Factorization, Inference and Parameter Learning in Discrete AMP Chain Graphs

Jose M. Peña
ADIT, IDA, Linköping University, SE-58183 Linköping, Sweden
jose.m.pena@liu.se

Abstract. We address some computational issues that may hinder the use of AMP chain graphs in practice. Specifically, we show how a discrete probability distribution that satisfies all the independencies represented by an AMP chain graph factorizes according to it. We show how this factorization makes it possible to perform inference and parameter learning efficiently, by adapting existing algorithms for Markov and Bayesian networks. Finally, we turn our attention to another issue that may hinder the use of AMP CGs, namely the lack of an intuitive interpretation of their edges. We provide one such interpretation.

1 Introduction

Chain graphs (CGs) are graphs with possibly directed and undirected edges, and no semidirected cycle. They have been extensively studied as a formalism to represent independence models, because they can model symmetric and asymmetric relationships between random variables. There are three different interpretations of CGs as independence models: The Lauritzen-Wermuth-Frydenberg (LWF) interpretation [6], the multivariate regression (MVR) interpretation [3], and the Andersson-Madigan-Perlman (AMP) interpretation [1]. No interpretation subsumes another [1,11].

In this paper, we focus on AMP CGs. Despite being much more expressive than Markov and Bayesian networks [10], AMP CGs have not enjoyed much success in the literature or in practice. We believe this is due to mainly two reasons. First, it is not known how to perform inference and parameter learning for AMP CGs efficiently, because it is not known how to factorize a probability distribution that satisfies all the independencies represented by an AMP CG. Compare this situation to that of LWF CGs, where such a factorization exists [4, Theorem 4.1] and thus inference can be performed efficiently [2, Section 6.5]. Second, AMP CGs do not appeal to intuition: Whereas the directed edges in a Bayesian network may be interpreted as causal relationships and the undirected edges in a Markov network as correlation relationships, it is not clear how to combine these two interpretations to produce an intuitive interpretation of the edges in an AMP CG.

In this paper, we address the two problems mentioned above. First, we introduce a factorization for AMP CGs and show how it makes it possible to perform
inference and parameter learning efficiently, by adapting existing algorithms for Markov and Bayesian networks. Second, we propose an intuitive interpretation of the edges in an AMP CG. We start with some notation and definitions.

2 Preliminaries

Unless otherwise stated, all the graphs and probability distributions in this paper are defined over a finite set of discrete random variables $V$. We use uppercase letters to denote random variables and lowercase letters to denote their states. The elements of $V$ are not distinguished from singletons. If a graph $G$ contains an undirected or directed edge between two nodes $V_1$ and $V_2$, then we write that $V_1 \xrightarrow{} V_2$ or $V_1 \rightarrow V_2$ is in $G$. The parents of a set of nodes $X$ of $G$ is the set $Pa_G(X) = \{V_i | V_1 \xrightarrow{} V_2 \text{ is in } G, V_i \notin X \text{ and } V_2 \in X\}$. The adjacents of $X$ is the set $Ad_G(X) = \{V_i | V_1 \xrightarrow{} V_2, V_1 \rightarrow V_2 \text{ or } V_1 \xrightarrow{} V_2 \text{ is in } G, V_i \notin X \text{ and } V_2 \in X\}$. A route between a node $V_1$ and a node $V_n$ in $G$ is a sequence of (not necessarily distinct) nodes $V_1, \ldots, V_n$ st $V_i \in Ad_G(V_{i+1})$ for all $1 \leq i < n$. If the nodes in the route are all distinct, then the route is called a path. A route is called descending if $V_i \rightarrow V_{i+1}$ or $V_i \leftarrow V_{i+1}$ is in $G$ for all $1 \leq i < n$. A route is called strictly descending if $V_i \rightarrow V_{i+1}$ is in $G$ for all $1 \leq i < n$. The descendants of a set of nodes $X$ of $G$ is the set $Dc_G(X) = \{V_n | \text{there is a descending route from } V_1 \text{ to } V_n \text{ in } G, V_1 \in X \text{ and } V_n \notin X\}$. The non-descendants of $X$ is the set $Nd_G(X) = V \setminus X \setminus Dc_G(X)$. The strict ascendants of $X$ is the set $Sa_G(X) = \{V_i | \text{there is a strictly descending route from } V_1 \text{ to } V_n \text{ in } G, V_1 \notin X \text{ and } V_n \notin X\}$.

The elements of $G$ are defined over a finite set of discrete random variables $X$. Unless otherwise stated, all the graphs and probability distributions in this paper are defined over a finite set of discrete random variables $V$. We use uppercase letters to denote random variables and lowercase letters to denote their states. The elements of $V$ are not distinguished from singletons. If a graph $G$ contains an undirected or directed edge between two nodes $V_1$ and $V_2$, then we write that $V_1 \xrightarrow{} V_2$ or $V_1 \rightarrow V_2$ is in $G$. The parents of a set of nodes $X$ of $G$ is the set $Pa_G(X) = \{V_i | V_1 \xrightarrow{} V_2 \text{ is in } G, V_i \notin X \text{ and } V_2 \in X\}$. The adjacents of $X$ is the set $Ad_G(X) = \{V_i | V_1 \xrightarrow{} V_2, V_1 \rightarrow V_2 \text{ or } V_1 \xrightarrow{} V_2 \text{ is in } G, V_i \notin X \text{ and } V_2 \in X\}$. A route between a node $V_1$ and a node $V_n$ in $G$ is a sequence of (not necessarily distinct) nodes $V_1, \ldots, V_n$ st $V_i \in Ad_G(V_{i+1})$ for all $1 \leq i < n$. If the nodes in the route are all distinct, then the route is called a path. A route is called descending if $V_i \rightarrow V_{i+1}$ or $V_i \leftarrow V_{i+1}$ is in $G$ for all $1 \leq i < n$. A route is called strictly descending if $V_i \rightarrow V_{i+1}$ is in $G$ for all $1 \leq i < n$. The descendants of a set of nodes $X$ of $G$ is the set $Dc_G(X) = \{V_n | \text{there is a descending route from } V_1 \text{ to } V_n \text{ in } G, V_1 \in X \text{ and } V_n \notin X\}$. The non-descendants of $X$ is the set $Nd_G(X) = V \setminus X \setminus Dc_G(X)$. The strict ascendants of $X$ is the set $Sa_G(X) = \{V_i | \text{there is a strictly descending route from } V_1 \text{ to } V_n \text{ in } G, V_1 \notin X \text{ and } V_n \notin X\}$.

The elements of $G$ are defined over a finite set of discrete random variables $X$. Unless otherwise stated, all the graphs and probability distributions in this paper are defined over a finite set of discrete random variables $V$. We use uppercase letters to denote random variables and lowercase letters to denote their states. The elements of $V$ are not distinguished from singletons. If a graph $G$ contains an undirected or directed edge between two nodes $V_1$ and $V_2$, then we write that $V_1 \xrightarrow{} V_2$ or $V_1 \rightarrow V_2$ is in $G$. The parents of a set of nodes $X$ of $G$ is the set $Pa_G(X) = \{V_i | V_1 \xrightarrow{} V_2 \text{ is in } G, V_i \notin X \text{ and } V_2 \in X\}$. The adjacents of $X$ is the set $Ad_G(X) = \{V_i | V_1 \xrightarrow{} V_2, V_1 \rightarrow V_2 \text{ or } V_1 \xrightarrow{} V_2 \text{ is in } G, V_i \notin X \text{ and } V_2 \in X\}$. A route between a node $V_1$ and a node $V_n$ in $G$ is a sequence of (not necessarily distinct) nodes $V_1, \ldots, V_n$ st $V_i \in Ad_G(V_{i+1})$ for all $1 \leq i < n$. If the nodes in the route are all distinct, then the route is called a path. A route is called descending if $V_i \rightarrow V_{i+1}$ or $V_i \leftarrow V_{i+1}$ is in $G$ for all $1 \leq i < n$. A route is called strictly descending if $V_i \rightarrow V_{i+1}$ is in $G$ for all $1 \leq i < n$. The descendants of a set of nodes $X$ of $G$ is the set $Dc_G(X) = \{V_n | \text{there is a descending route from } V_1 \text{ to } V_n \text{ in } G, V_1 \in X \text{ and } V_n \notin X\}$. The non-descendants of $X$ is the set $Nd_G(X) = V \setminus X \setminus Dc_G(X)$. The strict ascendants of $X$ is the set $Sa_G(X) = \{V_i | \text{there is a strictly descending route from } V_1 \text{ to } V_n \text{ in } G, V_1 \notin X \text{ and } V_n \notin X\}$.

The elements of $G$ are defined over a finite set of discrete random variables $X$. Unless otherwise stated, all the graphs and probability distributions in this paper are defined over a finite set of discrete random variables $V$. We use uppercase letters to denote random variables and lowercase letters to denote their states. The elements of $V$ are not distinguished from singletons. If a graph $G$ contains an undirected or directed edge between two nodes $V_1$ and $V_2$, then we write that $V_1 \xrightarrow{} V_2$ or $V_1 \rightarrow V_2$ is in $G$. The parents of a set of nodes $X$ of $G$ is the set $Pa_G(X) = \{V_i | V_1 \xrightarrow{} V_2 \text{ is in } G, V_i \notin X \text{ and } V_2 \in X\}$. The adjacents of $X$ is the set $Ad_G(X) = \{V_i | V_1 \xrightarrow{} V_2, V_1 \rightarrow V_2 \text{ or } V_1 \xrightarrow{} V_2 \text{ is in } G, V_i \notin X \text{ and } V_2 \in X\}$. A route between a node $V_1$ and a node $V_n$ in $G$ is a sequence of (not necessarily distinct) nodes $V_1, \ldots, V_n$ st $V_i \in Ad_G(V_{i+1})$ for all $1 \leq i < n$. If the nodes in the route are all distinct, then the route is called a path. A route is called descending if $V_i \rightarrow V_{i+1}$ or $V_i \leftarrow V_{i+1}$ is in $G$ for all $1 \leq i < n$. A route is called strictly descending if $V_i \rightarrow V_{i+1}$ is in $G$ for all $1 \leq i < n$. The descendants of a set of nodes $X$ of $G$ is the set $Dc_G(X) = \{V_n | \text{there is a descending route from } V_1 \text{ to } V_n \text{ in } G, V_1 \in X \text{ and } V_n \notin X\}$. The non-descendants of $X$ is the set $Nd_G(X) = V \setminus X \setminus Dc_G(X)$. The strict ascendants of $X$ is the set $Sa_G(X) = \{V_i | \text{there is a strictly descending route from } V_1 \text{ to } V_n \text{ in } G, V_1 \notin X \text{ and } V_n \notin X\}$.
Let $X$, $Y$ and $Z$ denote three disjoint subsets of $V$. When there is no $Z$-open path in an AMP CG $G$ between a node in $X$ and a node in $Y$, we say that $X$ is separated from $Y$ given $Z$ in $G$ and denote it as $X \perp_G Y | Z$. The independence model represented by $G$ is the set of separations $X \perp_G Y | Z$. The independence model represented by $G$ under marginalization of some nodes $L \subseteq V$ is the set of separations $X \perp_G Y | Z$ with $X, Y, Z \subseteq V \setminus L$. Finally, we denote by $X \perp_p Y | Z$ that $X$ is independent of $Y$ given $Z$ in a probability distribution $p$. We say that $p$ is Markovian wrt an AMP CG $G$ when, for all $X$, $Y$ and $Z$ disjoint subsets of $V$, if $X \perp_G Y | Z$ then $X \perp_p Y | Z$.

## 3 Factorization

A probability distribution $p$ is Markovian wrt an AMP CG $G$ iff the following three properties hold for all $C \in C_c(G)$ [1, Theorem 2]:

- C1: $C \perp_p \cup G(C) \setminus C(c(C))\cup(G(C))$.
- C2: $p(C|C_c(G)(P_a(G(C))))$ is Markovian wrt $G_C$.
- C3*: For all $D \subseteq C$, $D \perp_p C_c(G(P_a(G(C))) \setminus P_a(G(D))|P_a(G(D))$.

Then, C1 implies that $p$ factorizes as

$$p = \prod_{C \in C_c(G)} p(C|C_c(G)(P_a(G(C))))$$

The authors of [1, p. 50] note that if $p$ were strictly positive and $G$ were a LWF CG, then each conditional distribution above would factorize further into a product of potentials over certain subsets of the nodes in $C \cup P_a(G(C))$, as shown in [4, Theorem 4.1]. However, the authors state that no such further factorization appears to hold in general if $G$ is an AMP CG. We show that this is not true if $p$ is strictly positive. Specifically, C2 together with [6, Theorems 3 and 3.9] imply that

$$p(C|C_c(G)(P_a(G(C)))) = \prod_{K \in C(G)} \varphi(K,C_c(G)(P_a(G(C))))$$

However, one can show that $\varphi(K,C_c(G)(P_a(G(C))))$ is actually a function of $K \cup P_a(G(K))$, i.e. $\varphi(K,C_c(G)(P_a(G(C)))) = \varphi(K,P_a(G(K)))$. It suffices to recall from the proof of [6, Theorem 3.9] how $\varphi(K,C_c(G)(P_a(G(C))))$ can be obtained from $p(C|C_c(G)(P_a(G(C))))$, a method also known as canonical parameterization [5, Section 4.4.2.1]. Specifically, let $\phi(k,C_c(G)(P_a(G(C)))) = \log \varphi(K,C_c(G)(P_a(G(C))))$. Choose a fixed but arbitrary state $k_s$ of $K$. Then,

$$\phi(k,C_c(G)(P_a(G(C)))) = \sum_{\eta \in k_s} (-1)^{|k_s - \eta|} \log p(q,\eta_s|C_c(G)(P_a(G(C))))$$

where $\eta_s$ denotes the elements of $k_s$ corresponding to the elements of $K \setminus Q$. Now, note that $p(q,\eta_s|C_c(G)(P_a(G(C)))) = p(q,\eta_s|P_a(G(K)))$ by C3*, because $Q \subseteq K$. Then, $\varphi(K,C_c(G)(P_a(G(C))))$ is actually a function of $K \cup P_a(G(K))$. 
Putting together the results above, we have that $p$ factorizes as
\[
p = \prod_{C \in Cc(G)} \prod_{K \in Cl(G_C)} \varphi(K, Pa_C(K)) = \prod_{C \in Cc(G)} \prod_{K \in Cl(G_C)} \psi(K, Pa_C(K)).
\]

Note that the well-known factorizations induced by DAGs and UGs (see [6, Sections 3.2.1 and 3.2.2]) are special cases of Equation 1.

4 Parameter Learning

The factorization in Equation 1 enables us to perform parameter learning for AMP CGs efficiently by deploying the iterative proportional fitting procedure (IPFP) [8, Section 19.5.7], which returns the maximum likelihood estimates of the entries of the potentials for some given data. Specifically, we first simplify further the factorization by multiplying its potentials until no potential domain is included in another potential domain. Let $Q_1, \ldots, Q_n$ denote the potential domains in the resulting factorization. Note that each domain $Q_i$ is of the form $K \cup Pa_C(K)$ with $K \in Cl(G_C)$ and $C \in Cc(G)$. Then, we run the IPFP per se:

1. For each potential $\psi(Q_i)$
2. Set $\psi^0(Q_i) = 1$
3. Repeat until convergence
4. For each potential $\psi^t(Q_i)$
5. Set $\psi^{t+1}(Q_i) = \psi^t(Q_i) \frac{p_e(Q_i)}{p^t(Q_i)}$

where $p^t = \prod_{i=1}^n \psi^t(Q_i)$, and $p_e$ is the empirical probability distribution over $V$ obtained from the given data.

5 Inference

The factorization in Equation 1 also enables us to perform inference in AMP CGs efficiently by deploying the algorithm for inference in DAGs developed by [7], and upon which most other inference algorithms build. Specifically, we start by transforming $G$ into its moral graph $G^m$ by running the procedure below. This procedure differs from the one in [7], because $G$ is an AMP CG and not a DAG. In any case, the moralization procedure in [7] is a special case of the procedure below.

1. Set $G^m = G$
2. For each connectivity component $C \in Cc(G)$
3. For each clique $K \in Cl(G_C)$
4. Add the edge $X \to Y$ to $G^m$ for all $X \in Pa_C(K)$ and $Y \in K$
5. Add the edge $X \to Y$ to $G^m$ for all $X,Y \in Pa_C(K)$ s.t. $X \neq Y$
6. Replace all the directed edges in $G^m$ with undirected edges
The reason of why $G^m$ has the edges it has will become clear later. We continue by transforming $G^m$ into a triangulated graph $G^t$, and sorting its cliques to satisfy the so-called running intersection property. The procedure below accomplishes these two objectives. An UG is triangulated when every cycle in it contains a chord, i.e. an edge between two non-consecutive nodes in the cycle. The cliques of a triangulated graph can be ordered as $Q_1, \ldots, Q_n$ so that for all $1 < j \leq n$, $Q_j \cap (Q_1 \cup \ldots \cup Q_{j-1}) \subseteq Q_i$ for some $1 \leq i < j$. This is known as the running intersection property (RIP).

1. Set $G^t = G^m$
2. Repeat until all the nodes in $G^t$ are marked
3. Select an unmarked node in $G^t$ with the largest number of marked neighbours
4. Mark the node and make its marked neighbours form a complete set in $G^t$ by adding undirected edges
5. Save the node plus its marked neighbours as a candidate clique
6. Remove every candidate clique that is included in another
7. Label every clique with the last iteration that marked one of its nodes
8. Sort the cliques in ascending order of their labels

Finally, let $Q_1, \ldots, Q_n$ denote the ordering of the cliques of $G^t$ returned by the procedure above. Let $S_j = Q_j \cap (Q_1 \cup \ldots \cup Q_{j-1})$ and $R_j = Q_j \setminus S_j$. Note that for every $K \in \text{Cl}(G_C)$ with $C \in \text{Cc}(G)$, there is some $Q$, st $K \cup Pa_G(K) \subseteq Q$, because the moralization procedure above made $K \cup Pa_G(K)$ a complete set in $G^m$ and thus in $G^t$. Then,

$$p(V) = \prod_{C \in \text{Cc}(G)} \prod_{K \in \text{Cl}(G_C)} \psi(K, Pa_G(K)) = \prod_{i=1}^{n} \phi(Q_i)$$

and thus

$$p(V) = f([Q_1 \cup \ldots \cup Q_{n-1}] \setminus S_n, S_n)g(S_n, R_n)$$

and thus

$$R_n \perp p[Q_1 \cup \ldots \cup Q_{n-1}] \setminus S_n|S_n$$

by [6] p. 29, and thus

$$p(V) = p(Q_1 \cup \ldots \cup Q_{n-1})p(R_n|Q_1 \cup \ldots \cup Q_{n-1}) = p(Q_1 \cup \ldots \cup Q_{n-1})p(R_n|S_n).$$

Note also that

$$p(Q_1 \cup \ldots \cup Q_{n-1}) = \sum_{r_n} p(Q_1 \cup \ldots \cup Q_{n-1}, r_n) = \left[ \prod_{i=1}^{n-1} \phi(Q_i) \right] \sum_{r_n} \phi(S_n, r_n).$$

Then, Equations 2-4 imply that

$$p(R_n|S_n) = \phi(Q_n)/\sum_{r_n} \phi(S_n, r_n).$$
Note that $S_n \subseteq Q_j$ for some $1 \leq j < n$ by the RIP. Then, we replace $\phi(Q_j)$ with $\phi(Q_j) \sum_{r_n} \phi(S_n, r_n)$, after which Equation 4 implies that

$$p(Q_1 \cup \ldots \cup Q_{n-1}) = \prod_{i=1}^{n-1} \phi(Q_i).$$

We repeat the steps above for $p(Q_1 \cup \ldots \cup Q_{n-1})$ and so we obtain $p(R_i|S_i)$ for all $1 \leq i \leq n$. Now, note that $S_1 = \emptyset$ and, thus, $p(Q_1) = p(R_1|S_1)$. Moreover, since $S_2 \subseteq Q_1$ by the RIP, then

$$p(S_2) = \sum_{q_1 \setminus s_2} p(S_2, q_1 \setminus s_2)$$

and thus

$$p(Q_2) = p(R_2|S_2)p(S_2).$$

We repeat the steps above for $Q_3, \ldots, Q_n$ and so we obtain $p(Q_i)$ for all $1 \leq i \leq n$. To obtain $p(Q_i|o)$ where $o$ denotes some observations or evidence, we first remove all the entries of $\phi(Q_j)$ that are inconsistent with $o$ for all $1 \leq j \leq n$, then we repeat the steps above to get $p(Q_i, o)$ and, finally, we normalize by $p(o) = \sum_{q_i} p(q_i, o)$. To obtain $p(X|o)$ where $X \notin Q_i$ for all $1 \leq i \leq n$, we compute $p(x, o)$ for all $x$ as if $\{x, o\}$ were the observations and, then, we normalize by $p(o) = \sum_x p(x, o)$.

### 6 Error AMP CGs

So far in this article, we have shown how an AMP CG factorizes a probability distribution, and how this helps in performing parameter learning and inference efficiently. We believe that our findings solve some computational issues that have hindered the use of AMP CGs in practice. In this section, we turn our attention to another issue that may have also hindered the use of AMP CGs, namely the lack of an intuitive interpretation of their edges. Whereas the directed edges in a DAG may be interpreted as causal relationships and the undirected edges in an UG as correlation relationships, it is not clear how to combine these two interpretations to produce an intuitive interpretation of the edges in an AMP CG. We propose here a way to do it by adapting to discrete AMP CGs the interpretation for Gaussian AMP CGs presented in [1, Section 5] and further studied in [9, Section 3]. Specifically, we propose to interpret the directed edges in an AMP CG as causal relationships. In other words, the parents of a node represent its causal mechanism. We propose to assume that this mechanism is deterministic but it may sometimes work erroneously. We propose to interpret the undirected edges in the AMP CG as the correlation structure of the errors of the causal mechanisms of the different nodes. To show the validity of this interpretation, we will first modify the AMP CG by adding a deterministic node for each original node to represent explicitly the occurrence or not of an error in its causal mechanism and, then, we will show that the original and the modified AMP CGs are equivalent in some sense. We call the modified CG an error AMP (EAMP) CG. Since an EAMP CG is an AMP CG with deterministic nodes, we discuss these first.
6.1 AMP CGs with Deterministic Nodes

We say that a node $X$ of an AMP CG is determined by some $Z \subseteq V$ when $X \in Z$ or $X$ is a function of $Z$ in each probability distribution that is Markovian wrt the CG. In that case, we also say that $X$ is a deterministic node. We use $D(Z)$ to denote all the nodes that are determined by $Z$. From the point of view of the separations in an AMP CG, that a node outside the conditioning set of a separation is determined by it, has the same effect as if the node were actually in the conditioning set. We extend accordingly the definition of separation for AMP CGs to the case where deterministic nodes may exist. Given an AMP CG $G$, a path $\rho$ in $G$ is said to be $Z$-open when
- every triplex node in $\rho$ is in $D(Z) \cup S_G(D(Z))$, and
- no non-triplex node $B$ in $\rho$ is in $D(Z)$, unless $A - B - C$ is a subpath of $\rho$ and $Pa_G(B) \setminus D(Z) \neq \emptyset$.

6.2 EAMP CGs

The EAMP CG $H$ corresponding to an AMP CG $G$ is an AMP CG over $V \cup E$, where $E$ denotes the error nodes. Specifically, there is an error node $E_X$ for every node $X \in V$, and it represents whether an error in the causal mechanism of $X$ occurs or not. We set $Pa_H(X) = Pa_G(X) \cup E_X$ to represent that $E_X$ is part of the causal mechanism of $X$ in $H$. This causal mechanism works as follows: If $E_X = 0$ (i.e. no error) then $pa_G(X)$ determines the state of $X$ to be the distinguished state $x^*_{Pa_G(X)}$, else $X$ may take any state but the distinguished one. The undirected edges in $H$ are all between error nodes, and they represent the correlation structure of the error nodes. Specifically, the undirected edge $E_X - E_Y$ is in $H$ iff the undirected edge $X - Y$ is in $G$. Note that the error nodes are never observed, i.e. they are latent. The procedure below formalizes the transformation just described. See Figure 1 for an example.

1. Set $H = G$
2. For each node $X \in V$
3. Add the node $E_X$ and the edge $E_X \rightarrow X$ to $H$
4. Replace every edge $X - Y$ in $H$ by $X, Y \in V$ with an edge $E_X - E_Y$
Now, consider a probability distribution $p(V, E)$ that is Markovian wrt the EAMP CG $H$. Then,

$$p(V, E) = p(V|E)p(E) = \prod_{X \in V} p(X|Pa_G(X), E_X)p(E) \quad (5)$$

by C1 and C3*. Moreover, in order for the causal mechanism of $X$ in $H$ to match the description above, we restrict $p(X|Pa_G(X), E_X)$ to be of the following form:

$$p(X|Pa_G(X), E_X) = \begin{cases} 
1 & \text{if } E_X = 0 \text{ and } X = x_{pa_G(X)}^* \\
0 & \text{if } E_X = 0 \text{ and } X \neq x_{pa_G(X)}^* \\
q(X|pa_G(X)) & \text{if } E_X = 1
\end{cases} \quad (6)$$

where $q(X|pa_G(X))$ is an arbitrary conditional probability distribution with the only constraints that $q(X|pa_G(X)) = 0$ if $X = x_{pa_G(X)}^*$, and $q(X|pa_G(X)) > 0$ otherwise. The first constraint follows from the description above of the causal mechanism of $X$ in $H$, whereas the second is necessary for $p(V)$ being strictly positive. Note that $E_X$ is determined by $Pa_G(X) \cup X$. Specifically, if $X = x_{pa_G(X)}^*$ then $E_X = 0$, else $E_X = 1$. Then, $E$ is determined by $V$. Hereinafter, when we say that a probability distribution is Markovian wrt an EAMP CG, it should be understood that it also satisfies the constraint in Equation (6).

We assume that $p(E)$ is strictly positive, as a way to ensure that $p(V)$ is strictly positive. This together with the fact that $p(E)$ is Markovian wrt $H_E$, which follows from $p(V, E)$ being Markovian wrt $H$, implies that $p(E)$ factorizes as shown in Equation (1) and, thus, Equation (5) becomes

$$p(V, E) = \prod_{X \in V} p(X|Pa_G(X), E_X) [\prod_{E_C \in Cc(H_E)} \prod_{E_K \in Cl(H_{E_C})} \phi(E_K)]. \quad (7)$$

Thus, it is clear that the EAMP CG $H$ can be interpreted as we wanted: Each node is controlled by the causal mechanism specified in the AMP CG $G$, the mechanism is deterministic if no error occurs and it is random otherwise, and the errors of the different mechanisms obey the correlation structure specified in $G$. To see the last point, note that $E_C \in Cc(H_E)$ iff $C \in Cc(G)$, and $E_K \in Cl(H_{E_C})$ iff $K \in Cl(G_{E_C})$. Thus, $H$ somehow keeps the structural information in $G$. To make this claim more specific, note that the independence model represented by $G$ coincides with that represented by $H$ under marginalization of the error nodes which, recall from above, are latent [9, Theorem 1].

Recall that the independence

1 Unlike in this work, $V$ is a Gaussian random variable in [9]. However, that is irrelevant in the proof of [9, Theorem 1]. The proof builds upon the following two properties which, as we show, also hold for the framework in this work:

- A node $E_X \in E$ is determined by some $Z \subseteq V$ iff $Pa_G(X) \cup X \subseteq Z$. The if part follows from the fact shown above that $E_X$ is determined by $Pa_G(X) \cup X$. To see the only if part, assume to the contrary that $Z$ determines $E_X$ but $Pa_G(X) \cup X \not\subseteq Z$. Then, $X \notin Z$ or there is some $Y \in Pa_G(X) \setminus Z$. If $X \notin Z$, then let $H'$ be the EAMP CG $H'$ over $V \cup E$ whose only edge is $E_X \to X$, and let $p'$ be a probability distribution
model represented by \( H \) can be read off as shown in Section 6.1. Note that that the independence model represented by \( G \) coincides with that represented by \( H \) under marginalization of the error nodes implies that the probability distribution resulting from marginalizing \( E \) out of a distribution \( p(V, E) \) that is Markovian wrt \( H \) is Markovian wrt \( G \) and, thus, it factorizes as shown in Equation 1. Specifically, recall that \( E \) is determined by \( V \) and, thus, \( p(V, E) \) is actually a function of \( V \). Then, it suffices to set each potential \( \psi(K, Pa_G(K)) \) in Equation 1 equal to the following product of the terms in Equation 2

\[
\psi(K, Pa_G(K)) = \prod_{X \in K} p(X|Pa_G(X), E_X) \phi(E_K)
\]

bearing in mind that if \( X \) belongs to several cliques \( K \), then \( p(X|Pa_G(X), E_X) \) is assigned to only one (any) of the potentials \( \psi(K, Pa_G(K)) \). For instance, the following is a valid assignment for the AMP and EAMP CGs in Figure 1

\[
\psi(A) = p(A|E_A)\phi(E_A) \quad \psi(C, F, A) = p(F|E_F)\phi(E_C, E_F) \\
\psi(B, A) = p(B|A, E_B)\phi(E_B) \quad \psi(D, I, A, B) = p(D|A, B, E_D)\phi(E_D, I) \\
\psi(C, D, A, B) = p(C|A, E_C)\phi(E_C, E_D) \quad \psi(F, I) = p(I|E_I)\phi(E_F, E_I)
\]

Unfortunately, the opposite of the last result above does not hold. That is, not every probability distribution that factorizes according to an AMP CG coincides with the marginal of a distribution that is Markovian wrt the corresponding EAMP CG. To see it, let \( G \) be the AMP CG \( A \rightarrow B \rightarrow C \). Let \( H \) be the EAMP CG corresponding to \( G \), i.e. \( E_A \rightarrow A \rightarrow B \leftarrow E_B \rightarrow E_C \rightarrow C \). Consider a probability distribution \( p(A, B, C, E_A, E_B, E_C) \) that is Markovian wrt \( H \). Since as shown above \( \{E_A, E_B, E_C\} \) is determined by \( \{A, B, C\} \), Equation 4 implies that

\[
p(a_0, b^{a_0}_{E_A}, C) \begin{array}{c}
p(a_0|E_A)p(b^{a_0}_{E_A}|a_0, E_B)p(c|E_C)\phi(E_A)\phi(E_B, E_C) \end{array} = p(a_0|E_A)\phi(E_A) = \begin{array}{c}
p(a_1|E_A)p(b^{a_1}_{E_A}|a_1, E_B)p(c|E_C)\phi(E_A)\phi(E_B, E_C) \end{array}
\]

because both \( \{a_0, b^{a_0}_{E_A}\} \) and \( \{a_1, b^{a_1}_{E_A}\} \) determine that \( E_B = 0 \), which implies that \( p(b^{a_0}_{E_A}|a_0, E_B) = p(b^{a_1}_{E_A}|a_1, E_B) = 1 \). Now, consider a probability distribution \( p'(A, B, C) \) that factorizes according to \( G \). Then, Equation 1 implies that

\[
p'(a_0, b^{a_0}_{E_A}, C) = p(a_0)\psi(a_0, b^{a_0}_{E_A}, C) \\
p'(a_1, b^{a_1}_{E_A}, C) = p(a_1)\psi(a_1, b^{a_1}_{E_A}, C)
\]

Note that the ratio in Equation 8 is a function of \( C \) whereas the ratio in Equation 5 is not. Therefore, \( p(A, B, C) \neq p'(A, B, C) \) in general. 

that is Markovian wrt \( H' \). Note that \( E_X \) is a function of just \( X \) in \( p' \). If \( X \in Z \), then let \( H' \) have the edges \( E_X \rightarrow X \leftarrow Y \leftarrow E_Y \), and let \( p' \) be Markovian wrt \( H' \) st \( x^Z \neq x^Y \). Note that \( E_X \) is a function of just \( X \cup Y \) in \( p' \). Note also that in either case \( p' \) is Markovian wrt \( H \), because \( H' \) is a subgraph of \( H \). Note also that in neither case \( E_X \) is a function of \( Z \) in \( p' \). This contradicts that \( Z \) determines \( E_X \).

- A node \( X \in V \) is determined by some \( Z \subseteq V \) iff \( X \in Z \). If the part is trivial. To see the only if part, note that \( X \) is determined by \( Z \) only if \( X \in Z \) or \( E_X \) is determined by \( Z \). However, \( E_X \) is determined by \( Z \) only if \( X \in Z \) by the previous property.
Finally, note that every node $X \in V$ in an EAMP CG $H$ forms a connectivity component on its own. Therefore, the factorization in Equation 7 is actually of the same form as the factorization in Equation 1. This comes as no surprise because, after all, $H$ is an AMP CG over $V \cup E$.

7 Discussion

We have addressed some issues that may hinder the use of AMP CGs in practice. We hope that the results reported in this paper help others to deploy AMP CGs in practical applications. Specifically, we have shown how a discrete probability distribution that is Markovian wrt an AMP CG factorizes according to it. We have also shown how this factorization makes it possible to perform inference and parameter learning efficiently. Finally, we have provided an intuitive interpretation of AMP CGs that sheds some light on what the different edges may mean. Unfortunately, the interpretation provided is not perfect, i.e. not every probability distribution that factorizes according to an AMP CG coincides with the marginal of a distribution that is Markovian wrt the corresponding EAMP CG. We are working to solve this problem. We are also working on proving the opposite of the result in Section 3, i.e. proving that every probability distribution that factorizes according to an AMP CG is Markovian wrt it.

Acknowledgments. This work is funded by the Swedish Research Council (ref. 2010-4808), and by a so-called career contract at Linköping University.

References

1. Andersson, S. A., Madigan, D. and Perlman, M. D. Alternative Markov Properties for Chain Graphs. *Scandinavian Journal of Statistics*, 28:33-85, 2001.
2. Cowell, R. G., Dawid, P., Lauritzen, S. L. and Spiegelhalter, D. J. *Probabilistic Networks and Expert Systems*. Springer, 1999.
3. Cox, D. R. and Wermuth, N. *Multivariate Dependencies - Models, Analysis and Interpretation*. Chapman & Hall, 1996.
4. Frydenberg, M. The Chain Graph Markov Property. *Scandinavian Journal of Statistics*, 17:333-353, 1990.
5. Koller, D. and Friedman, N. *Probabilistic Graphical Models*. MIT Press, 2009.
6. Lauritzen, S. L. *Graphical Models*. Oxford University Press, 1996.
7. Lauritzen, S. L. and Spiegelhalter, D. J. Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. *Journal of the Royal Statistical Society B*, 50:157-224, 1988.
8. Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
9. Peña, J. M. Marginal AMP Chain Graphs. *International Journal of Approximate Reasoning*, 55:1185-1206, 2014.
10. Sonntag, D. On Expressiveness of the AMP Chain Graph Interpretation. In *Proceedings of the 7th European Workshop on Probabilistic Graphical Models*, 458-470, 2014.
11. Sonntag, D. and Peña, J. M. Chain Graph Interpretations and their Relations Revisited. *International Journal of Approximate Reasoning*, to appear.