Random Effects Graphical Models for Multiple Site Sampling*†

Devin S. Johnson
University of Alaska

Jennifer A. Hoeting
Colorado State University

December 23, 2003

Abstract

We propose a two component graphical chain model, the discrete regression distribution, in which a set of categorical (or discrete) random variables is modeled as a response to a set of categorical and continuous covariates. We examine necessary and sufficient conditions for a discrete regression distribution to be described by a given graph. The discrete regression formulation is extended to a state-space representation for the analysis of data collected at many random sites. In addition, some new results concerning marginalization in chain graph models are explored. Using the new results, we examine the Markov properties of the extended model as well as the marginal model of covariates and responses.

Key words and phrases: chain graph, contingency table, discrete regression model, graphical models, marginalization, random effects

1 Introduction

A graphical model, or more loosely a conditional independence model, is a probability density function for a multivariate vector that is parameterized in such a way that a complex independence structure can be characterized by a mathematical graph. A mathematical graph involves a set of vertices, one for each element of the vector, and a set of edges connecting some of the vertices. Edges can either be *undirected* or *directed*. An undirected edge between two vertices indicates dependence between the elements while

*This work was partially funded by STAR Research Assistance Agreement CR-829095 awarded to Colorado State University by the U.S. Environmental Protection Agency. The views expressed here are solely those of authors. EPA does not endorse any products or commercial services mentioned here.

†*Address for correspondence:* Department of Mathematical Sciences, University of Alaska Fairbanks, Fairbanks, AK 99775, U.S.A., E-mail: ffdsj@uaf.edu

a directed edge signifies a "causal" or "influential" effect from one element to another. Herein, we consider chain graph models. Chain graphs are contain both undirected and directed edges. Frydenberg (1990a) and Studený and Bouckaert (1998) provide thorough descriptions of chain graph properties and models.

Research in graphical models has grown considerably over recent years. Whittaker (1990) and Lauritzen (1996) offer comprehensive overviews of the field. This growth in research effort is undoubtedly due to wide applicability of graphical models in many areas of statistics. These models provide methods for examining complex relationships in multivariate distributional models. These complex relationships are described by the Markov properties of the distribution.

In classical inference for graphical models it is standard to assume that measured individuals essentially represent a sample from a single population, such as a city population or geographic site. If multiple individuals are sampled from each of several sites, the likelihood should be constructed to allow for the possibility that individuals from the same site may not be independent. If the number of locations is small one could account for this site effect by the addition of a "site" covariate or margin in a contingency table. A large number of sites, and hence a large number of parameters, precludes this solution.

Discrete compositional data provide a ready example of multiple site discrete data. Each compositional observation represents a multivariate vector of counts where each count represents the number of individuals at a randomly selected sampling site that possess a certain trait. Essentially, each site possess a contingency table of counts. Aitchison (1986, pg. 328), Billheimer and Guttorp (1997), Billheimer et al. (2001), and Johnson (2003) provide examples of this type of data. For the remainder of the paper we will use the term site to refer to a compositional observation, however, sampling does not have to take place over geographically referenced sites. Sampling could take place over time, for example.

In his book on compositional data, Aitchison (1986, pg. 327) notes that compositions originating from discrete counts could be thought of as "random effects" for standard categorical data analysis techniques such as log-linear modeling. Aitchison's brief comment concerning the link between categorical data analysis and discrete com-

positional data motivates this research.

In order to account for many random sites, we propose a new chain graph model for a single site, the discrete regression (DR) model. This new model allows conditional modeling of discrete variables comprising a contingency table given a set of site specific covariates. The Markov properties of this chain graph model are examined and necessary and sufficient conditions for the parameters are provided to ensure a given chain graph, $\mathcal{G}$ describes the independence structure of the model.

Following examination of the DR model, we extend the model to allow for "unobserved site effects" through the addition of random effects. The effect of latent variables on graphical associations has been investigated previously by a several investigators. Geiger and Meek (1998) illustrates a procedure for generating dependence constraints on the distribution for the variables that are directly observable. Smith and Croft (2003) investigate the geometry of the probability space for directed acyclic (DAG) graphical models when latent variables are present. Tian and Pearl (2002) investigates a methodology for testing the validity of a DAG graph model in the presence of latent variables. Lacruz et al. (2001) uses a chain model formulation to model hidden variables in a dynamic graphical model. Herein, the Markov properties of the random effects discrete regression (REDR) model are examined with respect to an *extended graph*, $\mathcal{G}^{\epsilon}$, which includes vertices for the random effects. In addition, we also provide a class of REDR models for which marginalization over the random effects preserves the conditional independencies described by the subgraph of covariates and response variables. Some new results concerning marginalization in chain graph models are proposed to facilitate investigation of the properties of the marginal model.

2 Preliminaries and notation

In this section, we review some of the basic concepts concerning general graphs and chain graphs. First, we will provide some notation for graphical terms. Throughout the article, we generally follow the notation given by Lauritzen (1996). Secondly, will provide some of the basic results concerning Markov properties of chain graph models.

2.1 Notation

A graph, $\mathcal{G} = (V, E)$, is a pair of sets, where V is a finite set of vertices and $E \subseteq V \times V$ is a set of edges. The graphs here are simple in that they contain no loops (an edge with the same beginning and end vertex) or multiple edges (all members of E are unique). Let $(\alpha, \beta) \in E$. If $(\beta, \alpha) \in E$, the edge is referred to as *undirected*. If $(\beta, \alpha) \notin E$, the edge is *directed*.

The usefulness of graphical models is due to the fact that a graph is a visual object and can be represented by a picture. Vertices are represented by dots and edges are represented by lines between the dots. For $\beta \in V$ and $\gamma \in V$, an undirected edge is represented by a line between the dots associated with β and γ. In text, an undirected edge is denoted by $\beta \sim \gamma$. If α is also an element of V and $(\alpha, \beta) \in E$ but $(\beta, \alpha) \notin E$, then, an arrow is drawn from α to β on the graph. In text this is denoted as $\alpha \rightarrow \beta$. If $(\alpha, \delta) \notin E$ and $(\delta, \alpha) \notin E$, for $\delta \in V$, then there is neither an arrow nor a line between α and δ. Absence of a line is denoted by $\alpha \nsim \delta$.

In a chain graph, there exists a known ordering of the vertices. If it is possible for α to influence or "cause" β , then α is said to precede β in this ordering. In an independence graph, an arrow may be present from α to β ($\alpha \rightarrow \beta$). When following a path from one vertex to the next, as with directed graphs, one must obey the direction of the arrow, if present. A *cycle* is a path for which the end vertex is also the starting vertex. A requirement of chain graphs is that they have no directed cycles. The set of *descendants*, $de(\alpha)$, of a vertex $\alpha \in V$, is the set of vertices, β, such that there exists a directed path from α to β, but not from β to α. The non-descendants are $nd(\alpha) = V \setminus [de(\alpha) \cup \{\alpha\}]$. The non-descendants of a vertex are those vertices in the "causal" present and past.

Suppose $A \subseteq V$ is a subset of vertices. Then A induces the subgraph $\mathcal{G}_A = (A, E_A)$ from $\mathcal{G}$, where $E_A = E \cap (A \times A)$. All edges with starting and ending points in A remain in the new graph, every edge with the vertices in that are not in A, $V \setminus A$ are removed. Throughout this article, we use the set notation $A \setminus B$ to refer to the elements in A that are not in B. Additionally, we use $A^c = V \setminus A$ to represent the compliment of a set. We will also use the notation $|A|$ to refer to the cardinality of the set A.

If an arrow is present from α to β , then α is said to be a *parent* of β . Conversely, β is said to be a *child* of α . The set of parents of β is denoted by $pa(\beta)$ and the set of children of α as $ch(\alpha)$. If $\alpha \sim \beta$, then α and β are said to be *neighbors* and $ne(\alpha)$ represents the set of neighbors of α . For a subset $A \subseteq V$, $pa(A)$, $ch(A)$, and $ne(A)$ denote the vertices that are not themselves members of A, but are parents, children, and neighbors of the vertices in A:

$$\begin{aligned} pa(A) &= \bigcup_{\alpha \in A} pa(\alpha) \backslash A \\ ch(A) &= \bigcup_{\alpha \in A} ch(\alpha) \backslash A \\ ne(A) &= \bigcup_{\alpha \in A} ne(\alpha) \backslash A. \end{aligned}$$

The *boundary* of the subset of vertices, A, is the set of all neighbors and parents of A, $bd(A) = pa(A) \cup ne(A)$. The *closure* of A is given by $A \cup bd(A)$. A graph is said to be *complete* if all vertices are joined by a line or arrow. A subset of vertices is complete if it induces a complete subgraph. In an undirected graph, a set B is said to be *simplicial* if $bd(B)$ is complete. A collection of connected sets $\{B_i\}$ is said to be a *simplicial collection* if all sets B_i are simplicial.

Chain graph models are constructed by conditioning random variables based on previous elements of the chain graph. These previous random variables are designated in a chain graph by the term *chain components*. The chain components can be seen in a chain graph by removing all of the directed edges. A chain component τ is called *terminal* if for every $\alpha \in \tau$, $ch(\alpha) = \emptyset$. A subset of vertices is called an *anterior* set if it can be generated by successive removal of terminal chain components.

Finally, using the notation of Frydenberg (1990a), we define the *moral graph*, $\mathcal{G}^m = (V, E^m)$, generated from $\mathcal{G}$. First, recall notation for the undirected version $\mathcal{G}^\sim$ of a graph $\mathcal{G} = (V, E)$. The undirected version $\mathcal{G}^\sim = (V, E^\sim)$, where $E^\sim$ is the same as E except that it is augmented so that all of the edges are undirected. The moral graph is generated from $\mathcal{G}$ by $E^m = E^\sim \bigcup_{\tau \in \mathcal{T}} E^*\{pa(\tau)\}$, where $E^*\{A\}$ is a complete collection of undirected edges for vertices in A. In other words, $\mathcal{G}^m$ is the undirected version of $\mathcal{G}$ with the addition that the parents of each chain component τ are made complete.

2.2 Markov Properties

With the previous collection of tools, we are now able to describe the Markov properties of chain graphs. Frydenberg (1990a) gives extensions to the definitions of the undirected Markov properties for use with chain graph models.

Markov Properties. *Let $\mathcal{G}$ be a chain graph that indexes a set of variables with probability measure P. The probability measure P is then said to be*

(P) *Pairwise Markovian with respect to $\mathcal{G}$ if, for any pair (α, β) of non-adjacent vertices with $\beta \in nd(\alpha)$*

$$\alpha \perp \beta \mid nd(\alpha) \setminus \{\beta\};$$

(L) *Local Markovian with respect to $\mathcal{G}$ if if for any vertex $\alpha \in V$*

$$\alpha \perp nd(\alpha) \mid bd(\alpha);$$

(G) *Global Markovian with respect to $\mathcal{G}$ if for any triple (A, B, S) of disjoint subsets of V such that S separates A from B in $(\mathcal{G}_{an(A\cup B\cup S)})^m$, the moral graph of the smallest ancestral set containing $A \cup B \cup S$, we have*

$$A \perp B \mid S.$$

If P possesses all three of the properties, it is said that P is Markovian with respect to $\mathcal{G}$. These properties are in fact, generalizations of the undirected Markov properties in that if $\mathcal{G}$ is an undirected graph, then the chain Markov properties are equivalent to the undirected Markov properties (Frydenberg, 1990a).

Frydenberg (1990a) also provides a theorem for distributions with a positive and continuous density that relates the Markov properties to Gibbs factorization of the density.

Theorem 2.1. *For any distribution P which has a positive and continuous density $p(\mathbf{x})$ with respect to a product measure, the following four statements are equivalent for any chain graph $\mathcal{G}$:*

i. P is Markovian with respect to $\mathcal{G}$

ii. $p(\mathbf{x}) = \prod_{\tau \in \mathcal{T}} p_{\tau|bd(\tau)}(\mathbf{x}_\tau | \mathbf{x}_{bd(\tau)})$ *where* $P_{cl(\tau)}$ *is Markovian w.r.t* $(\mathcal{G}_{cl(\tau)})^m$ *for all* $\tau \in \mathcal{T}$

iii. p can be factorized

$$p(\mathbf{x}) = \prod_{\tau \in \mathcal{T}} \prod_{C \in \mathcal{C}_\tau} \psi_{C,\tau}(\mathbf{x}_{C,\tau})$$

such that

$$\int_{\mathcal{X}_\tau} \prod_{C \in \mathcal{C}_\tau} \psi_{C,\tau}(\mathbf{x}_{C,\tau}) \mu_\tau(d\mathbf{x}_\tau) \equiv 1$$

for all $\tau \in \mathcal{T}$ where $\mathcal{T}$ denotes the set of chain components in $\mathcal{G}$ and $\mathcal{C}_\tau$ denotes the collection of cliques in $(\mathcal{G}_{cl(\tau)})^m$

iv. (iiii) If A is an anterior set then f_A has a Gibbs factorization w.r.t. $(\mathcal{G}_A)^m$.

We provide a fifth condition, which is not mentioned by Frydenberg (1990a), but is also equivalent to the statements in Theorem 2.1

Lemma 2.1. *Statements (i) - (iv) in Theorem 2.1 are equivalent to the statement*

v. $p(\mathbf{x})$ can be factorized

$p(\mathbf{x}) = \prod_{\tau \in \mathcal{T}} \prod_{B \in \mathcal{B}_\tau} \psi_{B,\tau}(\mathbf{x}_{B,\tau})(0)$*such that*

$$\int_{\mathcal{X}_\tau} \prod_{B \in \mathcal{B}_\tau} \psi_{B,\tau}(\mathbf{x}_{B,\tau}) \mu_\tau(d\mathbf{x}_\tau) \equiv 1$$

for all $\tau \in \mathcal{T}$ where $\mathcal{T}$ denotes the set of chain components in $\mathcal{G}$ and $\mathcal{B}_\tau$ denotes a collection of complete subsets in $(\mathcal{G}_{cl(\tau)})^m$.

Proof. Suppose that p factorizes as (v), then one can rewrite $p(\mathbf{x})$ as in Theorem 2.1(iii) by defining

$$\psi_{C,\tau}(\mathbf{x}_{C,\tau}) = \prod_{B \subseteq C} \psi_{B,\tau}(\mathbf{x}_{B,\tau}).$$

Now, suppose$p$ factorizes according to Theorem 2.1(iii), then, it factorizes according to (v), since, $\mathcal{C}_\tau$ is a collection of complete subsets in $(\mathcal{G}_{cl(\tau)})^m$. □

Lemma 2.1 essentially weakens statement (iii) of Theorem 2.1. The density p for probability distribution P does not have to factorize according to the cliques of a graph in order to be Markovian, it is enough for p to factorize according to complete sets which are not necessarily maximally complete. The condition in Proposition 2.1 is often easier to check than Theorem 2.1 (iii) when trying to ascertain whether or not a distribution is Markovian with respect to a given graph because there is no need to find the maximally complete subsets. In complicated graphs, the cliques may be hard to determine.

3 Marginalization in Chain Graph Models

In this section we describe some marginalization results for general chain graphs. These will be used to examine the consequences of marginalizing the random effects model, proposed in Section 5, over the unobserved latent effects. Specifically, we will use these results to examine the marginal model to determine whether conditional independencies between the covariates and responses are preserved under integration.

Marginalization in the case of undirected graphical models was first investigated by Asmussen and Edwards (1983) in the discrete variable setting. Frydenberg (1990b) continued investigation of marginalization of undirected graphical models in the mixed variable case. Typically, when considering marginalization of graphical models, interest lies in whether the marginal distribution P_A of a subset of variables A is Markovian with respect to the subgraph $\mathcal{G}_A$. Frydenberg (1990b) provides necessary and sufficient conditions for this to be true for the conditional Gaussian graphical model. Castillo et al. (1998, 1999) introduce marginalization operators for determining the subgraph for which a marginal distribution is Markovian.

There has been relatively little investigation into marginalization in chain graphs. Koster (2002) investigates the marginalization of MC graphical models, of which chain graphs are a subset, according to a separation criterion defined within. Koster notes that the defined separation criterion is not equivalent to the classic *moralization* criterion of Lauritzen and Wermuth (1989) and Frydenberg (1990a). In this section, we extend some of the definitions and results contained in Frydenberg (1990b) to chain graph models.

First, many of the marginalization results of Frydenberg (1990b) are based on the notion of simplicial sets in undirected graphs. Therefore, an extension of the definition of a simplicial set is needed for chain graphs.

Definition 3.1 (chain simplicial set). Let $\mathcal{G} = (V, E)$ be a chain graph. A set $B \subseteq V$ is said to be a *chain simplicial* set if $bd(B)$ is complete in $\left(\mathcal{G}_{an\{B\cup ch(B)\}}\right)^m$.

Correspondingly, a collection of unconnected subsets $B = \{B_1, \ldots, B_K\}$ in $\left(\mathcal{G}_{an\{B\cup ch(B)\}}\right)^m$ is a *chain simplicial collection* if B_i is a chain simplicial set for each $i = 1, \ldots, K$.

Now, we present some preliminary results concerning subsets and subgraphs when collapsing over chain simplicial sets. The first result allows examination of anterior sets.

Lemma 3.1. *Let $A \subseteq V$ be a subset of vertices of a chain graph $\mathcal{G}$ such that A^c is a chain simplicial collection. Then, for any $B \subseteq A$, $A^c \cap an(B)$ is a chain simplicial set with respect to the chain graph $\mathcal{G}_{an(B)}$.*

Proof. Consider any one of the chain simplicial sets, say A_i^c, that compose the collection A^c, where $A_i^c \cap an(B) \neq \emptyset$. The sets $ne\{A_i^c \cap an(B)\}$ and $ch\{A_i^c \cap an(B)\}$ remain complete in the appropriate moral graph as they are defined by undirected edges. Since A_i^c is chain simplicial in $\mathcal{G}$, every parent in $pa\{A_i^c \cap an(B)\}$ is a parent of each member of $ne\{A_i^c \cap an(B)\}$ and $ch\{A_i^c \cap an(B)\}$ respectively. It only remains to be shown that $pa\{A_i^c \cap an(B)\}$ is complete in the appropriate moral graph. Since $B \subseteq A$ there exists at least one chain component that intersects $ne\{A_i^c \cap an(B)\}$ or $ch\{A_i^c \cap an(B)\}$. Due to the previous statement and the fact that every member of $pa\{A_i^c \cap an(B)\}$ is also a parent of every member of $ne\{A_i^c \cap an(B)\}$ and $ch\{A_i^c \cap an(B)\}$, $pa\{A_i^c \cap an(B)\}$ is also complete in the appropriate moral graph. □

The second result allows for exchanging the order of moralization and subsetting operations when collapsing over chain simplicial sets. We introduce a new set notation here. Frydenberg (1990a) denotes $\phi(A)$ to be the set of vertices, not in A, which can be reached by directed path from $\alpha \in A$. We will augment this set by the addition of the neighbors of A to give $\phi^*(A) = \phi(A) \cup ne(A)$.

Proposition 3.1. *Let A be a subset of V such that A^c is a chain simplicial collection and $\phi^*(A^c)$ is not empty, then the following equality holds:*

$$(\mathcal{G}_A)^m = (\mathcal{G}^m)_A.$$

Proof. To begin, first note that the edge set for the left hand graph can be written as

$$\begin{aligned} E^{(L)} &= (E_A)^{\sim} \cup \bigcup_{j=1}^{K} E^*\{pa(\tau_j \cap A)\} \cap (A \times A) \\ &= (E_A)^{\sim} \cup \bigcup_{j=1}^{K} E^*\{pa(\tau_j \cap A) \cap A\} \end{aligned} \tag{1}$$

and correspondingly, the right hand edge set as

$$E^{(R)} = (E^{\sim})_A \cup \bigcup_{j=1}^{K} E^*\{pa(\tau_j) \cap A\}, \tag{2}$$

where $\tau_1, \ldots, \tau_K$ are the causal ordered chain components of $\mathcal{G}$.

First, note that for any chain graph, we have $E^{(L)} \subseteq E^{(R)}$. This is due to the fact that $pa(\tau_j \cap A) \cap A \subseteq pa(\tau_j) \cap A$ for all j and $(E_A)^{\sim} = (E^{\sim})_A$. Now since A^c is chain simplicial collection and $\phi^*(A^c) \neq \emptyset$,

$$pa(\tau_j) \cap A \subseteq pa(\tau_k \cap A) \cap A$$

for some $k \geq j = 1, \ldots, K$. If $\tau_j \cap A = \tau_j$ then the equality is trivial. Therefore, we have two additional conditions which need consideration. (1) First assume $\tau_j \cap A$ is not empty. Then, since A^c is chain simplicial, any parent of $\tau_j \cap A^c$ is also a parent of any $ne(\tau_j \cap A^c)$, therefore $k = j$ and the containment is an equality. (2) Now suppose τ_j is composed entirely of A^c vertices. Since $\phi^*(A^c) \neq \emptyset$ there exists a future chain component $\tau_k \subseteq \phi(\tau_j)$ such that τ_k intersects A. Now, since A^c is a chain simplicial, any A parent of τ_j must also be a parent of $\tau_k \cap A$. The preceding argument implies that $E^*\{pa(\tau_j) \cap A\} = E^*\{pa(\tau_j \cap A) \cap A\}$ or $E^*\{pa(\tau_j) \cap A\} \subseteq E^*\{pa(\tau_k \cap A) \cap A\}$ for some $k \geq j$ for $j = 1, \ldots, K$. Therefore, $E^{(R)} \subseteq E^{(L)}$, hence $(\mathcal{G}_A)^m = (\mathcal{G}^m)_A$. □

Now, using the preceding results, we propose the main theorem of this section, which states that a marginal distribution derived from a Markovian distribution is Markovian with respect to the subgraph of the remaining variables if the set which is collapsed over is a chain simplicial collection.

Theorem 3.1. *Let P be any Markovian probability distribution on $\mathcal{X}$ with respect to chain graph $\mathcal{G}$ and let $A \subseteq V$. If A^c is a chain simplicial collection, then P_A is $\mathcal{G}_A$ Markovian.*

Proof. If A is an anterior set then Theorem 2.1 (4) proves the result since P is $\mathcal{G}$ Markovian. Therefore, the remainder of the proof will concentrate on the case where $\phi^*(A^c)$ is not empty, and hence A is not an anterior set.

First, choose sets A_1, A_2, and S contained in A such that S separates A_1 from A_2 in $(\mathcal{G}_{A\cap an(A_1\cup A_2\cup S)})^m$. If $an(A_1 \cup A_2 \cup S) \cap A^c = \emptyset$, then $A_1 \perp A_2 \mid S$ since P is $\mathcal{G}$ Markovian. Now, if the set $an(A_1 \cup A_2 \cup S) \cap A^c \neq \emptyset$, then from Proposition 3.1 and Lemma 3.1 we have

$$\left\{\mathcal{G}_{A\cap an(A_1\cup A_2\cup S)}\right\}^m = \left[\left\{\mathcal{G}_{an(A_1\cup A_2\cup S)}\right\}^m\right]_A.$$

Since $an(A_1 \cup A_2 \cup S) \cap A^c$ is chain simplicial with respect to $\mathcal{G}_{an(A_1\cup A_2\cup S)}$, and hence is simplicial in $\left[\left\{\mathcal{G}_{an(A_1\cup A_2\cup S)}\right\}^m\right]_A$, S must separate A_1 from A_2 in $\left\{\mathcal{G}_{an(A_1\cup A_2\cup S)}\right\}^m$ (see proof of Corollary 2.5 in Asmussen and Edwards (1983) and Frydenberg (1990b)). Therefore, $A_1 \perp A_2 \mid S$ for any such A_1, A_2, and S contained in A and hence, P_A is $\mathcal{G}_A$ Markovian. □

4 Discrete regression model

In this section, we propose a new chain model distribution for random vectors containing both continuous and discrete components. We title this distribution the *discrete regression distribution* (DR) due to its similarity to the conditional Gaussian (CG) regression distribution of Lauritzen and Wermuth (1989) for a discrete response. The DR distribution is constructed by assuming that there exists a set $\Gamma \cup \Delta$ of continuous (Γ) and discrete (Δ) *predictor* variables which follow a CG distribution (Lauritzen and Wermuth, 1989). A set of discrete *response* variables Φ is then distributed according to a log-linear model based on the predictor variables. Our desire to model the variables in Φ as a response to the variables in $\Gamma \cup \Delta$ precludes the use of a CG distribution for the entire set of variables as proposed by Anderson and Bockenholt (2000). The problem with

the CG distribution is that it is not closed under conditioning (Lauritzen and Wermuth, 1989). Therefore, one would need to restrict the types of graphs used so that the Markov properties of the CG distribution for $V = \Phi \cup \Gamma \cup \Delta$ will match those of a chain graph with Φ as the terminal component. The discrete regression distribution eliminates the need for this restriction by building a model based on the desired conditioning.

4.1 Model Formulation

The full joint distribution of the predictor variables $\mathbf{X} = (\mathbf{X}_\Gamma, \mathbf{X}_\Delta)$ and discrete response variables $\mathbf{Y}_\Phi$ is given as the product density $p(\mathbf{y}_\Phi, \mathbf{x}) = p(\mathbf{y}_\Phi|\mathbf{x})p(\mathbf{x})$. More specifically, we begin by considering the conditional density of $\mathbf{Y}_\Phi|\mathbf{x}$ as a log-linear model. First, without worrying about constraining the cell probabilities to the interval [0, 1] or the sum of the probabilities to 1, we model the log probability for each cell $\mathbf{y}_\Phi$ of the response set with the linear model

$$l(\mathbf{y}_\Phi|\mathbf{x}) = \sum_{c\subseteq\Gamma} g_c(\mathbf{y}_\Phi|\mathbf{x}_\Delta) \prod_{\gamma\in c} x_\gamma + \sum_{m=2}^{M} \mathbf{h}_m(\mathbf{y}_\Phi|\mathbf{x}_\Delta)'\mathbf{x}_\Gamma^m, \tag{3}$$

where for every $\mathbf{y}_\Phi$ and $\mathbf{x}_\Delta$, $g_c,\ c \subseteq \Gamma$ is a real number, $\mathbf{h}_m,\ m = 1, \ldots, M$, is a vector in $\mathbb{R}^{|\Gamma|}$, and $\mathbf{x}_\Gamma^m = (x_1^m, \ldots, x_{|\Gamma|}^m)$. The set notation may appear unusual at first, but, $l(\mathbf{y}_\Phi|\mathbf{x})$ has the same structural formulation as a regression model that includes continuous and categorical covariates. The set notation provides a straightforward method for describing a general regression model with all levels of interaction between continuous and categorical variables. In addition, the model also includes polynomial terms, up to some finite power M, of the continuous variables.

Now, if we exponentiate $l(\mathbf{y}_\Phi|\mathbf{x})$, normalize the response cell probabilities, and assume the marginal predictor density has the CG form (Lauritzen and Wermuth, 1989), we obtain the DR joint density

$$\begin{aligned} p(\mathbf{y}_\Phi, \mathbf{x}) =& p(\mathbf{y}_\Phi|\mathbf{x})p_{CG}(\mathbf{x}) \\ =& \exp\left\{\alpha_\Phi(\mathbf{x}) + \sum_{c\subseteq\Gamma} g_c(\mathbf{y}_\Phi|\mathbf{x}_\Delta) \prod_{\gamma\in C} x_\gamma + \sum_{m=2}^{M} \mathbf{h}_m(\mathbf{y}_\Phi|\mathbf{x}_\Delta)'\mathbf{x}_\Gamma^m\right\} \\ & \times \exp\left\{g(\mathbf{x}_\Delta) + \mathbf{h}(\mathbf{x}_\Delta)'\mathbf{x}_\Gamma - \frac{1}{2}\mathbf{x}_\Gamma'\mathbf{T}(\mathbf{x}_\Delta)\mathbf{x}_\Gamma\right\} \end{aligned} \tag{4}$$

where $\alpha_\Phi(\mathbf{x})$ is a normalizing constant with respect to the response cells $\mathbf{y}_\Phi$, for all covariate cells $\mathbf{x}_\Delta$, g is a real number, h is a real vector in $\mathbb{R}^{|\Gamma|}$, and $\mathbf{T}$ is a real, positive definite matrix. The functions $g(\mathbf{x}_\Delta)$, $\mathbf{h}(\mathbf{x}_\Delta)$, $g_c(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$, and $\mathbf{h}_m(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$ are independent from one another for all $\mathbf{y}_\Phi$, $\mathbf{x}_\Delta$, $c \subseteq \Gamma$, and $m = 1, \ldots, M$ in the sense that they are functionally unrelated to each other.

Now, in the same manner as Lauritzen and Wermuth (1989), we will reparameterize the DR response density in terms of interaction effects. As in the proof of the Hammersley-Clifford Theorem (Appendix A), we will define interactions terms relative to an arbitrary but fixed value $(\mathbf{y}^*_\Phi, \mathbf{x}^*) = (\mathbf{y}^*_\Phi, \mathbf{0}_\Gamma, \mathbf{x}^*_\Delta)$ where $\mathbf{0}_\Gamma$ is a $|\Gamma|$ vector of zeros (see Lauritzen, 1996, pg 173). For $f \subseteq \Phi$, $c \subseteq \Gamma$, and $d \subseteq \Delta$, define the interactions

$$\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) = \sum_{a \subseteq d} \sum_{e \subseteq f} (-1)^{|d\setminus a|+|f\setminus e|} g_c(\mathbf{y}_e, \mathbf{y}^*_{\Phi\setminus e}|\mathbf{x}_a, \mathbf{x}^*_{\Delta\setminus a}) \tag{5}$$

and

$$\boldsymbol{\omega}_{fdm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) = \sum_{a \subseteq d} \sum_{e \subseteq f} (-1)^{|d\setminus a|+|f\setminus e|} \mathbf{h}_m(\mathbf{y}_e, \mathbf{y}^*_{\Phi\setminus e}|\mathbf{x}_a, \mathbf{x}^*_{\Delta\setminus a}) \tag{6}$$

The following lemma shows that the classic identifiability constraint for interaction effects in ANOVA or log-linear models is satisfied by (5) and (6).

Lemma 4.1. *The interaction terms $\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$ and $\boldsymbol{\omega}_{fdm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$ defined by (5) and (6), respectively, satisfy the two identifiability constraints*

(1) $\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) = 0$ if $y_\phi = y^*_\phi$ or $x_\delta = x^*_\delta$ for any $\phi \in f$ or $\delta \in d$;

(2) $\boldsymbol{\omega}_{fdm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) = \mathbf{0}$ if $y_\phi = y^*_\phi$ or $x_\delta = x^*_\delta$ for any $\phi \in f$ or $\delta \in d$.

Proof. In order to prove the proposition, we only need calculate the interaction terms under the assumption that either $y_\phi = y^*_\phi$ for some $\phi \in f \subseteq \Phi$, or $x_\delta = x^*_\delta$ for some $\delta \in d \subseteq \Delta$. Therefore, first assume $y_\phi = y_\phi*$ for some $\phi \in f \subseteq \Phi$. Then, the first

interaction term in the proposition is

$$\begin{aligned}\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) &= \sum_{a\subseteq d}\sum_{e\subseteq f}(-1)^{|d\setminus a|+|f\setminus e|}g_c(\mathbf{y}_e,\mathbf{y}^*_{\Phi\setminus e}|\mathbf{x}_a,\mathbf{x}^*_{\Delta\setminus a})\\ &= \sum_{a\subseteq d}(-1)^{|d\setminus a|}\sum_{e\subseteq f\setminus\{\phi\}}(-1)^{|f\setminus\{e\cup\phi\}|}\{g_c(\mathbf{y}_e,y^*_\phi,\mathbf{y}^*_{\Phi\setminus\{e\cup\phi\}}|\mathbf{x}_a,\mathbf{x}^*_{\Delta\setminus a})\\ &\qquad - g_c(\mathbf{y}_e,y^*_\phi,\mathbf{y}^*_{\Phi\setminus\{e\cup\phi\}}|\mathbf{x}_a,\mathbf{x}^*_{\Delta\setminus a})\}\\ &= 0.\end{aligned}$$

A completely analogous calculation holds by assuming $x_\delta = x^*_\delta$ for some $\delta \in d \subseteq \Delta$. Then one can repeat the calculations, replacing g_c with $\mathbf{h}_m$ to prove statement (2) of the lemma. □

Using Möbius inversion (Appendix A), we can rewrite (4) as a function of interaction terms. The DR density can now be rewritten using (5) and (6) and reparameterizing the CG density for the predictor variables as shown in Lauritzen and Wermuth (1989) to give,

$$\begin{aligned}p(\mathbf{y}_\Phi,\mathbf{x}) = \exp&\left[\alpha_\Phi(\mathbf{x}) + \sum_{f\subseteq\Phi}\sum_{c\subseteq\Gamma}\sum_{d\subseteq\Delta}\left\{\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)\prod_{\gamma\in c}x_\gamma\right\}\right.\\ &\left.+\sum_{f\subseteq\Phi}\sum_{\gamma\in\Gamma}\sum_{d\subseteq\Delta}\sum_{m=2}^{M}\omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)x_\gamma^m\right]\\ &\times\exp\left\{\sum_{d\subseteq\Delta}\lambda_d(\mathbf{x}_\Delta)+\sum_{d\subseteq\Delta}\sum_{\gamma\in\Gamma}\eta_{d\gamma}(\mathbf{x}_\Delta)x_\gamma\right.\\ &\left.-\frac{1}{2}\sum_{d\subseteq\Delta}\sum_{\gamma,\mu\in\Gamma}\psi_{d\mu\gamma}(\mathbf{x}_\Delta)x_\gamma x_\mu\right\},\end{aligned} \tag{7}$$

where $\omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$ is the γ element of the vector $\boldsymbol{\omega}_{fdm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$. Without loss of generality we can assume that $\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) = \omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) = 0$ if $f = \emptyset$. Any interaction term that is not a function of $\mathbf{y}_\Phi$ will cancel with the normalizing function $\alpha_\Phi(\mathbf{x})$.

The the response model in the DR distribution intersects the class of CG regression distributions given by Lauritzen and Wermuth (1989). That is to say, if we restrict the DR distribution to contain only quadratic power terms as well as only first and second order interaction terms for the continuous variables, we obtain a CG regression for a purely discrete response. While these restrictions include a wide range of useful models,

we find the restriction unnecessarily confining. So, we propose the DR model as a more flexible model for purely discrete response variables.

4.2 Markov Properties of the DR Distribution

In order to make inference concerning the conditional independence of variables in discrete regressions, we need to determine the Markov properties of the DR chain model. So, we provide the following proposition.

Proposition 4.1. *A DR distribution is Markovian with respect to a chain graph $\mathcal{G}$, with terminal chain component Δ and initial component $\Gamma \cup \Phi$, if and only if the interaction terms in (7) satisfy*

$$\begin{aligned} \beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) &\equiv 0 \text{ unless } f \cup c \cup d \text{ is complete in } \mathcal{G}, \\ \omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) &\equiv 0 \text{ unless } f \cup \{\gamma\} \cup d \text{ is complete in } \mathcal{G}, \end{aligned}$$

and

$$\begin{aligned} \lambda_d(\mathbf{x}_\Delta) &\equiv 0 \text{ unless } d \text{ is complete in } \mathcal{G}, \\ \eta_{d\gamma}(\mathbf{x}_\Delta) &\equiv 0 \text{ unless } d \cup \{\gamma\} \text{ is complete in } \mathcal{G}, \\ \psi_{d\mu\gamma}(\mathbf{x}_\Delta) &\equiv 0 \text{ unless } d \cup \{\mu, \gamma\} \text{ is complete in } \mathcal{G}. \end{aligned}$$

Proof. In order to prove Proposition 4.1 we will give a specialized version of the proof of the Hammersley-Clifford Theorem (Appendix A) for the factorization of each of the chain components so that Theorem 2.1 (iii) is satisfied.

We need only be concerned with the terminal chain component Φ. Lauritzen and Wermuth (1989) prove that the conditions concerning the interaction terms of the CG density are necessary and sufficient for the initial chain component to factorize according to $\mathcal{G}_{\Gamma\cup\Delta} = (\mathcal{G}_{\Gamma\cup\Delta})^m$.

Suppose that the interaction terms $\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$ and $\omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta),\ m = 1, \ldots, M$ are equal to zero for all subsets $f \cup c \cup d$ and $f \cup \{\gamma\} \cup d$ that are not complete. Then, it is easy to observe that $p(\mathbf{y}_\Phi|\mathbf{x})$ factorizes according to complete sets in $\mathcal{G}^m_{cl(\Phi)}$, since $p(\mathbf{y}_\Phi|\mathbf{x})$ is a function only of complete factors in $\mathcal{G}^m_{cl(\Phi)}$. Since the density $p(\mathbf{y}_\Phi|\mathbf{x})$ factorizes

according to complete sets in $\mathcal{G}^m_{cl(\phi)}$, it factorizes according to the cliques by Proposition 2.1.

Now, suppose that the DR distribution is Markovian with respect to $\mathcal{G}$. Then, for $f \subseteq \Phi$, $c \subseteq \Gamma$, and $d \subseteq \Delta$, the interaction term

$$\begin{aligned}\phi_{f\cup c\cup d}(\mathbf{y}_\Phi, \mathbf{x}) &= \sum_{e\subseteq f}\sum_{b\subseteq c}\sum_{a\subseteq d}(-1)^{|e\setminus f|+|c\setminus b|+|d\setminus a|}\log f(\mathbf{y}_e, \mathbf{y}^*_{\Phi\setminus e}, \mathbf{x}_b, \mathbf{0}^*_{\Gamma\setminus b}, \mathbf{x}_a, x^*_{\Delta\setminus a}) \\ &= 0\end{aligned} \tag{8}$$

if the DR distribution is Markov and $d\cup c\cup f$ is not complete in $\mathcal{G}$ (Appendix A). Therefore, we only need to show that $\phi_{f\cup c\cup d}(\mathbf{y}_\Phi, \mathbf{x}) \equiv 0 \Rightarrow \beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) = \omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) = 0$ for $m = 1, \ldots, M$.

Through use of the Möbius inversion theorem and Lemma 4.1, calculation of the Hammersley-Clifford interaction terms are as follows for $d \neq \emptyset$,

$$\begin{aligned}&\phi_{f\cup c\cup d}(\mathbf{y}_\Phi, \mathbf{x}) \\ &\quad= \sum_{e\subseteq f}\sum_{b\subseteq c}\sum_{a\subseteq d}(-1)^{|e\setminus f|+|c\setminus b|+|d\setminus a|}\log f(\mathbf{y}_e, \mathbf{y}^*_{\Phi\setminus e}, \mathbf{x}_b, \mathbf{0}^*_{\Gamma\setminus b}, \mathbf{x}_a, x^*_{\Delta\setminus a}) \\ &\quad= \sum_{b\subseteq c}(-1)^{|c\setminus b|}\sum_{e\subseteq f}\sum_{a\subseteq d}(-1)^{|d\setminus a|+|f\setminus e|}\left[\sum_{a\subseteq d}\sum_{b\subseteq c}\sum_{e\subseteq f}\beta_{eba}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)\prod_{\gamma\in b}x_\gamma\right. \\ &\qquad\qquad\left.+\sum_{a\subseteq d}\sum_{e\subseteq f}\sum_{\gamma\in b}\sum_{j=2}^{m}\omega_{e\gamma am}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)x^m_\gamma\right] \\ &\quad= \sum_{b\subseteq c}(-1)^{|c\setminus b|}\left[\beta_{dbf}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)\prod_{\gamma\in b}x_\gamma + \sum_{\gamma\in b}\sum_{m=2}^{M}\omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)x^j_\gamma\right] \\ &\quad= \beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)\prod_{\gamma\in c}x_\gamma + \sum_{\gamma\in c}\sum_{m=2}^{M}\omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)x_\gamma\sum_{b\subseteq c}(-1)^{|c\setminus b|}1_{[\gamma\in b]} \\ &\quad= \beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)\prod_{\gamma\in c}x_\gamma + \sum_{\gamma\in c}\sum_{m=2}^{M}\omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)x_\gamma\sum_{b\subseteq c\setminus\gamma}(-1)^{|[c\setminus\gamma]\setminus b|} \\ &\quad= \beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)\prod_{\gamma\in c}x_\gamma + \sum_{j=2}^{m}\omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)x_\gamma 1_{[c=\{\gamma\}]}\end{aligned} \tag{9}$$

If $c = \{\gamma\}$ then the interaction term $\phi_{d\cup c\cup f}(\mathbf{y}_\Phi, \mathbf{x})$ is a polynomial of order M in x_γ with coefficients $\omega_{f\gamma d2}(\mathbf{y}_\Phi|\mathbf{x}_\Delta), \ldots, \omega_{f\gamma dM}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$, and $\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$. Therefore, we have that $\phi_{d\cup c\cup f}(\mathbf{y}_\Phi, \mathbf{x}) = 0$ for any $\mathbf{x}$ implies that $\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$ and $\omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta),\ m = 1, \ldots, M$,

must be zero. If $|c| \geq 2$, then we have the single interaction $\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)\prod_{\gamma\in c} x_\gamma$ which equals zero if $\phi_{d\cup c\cup f}(\mathbf{y}_\Phi, \mathbf{x}) = 0$. □

5 Random effects discrete regression

In Section 4.1 we describe a model for a single randomly sampled site. Now, we will extend this model to account for possibly hundreds of randomly selected sites. Here, we are constructing a model for multidimensional contingency table data collected at many sites. For each site, a separate graphical model could be constructed, but this would increase the number of parameters to be estimated to an unmanageable level. Therefore, we propose a global graphical model for all sites that allows site-to-site flexibility. In order to add this flexibility, as well as model the randomness in site selection, we introduce a random error term to the response portion of the DR model in (7). In addition, through the use of the DR framework we can include site specific covariates to the contingency table model, obtaining a chain model for each site.

5.1 Random effects model construction

The addition of a random effect to the response model (7) produces a full model for $\mathbf{Y}_\Phi$, $\mathbf{X}$, and the random effects $\boldsymbol{\epsilon}$ of the form

$$p(\mathbf{y}_\Phi, \mathbf{x}, \boldsymbol{\epsilon}) = p_{RE}(\mathbf{y}_\Phi|\mathbf{x}, \boldsymbol{\epsilon})p_{CG}(\mathbf{x})p(\boldsymbol{\epsilon}). \tag{10}$$

Since there is usually only one observation of the explanatory variables per site, we will leave the model for the covariates, $p_{CG}(\mathbf{x})$, as it is given in (7). The response portion $p_{RE}(\mathbf{y}_\Phi|\mathbf{x}, \boldsymbol{\epsilon})$ of the random effects Discrete Regression (REDR) model is modified by the addition of a random intercept term to give,

$$\begin{aligned} p_{RE}(\mathbf{y}_\Phi|\mathbf{x}, \boldsymbol{\epsilon}) = \exp\Bigg\{ & \alpha_\Phi(\mathbf{x}) + \sum_{f\subseteq\Phi}\sum_{c\subseteq\Gamma}\sum_{d\subseteq\Delta} \beta_{fcd}(\mathbf{y}_\Phi, \mathbf{x}_\Delta)\prod_{\gamma\in c} x_\gamma \\ & + \sum_{f\subseteq\Phi}\sum_{\gamma\in\Gamma}\sum_{d\subseteq\Delta}\sum_{m=2}^{M} \omega_{f\gamma dm}(\mathbf{y}_\Phi, \mathbf{x}_\Delta)x_\gamma^m + \sum_{f\subseteq\Phi}\epsilon_f(\mathbf{y}_\Phi)\Bigg\}, \end{aligned} \tag{11}$$

where $\epsilon_f(y_\Phi) = 0$, if $y_\phi = y^*_\phi$ for any $\phi \in f$, to ensure identifiability. In order to allow modeling of a given independence structure for the multi-way response, we also introduce one other constraint on the random effects. If f is not complete in the graphical representation of the desired independence structure, then all of the random effect terms in $\boldsymbol{\epsilon}_f(\mathbf{y}_\Phi)$ are defined to be 0 for all cells $\mathbf{y}_\Phi$. The remaining random interactions $\boldsymbol{\epsilon}_f = \{\epsilon_f(\mathbf{y}_\Phi) : y_\phi \neq y^*_\phi \text{ for any } \phi \in f\}$, for f complete in the graphical representation of the desired independence structure, are given a multivariate distribution with mean $\mathbf{0}$ and covariance (or scale parameter) $\boldsymbol{\Sigma}_f$. For now, we will consider each of the random effects vectors as being independently distributed. We will show, however, that this restriction can be relaxed to some degree.

The introduction of random error terms in the manner given in (11) has three benefits. First, the model can adjust for site-to-site variability. Secondly, the model will automatically add some level of overdispersion to cell counts. Finally, every realization of the collection of random effects provides cell probabilities that maintain the desired independence relationships among the response variable, as well as between the response variables and the site covariates. To see this one can simply calculate the Hammersley-Clifford interaction terms described in Appendix A and illustrated for DR models in the proof of Proposition 4.1. For each realization of the random effects in (11), all of the interaction terms will remain the same as those calculated in Proposition 4.1, except $\phi_{f\cup\emptyset\cup\emptyset}(\mathbf{y}_\Phi, \mathbf{x}_\Delta) = \beta_{f\emptyset\emptyset}(\mathbf{y}_\Phi) + \epsilon_f(\mathbf{y}_\Phi)$. Therefore, since $\epsilon_f(\mathbf{y}_\Phi)$ is set to zero for sets f that are not complete in the graphical representation of the desired independence structure, the cell probabilities will factor according to that structure with probability 1.

In the REDR model description we have left the error distribution vague. We believe that different situations may necessitate different error structures. If it is reasonable to assume that the error structure is symmetric with few outliers, then a multivariate normal (MVN) distribution may be reasonable. In this case, the cell compositions will have a logistic normal (LN) distribution (Aitchison and Shen, 1980). However, other distributions could be used. For example a multivariate t distribution with k degrees of freedom could be used if it is desirable to have an error with heavier tails. It may be desirable to use the t errors instead of normal errors if there is a high level of overdis-

persion in the cell counts. For the remaining discussion of the random effects DR model we will assume a MVN distribution for the random effects (i.e., $p(\boldsymbol{\epsilon}_f) = p_N(\boldsymbol{\epsilon}_f;\ \mathbf{0}, \boldsymbol{\Sigma}_f)$); however, the theoretical results will remain the same for the t error model.

5.2 Likelihood formulation for multiple site sampling

The likelihood model for the multiple site sampling scheme is slightly different than the standard likelihood for mixed variable graphical models. For each site $i = 1, \ldots, S$, N_i individuals are sampled, while the explanatory variables are sampled only once. Therefore, the likelihood model for the response variable cell counts $\mathbf{c}$ given the covariates $\mathbf{x}$, the total number of individuals observed at a site, and the random effects is

$$p_M(\mathbf{c}|\mathbf{x}, \boldsymbol{\epsilon}) = \frac{N!}{\prod_{\mathbf{y}_\Phi} c(\mathbf{y}_\Phi)!} \prod_{\mathbf{y}_\Phi} p_{RE}(\mathbf{y}_\Phi|\mathbf{x}, \boldsymbol{\epsilon})^{c(\mathbf{y}_\Phi)}, \tag{12}$$

where $f_{RE}(\mathbf{y}_\Phi|\mathbf{x}, \boldsymbol{\epsilon})$ is given by (11).

Now, we focus on the multiple site likelihood for the explanatory variables. Assuming that the covariate observations are independently distributed and follow a CG distribution, we obtain the multiple site explanatory density

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_S) = \prod_{i=1}^{S} p_{CG}(\mathbf{x}_i|\boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\Psi}), \tag{13}$$

where $\mathbf{x}_i$ denotes the set of observed covariates for $i = 1, \ldots, S$ and $\boldsymbol{\lambda}$ $\boldsymbol{\eta}$, and $\boldsymbol{\Psi}$ represent the collected parameter sets $\{\lambda_d(\mathbf{x}_\Delta) : d \subseteq \Delta\}$, $\{\boldsymbol{\lambda}_d(\mathbf{x}_\Delta) : d \subseteq \Delta\}$, $\{\boldsymbol{\eta}_d(\mathbf{x}_\Delta) : d \subseteq \Delta\}$, $\{\boldsymbol{\Psi}_d(\mathbf{x}_\Delta) : d \subseteq \Delta\}$.

We now re-parameterize the homogeneous CG density in (13) into a more useful form. First, we break the CG density into a marginal model for the categorical components of the explanatory variable set and a conditional model for the continuous components. We then re-parameterize the conditional Gaussian distribution into an ANOVA like form. This re-parameterization gives the following form for the homogeneous CG

density,

$$\begin{aligned} p_{CG}(\mathbf{x}) &= p(\mathbf{x}_\Delta)p(\mathbf{x}_\Gamma|\mathbf{x}_\Delta) \\ &= \exp\left\{\sum_{d\subseteq\Delta}\lambda_d(\mathbf{x}_\Delta)\right\} \times \frac{1}{\sqrt{2\pi}}|\mathbf{\Psi}|^{-1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}\left(\mathbf{x}_\Gamma - \sum_{d\subseteq\Delta}\boldsymbol{\tau}_d(\mathbf{x}_\Delta)\right)'\mathbf{\Psi}\left(\mathbf{x}_\Gamma - \sum_{d\subseteq\Delta}\boldsymbol{\tau}_d(\mathbf{x}_\Delta)\right)\right\}, \end{aligned} \tag{14}$$

where $\mathbf{\Psi} = \sum_{d\subseteq\Delta}\mathbf{\Psi}_d(\mathbf{x}_\Delta)$ now represents the inverse covariance matrix for the continuous variables, which have a MVN distribution, $\boldsymbol{\tau}_d(\mathbf{x}_\Delta) = \mathbf{\Psi}^{-1}\boldsymbol{\eta}_d(\mathbf{x}_\Delta)$, and $\lambda_\emptyset(\mathbf{x}_\Delta)$ represents a normalizing constant in the log-linear model for $\mathbf{x}_\Delta$.

Define the vector of cell counts $\mathbf{c}_\Delta = [c(\mathbf{x}_\Delta)]$, where $c(\mathbf{x}_\Delta)$ is the number of sites for which the categorical covariates $\mathbf{X}_\Delta = \mathbf{x}_\Delta$. Using the re-parameterization of the CG density in (14) we can write the joint density of the covariates over all sites (13) as

$$\begin{aligned} p(\mathbf{x}_1,\ldots,\mathbf{x}_S) &= \left\{\prod_{i=1}^S p(\mathbf{x}_{\Delta i})\right\} \times \left\{\prod_{i=1}^S p(\mathbf{x}_{\Gamma i}|\mathbf{x}_{\Delta i})\right\} \\ &\propto p_M(\mathbf{c}_\Delta|\boldsymbol{\lambda})\prod_{i=1}^S p_N\left(\mathbf{x}_{\Gamma i};\ \sum_{d\subseteq\Delta}\tau_d(\mathbf{x}_{\Delta i}),\ \mathbf{\Psi}^{-1}\right), \end{aligned} \tag{15}$$

where the explanatory observation at the ith site is given by $\mathbf{x}_i = (\mathbf{x}_{\Delta i}, \mathbf{x}_{\Gamma i})$ and $p_M(\mathbf{c}_\Delta|\boldsymbol{\lambda})$ is the multinomial density

$$p_M(\mathbf{c}_\Delta|\boldsymbol{\lambda}) = \frac{S!}{\prod_{\mathbf{x}_\Delta} c(\mathbf{x}_\Delta)!}\prod_{\mathbf{x}_\Delta}\exp\left\{\sum_{d\subseteq\Delta}\lambda_d(\mathbf{x}_\Delta)\right\}^{c(\mathbf{x}_\Delta)}. \tag{16}$$

The full likelihood for parameter estimation in the REDR model (10) is obtained by combining the likelihood for response variable cell counts at each site (12), the random effects density, and the explanatory variable likelihood (13).

$$\begin{aligned} p\left(\{\mathbf{c}_i\},\{\mathbf{x}_i\},\{\boldsymbol{\epsilon}_i\}\right) &= \prod_{i=1}^S p_M(\mathbf{c}_i|\mathbf{x}_i,\boldsymbol{\epsilon}_i)p_{CG}(\mathbf{x}_i)p_N(\boldsymbol{\epsilon}_i) \\ &\propto \prod_{i=1}^S p_M(\mathbf{c}_i|\mathbf{x}_i)p_M(\mathbf{c}_\Delta|\boldsymbol{\lambda}) \times \prod_{i=1}^S p_N\left(\mathbf{x}_{\Gamma i};\ \sum_{d\subseteq\Delta}\tau_d(\mathbf{x}_{\Delta i}),\ \mathbf{\Psi}_\emptyset\right) \\ &\quad \times \prod_{i=1}^S\prod_{f\subseteq\Phi} p_N\left(\boldsymbol{\epsilon}_{f,i};\ \mathbf{0},\mathbf{\Sigma}_f\right), \end{aligned} \tag{17}$$

where $\mathbf{c}_i$ is a D vector of response variable cell counts for site i, $\mathbf{x}_i$ is a vector of observed covariates, $\mathbf{c}_\Delta$ is a vector of cell counts for the categorical covariates, and $\boldsymbol{\epsilon}_i$ represents the collection of random effects vectors $\{\boldsymbol{\epsilon}_{f,i} : f \subseteq \Phi \text{ and } f \text{ complete}\}$ for the ith site. It should be noted that all of the random effects vectors for a given site $\boldsymbol{\epsilon}_{f,i}$ are modeled as independent random variables in (17). This can be relaxed to some degree and will be discussed in Section 6.2.

6 Markov Properties of REDR Models

Now that we have defined the Random Effects DR (REDR) model (10), it is of interest to know what conditions determine the Markov properties of this distribution. In addition, it is also of interest to determine how these properties change when the distribution is marginalized over the random effects.

6.1 Extended chain graph

So, we begin with the first question. What conditions determine the Markov properties of the complete REDR distribution. One can observe from (11) that the random effects have exactly the same mathematical effect on the response as the interaction terms of the observed covariates. Therefore, we define an *extended chain graph* $\mathcal{G}^\epsilon$ where the random effects are included as parents of the response variables and are marginally independent of the observed covariates. The graph is represented by a set of vertices for the observable covariates and a set of vertices for the random effects and a set of vertices for the response variables. There are directed edges from observable covariates to the response. There are undirected edges between the response variables as well, depending on their conditional dependence structure. Also, for each $f \subseteq\in \Phi$ there is a directed edge from $\boldsymbol{\epsilon}_f$ to y_ϕ for every $\phi \in f$. Here, the $\boldsymbol{\epsilon}_f$ vertices represent a vector instead of a single random variable. While, this is generally not how a Markov random field is represented by a graph, the theoretical results will remain the same if we represent each element of $\boldsymbol{\epsilon}_f$ with a separate vertex. For ease of notation, we choose the former graphical representation. The random effects are marginally independent of the observed covariates,

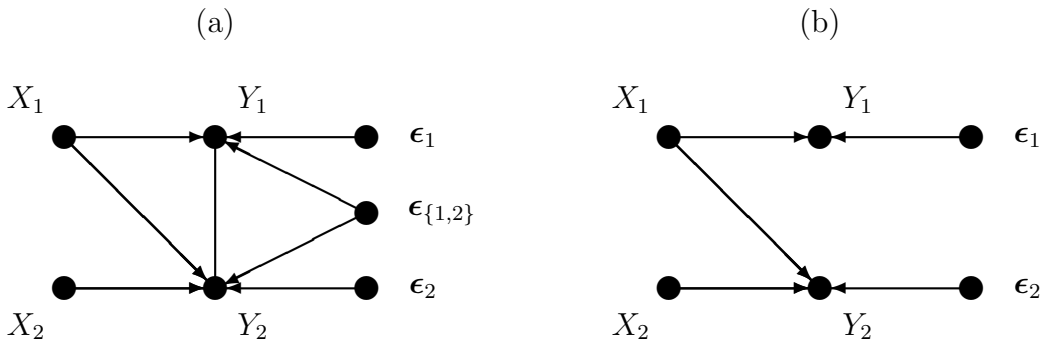

Figure 1: Example of an extended graph $\mathcal{G}^\epsilon$ for random effects Discrete Regression models. Here, (a) represents an extended graph for two response vertices that are connected, therefore, $\boldsymbol{\epsilon}_1$, $\boldsymbol{\epsilon}_2$, and $\boldsymbol{\epsilon}_{1,2}$ are all parents of Φ and (b) is an example of an extended graph for vertices in Φ that are unconnected, therefore, only $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ are parents of Φ. In, (b), by definition, $\boldsymbol{\epsilon}_{\{1,2\}}$ is defined to be a zero vector.

hence, there are no edges between the covariates and the random effects. Figure 1 gives an illustration of two possible extended graphs for a two variable response vector, (a) one where both components of Φ are connected, and (b) one where the components of Φ are not connected. In addition, for now we will consider each $\boldsymbol{\epsilon}_f$, $f \subseteq \Phi$ to be independently distributed, so, there are no edges between the random effects. As mentioned earlier this will be relaxed to some degree at a later point.

We now provide a proposition that describes necessary and sufficient conditions for a REDR distribution to be Markovian with respect to a given extended graph $\mathcal{G}^\epsilon$.

Proposition 6.1. *A REDR distribution P, given by (10) is $\mathcal{G}^\epsilon$ Markovian for a given extended chain graph $\mathcal{G}^\epsilon$, if and only if the interaction terms and random effects in (10) satisfy the following conditions where $f \subseteq \Phi$, $c \subseteq \Gamma$, and $d \subseteq \Delta$:*

1. *(a) $\beta_{fcd}(\mathbf{y}_\Phi, \mathbf{x}_\Delta) \equiv 0$ in (11) unless $f \cup c \cup d$ is complete in $\mathcal{G}^\epsilon$ for $c,\ d \neq \emptyset$,,*

   (b) $\omega_{f\gamma dm}(\mathbf{y}_\Phi, \mathbf{x}_\Delta) \equiv 0$ in (11), for $m = 1, \ldots, M$, unless $f \cup \{\gamma\} \cup d$ is complete in $\mathcal{G}^\epsilon$,

   (c) $\epsilon_f(\mathbf{y}_\Phi) = -\beta_{f\emptyset\emptyset}(\mathbf{y}_\Phi)$ in (11), with probability 1, for all cells $\mathbf{y}_\Phi$ if f is not complete in $\mathcal{G}^\epsilon$.

2. *(a) $\lambda_d(\mathbf{x}_\Delta) \equiv 0$ in (14) unless d is complete in $\mathcal{G}^\epsilon$,*

(b) $\tau_{d\gamma}(\mathbf{x}_\Delta) \equiv 0$ *unless* $d \cup \{\gamma\}$ *is complete in* $\mathcal{G}^\epsilon$*, where* $\tau_{d\gamma}(\mathbf{x}_\Delta)$ *is the element of the vector* $\boldsymbol{\tau}_d(\mathbf{x}_\Delta)$ *in (14) associated with* X_γ *and* $\gamma \in \Gamma$*,*

(c) $\psi_{d\gamma\mu} \equiv 0$ *unless* $d \cup \mu \cup \gamma\}$ *is complete in* $\mathcal{G}^\epsilon$*, where* $\psi_{d\gamma\mu}$ *is the* (μ, γ) *off-diagonal element of* $\boldsymbol{\Psi}_d(\mathbf{x}_\Delta)$ *in (14) and* $\mu \in \Gamma$*.*

Proof. To prove Proposition 6.1 we will show that the conditions presented are necessary and sufficient for a REDR model P (10) to have Gibbs factorization with respect to $\mathcal{G}^\epsilon$ according to Proposition 2.1 and hence show that P is $\mathcal{G}^\epsilon$ Markovian.

Upon examination of the second set of conditions 2(a) through 2(c), one can observe that they are necessary and sufficient for factorization of the marginal density of $\mathbf{X}$ and $\boldsymbol{\epsilon}$ on the subgraph $(\mathcal{G}^\epsilon)_{\{\Gamma \cup \Delta \cup \boldsymbol{\epsilon}\}}$. The second set of conditions are essentially a re-parameterization of the necessary and sufficient factorization criteria for the CG density given by Lauritzen and Wermuth (1989). Conditions 2(a) - 2(c) satisfy Proposition 2.1 for the initial chain component $\Gamma \cup \Delta \cup \boldsymbol{\epsilon}$. Since, the random effects $\{\boldsymbol{\epsilon}_f : f \subseteq \Phi\}$ are marginally independent from each other and $\mathbf{X}$ by construction, $p(\mathbf{x}, \boldsymbol{\epsilon}) = p_{CG}(\mathbf{x}) \prod_{f \subseteq \Phi} p(\boldsymbol{\epsilon}_f)$ factors according to $(\mathcal{G}^\epsilon)_{\{\Gamma \cup \Delta \cup \boldsymbol{\epsilon}\}}$ if and only if $p_{CG}(\mathbf{x})$ factorizes according to $(\mathcal{G}^\epsilon)_{\{\Gamma \cup \Delta\}}$.

Now, all we need to show is that $p_{RE}(\mathbf{y}_\Phi | \mathbf{x}, \boldsymbol{\epsilon})$ factorizes according to complete sets in $\{(\mathcal{G}^\epsilon)_{cl(\Phi)}\}^m$ to complete the factorization of the REDR distribution P according to Proposition 2.1. In order to show this we will follow a similar approach as in the proof of Proposition 4.1. First, note that if conditions 1(a) though 1(c) hold, then P is a function only of complete sets in $\{(\mathcal{G}^\epsilon)_{cl(\Phi)}\}^m$, since the parents of Φ are complete. Now, if we calculate the Hammersley-Clifford interaction terms for the REDR model P in the same manner as in the proof of Proposition 4.1 for the DR model, we see that the interaction terms are identical except for $\phi_{f \cup \emptyset \cup \emptyset}(\mathbf{y}_\Phi, \boldsymbol{\epsilon}) = \beta_{f\emptyset\emptyset}(\mathbf{y}_\Phi) + \epsilon_f(\mathbf{y}_\Phi)$. Now, if it is assumed that P is $\mathcal{G}^\epsilon$ Markovian, then $\phi_{f \cup \emptyset \cup \emptyset}(\mathbf{y}_\Phi, \boldsymbol{\epsilon}) = 0$ for all values of $\mathbf{y}_\Phi$ and $\epsilon_f(\mathbf{y}_\Phi)$. Therefore, $\epsilon_f(\mathbf{y}_\Phi) = -\beta_{f\emptyset\emptyset}(\mathbf{y}_\Phi)$ with probability 1. □

The second question, how do the Markov properties change by marginalizing over the random effects, is a more challenging question due to the fact that the model form prohibits analytical integration. Unfortunately, independence relationships between the

response variables are not generally preserved when marginalizing over all of the random effects. In certain instances, the random effects can act as a mixing distribution. Marginalizing over the random effects has the potential to destroy conditional independencies between response variables. Some model structures, however, are "preservative" in the sense that when one marginalizes over the random effects, the independence relationships between and within the covariates in $\Gamma \cup \Delta$ and responses in Φ are preserved. These model structures are explored further in the next section.

6.2 Preservative REDR Models

There is a sizable class of models for which a specified independence structure is guaranteed to be preserved when integrating the multi-way REDR density (10) over the random effects in (11). We term this class of models *preservative* due to this property. The class of preservative REDR graphs is defined by the following two conditions,

(1) All connected components $a_q,\ q = 1, \ldots, Q$, of Φ in $\mathcal{G}^\epsilon$ are complete, where Q represents the number of connected components in Φ,
and
(2) Any $\delta \in \Gamma \cup \Delta$ that is a parent of $\phi \in a_q$ is also a parent of every other $\phi \in a_q,\ q = 1, \ldots, Q$.

Formulation of REDR models in this fashion essentially allows the vector variables $\mathbf{Y}_{a_q},\ q = 1, \ldots, Q$, to function as a single unit when examining independence relationships.

Some useful independence models are members of the class of preservative REDR models. The first, obviously, is the completely saturated model, where all of the members of Φ are connected and complete and any parents of Φ are parents of every member. This does not mean that all covariates must be parents of every response variable, only that those covariates that *are* parents must be parents to every response variable. Another useful set of models is the completely independent response model. If all responses are conditionally independent of one another, then there is no restriction on the covariates as to whom they must be parents of in order to preserve response conditional independence.

6.3 Markov Properties of Preservative REDR Models

Now, we will demonstrate that if the conditions of Proposition 6.1 are satisfied for a REDR model with respect to a preservative extended graph $\mathcal{G}^\epsilon$, then the marginal distribution $(\mathbf{Y}_\Phi,\ \mathbf{X})$ is $\mathcal{G} = (\mathcal{G}^\epsilon)_{V\setminus\epsilon}$ Markovian. Therefore, when interest lies only in the inference of dependence relationships between and within the covariates and response, the random effects can simply be ignored in the graphical representation.

Proposition 6.2. *If P is an REDR model as described in Section 5.1, and P is Markovian with respect to a preservative, extended graph $\mathcal{G}^\epsilon$, then the marginal distribution of the covariates and responses, $P_{\Phi\cup\Gamma\cup\Delta}$, is $\mathcal{G} = (\mathcal{G}^\epsilon)_{V\setminus\epsilon}$ Markovian.*

Proof. By construction, the random effects are a chain simplicial set in $\mathcal{G}^\epsilon$. Therefore, by Theorem 3.1, if P is $\mathcal{G}^\epsilon$ Markovian then, $(\mathbf{Y}_\Phi,\ \mathbf{X})$ is $\mathcal{G} = (\mathcal{G}^\epsilon)_{V\setminus\epsilon}$ Markovian. □

In the situation where the model of interest is not a preservative model, integration over the random effects can still be carried out. Any REDR model is Markovian with respect to a "preservative" graph. All that needs to be done is to create a graph from the graph $\mathcal{G}^\epsilon$ for which the non-preservative model is Markovian by completing all connected response variable components and adding a directed edge from every parent of a connected response component to every other member of that component. Since we are adding edges, the original REDR model will be Markovian with respect to this new supplemental graph, since it will still factorize according to complete vertex sets. We can then proceed as shown in the proof.

It is possible to generalize the REDR model to allow some degree of association between the random effects for preservative models. In the proof of Proposition 6.2 we used the fact that the random effects vectors were independent from one another. Upon examination of the proof, however, it can be observed that if elements of $\boldsymbol{\epsilon}_f$ and $\boldsymbol{\epsilon}_{f'}$ are correlated for each f and $f' \subseteq a_q$, then the results will remain the same. We can therefore marginalize over random effects that are correlated, and still preserve the associations between the responses and covariates, if all of the correlation occurs between random effects associated with the same complete response component in a preservative model.

7 Conclusions

We have presented an analysis of the Markov properties for a broad class of models for spatially variable graphical models. The models allow modeling of multiple discrete response data sampled at many sites. Site to site variation in cell probabilities can be included without a large increase in the number of parameters to estimate. In addition, site level covariates can be included in the model and inference made between the covariates and response through the conditional independence relationships of a graphical association model. In order to examine the conditional independencies of the marginal model, we provide some results concerning marginalization in chain graph models. These results represent generalizations to the results for undirected graph models given by Frydenberg (1990b).

An obvious extension to the work presented here is to expand the ideas presented here to account for spatial dependencies. Such work will provide insight into spatial graphical models for contingency table data. One possible approach is to consider a *super* graphical model. The super interaction graph is constructed by constructing an extended independence graph for each site, then the random effects are connected by a model such as a lattice spatial model. One could analyze the properties of this giant graphical model. This approach is similar to the methodology proposed by Fienberg and Kim (1999) for combining log-linear graphical models. This is essentially the task we are performing by integrating a spatial process(es) as part of the state-space model. A graphical model describing spatial between site relationships of the random effects and covariates is combined with a graphical model for the random effects, covariates, and response variables within each site. Therefore, this methodology provides promise for either lattice or continuous spatial models.

Dahlhaus (2000) and Fried and Didelez (2003) provide a methodology for analyzing graphical models for multivariate Gaussian data collected in a vector time series. When data are collected as a time series, the spectral matrix of the vector process plays the same role as the inverse correlation matrix. These same spectral methods may be extended to analyze spatial models within a discrete regression framework.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, New York.

Aitchison, J. and Shen, S. M. (1980). Logistic-normal distribution: Some properties and uses. *Biometrika*, 67:261–272.

Anderson, C. and Bockenholt, U. (2000). Graphical regression models for polytomous data. *Psychometrika*, 65:497–509.

Asmussen, S. and Edwards, D. (1983). Collapsibility and response variables in contigency tables. *Biometrika*, 70:567–578.

Billheimer, D. and Guttorp, P. (1997). Natural variablity in benthis species composition in the Delaware Bay. *Environmental and Ecological Statistics*, 4:95–115.

Billheimer, D., Guttorp, P., and Fagen, W. F. (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association*, 96:1205–1214.

Castillo, E. F., Ferrández, J., and Sanmartín, P. (1998). Marginalizing in undirected graph and hypergraph models. *Uncertainty in Artificial Intelligence, Proceedings*, 14:69–78.

Castillo, E. F., Ferrández, J., and Sanmartín, P. (1999). Potential approach in marginalizing Gibbs models. *International Journal of Approximate Reasoning*, 21:257–290.

Dahlhaus, R. (2000). Graphical interaction models for multivariate time series. *Metrika*, 51:157–172.

Fienberg, S. E. and Kim, S. H. (1999). Combining conditional log-linear structures. *Journal of the American Statistical Association*, 94:229–239.

Fried, R. and Didelez, V. (2003). Decomposability and selection of graphical models for multivariate time series. *Biometrika*, 90:251–267.

Frydenberg, M. (1990a). The chain graph Markov property. *Scandinavian Journal of Statistics*, 17:333–353.

Frydenberg, M. (1990b). Marginalization and collapsibility in graphical interaction models. *The Annals of Statistics*, 18:790–805.

Geiger, D. and Meek, C. (1998). Graphical models and exponential families. *Uncertainty in Artificial Intelligence, Proceedings*, 14:156–165.

Johnson, D. (2003). *Random Effects Graphical Models for Discrete Compositional Data.* PhD thesis, Department of Statistics, Colorado State University.

Koster, J. T. A. (2002). Marginalization and conditioning in graphical models. *Bernoulli*, 8:817–840.

Lacruz, B., Lasala, P., and Lekuona, A. (2001). Selecting dynamic graphical models with hidden variables from data. *Computational Statistics*, 16:173–194.

Lauritzen, S. L. (1996). *Graphical Models.* Claredon Press.

Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 17:31–57.

Smith, J. Q. and Croft, J. (2003). Bayesian networks for discrete multivariate data: an algebraic approach to inference. *Journal of Multivariate Analysis*, 84:387–402.

Studený, M. and Bouckaert, R. R. (1998). On chain graph models for description of conditional independence. *The Annals of Statistics*, 26:1434–1495.

Tian, J. and Pearl, J. (2002). On testable implications of causal models with hidden variables. *Uncertainty in Artificial Intelligence, Proceedings*, 18:574–579.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Wiley, Chichster.

A Appendix: Hammersly-Clifford Theorem

Below we present the Hammersley-Clifford Theorem and a sketch of the proof. First, provide a lemma that is used in the proof of the theorem.

Lemma A.1 (Möbius inversion). *For two real valued functions, g and h, defined on the set of all subsets of a finite set V we have that the following two statements are equivalent*

(1) for all $a \subseteq V$: $g(a) = \sum_{b \subseteq a} h(b)$;

(2) for all $a \subseteq V$: $h(a) = \sum_{b \subseteq a} (-1)^{|a \setminus b|} g(b)$,

where $|a|$ represents the cardinality of the set a.

Theorem A.1 (Hammersley-Clifford). *A probability distribution P with positive and continuous density f with respect to a product measure μ satisfies the pairwise Markov property (UP) with respect to an undirected graph $\mathcal{G}$ if and only if it has Gibbs factorization (F) according to $\mathcal{G}$.*

Proof. Here we present a sketch of the proof given by Lauritzen (1996, pg. 36). First, it is fairly simple to show that (F) $\Rightarrow$ (UP), therefore, we will concentrate on the necessity portion of the proof. Now, assume that P is pairwise Markov with respect to $\mathcal{G}$. Since f is assumed positive, we will work with the log density. The definition of (F) can be written according to the log density as

$$\log f(\mathbf{x}) = \sum_{a \subseteq V} \phi_a(\mathbf{x}),$$

where $\phi_a(\mathbf{x})$ depends on $\mathbf{x}$ only through $\mathbf{x}_a$ and $\phi_a(\mathbf{x}) \equiv 0$ unless a is complete is a compete subset of V.

First, assume that P possesses (UP) and choose a fixed but arbitrary element $\mathbf{x}^*$ of $\mathcal{X}$. For all $a \subseteq V$ define,

$$H_a(\mathbf{x}) = \log f(\mathbf{x}_a, \mathbf{x}^*_{a^c}),$$

where $(\mathbf{x}_a, \mathbf{x}^*_{a^c})$ is the element $\mathbf{y}$ of $\mathcal{X}$ such that $y_\gamma = x_\gamma$ for $\gamma \in a$ and $y_\gamma = x^*_\gamma$ for $\gamma \notin a$. Since $\mathbf{x}^*$ is fixed, $H_a(\mathbf{x})$ depends on $\mathbf{x}$ only through $\mathbf{x}_a$. Now, for all $a \subseteq V$, define the interaction term

$$\phi_a(\mathbf{x}) = \sum_{b \subseteq a} (-1)^{|a \setminus b|} H_b(\mathbf{x}), \tag{18}$$

where $|a|$ represents the cardinality of the set a. Next, using Möbius inversion (Lemma A.1) the log density of P can be rewritten as

$$\log f(\mathbf{x}) = H_V(\mathbf{x}) = \sum_{a \subseteq V} \phi_a(\mathbf{x}),$$

It can be observed that $\phi_a(\mathbf{x})$ depends only on the components in $\mathbf{x}$ denoted by the subset a. Thus, it only needs to be shown that $\phi_a(\mathbf{x}) = 0$ whenever a is not a complete subset of V.

Choose $\alpha \in a$ and $\beta \in a$ such that $\alpha \not\sim \beta$. Now, letting $c = V \backslash \{\alpha, \beta\}$ and using the shorthand notation $H_a = H_a(\mathbf{x})$, we have

$$\phi_a(\mathbf{x}) = \sum_{b \subseteq c} (-1)^{|c \backslash b|} \left\{ H_b - H_{b \cup \{\alpha\}} - H_{b \cup \{\beta\}} + H_{b \cup \{\alpha, \beta\}} \right\}$$

Using the definition of $H_a(\mathbf{x})$ it can be shown that all of the terms in the curly brackets add to zero, hence $\phi_a(\mathbf{x}) = 0$ if there are members of a that are not complete in $\mathcal{G}$ and P possesses (UP). Therefore, (UP) $\Longrightarrow$ (F). $\square$