Graphical Models - Representing Multivariate Distributions

Amir Globerson

Many problems in machine learning involve the prediction of a large number of variables. Some examples are:

- Semantic Segmentation - Given an image x, assign a label $y_i \in \{1, \ldots, C\}$ to the i^{th} pixel in the image. This label corresponds to the identity of the object in this location in the image (e.g., if $y_i = 2$ then the i^{th} pixel is part of a car). Thus, if there are n pixels there are n^C possible different segmentation.
- Syntactic Parsing - Given a sentence x, provide a syntactic analysis y. There are various formalisms for syntax. For example in dependency parsing y is a spanning tree whose nodes are the words in x, and edges describe word dependencies. In constituent parsing, y is a derivation tree under some CFG grammar. Here again the number of labels is very large, for example the number of spanning trees in the dependency parsing case, which is exponential in n.
- Multilabel prediction - Given an object x (e.g., a document) predict n of its properties $y_1, \ldots, y_n$. For example x can be a document and $y_i = 1$ indicates that the document covers topic i. Or x is a patient, and $y_i = 1$ indicates the patient has the disease i. Here again the number of labels is exponential in n.
- Medical Diagnostics - Given symptoms of a patient what are the set of possible causes.

In all the above, we had set of variables $x_1, \ldots, x_m, y_1, \ldots, y_n$ and we want to make predictions. An elegant approach to do this is using a probabilistic model. If we had a distribution $p(x_1, \ldots, x_m, y_1, \ldots, y_n)$ we could use it to predict $y_1, \ldots, y_n$ by considering $p(y_1, \ldots, y_n | x_1, \ldots, x_m)$ and maximizing it over $y_1, \ldots, y_n$.

In other cases, we have a set of variables $x_1, \ldots, x_n$ (say intensities of pixels in an image) and we want to understand how these are distributed. This can be used for example for generating images, or for completing parts of an image. Similarly if $x_1, \ldots, x_n$ are words in a sentence, there are many advantages to modeling such distributions.

In the next few classes we will consider different approaches to modeling distributions over variables $X_1, \ldots, X_n$. We shall be particularly interested in models where certain independence or conditional independence properties hold. It turns out that such models can be specified via graphs, which is why they are called graphical models.

1 Definitions and Notations

We begin with a few definitions and results that will be needed later.

1.1 Directed Graphs

Definition 1. *A* **directed graph** *is a pair (V, E) where V is a set of nodes (which we will typically denote by $\{1, \ldots, n\}$ and ordered pairs of vertices $(i, j) \in E$ (where (i, j) corresponds graphically to an edge $i \to j$). A Directed Acyclic Graph (DAG) is a directed graph that has no directed cycles.*

Definition 2. *A* **topological ordering** *of a DAG is an ordering of its nodes $i_1, \ldots, i_n$ where i_k appears before all the nodes j such that $(i_k, j) \in E$. In other words all the children of i_k appear after it in the ordering. It's easy to see that each DAG has such an ordering and it can be found efficiently. Note that it is not necessarily unique.*

Definition 3. *Given a node i we shall be specifically interested in several sets of nodes. The parents of i are all the nodes with edges going into i. We denote it by $Pa(i)$. The descendants of i are all the nodes with a directed path* from *i. The ancestors of i are all the nodes with a directed path* to *i. The non-descendants of i are all the nodes that are not descendants of i. We denote those by $ND(i)$.*

In what follows we will use graphs to define distributions on variables $x_1, \ldots, x_n$. We will identify variables with nodes in the graph. Thus, the set of variables that corresponds to the parents of i will be denoted by $X_{Pa(i)}$.

1.2 Conditional Independence

Independence is one of the most important properties of a distribution. It tells us something very important about the relations between variables. For example if we know that X_i is independent of X_j it means that setting the value of X_i will not change the distribution of X_j. However, in many cases variables are not independent but rather independent given other variables.

Definition 4. *Given three random variables X, Y, Z we say that X is conditionally independent of Y given Z if for all x, y, z values it holds that:*

$$p(x, y|z) = p(x|z)p(y|z) \tag{1}$$

The above can be seen to be equivalent to the condition $p(x|y, z) = p(x|z)$.[1] *We denote this relation by $X \perp Y | Z$.*

Conditional independence has several useful properties. We mention two below and will use others during the course.

Lemma 1. *The Decomposition lemma for CI: Let W, X, Y, Z be random variables. Then $X \perp Z, W|Y$ implies $X \perp Z|Y$. The proof is straightforward.*

Another almost trivial result is:

Lemma 2. *If $X \perp Y|Z$ then $X \perp Y, Z|Z$. This follows since:*

$$p(x|y, z, z) = p(x|y, z) = p(x|z) \tag{2}$$

Where the last equality follows from $X \perp Y|Z$.

2 Bayesian Networks

We begin by describing a class of models called Bayesian Networks (BN).[2] A BN is defined using a directed graph and the properties of the directed graph imply the conditional independence properties of the BN.

A Bayesian Network is a model of a distribution $p(x_1, \ldots, x_n)$ that is constructed via a DAG. It has some very nice provable conditional independence properties as we show later.

[1] Some care should be taken for the case where $p(y|z) = 0$ to show the equivalence.

[2] Introduced by Judea Pearl in his 1988 book: "Probabilistic Reasoning in Intelligent Systems"

Assume we are given a DAG G and a set of distributions $p(x_i|x_{Pa(i)})$. The Bayesian network $B = (G, p)$ is a distribution over $x_1, \ldots, x_n$ defined as:

$$p_B(x_1, \ldots, x_n) = \prod_i p(x_i|x_{Pa(i)}) \tag{3}$$

First, we need to prove it is indeed a distribution. Clearly it is non-negative. We also want to show that

$$\sum_{x_1,\ldots,x_n} p_B(x_1, \ldots, x_n) = 1 \tag{4}$$

To show this assume wlog that $1, \ldots, n$ is a topological ordering. Then:

$$p_B(x_1, \ldots, x_n) = p(x_1|x_{Pa(1)})p(x_2|x_{Pa(2)}) \ldots p(x_n|x_{Pa(n)}) \tag{5}$$

where x_n appears only in the last term. Thus:

$$\begin{aligned}\sum_{x_1,\ldots,x_n} p_B(x_1, \ldots, x_n) &= \sum_{x_1,\ldots,x_{n-1}} p(x_1|x_{Pa(1)})p(x_2|x_{Pa(2)}) \ldots \sum_{x_n} p(x_n|x_{Pa(n)}) \\ &= \sum_{x_1,\ldots,x_{n-1}} p(x_1|x_{Pa(1)})p(x_2|x_{Pa(2)}) \ldots p(x_{n-1}|x_{Pa(n-1)})\end{aligned}$$

Continuing this recursively we have that the overall sum is one.

2.1 Local Independence Properties of BNs

Next, we can ask what CI properties such a distribution may have. In other words if we know that p_B is a Bayesian Network for a graph G, what CI properties must it satisfy (regardless of the choice of p).

We need the following definition: the set of non-descendants of i are all the nodes j for which there is no directed path from i to j. Denote these by $ND(i)$.

Theorem 1. *Given a Bayes net (G, p) the distribution p_B has the following properties:*

- *For all i, the variable X_i is conditionally independent of its non-descendants that are not his parents, given its parents:*

$$X_i \perp X_{ND(i) \setminus Pa(i)} | X_{Pa(i)} \tag{6}$$

- *It holds that:*

$$p_B(x_i|x_{Pa(i)}) = p(x_i|x_{Pa(i)}) \tag{7}$$

We note that from Lemma 2 it also follows that: $X_i \perp X_{ND(i)} | X_{Pa(i)}$. This is the common form in the literature but can be confusing.

Proof. First, consider the distribution $p_B(x_i, x_{ND(i)})$ which can easily be shown to have this form (after summing over all descendants of x_i)

$$p_B(x_i, x_{ND(i)}) = p(x_i|x_{Pa(i)}) \prod_{k \in ND(i)} p(x_k|x_{Pa(k)}) \tag{8}$$

So that:

$$p_B(x_i|x_{ND(i)}) = \frac{p_B(x_i, x_{ND(i)})}{\sum_{x_i} p_B(x_i, x_{ND(i)})} = p(x_i|x_{Pa(i)}) \tag{9}$$

Or in other words:

$$p_B(x_i|x_{ND(i) \setminus Pa(i)}, x_{Pa(i)}) = p(x_i|x_{Pa(i)}) \tag{10}$$

Multiply by $p_B(x_{ND(i)\setminus Pa(i)}|x_{Pa(i)})$ to get:

$$p_B(x_i|x_{ND(i)\setminus Pa(i)}, x_{Pa(i)})p_B(x_{ND(i)\setminus Pa(i)}|x_{Pa(i)}) = p(x_i|x_{Pa(i)})p_B(x_{ND(i)\setminus Pa(i)}|x_{Pa(i)}) \quad (11)$$

Equivalent to:

$$p_B(x_i, x_{ND(i)\setminus Pa(i)}|x_{Pa(i)}) = p(x_i|x_{Pa(i)})p_B(x_{ND(i)\setminus Pa(i)}|x_{Pa(i)}) \quad (12)$$

Sum over all value of $x_{ND(i)\setminus Pa(i)}$ to get:

$$p_B(x_i|x_{Pa(i)}) = p(x_i|x_{Pa(i)}) \quad (13)$$

Substitute this in Eq. 10 to get:

$$p_B(x_i|x_{ND(i)\setminus Pa(i)}, x_{Pa(i)}) = p_B(x_i|x_{Pa(i)}) \quad (14)$$

From which it follows that $X_i \perp X_{ND(i)\setminus Pa(i)}|X_{Pa(i)}$.
To obtain $X_i \perp X_{ND(i)}|X_{Pa(i)}$ we just need to note that $X \perp Y|Z$ implies $X \perp Y, Z|Z$. □

The above theorem implies several things. First it means that p_B satisfies the following:

$$p_B(x_1, \dots, x_n) = \prod_i p_B(x_i|x_{Pa(i)}) \quad (15)$$

We say that a distribution satisfying the above **factors according to** G.

Definition 1. *The Local Markov CI properties of DAG G are a set of n conditional independence properties given by:*

$$I_{LM}(G) = \{(X_i \perp X_{ND(i)}|X_{Pa(i)})\}_{i=1}^n \quad (16)$$

Definition 2. *Let p be a distribution over $X_1, \dots, X_n$. We denote by $I(p)$ the set of* all *conditional independence properties that are true for p*

For example $I(p)$ can look something like:

$$I(p) = \{X_1, X_2 \perp X_3, X_4|X_5, X_7, X_2, X_4 \perp X_5|X_6\} \quad (17)$$

Using the above definitions Theorem 1 can be expressed as the following assertion: If p factorizes according to G then $I_{LM}(G) \subseteq I(p)$.

2.2 Local Markov leads to factorization

The previous section showed that ***any*** distribution that factorizes has the $I_{LM}(G)$ properties. Now, suppose that we would like to construct a distribution that has $I_{LM}(G)$ properties. We know we can use a Bayesian Network to obtain a distribution with these properties. But maybe there are other, more complex distributions that do not factor and have this property. The following theorem states that there are not (it is essentially the converse of Theorem 1).

Theorem 2. *If p is a distribution such that $I_{LM}(G) \subseteq I(p)$, then p factorizes according to G.*

Proof. Assume wlog that $1, \dots, n$ is a topological ordering. Start by writing the chain rule:

$$p(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i|x_1, \dots, x_{i-1}) \quad (18)$$

We will be able to show the result if we show that $p(x_i|x_1,\ldots,x_{i-1}) = p(x_i|x_{Pa(i)})$.

Because of the topological ordering, all of the parents of x_i are in $x_1,\ldots,x_{i-1}$. Also, none of its descendants can be in this set. So we can write:

$$1,\ldots,i-1 = Pa(i) \cup S \tag{19}$$

where $S \subseteq ND(i) \setminus Pa(i)$. Thus what we want to show can be rephrased as $p(x_i|x_{Pa(i)}, x_S) = p(x_i|x_{Pa(i)})$. In other words, we want to show that $X_i \bot X_S | X_{Pa(i)}$.

We know from the LM property that:

$$X_i \bot X_{ND(i)\setminus Pa(i)} | X_{Pa(i)} \tag{20}$$

This is almost what we need, but remember that $S \subseteq ND(i)$. Fortunately, we have the Decomposition Lemma for CI (see above). Thus we can rewrite Eq. 20 as:

$$X_i \bot X_S, X_{ND(i)\setminus\{Pa(i),S\}} | X_{Pa(i)} \tag{21}$$

And from the lemma we conclude:

$$X_i \bot X_S | X_{Pa(i)} \tag{22}$$

which is what we needed. □

2.3 What other conditional independencies hold for Bayesian Networks?

The LM properties only tell us about n specific CIs. What about any other property $X \bot Y | Z$. How can we tell if it holds in a given BN or not? We may suspect that $X \bot Y | Z$ in a BN based on a graph if Z somehow blocks the influence of X on Y. The LM is an instance of this. We may thus first hypothesize that if there is no path between X and Y in the graph that does not go through Z the property will hold. This in fact is not true since the explaining away phenomenon is a counter example.

We need the following definition:

Definition 5. *An undirected path in a DAG G is active given a node set E if:*

- *For every V structure $i \to j \leftarrow k$ in the path either j or one of its descendants is in E.*
- *Every other node in the path is not in E*

The idea behind an active path is that if there is an active path between X_i and X_j given E then X_i is potentially dependent on X_j given E. This is extended to larger variable sets via the following definition.

Definition 6. *Given variables sets X, Y, Z in the graph G we say that X is d-separated from Y given Z if there is no active trail between a node in X and a node in Y given Z.*

The set of CI properties $I_{d-sep}(G)$ corresponds to all triplets $X \bot Y | Z$ where X is d-separated from Y given Z. In other words $X \bot Y | Z$ is in $d - sep(G)$ if every path between X and Y either has a node in Z that is not on a v-structure or it has a node outside Z that is in a v-structure or a descendant of a node in a v structure.

What can we say about d-separation and Bayesian networks?

Proposition 1. *For any BN on graph G it holds that $I(p) \supseteq I_{d-sep}(G)$.*

We will not show this here. It will wait until we talk about undirected models.

Since $I_{d-sep}(G)$ is quite a large set, we might expect that it captures all CI properties in a BN on G. This is clearly wrong however since for every graph G we can define a BN such that all variables are independent (i.e., $p_B = \prod_i p(x_i)$). Thus $I(p)$ will clearly contain more properties than those specified in $I_{d-sep}(G)$. So clearly $I_{d-sep}(G) \not\supseteq I(p)$.

However, you may suspect that this example is very specific. In other words we had to choose very specific p distributions to construct this p_B such that it will have properties not in $I_{d-sep}(G)$. In fact, it can be shown that except for a set of measure zero all BN on G will have satisfy $I(p) = I_{d-sep}(G)$.

In summary the following properties are equivalent:

- The distribution p factorizes according to G (i.e., it is a BN for G).
- $I(p) \supseteq I_{LM}(G)$
- $I(p) \supseteq I_{d-sep}(G)$

3 Markov Networks

Bayesian networks can describe a wide array of distributions. As we have seen these are almost precisely the distributions that satisfy the $I_{d-sep}(G)$ properties (*almost* because they may satisfy more CI properties, but this requires a very particular choice of parameters). It is clear that there are other set of CI properties that cannot be captured by Bayes nets.

We now present a different type of model family for distributions $p(x_1, \ldots, x_n)$ that satisfy CI properties that cannot be captured by BNs. Assume we are given an undirected graph with nodes $V = 1, \ldots, n$ and edges E where $ij \in E$ is an unordered pair. Denote the set of cliques in the graph G by $\mathcal{C}$. Give a set of non-negative functions $\phi_c(\boldsymbol{x}_c)$ where $c \in \mathcal{C}$ we have the following definition of a model distribution:

Definition 1. *The Markov Network (G, ϕ) is the distribution given by:*

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(\boldsymbol{x}_c) \tag{23}$$

Here Z is a normalization constant given by $Z = \sum_{\boldsymbol{x}} \prod_{c \in \mathcal{C}} \phi_c(\boldsymbol{x}_c)$. It is also called the partition function, a term originating in statistical physics. We shall say that if $p(\boldsymbol{x})$ can be written in such a form then it factors according to the undirected graph G.

Note that there may be some cliques for which $\phi_c(\boldsymbol{x}_c) = 1$ (in other words this clique is not *used*). If the cliques used correspond to a smaller graph G' then we can also say that the model factorizes according to G' and may obtain stronger results.

4 Conditional Independence Properties in Markov Networks

In BNs we showed that the fact that a distribution is a BN implies it has a set of conditional independence properties (namely $I_{d-sep}(G)$). Perhaps more surprisingly we also showed that the converse holds: namely, that if it has the properties $I_{d-sep}(G)$ it must be a Bayesian network for the graph G. Here we will follow a similar path for Markov Networks (MNs in what follows). We shall see that not all the results carry over simply.

We begin with a set of CI properties that is the equivalent of $I_{d-sep}(G)$. First, a lemma:

Proposition 1. *Let X be a set of random variables and W, Y, Z be sets of variables such that $X = (W, Y, Z)$. Then given a distribution $p(x)$ we have that $W \perp Y | Z$ if and only if $p(x)$ can be written as $p(x) = \phi_1(w, z)\phi_2(y, z)$.*

Can be proved directly using our results for Bayesian networks.

Definition 2. *Given an undirected graph G and three sets of variables W, Y, Z corresponding to nodes in the graph, we say that Z separates W and Y in G if every path between W and Y has a node in Z. We will denote this by $sep_G(W;Y|Z)$. We now define the following conditional independence properties based on G:*

$$\begin{aligned} I_{sep}(G) &= \{W \perp Y|Z : Z \text{ separates W and Y in } G\} \\ &= \{W \perp Y|Z : \quad sep_G(W;Y|Z)\} \end{aligned}$$

Some useful properties of separation:

- If Z separates Y and W, then for any $Z' \supset Z$ it holds that Z' separates Y and W. In other words $\text{sep}_G(W;Y|Z)$ implies $\text{sep}_G(W;Y|Z')$.
- If Z separates Y and W, then for any $W' \subset W$ it holds that Z separates Y and W'.

We then have the following result:

Theorem 1. *If p factorizes according to G then $I(p) \supseteq I_{sep}(G)$.*

The proof relies on the factorization lemma we showed in class and the decomposition lemma for CI (see book). We want to show that for every property $W \perp Y|Z \in I_{sep}(G)$, the distribution $p(x)$ satisfies it. Begin with the case that $X = W, Y, Z$. Now, we know that Z separates W and Y. This means there is no direct edge from W to Y. Thus there is no clique which includes nodes from W and Y. By the fact that p factorizes we have it must be writen as $p(x) = \phi_1(W, Z)\phi_2(Y, Z)$ and by Prop. 1 the CI follows.

Now, assume that $X = V, W, Y, Z$. The nodes in V can be partitioned into two sets. Denote by V_1, V_2 the maximum set of nodes such that Y, V_1 is separated from W, V_2 given Z. Clearly $V_1, V_2 = V$.[3] Now we are the previous case and we can deduce that $V_1, Y \perp V_2, W|Z$. Now by the decomposition lemma we have $Y \perp W|Z$. □

You may wonder if the converse holds. The short answer is no (the long answer is yes, if $p(\boldsymbol{x}) > 0$ for all $\boldsymbol{x}$. See Sec. 4.2 below).

4.1 Other (smaller) CI sets and relation to $I_{sep}(G)$

Recall that in BNs we discussed the Local Markov (LM) property which was a subset of $I_{d-sep}(G)$. But, in fact we showed that LM was not a weaker property than $I_{d-sep}(G)$ in the sense that any distribution that satisfies the former satisfies the latter. Below we show a related result for MNs.

Define the two following sets of CI properties:

Definition 3. *Define $I_{pair}(G)$ as the following sets of CI properties*

$$I_{pair}(G) = \{X_i \perp X_j | X_{V \setminus \{i,j\}} : ij \notin E\} \tag{24}$$

Note that there are $O(n^2)$ such properties and that $I_{pair}(G) \subseteq I_{sep}(G)$.

In words, $I_{pair}(G)$ states that two non-adjacent variables are conditionally independent given the rest of the graph.

The next property says that a variable X_i is independent of the rest of the graph given its neighbors. By neighbors of i we mean the nodes j such that $ij \in E$. We denote this set by $Nbr(i)$.

Definition 4. *Define $I_{LM}(G)$ as the following sets of CI properties*

$$I_{LM}(G) = \{X_i \perp X_{V \setminus \{Nbr(i), i\}} | X_{Nbr(i)} : i = 1, \ldots, n\} \tag{25}$$

Note that there are $O(n)$ such properties and that $I_{LM}(G) \subseteq I_{sep}(G)$. However, there is no strict relation between $I_{pair}(G)$ and $I_{LM}(G)$.

[3] To see this, note that for each variable A (not in W, Y, Z) it hold that either $sep_G(W, A; Y|Z)$ or $sep_G(W, A; Y|Z)$. Otherwise there are paths from A to Y and W that do not go through Z, which is impossible by the assumption that $sep_G(W; Y|Z)$.

What is the relation between these three properties? Clearly $I_{sep}(G)$ is stronger than the other two (since it contains them).

The next theorem shows that $I_{LM}(G)$ is stronger than $I_{pair}(G)$:

Theorem 3. *If $I(p) \supseteq I_{LM}(G)$ then $I(p) \supseteq I_{pair}(G)$ (this is sometimes denoted by $I_{LM}(G) \Rightarrow I_{pair}(G)$)*

The proof is in the book and involves use of the weak union property:

$$X \perp Y, W | Z \Rightarrow X \perp Y | Z, W \tag{26}$$

Proof. We know that the LM properties hold. Now we want want to show that all pairwise properties must hold. For a given $ij \notin E$, start with the LM property for i:

$$\begin{aligned} X_i &\perp X_{V \setminus i, N(i)} | X_{N(i)} \\ X_i &\perp X_j, X_{V \setminus i,j,N(i)} | X_{N(i)} \\ X_i &\perp X_j | X_{N(i)}, X_{V \setminus i,j,N(i)} \\ X_i &\perp X_j | X_{V \setminus i,j} \end{aligned}$$

□

To understand the converse, we shall use a stronger result which says that $I_{pair}(G) \Rightarrow I_{sep}(G)$ when $p(\boldsymbol{x})$ is strictly positive.

Theorem 4. *If $p(\boldsymbol{x})$ is a strictly positive distribution and $I(p) \supseteq I_{pair}(G)$ then $I(p) \supseteq I_{sep}(G)$.*

The proof is in the book and involves use of the following *intersection* property which holds for strictly positive distributions:

$$U \perp Y | Z, W \quad \& \quad U \perp W | Z, Y \Rightarrow U \perp Y, W | Z \tag{27}$$

Proof is by reverse induction. The base ($|Z| = n - 2$) is clear from the pairwise property assumption. Now assume for all Z' such that $|Z'| = k$ and prove for $k - 1$. First assume that $X = W \cup Z \cup Y$ and that $sep_G(W; Y|Z)$. We would like to show that the property $W \perp Y | Z$ is in $I(p)$. Take some $A \in Y$ and denote $Y' = Y \setminus A$.

Then:

$$\begin{aligned} sep_G(W; Y|Z) &\implies sep_G(W; Y'|Z) \\ sep_G(W; Y|Z) &\implies sep_G(W; A|Z) \end{aligned}$$

(since the sets on the LHS of the separation can always be shrunk). And the following can also be deduced (since the RHS can always be expanded):

$$\begin{aligned} sep_G(W; A|Z) &\implies sep_G(W; A|Z, Y') \\ sep_G(W; Y'|Z) &\implies sep_G(W; Y'|Z, A) \end{aligned}$$

Because of the induction assumption we have that:

$$\begin{aligned} &W \perp A | Z, Y' \\ &W \perp Y' | Z, A \end{aligned}$$

And from the intersection property Eq. 27 we have:

$$W \perp A, Y' | Z \implies W \perp Y | Z \tag{28}$$

as required.

To complete the proof we need the case that $X \supset W \cup Z \cup Y$. This is left as an exercise.

4.2 Does $I_{sep}(G)$ imply factorization?

When discussing Bayesian networks we saw that if a distribution $p(x)$ satisfies $I_{LM}(G)$ for a DAG G, then $p(x)$ has to factorize according to G. We can now ask a similar question for Markov networks. Say we know a distribution $p(x)$ satisfies an independence property like $I_{sep}(G)$, does it factorize according to G?

The general answer is no. i.e., there are distributions $p(x)$ that satisfy $I_{sep}(G)$, and which cannot be written as a product of function over the cliques of G. However, it becomes true with one extra condition on $p(x)$. We say that $p(x)$ is strictly positive if $p(x) > 0$ for all x. In 1971 Hammersley and Clifford showed that if $p(x)$ is strictly positive and satisfies $I_{pair}(G)$ then it factors according to G. Clearly this implies that the same holds for $I_{sep}(G)$ and $I_{LM}(G)$ since these are stronger properties than $I_{pair}(G)$.