Faithfulness in Chain Graphs: The Discrete Case

Jose M. Peña JOSPE@IDA.LIU.SE
Laboratory for Intelligent Information Systems
Department of Computer and Information Science
Linköping University, SE-58183 Linköping, Sweden

Abstract

This paper deals with chain graphs under the classic Lauritzen-Wermuth-Frydenberg interpretation. We prove that the strictly positive discrete probability distributions with the prescribed sample space that factorize according to a chain graph G with dimension d have positive Lebesgue measure wrt $\mathbb{R}^d$, whereas those that factorize according to G but are not faithful to it have zero Lebesgue measure wrt $\mathbb{R}^d$. This means that, in the measure-theoretic sense described, almost all the strictly positive discrete probability distributions with the prescribed sample space that factorize according to G are faithful to it.

Keywords: Chain graphs, faithfulness, Markov equivalence, factorization equivalence, largest chain graphs

1. Introduction

This paper deals with chain graphs under the classic Lauritzen-Wermuth-Frydenberg interpretation. The use of chain graphs to represent independence models in artificial intelligence and statistics has increased over the years, particularly in the case of undirected graphs and acyclic directed graphs. However, the vast majority of independence models that can be represented by chain graphs cannot be represented by undirected graphs or acyclic directed graphs (Peña, 2007). As Studený (2005, Section 1.1) points out, something that would help to judge whether this is an advantage of chain graphs would be proving that any independence model represented by a chain graph exists within an uncertainty calculus of artificial intelligence, e.g. a class of probability distributions or a class of relational databases. In this paper, we prove that for any chain graph there exists a discrete probability distribution with the prescribed sample space that is faithful to it. Actually, we prove a stronger result, namely that the strictly positive discrete probability distributions with the prescribed sample space that factorize according to a chain graph G with dimension d have positive Lebesgue measure wrt $\mathbb{R}^d$, whereas those that factorize according to G but are not faithful to it have zero Lebesgue measure wrt $\mathbb{R}^d$. This means that, in the measure-theoretic sense described, almost all the strictly positive discrete probability distributions with the prescribed sample space that factorize according to G are faithful to it. Previously, it has only been proven that for any chain graph there exists a discrete probability distribution that is faithful to it for some sample space, most likely different from the prescribed sample space (Studený & Bouckaert, 1998, Theorem 7.2). Another related result is that in (Peña et al., 2009, Theorem 3), where it is proven that for any undirected graph there exists a discrete probability distribution with the prescribed sample space that is faithful to it. This result

has also been proven for acyclic directed graphs (Meek, 1995, Theorem 7). The result in Meek (1995) is actually stronger, as it proves that, in a certain measure-theoretic sense, almost all the discrete probability distributions with the prescribed sample space that factorize according to an acyclic directed graph are faithful to it. This paper extends that result to chain graphs.

The rest of the paper is organized as follows. We start by reviewing some concepts in Section 2. In Section 3, we describe how we parameterize the strictly positive discrete probability distributions with the prescribed sample space that factorize according to a chain graph. We present our results on faithfulness in Section 4. In Section 5, we present some results about chain graph equivalence that follow from the results in Section 4. Finally, we close with some discussion in Section 6.

2. Preliminaries

In this section, we define some concepts used later in this paper. The definitions are based upon those in Lauritzen (1996) and Studený (2005). Let X denote a set of discrete random variables. We assume that every random variable $A \in X$ has a prescribed finite sample space of cardinality n_A ($n_A \geq 2$). For simplicity, we assume that the sample space of A are the integer numbers $0, 1, \ldots, n_A - 1$. We denote the number of random variables in $U \subseteq X$ as $|U|$. The elements of X are not distinguished from singletons and the union of the sets $U_1, \ldots, U_n \subseteq X$ is written as the juxtaposition $U_1 \ldots U_n$. We assume throughout the paper that the union of sets precedes the set difference when evaluating an expression. We use upper-case letters to denote random variables and the same letters in lower-case to denote their states. We use x_U with $U \subseteq X$ to denote the projection of x onto U. Unless otherwise stated, all the probability distributions and graphs in this paper are defined over X.

If a graph G contains an undirected (resp. directed) edge between two nodes A_1 and A_2, then we say that $A_1 - A_2$ (resp. $A_1 \rightarrow A_2$) is in G. A path from a node A_1 to a node A_n in a graph G is a sequence of distinct nodes $A_1, \ldots, A_n$ such that there exists an edge in G between A_i and A_{i+1} for all $1 \leq i < n$. A path from A_1 to A_n in G is called descending if $A_i - A_{i+1}$ or $A_i \rightarrow A_{i+1}$ is in G for all $1 \leq i < n$. If there is a descending path from A_1 to A_n in G, then A_1 is called an ancestor of A_n. Let $An_G(U)$ denote the set of ancestors of the nodes in $U \subseteq X$. If $A_1 \rightarrow A_2$ is in G then A_1 is called a parent of A_2. Let $Pa_G(U)$ denote the set of parents of the nodes in $U \subseteq X$. When G is evident from the context, we drop the G from $An_G(U)$ and $Pa_G(U)$, and use $An(U)$ and $Pa(U)$ instead. A directed cycle in G is a sequence of nodes $A_1, \ldots, A_n$ such that $A_i - A_{i+1}$ or $A_i \rightarrow A_{i+1}$ is in G for all $1 \leq i < n$, $A_i \rightarrow A_{i+1}$ is in G for some $1 \leq i < n$, $A_1, \ldots, A_{n-1}$ are distinct, and $A_n = A_1$. A chain graph (CG) is a graph (possibly) containing both undirected and directed edges and no directed cycles. An undirected graph (UG) is a CG containing only undirected edges. The nodes of a CG G can be partitioned into sets $B_1, \ldots, B_n$ called blocks that are well-ordered, i.e. if $A_1 - A_2$ is in G then $A_1, A_2 \in B_i$ for some $1 \leq i \leq n$, whereas if $A_1 \rightarrow A_2$ is in G then $A_1 \in B_i$ and $A_2 \in B_j$ for some $1 \leq i < j \leq n$. The moral graph of a CG G, denoted G^m, is the undirected graph where two nodes are adjacent iff they are adjacent in G or they are both in $Pa(B_i)$ for some block B_i of G. The subgraph of G induced by $U \subseteq X$, denoted G_U, is the graph over U where two nodes are connected by an (un)directed edge if that edge is in G. A set $U \subseteq X$ is complete in an UG G if there is an undirected edge in

G between every pair of distinct nodes in U. We denote the set of complete sets in G by $\mathcal{C}(G)$. We treat all singletons as complete sets and, thus, they are included in $\mathcal{C}(G)$.

Let U, V and W denote three disjoint subsets of X. We represent by $U \perp\!\!\!\perp_p V|W$ that U is independent of V given W in a probability distribution p. Likewise, we represent by $U \perp\!\!\!\perp_G V|W$ that U is separated from V given W in a CG G. Specifically, $U \perp\!\!\!\perp_G V|W$ holds when every path in $(G_{An(UVW)})^m$ from a node in U to a node in V contains a node from W. The independence model represented by a CG G is the set of separation statements $U \perp\!\!\!\perp_G V|W$. We say that a probability distribution p is Markovian wrt a CG G when $U \perp\!\!\!\perp_p V|W$ if $U \perp\!\!\!\perp_G V|W$ for all U, V and W disjoint subsets of X. We say that p is faithful to G when $U \perp\!\!\!\perp_p V|W$ iff $U \perp\!\!\!\perp_G V|W$ for all U, V and W disjoint subsets of X. We represent by $U \not\perp\!\!\!\perp_p V|W$ and $U \not\perp\!\!\!\perp_G V|W$ that $U \perp\!\!\!\perp_p V|W$ and $U \perp\!\!\!\perp_G V|W$ do not hold, respectively. Given $U, V \subseteq X$ such that $UV = X$, we say that an UG G decomposes into G_U and G_V if $U \cap V$ is a complete set in G and $U \setminus V \perp\!\!\!\perp_G V \setminus U|U \cap V$.

3. Parameterization of Chain Graphs

In this section, we describe how we parameterize the strictly positive probability distributions that factorize according to a CG. This is a key issue, because our results about faithfulness are not only relative to the CG at hand and the measure considered, the Lebesgue measure, but also to the dimension of the strictly positive probability distributions that factorize according to the CG at hand. Our parameterization is inspired by Besag (1974, p. 197).

We say that a strictly positive probability distribution p factorizes according to a CG G with n blocks if the following two conditions are met (Lauritzen, 1996, p. 53):

1. $p(x) = \prod_{i=1}^{n} p(x_{B_i}|x_{Pa(B_i)})$ where
2. $p(x_{B_i}|x_{Pa(B_i)}) = \prod_{C \in \mathcal{C}((G_{B_i Pa(B_i)})^m)} \psi_C^i(x_C)$ where $\psi_C^i(x_C)$ are positive real functions.

Let $\mathbf{0}_U$ denote that every random variable in $U \subseteq X$ takes value 0. Then, Condition 2 above is equivalent to the following condition:

2′. $p(x_{B_i}|x_{Pa(B_i)}) = p(\mathbf{0}_{B_i}|x_{Pa(B_i)}) \prod_{C \in \mathcal{C}((G_{B_i Pa(B_i)})^m)} \phi_C^i(x_C)$ where $\phi_C^i(x_C)$ are positive real functions.

To see it, let $\psi_C^i(x_C) = p(\mathbf{0}_{B_i}|x_{Pa(B_i)})\phi_C^i(x_C)$ if $C = Pa(B_i)$, and $\psi_C^i(x_C) = \phi_C^i(x_C)$ otherwise. Note that $Pa(B_i) \in \mathcal{C}((G_{B_i Pa(B_i)})^m)$. The motivation behind this redefinition of the factorization according to a CG is to use $p(\mathbf{0}_{B_i}|x_{Pa(B_i)})$ as a reference probability and define the rest of the probabilities $p(x_{B_i}|x_{Pa(B_i)})$ relative to it.

Since $\sum_{x_{B_i}} p(x_{B_i}|x_{Pa(B_i)}) = 1$ for all $x_{Pa(B_i)}$, it follows from Condition 2′ above that

$$p(\mathbf{0}_{B_i}|x_{Pa(B_i)}) = \frac{1}{\sum_{x_{B_i}} \prod_{C \in \mathcal{C}((G_{B_i Pa(B_i)})^m)} \phi_C^i(x_C)}.$$

Thus,

$$p(x_{B_i}|x_{Pa(B_i)}) = \frac{\prod_{C \in \mathcal{C}((G_{B_i Pa(B_i)})^m)} \phi_C^i(x_C)}{\sum_{x_{B_i}} \prod_{C \in \mathcal{C}((G_{B_i Pa(B_i)})^m)} \phi_C^i(x_C)}.$$

Besides, let $\mathcal{K}((G_{B_iPa(B_i)})^m) = \{C \in \mathcal{C}((G_{B_iPa(B_i)})^m) : C \cap B_i \neq \emptyset\} = \{C \in \mathcal{C}((G_{B_iPa(B_i)})^m) : C \not\subseteq Pa(B_i)\}$. Then,

$$p(x_{B_i}|x_{Pa(B_i)})$$

$$= \frac{[\prod_{C \in \mathcal{C}((G_{B_iPa(B_i)})^m) \setminus \mathcal{K}((G_{B_iPa(B_i)})^m)} \phi^i_C(x_C)][\prod_{C \in \mathcal{K}((G_{B_iPa(B_i)})^m)} \phi^i_C(x_C)]}{[\prod_{C \in \mathcal{C}((G_{B_iPa(B_i)})^m) \setminus \mathcal{K}((G_{B_iPa(B_i)})^m)} \phi^i_C(x_C)][\sum_{x_{B_i}} \prod_{C \in \mathcal{K}((G_{B_iPa(B_i)})^m)} \phi^i_C(x_C)]}$$

$$= \frac{\prod_{C \in \mathcal{K}((G_{B_iPa(B_i)})^m)} \phi^i_C(x_C)}{\sum_{x_{B_i}} \prod_{C \in \mathcal{K}((G_{B_iPa(B_i)})^m)} \phi^i_C(x_C)}. \quad (1)$$

Let $\mathcal{D}(G)^+$ denote the set of strictly positive probability distributions that factorize according to G. We parameterize the probability distributions in $\mathcal{D}(G)^+$ by parameterizing the functions $\phi^i_C(x_C)$ in Equation 1 for all $1 \leq i \leq n$. In particular, let $\theta^i_C(x_C)$ denote the parameter corresponding to the value $\phi^i_C(x_C)$. Thus, we can express any probability distribution $p \in \mathcal{D}(G)^+$ in terms of the parameters as $p(x) = \prod_{i=1}^n p(x_{B_i}|x_{Pa(B_i)})$ where

$$p(x_{B_i}|x_{Pa(B_i)}) = \frac{\prod_{C \in \mathcal{K}((G_{B_iPa(B_i)})^m)} \theta^i_C(x_C)}{\sum_{x_{B_i}} \prod_{C \in \mathcal{K}((G_{B_iPa(B_i)})^m)} \theta^i_C(x_C)}. \quad (2)$$

We say that a parameter is freely assignable (fa) if it can take different values independently of the other parameters. We restrict the parameters $\theta^i_C(x_C)$ where $x_A = 0$ for some $A \in C$ to take value one and, thus, these parameters are not fa. The rest of the parameters are fa, as we will show in Lemma 1. For example, consider $C = AB$ where A and B are two random variables that can take three possible values (0, 1 and 2). Then, $\theta^i_{AB}(ab)$ with $a, b \in \{1, 2\}$ are fa, whereas $\theta^i_{AB}(0b)$ and $\theta^i_{AB}(a0)$ with $a, b \in \{0, 1, 2\}$ are not fa because we have restricted them to take value one. Lemma 1 describes the values that the fa parameters can take. The goal with our parameterization is to use as few parameters as possible so that the parameter values corresponding to each probability distribution in $\mathcal{D}(G)^+$ are uniquely determined. We show that we achieve this goal in Lemma 2, which is crucial in proving the main result in this paper (Theorems 3 and 5). It is worth mentioning that if we did not restrict the parameters $\theta^i_C(x_C)$ where $x_A = 0$ for some $A \in C$ to take value one, then the parameter values corresponding to each probability distribution in $\mathcal{D}(G)^+$ would not be uniquely determined. An example follows. Consider $X = AB$ where A and B are two random variables that can take two possible values (0 and 1). Consider $G = \{A - B\}$. Then, the parameters involved in the example are $\theta^1_{AB}(00), \theta^1_{AB}(01), \theta^1_{AB}(10), \theta^1_{AB}(11), \theta^1_A(0), \theta^1_A(1), \theta^1_B(0)$ and $\theta^1_B(1)$. Consider the uniform probability distribution, which obviously is in $\mathcal{D}(G)^+$. Then, there are at least two sets of parameter values that give rise to this probability distribution: One, the set where every parameter takes value one and, two, the set where every parameter takes value one except $\theta^1_{AB}(01) = \theta^1_{AB}(10) = 2$, $\theta^1_{AB}(11) = 4$, and $\theta^1_A(1) = \theta^1_B(1) = 1/2$. This can easily be checked with the help of Equation 2. It is also worth mentioning that parameterizing the functions in Condition 2 above directly also results in overparameterization and arbitrariness in parameter values (Lauritzen, 1996, p. 35).

We define the dimension of G as the number of fa parameters in our parameterization of the probability distributions in $\mathcal{D}(G)^+$ and we denote it as d. Then, $d =$

$\sum_{i=1}^{n} \sum_{C \in \mathcal{K}((G_{B_i Pa(B_i)})^m)} \prod_{A \in C} (n_A - 1)$. We define the fa parameter space for $\mathcal{D}(G)^+$ as the set of values that the fa parameters are allowed to take. In this paper, the fa parameter space for $\mathcal{D}(G)^+$ is $(0, \infty)^d$. The following lemma supports this choice.

Lemma 1 *Let G be a CG of dimension d. Any element of the fa parameter space for $\mathcal{D}(G)^+$ corresponds to some probability distribution in $\mathcal{D}(G)^+$.*

Proof First, we show that any element of $(0, \infty)^d$ gives rise to some strictly positive probability distribution. Note that any such element gives rise to some strictly positive probability distribution $q^i(x_{B_i}|x_{Pa(B_i)})$ for all $1 \leq i \leq n$ by Equation 2, because

$$0 < \frac{\prod_{C \in \mathcal{K}((G_{B_i Pa(B_i)})^m)} \theta^i_C(x_C)}{\sum_{x_{B_i}} \prod_{C \in \mathcal{K}((G_{B_i Pa(B_i)})^m)} \theta^i_C(x_C)} < 1$$

for all x_{B_i} and $x_{Pa(B_i)}$, and

$$\sum_{x_{B_i}} \frac{\prod_{C \in \mathcal{K}((G_{B_i Pa(B_i)})^m)} \theta^i_C(x_C)}{\sum_{x_{B_i}} \prod_{C \in \mathcal{K}((G_{B_i Pa(B_i)})^m)} \theta^i_C(x_C)} = 1$$

for all $x_{Pa(B_i)}$. Therefore, any element of $(0, \infty)^d$ gives rise to some strictly positive probability distribution $q(x) = \prod_{i=1}^{n} q^i(x_{B_i}|x_{Pa(B_i)})$, because

$$0 < \prod_{i=1}^{n} q^i(x_{B_i}|x_{Pa(B_i)}) < 1$$

for all x, and

$$\sum_{x_{B_1 \ldots B_n}} \prod_{i=1}^{n} q^i(x_{B_i}|x_{Pa(B_i)})$$

$$= \sum_{x_{B_1}} q^1(x_{B_1}|x_{Pa(B_1)})[\sum_{x_{B_2}} q^2(x_{B_2}|x_{Pa(B_2)})[\ldots[\sum_{x_{B_n}} q^n(x_{B_n}|x_{Pa(B_n)})]\ldots]] = 1.$$

Now, we show that $q \in \mathcal{D}(G)^+$. Note that for all $1 \leq i < n$

$$\sum_{x_{B_{i+1} \ldots B_n}} \prod_{l=i+1}^{n} q^l(x_{B_l}|x_{Pa(B_l)})$$

$$= \sum_{x_{B_{i+1}}} q^{i+1}(x_{B_{i+1}}|x_{Pa(B_{i+1})})[\sum_{x_{B_{i+2}}} q^{i+2}(x_{B_{i+2}}|x_{Pa(B_{i+2})})[\ldots[\sum_{x_{B_n}} q^n(x_{B_n}|x_{Pa(B_n)})]\ldots]] = 1.$$

Thus, for all $1 \leq i \leq n$, it follows from the equation above that

$$q(x_{B_i Pa(B_i)}) = \sum_{x_{B_1 \ldots B_n \setminus B_i Pa(B_i)}} \prod_{l=1}^{n} q^l(x_{B_l}|x_{Pa(B_l)})$$

$$= \sum_{x_{B_1 \ldots B_{i-1} \setminus Pa(B_i)}} [[\prod_{l=1}^{i} q^l(x_{B_l}|x_{Pa(B_l)})][\sum_{x_{B_{i+1} \ldots B_n}} \prod_{l=i+1}^{n} q^l(x_{B_l}|x_{Pa(B_l)})]]$$

$$= \sum_{x_{B_1 \ldots B_{i-1} \setminus Pa(B_i)}} \prod_{l=1}^{i} q^l(x_{B_l}|x_{Pa(B_l)})$$

$$= q^i(x_{B_i}|x_{Pa(B_i)}) \sum_{x_{B_1 \ldots B_{i-1} \setminus Pa(B_i)}} \prod_{l=1}^{i-1} q^l(x_{B_l}|x_{Pa(B_l)}). \quad (3)$$

Thus, for all $1 \leq i \leq n$, it follows from the equation above that

$$q(x_{Pa(B_i)}) = \sum_{x_{B_i}} q(x_{B_i Pa(B_i)}) = [\sum_{x_{B_i}} q^i(x_{B_i}|x_{Pa(B_i)})][\sum_{x_{B_1 \ldots B_{i-1} \setminus Pa(B_i)}} \prod_{l=1}^{i-1} q^l(x_{B_l}|x_{Pa(B_l)})]$$

$$= \sum_{x_{B_1 \ldots B_{i-1} \setminus Pa(B_i)}} \prod_{l=1}^{i-1} q^l(x_{B_l}|x_{Pa(B_l)}). \quad (4)$$

Then, for all $1 \leq i \leq n$

$$q(x_{B_i}|x_{Pa(B_i)}) = \frac{q(x_{B_i Pa(B_i)})}{q(x_{Pa(B_i)})} = q^i(x_{B_i}|x_{Pa(B_i)}) \quad (5)$$

due to Equations 3 and 4. Therefore,

$$q(x) = \prod_{i=1}^{n} q^i(x_{B_i}|x_{Pa(B_i)}) = \prod_{i=1}^{n} q(x_{B_i}|x_{Pa(B_i)})$$

and, thus, q satisfies Condition 1 above. Moreover, $q(x_{B_i}|x_{Pa(B_i)})$ satisfies Condition 2 above for all $1 \leq i \leq n$: It suffices to note that $q(x_{B_i}|x_{Pa(B_i)}) = q^i(x_{B_i}|x_{Pa(B_i)})$ by Equation 5, and that $q^i(x_{B_i}|x_{Pa(B_i)})$ satisfies Condition 2 because

$$q^i(x_{B_i}|x_{Pa(B_i)}) = \frac{\prod_{C \in \mathcal{K}((G_{B_i Pa(B_i)})^m)} \theta^i_C(x_C)}{\sum_{x_{B_i}} \prod_{C \in \mathcal{K}((G_{B_i Pa(B_i)})^m)} \theta^i_C(x_C)} = \prod_{C \in \mathcal{C}((G_{B_i Pa(B_i)})^m)} \psi^i_C(x_C)$$

where $\psi^i_C(x_C) = 1$ if $C \subset Pa(B_i)$,

$$\psi^i_C(x_C) = \frac{1}{\sum_{x_{B_i}} \prod_{C \in \mathcal{K}((G_{B_i Pa(B_i)})^m)} \theta^i_C(x_C)}$$

if $C = Pa(B_i)$, and $\psi^i_C(x_C) = \theta^i_C(x_C)$ otherwise. Note that $Pa(B_i) \in \mathcal{C}((G_{B_i Pa(B_i)})^m)$. ■

Lemma 2 *Let G be a CG. There is a one-to-one correspondence between the elements of the fa parameter space for $\mathcal{D}(G)^+$ and the probability distributions in $\mathcal{D}(G)^+$.*

Proof First, we show that any probability distribution $p \in \mathcal{D}(G)^+$ corresponds to some element of the fa parameter space for $\mathcal{D}(G)^+$. The fa parameter values corresponding to p can be computed from p as follows. Consider any fa parameter $\theta^i_K(x_K)$ such that the values of the fa parameters $\theta^i_C(x_C)$ for all $C \in \mathcal{K}((G_{B_i Pa(B_i)})^m)$ such that $C \subset K$ have already been computed. The value of $\theta^i_K(x_K)$ can be computed as follows. Note that

$$p(\mathbf{0}_{B_i}|x_{Pa(B_i)}) = \frac{1}{\sum_{x_{B_i}} \prod_{C \in \mathcal{K}((G_{B_i Pa(B_i)})^m)} \theta^i_C(x_C)}$$

by Equation 2. To see this, recall from above that $C \cap B_i \neq \emptyset$ for all $C \in \mathcal{K}((G_{B_i Pa(B_i)})^m)$. Thus, every fa parameter in the numerator of Equation 2 takes value one due to our restriction that $\theta^i_C(x_C) = 1$ if $x_A = 0$ for some $A \in C$. Then, it follows from the equation above and Equation 2 that

$$p(x_{B_i}|x_{Pa(B_i)}) = p(\mathbf{0}_{B_i}|x_{Pa(B_i)}) \prod_{C \in \mathcal{K}((G_{B_i Pa(B_i)})^m)} \theta^i_C(x_C).$$

Therefore,

$$p(x_{B_i \cap K}, \mathbf{0}_{B_i \setminus K}|x_{Pa(B_i) \cap K}, \mathbf{0}_{Pa(B_i) \setminus K}) =$$
$$p(\mathbf{0}_{B_i}|x_{Pa(B_i) \cap K}, \mathbf{0}_{Pa(B_i) \setminus K}) \prod_{\{C \in \mathcal{K}((G_{B_i Pa(B_i)})^m) : C \subseteq K\}} \theta^i_C(x_C).$$

Consequently,

$$\theta^i_K(x_K) = \frac{p(x_{B_i \cap K}, \mathbf{0}_{B_i \setminus K}|x_{Pa(B_i) \cap K}, \mathbf{0}_{Pa(B_i) \setminus K})}{p(\mathbf{0}_{B_i}|x_{Pa(B_i) \cap K}, \mathbf{0}_{Pa(B_i) \setminus K}) \prod_{\{C \in \mathcal{K}((G_{B_i Pa(B_i)})^m) : C \subset K\}} \theta^i_C(x_C)}.$$

Note that $\theta^i_K(x_K)$ always takes a positive real value because p is strictly positive. Moreover, different probability distributions in $\mathcal{D}(G)^+$ correspond to different elements of the fa parameter space for $\mathcal{D}(G)^+$. To see it, assume to the contrary that there exist two distinct probability distributions $p, p' \in \mathcal{D}(G)^+$ that correspond to the same element. Note that this element uniquely identifies $p(x_{B_i}|x_{Pa(B_i)})$ by Equation 2 for all $1 \leq i \leq n$. Likewise, it uniquely identifies $p'(x_{B_i}|x_{Pa(B_i)})$ for all $1 \leq i \leq n$. Then, $p(x_{B_i}|x_{Pa(B_i)}) = p'(x_{B_i}|x_{Pa(B_i)})$ for all $1 \leq i \leq n$. However, this contradicts the assumption that p and p' are distinct by Condition 1 above.

Finally, we show that different elements of the fa parameter space for $\mathcal{D}(G)^+$ correspond to different probability distributions in $\mathcal{D}(G)^+$. Assume to the contrary that two distinct elements of the fa parameter space for $\mathcal{D}(G)^+$ correspond to the same probability distribution $q \in \mathcal{D}(G)^+$. There are two scenarios to consider:

- If there exists some $x_{Pa(B_i)}$ such that $\sum_{x_{B_i}} \prod_{C \in \mathcal{K}((G_{B_i Pa(B_i)})^m)} \theta^i_C(x_C)$ takes different value for the two elements, then the two elements differ in the value of $q^i(\mathbf{0}_{B_i}|x_{Pa(B_i)})$, because
$$q^i(\mathbf{0}_{B_i}|x_{Pa(B_i)}) = \frac{1}{\sum_{x_{B_i}} \prod_{C \in \mathcal{K}((G_{B_i Pa(B_i)})^m)} \theta^i_C(x_C)}$$
by Equation 2. To see it, recall from above that $C \cap B_i \neq \emptyset$ for all $C \in \mathcal{K}((G_{B_i Pa(B_i)})^m)$. Thus, every parameter in the numerator of Equation 2 takes value one due to our restriction that $\theta^i_C(x_C) = 1$ if $x_A = 0$ for some $A \in C$.

- Assume that $\sum_{x_{B_i}} \prod_{C \in \mathcal{K}((G_{B_i Pa(B_i)})^m)} \theta^i_C(x_C)$ takes the same value for the two elements for all $x_{Pa(B_i)}$. Since the two elements are different, it is possible to find a fa parameter $\theta^i_K(x_K)$ that takes different value in the two elements whereas for all $C \subset K$ the fa parameter $\theta^i_C(x_C)$ takes the same value in the two elements. Then, the two elements differ in the value of $q^i(x_{B_i \cap K}, \mathbf{0}_{B_i \setminus K} | x_{Pa(B_i) \cap K}, \mathbf{0}_{Pa(B_i) \setminus K})$, because

$$q^i(x_{B_i \cap K}, \mathbf{0}_{B_i \setminus K} | x_{Pa(B_i) \cap K}, \mathbf{0}_{Pa(B_i) \setminus K}) = \frac{\prod_{\{C \in \mathcal{K}((G_{B_i Pa(B_i)})^m) : C \subseteq K\}} \theta^i_C(x_C)}{\sum_{x_{B_i}} \prod_{C \in \mathcal{K}((G_{B_i Pa(B_i)})^m)} \theta^i_C(x_C)}$$

  by Equation 2.

Either scenario implies that the two elements differ in $q(x_{B_i} | x_{Pa(B_i)})$ by Equation 5, which is a contradiction. ■

4. Faithfulness in Chain Graphs

The two theorems below are the main contribution of this manuscript. They prove that for any CG G, in the measure-theoretic sense described below, almost all the probability distributions in $\mathcal{D}(G)^+$ are faithful to G.

Theorem 3 *Let G be a CG of dimension d. $\mathcal{D}(G)^+$ has positive Lebesgue measure wrt $\mathbb{R}^d$.*

Proof The one-to-one correspondence proved in Lemma 2 enables us to compute the Lebesgue measure wrt $\mathbb{R}^d$ of $\mathcal{D}(G)^+$ as the Lebesgue measure wrt $\mathbb{R}^d$ of the fa parameter space for $\mathcal{D}(G)^+$. It follows from Lemma 1 that the fa parameter space for $\mathcal{D}(G)^+$ has positive volume wrt $\mathbb{R}^d$ and, thus, that it has positive Lebesgue measure wrt $\mathbb{R}^d$. ■

Lemma 4 *Let U, V and W denote three disjoint subsets of X such that $U \not\perp_G V | W$ where G is a CG. Then, there exists a probability distribution $p \in \mathcal{D}(G)^+$ such that $U \not\perp_p V | W$.*

Proof Assume for a moment that the random variables in X are all binary. Then, there exists a strictly positive binary probability distribution q that is Markovian wrt G and such that $U \not\perp_q V | W$ (Studený & Bouckaert, 1998, Consequence 5.2). Then, there is some state x_{UVW} of UVW such that

$$q(x_{UVW}) q(x_W) - q(x_{UW}) q(x_{VW}) \neq 0. \tag{6}$$

Moreover, $q \in \mathcal{D}(G)^+$ (Frydenberg, 1990, Theorem 4.1). Then, q corresponds to some fa parameter values by Lemma 2. We can expand q to a probability distribution $p \in \mathcal{D}(G)^+$ over the original cardinalities of the random variables in X by assigning an arbitrarily small positive real value ϵ to the additional fa parameters, i.e. $\theta^i_C(x_C) = \epsilon$ where $x_A \geq 1$ for all $A \in C$ and $x_A > 1$ for some $A \in C$. This implies that $p(x)$ can be made arbitrarily close to $q(x)$ for all $x \in \{0, 1\}^{|X|}$ which, in turn, implies that $p(x_{UVW}), p(x_W), p(x_{UW})$ and $p(x_{VW})$

can jointly be made arbitrarily close to $q(x_{UVW}), q(x_W), q(x_{UW})$ and $q(x_{VW})$, respectively. This implies that, for some ϵ,

$$p(x_{UVW})p(x_W) - p(x_{UW})p(x_{VW}) \neq 0$$

due to Equation 6. Hence, $U \not\perp\!\!\!\perp_p V|W$. ■

Theorem 5 *Let G be a CG of dimension d. The set of probability distributions in $\mathcal{D}(G)^+$ that are not faithful to G has zero Lebesgue measure wrt $\mathbb{R}^d$.*

Proof The proof basically proceeds in the same way as that of Meek (1995, Theorem 7). Since any probability distribution $p \in \mathcal{D}(G)^+$ is Markovian wrt G (Frydenberg, 1990, Theorem 4.1), for p not to be faithful to G, p must satisfy some independence that is not entailed by G. That is, there must exist three disjoint subsets of X, here denoted as U, V and W, such that $U \not\perp\!\!\!\perp_G V|W$ but $U \perp\!\!\!\perp_p V|W$. Now, note that $U \perp\!\!\!\perp_p V|W$ iff

$$p(x_{UVW})p(x_W) - p(x_{UW})p(x_{VW}) = 0 \tag{7}$$

for all x_{UVW}. Note that for all x_{UVW}, each term $p(x_Z)$ in the left-hand side of the equation above can be expressed as

$$p(x_Z) = \sum_{x_{X \setminus Z}} \prod_{i=1}^{n} p(x_{B_i}|x_{Pa(B_i)}).$$

Recall from Equation 2 that, for all $1 \leq i \leq n$, $p(x_{B_i}|x_{Pa(B_i)})$ is a fraction of polynomials in the fa parameters in the parameterization of the probability distributions in $\mathcal{D}(G)^+$. Therefore, by simple algebraic manipulation, we can express the left-hand side of Equation 7 as a fraction of polynomials in the fa parameters. Consequently, for $U \perp\!\!\!\perp_p V|W$ to hold, the polynomial in the numerator of such a fraction must be zero. Note that some fa parameters may not appear in this polynomial. Each of these fa parameters can be added to it as a term with coefficient equal to zero. Let us allow the fa parameters to take any real value (originally, only positive real values were allowed). Then, for every x_{UVW} we have a real polynomial in real variables (i.e. the fa parameters) that should be satisfied for p not to be faithful to G. We interpret each of these polynomials as a real function on a real Euclidean space that includes the fa parameter space for $\mathcal{D}(G)^+$. Furthermore, each of these polynomials is non-trivial, that is, not all the values of the fa parameters are solutions to the polynomial. To prove this, it suffices to prove that there exists a probability distribution $p'' \in \mathcal{D}(G)^+$ for which the polynomial does not hold. Consider the polynomial for x_{UVW}. Note that, by Lemma 4, there exits a probability distribution $p' \in \mathcal{D}(G)^+$ such that $U \not\perp\!\!\!\perp_{p'} V|W$. Then, there is some state x'_{UVW} of UVW such that

$$p'(x'_{UVW})p'(x'_W) - p'(x'_{UW})p'(x'_{VW}) \neq 0.$$

Then, by renaming the possible states of the random variables in UVW appropriately, we can transform p' into the desired p''.

Let $sol(x_{UVW})$ denote the set of solutions to the polynomial for x_{UVW} referred above. Then, $sol(x_{UVW})$ has zero Lebesgue measure wrt $\mathbb{R}^d$ because it consists of the solutions to a non-trivial polynomial in real variables (i.e. the fa parameters) (Okamoto, 1973). Let $sol = \bigcup_{\{U,V,W \subseteq X \text{ disjoint} : U \not\perp_G V|W\}} \bigcap_{x_{UVW}} sol(x_{UVW})$. Then, sol has zero Lebesgue measure wrt $\mathbb{R}^d$, because the finite union and intersection of sets of zero Lebesgue measure has zero Lebesgue measure too. Consequently, the probability distributions in $\mathcal{D}(G)^+$ that are not faithful to G correspond to a set of elements of the fa parameter space for $\mathcal{D}(G)^+$ that has zero Lebesgue measure wrt $\mathbb{R}^d$ because it is contained in sol. Since this correspondence is one-to-one by Lemma 2, the probability distributions in $\mathcal{D}(G)^+$ that are not faithful to G also have zero Lebesgue measure wrt $\mathbb{R}^d$. ■

The only previous result on faithfulness in CGs that we are aware of is Theorem 7.2 in Studený & Bouckaert (1998), where it is proven that for any CG there exists a probability distribution that is faithful to it for some sample space. The two theorems above imply a stronger result, namely that for any CG G and any sample space there exists a probability distribution that is faithful to G.

5. Equivalence in Chain Graphs

The space of CGs can be divided in classes of equivalent CGs according to criteria such as Markov independence equivalence, Markov distribution equivalence, or factorization equivalence. As we prove below with the help of the theorems above, these criteria actually coincide in some cases. This result is important because the classes of Markov distribution equivalent CGs have a simple graphical characterization and a natural representative, the so-called largest CG, which now also apply to the classes of equivalence induced by the other criteria. We also prove below that all equivalent CGs have the same dimension wrt the parameterization introduced in Section 3.

Before proving our results, we formally define the equivalence criteria discussed in the paragraph above. Recall that, unless otherwise stated, all the probability distributions in this paper are over X. We say that two CGs are Markov independence equivalent if they represent the same independence model. We say that two CGs are Markov distribution equivalent wrt a class of probability distributions if every probability distribution in the class is Markovian wrt both CGs or wrt neither of them. We say that two CGs G and H are factorization equivalent if $\mathcal{D}(G)^+ = \mathcal{D}(H)^+$. The corollary below proves that these definitions coincide in some cases.

Corollary 6 *Let G and H denote two CGs. The following statements are equivalent:*

1. *G and H are Markov independence equivalent,*
2. *G and H are Markov distribution equivalent wrt the class of strictly positive probability distributions,*
3. *G and H are Markov distribution equivalent wrt any superset of the class of strictly positive probability distributions, and*
4. *G and H are factorization equivalent.*

Proof First, we prove that Statements 1 and 2 are equivalent. By definition, Markov independence equivalence implies Markov distribution equivalence wrt any class of probability distributions. To see the opposite implication, note that if G and H are not Markov independence equivalent, then one of them, say G, must represent a separation statement $U \perp\!\!\!\perp_G V|W$ that is not represented by H. Consider a probability distribution $p \in \mathcal{D}(H)^+$ faithful to H. Such a probability distribution exists due to Theorems 3 and 5, and it is Markovian wrt H (Frydenberg, 1990, Theorem 4.1). However, p cannot be Markovian wrt G, because $U \not\perp\!\!\!\perp_H V|W$ implies $U \not\perp\!\!\!\perp_p V|W$.

Now, we prove that Statements 1 and 3 are equivalent. By definition, Markov independence equivalence implies Markov distribution equivalence wrt any class of probability distributions. To see the opposite implication, note that if G and H are Markov distribution equivalent wrt a superset of the class of strictly positive probability distributions, then they also are Markov distribution equivalent wrt the class of strictly positive probability distributions and, thus, they are Markov independence equivalent by the paragraph above.

Finally, the equivalence of Statements 2 and 4 follows from Frydenberg (1990, Theorem 4.1). ■

Frydenberg (1990, Theorem 5.6) gives a straightforward graphical characterization of Markov distribution equivalence wrt a superset of the class of strictly positive probability distributions. Due to the corollary above, that is also a graphical characterization of the other three types of equivalence discussed in there. Hereinafter, we do not distinguish anymore between the different types of equivalence discussed in the corollary above because they coincide and, thus, we simply refer to them as equivalence. It is worth mentioning that the corollary above has also been proven in (Studený et al., 2009, Theorem 16). However, our proof is completely different: Our proof builds upon the fact that for any CG there exists a strictly positive probability distribution with the prescribed sample space that is faithful to it due to Theorems 3 and 5, whereas the proof in (Studený et al., 2009) does not build upon this fact because it has not been proven before.

Frydenberg (1990, Proposition 5.7) shows that every class of equivalent CGs contains a unique CG that has more undirected edges than any other CG in the class. Such a CG is called the largest CG (LCG) in the class, and it is usually considered a natural representative of the class. Studený (1998, Section 4.2) conjectures that the LCG G in a class of equivalent CGs has fewer fa parameters than any other CG in the class. This would imply that the most space efficient way of storing the probability distributions in $\mathcal{D}(G)^+$ is by factorizing them according to G rather than according to any other CG equivalent to G. The conjecture is stated in Studený (1998) with no particular parameterization in mind. The corollary below disproves the conjecture for the parameterization proposed in Section 3.

Corollary 7 *All equivalent CGs have the same dimension wrt the parameterization proposed in Section 3.*

Proof Studený et al. (2009) study the so-called feasible merging operation, which merges two blocks of a CG that satisfy certain conditions into a larger block by dropping the direction of the edges between the former two blocks. Let H and H' denote the CG before

and after, respectively, merging the blocks U and L into the block M. It is proven in (Studený et al., 2009, Lemma 32) that $(H'_{MPa_{H'}(M)})^m$ decomposes into $(H_{UPa_H(U)})^m$ and $(H_{LPa_H(L)})^m$, with a shared set of nodes $Pa_H(L)$. It follows from this decomposition that

$$\mathcal{C}((H'_{MPa_{H'}(M)})^m) = \mathcal{C}((H_{UPa_H(U)})^m) \cup \mathcal{C}((H_{LPa_H(L)})^m).$$

Note, however, that $\mathcal{C}((H_{UPa_H(U)})^m) \cap \mathcal{C}((H_{LPa_H(L)})^m) = \{C : C \subseteq Pa_H(L)\}$ due to the shared set of nodes in the decomposition. Then,

$$\mathcal{C}((H'_{MPa_{H'}(M)})^m) = \mathcal{C}((H_{UPa_H(U)})^m) \cup \mathcal{K}((H_{LPa_H(L)})^m).$$

Moreover, $\mathcal{C}((H_{UPa_H(U)})^m) \cap \mathcal{K}((H_{LPa_H(L)})^m) = \emptyset$.

It is also proven in (Studený et al., 2009, Lemma 32) that $Pa_{H'}(M) = Pa_H(U)$. Then,

$$\mathcal{K}((H'_{MPa_{H'}(M)})^m) = \mathcal{K}((H_{UPa_H(U)})^m) \cup \mathcal{K}((H_{LPa_H(L)})^m).$$

Moreover, $\mathcal{K}((H_{UPa_H(U)})^m) \cap \mathcal{K}((H_{LPa_H(L)})^m) = \emptyset$. Consequently, the contribution of M to the dimension of H' is the same as the sum of the contributions of U and L to the dimension of H. Thus, H and H' have the same dimension.

Let G denote the LCG in a class of equivalent CGs. Let G' denote any other CG in the class. Then, there exists a sequence of equivalent CGs $G' = G_1, \ldots, G_r = G$ such that each G_{i+1} is obtained from G_i by a feasible merging operation (Studený et al., 2009, Lemma 5 and Corollary 7). It follows from the paragraphs above that all the CGs in the sequence have the same dimension. Consequently, any two CGs in the equivalence class of G have the same dimension. ■

6. Conclusions

In this paper, we have proven that, in a certain measure-theoretic sense, almost all the strictly positive discrete probability distributions with the prescribed sample space that factorize according to a chain graph are faithful to it. This result extends previous results such as

- (Studený & Bouckaert, 1998, Theorem 7.2) where it is proven that for any chain graph there exists a discrete probability distribution that is faithful to it for some sample space, most likely different from the prescribed sample space,
- (Peña et al., 2009, Theorem 3) where it is proven that for any undirected graph there exists a discrete probability distribution with the prescribed sample space that is faithful to it, and
- (Meek, 1995, Theorem 7) where it is proven that, in a certain measure-theoretic sense, almost all the discrete probability distributions with the prescribed sample space that factorize according to an acyclic directed graph are faithful to it.

In this paper, we have also proven a number of consequences that follow from the result discussed above:

- The fact that the vast majority of independence models that can be represented by chain graphs cannot be represented by undirected graphs or acyclic directed graphs is an advantage of chain graphs, because such models exist within an uncertainty calculus of artificial intelligence: The class of strictly positive discrete probability distributions with any prescribed sample space.
- Some definitions of equivalence in chain graphs coincide, which implies that the graphical characterization of Markov distribution equivalence in Frydenberg (1990, Theorem 5.6) also applies to other definitions of equivalence.
- For the parameterization introduced in this paper, all the chain graphs in a class of equivalence have the same dimension and, thus, the factorizations they induce are equally space efficient for storing the strictly positive discrete probability distributions that factorize according to the chain graphs in the class.

We are currently investigating whether the results in this paper can be extended to the class of regular Gaussian distributions. The main problem lies in the derivation of a result analogous to that in Section 3.

Acknowledgments

This work is funded by the Swedish Research Council (ref. VR-621-2005-4202). We thank the Editor and the anonymous Referees for their insightful comments. We are specially grateful to one of the Referees for spotting a flaw in an earlier version of Section 3.

References

Julian Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society B*, 36:192-236, 1974.

Morten Frydenberg. The Chain Graph Markov Property. *Scandinavian Journal of Statistics*, 17:333-353, 1990.

Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

Christopher Meek. Strong Completeness and Faithfulness in Bayesian Networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 411-418, 1995.

Masashi Okamoto. Distinctness of the Eigenvalues of a Quadratic Form in a Multivariate Sample. *The Annals of Statistics*, 1:763-765, 1973.

Jose M. Peña. Approximate Counting of Graphical Models Via MCMC. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, pages 352-359, 2007.

Jose M. Peña, Roland Nilsson, Johan Björkegren and Jesper Tegnér. An Algorithm for Reading Dependencies from the Minimal Undirected Independence Map of a Graphoid

that Satisfies Weak Transitivity. *Journal of Machine Learning Research*, 10:1071-1094, 2009.

Milan Studený. Bayesian Networks from the Point of View of Chain Graphs. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 496-503, 1998.

Milan Studený. *Probabilistic Conditional Independence Structures*. Springer, 2005.

Milan Studený and Remco R. Bouckaert. On Chain Graph Models for Description of Conditional Independence Structures. *The Annals of Statistics*, 26:1434-1495, 1998.

Milan Studený, Alberto Roverato and Šárka Štěpánová. Two Operations of Merging and Splitting Components in a Chain Graph. *Kybernetika*, to appear, 2009. Available at http://staff.utia.cas.cz/studeny/aa23.html.