes, they come from physical models, where $(-\log\phi)$ represents an energy. This is the case of the Ising model in lecture 1. Note that the potential energy is defined up an additive constant; this fits with the ϕ potential being defined up to multiplicative constants.

Other times, they are "made up" by e.g engineers who want to represent a problem. For example the lattice models representing images, are MRF's where the graph and the potential functions are artificial (but useful) representations for images.

Sometimes, the potentials are obtained by a combination of scientific grounds, estimation from data, and convenience consideration. This is the case in the modeling of spatial processes. In such processes, the graph can represent: a grid of locations where weather measurements are taken (an irregular network), the states of the US, with edges between neighboring states (for the study of e.g ecological processes), a watershed (points and edges along rivers), a transportation network or a social contacts network (in epidemiology), etc.

In a factored representation the savings in terms of number of parameters w.r.t the multidimensional table representation are significant. Assume that all variables are binary, and all potentials are represented by (unnormalized) tables. Then for the graph in figure 1 the total number of parameters is

$$3 \times 2^2 + 2 \times 2^3 \;=\; 28$$

The size of a probability table over 7 binary variables is $2^7 - 1 = 127$ thus in this example we save 99 parameters (almost 80%).

2.6 The example continued: factorization (1) and the independencies prescribed by G

We will consider the example in figure 1. Let's prove that the factorization (1) implies that

$$\{A, B\} \perp F \,|\, C, D \quad \text{in } P$$

if P factors according to the graph.

We will use the following fact about distributions:

Lemma 3 *$X \perp Y \,|\, Z$ under P iff there exist functions $h_1(x,z), h_2(y,z)$ such that $P_{XYZ}(x,y,z) = h_1(x,z)h_2(y,z)$.*

Exercise Prove the lemma.

To use Lemma 3 in our proof, we write P_{ABCDF} in product form.

$$\begin{aligned}
& P_{ABCDF}(a,b,c,d,f) \;= && (2)\\
&\quad= \sum_{e,g} P_{ABCDEFG} && (3)\\
&\quad= \sum_{e,g} \phi_{AB}(a,b)\phi_{AC}(a,c)\phi_{BD}(b,d)\phi_{CDE}(c,d,e)\phi_{EFG}(e,f,g) && (4)\\
&\quad= \phi_{AB}(a,b)\phi_{AC}(a,c)\phi_{BD}(b,d)\sum_{e}\phi_{CDE}(c,d,e)\underbrace{\sum_{g}\phi_{EFG}(e,f,g)}_{\psi_{EF}(e,f)} && (5)
\end{aligned}$$

$$
\begin{aligned}
&= \phi_{AB}(a,b)\phi_{AC}(a,c)\phi_{BD}(b,d)\underbrace{\sum_e \phi_{CDE}(c,d,e)\psi_{EF}(e,f)}_{\psi_{CDF}(c,d,f)} && (6)\\
&= \underbrace{[\phi_{AB}(a,b)\phi_{AC}(a,c)\phi_{BD}(b,d)]}_{h_1(a,b,c,d)}\underbrace{\psi_{CDF}(c,d,f)}_{h_2(c,d,f)} && (7)
\end{aligned}
$$

From the last form of the marginal P_{ABCDF} we can conclude that the relationship $\{A,B\} \perp F \mid C,D$ is true under P.

From the factorization, we can also see that $\{A,B\} \not\perp F \mid D$ in general. For this, we start from the expression of the joint P_{ABCDF} and marginalize over C to obtain P_{ABDF}.

$$
\begin{aligned}
P_{ABDF}(a,b,d,f) &= && (8)\\
&= \sum_c P_{ABCDF} && (9)\\
&= \sum_c \phi_{AB}(a,b)\phi_{AC}(a,c)\phi_{BD}(b,d)\psi_{CDF}(c,d,f) && (10)\\
&= \phi_{AB}(a,b)\phi_{BD}(b,d)\underbrace{\sum_c[\phi_{AC}(a,c)\psi_{CDF}(c,d,f)]}_{\psi_{ADF}(a,d,f)} && (11)\\
&= \phi_{AB}(a,b)\phi_{BD}(b,d)\psi_{ADF}(a,d,f) && (12)
\end{aligned}
$$

This expression contains the factor ψ_{ADF} that depends on both A, and F. Thus, variable sets A, B and F cannot in general be separated if we conditon on D only. Hence, this independence relationships does not hold in all distributions that factor according to the graph in figure 1. (However, there *may be* clique potentials for which the function ψ_{ADF} can be decomposed into a product for which A, F are separated. This set of distributions is a measure zero set in the set of all MRF's that factor according to this graph.)

2.7 Gaussian Markov fields

A multivariate normal distribution over has the form

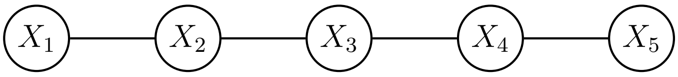
$$
P_V(x_V) \propto e^{-\frac{1}{2}(x_V-\mu_V)^T\Sigma^{-1}(x_V-\mu_V)} \qquad (13)
$$

If we denote the inverse covariance by $D = \Sigma^{-1}$ and if, for simplicity, we assume $\mu_V \equiv 0$, then the multivariate normal over V can be written as

$$\begin{aligned} P_V(x_V) &\propto e^{-\frac{1}{2}x_V^T D x_V} && (14) \\ &= e^{-\frac{1}{2}\sum_{i\in V} D_{ii}x_i^2 - \sum_{i<j} D_{ij}x_i x_j} && (15) \\ &= \prod_{i\in V} \underbrace{e^{-\frac{1}{2}D_{ii}x_i^2}}_{\psi_i} \prod_{i<j} \underbrace{e^{-D_{ij}x_i x_j}}_{\phi_{ij}} && (16) \end{aligned}$$

This is equivalent with a MRF that has an edge for every non-zero D_{ij}. Conversely, in multivariate Gaussian, the zeros in the inverse covariance matrix encode the conditional independencies.

2.8 Markov chains



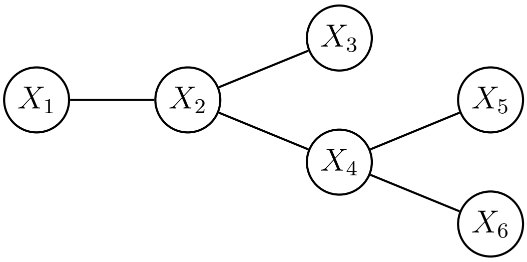
The joint distribution

$$P_{X_1X_2X_3X_4X_5} = (P_{X_1}P_{X_2|X_1})P_{X_3|X_2}P_{X_4|X_3}P_{X_5|X_4}$$

is a product of conditional distributions involving X_{t+1}, X_t. X_{t+1}, X_t are neighbors in the chain. In this case the clique potentials, one for each edge, are $\phi_{1,2} = P_{X_1}P_{X_2|X_1}$, $\phi_{i,i+1} = P_{X_{i+1}|X_i}$, for $i > 1$, and have a probabilistic interpretation.

The same can be said about the Hidden Markov Model.

2.9 Trees



Tree = connected graph with no cycles (we also call it **spanning** tree). If disconnected and no cycles, we call it a **forest**. Sometimes we use the term tree to mean either a spanning tree or a forest.

Property: between every two variables in a spanning tree there is exactly one path (at most one path for forests).

All cliques in a spanning tree have size 2.