Structure of the High-Order Boltzmann Machine from Independence Maps

F. Xabier Albizuri, Alicia d’Anjou, Manuel Graña, Member, IEEE, and Pedro Larrañaga, Member, IEEE

Abstract—In this paper we consider the determination of the structure of the high-order Boltzmann machine (HOBM), a stochastic recurrent network for approximating probability distributions. We obtain the structure of the HOBM, the hypergraph of connections, from conditional independences of the probability distribution to model. We assume that an expert provides these conditional independences and from them we build independence maps, Markov and Bayesian networks, which represent conditional independences through undirected graphs and directed acyclic graphs respectively. From these independence maps we construct the HOBM hypergraph. The central aim of this paper is to obtain a minimal hypergraph. Given that different orderings of the variables provide in general different Bayesian networks, we define their intersection hypergraph. We prove that the intersection hypergraph of all the Bayesian networks \((\mathcal{N}')\) of the distribution is contained by the hypergraph of the Markov network, it is more simple, and we give a procedure to determine a subset of the Bayesian networks that verifies this property. We also prove that the Markov network graph establishes a minimum connectivity for the hypergraphs from Bayesian networks.

Index Terms—Bayesian networks, Boltzmann machines, independence maps, graphical models, log-linear models, neural networks.

I. INTRODUCTION

T

HE conventional Boltzmann machine (BM) [1], [9], as well as the high-order Boltzmann machine (HOBM) [15], [3], is a technique whose purpose is, in its fundamental formulation, to describe and model probability distributions defined on a set of binary random variables.

The BM approximates a distribution with a model where the probability function is defined as the normalized exponential of a consensus function. The learning algorithm is a steepest descent of the Kullback–Leibler divergence between the distribution to learn and the approximation distribution. In the conventional BM there are hidden units, the consensus function is formed by first and second-degree terms on the variables, i.e., connections up to order two between units, and the approximation distribution is the marginal distribution on the visible units. The HOBM is a variation of the conventional BM where we consider higher order connections and do not use hidden units. The HOBM allows us to undertake the problem considered in this paper, that is the determination of the structure of the BM.

The structure of the HOBM is given by the set of connections. We will call this connection set the hypergraph of the HOBM. We study the determination of the hypergraph from independence maps [5], [18]. The books by Pearl [13], Whittaker [19], and Lauritzen [11] are good comprehensive works on the subject of independence maps. An independence map is a graph with a vertex separation criterion for inferring conditional independences between random variables, the vertices of the graph, and it represents part of the independences of a probability distribution. There are basically two kinds of independence maps: Markov networks, which are undirected graphs, and Bayesian networks, which are directed acyclic graphs. Some attempts to link independence maps and BM’s have been carried out [6], [10], [12]. In this paper we give a systematic solution to the problem of determining the structure of the HOBM from conditional independences.

Our approach assumes that an expert provides us conditional independences of the probability distribution to learn. Markov and Bayesian networks are built from these conditional independences. The structure of the HOBM will be fixed according to the probability function factorizations associated with these independence maps. In this paper we will show how to construct the hypergraph of connections of the HOBM from Markov and Bayesian networks. Given a probability distribution, i.e., its conditional independences, the Markov network is unique, but we have in general different Bayesian networks for different orderings of the variables. We will introduce the notion of intersection hypergraph of the Bayesian networks corresponding to different orderings of the variables. The central aim of our paper is to get a minimal hypergraph, and we have investigated whether the Markov network provides this minimal hypergraph, or we can optimize it by means of Bayesian networks, searching for suitable variable orderings. In this study we will prove that we can determine a set of Bayesian networks such that the intersection hypergraph is contained by the hypergraph of the Markov network, it is more simple. We will establish this result through chordal independence maps, which will be an intermediate step between the Markov network and the Bayesian networks that we search for. We will also prove that the Markov network determines a minimum of connectivity for the hypergraphs constructed from Bayesian networks.

In this paper we will show how we can define the structure of the HOBM, i.e., the set of weighted connections, from the qualitative information provided by conditional indepen-
ences. Once the structure is defined, the learning algorithm of the HOBM can be applied to a set of samples from the distribution to learn, obtaining the connection weights that minimize the divergence between the approximation distribution and the distribution of the sample set [2].

This paper is organized as follows. In Section II we introduce the HOBM. In Section III we study the determination of the structure of the HOBM, the hypergraph of connections, from factorizations of the probability function to learn. In Sections IV and V we get hypergraphs from Markov and Bayesian networks respectively. In Section VI we present the results that relate the hypergraphs obtained from these independence maps. In Section VII we end with the conclusions.

II. THE HIGH-ORDER BOLTZMANN MACHINE

A BM is a stochastic recurrent network with a local learning algorithm. The configuration of the HOBM with $N$ units is defined by $u \in \{0,1\}^N$ and the state of the unit $i \in \{1, \ldots, N\}$ by $x_i$. A connection $\lambda = \{i_1, \ldots, i_m\}$ is a nonempty subset of $\{1, N\} = \{1, \ldots, N\}$, that is $\lambda \in P([1, N])$, and $[\lambda]$ denotes the order $m$ of the connection $\lambda$, the number of its ends. Some connections have an associated weight, a real number $w_{\lambda}$, modified by the learning algorithm. We define the consensus function

$$ C^*(u) = \sum_{\lambda \in L^*} w_{\lambda} a(\lambda | u) $$

where $L^* \subseteq P([1, N])$ is the set of weighted connections. The function $a(\lambda | u)$ is defined as

$$ a(\lambda | u) = \prod_{i \in \lambda} x_i $$

so $a(\lambda | u) = 1$ if every end of $\lambda$ takes value one (the connection is activated) and $a(\lambda | u) = 0$ otherwise.

The stochastic transition law is the following: given a configuration $u$ of the HOBM, we choose at random an unit $j$ and change its state $x_j$ to $x'_j = 1 - x_j$ with probability

$$ \frac{1}{1 + \exp(-\Delta C^*(u))} $$

where

$$ \Delta C^*(u) = (1 - 2x_j) \sum_{\lambda \in L^*/j \not\in \lambda} w_{\lambda} a(\lambda \setminus \{j\} | u). $$

The dynamics defined corresponds to a Markov chain where the stationary probability distribution is the Boltzmann–Gibbs distribution

$$ P^*(u) = \frac{1}{Z} \exp C^*(u) $$

where $Z = \sum_u \exp C^*(u)$.

The purpose of the HOBM is to approximate a positive probability distribution $P(u)$ on $\{0,1\}^N$, usually given by the frequency distribution of a set of samples, with the distribution $P^*(u)$. The learning algorithm is a steepest descent of the Kullback–Leibler divergence

$$ D = \sum_u P(u) \ln \frac{P(u)}{P^*(u)}. $$

The weights $\{w_{\lambda}/\lambda \in L^*\}$ are modified by the iterative rule

$$ w_{\lambda}^{t+1} = w_{\lambda}^t - \alpha \cdot (p_{\lambda}^t - p_{\lambda}) $$

where

$$ p_{\lambda}^t = \sum_u P^*(u) a(\lambda | u) \quad p_{\lambda} = \sum_u P(u) a(\lambda | u) $$

are the activation probabilities of the connection $\lambda$ under the approximation distribution and under the distribution to learn.

The learning algorithm of the HOBM converges to the strict global minimum of the divergence (1), which corresponds to the maximum likelihood estimate of the connection weights [2].

III. STRUCTURE OF THE HOBM FROM CONDITIONAL INDEPENDENCES

We will determine the structure of the HOBM, the hypergraph of weighted connections $L^*$, from conditional independences. We assume that an expert provides the conditional independences of the probability distribution to learn $P(u)$ and from these independences we will establish the structure of the HOBM. Once the structure is determined the learning algorithm of the HOBM is applied to a set of samples from $P(u)$ obtaining the distribution $P^*(u)$, the estimation of the distribution to learn.

Given the conditional independences, the structure of the HOBM is obtained through the factorizations of the probability function $P(u)$ provided by these independences. In order to determine the structure from factorizations of the distribution to learn we begin defining the factorization hypergraphs.

We start considering the factorization of a probability function. Let $P(u)$ be a positive probability function on $\{0,1\}^N$ that admits a factorization

$$ P(u) = \prod_{i=1}^m Q_i(U_i) $$

where $U_i \subseteq U = \{X_1, \ldots, X_N\}$. It can be shown that the probability function $P(u)$ can be written as

$$ P(u) = \frac{1}{Z} \exp C(u) $$

through a consensus function

$$ C(u) = \sum_{\lambda \in L} w_{\lambda} a(\lambda | u) $$

where the weights $w_{\lambda}$ are determined and the set of weighted connections is

$$ L = \bigcup_{i=1}^m \{\lambda \in P([1, N]) / \lambda \subseteq I(U_i)\}. $$

We name $L$ the hypergraph of the factorization (2). In this notation $I(U_i)$ is the set of indexes of the variables in $U_i$. A detailed proof is in Appendix A.

1 We use capital letters for variables, lower-case letters for values taken by variables, boldfaced capital letters for (ordered) sets of variables, and boldfaced lower-case letters for assignments of values to the variables in these sets. So a boldfaced lower-case letter, with or without subscript, is a vector with dimension the number of variables in the set.
So, if a probability function \( P(\mathbf{u}) \) admits a factorization (2) it can be written as the normalized exponential of a consensus function whose terms correspond with the connections in (3), that is the connections in \( L \) correspond to the subsets of the variables in the argument of each function \( Q_j(\mathbf{u}_j) \).

If \( P(\mathbf{u}) \) admits various factorizations, \( L_1, \ldots, L_p \) being the corresponding factorization hypergraphs, then \( P(\mathbf{u}) \) is the normalized exponential of a consensus function whose terms correspond with the connections in the intersection hypergraph

\[
L = \bigcap_{j=1}^{p} L_j.
\]

This is clear since given \( P(\mathbf{u}) \) the \( w_\lambda \) are determined.

Establishing the structure of the HOBM from factorization hypergraphs is justified as follows. Let \( L \) be the intersection hypergraph of some factorizations of the distribution to learn \( P(\mathbf{u}) \) and let \( S \) be a set of samples whose frequency distribution is \( P(\mathbf{u}) \). We define the set of weighted connections \( L^* = L \). The learning algorithm of the HOBM converges to the global minimum of the divergence (1). The divergence \( D \) is nonnegative \[19\], and \( D = 0 \) if and only if \( P(\mathbf{u}) = P^*(\mathbf{u}) \). Since \( L^* = L \) there exist connection weights such that \( P^*(\mathbf{u}) = P(\mathbf{u}) \). Therefore the learning algorithm of the HOBM converges to the connection weights for which \( P^*(\mathbf{u}) = P(\mathbf{u}) \).

In general we have a set of samples from the distribution to learn \( P(\mathbf{u}) \), we define \( L^* = L \) and the learning algorithm converges to the maximum likelihood estimate of the connection weights. So we obtain, among the distributions that can be written through the hypergraph \( L^* = L \) provided by the factorizations, the distribution \( P^*(\mathbf{u}) \) that best fits the sample set.

We consider now the factorization of a probability function \( P(\mathbf{u}) \) from conditional independences. Let \( \mathbf{X}, \mathbf{Y}, \) and \( \mathbf{Z} \) be three disjoint subsets of \( \mathbf{U} \), being \( \mathbf{X}, \mathbf{Y} \neq \emptyset \). The subsets of variables \( \mathbf{X} \) and \( \mathbf{Y} \) are conditionally independent given \( \mathbf{Z} \), we write \( \mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z} \), if

\[
P(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) = P(\mathbf{x} \mid \mathbf{z})P(\mathbf{y} \mid \mathbf{z})
\]

whenever \( P(\mathbf{z}) > 0 \). Equivalently \( \mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z} \) if \( P(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) = P(\mathbf{x} \mid \mathbf{z}) \) whenever \( P(\mathbf{y}, \mathbf{z}) > 0 \). We will study the factorizations provided by independence maps, Markov and Bayesian networks that represent graphically (part of the) conditional independences of a probability distribution. The Markov networks are undirected graphs and the Bayesian networks are directed acyclic graphs (DAG’s). A separation criterion allows us to infer the conditional independences from the graph.

IV. HYPERGRAPH FROM THE MARKOV NETWORK

Let \( P(\mathbf{u}) \) be a probability distribution on a set of variables \( \mathbf{U} \). An undirected graph \( G = (\mathbf{U}, \mathcal{E}) \) can be used to represent (part of) its conditional independences, where the set of vertices is \( \mathbf{U} \) and \( \mathcal{E} \) is the set of edges between vertices. If a subset \( \mathbf{Z} \) of vertices intercepts every path\(^2\) between the vertices of \( \mathbf{X} \) and those of \( \mathbf{Y} \) then we write \( \langle \mathbf{X} \mid \mathbf{Z} \mid \mathbf{Y} \rangle \), that is \( \mathbf{Z} \) separates \( \mathbf{X} \) and \( \mathbf{Y} \). An undirected graph \( G \) is an independence map or I-map of \( P(\mathbf{u}) \) if

\[
\langle \mathbf{X} \mid \mathbf{Z} \mid \mathbf{Y} \rangle \Rightarrow \mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}.
\]

A Markov network of \( P(\mathbf{u}) \) is a minimal I-map of \( P(\mathbf{u}) \). It is constructed according to the following result proved by Pearl and Paz\[14\], \[13\]: every positive distribution \( P(\mathbf{u}) \) has a unique Markov network \( G_0 = (\mathbf{U}, \mathcal{E}_0) \), where

\[
(\alpha, \beta) \notin \mathcal{E}_0 \iff \alpha \perp \beta \mid U - \alpha - \beta.
\]

It follows that the graph \( G = (\mathbf{U}, \mathcal{E}) \) is an I-map of \( P(\mathbf{u}) \) if and only if the Markov network of \( P(\mathbf{u}) \), \( G_0 = (\mathbf{U}, \mathcal{E}_0) \), is a partial graph of \( G \), i.e., \( \mathcal{E}_0 \subseteq \mathcal{E} \).

We are interested in the factorizations obtained from the conditional independences represented in the Markov network of a distribution. Hammersley and Clifford\[7\] showed that a graph \( G = (\mathbf{U}, \mathcal{E}) \) is an I-map of \( P(\mathbf{u}) \) if and only if \( P(\mathbf{u}) \) is a normalized product of nonnegative functions on the cliques\(^3\) of \( G \) (also proved in \[11\]). Therefore given the Markov network \( G_0 = (\mathbf{U}, \mathcal{E}_0) \) of a positive distribution \( P(\mathbf{u}) \), it admits a factorization

\[
P(\mathbf{u}) = \prod_{i=1}^{m} Q_i(\mathbf{c}_i)
\]

where \( \mathbf{c}_1, \ldots, \mathbf{c}_m \) are the cliques of \( G_0 \). The hypergraph of this factorization is

\[
L_{G_0} = \bigcup_{i=1}^{m} \{ \lambda \in P^*([1,M]) / \lambda \subseteq I(\mathbf{c}_i) \}.
\]

The connections in \( L_{G_0} \) correspond to the subsets of the cliques. We will name it the hypergraph of the Markov network of \( P(\mathbf{u}) \).

We present an example where the structure of the HOBM will be determined from the Markov network. Let \( X, Y, Z, U, V, \) and \( W \) be variables corresponding to units that are fixed according to a probability law defined as follows. The values zero or one for \( X, Y, \) and \( Z \) are fixed randomly and independently. When \( X = Y \) we fix \( U = 1 \) with probability 0.9 (\( U = 0 \) with probability 0.1) and when \( X \neq Y \) we fix \( U = 0 \) with probability 0.9. Likewise, if \( Y = Z \) the variable \( V \) tends to one and otherwise to zero, and if \( X = Z \) the variable \( W \) tends to one and otherwise to zero.

In Fig. 1 we show the Markov network \( G_0 \) of this probability distribution. For instance, knowing \( Y, Z, V, \) and \( W \), the variables \( X \) and \( U \) are not independent, then the edge \( (X, U) \) is in \( G_0 \). Knowing \( Z, U, V, \) and \( W \), the variables \( X \) and \( Y \) are not independent (we know \( U \)), then \( (X, Y) \) is in \( G_0 \). Knowing \( Y, Z, U, \) and \( W \), the variables \( X \) and \( V \) are independent, \( (X, V) \) is not in \( G_0 \). The cliques of \( G_0 \) are \( \{X, Y, Z\}, \{X, Y, U\} \), etc. and therefore the hypergraph of the Markov network is \( L_{G_0} \).

\(^2\)We follow generally the terminology of \[8\].

\(^3\)Given an undirected graph, a subset of vertices is complete if all its vertices are adjacent to each other. A clique is a maximal complete subset.
Fig. 1. In our example $G_0$ is the Markov network and $L_{G_0}$ its hypergraph. For the ordering $\{X, Y, Z, U, V, W\}$ we have the Bayesian network $D$ and its hypergraph $L_D$, $D$ being a Bayesian network on $G_0 = G^\circ$. The intersection hypergraph of all the Bayesian networks is $L_{\text{DAG}}$. We note that $L_{\text{DAG}} \subseteq L_{G_0}$ and $G(L_{\text{DAG}}) = G_0$. The architecture of the HOBM would be given by $L_D$.

V. HYPERGRAPHS FROM BAYESIAN NETWORKS

Besides the undirected graphs, with their straightforward separation criterion for inferring conditional independences, DAG’s can be used to represent conditional independences of a probability distribution. The $d$-separation criterion in a DAG is as follows. Given a DAG $D = (V, E)$, where $V$ is the vertex set and $E$ the arrow set, the vertex subset $Z$ is said to activate a path\(^4\) between a vertex of $X$ and a vertex of $Y$ if, for vertices in this path:

1) every vertex with two incident arrows (of the path) is in $Z$ or has some descendant\(^5\) in $Z$.

2) the remaining vertices are not in $Z$.

The subset $Z$ is said to $d$-separate $X$ and $Y$, $(X \perp Y \mid Z)_D$, if there are no paths between vertices in $X$ and in $Y$ activated by $Z$, i.e., every path is blocked by $Z$. A DAG $D = (V, E)$ is an $I$-map of $P(u)$ if $(X \perp Z \mid Y)_D$ implies $X \perp Y \mid Z$.

A Bayesian network of $P(u)$ is a DAG that is a minimal $I$-map. It is constructed as follows. Given $P(u)$, let us consider an ordering of the variables $U = \{X_1, X_2, \ldots, X_N\}$. Let $F_{X_i}$ be a minimal subset of $U_{X_i} = \{X_1, X_2, \ldots, X_i\}$ such that

$$X_i \perp U_{X_i} \perp F_{X_i} \mid F_{X_i},$$

(5)

(true if $U_{X_i} = F_{X_i} = \emptyset$). The boundary DAG of $P(u)$ relative to the ordering of $U$ is the DAG obtained assigning the vertices of $F_{X_i}$ as the parents of $X_i$, for $i = 1, \ldots, N$. It is unique given an ordering if $P(u)$ is positive [13], and\(^6\) we will obtain $F_{X_i}$ starting with $F_{X_i} = U_{X_i}$ and eliminating successively vertices so that the conditional independence (5) holds. Verma [17] proved that if $D$ is a boundary DAG of $P(u)$, then $D$ is a Bayesian network of $P(u)$. Conversely, given a DAG $D$ that is $I$-map of $P(u)$, from $d$-separation, every variable $X$ is independent of all its nondescendants given its parents $F_X$.

So given $P(u)$ we must fix an ordering of the variables in $U$ to obtain the Bayesian network, constructing the boundary DAG. We have in general different Bayesian networks of $P(u)$ for different orderings.

We consider now the factorization provided by a Bayesian network. Let $D$ be a Bayesian network of a positive distribution $P(u)$ on $\{0, 1\}^N$ and $U = \{X_1, X_2, \ldots, X_N\}$ an ordering consistent with the DAG $D$, that is the parents $F_{X_i}$ of a variable $X_i$ come before $X_i$ in the ordering (such ordering always exists). Writing the chain rule

$$P(x_1, x_2, \ldots, x_N) = P(x_N \mid x_{N-1}, \ldots, x_2)P(x_{N-1} \mid x_{N-2}, \ldots, x_1) \cdots P(x_3 \mid x_2, x_1)P(x_2 \mid x_1)$$

we have that each vertex is independent of its predecessors in the ordering given its parents, therefore

$$P(u) = \prod_{i=1}^{N} P(x_i \mid f_{X_i}).$$

The hypergraph of this factorization is

$$L = \bigcup_{i=1}^{N} \{\lambda \in \mathcal{P}([1, N]) / \lambda \subseteq I(\{X_i\} \cup F_{X_i})\}.$$
of Bayesian networks such that the intersection hypergraph is contained by the hypergraph of the Markov network, therefore the intersection hypergraph of all the Bayesian networks is more simple than the Markov network hypergraph. In order to obtain this result we will start from the Markov network and we will construct a determined set of chordal I-maps triangulating it. The intersection hypergraph of Bayesian networks on these chordal I-maps will minimize the Markov network hypergraph.

The structure of our argument is illustrated in Fig. 2. From the Markov network $G_0$ a chordal I-map $G^*$ can be obtained. By means of a join tree $T$ associated with $G^*$ we define a DAG $D^*$ on $G^*$ and only if there exists a tree $T$ (join tree) with the cliques of $G$ as vertices, such that for every vertex $\alpha$ of $G$, any two cliques containing $\alpha$ are either adjacent in $T$ or connected by a path made of cliques that contain $\alpha$.

![Diagram](image)

Fig. 2. From the independences of $P$ we have the Markov network $G_0$ and from a chordal I-map $G^*$. Through a join tree $T$ we define $D^*$. By Proposition 2, $D^*$ is an I-map. From $D^*$ we get a Bayesian network $D$, Theorem 3 stating that $L_{G_0} \subseteq L_{G^*}$. If the Markov network $G_0$ of $P(u)$ is not chordal we will add edges to get a chordal I-map of $P(u)$, which is not unique in general. An algorithm for triangulating a graph and getting a join tree is described in [13] and [16]. We will prove that given a chordal I-map $G^*$ of $P(u)$ there exists a Bayesian network whose hypergraph is contained in the hypergraph of $G^*$ (defined as in (4), by means of its cliques).

If $G^*$ is a chordal graph and $T$ a join tree of $G^*$, we direct acyclicly the edges of $G^*$ as follows. Let us consider a directing of the tree $T$ and an ordering of the vertices $C$ of $T$ consistent with $T$, and by Theorem 3, $L_{G^*} \subseteq L_{G_0}$, so $L_{G^*} \subseteq L_{G_0}$ obtaining Theorem 4.

A detailed proof of this Theorem is found in Appendix B.

A detailed proof of this Proposition is found in Appendix B.

Let $D^*$ be a DAG on a chordal I-map $G^*$ of $P(u)$ consistent with a join tree $T$ of $G^*$, then $D^*$ is an I-map of $P(u)$.

A detailed proof of this Theorem is found in Appendix B.

We have proved that given a chordal I-map of a distribution, obtained triangulating its Markov network, we can define an ordering of variables such that the corresponding Bayesian network has a hypergraph contained in the hypergraph of the chordal I-map. Now we will prove that the intersection hypergraph of (some of) the Bayesian networks is contained in the hypergraph of $G^*$.

Let $G_0$ be the Markov network of $P(u)$ and $G^*$ a chordal I-map obtained adding the edges $I_{1}, \ldots, I_{q}$ to $G_0$. We define $G_j^*$ for $j = 1, \ldots, q$ as follows. Let $G_j$ be the graph obtained

An undirected graph $T$ is a tree if it is connected and acyclic.
from $G^*$ eliminating the edge $l_j$. We define $G^{*j}_j$ as a chordal graph that we get adding edges different from $l_j$ to $G_j$ (this is always possible).

The intersection hypergraph $L^{*}_i$ of $G^{*1}_i, \ldots , G^{*q}_i, G^*$ is contained in the hypergraph $L^*_i$ of the Markov network. Effectively, if $\lambda = \{i_1, \ldots , i_p\} \in L^{*}_i$, the set of vertices $U_\lambda = \{x_{i_1}, \ldots , x_{i_p}\}$ is complete in $G^{*1}_i, \ldots , G^{*q}_i, G^*$, then in $U_\lambda$ there is at most one end of $l_j$, for $j = 1, \ldots , q$. but $U_\lambda$ is complete in $G^*$, which has not more edges than $G_0$. To save $l_1, \ldots , l_p$ then $U_\lambda$ is complete in $G_0$, $\lambda \in L^*_i$. Therefore $L^{*}_i \subseteq L^*_i$, (in fact $L^{*}_i = L^*_i$).

Let $D_1, \ldots , D_q, D$ be, respectively, Bayesian networks on $G^{*1}_i, \ldots , G^{*q}_i, G^*$ consistent with respective join trees. From the preceding result and Theorem 3 we have that the intersection hypergraph $L^{*}_i$ of $D_1, \ldots , D_q, D$ is contained in $L^*_i$. Consequently we can state this theorem.

Theorem 4: Let $P(\mathbf{u})$ be a positive distribution on $\{0,1\}^N$. The intersection hypergraph of the Bayesian networks is contained in the hypergraph of the Markov network.

It is not necessary to consider all $(N!)$ the Bayesian networks in order to obtain a hypergraph more simple than the hypergraph of the Markov network. We can obtain such hypergraph following the procedure described above.

B. Minimum Connectivity of an Intersection Hypergraph of Bayesian Networks

We will terminate this study showing that the Markov network of a distribution establishes a minimum of connectivity for the hypergraphs of Bayesian networks. Let $L$ be the intersection hypergraph of some Bayesian networks. We define the (undirected) graph associated with $L$, $G(L)$, saying that the adjacency in $L$ and in $G(L)$ coincide.

Theorem 5: Let $L$ be the intersection hypergraph of some Bayesian networks of a positive distribution $P(\mathbf{u})$ on $\{0,1\}^N$. The Markov network $G_0$ is a partial graph of $G(L)$.

Corollary 6: If $L$ is contained in the hypergraph of the Markov network $G_0$, then $G(L) = G_0$.

Let us prove the theorem. If $L$ is the intersection hypergraph of some Bayesian networks of $P(\mathbf{u})$, then $P(\mathbf{u}) = Z^{-1}\exp C(\mathbf{u})$, where

$$C(\mathbf{u}) = \sum_{\lambda \in L} u_{\lambda} \alpha(\lambda | \mathbf{u}),$$

(6)

Each $\lambda \in L$ is included in some clique of $G(L)$ so that we can group the terms in (6) according to the cliques of $G(L)$ (there will be several possible groupings). Therefore $P(\mathbf{u})$ is a product of function on the cliques of $G(L)$, then by Section IV, $G(L)$ is an I-map of $P(\mathbf{u})$, i.e., $G_0$ is a partial graph of $G(L)$. The corollary follows directly from the theorem.

Consequently, given a probability distribution, if $L$ is the intersection hypergraph of all the Bayesian networks or the intersection hypergraph of the subset of Bayesian networks provided by the procedure described above, we know that $L$ is more simple than the Markov network hypergraph $L \subseteq L_0$, but the Markov network $G_0$ establishes a minimum connectivity for $L$ in the sense that we have defined, that is $G(L) = G_0$. We conclude with the example studied in Sections IV and V, in Fig. 1. The Markov network $G_0$ is chordal, $G_0 = G^*$. A (directed) join tree is formed by the cliques $C_1 = \{X, Y, Z\}$, $C_2 = \{X, Y, U\}$, $C_3 = \{Y, Z, V\}$, and $C_4 = \{X, Z, W\}$, the clique $C_1$ being the parent of $C_2$, $C_3$ and $C_4$. Ordering the vertices as in $\{X, Y, Z, U, V, W\}$ we have a DAG $D^*$ on $G_0 = G^*$ like $D$ but $X$ being parent of $Y$ and $Z$, and $Y$ parent of $Z$. From it we get $D$, a Bayesian network on $G_0 = G^*$, $L_D$ being the corresponding hypergraph. Besides, the Bayesian networks obtained for other orderings of the variables do not simplify the connections in $L_D$, therefore $L_D$ is the intersection hypergraph of all the Bayesian networks. The hypergraph $L_D$ is more simple than $L_G$, that is $L_D \subseteq L_G$. And $G(L_D) = G_0$. The structure of the HOBM would be given by $L_D$.

VII. CONCLUSION

In this paper we have studied the determination of the structure of the HOBM. We have started observing that if a factorization of the probability function to learn is known, this function can be written through a consensus function whose terms correspond with a determined hypergraph of connections. If various factorizations are known, these terms correspond with the intersection of the respective hypergraphs. Assuming that an expert provides conditional independences of the probability distribution to learn, factorizations are obtained from these conditional independences, and from the factorizations we determine the structure of the HOBM. Next, the learning algorithm of the HOBM can be applied to a set of samples from the distribution to learn, obtaining the approximation distribution that best fits the sample set.

We have considered independence maps, Markov and Bayesian networks that represent conditional independences through undirected graphs and directed acyclic graphs respectively. We have got from these independence maps the hypergraphs that determine the structure of the HOBM.

Given a positive distribution $P(\mathbf{u})$ its Markov network is unique, so the corresponding hypergraph too. In order to define the Bayesian network we have to give an ordering of the variables in $\mathbf{U}$. Therefore we have different Bayesian networks and different hypergraphs. We have established a link between these hypergraphs. We have proved that the intersection hypergraph of all the Bayesian networks is contained by the hypergraph of the Markov network, i.e., it is more simple. The number of Bayesian networks is nonpolynomial ($\mathcal{N}^N$) in the number of variables, and if the computation of the $\mathcal{N}!$ Bayesian networks is not feasible the procedure described in Section VI for obtaining the Bayesian networks $D_1, \ldots , D_q, D$ permits us to define a subset of Bayesian networks whose intersection hypergraph is contained in the hypergraph of the Markov network. Finally, we have proved that although we can get from Bayesian networks a hypergraph more simple than the hypergraph of the Markov network, the Markov network

---

Two vertices are adjacent in the hypergraph if there is a connection of some order between them.
establishes a minimum of connectivity for the hypergraphs
of Bayesian networks.

In practice, the choice between the Bayesian construction
and the Markov construction depends on each problem. The
intersection hypergraph of the \( \Omega \) Bayesian networks is the
minimal one but, if the computation of all the Bayesian
networks is not feasible we can determine an adequate subset
with the procedure described in our work, or we can simply
use the hypergraph of the Markov network. The complexity
of the hypergraph provided by each method depends on the
problem.

In our study several questions remain open for further
investigation. The set of Bayesian networks \( D_1, \ldots, D_q, D \)
provided by the procedure of Section VI is not unique. A
line of research is to get a procedure that obtains a subset
of Bayesian networks such that the intersection hypergraph
is the minimal one. Another line of research is to establish
the conditions for which the hypergraph given by the proce-
dure described in Section VI coincides with the intersection
hypergraph of all the \( \Omega \) Bayesian networks, the minimal
hypergraph.

APPENDIX A

FACTORIZATION HYPERGRAPHS

The hypergraph (3) is obtained from this lemma (proved,
e.g., in [11]).

Lemma 7 (Möbius Inversion): Let \( V \) and \( \Phi \) be real func-
tions defined on the set of all subsets of a finite set \( U \). Then
the following statements are equivalent:

1) \( \forall A \subseteq U : V(A) = \sum_{B \subseteq A} \Phi(B) \)
2) \( \forall B \subseteq U : \Phi(B) = \sum_{A \supseteq B} (-1)^{|A|-|B|} V(A) \)

Given a positive distribution \( P(u) \) on \( \{0,1\}^N \) and defining
for \( \lambda \in \mathcal{P}(\{1, N\}] \)

\[
W_{\lambda} = \sum_{u \in D(\lambda)} (-1)^{|\lambda|-|u|} \ln \frac{P(u)}{P(0)}
\]

(7)

where \( D(\lambda) = \{ u/j \notin \lambda \Rightarrow x_j = 0 \} \) and \( |u| \) is the number of variables with value one, we have from Lemma 7 that

\[
\ln \frac{P(u)}{P(0)} = \sum_{\lambda \in \mathcal{P}(\{1, N\}] W_{\lambda} \alpha(\lambda \mid u)
\]

that is \( P(u) = Z^{-1} \exp(C(u)) \) where \( C(u) = \sum_{\lambda \in \mathcal{P}(\{1, N\}] W_{\lambda} \alpha(\lambda \mid u) \) and \( Z = P(0)^{-1} = \sum_{u} \exp(C(u)) \). Therefore any positive
distribution can be written as the normalized exponential of
a consensus function. Conversely, if a distribution is the
normalized exponential of a consensus function then the
connection weights are given by (7), according to Lemma 7.

If \( P(u) \) admits the factorization (2), by Lemma 7, defining
for \( \lambda \subseteq I(U_2) \)

\[
w_{\lambda} = \sum_{u \in D(\lambda)} (-1)^{|\lambda|-|u|} \ln Q_{\lambda}(u)
\]

where \( D(\lambda) = \{ u/j \notin \lambda \Rightarrow x_j = 0, \forall X_j \in U_2 \} \) and \( |u| \) is the number of variables in \( U_2 \) with value one, we have

\[
\ln Q_{\lambda}(u) = \sum_{\lambda \subseteq I(U_2)} w_{\lambda} \alpha(\lambda \mid u)
\]

being \( a(\lambda \mid u) = \prod_{j \in \lambda} x_j(1) \) if \( \lambda = \emptyset \). Then \( P(u) = Z^{-1} \exp(C(u)) \), where \( Z = P(0)^{-1} = \sum_{u} \exp(C(u)) \) and

\[
C(u) = \sum_{\lambda \subseteq I(U_2)} w_{\lambda} \alpha(\lambda \mid u), \text{ where the connection weights are }
\]

\[
w_{\lambda} = \sum_{\lambda \subseteq I(U_2)} w_{\lambda}
\]

and \( L \) is given by (3).

APPENDIX B

BAYESIAN NETWORKS ON CHORDAL INDEPENDENCE MAPS

We prove now Proposition 2 and Theorem 3. First we
establish the following lemma.

Lemma 8: If \( D^* \) is a DAG on a chordal graph \( G^* \) consistent
with a join tree \( T \) of \( G^* \), then every pair of arrows incident
on a vertex emanate from two adjacent vertices.

Proof: We consider the ordering of vertices of \( T \) and \( G^* \)
used to build \( D^* \). If two arrows incident on \( \gamma \) emanate
from \( \alpha \) and \( \beta \), the indexes of \( \alpha \) and \( \beta \) are smaller than the
index of \( \gamma \). Let \( C_i \) be the first clique where both \( \alpha \) and \( \beta \) appear.
If \( \beta \in C_i \) then \( \alpha \) and \( \beta \) are adjacent. Suppose \( \beta \notin C_i \). Then i \( \neq 1 \) since the index of \( \beta \) is smaller than the index of \( \gamma \). Let \( U_1 \) and \( U_2 \) be, respectively, the union of the cliques of the two
subtrees of \( T \) obtained removing the edge between \( C_i \) and \( C_i \), being \( C_i \in \subset \subseteq U_1 \) and \( C_i \in \subset \subseteq U_2 \). \( \beta \) belongs to some clique with smaller index than \( i \), then \( \beta \in U_1 \). From Theorem 1 we have that \( \cup U_1 \cap U_2 = C_i \cap C_i \), then \( \beta \notin U_2 \), therefore as \( \beta \) and \( \gamma \) are adjacent, \( \gamma \in \subset \subseteq U_2 \). Since the index of \( \alpha \) is smaller than the index of \( \gamma \), \( \alpha \in U_1 \) and \( \alpha \in \subset \subseteq U_2 \), \( \alpha \in C_i \cap C_i \) being contradictory to the definition of \( C_i \). Then \( \alpha \) and \( \beta \) are adjacent.

We prove Proposition 2: If \( D^* \) is a DAG on a chordal I-map
\( G^* \) of \( P(u) \) consistent with a join tree \( T \) of \( G^* \), then \( D^* \) is
an I-map of \( P(u) \).

Proof: Suppose that \( (X \mid Z \mid Y)_{D^*} \) and that \( \alpha \in X \) and \( \beta \in Y \). Let \( \mu(\alpha \mid \beta) \) be a path between \( \alpha \) and \( \beta \). If there
is no vertices of \( Z \) in \( \mu(\alpha \mid \beta) \), since it is blocked, there are
vertices in \( \mu(\alpha \mid \beta) \) with two incident arrows. From Lemma 8,
removing the vertices with two incident vertices of \( \mu(\alpha \mid \beta) \), we get a shorter path \( \mu(\alpha \mid \beta) \). \( \mu(\alpha \mid \beta) \) has no vertices of \( Z \) and it is blocked. Repeating the process, we will conclude that
there is an arrow between \( \alpha \) and \( \beta \), against the hypothesis.
Therefore, there exists some vertex of \( Z \) in \( \mu(\alpha \mid \beta) \). Then
\( (X \mid Z \mid Y)_{G^*} \), so \( X \perp \perp Y \mid Z \).

Finally, we prove Theorem 3: Let \( P(u) \) be a positive distribu-
tion on \( \{0,1\}^N \). Given a chordal I-map \( G^* \), the hypergraph
of any Bayesian network on \( G^* \) consistent with a join tree \( T \)
of \( G^* \) is contained in the hypergraph of \( G^* \).

Proof: Let \( D^* \) be a DAG on \( G^* \) consistent with \( T \) and \( D \) a Bayesian network consistent with \( T \); \( F_X \subseteq F_C \).
Given a vertex \( X_{i} \), let \( C_k \) be the first clique where it appears. If
\( k = 1 \), \( \{X_1\} \cup F_X \subseteq C_k \). Suppose \( k \neq 1 \). \( X_i \notin C_{\{X_i\}} \). Let \( U_1 \) and \( U_2 \) be the unions of the cliques of the two subtrees of

Authorized licensed use limited to: Univ Politecnica de Madrid. Downloaded on March 22,2022 at 12:37:46 UTC from IEEE Xplore. Restrictions apply.
and are adjacent. From Theorem 1 we have that $U_1 \cup V_2 \subseteq C_k$. Therefore $\{X_i\} \cup F_{X_i} \subseteq C_k$. Therefore the hypergraph of the Bayesian network $\bar{D}$ is contained by the hypergraph of $G^*$. 

**REFERENCES**


