UGMs: Markov Properties
10708, Fall 2020
Pradeep Ravikumar

1 Global Markov Properties

Consider an UG $G = (V, E)$, and where we associate the nodes V with variables in a random vector X. In the sequel, we will use nodes $s \in V$ and RVs X_s interachangeably. We now set out to — just by inspecting the graph — specify a set of conditional independence "Markov" properties that X could satisfy.

We first recall some graph-theoretic notation. Given a graph G, $X - X_{v_1} - X_{v_2} - \ldots - X_{v_k} - Y$ is said to be a path of length k between two nodes X and Y iff $(X_i, X_{i+1}) \in E(G)$, for $i \in [k-1]$. It is said to be an active path given a set of nodes $\mathbf{Z}$ if none of the nodes in the path are contained in $\mathbf{Z}$. We say a set of nodes $\mathbf{Z}$ **separate** $\mathbf{X}$ from $\mathbf{Y}$ in UG G iff there is no active path from any node $X \in \mathbf{X}$ to $\mathbf{Y}$ given $\mathbf{Z}$. It is a very natural notion: if we remove nodes in $\mathbf{Z}$ from the graph, $\mathbf{X}$ and $\mathbf{Y}$ lie in disconnected components. We term this $\text{SEP}_G(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})$.

Given an UG G, we can then write down a list of conditional independencies:

$$\mathbb{I}(G) = \{\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \,:\, \text{SEP}_G(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})\}.$$

These are sometimes also called the "global Markov properties" associated with the UG G. An interesting question is whether a UGM distribution P that *factors according to* G satisfies the global Markov properties associated with G.

1.1 Soundness

Proposition 1 *Any distribution P that* factors according to *G satisfies the global Markov properties associated with G i.e. $I(G) \subseteq I(P)$, where $I(P)$is the set of all conditional independencies satisfied by P.*

Proof. Consider three disjoint subsets $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ such that $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z} = \mathcal{X}$. Suppose we have that $\text{SEP}_G(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})$. So there are no direct edges between nodes in $\mathbf{X}$ and nodes in $\mathbf{Y}$. So any clique in $\mathcal{C}(G)$ is contained in

$$\mathbb{I}_{\mathbf{X}} = \{C \in \mathcal{C}(G) \;\; C \subseteq \mathbf{X} \cup \mathbf{Z}\},$$

or in

$$\mathbb{I}_{\mathbf{Y}} = \{C \in \mathcal{C}(G) \;\; C \subseteq \mathbf{Y} \cup \mathbf{Z}\}.$$

Since P factors according to G, we have that:

$$P(X) = \frac{1}{Z} \prod_{C \in \mathbb{I}_{\mathbf{X}}} \Phi_C(X_C) \prod_{C' \in \mathbb{I}_{\mathbf{X}}} \Phi_{C'}(X_{C'}).$$

From the definition of $\mathbb{I}_{\mathbf{X}}$ and $\mathbb{I}_{\mathbf{Y}}$, we can rewrite this simply as:

$$P(X) = \frac{1}{Z} f(\mathbf{X}, \mathbf{Z})\, g(\mathbf{Y}, \mathbf{Z}),$$

for some functions $f(\cdot), g(\cdot)$. From this factorized form it follows from a simple algebraic calculation that P satisfies the conditional independence $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$.

The proof for the case $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z} \subset \mathcal{X}$ follows from a simple extension of above argument. □

It is worth pausing here to reflect on the remarkable result above: any distribution that factors according to the cliques of a graph satisfies the set of all global Markov properties associated with that graph. We thus obtain a very simple characterization of (at least some of) the *constraints* satisfied by any UGM P associated with a graph G. The above property is thus also called *soundness* of the global Markov properties associated with an UG.

A natural question to ask is if the converse holds. That is, would any distribution that satisfies global Markov properties associated with G also factor according to G. It turns out this does not hold in general. But it does hold for positive distributions, a famous result also known as the Hammersley-Clifford Theorem.

Theorem 2 (Hammersley-Clifford) *Any positive distribution P that satisfies the global Markov properties associated with G (so that $I(G) \subseteq I(P)$) also* factors according to G.

Proof. The proof relies on the so-called Mobius inversion lemma applied to set functions. Suppose $v : 2^{[}p] \mapsto \mathbb{R}$ is some function over sets $S \subseteq [p]$. Then there is a unique set function $u : 2^{[}p] \mapsto \mathbb{R}$ such that:

$$v(S) = \sum_{T \subseteq S} u(T),$$

and

$$u(S) = \sum_{T \subseteq S} (-1)^{|T|-|S|} v(T).$$

Thus, the set function $u(S)$ consists of "marginal" contribution of the set S, with respect to v, after marginalizing out the contributions of smaller subsets. And conversely, the set function $v(S)$ can be written as the sum of the marginal contributions of all its subsets.

Let $\mathbf{x}^*$ be some fixed "reference" value in $\mathcal{X}$. Then, for any value $\mathbf{x} \in \mathcal{X}$, and any distribution P, define the set function

$$v_P(S) = \ln P(\mathbf{x}_S, \mathbf{x}^*_{-S}),$$

and define the marginal contribution set function $u_P(\cdot)$ as above via Mobius inversion. We then have that:

$$\begin{aligned}\ln P(\mathbf{X}) &= v_P([p]) \\ &= \sum_{T \subseteq [p]} u_P(T).\end{aligned}$$

If we can then show that for any P that satisfies global Markov properties wrt G, has factors $u_p(T) = 0$ for any subset T that does not correspond to a clique in G, then we have shown that P factors according to G. To show this, consider a pair of nodes X, Y not connected by an edge. We will then show that any subset $T \subseteq [p]$ that contains both X and Y (and hence is not a clique), will have $u_P(T) = 0$. Now any such subset containing both X and Y consists of four groups of subsets: for any $W \subseteq T - \{X, Y\}$, we see that W, $W \cup \{X\}$, $W \cup \{Y\}$, and $W \cup \{X, Y\}$ also lie in T. We can thus write:

$$u_P(T) = \sum_{W \subseteq T - \{X,Y\}} (-1)^{|T| - |\{X,Y\}| - |W|} \left(v_P(W) - v_P(W \cup \{X\}) - v_P(W \cup \{Y\}) + v_P(W \cup \{X, Y\}) \right).$$

Now, let $U = [p] - T$. Denote: $\mathbf{u}^* = \mathbf{x}^*_U, x^* = \mathbf{x}^*_{\{}x\}, y^* = \mathbf{x}^*_{\{}y\}, \mathbf{w}^* = \mathbf{x}^*_W$. And correspondingly: $\mathbf{u} = \mathbf{x}_U, x = \mathbf{x}_{\{}x\}, y = \mathbf{x}_{\{}y\}, \mathbf{w} = \mathbf{x}_W$. We then have:

$$\begin{aligned}v_P(W \cup \{X, Y\} - v_P(W \cup \{X\}) &= \ln \frac{P(x, y, \mathbf{w}, \mathbf{u}^*)}{P(x, y^*, \mathbf{w}, bu^*)} \\ &= \ln \frac{P(y|x, \mathbf{w}, \mathbf{u}^*) P(x, \mathbf{w}, \mathbf{u}^*)}{P(y^*|x, \mathbf{w}, bu^*) P(x, \mathbf{w}, \mathbf{u}^*)} \\ &= \ln \frac{P(y|x^*, \mathbf{w}, \mathbf{u}^*) P(x, \mathbf{w}, \mathbf{u}^*)}{P(y^*|x^*, \mathbf{w}, bu^*) P(x, \mathbf{w}, \mathbf{u}^*)} \\ &= \ln \frac{P(y|x^*, \mathbf{w}, \mathbf{u}^*) P(x^*, \mathbf{w}, \mathbf{u}^*)}{P(y^*|x^*, \mathbf{w}, bu^*) P(x^*, \mathbf{w}, \mathbf{u}^*)} \\ &= \ln \frac{P(x^*, y, \mathbf{w}, \mathbf{u}^*)}{P(x^*, y^*, \mathbf{w}, bu^*)} \\ &= v_P(W \cup \{Y\}) - v_P(W),\end{aligned}$$

where the third equality follows from the conditional independence satisfied by P: $X \perp\!\!\!\perp Y \,|\, \mathcal{X} - \{X, Y\}$. It thus follows that $u_P(T) = \sum_{W \subseteq T - \{X,Y\}} 0 = 0$. $\square$

1.2 Completeness

Note that the previous "soundness" theorem entailed that any distribution that factors according to an UG G also satisfies the global Markov properties associated with G (and the converse holds as well for positive distributions). Now these global Markov properties

are a set of conditional **independencies**. A key question is whether we can also say the same thing for the corresponding set of conditional **dependencies**, that is any conditional independence assertion that does not lie in the set of global Markov properties associated with G.

That is, whether any distribution that factors according to UG G satisfies not just conditional independencies wrt G but also conditional *dependencies*. Formally, if it does not hold that $\text{SEP}_G(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, then $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z}$ is not satisfied by any distribution P that factors according to G. But this turns out to be too strong, and does not actually hold. Instead, we have the following weaker version of "completeness":

Theorem 3 *Suppose X and Y are not separated given $\mathbf{Z}$ in UG G. Then there exists a distribution P that factors according to G, where X and Y are dependent given $\mathbf{Z}$.*

2 Other Markov Properties

We can also associate other sets of conditional independencies with an UG G, which might be even more natural than the global Markov properties.

Definition 4 (Pairwise Markov properties) *Given a UG G, we define the set of pairwise Markov independencies as the set:*

$$\mathbb{I}_p(G) = \{X \perp\!\!\!\perp Y \,|\, \mathcal{X} - \{X, Y\} \,:\, (X, Y) \notin E(G)\}.$$

This is thus a very local pairwise restriction of the set of global Markov properties. Towards another natural local restriction, let us define the Markov blanket $\text{MB}_G(X)$of any node X wrt G as:

$$\text{MB}_G(X) = \text{NBRS}(X).$$

Definition 5 (Local Markov properties) *Given a UG G, we define the set of local Markov independencies as the set:*

$$\mathbb{I}_\ell(G) = \{X \perp\!\!\!\perp \mathcal{X} - \{X\} - \text{MB}_G(X) \,|\, \text{MB}_G(X)\}.$$

The following relationship can be easily shown to hold over these varied sets of Markov properties:

Proposition 6

$$\mathbb{I}_p(G) \subseteq I_\ell(G) \subseteq I(G).$$

A more interesting question is whether the corresponding entailed distributions are also strict subsets. It turns out that for general distributions these could indeed be strict subset relationships.

Example 7 *Let P be any distribution over $\mathcal{X} = \{X_1, \ldots, X_p\}$. Let $\mathcal{X}' = \{X'_1, \ldots, X'_p\}$, and $\mathcal{X}'' = \{X''_1, \ldots, X''_p\}$ Consider the distribution P' over $(\mathcal{X}, \mathcal{X}', \mathcal{X}'')$ such that $X''_s = X'_s = X_s$, for all $s \in [p]$. And consider the empty graph G' over $(\mathcal{X}, \mathcal{X}', \mathcal{X}'')$ with no edges. Clearly, for any pair of variables $Y_s, Z_{s'}$, they are conditionally independent given the rest of the nodes, since one of the nodes in the remainder exactly specifies Y_s. So P' satisfies pairwise Markov properties, but not all local or global Markov properties wrt G'.*

Example 8 *Let P be any distribution over $\mathcal{X} = \{X_1, \ldots, X_p\}$. Let $\mathcal{X}' = \{X'_1, \ldots, X'_p\}$. Consider the distribution P' over $(\mathcal{X}, \mathcal{X}')$ such that $X'_s = X_s$, for all $s \in [p]$. And consider the graph G' over $(\mathcal{X}, \mathcal{X}')$ such that the only edges are between (X_s, X'_s) for all $s \in [p]$. For any variable X_s, conditioned on its neighbor X'_s, it is independent of its non-neighbors. Then, P' satisfies the local Markov properties wrt G'. But it clearly does not satisfy all global Markov properties wrt G'.*

But for *positive* distributions, these are all equivalent.

Proposition 9 *For any positive distribution P, the following statements are equivalent:*

1. *P satisfies cond. independencies in $\mathbb{I}_p(G)$*
2. *P satisfies cond. independencies in $\mathbb{I}_\ell(G)$*
3. *P satisfies cond. independencies in $I(G)$*

2.1 From Distributions to Graphs

So far we have gone from graphs to distributions: given an UG G, we then specify distributions P that factor wrt G or satisfies Markov properties wrt G (and where the latter are equivalent for positive distributions). Now suppose we are given a distribution P over $\mathcal{X}$. Then, what is an UG G s.t. P satisfies Markov properties wrt G? This is a very easy question: we can simply output the complete graph G which has the empty set as its set of Markov properties. Which is obviously satisfied by the given distribution P.

Definition 10 (Minimal UGM) *Given a distribution P, we say that G is its minimal UG when*

$$I(G) \subseteq I(P),$$

and where the above is not satisfied for any graph G' with $E(G') \subset E(G)$.

Thus P factors wrt its minimal UG G, but removing any edge from G would entail that P no longer factors wrt G. It could nonetheless be the case that P satisfies additional conditional independencies not entailed by the minimal UG G.

Definition 11 (Perfect UGM) *Given a distribution P, we say that G is its perfect UG when $I(P) = I(G)$.*

For arbitrary distributions, it might not be possible to find a perfect UG.

3 Algebraic Extensions of UGMs

3.1 Factor Graphs

Consider a fully connected UG G. Then, its maximal clique is simply all of $\mathcal{X}$, so that any distribution P that factors wrt G just says that P is a multivariate function over all of $\mathcal{X}$, i.e. does not impose any restriction at all. Consider on the other hand, a "pairwise" factorization:

$$P(X) = \frac{1}{Z} \prod_{s \neq t} \Phi(X_s, X_t),$$

just over cliques of size two (i.e. edges). This clearly has a compact representation, but its minimal UG can be seen to be the fully connected graph G. Thus, P is a UGM wrt the fully connected graph G, which moreover is the minimal UG. So P has an even more compact representation than the UGM based representation!

This is where we part ways with requiring an equivalence between factorization and Markov properties, as with UGMs, and only focus on a compact factorization with respect to the graph G, such as the pairwise model above. Interestingly, this can in turn be modeled as a UGM, but over an extended graph called a factor graph.

Definition 12 *A factor graph $\mathcal{F}$ consists of two sets of nodes: variable nodes V, and factor nodes F, where each factor node $f \in F$ is connected to some subset of variable nodes, and is associated with a factor function Φ_f that only depends on the corresponding neighboring variable nodes. A distribution P that factors wrt a factor graph $\mathcal{F}$ is simply a normalization of the product of factor functions associated with its factor nodes:*

$$P(X) = \frac{1}{Z} \prod_{f \in F} \Phi_f(x_f).$$

Thus, the factor graph makes the factorized form of the distribution clear. For instance, with the pairwise factorized distribution example earlier, the corresponding factor graph has $\binom{p}{2}$ factors, one for each node pair.

3.2 Log-Linear Models

Recall the Gibbs distribution form of a UGM:

$$P(X) = \frac{1}{Z} \exp(-\sum_{C \in \mathcal{C}} \phi_C(X_C)).$$

More generally, suppose we have access to some "feature functions" $\{f_i : \mathcal{X} \mapsto \mathbb{R}\}_{i=1}^k$. We can then consider the parametric family of so-called log-linear models:

$$P_\theta(X) = \frac{1}{Z} \exp(-\sum_{i=1}^{k} \theta_i f_i(X)).$$

We say that P is a log-linear UGM wrt UG graph G if it belongs to a log-linear model family where the feature functions $\{f_i\}_{i=1}^k$ each only depend on variables within a clique of the graph G. We will discuss these at greater length when we discuss inference procedures for PGMs.

Ising Models. A popular class of such models are Ising models, where each variable $X_s \in \{-1, 1\}$ corresponds to the direction of an atom's spin. The joint distribution is then specified via the log-linear model:

$$P(X) = \frac{1}{Z} \exp(\sum_{s \in V} \theta_s X_s + \sum_{(s,t) \in E} \theta_{st} X_s X_t)$$

When all $\theta_{st} > 0$, the model prefers neighboring atoms to have the same spin, and is called a Ferromagnetic Ising model. Learning and inference with such models is typically easier, as we will see in the sequel. When all $\theta_{st} < 0$, it is then called an Anti-Ferromagnetic Ising model.

Boltzmann Model. An equivalent reparameterization of the Ising model is when $X_s \in \{0, 1\}$, but with the same parameterization as the Ising model:

$$P(X) = \frac{1}{Z} \exp(\sum_{s \in V} \theta_s X_s + \sum_{(s,t) \in E} \theta_{st} X_s X_t)$$

For this model, the conditional distribution of a node X_s given the rest of the nodes is given by the logistic regression model:

$$P(X_s|X_{-s}) = \text{sigmoid}(\theta_s + \sum_t \theta_{st} X_t),$$

which mimics the behavior of a neuron. The Boltzmann distribution could thus be viewed as a simplistic characterization of a neuron network.

Potts Model. The Boltzmann model interactions could equivalently be written as $\theta_{st} I[x_s == x_t]$. One could also use such interactions when each variable $X_s \in [K]$ takes values in some discrete set, where it is called a Potts model.

Metric MRFs Another important instance of UGMs are where:

$$P(X) = \frac{1}{Z} \exp\left(\sum_{s \in V} f_s(X_s) + \sum_{(s,t) \in E} \theta_{st} f(X_s, X_t) \right),$$

where $f_s : \mathcal{X}_s \mapsto \mathbb{R}$ are arbitrary functions, and $f : \mathcal{X}_s \times \mathcal{X}_t \mapsto \mathbb{R}$ is a metric function that satisfies:

- **Reflexivity:** $f(x_s, x_t) = 0$ iff $x_s == x_t$.
- **Symmetry:** $f(x_s, x_t) = f(x_t, x_s)$.
- **Triangle Inequality:** $f(x_s, x_u) \leq f(x_s, x_t) + f(x_t, x_u)$.

Such metric interactions occur frequently in UGM applications to computer vision, where they encourage "smoothness". Specifically, each $f_s(\cdot)$ specifies a node-wise preference, while $f(x_s, x_t)$ encourages nearby nodes to have similar values when $\theta_{st} > 0$.