The Lauritzen-Chen Likelihood For Graphical Models

Ilya Shpitser
Department of Computer Science
Johns Hopkins University
Baltimore, MD, 21218
ilyas@cs.jhu.edu

Abstract

Graphical models such as Markov random fields (MRFs) that are associated with undirected graphs, and Bayesian networks (BNs) that are associated with directed acyclic graphs, have proven to be a very popular approach for reasoning under uncertainty, prediction problems and causal inference.

Parametric MRF likelihoods are well-studied for Gaussian and categorical data. However, in more complicated parametric and semi-parametric settings, likelihoods specified via clique potential functions are generally not known to be congenial (jointly well-specified) or non-redundant. Congenial and non-redundant DAG likelihoods are far simpler to specify in both parametric and semi-parametric settings by modeling Markov factors in the DAG factorization. However, DAG likelihoods specified in this way are not guaranteed to coincide in distinct DAGs within the same Markov equivalence class. This complicates likelihoods based model selection procedures for DAGs by "sneaking in" potentially unwarranted assumptions about edge orientations.

In this paper we link a density function decomposition due to Chen with the clique factorization of MRFs described by Lauritzen to provide a general likelihood for MRF models. The proposed likelihood is composed of variationally independent, and non-redundant closed form functionals of the observed data distribution, and is sufficiently general to apply to arbitrary parametric and semi-parametric models. We use an extension of our developments to give a general likelihood for DAG models that is guaranteed to co-incide for all members of a Markov equivalence class. Our results have direct applications for model selection and semi-parametric inference.

1 INTRODUCTION

Graphical Markov models are a widely used approach for probabilistic modeling tasks of all types, including natural language and speech processing (Bilmes and Bartels, 2005; Smith, 2011), computational biology (Sachs et al., 2005), computer vision (Ji, 2019), and causal inference (Spirtes et al., 2001; Pearl, 2009), among many others. The popularity of graphical models stems from their tractable likelihoods expressed via factorizations, and intuitive graphical visualization of conditional independence restrictions in the model. Common graphical models include Markov random fields (MRFs) (Ising, 1925; Sherington and Kirkpatrick, 1975), associated with undirected graphs (UGs), and Bayesian networks (BNs) (Pearl, 1988), associated with directed acyclic graphs (DAGs).

MRF likelihoods for Gaussian and categorical data are very well studied (Lauritzen, 1996). In modern high dimensional applications (Maathuis et al., 2019), MRF likelihoods are specified by modeling terms in the clique factorization of the model: $p(\vec{v}) = \prod_{\vec{C}} \phi_{\vec{C}}(\vec{v}_{\vec{C}})$, where the product is over potential functions corresponding to cliques in an undirected graph, and $\vec{v}_{\vec{C}}$ is the value assignment $\vec{v}$ restricted to $\vec{C} \subseteq \vec{V}$. Terms $\phi_{\vec{C}}$ in such likelihoods are not, in general, known to be congenial and non-redundant.

Congeniality refers to the property that different parts of the model are not specified in a mutually contradictory way. Congeniality was first described in the context of multiple imputation in missing data problems, where careless use of analysis and missing data models may impose contradictory restrictions, guaranteeing model misspecification (Meng, 1994). However, lack of congeniality becomes a concern whenever different parts of the model specification "overlap" and impose different restrictions on the same set of variables, as may happen with clique factors of an MRF.

A likelihood is non-redundant if different parts of the likelihood specification do not carry the same information about the model. Redundant likelihoods may be undesirable as they do not have identified parameterizations, and do not yield insight about the model dimension or the model tangent space. Incongenial and redundant likelihoods are described in more detail below.

Though in many applications congenial and non-redundant likelihoods are not needed, they are crucial if model parameters are of primary interest. In addition, such likelihoods are important for deriving semi-parametrically efficient estimators based on influence functions that take advantage of Markov restrictions in graphical models (Tsiatis, 2006), and for score based model selection algorithms (Chickering, 2002).

In this paper, we describe general likelihoods for MRFs based on terms of the clique factorization that are guaranteed to be congenial and non-redundant. These terms are defined algebraic functions of the observed data distributions termed *partial cross-product ratios* by Lauritzen (1996) that may be viewed as generalizations of odds ratio functions, or as generalized higher order interaction functions. The likelihood is based on a synthesis of ideas on clique factorizations presented by Lauritzen (1996), and the odds ratio decomposition derived by Chen (2007). We use this likelihood to define parametric and semi-parametric likelihoods for BN models that are guaranteed to be identical within distinct DAGs that are Markov equivalent (obey the same set of marginal and conditional independence restrictions).

After introducing some preliminaries in Section 2 and illustrating the concepts of congeniality, non-redundance, and variational independence in Section 3, we describe the Chen and Lauritzen decompositions of conditional distributions in Section 4, describe the main results showing the close relationship of these decompositions in Section 5, and describe BN likelihoods in Section 6. We conclude with some examples of the derived likelihoods, and discuss limitations and areas of future work.

2 PRELIMINARIES

We first introduce necessary graphical modeling preliminaries. Graphs are assumed to have a vertex set $\vec{V}$, and we will restrict attention to positive distributions. Given any graph $\mathcal{G}$, for $\vec{S} \subseteq \vec{V}$, an induced subgraph $\mathcal{G}_{\vec{S}}$ of $\mathcal{G}$ is defined as the graph with a vertex set $\vec{S}$ and all edges in $\mathcal{G}$ connecting elements in $\vec{S}$.

Given an undirected graph (UG) $\mathcal{G}$, a clique $\vec{C}$ is a (possibly empty) subset of vertices in $\vec{V}$ that are pairwise connected in $\mathcal{G}$. The set of all cliques in $\mathcal{G}$ is denoted by $\mathcal{C}(\mathcal{G})$, while the set of all *maximal* cliques is denoted by $\bar{\mathcal{C}}(\mathcal{G})$. Note that, in general, neither $\mathcal{C}(\mathcal{G})$ nor $\bar{\mathcal{C}}(\mathcal{G})$ will

form a partition of $\vec{V}$ in $\mathcal{G}$. A joint distribution $p(\vec{v})$ is in the Markov random field (MRF) model of a UG $\mathcal{G}$ if for every value $\vec{v}$, $p(\vec{v}) = \prod_{\vec{C} \in \mathcal{C}(\mathcal{G})} \phi_{\vec{C}}(\vec{v}_{\vec{C}})$, where $\phi_{\vec{C}}$ are *potential functions* which map values of $\vec{C}$ to real numbers. Potential functions are *not* necessarily normalized probabilities. Equivalently, $p(\vec{v})$ is in the MRF model if $p(\vec{v}) = Z^{-1} \prod_{\vec{C} \in \bar{\mathcal{C}}(\mathcal{G})} \phi_{\vec{C}}(\vec{v}_{\vec{C}})$, where Z is a normalizing constant.

If we restrict attention to positive distributions, an MRF model may be equivalently defined as the set of distributions $p(\vec{v})$ that satisfy either the global or pairwise Markov property for $\mathcal{G}$. The global Markov property for $p(\vec{v})$ and a UG $\mathcal{G}$ states that for any disjoint subsets $\vec{A}, \vec{B}, \vec{C}$ of $\vec{V}$ whenever all paths from $\vec{A}$ to $\vec{B}$ in $\mathcal{G}$ are intercepted by $\vec{C}$, then $\vec{A} \perp\!\!\!\perp \vec{B} | \vec{C}$ in $p(\vec{v})$. The pairwise Markov property for $p(\vec{v})$ and $\mathcal{G}$ states that for any vertex pair A, B non-adjacent in $\mathcal{G}$, $A \perp\!\!\!\perp B | \vec{V} \setminus \{A, B\}$ in $p(\vec{v})$.

A joint distribution $p(\vec{v})$ is in the Bayesian network (BN) model of a directed acyclic graph (DAG) $\mathcal{G}$ if for every value $\vec{v}$, $p(\vec{v}) = \prod_{V \in \vec{V}} p(\vec{v}_{\{V\}} | \vec{v}_{\mathrm{pa}_{\mathcal{G}}(V)})$, where $\mathrm{pa}_{\mathcal{G}}(V)$ are the set of parents of V in $\mathcal{G}$. A BN model may equivalently defined as the set of distributions $p(\vec{v})$ that satisfy the global Markov property for $\mathcal{G}$ given by the d-separation criterion (Pearl, 1988), where for any disjoint subsets $\vec{A}, \vec{B}, \vec{C}$ of $\vec{V}$ whenever all paths from $\vec{A}$ to $\vec{B}$ in $\mathcal{G}$ are d-separated by $\vec{C}$, then $\vec{A} \perp\!\!\!\perp \vec{B} | \vec{C}$ in $p(\vec{v})$.

Both MRF and BN models may be generalized into *conditional MRF (CMRF)* (Lafferty et al., 2001) and *conditional BN (CBN)* models, associated with conditional undirected graphs (CUGs) and conditional directed acyclic graphs (CDAGs). A CUG $\mathcal{G}(\vec{V}, \vec{W})$ is a mixed graph containing *random vertices* $\vec{V}$ and *fixed vertices* $\vec{W}$ as well as undirected and directed edges, such that undirected edges are only among elements of $\vec{V}$ and directed edges are always from an element in $\vec{W}$ to an element in $\vec{V}$. Similarly, a CDAG $\mathcal{G}(\vec{V}, \vec{W})$ is a directed graph containing *random vertices* $\vec{V}$ and *fixed vertices* $\vec{W}$, such that there are no directed cycles, and the only allowed edges adjacent to $\vec{W}$ are out of elements in $\vec{W}$ and into elements in $\vec{V}$. The notion of parents generalizes to conditional graphs to potentially include fixed vertices in $\vec{W}$.

A distribution $p(\vec{v} | \vec{w})$ is in the CMRF model of a CUG $\mathcal{G}(\vec{V}, \vec{W})$ if for every $\vec{v}, \vec{w}$, $p(\vec{v} | \vec{w}) = \prod_{\vec{C} \in \mathcal{C}(\mathcal{G}_{\vec{V}})} \phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}_{\vec{C}^*})$, where for every element $\vec{C}$, $\vec{C}^* = \bigcap_{C \in \vec{C}} \mathrm{pa}_{\mathcal{G}}(C)$. Equivalently, $p(\vec{v} | \vec{w})$ is in the CMRF model of a CUG $\mathcal{G}(\vec{V}, \vec{W})$ if for every $\vec{v}, \vec{w}$, $p(\vec{v} | \vec{w}) = Z(\vec{w})^{-1} \prod_{\vec{C} \in \bar{\mathcal{C}}(\mathcal{G}_{\vec{V}})} \phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}_{\vec{C}^*})$, where $Z(\vec{w})$ is a *normalizing function*. A distribution $p(\vec{v} | \vec{w})$ is in the CBN model of a CDAG $\mathcal{G}(\vec{V}, \vec{W})$ if for every $\vec{v}, \vec{w}$, $p(\vec{v} | \vec{w}) = \prod_{V \in \vec{V}} p(\vec{v}_{\{V\}} | \vec{v}_{\mathrm{pa}_{\mathcal{G}}(V) \cap \vec{V}}, \vec{w}_{\mathrm{pa}_{\mathcal{G}}(V) \cap \vec{W}})$.

3 CONGENIALITY, NON-REDUNDANCE AND VARIATIONAL INDEPENDENCE

Consider a non-parametric (unrestricted) model on two binary random variables, A and B. A variationally independent parameterization for the likelihood of this model may be obtained by choosing any three joint distribution parameters, for example $p(A = 0, B = 0)$, $p(A = 0, B = 1)$, and $p(B = 1, A = 0)$. Variational independence means that any of these three parameters may be chosen to be any valid probability without restricting values of the other parameters.

An alternative parameterization for this likelihood is given by the following three parameters: $p(A = 0)$, $p(B = 0)$ and $p(A = 0, B = 0)$. This *Möbius parameterization* is variationally dependent, since allowed values of the marginal parameters $p(A = 0)$ and $p(B = 0)$ may be influenced by the value of the joint parameter $p(A = 0, B = 0)$. Despite being variationally dependent, this is a valid parameterization for categorical data and corresponds to a smooth parameter map given by the Möbius inversion formula. Variationally dependent parameterizations may require extensions to standard iterative optimization methods since any local step that changes parameter values (for instance via gradient descent) has to also ensure the newly chosen parameter values correspond to an element in the model. Möbius parameterizations have previously been provided for mixed graph models of categorical data (Evans and Richardson, 2014, 2019).

Now consider a model on two real-valued random variables A and B. A generalization of the Möbius parameterization would aim to specify certain parts $g_A(A)$, $g_B(B)$ and $g_{A,B}(A, B)$ of density functions $f(A)$, $f(B)$ and $f(A, B)$, respectively. An *incongenial* specification of objects g_A, g_B, and $g_{A,B}$ is one that imposes contradictory restrictions such that the resulting model (set of distributions) excludes the true distribution by construction. In other words, an incongenial specification results in a misspecified model. An extreme case of incongeniality is one where there is no valid density that satisfies all imposed restrictions, leading to the degenerate empty set model. Issues of congeniality were first described when discussing multiple imputation approaches to missing data problems (Meng, 1994).

A redundant specification of the model is one that uses "too many pieces," leading to non-identification. For example, a non-parametric model on two binary random variables A and B has dimension 3. A parameterization of this model that involves 4 unrestricted real-valued parameters leads to non-identification: there are infinitely many values of these 4 parameters that are consistent with any given distribution $p(A, B)$. The same issue may arise in semi-parametric models (despite such models being infinite dimensional). A redundant specification of such a model leads to a situation where many specifications are consistent with a particular observed data distribution in the model. In other words, a redundant model specification leads to a parameter map that is not bijective, which prevents identification of model specifications from the observed data distribution. While redundant specifications may be harmless in some applications, they present difficulties in settings where model parameters (and thus their identification) is of primary interest.

In the remainder of this paper, we will describe an approach that leads to congenial, non-redundant and variationally independent specifications of a wide class of graphical models, including those associated with undirected, directed, and mixed graphs.

4 LIKELIHOOD DECOMPOSITIONS

Chen (2007) has considered the following decomposition for an arbitrary conditional distribution $p(v_1, v_2|w)$, given an arbitrary set of reference values v_1^*, v_2^*:

$$p(v_1, v_2|\vec{w}) = \frac{p(v_1|v_2^*, \vec{w})p(v_2|v_1^*, \vec{w})OR(v_1, v_2; v_1^*, v_2^*|\vec{w})}{Z(\vec{w})},$$
(1)

where $OR(v_1, v_2; v_1^*, v_2^*|\vec{w}) = \frac{p(v_1, v_2|w)p(v_1^*, v_2^*|\vec{w})}{p(v_1, v_2^*|\vec{w})p(v_1^*, v_2|\vec{w})}$ is the conditional odds ratio function. In subsequent discussion, we will suppress the reference values from odds ratio functions, and denote them as e.g., $OR(v_1, v_2|\vec{w})$.

The following important result is derived by Chen (2003, 2007). In the interests of being self-contained, we reproduce the proof of this result in the Appendix.

Proposition 1 *The terms in the numerator of (1) are non-redundant and variationally independent.*

The decomposition in (1) can be generalized in a straightforward way to a multivariate conditional distribution $p(\vec{v}|\vec{w}) = p(v_1, \ldots, v_K|\vec{w})$ by inductively applying (1) while taking v_k as the first variable and $(v_1, \ldots, v_{k-1})$ as the second variable for all $k = 2, \ldots, K$. This yields the following decomposition for $p(v_1, \ldots, v_K|\vec{w})$:

$$Z(\vec{w})^{-1} \cdot \left(\prod_{k=1}^{K} p(v_k|\vec{v}_{-k}^*, \vec{w}) \right) \times$$

$$\times \left(\prod_{k=2}^{K} OR(v_k, (v_1, \ldots, v_{k-1})|v_{k+1}^*, \ldots, v_K^*, \vec{w}) \right), \quad (2)$$

where $\vec{v}_{-k} \equiv (v_1, \ldots, v_{k-1}, v_{k+1}, \ldots, v_K)$.

Two properties of the decomposition in (2) are worth noting. First, by inductively applying Proposition 1, it is straightforward to establish that all terms in the numerator of (2) are non-redundant, and variationally independent.

Second, this decomposition is valid for any order on variables in $\vec{V}$.

We now consider an alternative decomposition of a conditional distribution $p(\vec{v}|\vec{w})$. The presentation is a generalization of a decomposition of joint distributions $p(\vec{v})$ described by Lauritzen (1996). For any subset $\vec{C} \subseteq \vec{V}$, define

$$H_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}) \equiv \log p(\vec{v}_{\vec{C}}, \vec{v}^*_{\vec{V} \setminus \vec{C}}|\vec{w});$$

$$\tilde{\phi}_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}) \equiv \sum_{\vec{B} \subseteq \vec{C}} (-1)^{|\vec{C} \setminus \vec{B}|} H_{\vec{B}}(\vec{v}_{\vec{B}}, \vec{w}). \quad (3)$$

Since $H_{\vec{C}}$ and $\tilde{\phi}_{\vec{C}}$ are defined for every subset $\vec{C}$ of $\vec{V}$, they are related by the Möbius inversion formula, as follows:

$$\log p(\vec{v}|\vec{w}) = H_{\vec{V}}(\vec{v}, \vec{w}) = \sum_{\vec{C} \subseteq \vec{V}} \tilde{\phi}_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}). \quad (4)$$

For details, consult Appendix A.3 of Lauritzen (1996).

We can rewrite (4) as:

$$p(\vec{v}|\vec{w}) = \prod_{\vec{C} \subseteq \vec{V}} \exp\left\{ \tilde{\phi}_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}) \right\} = \prod_{\vec{C} \subseteq \vec{V}} \phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}),$$
$$(5)$$

where we define, for every $\vec{C}$:

$$\phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}) \equiv \exp\left\{ \tilde{\phi}_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}) \right\}$$

$$= \exp\left\{ \sum_{\vec{B} \subseteq \vec{C}} (-1)^{|\vec{C} \setminus \vec{B}|} H_{\vec{B}}(\vec{v}_{\vec{B}}, \vec{w}) \right\}. \quad (6)$$

We term (5) the Lauritzen decomposition of $p(\vec{v}|\vec{w})$. The CMRF factorization with respect to a CUG $\mathcal{G}(\vec{V}, \vec{W})$ is the Lauritzen decomposition where terms that do not correspond to cliques in $\mathcal{G}_{\vec{V}}$ disappear. It is simple to show that the CMRF factorization implies the CMRF version of the pairwise Markov property. In fact, for positive distributions the converse is also true, due to the following.

Theorem 1 (Hammersly-Clifford for conditional MRFs)
Assume a positive $p(\vec{v}|\vec{w})$ obeys the pairwise Markov property for a CUG $\mathcal{G}(\vec{V}, \vec{W})$. That is, for every $V \in \vec{V}$, $Z \in \vec{V} \cup \vec{W}$ such that Z is non-adjacent to V in $\mathcal{G}(\vec{V}, \vec{W})$, $p(v|(\vec{v} \cup \vec{w}) \setminus \{v\})$ is only a function of $(\vec{v} \cup \vec{w}) \setminus \{z\}$. Then $p(\vec{v}|\vec{w})$ factorizes with respect to $\mathcal{G}(\vec{V}, \vec{W})$. That is,

$$p(\vec{v}|\vec{w}) = \prod_{\vec{C} \subseteq \vec{c}(\mathcal{G}_{\vec{V}})} \phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}_{\vec{C}^*}), \quad (7)$$

where for every $\vec{C}$, $\vec{C}^* = \bigcap_{C \in \vec{C}} \text{pa}_{\mathcal{G}}(C)$, and $\phi_{\vec{C}}$ are terms defined as in (5).

The proof, which is a simple extension of the proof of the MRF version of the Theorem derived by Lauritzen (1996), is in the Appendix.

The two decompositions of a conditional distribution $p(\vec{v}|\vec{w})$ in the CMRF model of a CUG $\mathcal{G}(\vec{V}, \vec{W})$ have desirable and complementary properties. The Chen decomposition in (2) contains non-redundant and variationally independent terms, whereas the Lauritzen decomposition in (7) only has non-trivial terms corresponding to cliques of an induced subgraph $\mathcal{G}_{\vec{V}}$, thus allowing it to represent CMRF Markov restrictions. As we show in the next section, these two decompositions are related, allowing us to define general likelihoods for CMRF models by taking advantage of both properties at once.

5 CONNECTIONS BETWEEN CHEN AND LAURITZEN DECOMPOSITIONS

To illustrate the connection between the two decompositions described in the previous section, we compare the decompositions for a three-variable joint distribution $p(v_1, v_2, v_3)$. The Chen decomposition (2) is:

$$p(v_1, v_2, v_3) = Z^{-1} p(v_1|v_2^*, v_3^*) p(v_2|v_1^*, v_3^*) p(v_3|v_1^*, v_2^*)$$
$$OR(v_1, v_2|v_3^*) OR(v_3, (v_1, v_2)) \quad (8)$$

The Lauritzen decomposition (5) of $p(v_1, v_2, v_3)$ is:

$$(9)$$

The decompositions (8) and (9) are equivalent since $OR(v_1, v_3|v_2^*) OR(v_2, v_3|v_1^*) \frac{OR(v_1, v_2|v_3)}{OR(v_1, v_2|v_3^*)} = OR(v_3, (v_1, v_2))$ (see the Appendix), and consequently $\frac{p(v_1^*|v_2^*, v_3^*) p(v_2^*|v_1^*, v_3^*) p(v_3^*|v_1^*, v_2^*)}{p(v_1^*, v_2^*, v_3^*)} = Z$.

To see that the univariate conditional terms, such as $p(v_1|v_2^*, v_3^*)$, terms corresponding to two variable subsets of $\{V_1, V_2, V_3\}$, such as $\phi_{\{V_1, V_2\}}$, and the term corresponding to $\{V_1, V_2, V_3\}$, namely $\phi_{\{V_1, V_2, V_3\}}$ in (9) are non-redundant and variationally independent, we argue as follows. Note that the terms in the numerator of the decomposition (8) are non-redundant and variationally independent for any order (i, j, k) on indices of variables V_1, V_2, V_3. This implies the univariate conditional terms, any term of the form $OR(v_i, v_j|v_k^*)$, and the other OR terms grouped with $OR(v_k, (v_i, v_j))$ are non-redundant and variationally independent. This also implies these terms are non-redundant and variationally independent of the "remainder" of $OR(v_k, (v_i, v_j))$ (under any index order

i, j, k) once two variable OR terms are excluded, which is precisely $\phi_{\{V_1, V_2, V_3\}}$.

As the following results show, the relationship between the Chen and Lauritzen decompositions highlighted by this example is general.

Theorem 2 *Given a positive distribution $p(\vec{v}|\vec{w})$ on an ordered set of variables $\vec{V} \equiv \{V_1, \ldots, V_K\}$, for each $k = 2, \ldots, K$,*

$$OR(v_k, (v_1, \ldots, v_{k-1})|v_{k+1}^*, \ldots, v_K^*, \vec{w})$$
$$= \prod_{\{V_k\} \subset \vec{C} \subseteq \vec{V}_k} \phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}), \quad (10)$$

where $\vec{V}_k \equiv \{V_1, \ldots, V_k\}$, and $\phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w})$ is defined as in (6).

Proof: For any value vector $\{v_1, \ldots, v_K\}$, denote $\vec{v}_k \equiv \{v_1, \ldots, v_k\}$. Denote $OR(\vec{v}_{k-1}, v_k|v_{k+1}^*, \ldots, v_K^*, \vec{w})$ by η_k. By definition,

$$\eta_k = \frac{p(\vec{v}_{k-1}, v_k, v_{k+1}^*, \ldots, v_K^*|\vec{w}) p(\vec{v}_{k-1}^*, v_k^*, v_{k+1}^*, \ldots, v_K^*|\vec{w})}{p(\vec{v}_{k-1}, v_k^*, v_{k+1}^*, \ldots, v_K^*|\vec{w}) p(\vec{v}_{k-1}^*, v_k, v_{k+1}^*, \ldots, v_K^*|\vec{w})}$$

We claim that the following set of equalities hold.

$$\prod_{\{V_k\} \subset \vec{C} \subseteq \{V_1, \ldots, V_k\}} \phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w})$$
$$= \prod_{\{V_k\} \subset \vec{C} \subseteq \{V_1, \ldots, V_k\}} \exp\left\{ \sum_{\vec{B} \subseteq \vec{C}} (-1)^{|\vec{C} \setminus \vec{B}|} H_{\vec{B}}(\vec{v}_{\vec{B}}, \vec{w}) \right\}$$
$$(11)$$
$$= \exp\left\{ H_{\{V_1, \ldots, V_k\}} + H_\emptyset - H_{\{V_k\}} - H_{\{V_1, \ldots, V_{k-1}\}} \right\} = \eta_k,$$

where $H_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w})$ for any $\vec{C} \subseteq \vec{V}$ is defined as in (3).

The first equality holds by definition of $\phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w})$. To see that the last two equalities hold, we argue as follows. First, the number of subsets of $\{V_1, \ldots, V_k\}$ is 2^k, and the number of subsets of $\{V_1, \ldots, V_k\}$ that includes V_k is 2^{k-1}. Thus the number of terms in the first line of (11) is $2^{k-1} - 1$ (since the set $\{V_k\}$ does not have a corresponding term).

$H_{\{V_1, \ldots, V_k\}}$ occurs once, in the term corresponding to $\{V_1, \ldots, V_k\}$, with a positive sign.

Since all terms correspond to sets that must contain V_k, $H_{\{V_1, \ldots, V_{k-1}\}}$ occurs once, in the term corresponding to $\{V_1, \ldots, V_k\}$, with a negative sign, since $|\{V_1, \ldots, V_k\} \setminus \{V_1, \ldots, V_{k-1}\}| = 1$.

$H_\emptyset$ and $H_{\{V_k\}}$ occur in every term, and moreover $2^{k-1} - 2$ occurrences of these terms can be paired with opposite signs, and thus cancel. The last occurrence is in the term corresponding to $\{V_1, \ldots, V_k\}$, where $H_\emptyset$ occurs with a positive sign, and $H_{\{V_k\}}$ occurs with a negative sign.

Consider a set $\vec{B} \subseteq \{V_1, \ldots, V_k\}$ that is not equal to $\emptyset$, $\{V_k\}$, $\{V_1, \ldots, V_{k-1}\}$, or $\{V_1, \ldots, V_k\}$. The sets $\vec{C}$ such that $\{V_k\} \subset \vec{C} \subseteq \{V_1, \ldots, V_k\}$ where $\vec{B} \subseteq \vec{C}$ is $\vec{C}' \cup \vec{B}$, where $\vec{C}'$ is any subset of $\{V_1, \ldots, V_k\} \setminus \vec{B}$. It's clear that there is an even number of such subsets, specifically $2^{|\{V_1, \ldots, V_k\} \setminus \vec{B}|}$, and that the corresponding terms $H_{\vec{B}}$ cancel as they occur with alternating signs.

The last equality then holds by the definition of η_k. □

Corollary 1 (Lauritzen-Chen (L-C) decomposition)
Any positive distribution $p(v_1, \ldots, v_K|\vec{w})$ can be written as,

$$Z(\vec{w})^{-1} \left(\prod_{k=1}^{K} p(v_k|\vec{v}_{-k}^*, \vec{w}) \right) \times$$
$$\prod_{k=2}^{K} OR(v_k, (v_1, \ldots, v_{k-1})|v_{k+1}^*, \ldots, v_K^*, \vec{w}) \quad (12)$$

$$= \prod_{\vec{C} \subseteq \{V_1, \ldots, V_K\}} \phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}) \times$$
$$= \left(\frac{\phi_\emptyset(\vec{w})}{\prod_{k=1}^{K} p(v_k^*|\vec{v}_{-k}^*, \vec{w})} \right) \left(\prod_{k=1}^{K} p(v_k|\vec{v}_{-k}^*, \vec{w}) \right)$$
$$\prod_{\vec{C} \subseteq \{V_1, \ldots, V_K\}; |\vec{C}| \geq 2} \phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}). \quad (13)$$

Furthermore, $Z(\vec{w}) = \frac{\prod_{k=1}^{K} p(v_k^*|\vec{v}_{-k}^*, \vec{w})}{\phi_\emptyset(\vec{w})}$ for all $\vec{w}$, and all terms $p(v_k|\vec{v}_{-k}^*, \vec{w})$ (for all k) and $\phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w})$ if $|\vec{C}| \geq 2$ in (13) are non-redundant and variationally independent.

Proof: Equality of (12) and (13), as well as equality of $Z(\vec{w})$ and $\frac{\prod_{k=1}^{K} p(v_k^*|\vec{v}_{-k}^*, \vec{w})}{\phi_\emptyset(\vec{w})}$ follows by Theorem 2.

To show non-redundance and variational independence, we proceed by induction. Results of Chen (2003, 2007) show that all terms in the numerator of the right hand side of (12) are non-redundant and variationally independent. Since $OR(v_1, v_2|v_3^*, \ldots, v_k^*, \vec{w}) = \phi_{\{V_1, V_2\}}(\vec{v}_{\{V_1, V_2\}}, \vec{w})$, and since the (12) holds for any ordering on variables $\vec{V}$, we conclude that terms $p(v_k|\vec{v}_{-k}^*, \vec{w})$, $\phi_{\{V_i, V_j\}}(\vec{v}_{\{V_i, V_j\}}, \vec{w})$ are non-redundant and variationally independent for any i, j, k. This establishes the base case.

Assume we have shown that terms $p(v_k|\vec{v}_{-k}^*, \vec{w})$ (for any $k \in \{1, \ldots, K\}$) and $\phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w})$ (for all $\vec{C}$ such that $|\vec{C}| < i$) are non-redundant and variationally independent.

Let $\eta_i \equiv OR(v_i, (v_1, \ldots, v_{i-1})|v_{i+1}^*, \ldots, v_K^*, \vec{w})$. Under the given variable ordering, η_i is non-redundant and variationally independent of all other terms in (12). By Theorem 2, η_i contains a single term $\phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w})$ of size i, and other terms of size smaller than i. These terms, along with univariate conditionals in (12), are all non-redundant and

variationally independent of $\phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w})$ by the inductive hypothesis. Furthermore, $\phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w})$ is non-redundant and variationally independent of η_j for all $j > i$. Since variation independence in (12) is order-independent, the same argument can be repeated for any set $\vec{C}$ of size i, and the corresponding term $\phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w})$. This establishes the induction, and thus the result. $\qquad\square$

The Lauritzen-Chen decomposition immediately yields a non-redundant and variationally independent factorization of CMRFs, as follows.

Corollary 2 (Lauritzen-Chen (L-C) factorization of conditional MRFs) *For any positive $p(\vec{v}|\vec{w})$ in the CMRF model of a CUG $\mathcal{G}(\vec{V}, \vec{W})$,*

$$p(\vec{v}|\vec{w}) = \left(\frac{\phi_{\emptyset}(\vec{w}) \left(\prod_{k=1}^{K} p(v_k | \vec{v}^*_{\mathrm{nb}_{\mathcal{G}}(V_k)}, \vec{w}_{\mathrm{pa}_{\mathcal{G}}(V_k)}) \right)}{\prod_{k=1}^{K} p(v_k^* | \vec{v}^*_{\mathrm{nb}_{\mathcal{G}}(V_k)}, \vec{w}_{\mathrm{pa}_{\mathcal{G}}(V_k)})} \right)$$
$$\times \prod_{\vec{C} \in \mathcal{C}(\mathcal{G}_{\vec{V}}); |\vec{C}| \geq 2} \phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}_{\vec{C}^*}), \qquad (14)$$

where $\mathrm{nb}_{\mathcal{G}}(V_k)$ is the set of vertices in $\mathcal{G}$ with adjacencies to V_k via undirected edges, $\vec{C}^* = \bigcap_{C \in \vec{C}} \mathrm{pa}_{\mathcal{G}}(C)$ for every $\vec{C}$, and $\phi_{\vec{C}}$ are terms defined as in (6). Furthermore, all terms $p(v_k | \vec{v}^*_{\mathrm{nb}_{\mathcal{G}}(V_k)}, \vec{w}_{\mathrm{pa}_{\mathcal{G}}(V_k)})$ (for all k) and $\phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}_{\vec{C}^*})$ for all $\vec{C} \in \mathcal{G}(\mathcal{G}_{\vec{V}})$ where $|\vec{C}| \geq 2$ in (14) are non-redundant and variationally independent.

Proof: This follows immediately by Theorem 2 and Theorem 1. $\qquad\square$

In other words, (14) is obtained from (13) due to the fact that any term in (13) that does not correspond to a clique in $\mathcal{G}_{\vec{V}}$ becomes trivial (equal to 1) whenever $p(\vec{v}|\vec{w})$ is in a CMRF model of a CUG $\mathcal{G}(\vec{V}, \vec{W})$ (or an MRF model of a UG $\mathcal{G}(\vec{V})$ by taking $\vec{W} = \emptyset$).

Corollary 2 allows a non-redundant likelihood specification for parametric CMRF (or MRF) models that do not correspond to categorical or Gaussian data. In addition, flexible semi-parametric likelihoods for CMRFs (or MRFs) can be specified as well. However, while the terms $p(v_k | \vec{v}^*_{\mathrm{nb}_{\mathcal{G}}(V_k)}, \vec{w}_{\mathrm{pa}_{\mathcal{G}}(V_k)})$ and $\phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}_{\vec{V}^*})$ can be specified independently of each other, the fact that they are specific closed-form functionals of the distribution $p(\vec{v}|\vec{w})$ implies that they must satisfy certain restrictions in order for the overall model likelihood to be well-defined. The restrictions on terms in (14) are given by the following result.

Lemma 1 *For any positive $p(\vec{v}|\vec{w})$ in the CMRF model of a CUG $\mathcal{G}(\vec{V}, \vec{W})$,*

- *for any $V_k \in \vec{V}$, the term $p(v_k | \vec{v}^*_{\mathrm{nb}_{\mathcal{G}}(V_k)}, \vec{w}_{\mathrm{pa}_{\mathcal{G}}(V_k)})$ in (13) must be non-negative, and integrate to 1,*

- *for any $\vec{C} \subseteq \vec{V}$ such that $\vec{C} \in \mathcal{C}(\mathcal{G}_{\vec{V}})$ and $|\vec{C}| \geq 2$, the term $\phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}_{\vec{C}^*})$ in (13) must be non-negative and satisfy the following:*

$$(\forall V_i \in \vec{C}, \vec{v}_{\vec{C}\backslash\{V_i\}}, \vec{w}_{\vec{C}^*}) \ \ \phi_{\vec{C}}(v_i^*, \vec{v}_{\vec{C}\backslash\{V_i\}}, \vec{w}_{\vec{C}^*}) = 1. \qquad (15)$$

The terms are otherwise unrestricted.

Proof: The condition on terms $p(v_k | \vec{v}^*_{\mathrm{nb}_{\mathcal{G}}(V_k)}, \vec{w}_{\mathrm{pa}_{\mathcal{G}}(V_k)})$ follows since they are conditional probabilities, and on terms $\phi_{\{V_i, V_j\}}(\vec{v}_{\{V_i, V_j\}}, \vec{w}_{\vec{C}^*})$ follows by properties of conditional odds ratio functions.

Assume, by induction, the result holds for all $\phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}_{\vec{C}^*})$, where $|\vec{C}| = i$. To establish the conclusion we show that for all $\phi_{\vec{D}}(\vec{v}_{\vec{D}}, \vec{w}_{\vec{D}^*})$, where $\vec{D} = \vec{C} \cup \{V_i\}$, and for all $V_i \in \vec{D}$, $\phi_{\vec{D}}(\vec{v}_{\vec{D}}, \vec{w}_{\vec{D}^*}) = \frac{\phi_{\vec{C}}(\vec{v}_{\vec{C}}, v_i, \vec{w}_{\vec{C}^*})}{\phi_{\vec{C}}(\vec{v}_{\vec{C}}, v_i^*, \vec{w}_{\vec{C}^*})}$. We have:

$$\frac{\phi_{\vec{C}}(\vec{v}_{\vec{C}}, v_i, \vec{w}_{\vec{C}^*})}{\phi_{\vec{C}}(\vec{v}_{\vec{C}}, v_i^*, \vec{w}_{\vec{C}^*})} =$$

$$= \frac{\exp\left\{ \sum_{\vec{B} \subseteq \vec{C}} (-1)^{|\vec{C}\backslash\vec{B}|} H_{\vec{B}}(\vec{c}_{\vec{B} \cup \{V_i\}}, \vec{w}_{\vec{C}^*}) \right\}}{\exp\left\{ \sum_{\vec{B} \subseteq \vec{C}} (-1)^{|\vec{C}\backslash\vec{B}|} H_{\vec{B}}(\vec{c}_{\vec{B}}, \vec{w}_{\vec{C}^*}) \right\}}$$

$$= \frac{\exp\left\{ \sum_{\{R_i\} \subseteq \vec{B} \subseteq \vec{D}} (-1)^{|\vec{D}\backslash\vec{B}|} H_{\vec{B}}(\vec{c}_{\vec{B}}, \vec{w}_{\vec{C}^*}) \right\}}{\exp\left\{ \sum_{\{R_i\} \not\subseteq \vec{B} \subseteq \vec{D}} (-1)^{|\vec{D}\backslash\vec{B}|-1} H_{\vec{B}}(\vec{c}_{\vec{B}}, \vec{w}_{\vec{C}^*}) \right\}}$$

$$= \exp\left\{ \sum_{\vec{B} \subseteq \vec{D}} (-1)^{|\vec{D}\backslash\vec{B}|} H_{\vec{B}}(\vec{c}_{\vec{B}}, \vec{w}_{\vec{C}^*}) \right\} = \phi_{\vec{D}}(\vec{v}_{\vec{D}}, \vec{w}_{\vec{C}^*}).$$

That $\phi_{\vec{D}}(\vec{v}_{\vec{D}}, \vec{w}_{\vec{C}^*})$ is only a function of $\vec{v}_{\vec{D}}$ and $\vec{w}_{\vec{D}^*}$ (note that $\vec{D}^* \subseteq \vec{C}^*$ by definition) follows since $p(\vec{v}|\vec{w})$ is in the CMRF model for $\mathcal{G}(\vec{V}, \vec{W})$, and Theorem 1. $\qquad\square$

There are two ways to view this result. If we are given a conditional distribution $p(\vec{v}|\vec{w})$ in the CMRF model for $\mathcal{G}(\vec{V}, \vec{W})$, Lemma 1 will tell us the properties that pieces of the L-C factorization of $p(\vec{v}|\vec{w})$ must satisfy. Conversely, if we are interested in constructing a conditional distribution that lies in the CMRF model for a particular $\mathcal{G}(\vec{V}, \vec{W})$, Lemma 1 will tell us what restrictions to impose on pieces of an L-C factorization such that they form an element of the model. See also Lemmas 1 and 2 in prior work by Chen (2010).

6 LIKELIHOODS FOR BAYESIAN NETWORK MODELS

Bayesian network (BN) models admit natural likelihoods which parameterize each Markov factor $p(\vec{v}_{\{V\}} | \vec{v}_{\mathrm{pa}_{\mathcal{G}}(V)})$. Two distinct DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ may imply the same BN

model in the sense that d-separation statements in $\mathcal{G}_1$ and $\mathcal{G}_2$ imply the same list of conditional independences. In this case, DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are said to lie in the same Markov equivalence class. An elegant result states that $\mathcal{G}_1$ and $\mathcal{G}_2$ are equivalent in this sense if and only if they share edge adjacencies, and the same unshielded colliders (vertex triplets of the form $A \rightarrow C \leftarrow B$, where A and B are not adjacent) (Verma and Pearl, 1990).

Likelihoods that parameterize $p(\vec{v}_{\{V\}}|\vec{v}_{\mathrm{pa}_{\mathcal{G}}(V)})$ terms for Gaussian or categorical data are coherent for the Markov structure implied by the BN model in the sense that they will coincide for distinct DAGs within the Markov equivalence class. However, general likelihoods do not have this property. which complicates their use in applications where this coherence with respect to Markov structure is desirable, such as score based model selection algorithms that use data to select the best Markov model, or efficient semi-parametric estimators, with efficiency gains implied by Markov restrictions imposed on the data. Huang et al. (2018) provide an excellent discussion of this issue in the context of model selection.

We show how to use the Lauritzen-Chen likelihood to derive general parametric and semi-parametric likelihoods for DAG models that coincide within the Markov equivalence class. We will do so by imposing the likelihood on a special type of mixed graph called the *essential graph* (Andersson et al., 1997), which represents the Markov equivalent class of DAGs. Essential graphs are a special case of mixed graph graphical models known as chain graph models (Lauritzen, 1996), and we will derive likelihoods for them first. We will restriction attention to chain graph models under the Lauritzen-Wermuth-Frydenberg (LWF) interpretation (Lauritzen, 1996).

A chain graph (CG) is a mixed graph containing directed and undirected edges with the property that no partially directed cycles are present. A partially directed cycle is a sequence of consecutive edges that contains at least one directed edge with the property that undirected edges in this sequence can be oriented in such a way as to create a directed cycle.

Given a CG $\mathcal{G}$, a *block* is an undirected connected component. The set of blocks $\mathcal{B}(\mathcal{G})$ in a CG $\mathcal{G}$ form a partition of vertices $\vec{V}$ in $\mathcal{G}$. The CG model under the Lauritzen-Wermuth-Frydenberg (LWF) interpretation may be defined by the following factorization which generalizes the BN and MRF factorizations.

$$p(\vec{v}) = \prod_{\vec{B} \in \mathcal{B}(\mathcal{G})} p(\vec{v}_{\vec{B}}|\vec{v}_{\mathrm{pa}_{\mathcal{G}}(\vec{B})})$$

$$= \prod_{\vec{B} \in \mathcal{B}(\mathcal{G})} \left(\prod_{\vec{C} \in \mathcal{C}(\mathcal{G}_{\vec{B}})} \phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{v}_{\vec{C}^*}) \right), \qquad (16)$$

where $\vec{C}^* = \bigcap_{C \in \vec{C}} \mathrm{pa}_{\mathcal{G}}(C)$ for every $\vec{C} \in \mathcal{G}_{\vec{B}}$ and $\vec{B} \in$ $\mathcal{B}(\mathcal{G})$.

In other words, the LWF CG factorization is a two level factorization. The outer factorization of $p(\vec{v})$ resembles a DAG factorization, but formulated on elements of $\mathcal{B}(\mathcal{G})$ rather than individual vertices. The inner factorization for every term $p(\vec{v}_{\vec{B}}|\vec{v}_{\mathrm{pa}_{\mathcal{G}}(\vec{B})})$ of the outer factorization is a CMRF factorization with respect to the conditional graph $\mathcal{G}(\vec{B}, \mathrm{pa}_{\mathcal{G}}(\vec{B}))$ obtained from the CG $\mathcal{G}$ by restricting to vertices in $\vec{B} \cup \mathrm{pa}_{\mathcal{G}}(\vec{B})$, dropping all edges among $\mathrm{pa}_{\mathcal{G}}(\vec{B})$, and treating $\mathrm{pa}_{\mathcal{G}}(\vec{B})$ as fixed vertices.

This immediately implies the following result.

Corollary 3 (Lauritzen-Chen (L-C) factorization of LWF chain graph models) *For any positive $p(\vec{v})$ that factorizes with respect to a CG $\mathcal{G}$, $p(\vec{v})$ also obeys the following factorization:*

$$\prod_{\vec{B} \in \mathcal{B}(\mathcal{G})} \frac{\phi_{\emptyset}(\vec{v}_{\mathrm{pa}_{\mathcal{G}}(\vec{B})}) \prod_{B \in \vec{B}} p(\vec{v}_{\{B\}}|\vec{v}^*_{\mathrm{nb}_{\mathcal{G}_{\vec{B}}}(B)}, \vec{v}_{\mathrm{pa}_{\mathcal{G}}(B)})}{\prod_{B \in \vec{B}} p(\vec{v}^*_{\{B\}}|\vec{v}^*_{\mathrm{nb}_{\mathcal{G}_{\vec{B}}}(B)}, \vec{v}_{\mathrm{pa}_{\mathcal{G}}(B)})}$$

$$\times \prod_{\vec{C} \in \mathcal{C}(\mathcal{G}_{\vec{B}}); |\vec{C}| \geq 2} \phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{v}_{\vec{C}^*}) \quad (17)$$

where for every $\vec{B} \in \mathcal{B}(\mathcal{G})$, and every $\vec{C} \in \mathcal{C}(\mathcal{G}_{\vec{B}})$, the term $\phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{v}_{\vec{C}^*})$ is defined as in (3) and (6) using $H_{\vec{D}}(\vec{v}_{\vec{D}}, \vec{v}_{\vec{C}^*}) = \log p(\vec{v}_{\vec{D}}, \vec{v}^*_{\vec{B} \setminus \vec{D}}|\vec{v}_{\mathrm{pa}_{\mathcal{G}}(\vec{B})})$ for every $\vec{D} \subseteq \vec{C}$.

Furthermore, all terms $\{\phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{v}_{\vec{C}^*}) : \vec{C} \in \mathcal{C}(\mathcal{G}_{\vec{B}}), \vec{B} \in \mathcal{B}(\mathcal{G})\}$, and $\{p(\vec{v}_{\{B\}}|\vec{v}^*_{\mathrm{nb}_{\mathcal{G}_{\vec{B}}}(B)}, \vec{v}_{\mathrm{pa}_{\mathcal{G}}(B)}) : B \in \vec{B} \in \mathcal{B}(\mathcal{G})\}$ in (17) are non-redundant and variationally independent.

Proof: This follows from the fact that the inner LWF CG factorization is a CMRF factorization, Corollary 2, and that terms in the outer CG factorization are non-redundant and variationally independent. □

Given a DAG $\mathcal{G}$, an *essential graph* $\mathcal{G}^e$ is a mixed graph with a directed edge $V_i \rightarrow V_j$ whenever every DAG in the Markov equivalence class contains the edge $V_i \rightarrow V_j$, and an undirected edge $V_i - V_j$ whenever V_i and V_j are adjacent in every DAG in the Markov equivalence class, but the orientation of the edge differs for different elements of the equivalence class. Results by Andersson et al. (1997) show that the essential graph $\mathcal{G}^e$ is a CG, and give a polynomial time algorithm for its construction, given any DAG $\mathcal{G}$.

The following result is an immediate consequence of the results derived by Andersson et al. (1997).

Proposition 2 *For any DAG $\mathcal{G}$, a distribution $p(\vec{v})$ obeys the DAG factorization respect to $\mathcal{G}$ if and only if $p(\vec{v})$ obeys the LWF CG factorization with respect to $\mathcal{G}^e$.*

Proposition 2 and Corollary 3 immediately yield coherent likelihoods for DAG models that are guaranteed to coincide for distinct DAGs in the same equivalence class. As was the case for MRFs, these likelihoods can be specified for both parametric and semi-parametric models. Such likelihoods have previously been derived in the special case of categorical data (Castelo and Perlman, 2004).

7 EXAMPLES

Figure 1 (a) shows an MRF with two restrictions: $(A \perp\!\!\!\perp D|B,C)$, and $(B \perp\!\!\!\perp C|A,D)$. The L-C factorization of any positive distribution $p(a,b,c,d)$ in this MRF model is:

$$Z^{-1}p(a|b^*,c^*)p(b|a^*,d^*)p(c|a^*,d^*)p(d|b^*,c^*)$$
$$\underbrace{OR(a,b|c^*,d^*)}_{\phi_{\{A,B\}}}\underbrace{OR(a,c|b^*,d^*)}_{\phi_{\{A,C\}}}\underbrace{OR(b,d|a^*,c^*)}_{\phi_{\{B,D\}}}$$
$$\underbrace{OR(c,d|a^*,b^*)}_{\phi_{\{C,D\}}}$$

where

$$Z = \frac{p(a^*|b^*,c^*)p(b^*|a^*,d^*)p(c^*|a^*,d^*)p(d^*|b^*,c^*)}{p(a^*,b^*,c^*,d^*)}.$$

Note that there are no terms corresponding to any subset of $\{A,B,C,D\}$ of size 3 or 4, as no such set corresponds to a clique in the graph in Figure 1 (a). This is the simplest MRF model not observationally equivalent to any DAG model. This model may be generalized to MRFs with graphs forming a square grid. Square grid MRFs have been used as spin models in physics (Nishimori, 2001), the original application of MRF models due to Ising (1925), and in computer vision to model images composed of pixels (Szeliski, 2022). An example of such an MRF with the corresponding factorization is given in the Appendix.

Figures 1 (b), (c), and (d) show three (out of eighteen) distinct DAGs that are in the same Markov equivalence class, and thus imply the same BN model. The essential graph corresponding to this Markov equivalence class is shown in Figure 1 (e). Note that all edges that differ in their orientation in Figures 1 (b), (c), and (d) are represented by undirected edges in Figure 1 (e). The factorization for $p(a,b,c,d,e,f)$ associated with Figure 1 (e) yields a likelihood that obeys Markov restrictions in the equivalence class, and coincides for all DAGs in the class by construction. It is given as follows:

$$Z_1^{-1}p(a|b^*,c^*)p(c|a^*)p(b|a^*)OR(a,b|c^*)OR(a,c|b^*)\times$$
$$Z_2(b,c)^{-1}p(d|b,c,e^*,f^*)p(e|b,c,d^*,f^*)p(f|b,c,d^*,e^*)\times$$
$$OR(d,e|b,c,f^*)OR(d,f|b,c,e^*)OR(e,f|b,c,d^*)\times$$
$$\frac{OR(d,e|f,b,c)}{OR(d,e|f^*,b,c)}.$$

Additional examples of the L-C factorization in more complex models, arising in areas such as philogenetic analysis and computer vision, are discussed in the Appendix.

8 DISCUSSION

In this paper we have shown that decompositions of joint distributions described by Chen (2003, 2007, 2010) and Lauritzen (1996) are closely connected. This allowed us to show that the clique factorization for Markov random field graphical models may be specified using univariate conditional distribution terms, and alternate product terms that arise from the Möbius inversion formula and that generalize the odds ratio function. We show that these terms are *variationally independent*, and *non-redundant*. This result is also generalized to conditional Markov random field models, directed acyclic graph (DAG) models, and chain graph (CG) models. In particular, these results allow parametric and semi-parametric likelihoods for DAG models to be specified that coincide, by construction, among all elements of the Markov equivalence class of any DAG model. This solves a long-standing open issue in the score based causal discovery literature (Huang et al., 2018).

The proposed MRF likelihoods may be viewed as a generalization of hierarchical log-linear models for categorical data (Lauritzen, 1996), where univariate conditional terms $p(v_k|\vec{v}_{-k},\vec{w})$ generalize main effect parameters, and generalized odds ratio terms $\phi_{\vec{C}}(\vec{v}_{\vec{C}},\vec{w}_{\vec{C}^*})$ generalize higher order ($|\vec{C}|$-way) interaction parameters in such models.

A natural semi-parametric modeling approach for likelihoods we describe is to use flexible methods for univariate conditional terms, such as those based on kernel regressions (Nadaraya, 1964; Watson, 1964), and parametric forms for generalized interaction terms, for example the transformed linear generalized odds ratio model: $\log \phi_{\vec{C}}(\vec{v}_{\vec{C}},\vec{w}_{\vec{C}^*};\gamma) = \gamma \prod_{C \in \vec{C}}(g_C(\vec{v}_C,\vec{w}_{\vec{C}^*}) - g_C(\vec{v}_C^*,\vec{w}_{\vec{C}^*}))$ for fixed or unknown functions g_C. Such likelihoods may be optimized by iterative methods similarly to other semi-parametric likelihood models, such as the projection pursuit model (Friedman and Stuetzle, 1981).

We note that the proposed likelihood differs from copula models (Sklar, 1959). The latter aims to link together univariate *marginal* distribution by a (typically parametric) object called the *copula* that captures joint behavior of the random variables in the model. On the other hand, the univariate terms in the proposed likelihoods are univariate *conditional* distributions, with the other terms capturing interactions among variables.

If the terms in the proposed likelihoods are kept unrestricted, aside from necessary restrictions imposed by Lemma 1, the result yields a useful view on the tangent space of the corresponding Markov model, which is useful
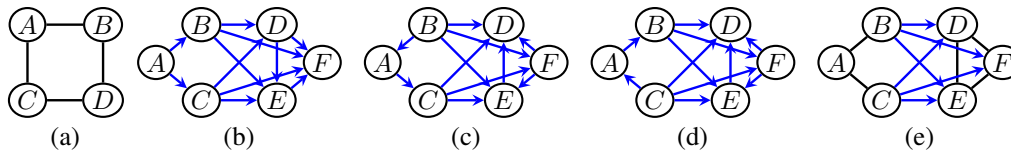
Figure 1: (a) An MRF model with two restrictions: $(A \perp\!\!\!\perp D|B, C)$, and $(B \perp\!\!\!\perp C|A, D)$. (b), (c), (d) Three distinct DAGs in the same Markov equivalence class corresponding to the Bayesian network (BN) model with restrictions $(B \perp\!\!\!\perp C|A)$, $(D, E, F \perp\!\!\!\perp A|B, C)$. (e) The essential graph CG corresponding to the Markov equivalence class containing DAGs in (b), (c), and (d).

for deriving estimators based on influence functions that attain the semi-parametric efficiency bound. Indeed, a special case of the Chen decomposition for a particular class of Markov random fields has already been used to derive an efficient influence function in a missing data model (Malinsky et al., 2021).

Additional work is needed to obtain practical estimators for graphical model parameters using the proposed likelihoods. This is because the likelihoods we describe inherit the usual difficulties of known parametric likelihoods for MRFs and CMRFs: the computation of the normalizing constant Z or normalizing function $Z(\vec{w})$. Existing approaches to this problem include approximation of these objects by numerical integration, sum-product algorithms in sufficiently sparse graphs (Koller and Friedman, 2009), variational inference (Wainwright and Jordan, 2008), or reformulation of the problem via composite likelihoods (Varin et al., 2011) where the need for normalization disappears. For a review of these methods, see (Wainwright and Jordan, 2008). Application of these approaches to the proposed likelihoods is an interesting area of future work. In particular, the application of sum-product algorithms would entail developing classes of likelihoods where intermediate objects obtained by marginalization and products of clique factors maintain a particular form.

Acknowledgements

References

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics*, 25(2):505–541.

Bilmes, J. and Bartels, C. (2005). Graphical model architectures for speech recognition. *IEEE Signal Processing Magazine*, 22:89–100.

Castelo, R. and Perlman, M. (2004). Learning essential graph Markov models from data. In Gamez, J., Moral, S., and Salmeron, A., editors, *Advances in Bayesian Networks. Studies in Fuzziness and Soft Computing, volume 146*, pages 255–269. Springer, Berlin, Heidelberg.

Chen, H. Y. (2003). A note on the prospective analysis of outcome-dependent samples. *Journal of the Royal Statistical Society (Series B)*, 65(2):575–584.

Chen, H. Y. (2007). A semiparametric odds ratio model for measuring association. *Biometrics*, 63:413–421.

Chen, H. Y. (2010). Compatibility of conditionally specified models. *Statistics and Probability Letters*, 80(7-8):670–677.

Chickering, D. M. (2002). Learning equivalence classes of Bayesian network structures. *Journal of Machine Learning Research*, 2:445–498.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (2002). *Biological Sequence Analysis*. Cambridge University Press.

Evans, R. J. and Richardson, T. S. (2014). Markovian acyclic directed mixed graphs for discrete data. *Annals of Statistics*, pages 1–30.

Evans, R. J. and Richardson, T. S. (2019). Smooth, identifiable supermodels of discrete DAG models with latent variables. *Bernoulli*, 25(2):848–876.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823.

Friedman, N., Ninio, M., Pe'er, I., and Pupko, T. (2002). A structural EM algorithm for phylogenetic inference. *Journal of Computational Biology*, 9(2):331–353.

Huang, B., Zhang, K., Lin, Y., Scholkopf, B., and Glymour, C. (2018). Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1551–1560.

Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift fur Physik*, 31:253–258.

Ji, Q. (2019). *Probabilistic Graphical Models for Computer Vision*. Elsevier.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, Cambridge MA.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML-01)*, pages 282 – 289. Morgan Kaufmann.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford, U.K.: Clarendon.

Maathuis, M., Drton, M., Lauritzen, S., and Martin Wainwright, e. (2019). *Handbook of Graphical Models*. CRC Press.

Malinsky, D., Shpitser, I., and Tchetgen Tchetgen, E. J. (2021). Semiparametric inference for nonmonotone missing-not-at-random data: the no self-censoring model. *Journal of the American Statistical Association*, pages 1–9.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9:538–558.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9(1):141–142.

Nishimori, H. (2001). *Statistical Physics of Spin Glasses and Information Processing: An Introduction*. Oxford: Oxford University Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan and Kaufmann, San Mateo.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.

Sherington, D. and Kirkpatrick, S. (1975). Solvable model of a spin-glass. *Physical Review Letters*, 35(35):1792–1796.

Sklar, A. (1959). Fonctions de repartition a *n* dimensions et leurs marges. *Publ. Inst. Statist. Univ.*, 8:229–231.

Smith, N. A. (2011). Linguistic structure prediction. *Synthesis Lectures on Human Language Technologies*, 4(2):1–274.

Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, Prediction, and Search*. Springer Verlag, New York, 2 edition.

Szeliski, R. (2022). *Computer Vision: Algorithms and Applications*. Springer Nature.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. Springer-Verlag New York, 1st edition edition.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42.

Verma, T. S. and Pearl, J. (1990). Equivalence and synthesis of causal models. Technical Report R-150, Department of Computer Science, University of California, Los Angeles.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A*, 26(4):359–372.

A   PROOFS

Proposition 1 *The joint distribution $p(v_1, v_2 | \vec{w})$ admits the following decomposition.*

$$p(v_1, v_2 | \vec{w}) = \frac{p(v_1 | v_2^*, \vec{w}) p(v_2 | v_1^*, \vec{w}) OR(v_1, v_2 | \vec{w})}{Z(\vec{w})}$$

Moreover, all terms in the numerator of this decomposition are non-redundant *and* variationally independent.

Proof: Here we follow the structure of the arguments presented by Chen (2003), as well as the main body and the Appendix of the paper by Chen (2007). We generalize the argument of Chen (2003) somewhat to apply to conditional distributions. In the following, summation may be replaced by integration where appropriate.

We will first show that

$$p(v_1 | v_2, \vec{w}) = \frac{p(v_1 | v_2^*, \vec{w}) OR(v_1, v_2 | \vec{w})}{\sum_{v_1} p(v_1 | v_2^*, \vec{w}) OR(v_1, v_2 | \vec{w})},$$

$$p(v_2 | v_1, \vec{w}) = \frac{p(v_2 | v_1^*, \vec{w}) OR(v_1, v_2 | \vec{w})}{\sum_{v_2} p(v_2 | v_1^*, \vec{w}) OR(v_1, v_2 | \vec{w})}.$$

Next, we will show that $p(v_1 | v_2^*, \vec{w})$ may be specified in a non-redundant and variationally independent way from any function of v_2 and $\vec{w}$ and in particular from $p(v_2 | v_1^*, \vec{w})$. Symmetrically we will show that $p(v_2 | v_1^*, \vec{w})$ may be specified in a non-redundant and variationally independent way from any function of v_1 and $\vec{w}$ and in particular from $p(v_1 | v_2^*, \vec{w})$. Finally, we will derive that the numerator of the decomposition are non-redundant and variationally independent.

We start by noting that by Bayes rule, $p(v_1 | v_2^*, \vec{w}) = \frac{p(v_2^* | v_1, \vec{w}) p(v_1 | \vec{w})}{\sum_{v_1} p(v_2^* | v_1, \vec{w}) p(v_1 | \vec{w})}$. This implies

$$p(v_1 | \vec{w}) = \frac{p(v_1 | v_2^*, \vec{w}) \sum_{v_1} p(v_2^* | v_1 | \vec{w}) p(v_1 | \vec{w})}{p(v_2^* | v_1, \vec{w})}$$

$$= \frac{p(v_1 | v_2^*, \vec{w}) \sum_{v_1} p(v_2^* | v_1, \vec{w}) \left(p(v_1 | v_2^*, \vec{w}) p(v_2^* | \vec{w}) / p(v_2^* | v_1, \vec{w}) \right)}{p(v_2^* | v_1, \vec{w})}$$

$$= \frac{p(v_1 | v_2^*, \vec{w})}{p(v_2^* | v_1, \vec{w})} \sum_{v_1} p(v_1 | v_2^*, \vec{w}) p(v_2^* | \vec{w})$$

$$= \frac{p(v_1 | v_2^*, \vec{w})}{p(v_2^* | v_1, \vec{w})} \Big/ \sum_{v_1} \frac{p(v_1 | v_2^*, \vec{w})}{p(v_2^* | v_1, \vec{w})},$$

where the last equality follows from the fact that

$$\sum_{v_1} p(v_1 | v_2^*, \vec{w}) p(v_2^* | \vec{w}) \times \sum_{v_1} \frac{p(v_1 | v_2^*, \vec{w})}{p(v_2^* | v_1)} = \sum_{v_1} \frac{p(v_1 | v_2^*, \vec{w}) p(v_2^* | \vec{w})}{p(v_2^* | v_1, \vec{w})} = \sum_{v_1} \frac{p(v_2^*, v_1 | \vec{w})}{p(v_2^* | v_1, \vec{w})} = \sum_{v_1} p(v_1 | \vec{w}) = 1.$$

Applying Bayes rule again, we see that

$$p(v_1 | v_2, \vec{w}) = \frac{p(v_2 | v_1, \vec{w}) p(v_1 | \vec{w})}{\sum_{v_1} p(v_2 | v_1, \vec{w}) p(v_1 | \vec{w})} = \frac{p(v_2 | v_1, \vec{w}) \frac{p(v_1 | v_2^*, \vec{w})}{p(v_2^* | v_1, \vec{w})} \Big/ \sum_{v_1} \frac{p(v_1 | v_2^*, \vec{w})}{p(v_2^* | v_1, \vec{w})}}{\sum_{v_1} p(v_2 | v_1, \vec{w}) \frac{p(v_1 | v_2^*, \vec{w})}{p(v_2^* | v_1, \vec{w})} \Big/ \sum_{v_1} \frac{p(v_1 | v_2^*, \vec{w})}{p(v_2^* | v_1, \vec{w})}} = \frac{p(v_2 | v_1, \vec{w}) \frac{p(v_1 | v_2^*, \vec{w})}{p(v_2^* | v_1, \vec{w})}}{\sum_{v_1} p(v_2 | v_1, \vec{w}) \frac{p(v_1 | v_2^*, \vec{w})}{p(v_2^* | v_1, \vec{w})}}.$$

Since the conditional odds ratio function $OR(v_1, v_2; v_1^*, v_2^* | \vec{w})$ given reference values v_1^*, v_2^* is defined as $\frac{p(v_2 | v_1, \vec{w}) p(v_2 | v_1^*, \vec{w})}{p(v_2^* | v_1, \vec{w}) p(v_2 | v_1^*, \vec{w})}$ (or equivalently as $\frac{p(v_1 | v_2, \vec{w}) p(v_1^* | v_2^*, \vec{w})}{p(v_1^* | v_2, \vec{w}) p(v_1 | v_2^*, \vec{w})}$), we have:

$$p(v_1 | v_2, \vec{w}) = \frac{p(v_2 | v_1, \vec{w}) \frac{p(v_1 | v_2^*, \vec{w})}{p(v_2^* | v_1, \vec{w})}}{\sum_{v_1} p(v_2 | v_1, \vec{w}) \frac{p(v_1 | v_2^*, \vec{w})}{p(v_2^* | v_1, \vec{w})}} = \frac{p(v_2 | v_1, \vec{w}) \frac{p(v_1 | v_2^*, \vec{w})}{p(v_2^* | v_1, \vec{w})} \frac{p(v_2^* | v_1^*, \vec{w})}{p(v_2 | v_1^*, \vec{w})}}{\frac{p(v_2^* | v_1^*, \vec{w})}{p(v_2 | v_1^*, \vec{w})} \sum_{v_1} p(v_2 | v_1, \vec{w}) \frac{p(v_1 | v_2^*, \vec{w})}{p(v_2^* | v_1, \vec{w})}} = \frac{p(v_1 | v_2^*, \vec{w}) OR(v_1, v_2 | \vec{w})}{\sum_{v_1} p(v_1 | v_2^*, \vec{w}) OR(v_1, v_2 | \vec{w})},$$

$$\tag{18}$$

where we shorten the conditional odds ratio function notation to $OR(v_1, v_2|\vec{w})$ by keeping the reference values implicit. Since this argument is completely symmetric with respect to v_1 and v_2, we also have:

$$p(v_2|v_1, \vec{w}) = \frac{p(v_2|v_1^*, \vec{w})OR(v_1, v_2|\vec{w})}{\sum_{v_2} p(v_2|v_1^*, \vec{w})OR(v_1, v_2|\vec{w})}. \tag{19}$$

Next, note that (18), (19) and the chain rule of probability imply:

$$p(v_1, v_2|\vec{w}) = \frac{p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})}{\sum_{v_1} p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})}p(v_2|\vec{w}) = \frac{p(v_2|v_1^*, \vec{w})OR(v_1, v_2|\vec{w})}{\sum_{v_2} p(v_2|v_1^*, \vec{w})OR(v_1, v_2|\vec{w})}p(v_1|\vec{w}). \tag{20}$$

Thus,

$$p(v_1|\vec{w}) = \sum_{v_2} p(v_1, v_2|\vec{w}) = p(v_1|v_2^*, \vec{w}) \sum_{v_2} \frac{OR(v_1, v_2|\vec{w})}{\sum_{v_1} p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})}p(v_2|\vec{w}). \tag{21}$$

Next, note that (20) and (21) imply:

$$\frac{p(v_2|v_1^*, \vec{w})}{\sum_{v_2} p(v_2|v_1^*, \vec{w})OR(v_1, v_2|\vec{w})} = \frac{p(v_1|v_2^*, \vec{w})}{\sum_{v_1} p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})} \frac{p(v_2|\vec{w})}{p(v_1|\vec{w})}$$

$$= \frac{p(v_2|\vec{w})}{\sum_{v_1} p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})} \Big/ \sum_{v_2} \frac{OR(v_1, v_2|\vec{w})p(v_2|\vec{w})}{\sum_{v_1} p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})}.$$

This, in turn, implies that:

$$p(v_2|\vec{w}) = p(v_2|v_1^*, \vec{w}) \frac{\sum_{v_1} p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})}{\sum_{v_2} p(v_2|v_1^*, \vec{w})OR(v_1, v_2|\vec{w})} \sum_{v_2} \frac{OR(v_1, v_2|\vec{w})p(v_2|\vec{w})}{\sum_{v_1} p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})}. \tag{22}$$

Since $\sum_{v_2} p(v_2|\vec{w}) = 1$,

$$\sum_{v_2} p(v_2|v_1^*, \vec{w}) \frac{\sum_{v_1} p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})}{\sum_{v_2} p(v_2|v_1^*, \vec{w})OR(v_1, v_2|\vec{w})} \sum_{v_2} \frac{OR(v_1, v_2|\vec{w})p(v_2|\vec{w})}{\sum_{v_1} p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})} = 1$$

$$\Rightarrow \sum_{v_1, v_2} p(v_2|v_1^*, \vec{w})p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w}) = \frac{\sum_{v_2} p(v_2|v_1^*, \vec{w})OR(v_1, v_2|\vec{w})}{\sum_{v_2} \frac{OR(v_1, v_2|\vec{w})p(v_2|\vec{w})}{\sum_{v_1} p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})}}. \tag{23}$$

Equations (22) and (23) together imply:

$$p(v_2|\vec{w}) = \frac{p(v_2|v_1^*, \vec{w}) \sum_{v_1} p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})}{\sum_{v_1, v_2} p(v_2|v_1^*, \vec{w})p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})}, \tag{24}$$

Rearranging (24) yields:

$$p(v_2|v_1^*, \vec{w}) = \frac{p(v_2|\vec{w}) \sum_{v_1, v_2} p(v_2|v_1^*, \vec{w})p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})}{\sum_{v_1} p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})}.$$

Since $\sum_{v_2} p(v_2|v_1^*, \vec{w}) = 1$,

$$\sum_{v_2} \frac{p(v_2|\vec{w}) \sum_{v_1, v_2} p(v_2|v_1^*, \vec{w})p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})}{\sum_{v_1} p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})} = 1$$

$$\Rightarrow \sum_{v_2} \frac{p(v_2|\vec{w})}{\sum_{v_1} p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})} = \frac{1}{\sum_{v_1, v_2} p(v_2|v_1^*, \vec{w})p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})}$$

$$\Rightarrow p(v_2|v_1^*, \vec{w}) = \frac{p(v_2|\vec{w})}{\sum_{v_1} p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})} \Big/ \sum_{v_2} \frac{p(v_2|\vec{w})}{\sum_{v_1} p(v_1|v_2^*, \vec{w})OR(v_1, v_2|\vec{w})}. \tag{25}$$

Equations (24) and (25) imply that $p(v_2|\vec{w})$ and $p(v_2|v_1^*, \vec{w})$ are connected by an (algebraic) bijective mapping. Thus, since models for $p(v_1|v_2^*, \vec{w})$ and $p(v_2|\vec{w})$ admit a non-redundant and variationally independent parameterization, then so do $p(v_1|v_2^*, \vec{w})$ and $p(v_2|v_1^*, \vec{w})$.

A symmetric argument may be used to establish the following identities:

$$p(v_1|\vec{w}) = \frac{p(v_1|v_2^*, \vec{w}) \sum_{v_2} p(v_2|v_1^*, \vec{w}) OR(v_1, v_2|\vec{w})}{\sum_{v_1, v_2} p(v_1|v_2^*, \vec{w}) p(v_2|v_1^*, \vec{w}) OR(v_1, v_2|\vec{w})}, \tag{26}$$

and

$$p(v_1|v_2^*, \vec{w}) = \frac{p(v_1|\vec{w})}{\sum_{v_2} p(v_2|v_1^*, \vec{w}) OR(v_1, v_2|\vec{w})} \Big/ \sum_{v_1} \frac{p(v_1|\vec{w})}{\sum_{v_2} p(v_2|v_1^*, \vec{w}) OR(v_1, v_2|\vec{w})}. \tag{27}$$

Combining either (26) or (24) and (20) yields:

$$\begin{aligned} p(v_1, v_2|\vec{w}) &= \frac{p(v_1|v_2^*, \vec{w}) p(v_2|v_1^*, \vec{w}) OR(v_1, v_2|\vec{w})}{\sum_{v_1, v_2} p(v_1|v_2^*, \vec{w}) p(v_2|v_1^*, \vec{w}) OR(v_1, v_2|\vec{w})} \\ &= \frac{p(v_1|v_2^*, \vec{w}) p(v_2|v_1^*, \vec{w}) OR(v_1, v_2|\vec{w})}{Z(\vec{w})}. \end{aligned}$$

Finally, we note that $p(v_1|v_2, \vec{w})$ and $p(v_2|\vec{w})$ admit a non-redundant and variationally independent parameterization. However, (18) suggests $p(v_1|v_2, \vec{w})$ may be parameterized by $p(v_1|v_2^*, \vec{w})$ and $OR(v_1, v_2|\vec{w})$, and above argument suggests $p(v_2|\vec{w})$ has a bijective map with $p(v_2|v_1^*, \vec{w})$. Thus, $p(v_2|v_1^*, \vec{w})$ admits a parameterization that is non-redundant and variationally independent for $p(v_1|v_2^*, \vec{w})$ and $OR(v_1, v_2|\vec{w})$. A symmetric argument implies $p(v_1|v_2^*, \vec{w})$ admits a parameterization that is non-redundant and variationally independent for $p(v_2|v_1^*, \vec{w})$ and $OR(v_1, v_2|\vec{w})$. Thus, all terms in the numerator of the decomposition of $p(v_1, v_2|\vec{w})$ are non-redundant and variationally independent. □

Lemma 2 $OR(v_3, (v_1, v_3)) = OR(v_1, v_3|v_2^*) OR(v_2, v_3|v_1^*) \frac{OR(v_1, v_2|v_3)}{OR(v_1, v_2|v_3^*)}$

Proof: The conclusion follows by definition and term cancellation, as follows:

$$OR(v_3, (v_1, v_2)) = \frac{p(v_3, (v_1, v_2)) p(v_3^*, (v_1^*, v_2^*))}{p(v_3^*, (v_1, v_2)) p(v_3, (v_1^*, v_2^*))}$$

$$OR(v_1, v_3|v_2^*) = \frac{p(v_1, v_3, v_2^*) p(v_1^*, v_3^*, v_2^*)}{p(v_1^*, v_3, v_2^*) p(v_1, v_3^*, v_2^*)}$$

$$OR(v_2, v_3|v_1^*) = \frac{p(v_2, v_3, v_1^*) p(v_2^*, v_3^*, v_1^*)}{p(v_2^*, v_3, v_1^*) p(v_2, v_3^*, v_1^*)}$$

$$\frac{OR(v_1, v_2|v_3)}{OR(v_1, v_2|v_3^*)} = \frac{\frac{p(v_1, v_2, v_3) p(v_1^*, v_2^*, v_3)}{p(v_1^*, v_2, v_3) p(v_1, v_2^*, v_3)}}{\frac{p(v_1, v_2, v_3^*) p(v_1^*, v_2^*, v_3^*)}{p(v_1^*, v_2, v_3^*) p(v_1, v_2^*, v_3^*)}}$$

$$OR(v_1, v_3|v_2^*) OR(v_2, v_3|v_1^*) \frac{OR(v_1, v_2|v_3)}{OR(v_1, v_2|v_3^*)} = \frac{p(v_1, v_3, v_2^*) p(v_1^*, v_3^*, v_2^*)}{p(v_1^*, v_3, v_2^*) p(v_1, v_3^*, v_2^*)} \frac{p(v_2, v_3, v_1^*) p(v_2^*, v_3^*, v_1^*)}{p(v_2^*, v_3, v_1^*) p(v_2, v_3^*, v_1^*)}$$

$$\times \frac{\frac{p(v_1, v_2, v_3) p(v_1^*, v_2^*, v_3)}{p(v_1^*, v_2, v_3) p(v_1, v_2^*, v_3)}}{\frac{p(v_1, v_2, v_3^*) p(v_1^*, v_2^*, v_3^*)}{p(v_1^*, v_2, v_3^*) p(v_1, v_2^*, v_3^*)}}$$

$$= \frac{p(v_3, (v_1, v_2)) p(v_3^*, (v_1^*, v_2^*))}{p(v_3^*, (v_1, v_2)) p(v_3, (v_1^*, v_2^*))} = OR(v_3, (v_1, v_2))$$

□

Theorem 1 (Hammersly-Clifford for conditional MRFs) *Assume a positive $p(\vec{v}|\vec{w})$ obeys the pairwise Markov property for a CUG $\mathcal{G}(\vec{V}, \vec{W})$. That is, for every $V \in \vec{V}$, $Z \in \vec{V} \cup \vec{W}$ such that Z is non-adjacent to V in $\mathcal{G}(\vec{V}, \vec{W})$, $p(v|(\vec{v} \cup \vec{w}) \setminus \{v\})$ is only a function of $(\vec{v} \cup \vec{w}) \setminus \{z\}$. Then $p(\vec{v}|\vec{w})$ factorizes with respect to $\mathcal{G}(\vec{V}, \vec{W})$. That is,*

$$p(\vec{v}|\vec{w}) = \prod_{\vec{C} \subseteq \bar{\mathcal{C}}(\mathcal{G}_{\vec{V}})} \phi_{\vec{C}}(\vec{v}_{\vec{C}}, \vec{w}_{\vec{C}^*}),$$

where for every $\vec{C}$, $\vec{C}^* = \bigcap_{C \in \vec{C}} \mathrm{pa}_\mathcal{G}(C)$, and $\phi_{\vec{C}}$ are terms defined as in (5).

Proof: We generalize the structure of the proof for the MRF case found in the book by Lauritzen (1996) to CMRFs.

Fix distinct $Y, Z \in \vec{A} \subseteq \vec{V}$ that are not neighbors in $\mathcal{G}(\vec{V}, \vec{W})$, and let $\vec{C} = \vec{A} \setminus \{Y, Z\}$, Since $p(\vec{v}|\vec{w})$ is pairwise Markov with respect to $\mathcal{G}(\vec{V}, \vec{W})$, we have that $\tilde{\phi}_{\vec{A}}(\vec{v}_{\vec{A}}, \vec{w})$ is equal to

$$\sum_{\vec{b}: \vec{b} \subseteq \vec{c}} (-1)^{|\vec{c} \setminus \vec{b}|} \left\{ H_{\vec{B}}(\vec{v}_{\vec{B}}, \vec{w}) - H_{\vec{B} \cup \{Y\}}(\vec{v}_{\vec{B} \cup \{Y\}}, \vec{w}) \right.$$
$$\left. - H_{\vec{B} \cup \{Z\}}(\vec{v}_{\vec{B} \cup \{Z\}}, \vec{w}) + H_{\vec{B} \cup \{Y, Z\}}(\vec{v}_{\vec{B} \cup \{Y, Z\}}, \vec{w}) \right\}. \tag{28}$$

Let $\vec{D} = \vec{V} \setminus \{Y, Z\}$. Then we have

$$H_{\vec{B} \cup \{Y, Z\}}(\vec{v}_{\vec{B} \cup \{Y, Z\}}, \vec{w}) - H_{\vec{B} \cup \{Y\}}(\vec{v}, \vec{w})$$
$$= \log \frac{p(\vec{b}, y, z, \vec{v}^*_{\vec{D} \setminus \vec{B}} | \vec{w})}{p(\vec{b}, y, z^*, \vec{v}^*_{\vec{D} \setminus \vec{B}} | \vec{w})} \text{ (by definition)}$$
$$= \frac{p(y|\vec{b}, \vec{v}^*_{\vec{D} \setminus \vec{B}}, \vec{w}) p(z|\vec{b}, \vec{v}^*_{\vec{D} \setminus \vec{B}}, \vec{w})}{p(y|\vec{b}, \vec{v}^*_{\vec{D} \setminus \vec{B}}, \vec{w}) p(z^*|\vec{b}, \vec{v}^*_{\vec{D} \setminus \vec{B}}, \vec{w})} \text{ ($Y \perp\!\!\!\perp Z | \vec{D}, \vec{W}$ by the pairwise property)}$$
$$= \frac{p(y^*|\vec{b}, \vec{v}^*_{\vec{D} \setminus \vec{B}}, \vec{w}) p(z|\vec{b}, \vec{v}^*_{\vec{D} \setminus \vec{B}}, \vec{w})}{p(y^*|\vec{b}, \vec{v}^*_{\vec{D} \setminus \vec{B}}, \vec{w}) p(z^*|\vec{b}, \vec{v}^*_{\vec{D} \setminus \vec{B}}, \vec{w})} \text{ (the first top and bottom terms cancel)}$$
$$= \log \frac{p(\vec{b}, y^*, z, \vec{v}^*_{\vec{D} \setminus \vec{B}} | \vec{w})}{p(\vec{b}, y^*, z^*, \vec{v}^*_{\vec{D} \setminus \vec{B}} | \vec{w})} \text{ (by the chain rule of probability)}$$
$$= H_{\vec{B} \cup \{Z\}}(\vec{v}_{\vec{B} \cup \{Z\}}, \vec{w}) - H_{\vec{B}}(\vec{v}_{\vec{B}}, \vec{w}) \text{ (by definition)}.$$

Thus all terms in the curly brackets in (28) add to zero and henceforth the entire sum is zero whenever $\vec{A}$ is not a clique.

Next, fix a term $\tilde{\phi}_{\vec{A}}(\vec{v}_{\vec{A}}, \vec{w})$ such that $\vec{A} \in \mathcal{C}(\mathcal{G}_{\vec{V}})$, and fix $Y \in \vec{A}$ and $Z \in \vec{W}$ such that Y and Z are not adjacent in $\mathcal{G}(\vec{V}, \vec{W})$. We can express this term as follows:

$$\tilde{\phi}_{\vec{A}}(\vec{v}_{\vec{A}}, \vec{w}) = \sum_{\vec{B} \subseteq \vec{A}} (-1)^{|\vec{A} \setminus \vec{B}|} H_{\vec{B}}(\vec{v}_{\vec{B}}, \vec{w})$$
$$= \sum_{\vec{B} \subseteq \vec{A} \setminus \{Y\}} (-1)^{|(\vec{A} \setminus \{Y\}) \setminus \vec{B}|} \left\{ H_{\vec{B}}(\vec{v}_{\vec{B}}, \vec{w}) - H_{\vec{B} \cup \{Y\}}(\vec{v}_{\vec{B} \cup \{Y\}}, \vec{w}) \right\}.$$

We have

$$H_{\vec{B}}(\vec{v}_{\vec{B}}, \vec{w}) - H_{\vec{B} \cup \{Y\}}(\vec{v}_{\vec{B} \cup \{Y\}}, \vec{w}) = \log \frac{p(y^*, \vec{v}_{\vec{B}}, \vec{v}^*_{\vec{V} \setminus (\vec{B} \cup \{Y\})} | \vec{w})}{p(y, \vec{v}_{\vec{B}}, \vec{v}^*_{\vec{V} \setminus (\vec{B} \cup \{Y\})} | \vec{w})}$$

Since Y and Z are not adjacent in $\mathcal{G}(\vec{V}, \vec{W})$, by the pairwise Markov property, this object is only a function of $y, y^*, \vec{v}_{\vec{B}}$ and $\vec{w}_{\vec{W} \setminus \{Z\}}$. Applying this argument to every $Z \in \vec{W}$ that is not adjacent to some $Y \in \vec{V}$ yields that $\tilde{\phi}_{\vec{A}}(\vec{v}_{\vec{A}}, \vec{w})$ is only a function $\vec{v}_{\vec{A}}$ and $\vec{w}_{\bigcap_{C \in \vec{C}} \mathrm{pa}_\mathcal{G}(C)}$.

This establishes the result. $\qquad\square$

B ADDITIONAL EXAMPLES OF THE L-C FACTORIZATION

The MRF model in Fig. 1 (a) generalizes to MRFs associated with lattice graphs, in particular graphs that can be represented as square grids. Such graphs have been used to analyze particle spins (Nishimori, 2001), and in analysis of images composed of pixels (Szeliski, 2022). An example of a square grid MRF model is shown in Fig. 2. The L-C factorization
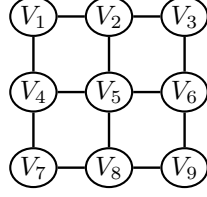
Figure 2: A lattice graph representing an MRF that arises in analysis of images composed of pixels, or in spin state modeling.
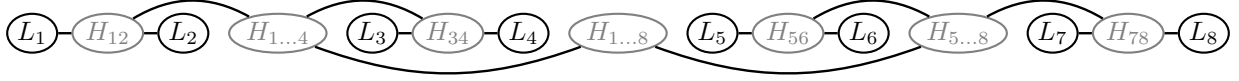

Figure 3: A binary tree graph with 8 leaf vertices corresponding to observed variables, and 7 vertices corresponding to latent variables.

of this MRF model is as follows:

$$Z^{-1}p(v_1 \mid v_2^*, v_4^*)p(v_2 \mid v_1^*, v_3^*, v_5^*)p(v_3 \mid v_2^*, v_6^*)p(v_4 \mid v_1^*, v_5^*, v_7^*)p(v_5 \mid v_4^*, v_2^*, v_6^*, v_8^*)p(v_6 \mid v_5^*, v_3^*, v_9^*)p(v_7 \mid v_4^*, v_8^*)\times$$
$$p(v_8 \mid v_7^*, v_5^*, v_9^*)p(v_9 \mid v_8^*, v_6^*)OR(v_1, v_2|v_3^*, v_4^*, v_5^*)OR(v_2, v_3|v_1^*, v_5^*, v_6^*)OR(v_1, v_4|v_2^*, v_5^*, v_7^*)OR(v_3, v_6|v_2^*, v_5^*, v_9^*)\times$$
$$OR(v_2, v_5|v_1^*, v_4^*, v_3^*, v_6^*, v_8^*)OR(v_4, v_5|v_1^*, v_2^*, v_6^*, v_7^*, v_8^*)OR(v_5, v_6|v_2^*, v_3^*, v_4^*, v_8^*, v_9^*)OR(v_4, v_7|v_1^*, v_5^*, v_8^*)\times$$
$$OR(v_5, v_8|v_2^*, v_4^*, v_6^*, v_7^*, v_9^*)OR(v_6, v_9|v_3^*, v_5^*, v_8^*)OR(v_7, v_8|v_4^*, v_5^*, v_9^*)OR(v_8, v_9|v_5^*, v_6^*, v_7^*).$$

Since only cliques of size 1 or 2 exist in this graph, the factorization only contains the normalizing constant, the odds ratio terms, and the univariate conditional distribution terms. Some reference values have been removed from these terms, without loss of generality, due to conditional independence restrictions implied by the global Markov property of this model.

Latent variable BN and MRF models have been used in the analysis of speech signals (Bilmes and Bartels, 2005; Smith, 2011), and philogeny (Durbin et al., 2002). A graph representing a philogenetic tree model, where only leaf nodes $L_1, \ldots, L_8$ are observed is shown in Fig. 3. The L-C factorization of this MRF model is as follows:

$$Z^{-1}p(l_1|h_{12}^*)p(l_2|h_{12}^*)p(l_3|h_{34}^*)p(l_4|h_{34}^*)p(l_5|h_{56}^*)p(l_6|h_{56}^*)p(l_7|h_{78}^*)p(l_8|h_{78}^*)p(h_{12}|l_1^*, l_2^*, h_{1\ldots4}^*)p(h_{34}|l_3^*, l_4^*, h_{1\ldots4}^*)\times$$
$$p(h_{56}|l_5^*, l_6^*, h_{5\ldots8}^*)p(h_{78}|l_7^*, l_8^*, h_{5\ldots8}^*)p(h_{1\ldots4}|h_{12}^*, h_{34}^*, h_{1\ldots8}^*)p(h_{5\ldots8}|h_{56}^*, h_{78}^*, h_{1\ldots8}^*)p(h_{1\ldots8}|h_{1\ldots4}^*, h_{5\ldots8}^*)\times$$
$$OR(l_1, h_{12}|l_2^*, h_{1\ldots4}^*)OR(l_2, h_{12}|l_1^*, h_{1\ldots4}^*)OR(l_3, h_{34}|l_4^*, h_{1\ldots4}^*)OR(l_4, h_{34}|l_3^*, h_{1\ldots4}^*)OR(l_5, h_{56}|l_6^*, h_{5\ldots8}^*)\times$$
$$OR(l_6, h_{56}|l_5^*, h_{5\ldots8}^*)OR(l_7, h_{78}|l_8^*, h_{5\ldots8}^*)OR(l_8, h_{78}|l_7^*, h_{5\ldots8}^*)OR(h_{12}, h_{1\ldots4}|l_1^*, l_2^*, h_{34}^*, h_{1\ldots8}^*)\times$$
$$OR(h_{34}, h_{1\ldots4}|l_3^*, l_4^*, h_{12}^*, h_{1\ldots8}^*)OR(h_{56}, h_{5\ldots8}|l_5^*, l_6^*, h_{78}^*, h_{1\ldots8}^*)OR(h_{78}, h_{5\ldots8}|l_7^*, l_8^*, h_{56}^*, h_{1\ldots8}^*)\times$$
$$OR(h_{1\ldots4}, h_{1\ldots8}|h_{12}^*, h_{34}^*, h_{5\ldots8}^*)OR(h_{5\ldots8}, h_{1\ldots8}|h_{56}^*, h_{78}^*, h_{1\ldots4}^*).$$

As before, since only cliques of size 1 or 2 exist in this graph, the factorization only contains the normalizing constant, the odds ratio terms, and the univariate conditional distribution terms. Some reference values have been removed from these terms without loss of generality due to the global Markov property. Prior work on philogenetic analysis has used procedures based on the structural EM algorithm to select among models of this type (Friedman et al., 2002). An interesting area of future work is extending such methods to more general likelihoods proposed here.