Faithfulness in Chain Graphs: The Gaussian Case

Jose M. Peña
ADIT, Department of Computer and Information Science
Linköping University, Sweden

Abstract

This paper deals with chain graphs under the classic Lauritzen-Wermuth-Frydenberg interpretation. We prove that almost all the regular Gaussian distributions that factorize with respect to a chain graph are faithful to it. This result has three important consequences. First, chain graphs are more powerful than undirected graphs and acyclic directed graphs for representing regular Gaussian distributions, as some of these distributions can be represented exactly by the former but not by the latter. Second, the moralization and c-separation criteria for reading independencies from a chain graph are complete, in the sense that they identify all the independencies that can be identified from the chain graph alone. Third, some definitions of equivalence in chain graphs coincide and, thus, they have the same graphical characterization.

1 INTRODUCTION

This paper deals with chain graphs (CGs) under the classic Lauritzen-Wermuth-Frydenberg (LWF) interpretation. We prove that almost all the regular Gaussian distributions that factorize with respect to a CG are faithful to it. Previously, it has been proven that for any undirected graph there exists a regular Gaussian distribution that is faithful to it (Lněnička & Matúš, 2007, Corollary 3). A stronger result has been proven for acyclic directed graphs: Almost all the regular Gaussian distributions that factorize with respect to an acyclic directed graph are faithful to it (Spirtes et al., 1993, Theorem 3.2). Therefore, this paper extends the latter result to CGs. It is worth noticing that a result analogous to the one in this paper has been proven in (Peña, 2009, Theorems 3 and 5) but for strictly positive discrete probability distributions with arbitrary prescribed sample space.

There are three important implications of the result proven in this paper. First, there are independence models that can be induced by some CG but that cannot be induced by any undirected graph or acyclic directed graph. As a matter of fact, the experimental results in (Peña, 2007) suggest that this may be the case for the vast majority of independence models induced by CGs. In other words, for most CGs, every undirected graph and acyclic directed graph either represents some separation statement that is not represented by the CG or does not represent some separation statement that is represented by the CG. As Studený (2005, Section 1.1) points out, something that would confirm that this is an advantage of CGs for modeling regular Gaussian distributions would be proving that any independence model induced by a CG is also induced by some regular Gaussian distribution. The result in this paper confirms this point. Second, in the literature, there exist two graphical criteria for identifying independencies holding in a probability distribution p that factorizes with respect to a CG G: The moralization criterion (Lauritzen, 1996) and the c-separation criterion (Studený, 1998). Both criteria are known to be equivalent (Studený, 1998, Lemma 5.1). Furthermore, both criteria are known to be sound, i.e. they only identify independencies in p (Lauritzen, 1996, Theorems 3.34 and 3.36). The result in this paper implies that both criteria are also complete for regular Gaussian distributions: If p is a regular Gaussian distribution, then both criteria identify all the independencies in p that can be identified on the sole basis of G, because there exists a regular Gaussian distribution that is faithful to G. Third, the result in this paper implies that, in the frame of regular Gaussian distributions, the definitions of Markovian distribution equivalent CGs, Markov independence equivalent CGs, and factorization equivalent CGs coincide, which implies that the graphical characterization of Markovian distribution equivalence in

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15. Copyright 2011 by the authors.

(Frydenberg, 1990, Theorem 5.6) also applies to the other definitions of equivalence considered.

It is worth mentioning that there is an alternative to the LWF interpretation of CGs: The Andersson-Madigan-Perlman (AMP) interpretation of CGs (Andersson et al., 2001). The two interpretations are sometimes considered as competing and, thus, their relative merits have been pointed out. For instance, the LWF interpretation is claimed to have a simpler structure of the classes of Markovian distribution equivalent CGs, and a natural representative of each class (the so-called largest CG), which makes this interpretation more amenable to causal discovery in CGs (Roverato and Studený, 2006). On the other hand, the AMP interpretation is claimed to have a simpler separation criterion (Levitz et al., 2001), a simpler characterization of Markov distribution equivalent CGs (Andersson et al., 2001), and simpler parameter estimation within the Gaussian distribution framework (Drton and Eichler, 2006). The reason why this paper deals with the LWF interpretation is not that we advocate its use. The reason is that a result analogous to the one in this paper is proven in (Levitz et al., 2001, Theorem 6.1) under the AMP interpretation, and we want to extend it to the LWF interpretation. In fact, we do not think it is possible to advocate the use of any of the two interpretations without having a specific application in mind to assess their relative merits for it, because no interpretation subsumes the other: There are many independence models that can be induced by some CG under one interpretation but that cannot be induced by any CG under the other interpretation (Andersson et al., 2001, Theorem 6). Then, at the level of generality of this paper, we do not see the two interpretations as competing but as complementary, and any new insight into any of them as relevant.

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 regular Gaussian distributions that factorize with respect to a CG. We present our results on faithfulness in Section 4. In Section 5, we present our results about CG equivalence. Finally, we close with some discussion in Section 6.

2 PRELIMINARIES

In this section, we define some concepts used later in this paper. We first recall some definitions from probabilistic graphical models. See, for instance, (Lauritzen, 1996) and (Studený, 2005) for further information. Let $V = \{1, \ldots, N\}$ be a finite set of size N. The elements of V are not distinguished from singletons and the union of the sets $I_1, \ldots, I_l \subseteq V$ is written as the juxtaposition $I_1 \ldots I_l$. We denote by $|I|$ the size or cardinality of a set $I \subseteq V$, e.g. $|V| = N$. We assume throughout the paper that the union of sets precedes the set difference when evaluating an expression. Unless otherwise stated, all the graphs in this paper are defined over V.

If a graph G contains an undirected (resp. directed) edge between two nodes v_1 and v_2, then we write that $v_1 - v_2$ (resp. $v_1 \rightarrow v_2$) is in G. If $v_1 \rightarrow v_2$ is in G then v_1 is called a parent of v_2. Let $Pa_G(I)$ denote the set of parents in G of the nodes in $I \subseteq V$. When G is evident from the context, we drop the G from $Pa_G(I)$ and use $Pa(I)$ instead. A route from a node v_1 to a node v_l in a graph G is a sequence of nodes $v_1, \ldots, v_l$ such that there exists an edge in G between v_i and v_{i+1} for all $1 \leq i < l$. The length of a route is the number of (not necessarily distinct) edges in the route, e.g. the length of the route $v_1, \ldots, v_l$ is $l - 1$. We treat all singletons as routes of length zero. A path is a route in which the nodes $v_1, \ldots, v_l$ are distinct. A route is called undirected if $v_i - v_{i+1}$ is in G for all $1 \leq i < l$. A route is called descending if $v_i - v_{i+1}$ or $v_i \rightarrow v_{i+1}$ is in G for all $1 \leq i < l$. If there is a descending route from v_1 to v_l in G, then v_1 is called an ancestor of v_l and v_l is called a descendant of v_1. Let $An_G(I)$ denote the set of ancestors in G of the nodes in $I \subseteq V$. A descending route $v_1, \ldots, v_l$ is called a directed pseudocycle if $v_i \rightarrow v_{i+1}$ is in G for some $1 \leq i < l$, and $v_l = v_1$. A chain graph (CG) is a graph (possibly) containing both undirected and directed edges and no directed pseudocycles. An undirected graph (UG) is a CG containing only undirected edges. The underlying UG of a CG is the UG resulting from replacing the directed edges in the CG by undirected edges. A set of nodes of a CG is connected if there exists an undirected route in the CG between every pair of nodes in the set. A connectivity component of a CG is a connected set that is maximal with respect to set inclusion. Hereinafter, we assume that the connectivity components $B_1, \ldots, B_n$ of a CG G are well-ordered, i.e. if $v_1 \rightarrow v_2$ is in G then $v_1 \in B_i$ and $v_2 \in B_j$ 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 connectivity component B_i of G. The subgraph of G induced by $I \subseteq V$, denoted G_I, is the graph over I where two nodes are connected by a (un)directed edge if that edge is in G. A path $v_1, \ldots, v_l$ in G is called a complex if the subgraph of G induced by the set of nodes in the path looks like $v_1 \rightarrow v_2 - \ldots - v_{l-1} \leftarrow v_l$. The path $v_2, \ldots, v_{l-1}$ is called the region of the complex. A section of a route ρ in a CG is a maximal subroute of ρ that only contains undirected edges. A section $v_2 - \ldots - v_{l-1}$ of ρ is a collider section of ρ if $v_1 \rightarrow v_2 - \ldots - v_{l-1} \leftarrow v_l$ is a subroute of ρ. Furthermore, a route ρ in a CG is said to be superactive with respect to $K \subseteq V$ when (i) every collider section of ρ has some node in K, and (ii) every other section of ρ has no node in K.

A set $I \subseteq V$ is complete in an UG G if there is an undirected edge in G between every pair of distinct nodes in I. 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 $X = (X_i)_{i \in V}$ denote a column vector of random variables and X_I ($I \subseteq V$) its subvector $(X_i)_{i \in I}$. We use upper-case letters to denote random variables and the same letters in lower-case to denote their states. Unless otherwise stated, all the probability distributions in this paper are defined on (state space) $\mathbb{R}^N$. Let I, J and K denote three disjoint subsets of V. We denote by $I \perp_p J|K$ that X_I is independent of X_J given X_K in a probability distribution p. Likewise, we denote by $I \perp_G J|K$ that I is separated from J given K in a CG G. Specifically, $I \perp_G J|K$ holds when there is no route in G from a node in I to a node in J that is superactive with respect to K. This is equivalent to say that $I \perp_G J|K$ holds when every path in $(G_{An_G(IJK)})^m$ from a node in I to a node in J has some node in K. The independence model induced by a CG G is the set of separation statements $I \perp_G J|K$. We say that a probability distribution p is Markovian with respect to a CG G when $I \perp_p J|K$ if $I \perp_G J|K$ for all I, J and K disjoint subsets of V. We say that p is faithful to G when $I \perp_p J|K$ iff $I \perp_G J|K$ for all I, J and K disjoint subsets of V. We denote by $I \not\perp_p J|K$ and $I \not\perp_G J|K$ that $I \perp_p J|K$ and $I \perp_G J|K$ do not hold, respectively.

We now recall some results from matrix theory. See, for instance, (Horn and Johnson, 1985) for more information. Let $A = (A_{i,j})_{i,j \in V}$ denote a square matrix. Let $A_{I,J}$ with $I, J \subseteq V$ denote its submatrix $(A_{i,j})_{i \in I, j \in J}$. The determinant of A can recursively be computed, for fixed $i \in V$, as $det(A) = \sum_{j \in V} (-1)^{i+j} A_{i,j} det(A_{\setminus(ij)})$, where $A_{\setminus(ij)}$ denotes the matrix produced by removing the row i and column j from A. If $det(A) \neq 0$ then the inverse of A can be computed as $(A^{-1})_{i,j} = (-1)^{i+j} det(A_{\setminus(ji)})/det(A)$ for all $i, j \in V$. We say that A is strictly diagonally dominant if $abs(A_{i,i}) > \sum_{\{j \in V \,:\, j \neq i\}} abs(A_{i,j})$ for all $i \in V$, where $abs()$ denotes absolute value. A matrix A is Hermitian if it is equal to the matrix resulting from, first, transposing A and, then, replacing each entry by its complex conjugate. Clearly, a real symmetric matrix is Hermitian. A real symmetric $N \times N$ matrix A is positive definite if $x^T A x > 0$ for all non-zero $x \in \mathbb{R}^N$.

Remark 1. *Note that $det(A)$ is a real polynomial in the entries of A, and that $(A^{-1})_{i,j}$ is then the restriction of a fraction of two real polynomials in the entries of A to the area where $det(A)$ is non-zero.*

Finally, we recall some results about Gaussian distributions. We represent a Gaussian distribution as $\mathcal{N}(\mu, \Sigma)$ where μ is its mean vector and Σ its covariance matrix. We say that a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ is regular if Σ is positive definite or, equivalently, invertible. In this paper, we often find more convenient to work with the inverse of the covariance matrix $\Omega = \Sigma^{-1}$, which is also known as the concentration matrix or precision matrix. Since $\Sigma = \Omega^{-1}$, we thus often write $\mathcal{N}(\mu, \Omega^{-1})$ instead of $\mathcal{N}(\mu, \Sigma)$. Let I, J, K and L denote four disjoint subsets of V. Any regular Gaussian distribution p satisfies, among others, the following properties: Symmetry $I \perp_p J|K \Rightarrow J \perp_p I|K$, decomposition $I \perp_p JL|K \Rightarrow I \perp_p J|K$, intersection $I \perp_p J|KL \wedge I \perp_p L|KJ \Rightarrow I \perp_p JL|K$, and weak transitivity $I \perp_p J|K \wedge I \perp_p J|Ku \Rightarrow I \perp_p u|K \vee u \perp_p J|K$ with $u \in V \setminus IJK$.

Let I and J denote two disjoint subsets of V. Let $p(x_{IJ}) = \mathcal{N}(\mu, \Omega^{-1})$ where Ω is positive definite. Then, as shown in (Bishop, 2006, Section 2.3.1), $p(x_J|x_I) = \mathcal{N}(\delta x_I + \gamma, \epsilon^{-1})$ where δ, γ and ϵ are the following real matrices of dimensions, respectively, $|J| \times |I|$, $|J| \times 1$ and $|J| \times |J|$:

$$\delta = -(\Omega_{J,J})^{-1} \Omega_{J,I}, \tag{1}$$

$$\gamma = \mu_J + (\Omega_{J,J})^{-1} \Omega_{J,I} \mu_I \tag{2}$$

and

$$\epsilon = \Omega_{J,J}. \tag{3}$$

Let $p(x_I) = \mathcal{N}(\alpha, \beta^{-1})$ and $q(x_J|x_I) = \mathcal{N}(\delta x_I + \gamma, \epsilon^{-1})$ where δ, γ and ϵ are real matrices of dimensions, respectively, $|J| \times |I|$, $|J| \times 1$ and $|J| \times |J|$, and β and ϵ are positive definite. Then, as shown in (Bishop, 2006, Section 2.3.3), $p(x_I)q(x_J|x_I)$ is a Gaussian distribution $\mathcal{N}(\lambda, \Lambda^{-1})$ over $\begin{pmatrix} x_I \\ x_J \end{pmatrix}$ where

$$\lambda = \begin{pmatrix} \alpha \\ \delta\alpha + \gamma \end{pmatrix} \tag{4}$$

and

$$\Lambda = \begin{pmatrix} \beta + \delta^T \epsilon \delta & -\delta^T \epsilon \\ -\epsilon\delta & \epsilon \end{pmatrix}. \tag{5}$$

Moreover, $p(x_I)q(x_J|x_I)$ is regular because

$$\Lambda^{-1} = \begin{pmatrix} \beta^{-1} & \beta^{-1}\delta^T \\ \delta\beta^{-1} & \epsilon^{-1} + \delta\beta^{-1}\delta^T \end{pmatrix}. \tag{6}$$

3 PARAMETERIZATION

In this section, we describe how we parameterize the regular Gaussian distributions that factorize with respect 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 number of parameters of the regular Gaussian distributions that factorize with respect to the CG at hand.

We say that a regular Gaussian distribution p factorizes with respect to a CG G with connectivity components $B_1, \ldots, B_n$ if the following two conditions are met (Lauritzen, 1996, Proposition 3.30):

F1. $p(x) = \prod_{i=1}^{n} p(x_{B_i} | x_{Pa(B_i)})$ where

F2. $p(x_{B_i Pa(B_i)}) = \prod_{C \in \mathcal{C}((G_{B_i Pa(B_i)})^m)} \psi^i_C(x_C)$ where each $\psi^i_C(x_C)$ is a non-negative real function.

Let $\mathcal{N}(G)$ denote the set of regular Gaussian distributions that factorize with respect to G. We parameterize each probability distribution $p \in \mathcal{N}(G)$ with the following parameters:

- The mean vector μ of p.
- The submatrices $\Omega^i_{B_i,B_i}$ and $\Omega^i_{B_i,Pa(B_i)}$ of the precision matrix Ω^i of $p(x_{B_i Pa(B_i)})$ for all $1 \leq i \leq n$.

We warn the reader that if Ω denotes the precision matrix of p, then Ω^i is not $\Omega_{B_i Pa(B_i), B_i Pa(B_i)}$ but $((\Omega^{-1})_{B_i Pa(B_i), B_i Pa(B_i)})^{-1}$. It is worth mentioning that an alternative parameterization of the probability distributions in $\mathcal{N}(G)$ is presented in (Wermuth, 1992). The main difference between our parameterization and the alternative one is that we parameterize certain concentration matrices whereas they parameterize certain partial concentration matrices. However, both parameterizations are equivalent. We omit the details of the equivalence because they are irrelevant for our purpose. We stick to our parameterization simply because it is more convenient for the calculations performed in this paper.

Note that the values of some of the parameters in the parameterization introduced above are determined by the values of the rest of the parameters. Specifically, for all $1 \leq i \leq n$, the following constraints apply:

C1. $(\Omega^i_{B_i,B_i})_{j,k} = (\Omega^i_{B_i,B_i})_{k,j}$ for all $j, k \in B_i$, because $\Omega^i_{j,k} = \Omega^i_{k,j}$ since Ω^i is symmetric.

C2. $(\Omega^i_{B_i,B_i})_{j,k} = 0$ for all $j, k \in B_i$ such that j and k are not adjacent in G. To see it, note that j and k are not adjacent in $(G_{B_i Pa(B_i)})^m$. Consequently, any path between j and k in $(G_{B_i Pa(B_i)})^m$ must pass through some node in $B_i \setminus jk$ or $Pa(B_i)$. Then, $j \perp_{(G_{B_i Pa(B_i)})^m} k | B_i Pa(B_i) \setminus jk$, which implies $j \perp_{p(x_{B_i Pa(B_i)})} k | B_i Pa(B_i) \setminus jk$ because $p(x_{B_i Pa(B_i)})$ is Markovian with respect to $(G_{B_i Pa(B_i)})^m$ due to the condition F2 above (Lauritzen, 1996, Proposition 3.30, Theorems 3.34 and 3.36). The latter independence statement implies $\Omega^i_{j,k} = 0$ and, thus, $(\Omega^i_{B_i,B_i})_{j,k} = 0$ (Lauritzen, 1996, Proposition 5.2).

C3. $(\Omega^i_{B_i,Pa(B_i)})_{j,k} = 0$ for all $j \in B_i$ and $k \in Pa(B_i)$ such that j and k are not adjacent in G, by a reasoning analogous to the one above.

Hereinafter, the parameters whose values are not determined by the constraints above are called non-determined (nd) parameters. However, the values the nd parameters can take are constrained by the fact that these values must correspond to some probability distribution in $\mathcal{N}(G)$. We prove in Lemma 1 that this is equivalent to requiring that the nd parameters can only take real values such that $\Omega^i_{B_i,B_i}$ is positive definite for all $1 \leq i \leq n$. That is why the set of nd parameter values satisfying this requirement are hereinafter called the nd parameter space for $\mathcal{N}(G)$. We do not work out the inequalities defining the nd parameter space because these are irrelevant for our purpose. The number of nd parameters is what we call the dimension of G, and we denote it as d. Specifically, $d = 2|V| + |G|$ where $|G|$ is the number of edges in G:

- $|V|$ due to μ.
- $|V|$ due to $(\Omega^i_{B_i,B_i})_{j,j}$ for all $1 \leq i \leq n$ and $j \in B_i$.
- $|G|$ due to the entries below the diagonal of $\Omega^i_{B_i,B_i}$ that are not identically zero and the entries of $\Omega^i_{B_i,Pa(B_i)}$ that are not identically zero for all $1 \leq i \leq n$. To see this, recall from the constraints C1-C3 above that there is one entry below the diagonal in some $\Omega^i_{B_i,B_i}$ that is not identically zero for each undirected edge in G, and one entry in some $\Omega^i_{B_i,Pa(B_i)}$ that is not identically zero for each directed edge in G.

Lemma 1. *Let G be a CG. There is a one-to-one correspondence between the probability distributions in $\mathcal{N}(G)$ and the elements of the nd parameter space for $\mathcal{N}(G)$.*

Proof. We first prove that the mapping of probability distributions into nd parameter values is injective. Obviously, any probability distribution in $p \in \mathcal{N}(G)$ is mapped into some real values of the nd parameters μ, $\Omega^i_{B_i,B_i}$ and $\Omega^i_{B_i,Pa(B_i)}$ for all $1 \leq i \leq n$. In particular, $\Omega^i_{B_i,B_i}$ takes value $(((\Omega^{-1})_{B_i Pa(B_i), B_i Pa(B_i)})^{-1})_{B_i,B_i}$ where Ω is the precision matrix of p. Then, that $\Omega^i_{B_i,B_i}$ is positive definite follows from the fact that Ω is positive definite (Studený, 2005, p. 237). Thus, p is mapped into some element of the nd parameter space for $\mathcal{N}(G)$.

Moreover, different probability distributions are mapped into different elements. To see it, assume to the contrary that there exist two distinct probability distributions $p, p' \in \mathcal{N}(G)$ that are mapped into the same element. Note that this element uniquely identifies $p(x_{B_i}|x_{Pa(B_i)})$ by Equations 1-3 for all $1 \leq i \leq n$, where $I = Pa(B_i)$ and $J = B_i$. 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 the condition F1 above.

We now prove in three steps that the mapping of nd parameter values into probability distributions is injective.

Step 1 We first show that any element of the nd parameter space for $\mathcal{N}(G)$ is mapped into some regular Gaussian distribution q. Note that any element of the nd parameter space for $\mathcal{N}(G)$ uniquely identifies a Gaussian distribution $q^i(x_{B_i}|x_{Pa(B_i)})$ for all $1 \leq i \leq n$ by Equations 1-3, where $I = Pa(B_i)$ and $J = B_i$. Specifically, $q^i(x_{B_i}|x_{Pa(B_i)}) = \mathcal{N}(\delta^i x_{Pa(B_i)} + \gamma^i, (\epsilon^i)^{-1})$ where

$$\delta^i = -(\Omega^i_{B_i,B_i})^{-1}\Omega^i_{B_i,Pa(B_i)}, \tag{7}$$

$$\gamma^i = \mu_{B_i} + (\Omega^i_{B_i,B_i})^{-1}\Omega^i_{B_i,Pa(B_i)}\mu_{Pa(B_i)} \tag{8}$$

and

$$\epsilon^i = \Omega^i_{B_i,B_i}. \tag{9}$$

In the equations above, we have assumed that the values of all the entries of $\Omega^i_{B_i,B_i}$ and $\Omega^i_{B_i,Pa(B_i)}$ have previously been determined from the element of the nd parameter space at hand and the constraints C1-C3 above. Furthermore, note that $q^i(x_{B_i}|x_{Pa(B_i)})$ is regular because, by definition, $\Omega^i_{B_i,B_i}$ is positive definite. Clearly, $q^i(x_{B_i}|x_{Pa(B_i)})$ can be rewritten as a regular Gaussian distribution $r^i(x_{B_i}|x_{B_1\ldots B_{i-1}})$: It suffices to take $r^i(x_{B_i}|x_{B_1\ldots B_{i-1}})$ equal to

$$\mathcal{N}((\delta^i, \mathbf{0})\begin{pmatrix} x_{Pa(B_i)} \\ x_{B_1\ldots B_{i-1}\setminus Pa(B_i)} \end{pmatrix} + \gamma^i, (\epsilon^i)^{-1})$$

where $\mathbf{0}$ is a matrix of zeroes of dimension $|B_i| \times |B_1 \ldots B_{i-1} \setminus Pa(B_i)|$. Then, $r^1(x_{B_1})r^2(x_{B_2}|x_{B_1})$ is a regular Gaussian distribution by Equations 4-6. Likewise, $r^1(x_{B_1})r^2(x_{B_2}|x_{B_1})r^3(x_{B_3}|x_{B_1B_2})$ is a regular Gaussian distribution. Continuing with this process for the rest of connectivity components proves that $\prod_{i=1}^n q^i(x_{B_i}|x_{Pa(B_i)}) = \prod_{i=1}^n r^i(x_{B_i}|x_{B_1\ldots B_{i-1}})$ is mapped into some regular Gaussian distribution q.

Step 2 We now show that $q \in \mathcal{N}(G)$. Note that for all $1 \leq i < n$ and any fixed value of $x_{B_1\ldots B_i}$

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

$$= \int q^{i+1}(x_{B_{i+1}}|x_{Pa(B_{i+1})})[\int q^{i+2}(x_{B_{i+2}}|x_{Pa(B_{i+2})})[\ldots$$

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

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

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

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

$$q^i(x_{B_i}|x_{Pa(B_i)}) \int \prod_{l=1}^{i-1} q^l(x_{B_l}|x_{Pa(B_l)})dx_{B_1\ldots B_{i-1}\setminus Pa(B_i)}. \tag{10}$$

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

$$q(x_{Pa(B_i)}) = \int q(x_{B_iPa(B_i)})dx_{B_i}$$

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

$$= \int [\int \prod_{l=1}^{i} q^l(x_{B_l}|x_{Pa(B_l)})dx_{B_i}]dx_{B_1\ldots B_{i-1}\setminus Pa(B_i)} \tag{11}$$

$$= \int \prod_{l=1}^{i-1} q^l(x_{B_l}|x_{Pa(B_l)})dx_{B_1\ldots B_{i-1}\setminus Pa(B_i)}. \tag{12}$$

Note the use of Fubini's theorem to change the order of integration and produce Equation 11. Consequently, for all $1 \leq i \leq n$

$$q(x_{B_i}|x_{Pa(B_i)}) = \frac{q(x_{B_iPa(B_i)})}{q(x_{Pa(B_i)})} = q^i(x_{B_i}|x_{Pa(B_i)}) \tag{13}$$

due to Equations 10 and 12. 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 the condition F1 above. Moreover, $q(x_{B_iPa(B_i)})$ satisfies the condition F2 for all $1 \leq i \leq n$. We show this by induction on i. Let Λ^i denote the precision matrix of $q(x_{B_iPa(B_i)})$, and note that

$$q(x_{B_iPa(B_i)}) = q^i(x_{B_i}|x_{Pa(B_i)})q(x_{Pa(B_i)})$$

by Equation 13. So, Λ^i can be calculated from $q^i(x_{B_i}|x_{Pa(B_i)})$ and $q(x_{Pa(B_i)})$ via Equation 5. Specifically, it follows from Equations 5 and 9, respectively 7, that

$$\Lambda^i_{B_i,B_i} = \epsilon^i = \Omega^i_{B_i,B_i} \tag{14}$$

and

$$\Lambda^i_{B_i,Pa(B_i)} = -\epsilon^i\delta^i = -\Omega^i_{B_i,B_i}[-(\Omega^i_{B_i,B_i})^{-1}\Omega^i_{B_i,Pa(B_i)}] = \Omega^i_{B_i,Pa(B_i)}. \quad (15)$$

Consequently, due to the constraints C2 and C3 above, $\Lambda^i_{j,k} = 0$ for all $j, k \in B_iPa(B_i)$ such that j and k are not adjacent in $(G_{B_iPa(B_i)})^m$. Moreover, $\Lambda^i_{j,k} = 0$ is equivalent to $j \perp_{q(x_{B_iPa(B_i)})} k | B_iPa(B_i) \setminus jk$ (Lauritzen, 1996, Proposition 5.2). This implies that $q(x_{B_iPa(B_i)})$ factorizes with respect to $(G_{B_iPa(B_i)})^m$ and, thus, that it satisfies the condition F2 above (Lauritzen, 1996, Proposition 3.30, Theorems 3.34 and 3.36). Consequently, $q \in \mathcal{N}(G)$.

Step 3 We finally show that different elements of the nd parameter space for $\mathcal{N}(G)$ are mapped into different probability distributions in $\mathcal{N}(G)$. Assume to the contrary that two distinct elements of the nd parameter space for $\mathcal{N}(G)$ are mapped into the same probability distribution $q \in \mathcal{N}(G)$. Assume that the two elements differ in the value for μ_{B_i}, $\Omega^i_{B_i,B_i}$ or $\Omega^i_{B_i,Pa(B_i)}$ but that they coincide in the values for μ_{B_l}, $\Omega^l_{B_l,B_l}$ and $\Omega^l_{B_l,Pa(B_l)}$ for all $1 \leq l < i$. There are two scenarios to consider:

- If the two elements differ in the value for $\Omega^i_{B_i,B_i}$ or $\Omega^i_{B_i,Pa(B_i)}$, then they are mapped into two different $q(x_{B_iPa(B_i)})$ by Equations 14 and 15, because two regular Gaussian distributions with different precision matrices are different. However, this contradicts the assumption that the two elements are mapped into the same q.
- If the two elements differ in the value for μ_{B_i} but they do not differ in the values for $\Omega^i_{B_i,B_i}$ and $\Omega^i_{B_i,Pa(B_i)}$, then the two elements do not differ in the value for $\mu_{Pa(B_i)}$ either, because $Pa(B_i) \subseteq B_1 \ldots B_{i-1}$ and we assumed above that the two elements coincide in the values for μ_{B_l} for all $1 \leq l < i$. Then, the two elements are mapped into the same δ^i but different γ^i in Equations 7 and 8. That is, the two elements are mapped into two different $q^i(x_{B_i}|x_{Pa(B_i)})$ and, thus, to two different $q(x_{B_i}|x_{Pa(B_i)})$ by Equation 13. However, this contradicts the assumption that the two elements are mapped into the same q.

□

Remark 2. *Note the following three observations:*

- *For all $1 \leq i \leq n$, according to the constraints C1-C3 above, every entry of $\Omega^i_{B_i,B_i}$ and $\Omega^i_{B_i,Pa(B_i)}$ is equal either to zero or to some nd parameter in the parameterization of the probability distributions in $\mathcal{N}(G)$.*
- *For all $1 \leq i \leq n$, by Remark 1, every entry of $(\Omega^i_{B_i,B_i})^{-1}$ is a fraction of real polynomials in the entries of $\Omega^i_{B_i,B_i}$ and, thus, a fraction of real polynomials in the nd parameters in the parameterization of the probability distributions in $\mathcal{N}(G)$. Thus, every entry of the matrices δ^i and ϵ^i in Equations 7 and 9 is also a fraction of real polynomials in the referred nd parameters.*
- *Every entry of the precision matrix of $r^1(x_{B_1})r^2(x_{B_2}|x_{B_1})$ in the proof above is, by Equation 5, a real polynomial in the entries of δ^2, ϵ^2 and the precision matrix of $r^1(x_{B_1})$, i.e. ϵ^1. Likewise, every entry of the precision matrix of $r^1(x_{B_1})r^2(x_{B_2}|x_{B_1})r^3(x_{B_3}|x_{B_1B_2})$ in the proof above is a real polynomial in the entries of δ^3, ϵ^3 and the precision matrix of $r^1(x_{B_1})r^2(x_{B_2}|x_{B_1})$, that is, a real polynomial in the entries of δ^3, ϵ^3, δ^2, ϵ^2 and ϵ^1. Continuing with this process for the rest of connectivity components shows that every entry of the precision matrix of $q(x) = \prod_{i=1}^n r^i(x_{B_i}|x_{B_1 \ldots B_{i-1}})$ in the proof above is a real polynomial in the entries of the matrices ϵ^1, and δ^i and ϵ^i for all $1 < i \leq n$.*

It follows from the observations above that every entry of the precision matrix of q in the proof above is a fraction of real polynomials in the nd parameters in the parameterization of the probability distributions in $\mathcal{N}(G)$. Consequently, by Remark 1, every entry of the covariance matrix of q is a fraction of real polynomials in the nd parameters in the parameterization of the probability distributions in $\mathcal{N}(G)$. Moreover, note the following two observations on the latter fractions:

- *Each of these fractions is defined on the whole nd parameter space for $\mathcal{N}(G)$: The polynomial in the denominator of the fraction is non-vanishing in the nd parameter space for $\mathcal{N}(G)$ because, as we have proven in Step 1 in the theorem above, q is a Gaussian distribution.*
- *Within the nd parameter space for $\mathcal{N}(G)$, each of these fractions vanishes only in the points where the polynomial in the numerator of the fraction vanishes because, as we have just seen, the denominator of the fraction is non-vanishing in the nd parameter space for $\mathcal{N}(G)$.*

4 FAITHFULNESS

The two theorems in this section are the main contribution of this manuscript. Together with the lemma in the previous section, they imply that almost all the regular Gaussian distributions that factorize with respect to a CG are faithful to it.

Theorem 1. *Let G be a CG of dimension d. The nd parameter space for $\mathcal{N}(G)$ has positive Lebesgue measure with respect to $\mathbb{R}^d$.*

Proof. Since we do not know a closed-form expression of the nd parameter space for $\mathcal{N}(G)$, we take an indirect approach to prove the lemma. Recall that, by definition, the nd parameter space for $\mathcal{N}(G)$ is the set of real values such that, after the extension determined by the constraints C1 and C2, $\Omega^i_{B_i,B_i}$ is positive definite for all $1 \leq i \leq n$. Therefore, all the nd parameters except those in $\Omega^i_{B_i,B_i}$ for all $1 \leq i \leq n$ can take values independently of the rest of the nd parameters. The nd parameters in $\Omega^i_{B_i,B_i}$ cannot take values independently one of another because, otherwise, $\Omega^i_{B_i,B_i}$ may not be positive definite. However, if the entries in the diagonal of $\Omega^i_{B_i,B_i}$ take values in $(|B_i|-1,\infty)$ and the rest of the nd parameters in $\Omega^i_{B_i,B_i}$ take values in $[-1,1]$, then the nd parameters in $\Omega^i_{B_i,B_i}$ can take values independently one of another. To see it, note that in this case $\Omega^i_{B_i,B_i}$ will always be Hermitian, strictly diagonally dominant, and with strictly positive diagonal entries, which implies that $\Omega^i_{B_i,B_i}$ will always be positive definite (Horn and Johnson, 1985, Corollary 7.2.3).

The subset of the nd parameter space of $\mathcal{N}(G)$ described in the paragraph above has positive volume in $\mathbb{R}^d$ and, thus, it has positive Lebesgue measure with respect to $\mathbb{R}^d$. Then, the nd parameter space of $\mathcal{N}(G)$ has positive Lebesgue measure with respect to $\mathbb{R}^d$. □

Before proving the second theorem, we introduce two auxiliary lemmas whose proofs are given in the appendix.

Lemma 2. *Let G be a CG. For every $i,j \in V$ and $Z \subseteq V \setminus ij$, there exists a real polynomial $S(i,j,Z)$ in the nd parameters in the parameterization of the probability distributions in $\mathcal{N}(G)$ such that, for every $p \in \mathcal{N}(G)$, $i \perp_p j|Z$ iff $S(i,j,Z)$ vanishes for the nd parameter values coding p.*

We interpret the polynomial in the lemma above as a real function on a real Euclidean space that includes the nd parameter space for $\mathcal{N}(G)$. We say that the polynomial in the lemma above is non-trivial if not all the values of the nd parameters are solutions to the polynomial. This is equivalent to the requirement that the polynomial is not identically zero.

Lemma 3. *Let G be a CG such that $i \not\perp_G j|Z$, where $i,j \in V$ and $Z \subseteq V \setminus ij$. Then, there exists a probability distribution $p \in \mathcal{N}(G)$ such that $i \not\perp_p j|Z$.*

Theorem 2. *Let G be a CG of dimension d. The subset of the nd parameter space for $\mathcal{N}(G)$ that corresponds to the probability distributions in $\mathcal{N}(G)$ that are not faithful to G has zero Lebesgue measure with respect to $\mathbb{R}^d$.*

Proof. Note that the probability distributions in $\mathcal{N}(G)$ are Markovian with respect to G (Lauritzen, 1996, Proposition 3.30, Theorems 3.34 and 3.36). Then, for any probability distribution $p \in \mathcal{N}(G)$ 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 V, here denoted as I, J and Z, such that $I \not\perp_G J|Z$ but $I \perp_p J|Z$. However, if $I \not\perp_G J|Z$ then $i \not\perp_G j|Z$ for some $i \in I$ and $j \in J$. Furthermore, if $I \perp_p J|Z$ then $i \perp_p j|Z$ by symmetry and decomposition. By Lemma 2, there exists a real polynomial $S(i,j,Z)$ in the nd parameters in the parameterization of the probability distributions in $\mathcal{N}(G)$ such that, for every $q \in \mathcal{N}(G)$, $i \perp_q j|Z$ iff $S(i,j,Z)$ vanishes for the nd parameter values coding q. Furthermore, $S(i,j,Z)$ is non-trivial by Lemma 3. Let $sol(i,j,Z)$ denote the set of solutions to the polynomial $S(i,j,Z)$. Then, $sol(i,j,Z)$ has zero Lebesgue measure with respect to $\mathbb{R}^d$ because it consists of the solutions to a non-trivial real polynomial in real variables (the nd parameters) (Okamoto, 1973). Then, $sol = \bigcup_{\{I,J,Z \subseteq V \text{ disjoint} : I \not\perp_G J|Z\}} \bigcup_{\{i \in I, j \in J : i \not\perp_G j|Z\}} sol(i,j,Z)$ has zero Lebesgue measure with respect to $\mathbb{R}^d$, because the finite union of sets of zero Lebesgue measure has zero Lebesgue measure too. Consequently, the subset of the nd parameter space for $\mathcal{N}(G)$ that corresponds to the probability distributions in $\mathcal{N}(G)$ that are not faithful to G has zero Lebesgue measure with respect to $\mathbb{R}^d$ because it is contained in sol. □

In summary, it follows from Theorems 1 and 2 that, in the measure-theoretic sense described, almost all the elements of the nd parameter space for $\mathcal{N}(G)$ correspond to probability distributions in $\mathcal{N}(G)$ that are faithful to G. Since this correspondence is one-to-one by Lemma 1, it follows that almost all the regular Gaussian distributions in $\mathcal{N}(G)$ are faithful to G.

5 EQUIVALENCE

In this section, we prove with the help of the theorems in the previous section that some definitions of equivalent CGs coincide. Recall that, unless otherwise stated, all the probability distributions in this paper are defined on (state space) $\mathbb{R}^N$, where $|V| = N$. We say that two CGs are Markov independence equivalent if they induce the same independence model. We say that two CGs are Gaussian Markovian distribution equivalent if every regular Gaussian distribution is Markovian with respect to both CGs or with respect to neither of them. We say that two CGs G and H are Gaussian factorization equivalent if $\mathcal{N}(G) = \mathcal{N}(H)$.

Corollary 1. *Let G and H denote two CGs. The following statements are equivalent:*

1. *G and H are Gaussian factorization equivalent.*
2. *G and H are Gaussian Markovian distribution equivalent.*
3. *G and H are Markov independence equivalent.*

Proof. The equivalence of Statements 1 and 2 follows from (Lauritzen, 1996, Proposition 3.30, Theorems 3.34 and 3.36). We now prove that Statements 2 and 3 are equivalent. By definition, Markov independence equivalence implies Gaussian Markovian distribution equivalence. 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 $I \perp_G J|K$ that is not represented by H. Consider a probability distribution $p \in \mathcal{N}(H)$ faithful to H. Such a probability distribution exists by Section 4, and it is Markovian with respect to H. However, p cannot be Markovian with respect to G, because $I \not\perp_H J|K$ implies $I \not\perp_p J|K$. □

The importance of the previous corollary lies in the fact that Frydenberg (1990, Theorem 5.6) gives a straightforward graphical characterization of Gaussian Markovian distribution equivalence: Two CGs are Gaussian Markovian distribution equivalent iff they have the same underlying UG and the same complexes. Due to the corollary above, that is also a graphical characterization of the other two types of equivalence discussed 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.

Finally, we prove that all equivalent CGs have the same dimension with respect to the parameterization introduced in Section 3. This result disproves the following conjecture. 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, for discrete probability distributions, the LCG in a class of equivalent CGs has fewer nd parameters than any other CG in the class. This would imply that the most space efficient way of storing the discrete probability distributions that factorize with respect to a class of equivalent CGs is by factorizing them with respect to the LCG in the class rather than with respect to any other CG in the class. The following corollary implies that an analogous conjecture for regular Gaussian distributions and the parameterization of them proposed in Section 3 would be false.

Corollary 2. *All equivalent CGs have the same dimension with respect to the parameterization proposed in Section 3.*

Proof. Let G denote the LCG in a class of equivalent CGs. Let H denote any other CG in the class. Recall that the dimensions of G and H with respect to the parameterization proposed in Section 3 are, respectively, $2|V|+|G|$ and $2|V|+|H|$. Note that H can be obtained from G by orienting some of the undirected edges in G (Volf & Studený, 1999, Theorem 3.9). Then, $|H| = |G|$ and, thus, $2|V| + |G| = 2|V| + |H|$. □

6 CONCLUSIONS

In this paper, we have proven that almost all the regular Gaussian distributions that factorize with respect to a CG are faithful to it. This result extends the results in (Spirtes et al., 1993, Theorem 3.2) and (Lněnička & Matúš, 2007, Corollary 3). There are four consequences that follow from the result proven in this paper. First, the experimental results in (Peña, 2007) suggest that the vast majority of independence models that can be induced by CGs cannot be induced by undirected graphs or acyclic directed graphs. This is an advantage of CGs when dealing with regular Gaussian distributions, because there exists a regular Gaussian distribution that is faithful to each of these independence models. Second, the moralization and c-separation criteria for reading independencies holding in the regular Gaussian distributions that factorize with respect to a CG are complete (i.e. they identify all the independencies that can be identified on the sole basis of the CG), because there exists a regular Gaussian distribution that is faithful to the CG. Third, the definitions of Gaussian Markovian distribution equivalent CGs, Markov independence equivalent CGs, and Gaussian factorization equivalent CGs coincide, which implies that the graphical characterization of Gaussian Markovian distribution equivalence in (Frydenberg, 1990, Theorem 5.6) also applies to the other definitions of equivalence considered. Four, for the parameterization introduced in this paper, all the CGs in a class of equivalence have the same dimension and, thus, their factorizations are equally space efficient for storing the regular Gaussian distribution that factorize with respect to the CGs in the class.

Appendix: Proofs of Lemmas 2 and 3

Proof of Lemma 2. Let Σ denote the covariance matrix of p. Note that $i \perp_p j|Z$ iff $((\Sigma_{ijZ,ijZ})^{-1})_{i,j} = 0$ (Lauritzen, 1996, Proposition 5.2). Recall that $((\Sigma_{ijZ,ijZ})^{-1})_{i,j} = (-1)^{\alpha} det(\Sigma_{iZ,jZ})/det(\Sigma_{ijZ,ijZ})$ with $\alpha \in \{0,1\}$. Note that $det(\Sigma_{ijZ,ijZ}) > 0$ because $\Sigma_{ijZ,ijZ}$ is positive definite (Studený, 2005, p. 237). Then, $i \perp_p j|Z$ iff $det(\Sigma_{iZ,jZ}) = 0$. Thus, $i \perp_p j|Z$ iff a real polynomial $R(i,j,Z)$ in the entries of Σ vanishes due to Remark 1. However, note that it follows from Lemma 1 and Remark 2 that each entry of Σ is a fraction of real polynomials in the nd parameters in the parameterization of the probability distributions in $\mathcal{N}(G)$. Recall also from Remark 2 that the polynomial in the denominator of each of these fractions is non-vanishing in the nd parameter space for $\mathcal{N}(G)$. Therefore, by simple algebraic manipulation, the polynomial $R(i,j,Z)$ can be expressed as a fraction $S(i,j,Z)/T(i,j,Z)$ of real polynomials in the nd parameters where $T(i,j,Z)$ is non-vanishing in the nd parameter space for $\mathcal{N}(G)$. Consequently, $i \perp_p j|Z$ iff the real polynomial $S(i,j,Z)$ in the nd parameters vanishes for the values coding p. □

Before proving Lemma 3, some auxiliary lemmas are proven.

Lemma 4. *Let G and H be two CGs such that the undirected (resp. directed) edges in H are a subset of the undirected (resp. directed) edges in G. Then, $\mathcal{N}(H) \subseteq \mathcal{N}(G)$.*

Proof. Note that a regular Gaussian distribution factorizes with respect to a CG iff it is Markovian with respect to the CG (Lauritzen, 1996, Proposition 3.30, Theorems 3.34 and 3.36). Then, $\mathcal{N}(H) \subseteq \mathcal{N}(G)$ because the independence model induced by H is a superset of that induced by G. □

Lemma 5. *Let G be a CG such that*

1. *G has a route between the nodes i and j that has no collider section, and*
2. *the route has no node in $Z \subseteq V \setminus ij$.*

Then, there exists a probability distribution $p \in \mathcal{N}(G)$ such that $i \not\perp_p j|Z$.

Proof. The route in the lemma can be converted into a path ρ between i and j in G as follows: Iteratively, remove from the route any subroute between a node and itself. Note that none of these removals produces a collider section: It suffices to note that if the route after the removal has a collider section, then the route before the removal must have a collider section, which is a contradiction. Consequently, ρ is a path between i and j in G that has no collider section. Therefore, ρ is superactive with respect to Z: Since the route in the lemma has no node in Z, ρ has no node in Z either. Now, remove from G all the edges that are not in ρ, and call the resulting CG H. Note that H has no complex since ρ has no collider section. Drop the direction of every edge in H and call the resulting UG L. Now, note that there exists a regular Gaussian distribution p that is faithful to L (Lněnička & Matúš, 2007, Corollary 3) and, thus, $i \not\perp_p j|Z$ because $i \not\perp_L j|Z$. Note also that the fact that p is faithful to L implies that p is Markovian with respect to L which, in turn, implies that p is also Markovian with respect to H, because H and L have the same underlying UG and complexes (Frydenberg, 1990, Theorem 5.6). Consequently, $p \in \mathcal{N}(H)$ (Lauritzen, 1996, Proposition 3.30, Theorems 3.34 and 3.36) and, thus, $p \in \mathcal{N}(G)$ because $\mathcal{N}(H) \subseteq \mathcal{N}(G)$ by Lemma 4. □

Let ν denote an undirected route $v_2 - \ldots - v_{l-1}$ in a CG. Hereinafter, we denote by $v_1 \rightarrow \nu \leftarrow v_l$ the route $v_1 \rightarrow v_2 - \ldots - v_{l-1} \leftarrow v_l$.

Lemma 6. *Let G be a CG such that*

1. *G has a route $i \rightarrow \nu \leftarrow j$ where $i, j \in V$ and ν is an undirected route, and*
2. *some node in ν is in Z or has a descendant in Z, where $Z \subseteq V \setminus ij$.*

Then, there exists a probability distribution $p \in \mathcal{N}(G)$ such that $i \not\perp_p j|Z$.

Proof. The route ν can be converted into a path ϑ in G as follows: Iteratively, remove from ν any subroute between a node and itself. Note that ν does not contain either i or j because, otherwise, G would have a directed pseudocycle between i and itself or between j and itself, which is a contradiction. Therefore, ϑ does not contain either i or j and, thus, $i \rightarrow \vartheta \leftarrow j$ is a path in G. Note that the subroutes removed from ν contain only undirected edges. Therefore, every node that is in ν but not in ϑ is a descendant of some node in ϑ. Consequently, some node in ϑ is in Z or has a descendant in Z, due to the assumptions in the lemma.

We first prove the lemma for the case where some node in ϑ is in Z. Remove from G all the edges that are not in $i \rightarrow \vartheta \leftarrow j$, and call the resulting CG H. Note that $i \rightarrow \vartheta \leftarrow j$ is a complex in H and, thus, that $i \perp_H j$. Let k denote the closest node to i that is in ϑ and in Z.

We prove in this paragraph that there exists a probability distribution $p \in \mathcal{N}(H)$ such that $i \not\perp_p jZ \setminus k|k$.

By Lemma 2, there exists a real polynomial $S(i,k,\emptyset)$ in the nd parameters in the parameterization of the probability distributions in $\mathcal{N}(H)$ such that, for every $q \in \mathcal{N}(H)$, $i \perp_q k$ iff $S(i,k,\emptyset)$ vanishes for the nd parameter values coding q. Furthermore, $S(i,k,\emptyset)$ is non-trivial. To see this, remove from H all the edges outside the path between i and k, and call the resulting CG L. Note that $\mathcal{N}(L) \subseteq \mathcal{N}(H)$ by Lemma 4. Now note that, by Lemma 5, there exists a probability distribution $r \in \mathcal{N}(L)$ such that $i \not\perp_r k$. By an analogous reasoning, we can conclude that there exists a non-trivial real polynomial $S(j,k,\emptyset)$ in the nd parameters in the parameterization of the probability distributions in $\mathcal{N}(H)$ such that, for every $q \in \mathcal{N}(H)$, $j \perp_q k$ iff $S(j,k,\emptyset)$ vanishes for the nd parameter values coding q. Let $sol(i,k,\emptyset)$ and $sol(j,k,\emptyset)$ denote the sets of solutions to the polynomials $S(i,k,\emptyset)$ and $S(j,k,\emptyset)$, respectively. Let d denote the dimension of H. Then, $sol(i,k,\emptyset)$ and $sol(j,k,\emptyset)$ have both zero Lebesgue measure with respect to $\mathbb{R}^d$ because they consist of the solutions to non-trivial real polynomials in real variables (the nd parameters) (Okamoto, 1973). Then, $sol = sol(i,k,\emptyset) \cup sol(j,k,\emptyset)$ also has zero Lebesgue measure with respect to $\mathbb{R}^d$, because the finite union of sets of zero Lebesgue measure has zero Lebesgue measure too. Consequently, the probability distributions $q \in \mathcal{N}(H)$ such that $i \perp_q k$ or $j \perp_q k$ correspond to a set of elements of the nd parameter space for $\mathcal{N}(H)$ that has zero Lebesgue measure with respect to $\mathbb{R}^d$ because it is contained in sol. Since this correspondence is one-to-one by Lemma 1, Theorem 1 implies that there exists a probability distribution $p \in \mathcal{N}(H)$ such that $i \not\perp_p k$ and $j \not\perp_p k$. Furthermore, as shown above $i \perp_H j$ and, thus, $i \perp_p j$ because p is Markovian with respect to H, since $p \in \mathcal{N}(H)$ (Lauritzen, 1996, Proposition 3.30, Theorems 3.34 and 3.36). Then, $i \not\perp_p j | k$ by symmetry and weak transitivity and, thus, $i \not\perp_p jZ \setminus k | k$ by decomposition.

Finally, recall that since k is the closest node to i that is in ϑ and in Z, then $i \perp_H Z \setminus k | jk$ and thus $i \perp_p Z \setminus k | jk$ because p is Markovian with respect to H. Then, $i \not\perp_p j | Z$ by intersection on $i \not\perp_p jZ \setminus k | k$ and $i \perp_p Z \setminus k | jk$. Consequently, we have proven that there exists a probability distribution $p \in \mathcal{N}(H)$ such that $i \not\perp_p j | Z$. Moreover, $p \in \mathcal{N}(G)$ because $\mathcal{N}(H) \subseteq \mathcal{N}(G)$ by Lemma 4.

We now prove the lemma for the case where no node in ϑ is in Z but some node in ϑ has a descendant in Z. Consider the shortest descending path between a node in ϑ and a node in Z. Let l and k denote the initial and final nodes of the path, i.e. $k \in Z$. Remove from G all the edges that are not in $i \rightarrow \vartheta \leftarrow j$ or in the path between l and k, and call the resulting CG H. Note that $i \rightarrow \vartheta \leftarrow j$ is a complex in H and, thus, that $i \perp_H j$. Therefore, we can follow the same steps as above to prove that there exists a probability distribution $p \in \mathcal{N}(H)$ such that $i \not\perp_p jZ \setminus k | k$. Finally, recall that there is no path between i and any node in $Z \setminus k$ in H, then $i \perp_H Z \setminus k | jk$ and thus $i \perp_p Z \setminus k | jk$ because p is Markovian with respect to H. Then, $i \not\perp_p j | Z$ by intersection on $i \not\perp_p jZ \setminus k | k$ and $i \perp_p Z \setminus k | jk$. Consequently, we have proven that there exists a probability distribution $p \in \mathcal{N}(H)$ such that $i \not\perp_p j | Z$. Moreover, $p \in \mathcal{N}(G)$ because $\mathcal{N}(H) \subseteq \mathcal{N}(G)$ by Lemma 4. □

Proof of Lemma 3. We prove the lemma in two steps. In the first step, we introduce some notation that we use in the second step, the actual proof of the lemma.

Step 1 Given a route ρ in a CG H, we define H_ρ as the CG resulting from removing from H all the edges that are not in ρ. We define the level of a node in H as the index of the connectivity component the node belongs to. We define the dlength of a route as the number of distinct edges in the route. Note the difference between the dlength and the length of a route: The former counts edges without repetition and the latter with repetition (recall Section 2). We say that a route is dshorter than another route if the former has smaller dlength than the latter. Likewise, we say that a route is dshortest if no other route is dshorter than it. Let $\mathfrak{N}$ denote any total order of the nodes in the CG H. Let $\mathfrak{R}$ denote any total order of all the routes between two nodes in H. Finally, if $a \not\perp_H b | C$ where $a, b \in V$ and $C \subseteq V \setminus ab$, then we define $splits(a,b,C,H)$ as follows:

S1. If there is a route in H like that in Lemma 5 or 6 for $i = a$, $j = b$ and $Z = C$, then we define $splits(a,b,C,H) = 0$.

S2. Otherwise, we define recursively $splits(a,b,C,H) = splits(a,k,C,H_\rho) + splits(b,k,C,H_\rho) + 1$, where ρ and k are selected as follows. Let Ψ denote the set of routes between a and b in H that are superactive with respect to C. Let Φ denote the dshortest routes in Ψ. Let Υ denote the shortest routes in Φ. Let ρ denote the route in Υ that comes first in $\mathfrak{R}$. We call ρ the splitting route. Furthermore, let K denote the set of nodes in ρ but not in Cab that have minimal level in H_ρ. Let k denote the node in K that comes first in $\mathfrak{N}$. Note that the only point with $\mathfrak{R}$ and $\mathfrak{N}$ is to select ρ and k unambiguously.

Note that we have implicitly assumed in the definition S2 that K is non-empty. We now prove that this is always true. Assume to the contrary that K is empty. This means that all the nodes in ρ are in Cab. Since the definition S1 did not apply, ρ must have some collider section ν. Moreover, $a = v_1 \rightarrow \nu \leftarrow v_l = b$ is a subroute of ρ: If $v_1 \notin \{a, b\}$ (resp. $v_l \notin \{a, b\}$) then v_1 (resp. v_l) must be outside C for ρ to be superactive with respect to C, which contradicts the assumption that all the nodes in ρ are in Cab. Moreover, some node in ν must be in C for ρ to be superactive with respect to C. However, this implies that $a \rightarrow \nu \leftarrow b$ is a route that satisfies the requirements of the definition S1, which is a contradiction.

Finally, we prove that $splits(a, k, C, H_\rho)$ and $splits(b, k, C, H_\rho)$ in the definition S2 are well-defined. Let ϱ denote the subroute of ρ between the first occurrences of a and k in ρ when going from a to b. Note that if ρ contains k only in non-collider sections, then none of the other nodes in those sections can be in C for ρ to be superactive with respect to C and, thus, ϱ is a route between a and k in H_ρ that is superactive with respect to C and, thus, $a \not\perp_{H_\rho} k|C$ and, thus, $splits(a, k, C, H_\rho)$ is defined. We now prove that ρ contains k only in non-collider sections. Assume the contrary and let ν denote any collider section of ρ that contains k. Note that $a = v_1 \rightarrow \nu \leftarrow v_l = b$ is a subroute of ρ, because if $v_1 \notin \{a, b\}$ or $v_l \notin \{a, b\}$ then there exists a node in ρ but not in Cab with smaller level than k in H_ρ, which is a contradiction. Moreover, some node in ν must be in C for ρ to be superactive with respect to C. However, this implies that $a \rightarrow \nu \leftarrow b$ is a route that satisfies the requirements of the definition S1, which is a contradiction. Now, let φ denote the subroute of ρ between the first occurrences of b and k in ρ when going from b to a. By repeating the reasoning above with φ instead of ϱ, we can conclude that $b \not\perp_{H_\rho} k|C$ and, thus, that $splits(b, k, C, H_\rho)$ is defined too. Moreover, note that ϱ and φ have dlength equal or smaller than ρ and length strictly smaller than ρ. Therefore, the splitting routes for $splits(a, k, C, H_\rho)$ and $splits(b, k, C, H_\rho)$ are each either dshorter or shorter than ρ. This guarantees that the recursive definition S2 eventually reaches the trivial case S1.

Step 2 We prove the lemma by induction over the value of $splits(i, j, Z, G)$. If $splits(i, j, Z, G) = 0$, then there exists a route in G like that in Lemma 5 or 6. Therefore, there exists a probability distribution $p \in \mathcal{N}(G)$ such that $i \not\perp_p j|Z$ by Lemma 5 or 6.

Assume as induction hypothesis that the lemma holds for any value of $splits(i, j, Z, G)$ smaller than m ($m > 0$). We now prove it for value m. Recall that $splits(i, j, Z, G) = splits(i, k, Z, G_\rho) + splits(j, k, Z, G_\rho) + 1$ where ρ is a dshortest route among all the routes between i and j in G that are superactive with respect to Z, and k is a node in ρ but not in Zij that has minimal level in G_ρ. Then, as shown in Step 1, $i \not\perp_{G_\rho} k|Z$ and $j \not\perp_{G_\rho} k|Z$. Moreover, $splits(i, k, Z, G_\rho)$ and $splits(j, k, Z, G_\rho)$ are both smaller than m. Then, by the induction hypothesis, there exist two probability distributions $r, s \in \mathcal{N}(G_\rho)$ such that $i \not\perp_r k|Z$ and $j \not\perp_s k|Z$. We prove below that there exists a probability distribution $p \in \mathcal{N}(G_\rho)$ such that $i \not\perp_p j|Z$. Note that $p \in \mathcal{N}(G)$ because $\mathcal{N}(G_\rho) \subseteq \mathcal{N}(G)$ by Lemma 4.

By Lemma 2, there exists a real polynomial $S(i, k, Z)$ in the nd parameters in the parameterization of the probability distributions in $\mathcal{N}(G_\rho)$ such that, for every $q \in \mathcal{N}(G_\rho)$, $i \perp_q k|Z$ iff $S(i, k, Z)$ vanishes for the nd parameter values coding q. Furthermore, $S(i, k, Z)$ is non-trivial due to the probability distribution r above. Similarly, there exists a real polynomial $S(j, k, Z)$ in the nd parameters in the parameterization of the probability distributions in $\mathcal{N}(G_\rho)$ such that, for every $q \in \mathcal{N}(G_\rho)$, $j \perp_q k|Z$ iff $S(j, k, Z)$ vanishes for the nd parameter values coding q. Furthermore, $S(j, k, Z)$ is also non-trivial due to the probability distribution s above. Let $sol(i, k, Z)$ and $sol(j, k, Z)$ denote the sets of solutions to the polynomials $S(i, k, Z)$ and $S(j, k, Z)$, respectively. Let d denote the dimension of G_ρ. Then, $sol(i, k, Z)$ and $sol(j, k, Z)$ have both zero Lebesgue measure with respect to $\mathbb{R}^d$ because they consist of the solutions to non-trivial real polynomials in real variables (the nd parameters) (Okamoto, 1973). Then, $sol = sol(i, k, Z) \cup sol(j, k, Z)$ also has zero Lebesgue measure with respect to $\mathbb{R}^d$, because the finite union of sets of zero Lebesgue measure has zero Lebesgue measure too. Consequently, the probability distributions $q \in \mathcal{N}(G_\rho)$ such that $i \perp_q k|Z$ or $j \perp_q k|Z$ correspond to a set of elements of the nd parameter space for $\mathcal{N}(G_\rho)$ that has zero Lebesgue measure with respect to $\mathbb{R}^d$ because it is contained in sol. Since this correspondence is one-to-one by Lemma 1, Theorem 1 implies that there exists a probability distribution $p \in \mathcal{N}(G_\rho)$ such that $i \not\perp_p k|Z$ and $j \not\perp_p k|Z$. Note that these two independence statements together with $i \perp_p j|Zk$ would imply the desired result by symmetry and weak transitivity. We prove below $i \perp_{G_\rho} j|Zk$ which, in turn, implies $i \perp_p j|Zk$ because p is Markovian with respect to G_ρ, since $p \in \mathcal{N}(G_\rho)$ (Lauritzen, 1996, Proposition 3.30, Theorems 3.34 and 3.36).

Assume to the contrary $i \not\perp_{G_\rho} j|Zk$. Let ϱ denote any route between i and j in G_ρ that is superactive with respect to Zk. Note that ϱ must contain k because, otherwise, ϱ would be a route between i and j in G that is superactive with respect to Z and that is dshorter than ρ, which is a contradiction. Furthermore, ϱ must contain k only in collider sections because, otherwise, ϱ would not be superactive with respect to Zk. Let ν denote any collider section of ϱ that contains k. Note that $i = v_1 \rightarrow \nu \leftarrow v_l = j$ is a subroute of ϱ, because if

$v_1 \notin \{i, j\}$ or $v_l \notin \{i, j\}$ then there exists a node in ϱ but not in Zij with smaller level than k in G_ρ. Since ϱ is a route in G_ρ, this implies that there exists a node in ρ but not in Zij with smaller level than k in G_ρ, which is a contradiction. Note also that no descendant of k in G can be in Z because, otherwise, $i \rightarrow \nu \leftarrow j$ would be a route that satisfies the requirements of the definition S1, which is a contradiction. However, if no descendant of k in G is in Z, then ρ must contain k only in non-collider sections because, otherwise, ρ would not be superactive with respect to Z. The last two observations imply that i or j is a descendant of k in G which, together with $i \rightarrow \nu \leftarrow j$, implies that G has a directed pseudocycle between i and itself or between j and itself, because ν contains k. This is a contradiction. □

Acknowledgements

This work is funded by the Center for Industrial Information Technology (CENIIT) and a so-called career contract at Linköping University. We are thankful to Prof. Jose A. Lozano for discussions on this work.

References

Steen A. Andersson, David Madigan and Michael D. Perlman. Alternative Markov Properties for Chain Graphs. *Scandinavian Journal of Statistics*, 28:33-85, 2001.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Mathias Drton and Michael Eichler. Maximum Likelihood Estimation in Gaussian Chain Graph Models under the Alternative Markov Property. *Scandinavian Journal of Statistics*, 33:247-257, 2006.

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

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

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

Michael Levitz, Michael D. Perlman and David Madigan. Separation and Completeness Properties for AMP Chain Graph Markov Models. *The Annals of Statistics*, 29:1751-1784, 2001.

Radim Lněnička and František Matúš. On Gaussian Conditional Independence Structures. *Kybernetika*, 43:327-342, 2007.

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. Faithfulness in Chain Graphs: The Discrete Case. *International Journal of Approximate Reasoning*, 50:1306-1313, 2009.

Alberto Roverato and Milan Studený. A Graphical Representation of Equivalence Classes of AMP Chain Graphs. *Journal of Machine Learning Research*, 7:1045-1078, 2006.

Peter Spirtes, Clark Glymour, Richard Scheines. *Causation, Prediction, and Search*. Springer-Verlag, 1993.

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.

Martin Volf and Milan Studený. A Graphical Characterization of the Largest Chain Graphs. *International Journal of Approximate Reasoning*, 20:209-236, 1999.

Nanny Wermuth. On Block-Recursive Linear Regression Equations (with Discussion). *Brazilian Journal of Probability and Statistics*, 6:1-56, 1992.