AN ABSTRACT OF THE DISSERTATION OF

Xuan Che for the degree of Doctor of Philosophy in Statistics presented on August 16, 2012.

Title: Spatial Graphical Models with Discrete and Continuous Components.

Abstract approved: _____

Alix I. Gitelman

Graphical models use Markov properties to establish associations among dependent variables. To estimate spatial correlation and other parameters in graphical models, the conditional independences and joint probability distribution of the graph need to be specified. We can rely on Gaussian multivariate models to derive the joint distribution when all the nodes of the graph are assumed to be normally distributed. However, when some of the nodes are discrete, the Gaussian model no longer affords an appropriate joint distribution function. We develop methods specifying the joint distribution of a chain graph with both discrete and continuous components, with spatial dependencies assumed among all variables on the graph. We propose a new group of chain graphs known as the generalized tree networks. Constructing the chain graph as a generalized tree network, we partition its joint distributions according to the maximal cliques. Copula models help us to model correlation among discrete variables in the cliques. We examine the method by analyzing datasets with simulated Gaussian and Bernoulli Markov random fields, as well as with a real dataset involving household income and election results. Estimates from the graphical models are compared with those from spatial random effects models and multivariate regression models.

Spatial Graphical Models with Discrete and Continuous Components

by

Xuan Che

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of

the requirement for the

degree of

Doctor of Philosophy

Presented August 16, 2012

Commencement June 2013

Doctor of Philosophy dissertation of Xuan Che presented on August 16, 2012.

APPROVED:

Major Professor, representing Statistics

Chair of the Department of Statistics

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Xuan Che, Author

ACKNOWLEDGEMENTS

Academic

I would like to express my most sincere thank to my major professor, Dr. Alix Gitelman. She is not only an advisor for my study, but also a mentor and model for my personality and my life. She is an inspirational teacher and warmhearted person, and even in the hardest times I knew that I can always seek her guidance and unconditional support to help me moving on. She showed me the world of statistics and made me realized what kind of a person I want to be. I cannot possibly imagine a better advisor and am forever in debt to her.

I feel fortunate to be in Dr. Lisa Madsen and Dr. Bob Smythe's Master's level classrooms, and years later to have them in my committee again. Their teachings and helps to my research are equally tremendous. I appreciate Dr. Dave Birkes, Dr. Bob Smythe, and Dr. Annie Qu for teaching my advanced level classes, and especially the numerous midnight office hours spent with Dr. Birkes. I thank Jack Mortenson for discussing his spatial data with me, Dr. Gerd Bobe for valiantly stepping in and saving me for my defense, and all the professors and friends who showed interest, supported, and helped me on my research.

Personal

My parents. Of course. They are the ones who love me the most, and the ones who sacrifice the most when their son is always away, years after years. Thank you Mom and Dad and I love you too! My friends, Lindy Hoppers and Oregon, you have given me the happiest years of my life, and I wouldn't miss it for the world.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Graphical models use Markov properties to establish associations among dependent variables. To estimate spatial correlation and other parameters in graphical models, the conditional independences and joint probability distribution of the graph need to be specified. For normal cases, we can rely on Gaussian multivariate models to derive the joint distribution. Gitelman and Herlihy (2007) proposed isomorphic chain graphs (ICG), where variables at each site of the graph are assumed to have the same joint distribution. Irvine (2007) associated different types of ICG with various estimable spatial models under normality and the Alternative Markov Property (AMP). We show that similar associations may also be established under normality and a second type of Markov property, the Lauritzen-Wermuth-Frydenberg (LWF) Markov property.

When some of the variables on the graph are discrete, the Gaussian model no longer affords an appropriate joint distribution function. We develop methods specifying the joint distribution with both discrete and continuous random variables, with spatial dependencies assumed among all variables. By converting the one-way between-variables associations into two-way associations, we can partition the joint distribution of the graph into smaller factors. The connections between these factors are determined by local spatial structures, and this allows our method to be applied to both regular and irregular lattices.

Spatial modeling has witnessed a steady broadening of its application in recent years. Many scientific areas now incorporate some kind of spatial statistical models in their toolbox. Associations between variables and their dependence structures on spatial domains may be collectively summarizes by the *First Law of Geography* (Tobler, 1970), which states that "everything is related to everything else, but near things are

more related than distant things."

Inferences of spatial statistics differ from those of traditional studies in nature. Researchers acquire knowledge of spatial relationships during one of three stages. The first stage is to present and display the spatial relationships, usually done with the aids of graphs and charts. The second stage is to understand the relationship. The third stage is to quantify the relationship, either for the sake of understanding the underlying spatial structure of the domain itself, and/or for making better modeling statements among variables after accounting for the spatial effects. These stages are ordered in the sense that inference on the current stage cannot be completed without the previous one. For instance, without graphing the data it is nearly impossible to visually picture and understand the spatial patterns. In the mean time, no modeling tools may be effectively selected without understanding the pattern first.

Due to the enormous variety of spatial data, the three stages are often treated very differently between datasets. It is therefore desirable to have an inferential method that can be applied to a large number of datasets without altering or calibrating too much for each specific data. Cressie (1993) categorized spatial data into three major families: geostatistical data, lattice data, and point patterns. Geostatistical data are defined on fixed, continuous spatial domains, while lattice data appear on fixed but discrete domains. In contrast, point patterns have random domains. Graphical models are a natural fit on lattice data, though also applicable to geostatistical data and point patterns through spatial aggregations (Schabenberger and Gotway, 2004). They have proved to be effective for displaying, understanding and quantifying multivariate spatial structures.

Graphical models were first developed in response to the questions in path analysis (Wright, 1921, 1934). They harness a node and edge diagram to explain dependency structures within multivariate systems. A graphical model, or a graph, $\mathcal{G}$, is defined by $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, with $\mathbf{V}$ being its node set and $\mathbf{E}$ being the edge set. An edge, (v_i, v_j), is an association between two nodes, $v_i, v_j \in \mathbf{V}$. The flexibility of the graphical model makes it an appealing analysis option for various dependent systems. It can be employed for macro-level inferences, such as image restoration processing (Geman and Geman, 1984) and total system energy calculation in statistical physics via the Ising model (Fisher and Burford, 1967). It has, however, seen limited applications on spatial dependence data, besides when applied to simple graphical structures such as Bayes networks (Haas et al., 1994; Adriaenssens et al., 2004). We believe that the graphical model is an equally cogent inference tool for gaining knowledge on spatial domains, complementing its utilization on the micro level. This thesis is set to verify this belief.

Since lattice data are defined as a countable series of observations on a fixed and discrete spatial domain (Cressie, 1993), observations on the lattice can be naturally projected as nodes in a graphical model. Such models are known as spatial graphical models. Each node in a spatial graphical model denotes a variable observed at a given site, while spatial associations between sites are represented by edges. A graphical model structures spatially dependent data using only these two fundamental elements, without refering to other structural measurements commonly found in other types of spatial models such as the correlation matrix. Through the display of nodes and edges, the diagram provides an intuitive tool to represent relationships in the dataset. Multiple types of graphical models also ensure that they accommodate a large variety of spatial associations. Typical types of graphs include undirected graphs (UG), acyclic

directed graphs (ADG), and chain graphs (CG). UG contains undirected edges only. ADG contains directed edges only and is without a directed cycle. CG may be considered as a combination of UG and ADG and has both directed and undirected edges but no semi-directed cycles. They correspond to different types of associations we may encounter. For individual data, specific graphs can be selected and combined to explain regression-type effects, usually denoted by directed edges, as well as two-way correlations, denoted by undirected edges.

Assuming identical site structures of the graph, isomorphic chain graph (ICG) reduces the number of unknown parameters needed to be estimated from the model. With a properly defined edge set or neighborhood structure, the ICG is structurally simple yet remains highly adaptive for graphs with more complicated and larger scale.

The spatial graphical model explains associations on lattices through a conditional route. Many spatial models rely on a covariance matrix or semivariograms to model spatial effects (Cressie, 1993). These methods estimate and summarize over the entire domain, and are usually expressed as functions of spatial lags, distances, and directions, *etc.* In spatial graphical models, the data is modeled conditionally. The same spatial effects are modeled by the edge set $\mathbf{E}$ of the graph, and the data structure is established only by these pairwise node associations. In this thesis we call the domain-wise spatial structure the global structure, in contrast to the local structure, which stands for the pairwise relationships associated with the edges.

The main benefit for specifying the spatial structure of a graph locally is that the absence of an edge between any pair of nodes induces conditional independence between the two nodes. It is usually easier to model a complicated system conditionally than jointly. To illustrate this, consider an elementary graph involving only three

nodes, $\mathbf{V} = \{v_1, v_2, v_3\}$. We say v_1 and v_2 are conditionally independent given v_3, and write $v_1 \perp\!\!\!\perp v_2 \mid v_3$, if $p(v_1, v_2|v_3) = p(v_1|v_3)p(v_2|v_3)$. Graphically, this conditional independence suggests that there is no edge between v_1 and v_2. When $v_1 \perp\!\!\!\perp v_2 \mid v_3$, we have

$$p(v_1, v_2, v_3) = p(v_1|v_3)p(v_2|v_3)p(v_3) = p(v_1, v_3)p(v_2, v_3)p(v_3)^{-1},$$

which reduces the trivariate joint distribution of $\mathbf{V}$ on the left hand side to a product of bivariate and univariate distributions on the right hand side. This dimension reduction becomes even more helpful for larger graphs.

Collectively, the conditional independences endowed within a graphical model may be extracted and translated from its diagram under a certain set of rules, known as the Markov property (Wermuth and Lauritzen, 1990). Each type of graph has its own Markov properties. The UG conditional independences follow the undirected graph Markov property, while on ADG and CG they are governed by the LWF and AMP Markov properties (Lauritzen, 1996; Andersson et al., 2001). Undirected graphs with Markov properties are known as Markov random fields, while Markov ADG is called Bayes network. The Markov properties equivalent theorems (Frydenberg, 1990; Matúš, 1992; Andersson et al., 1997c) established equivalences between the local and global structures on the graphs, which allow us to model the whole lattice based on conditional independences identified locally.

Graphical models also enjoy the flexibility representing discrete correlated nodes. Most spatial regression models carry a spatially dependent error term for the response variable only (Schabenberger and Gotway, 2004). There are few discussions and propos-

als for spatial regression models that focus on their multivariate aspect (Wartenberg, 1985; Thioulouse et al., 1995; Schmidt and Gelfand, 2003). However, for spatial data, this is hardly the norm. With the exception of carefully designed spatial experiments (Zimmerman and Harville, 1991; Fedorov, 1996; Quinn and Keough, 2002), more often than not, none of the variables is really spatially independent. We suspect that all variables recorded on the lattice follow the first law of geography and exhibit some level of spatial structure. This will be reflected in the spatial graphical models and ICG as they permit spatial autocorrelation for any of their variables.

Another advantage of using conditional independence and Markov properties is that by reading the dependence structure through the graph (rather than a variance-covariance matrix or a semivariogram), graphical models are less dependent upon the marginal distributions of the variables. In other words, a change of distribution in one node would not alter the structure of the entire model. This freedom allows us to accommodate non-Gaussian and especially discrete spatial variables on the graph, something that is of special importance in many studies.

With the spatial structure of the graphs expressed in terms of conditional independences and Markov properties, it is possible to associate graphical models with estimable spatial models for inference. For this purpose assorted graphical models were proposed, covering a variety of diagram structures ranging from tree networks, hierarchical graphs, chessboard grids, to spiderweb-like nets, among many others (Meilă and Jordan, 2001; Kato et al., 1996; Ripley and Kelly, 1977; Knorr-Held and Rue, 2002). For ICG, Irvine (2007) associated $\mathbf{G}_{XY}$ graphs under the AMP Markov property and Gaussian cases with the separable model and the linear model of coregionalization, and $\mathbf{G}_X$ with the separable model.

In this thesis we continue this endeavor by obtaining estimable models associated with chain graphs having both discrete and continuous components. Under the Gaussian assumption, the conditional independence between two nodes corresponds to a zero entry in the precision matrix of the graph (Rue, 1999). Since the precision matrix is the inverse of the variance-covariance matrix, we may determine the variance-covariance matrix through local Markov properties. Comparing the matrix obtained from the Markov property with the existing spatial model variance-covariance matrices, we may find an agreeing model that may be used to estimate the joint distribution on the graph. However, with discrete random variables in the graph we do not have the luxury to connect conditional independence directly with a precision matrix. Instead, we partition the ICG using the Hammersley-Clifford theorem (Hammersley and Clifford, 1971), junction trees (Jensen and Møller, 1991), and copula models (Nelsen, 1999). This partition does not require normality of the nodes, and therefore permits continuous and discrete components on the graph simultaneously.

The central idea of our approach is to first convert the ICG into UG, and then write the joint distribution of the lattice as a product of conditional distributions based on subsets of the lattice. Jensen and Møller (1991) had shown that many models using the conditional approach end up with a joint distribution in the form of

$$f(\mathbf{V}, \Theta) \quad \propto \quad \frac{g(\mathbf{V}, \Theta)}{Z(\Theta)}, \tag{1.1}$$

where $f(\mathbf{V}, \Theta)$ is the joint distribution function of random variables $\mathbf{V}$ and unknown parameters Θ, $g(\cdot)$ is an explicitly known function, and $Z(\cdot)$ is a normalizing function with no closed form except when the whole graph is Gaussian (Knorr-Held and Rue,

2002). Both the Ising model for bivariate statistical systems and Hammersley-Clifford theorem follow the form of Equation (1.1). The main challenge, and the difference between those models, is how to evaluate $Z(\cdot)$. Our approach has to meet this challenge as well.

There are generally two ways solving this problem of unknown $Z(\cdot)$. The first one estimates $Z(\cdot)$ using a numerical approximation algorithm (Ogata and Tanemura, 1984), while the second method does not try to estimate $Z(\cdot)$ at all but instead maximizes a log pseudolikelihood function of the graph (Besag, 1975; Strauss and Ikeda, 1990; Huang and Ogata, 2002):

$$l_{\mathcal{G}}^{PL}(\mathbf{V}, \Theta) = \sum_{v_i \in \mathbf{V}} \log f_{\Theta}(v_i \mid \mathbf{V}_{-i}), \tag{1.2}$$

where $\mathbf{V}_{-i}$ is all the nodes in $\mathbf{V}$ except v_i. Jensen and Møller (1991) showed that the maximum pseudolikelihood estimates (MPLE) based on Equation (1.2) are consistent and asymptotically normal around the true values of Θ for large sample sizes.

We take a third approach to the problem. Our approach is a continuation of the coding techniques on conditional Markov random fields introduced by Besag (1974). He suggested that if we may separate the node set, $\mathbf{V}$, into two groups, $\mathbf{V}_B$ and $\mathbf{V}_C$, such that any two nodes from $\mathbf{V}_B$ are conditionally independent given $\mathbf{V}_C$, then the estimates from

$$l_{\mathcal{G}}^{C}(\mathbf{V}_B, \Theta) = \sum_{v_i \in \mathbf{V}_B} \log f_{\Theta}(v_i), \tag{1.3}$$

approximate the true values of Θ. Graphically $\mathbf{V}_B$ is coded as the conditional Markov random field, and $\mathbf{V}_C$ is known as the conditioning set of nodes. The model has an

advantage by avoiding evaluating $Z(\cdot)$, but it comes with a cost of overlooking the local structure on the graph.

We continue this idea of finding a non-approximated joint distribution on a subset of the graph using conditional independence induced factorizations. What sets our method apart from Besag's is the choice of the conditioning and factor subsets. We choose to use the distribution function of maximal cliques on the UG as the factors of the partition, as opposed to conditional or marginal distributions of single nodes used in the coding methods and pseudolikelihood models. Maximal cliques are the largest sets of nodes that are fully connected. A set of nodes $\{v_1, \dots v_n\}$ is a clique when every pair of nodes $v_i - v_j, i, j \in 1 \dots n$ is connected in the set. A maximal clique means if there is any other node v_{n+1} added to the clique, then there is at least one pair of nodes in $\{v_1, \dots v_n, v_{n+1}\}$ that is not connected by an edge. Since there are no unconnected pairs in a clique, there is no conditional independence property that may be summarized from it. Maximal cliques represent the sets of nodes whose joint distribution evaluations are impossible to simplify by conditional independence and Markov properties. This means that they are the smallest "blocks" in the graph that have to be estimated as a whole. With the local structure inherited within the maximal cliques, the model may be broken down into clusters, and written as a product function of the maximal cliques, known as junction trees (Jensen et al., 1990). $Z(\cdot)$ will be evaluated on the junction trees, but with an explicitly known form. Each of the maximal clique distributions is modeled using a multivariate copula, and the various junction tree estimates on the same graph are eventually combined to form a final estimate.

The main body of this thesis is organized into three chapters. In Chapter 2, we introduce the graphical model, along with the basic terminology necessary for the narrative. We also define conditional independence, and show how distribution of graphs can be summarized using Markov property in undirected graphs (UG), and the LWF and AMP Markov properties in chain graphs (CG). We list the four categories of isomorphic chain graphs ($\mathbf{G}_{\varnothing}$, $\mathbf{G}_X$, $\mathbf{G}_Y$, $\mathbf{G}_{XY}$), and show that besides the AMP Markov property association between the graphs and spatial regression models discovered by Irvine (2007), $\mathbf{G}_{XY}$ ICG can also be associated with multivariate conditional autoregressive models under Gaussian cases and LWF Markov property. In Chapter 3, we explore further into the realm of discrete CG. We use the moralization process to convert CG into UG, and establish a group of CG known as generalized tree networks (GTN) that could benefit the most from the moralization. On regular lattices, the GTN takes on a second order nearest neighbor structure. Moralized GTN may be partitioned according to the Hammersley-Clifford theorem, and based on whether it is a junction tree, they will be either coded as a conditional Markov random fields or treated unconditionally. We use a copula to model the maximal clique marginal distribution, with both continuous and discrete multivariate copulas for different types of variables. In Chapter 4, we examine two example datasets, one simulated from Gaussian and Bernoulli regular lattices, and the other on household income and election result, an irregular lattice example aggregated from geostatistical data. We use Markov chain Monte Carlo methods to obtain Bayesian inferences for the two example, and compare them with existing multivariate spatial regression models. We conclude with discussions and suggestions for future study in Chapter 5.

Chapter 2

Graphical Models and Dependence Structures for Spatial Data

2.1 Introduction

In this chapter we introduce the fundamental concepts and theory behind graphical models, since this theory serves as the backbone of our research. It provides main tools we employ in later chapters to address questions and make inference in multivariate spatial systems.

Graphical models have a long history of relating complex, multivariate random variables with reliable estimates and predictions (Lauritzen, 1996). A graphical model is essentially a diagram that depicts a set of multivariate random variables. It incorporates rudimentary elements, nodes and edges, to denote relationships between variables. Its origin can be traced back to a handful of studies. Gibbs (1902) introduced one of the first concepts of graphical models when studying patterns in statistical physics and particle interactions. Similar ideas also arose in genetics (Wright, 1934), thermodynamics and contingency table analysis (Darroch et al., 1980). Over the years their application has reached into other areas involving multivariate data, and there has been great growth in recent years in machine learning and artificial intelligence applications corresponding to computational advancements (Andrieu et al., 2003; Bishop, 2006; Wainwright and Jordan, 2008).

Graphical models have proved to be a natural fit for many types of multivariate data, including in the realm of spatial statistics (Schabenberger and Gotway, 2004; Wainwright and Jordan, 2008). A spatial data set may comprise a large variety of different variables, characterized by within-variable autocorrelations, between-variable regression effects, or combination of these, which typically makes modeling a challenge. Several researchers have contributed to the application of graphical models for spatial

data. Examples include those from Besag (1972, 1974, 1975); Wermuth and Lauritzen (1989); Cressie (1993); Lauritzen (1996); Carlin et al. (2003); and Jordan (2004). These models primarily focus on Gaussian data or contingency tables.

Gitelman and Herlihy (2007) introduced isomorphic chain graphs (ICG) to represent within-site regression-type relationships and across-site autocorrelation. Irvine (2007) associated ICG under normality with known families of multivariate spatial statistical models. Irvine and Gitelman (2010) divided the ICG class into four categories based on different types of spatial autocorrelation. Under a special conditional independence identification, known as the AMP Markov property, they showed that certain ICG categories may relate with specific multivariate models, such as the separable models and the linear models of coregionalization (LMC). We extend their work by considering the ICG where some nodes are non-Gaussian.

This chapter is organized as follows. In Section 2.2 we provide the necessary definitions and terminology of graphical model theory. In Section 2.3 we introduce two important concepts, conditional independence and the Markov property, and explain why they are crucial in our modeling efforts. We examine different categories of ICG in Section 2.4 and show why they are especially well suited for some spatial data sets. Finally, in Section 2.5 we illustrate that while multivariate conditional autoregressive models (MCAR) may be linked with a particular category of ICG under LWF Markov properties and Gaussian variables, conventional families of spatial regression models, such as the simultaneous autoregressive model (SAR) and conditional autoregressive model (CAR), fail to connect any ICG type with an estimable joint distribution. This shortcoming drives us to search for alternative approaches, resulting the use of generalized tree networks for ICG and building conditional junction trees in Chapter 3.

2.2 Lexicon of graphical models

Before expanding our narrative on graphical models, we need to first establish some basic definitions. A few sets of slightly different terminology exist for graphical models, including Lauritzen and Wermuth (1989); Frydenberg (1990) and Andersson et al. (2001). We have chosen our definitions to closely follow those in Andersson et al. (2001).

A *graph* $\mathcal{G}$ is a pair of sets, $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ is a finite set of *nodes* and $\mathbf{E}$ is a set of *edges* that link these nodes. In the spatial setting, a node, also known as a *vertex*, can be thought of as a random variable observed at a location (or site). The edges can be considered a subset of all possible pairwise relationships among nodes, and hence we can write $\mathbf{E} \subseteq \{(v, w) \in \mathbf{V} \times \mathbf{V} \mid v \neq w\}$. Two types of edges may occur in a graph: an edge $(v, w) \in \mathbf{E}$ whose opposite (w, v) is not in $\mathbf{E}$ is called a *directed edge*. It is denoted by $v \rightarrow w$. An edge $(v, w) \in \mathbf{E}$ whose opposite is in $\mathbf{E}$ is an *undirected edge*. It is denoted by $v - w$.

A subset of the node set, $\mathbf{A} \subseteq \mathbf{V}$, may induce a *subgraph*, $\mathcal{G}_{\mathbf{A}}$, on $\mathcal{G}$. A subgraph is a graph in its own right, and can be written as $\mathcal{G}_{\mathbf{A}} = (\mathbf{A}, \mathbf{E} \cap \{\mathbf{A} \times \mathbf{A}\})$: its node set is $\mathbf{A}$, and its edge set is obtained from $\mathcal{G}$ by keeping all the edges with both end points in $\mathbf{A}$. A subgraph never converts the directed edges in its original graph into undirected edges, or vice-versa.

Based on the types of edges it possesses, a graph is grouped into several general possible categories. If a graph has only undirected edges, it is known as an *undirected graph* (UG; also called a *Markov random field* [MRF]); a graph with only directed edges is a *directed graph* (DG) or *digraph*. Graphs that have both directed and undirected

edges are called chain graphs; although this definition is a little bit more complicated and it will be given later.

The contiguity of a graph can be measured by its completeness. This definition is different from the completeness of an estimator. A graph $\mathcal{G}$ is *complete* if all its nodes are linked through directed or undirected edges. An undirected graph can be broken down into subsets known as cliques. A subset $\mathbf{A} \subseteq \mathbf{V}$ is a *clique* if $\mathcal{G}_{\mathbf{A}}$ is undirected and complete. If a clique cannot accommodate any more nodes without ceasing to be a clique, it is called *maximal clique*.

In a directed graph or chain graph, each directed edge connects a *parent* node to a *child* node. The directed edge originates from the parent and ends at the child. The *parent set* of a subset of nodes, denoted by $pa(\mathbf{A})$, is the collection of all parents that have children in $\mathbf{A}$. Conversely, the *children set* of $\mathbf{A}$, $ch(\mathbf{A})$, contains all the child nodes with a directed edge pointed out from $\mathbf{A}$.

In an undirected graph or chain graph, two nodes $v, w \in \mathcal{G}$ are *adjacent* or *neighbors* if there is an undirected edge between them. One should be aware that the neighbor definition in spatial graphical models is different from that in geography in the sense that not only could two nodes from two proximal locations be considered neighbors, but two nodes at the same location may also be neighbors if they measure two distinct, but related random variables at that location. The collection of all the neighboring pairs is called the *neighborhood structure* of the graph. Similar to the parent set, a *neighbor set*, $nb(\mathbf{A})$, is the collection of all neighbors of the nodes in $\mathbf{A}$. The *boundary*, $bd(\mathbf{A})$, of $\mathbf{A}$ includes all nodes in $\mathbf{V}$ that have either an undirected or directed edge pointed to $\mathbf{A}$. It is a combination of the parent and neighbor sets.

None of the parent set, children set, neighbor set, or the boundary of $\mathbf{A}$ includes $\mathbf{A}$ itself. They all dwell within the periphery of $\mathbf{A}$. On the other hand, a *closure* of $\mathbf{A}$ includes the subset itself and is defined as the union of $\mathbf{A}$ and its boundary.

We provide a mathematical specification of parent, child, neighbor, boundary and closure sets here:

Definition 2.2.1. *Parent, child, neighbor, boundary, and closure sets.*

$$\begin{aligned} pa(\mathbf{A}) &\equiv \{v \in \mathbf{V} \backslash \mathbf{A} \mid v \to a \in \mathbf{E}, a \in \mathbf{A}\}, \\ ch(\mathbf{A}) &\equiv \{v \in \mathbf{V} \backslash \mathbf{A} \mid a \to v \in \mathbf{E}, a \in \mathbf{A}\}, \\ nb(\mathbf{A}) &\equiv \{v \in \mathbf{V} \backslash \mathbf{A} \mid v - a \in \mathbf{E}, a \in \mathbf{A}\}, \\ bd(\mathbf{A}) &\equiv \{v \in \mathbf{V} \backslash \mathbf{A} \mid (v, a) \in \mathbf{E}, a \in \mathbf{A}\}, \\ cl(\mathbf{A}) &\equiv \mathbf{A} \cup bd(\mathbf{A}). \end{aligned} \tag{2.1}$$

A consecutive series of edges in the graph may form a path or cycle. A *path* of length $n \geqslant 1$ from v to w in $\mathcal{G}$ is a sequence of distinct nodes, $\{v_0, v_1, \ldots, v_n\}$, where $v_0 = v, v_n = w$ and $(v_{i-1}, v_i) \in \mathbf{E}$, $i = 1, \ldots, n$. An n-cycle is a path of length $n \geqslant 3$ with $v_0 = v_n$, *i.e.*, the two ends of the path are the same node.

Cycles and paths can be either directed, semi-directed, or undirected. A directed path/cycle is when $v_{i-1} \to v_i \in \mathbf{E}$ for all i, or all edges are arrows. A semi-directed path/cycle is when $v_{i-1} \to v_i \in \mathbf{E}$ for at least one $i \in 1, \ldots, n$. An undirected path/cycle is when $v_{i-1} - v_i \in \mathbf{E}$ for all i's; *i.e.*, all edges are undirected.

For a subset $\mathbf{A} \subseteq \mathbf{V}$, its *ancestors* in $\mathcal{G}$ are defined as the set of nodes in $\mathbf{V}$ that have directed paths to some $a \in \mathbf{A}$. The collection of ancestors of $\mathbf{A}$ forms the *ancestor*

set $an(\mathbf{A})$. The *descendants* of $\mathbf{A}$, on the other hand, are defined as the nodes that have a directed path from some $a \in \mathbf{A}$, and the *descendant set* is denoted as $de(\mathbf{A})$. The *non-descendants* of $\mathbf{A}$ are $nd(\mathbf{A}) = \mathbf{V} \backslash \{de(\mathbf{A}) \cup \mathbf{A}\}$.

If a directed graph has no (directed) cycles, it is an *acyclic digraph* (ADG). An acyclic digraph allows only directed paths and no cycles. A graph is *adicyclic*, or is a *chain graph* (CG), if it contains no semi-directed cycles. In other words, a chain graph allows all types of paths, as well as directed and undirected cycles. Both the directed graph and undirected graph can be considered as special cases of chain graphs. An undirected graph may only have undirected paths and cycles. If it does not have any cycles, the undirected graph is known as a *tree network* (Meilă and Jordan, 2001).

A *chord* is an edge between two non-consecutive nodes, in a path with length $n \geqslant 3$ or a cycle with $n \geqslant 4$. We say a UG is *chordal* if every cycle with $n \geqslant 4$ possesses a chord. A path or cycle is *chordless* or *non-chordal* if no non-consecutive nodes are linked by an edge.

The types of paths induce connected components in a graph. In a chain graph we say v *leads to* w, and write $v \mapsto w$, if there is a path that goes from v to w. When both $v \mapsto w$ and $w \mapsto v$, we say v to w are *connected* and write $v \rightleftharpoons w$. If every pair of nodes is connected in $\mathcal{G}$, the graph $\mathcal{G}$ is connected. *Connected components*, $[a]$, is the set of nodes connected to a, such that $[a] \equiv \{v \in \mathbf{V} \mid v \rightleftharpoons a\}$.

The contrary idea to the connectivity of a graph is separation. In a way, knowing the separation is more important to us because the direction of greater connectivity means the nodes are more entangled with each other and the whole graph is more complicated, whereas greater separation means more possibility for simplification. For

a triplet of disjoint, non-empty subsets $\mathbf{A}, \mathbf{B}, \mathbf{S}$ of $\mathbf{V}$, we say $\mathbf{S}$ separates $\mathbf{A}$ and $\mathbf{B}$ if every path in $\mathcal{G}$ between a node in $\mathbf{A}$ and a node $\mathbf{B}$ intersects with $\mathbf{S}$.

A few more terms are needed for chain graphs to express Markov properties on them. In a chain graph with both directed and undirected edges, a *chain component*, τ, is a set of nodes in $\mathbf{V}$ that are only connected to each other through undirected paths and to other parts of the graph through directed edges. In other words, there is no directed edge in a chain component. The collection of chain components is usually denoted by $\mathbf{T}$, with $\tau \in \mathbf{T}$. Each node in a chain graph lies in an unique chain component only, because all the chain components are disjoint. The chain components are connected to each other only by directed paths to form the whole graph $\mathcal{G}$. The subgraph induced by the chain components $\mathbf{T}$ is a directed graph.

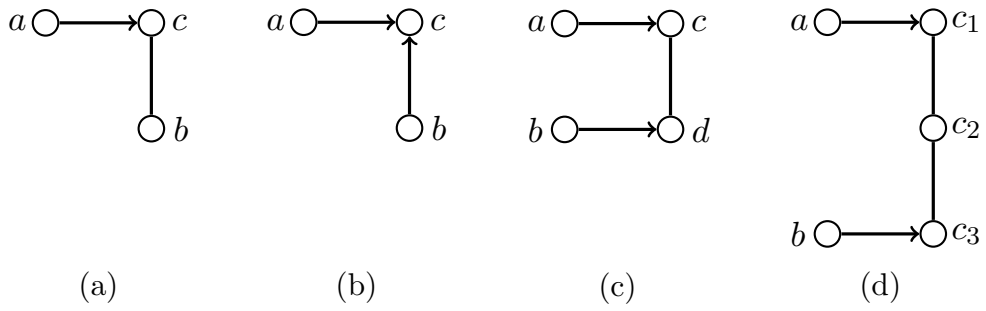
In our research it is important to recognize the neighborhood structures of chain graphs, so that we can convert them to undirected graphs for further processing. For this purpose we need to first categorize these neighborhoods, especially those subgraphs with both directed and undirected edges in them, so that the conversion can then be made possible.

A three node subgraph $a \to c - b$ is called a *flag* and is denoted as $[a, b; c]$. $a - c \leftarrow b$ is another flag, denoted as $[b, a; c]$. A *2-biflag* $[a, b; c, d]$ has the form $a \to c - d \leftarrow b$, with the linkage between a and b is one of undirected, directed or no edge.

A *k-complex* $(a, b; c_1 \ldots, c_k)$ is a subgraph of the form $a \to c_1 - \ldots - c_k \leftarrow b$. If $k = 1$, such that $a \to c \leftarrow b$, it is called an *immorality*. If $k \geqslant 2$, it is also known as a *multicomplex*. Examples of flag, immorality, 2-biflag and 3-complex can be viewed in Figure 2.1. A *moralized k-complex* is a chain component obtained from a k-complex

by first adding an undirected edge between a and b and then converting all directed edges in the subgraph into undirected ones. Essentially, a moralized k-complex is a chordless $(k+2)$-cycle. The process of moralizing all the k-complexes in a chain graph is also known as *moralization* (Cowell et al., 1999; Studený, 2001). The resulting graph is an undirected graph, which is known as the *Moralized graph* ($\mathcal{G}^M$) derived from the original chain graph $\mathcal{G}$. As we will see later, these moralized undirected graphs play an important role in our model development.

Figure 2.1: A flag (a), an immorality (b), a 2-biflag (c) and a 3-complex (d)



These definitions form the general lexicon of graphical model theory, but they are in no sense complete. Some additional terminology is of specific interest later, and we will give those definitions when we encounter the need for them. For now, our knowledge of graphical models is enough to introduce the conditional independence Markov properties, two important features of the graphs.

2.3 Conditional independence and Markov properties

2.3.1 Conditional independence

There are two approaches one may take to model the dependence structure of a complicated multivariate system: one may either choose to specify the joint distribution of all the variables in the system directly or to build it through the conditional distributions of individual variables or groups of variables. In many cases, especially with large samples, it is easier to work with the conditional distributions than with the joint distribution directly (Jensen and Møller, 1991; Drton and Eichler, 2006). For spatial data, the conditional perspective also makes more sense, since it intuitively explains variations and correlations at one location as dependent upon neighboring locations. Conditional distributions factor the joint probability distribution into smaller pieces to simplify the modeling. It is this second, conditional approach on which Irvine and Gitelman (2010) have taken to connect Gaussian ICG with multivariate spatial models.

At the center of the approach is the idea of conditional independence, which can be graphically represented. Dawid (1979, 1980) introduced and formalized the concept of conditional independence and its relations to the graphs. If three random variables X, Y and Z share a joint probability distribution P, we say *X is conditionally independent from Y given Z in P* if and only if

$$p(X, Y \mid Z) = p(X \mid Z)p(X \mid Z). \tag{2.2}$$

Equation (2.2) can be denoted as $X \perp\!\!\!\perp Y \mid Z[P]$. If it is clear which joint distribution we are referring to, then $[P]$ is usually omitted and we simply write $X \perp\!\!\!\perp Y \mid Z$. If

the three variables are considered to be continuous with respect to a product measure, and $f_Z(z) > 0$ for all z, then Equation (2.2) is equivalent to

$$X \perp\!\!\!\perp Y \mid Z \iff f_{XY|Z}(x, y \mid z) = f_{X|Z}(x \mid z) f_{Y|Z}(y \mid z) \tag{2.3}$$

$$\iff f_{XYZ}(x, y, z) = \frac{f_{XZ}(x, z) f_{YZ}(y, z)}{f_Z(z)}. \tag{2.4}$$

For a special case, when Z is the empty set $Z = \emptyset$, we say X and Y are independent and write $X \perp\!\!\!\perp Y \mid \emptyset \iff X \perp\!\!\!\perp Y$.

From the relationship $X \perp\!\!\!\perp Y \mid Z$, Dawid (1980) introduced the following properties, which were coined as the *axioms for conditional independence* by Lauritzen (1996, Chap. 3). Suppose X, Y, Z, U and W are random variables with probability distribution P, then:

Proposition 2.3.1. *(Axioms for conditional independence) For some measurable function, h, on the sample space of X,*

(C1) $X \perp\!\!\!\perp Y \mid Z \iff Y \perp\!\!\!\perp X \mid Z$;

(C2) $X \perp\!\!\!\perp Y \mid Z$ *and* $U = h(X) \implies U \perp\!\!\!\perp Y \mid Z$;

(C3) $X \perp\!\!\!\perp Y \mid Z$ *and* $U = h(X) \implies X \perp\!\!\!\perp Y \mid \{Z, U\}$;

(C4) $X \perp\!\!\!\perp Y \mid Z$ *and* $X \perp\!\!\!\perp W \mid \{Y, Z\} \implies X \perp\!\!\!\perp \{W, Y\} \mid Z$.

In addition to (C1) - (C4), there is yet another property (C5) which does not always hold true. However, if P is a positive and continuous density with respect to a product measure π, then

(C5) $X \perp\!\!\!\perp Y \mid \{Z, W\}$ and $X \perp\!\!\!\perp Z \mid \{Y, W\} \implies X \perp\!\!\!\perp \{Y, Z\} \mid W$.

(C5) is almost always true in practice except some trivial cases (such as when $X \equiv Y \equiv Z$). Based on these conditional independence axioms, the joint distribution of a multivariate stochastic system can be specified using its conditional distributions. The conditional independences drive our model development and allow us to correspond the joint distributions obtained from graphical structures with some specific spatial regression models, such as the autoregressive models (CAR and SAR, see Section 2.5) for spatial stochastic data. The theory behind the equivalence of the joint and conditional distributions is supported by Brook's Lemma (Brook, 1964), repeated here.

Lemma 2.3.2. *(Brook's Lemma) Let* $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$ *and* $\mathbf{y}' = \{y_1', y_2', \ldots, y_n'\}$ *be two different realizations of a random vector* $\mathbf{Y} = \{Y_1, Y_2, \ldots Y_n\}$. *Let* $\pi(\mathbf{Y})$ *be a strictly positive density function of* $\mathbf{Y} \in \mathbb{R}^n$, *then*

$$\begin{aligned} \frac{\pi(\mathbf{y})}{\pi(\mathbf{y}')} &= \prod_{i=1}^{n} \frac{\pi(y_i|y_1, \ldots, y_{i-1}, y_{i+1}', \ldots, y_n')}{\pi(y_i'|y_1, \ldots, y_{i-1}, y_{i+1}', \ldots, y_n')}; \\ &= \prod_{i=1}^{n} \frac{\pi(y_i|y_1', \ldots, y_{i-1}', y_{i+1}, \ldots, y_n)}{\pi(y_i'|y_1', \ldots, y_{i-1}', y_{i+1}, \ldots, y_n)}. \end{aligned} \tag{2.5}$$

Brook's Lemma does not deliver an exact form for the joint distribution on $\mathbf{Y}$. Rather, it provides the ratio between any joint density pair based on possible realizations of the graph or the model. The importance of Brook's Lemma is that, proportional to a constant (which cancels out in the ratio), the joint distribution of a random vector can be specified as the product of its conditional distributions. We may argue that this ratio is as good as the exact joint distribution because it gives us full access to model the unknown parameters based on maximum likelihood methods and to make spatial predictions, also known as kriging (Lauritzen, 1996). The only difference

between the denominator on the right of Equation 2.5 and the joint distribution $\pi(\mathbf{y}')$ is a normalizing constant multiplier, which is a nuisance in the model.

Although in Brook's Lemma the joint distribution can be uniquely determined by its conditional distributions, it does not guarantee that any set of arbitrary conditional distributions will define a proper joint distribution. A counter-example of this can be found in Section 3.3 (3.2, upon merging chapters).

In graphical models, the neighborhood structures of the graph help us "visualize" the conditional independence properties of the random variables represented in the graph, while the conditional independences in turn enable us to build the conditional distributions to be combined together to form the joint distribution. A formal summary of the relationships between the graph and its conditional independences are known as the Markov properties of the graph. They differ between directed, undirected and chain graphs, but in all cases are important representations of the graph structures.

2.3.2 Markov properties for undirected graphs

In this section, we consider conditional independences on undirected graphs with finite numbers of nodes. The probability space of the undirected graph is either a finite-dimensional real vector space, or a combination of real vector space and finite discrete sets. The conditional independence on the graph can be generalized using one of three Markov properties (Frydenberg, 1990; Lauritzen, 1996; Andersson et al., 2001).

Definition 2.3.3. *(Pairwise, local and global Markov properties on undirected graphs) A probability distribution P on an undirected graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is said to have*

(UP) The pairwise Markov property*, relative to $\mathcal{G}$, if for any two non-neighboring*

nodes $\alpha, \beta \in \mathbf{V}$*,*

$$\alpha \perp\!\!\!\perp \beta \mid \mathbf{V} \backslash \{\alpha, \beta\};$$

(UL) The local Markov property*, relative to* $\mathcal{G}$*, if for any node* $\alpha \in \mathbf{V}$*,*

$$\alpha \perp\!\!\!\perp \mathbf{V} \backslash cl(\alpha) \mid bd(\alpha);$$

(UG) The global Markov property*, relative to* $\mathcal{G}$*, if for any triplet of subsets (*$\mathbf{A}, \mathbf{B}, \mathbf{S}$*) of* $\mathbf{V}$ *such that* $\mathbf{S}$ *separates* $\mathbf{A}$ *from* $\mathbf{B}$ *in* $\mathcal{G}$*,*

$$\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{S}.$$

Under any circumstance, the relationships between these three Markov properties may be described by Proposition 2.3.4, due to Lauritzen (1996).

Proposition 2.3.4. *For any probability distribution* P *on any undirected graph* $\mathcal{G}$*,*

$$\text{(UG)} \quad \Longrightarrow \quad \text{(UL)} \quad \Longrightarrow \quad \text{(UP)}. \tag{2.6}$$

Equation (2.6) does not depend on the positive, continuous condition, which is stated as (C5) in Prop. 2.3.1. If, however, the (C5) condition is met, we can improve upon Proposition 2.3.4 to reach the stronger Pearl-Paz Theorem (Pearl and Paz, 1987).

Theorem 2.3.5. *(Pearl-Paz Theorem) If a probability distribution* P *holds for* (C5) *on an undirected graph* $\mathcal{G}$*, then*

$$\text{(UG)} \quad \Longleftrightarrow \quad \text{(UL)} \quad \Longleftrightarrow \quad \text{(UP)}. \tag{2.7}$$

The Pearl-Paz Theorem means that, under positive and continuous conditions, all three Markov properties are equivalent on undirected graphs. Therefore, we may start working from one of the properties on the graph and try to finish modeling the distribution of the graphical model based on another one. Incidentally, among the three properties, it is the global Markov property (UG) that is of the most importance to us. It serves as the general rule determining which sets of nodes are conditionally independent of each other and helps us facilitate the factorization of the joint distribution on the whole graph.

2.3.3 LWF Markov properties for chain graphs

When generalizing the Markov property to chain graphs, it has been noticed that multiple alternative Markov properties may exist for the same graph; *i.e.*, a chain graph is capable of inducing different covariance structures and statistical models (Cox and Wermuth, 1993; Wermuth et al., 1994; Andersson et al., 2001). Among them, perhaps the most studied are the Lauritzen-Wermuth-Frydenberg (LWF) and Andersson-Madigan-Perlman (AMP) Markov properties.

The reason for us to study these properties is to leverage the conditional independence structures to write estimable statistical models for spatial chain graphs. In this section we first introduce the LWF property on chain graphs as a generalization from undirected graphs, and show that the properties can also be stated in a block recursive context. We then give the definitions of the AMP property, followed by a comparison between these two properties, and a discussion of their similar but distinct applications in spatial chain graphs.

First attempts for the Markov properties on acyclic directed graphs and chain graphs were made by Kiiveri et al. (1984), joined by Pearl and Verma (1987) and Smith (1989) later. Frydenberg (1990) and Wermuth and Lauritzen (1990) summarized it systematically into the LWF Markov property, which provides conditional independence interpretations for nodes with both directed and undirected relationships.

Definition 2.3.6. *(LWF Markov properties on chain graphs) We say a probability distribution P on a chain graph $\mathcal{G}$ satisfies*

(CP) The pairwise chain Markov property*, relative to $\mathcal{G}$, if for any pair of non-neighboring nodes (α, β) with $\beta \in nd(\alpha)$*

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

(CL) The local chain Markov property*, relative to $\mathcal{G}$, if for any node $\alpha \in \mathbf{V}$,*

$$\alpha \perp\!\!\!\perp \{nd(\alpha) \backslash cl(\alpha)\} \mid bd(\alpha);$$

(CG) The global chain Markov property*, relative to $\mathcal{G}$, if for any triplet of subsets $(\mathbf{A}, \mathbf{B}, \mathbf{S})$ of $\mathbf{V}$ such that $\mathbf{S}$ separates $\mathbf{A}$ from $\mathbf{B}$ in $\mathcal{G}^M_{An(\mathbf{A},\mathbf{B},\mathbf{S})}$, the moralized undirected graph of the smallest ancestral set containing $\mathbf{A} \cup \mathbf{B} \cup \mathbf{S}$,*

$$\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{S}.$$

When there is no assumption of conditional independence of any kind imposed on the chain graph, we cannot make statements about the equivalence of these Markov

properties. However, when conditions (C1) to (C5) hold true, much like their counterparts for undirected graphs, the three LWF Markov properties in Definition 2.3.6 are indeed equivalent (Frydenberg, 1990; Lauritzen, 1996).

Theorem 2.3.7. *If a probability distribution P holds for* (C5) *for disjoint subsets of* $\mathbf{V}$ *on the chain graph* $\mathcal{G}$*, then*

$$(\text{CG}) \iff (\text{CL}) \iff (\text{CP}). \tag{2.8}$$

Proof of the theorem can be found in Lauritzen (1996). Similar to the Markov properties in undirected graphs, the equivalence here allows us to recognize the graph separations on which the conditional distribution model are based.

To facilitate the transition into the second Markov property, the AMP Markov property, we provide an equivalent definition of the LWF Markov property. This time, the Markov property is defined in a block recursive context. Unlike the univariate recursive definitions (CP and CL in Prop. 2.3.6), where conditional independence is specified on single nodes, block recursive definition specifies the conditional independence between chain components of the graph (Wermuth, 1991); *i.e.*, subsets of nodes connected by undirected paths. The reason behind this redefinition is that the block recursive AMP definition is almost the same as the LWF definition except for one small but important difference. The block recursive context allows the AMP properties to be conditioned on a subset smaller and included in the conditioning subset required by the LWF properties, thus permiting a different and sometime more flexible modeling method.

We consider the graph $\mathcal{D}$ induced by the chain components on $\mathcal{G}$. $\mathbf{T}$ is the collection

of all chain components of $\mathcal{G}$. The nodes of $\mathcal{D}$ are $\tau_i \in \mathbf{T}$, and its edges are $\mathcal{E} \subseteq \{(\tau_i, \tau_j) \in \mathbf{T} \times \mathbf{T}\}$, the directed edges connecting the chain components in $\mathcal{G}$. $\mathcal{D} = (\mathbf{T}, \mathcal{E})$ is an acyclic directed graph (ADG), because there are no undirected edges between chain components. The LWF Markov property defined on this ADG is the block recursive property of the original chain graph $\mathcal{G}$ (Andersson et al., 2001; Levitz et al., 2001).

Definition 2.3.8. *(LWF block recursive Markov property on chain graphs) Let* $\mathcal{D} = (\mathbf{T}, \mathcal{E})$ *be the chain component induced acyclic directed graph of chain graph* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$*. A probability measure* P *on* $\mathcal{G}$ *is said to have the LWF block recursive Markov property if and only if*

(B1) *the conditional distribution* $P_{\tau|pa_{\mathcal{D}}(\tau)}$ *on any chain component* τ *is global Markovian relative to the undirected graph* $\mathcal{G}_\tau$*, the subgraph of* $\mathcal{G}$ *induced by* τ*; and*

(B2) *for any chain component* $\tau \in \mathbf{T}$*,*

$$\tau \perp\!\!\!\perp \{nd_{\mathcal{D}}(\tau) \backslash pa_{\mathcal{D}}(\tau)\} \mid pa_{\mathcal{D}}(\tau); \text{ and}$$

(B3) *for any chain component* $\tau \in \mathbf{T}$*, and any node* $v \subseteq \tau$*,*

$$v \perp\!\!\!\perp \{pa_{\mathcal{D}}(\tau) \backslash pa_{\mathcal{G}}(v)\} \mid bd_{\mathcal{G}}(v).$$

Lauritzen et al. (1990) and Frydenberg (1990) showed that the block recursive and global Markov properties are equivalent on chain graphs. Hence these Markov properties can be used to summarize the conditional independence structures on chain graphs.

2.3.4 AMP Markov properties for chain graphs

Unlike the case for undirected graphs, Markov properties on chain graphs are not unique; *i.e.*, the same chain graph may represent different conditional independences. In a series of papers published by Andersson et al. (1997a,b, 2001), an alternative definition, the AMP Markov property, was proposed to provide another way of specifying the conditional independences of graphs and a different approach building models. The difference between LWF and AMP Markov properties is slight yet important. If we alter (B3) in Prop. 2.3.8 into (B3*) below, then the set of conditional independence conditions (B1), (B2), and (B3*) will induce the alternative AMP Markov property on the same graph.

(B3*) For any chain component $\tau \in \mathbf{T}$, and any node $v \subseteq \tau$,

$$v \perp\!\!\!\perp \{pa_{\mathcal{D}}(\tau) \backslash pa_{\mathcal{G}}(v)\} \mid pa_{\mathcal{G}}(v).$$

Definition 2.3.9. *(AMP block recursive Markov property on chain graphs) Let* $\mathcal{D}_{\mathcal{G}} = (\mathbf{T}, \mathcal{E})$ *be the chain component induced acyclic directed graph of chain graph* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$. *A probability measure* P *on* $\mathcal{G}$ *is said to have the AMP block recursive Markov property if and only if it holds for* (B1), (B2) *and* (B3*).

The nominal difference between (B3) and (B3*), or, the difference between LWF and AMP Markov properties, lies in the conditioning subsets of the block recursive conditions. While in LWF the subset is $bd_{\mathcal{G}}(v)$, in AMP the set of neighbors of v, $nb_{\mathcal{G}}(v)$, is deleted from the conditioning subset, leaving $pa_{\mathcal{G}}(v)$ only. This change alters the conditional independence and the structure of the graphical model substantially, even for very basic and simple graphs. Consider Figure 2.2 from Andersson et al.

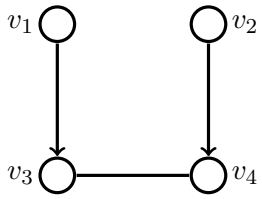
(2001) for example. It is a chain graph with four nodes, two directed edges and one undirected edge. Its chain components are $\mathbf{T} = \{v_1, v_2, \{v_3, v_4\}\}$. If we consider $v = v_3$, it belongs to the chain component $\tau = \{v_3, v_4\}$. In this case we may write its parent and boundary sets as $pa_{\mathcal{G}}(v_3) = v_1$, $bd_{\mathcal{G}}(v_3) = \{v_1, v_4\}$, and the parent set of the chain component $pa_{\mathcal{D}}(\tau) = pa_{\mathcal{D}}(v_3, v_4) = \{v_1, v_2\}$. Based on (B3) and the LWF property, we have

$$v_3 \perp\!\!\!\perp v_2 \mid \{v_1, v_4\}, \tag{2.9}$$

while based on (B3*) and the AMP property, we have

$$v_3 \perp\!\!\!\perp v_2 \mid v_1. \tag{2.10}$$

Figure 2.2: A simple chain graph with different LWF and AMP Markov properties



Although LWF and AMP properties are both useful for model development, Equation (2.9) and (2.10) suggest that, apart from a notable exception, the two Markov properties give, for the same graph, different statistical models. The exception occurs when there is no directed edge in the graph, or, when the graph is undirected. Since undirected graphs are special cases of chain graphs, the chain graph Markov properties also apply to undirected graphs and they correspond to the UG Markov property. We summarize this exception by a new result, Theorem 2.3.10.

Theorem 2.3.10. *When a chain graph does not possess a directed edge; i.e., when it is an undirected graph,*

$$\text{LWF} \iff \text{AMP} \iff \text{UG } \textit{Markov property.}$$

Proof. To show that LWF $\iff$ AMP, notice that in an undirected graph $pa_{\mathcal{D}}(\tau) = \emptyset$ so (B3) and (B3*) are irrelevant. Also the moralization of an undirected graph equals itself, $\mathcal{G}^M_{An(\mathbf{A},\mathbf{B},\mathbf{S})} = \mathcal{G}$, therefore (CG) $\iff$ (UG) and hence LWF $\iff$ UG Markov property. □

Apart from undirected graphs, the conditional independence disparity between LWF and AMP properties on chain graphs suggests the following: There is no one-to-one equivalence between a chain graph and a statistical model. This can be interpreted in two ways. First, we can claim that we may construct different statistical models based on the same chain graph using either LWF or AMP Markov properties. Again let us look at the example of Figure 2.2. Under the LWF Markov property and Equation (2.9), assuming the probability distribution π at each node is positive and continuous, we can build a model based on the conditional independence

$$\pi_{\mathcal{G}}(\mathbf{V}) = \pi(v_3 \mid v_1, v_4)\pi(v_2 \mid v_1, v_4)\pi(v_1, v_4). \tag{2.11}$$

However, this equation is not true under the AMP property. Instead, we have

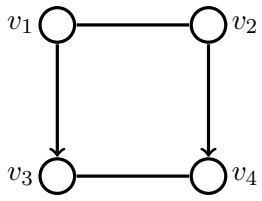
$$\pi^*_{\mathcal{G}}(\mathbf{V}) = \pi(v_4 \mid v_1, v_2, v_3)\pi(v_3 \mid v_1)\pi(v_2)\pi(v_1). \tag{2.12}$$

The two Markov properties have provided us alternative options to model the data

dependence structure. Both properties can be applied differently in practice, and in some cases it is more convenient to work with a particular one than the other. Equation (2.12) is also known as the *recursive linear representation* (Andersson and Perlman, 1998) to emphasize the fact that its nodes are added to the joint distribution recursively. For multivariate Gaussian models, Andersson et al. (2001) show that it is easier to specify "a direct mode of data generation" using the recursive linear representation. Hence for Gaussian cases, it is easier to generate the data based on Equation (2.12) under AMP property than Equation (2.11) under LWF property.

The second revelation we have from the discrepancy is that for both LWF and AMP properties, there can be more than one distinct chain graph that induces the same conditional model. For instance, although different by structure, both Figure 2.2 and Figure 2.3 imply $v_3 \perp\!\!\!\perp v_2 \mid \{v_2, v_4\}$ and Equation (2.11) under the LWF Markov property.

Figure 2.3: An LWF-equivalent chain graph to Figure 2.2.



This intriguing feature of equivalent classes of chain graphs has been studied by Cox and Wermuth (1993) and Drton and Eichler (2006). Although from a statistician's point of view, the subject and interest of the study is the actual statistical model, not the picture of the graph, we are still interested in the uniqueness of graph structures. For example, one may raise the following question: if there are multiple chain graphs

that may produce the same set of conditional independences, will these chain graphs have the same efficiency in modeling the data and the same risk in Bayesian prior specifications (Madigan et al., 1996)?

To answer these types of questions, Frydenberg (1990) introduced the LWF equivalence class of chain graphs represented by the *largest chain graph*, while Roverato (2006) and Andersson and Perlman (2006) implemented similar ideas of AMP-equivalence classes using the *largest deflagged graph*. We try to explore the equivalence class among chain graphs through the study of their moralizations. Because all the different chain graphs in an equivalence class may imply the same conditional model, and in the mean time, a single chain graph may be used to represent multiple models. This is the discrepancy we need to keep in mind when building spatial dependence models. In the next section, we demonstrate the use of LWF and AMP Markov properties on spatial graphical models with both Gaussian and discrete distributions, and illustrate the cases where we can write an estimable joint distribution function based on these properties.

2.4 Denoting spatial dependency using graphical models

Our continuing objective is to find a spatial model with which we can make inference for spatial dependences and within-location associations simultaneously. In its most general form the model should accommodate both the regression-like and other types of within-location associations. We also want it to be applicable when some of the nodes (random variables at locations) are discrete.

Collected by field scientists or remote sensing technology (Polastre et al., 2004;

Wang, 1994), spatial data typically resemble a collection of observations recorded on a finite set of distinct locations rather than on one location only. In some studies, measurements also repeat over time, leading to *spatial-temporal data*. In many studies involving spatial data, there will be various measurements at a single location for the purpose of regression-type inferences or predictions. Examples of this type of spatial research are numerous and cross many disciplines. To name a few, they vary from ecological studies such as sea lion abundance on the coast line of Alaska (Pitcher et al., 2001), to mapped incidences of health issues and alcoholic diseases in Finland (Vanhatalo and Vehtari, 2007), and even to spatial prediction of the hiding place of Osama bin Laden based on intelligence and environmental information (Gillespie and Agnew, 2009).

When data are collected at multiple locations, over various time points, and with various measurements at each location, it is only natural that we want to make inference while accounting for these associations; *i.e.*, to build a single model that may account for the between-location spatial association, within-location multivariate association, and any temporal effect simultaneously. In our work, we will not address the temporal component of the model, but rather the spatial and regression-type components only. We assume all models in the following narrative apply to spatial systems at a fixed time point.

In this section we consider isomorphic chain graphs (Gitelman and Herlihy, 2007), which incorporate spatial dependence while retaining an elegant and relatively simple form within sites. In Section 2.5, we demonstrate that for multivariate Gaussian cases these isomorphic chain graphs can be connected to well known multivariate spatial models for making inferences. However, for non-Gaussian cases, there might not always

be an existing, estimable spatial model that can be assembled from the graph. We will further discuss how to deal with these cases in Chapter 3.

2.4.1 Spatial graphical models

A *spatial graphical model* is a graphical model defined on a spatial domain. In this context, each node in the graph is used to represent a random variable measured at a particular location. We call the locations where data are observed *sites*. Spatial dependences and other types of associations between variables are denoted by edges in the graph. By incorporating between-site and within-site associations under an universal framework, the node-and-edge diagram, we will represent the whole spatial data set using a single graph.

We denote each site with observed data of the spatial domain as $s_1, s_2, \ldots, s_n, n < \infty$. At each site s_i we assume that the same number of variables are observed. Denote the variables (nodes) at site s_i as $Y(s_i), X_1(s_i), \ldots X_p(s_i)$. Using this notations, we define the following subsets:

$$\begin{aligned}
\mathbf{Y} &= \{Y(s_1), Y(s_2), \ldots, Y(s_n)\}; \\
\mathbf{X}_j &= \{X_j(s_1), X_j(s_2), \ldots, X_j(s_n)\}, \quad j = 1, \ldots, p; \\
\mathbf{X}(s_i) &= \{X_1(s_i), X_2(s_i), \ldots, X_p(s_i)\}; \\
\mathbf{Z}(s_i) &= \{Y(s_i), X_1(s_i), X_2(s_i), \ldots X_p(s_i)\}, \quad i = 1, \ldots, n.
\end{aligned}$$

Overall, $\mathbf{V} = \{\mathbf{Y}, \mathbf{X}_1, \ldots \mathbf{X}_p\}$ is the node set of a spatial graphical model with $n \times (p+1)$ nodes, and together with the edge set, $\mathbf{E}$, it forms a graph, $\mathcal{G} = (\mathbf{V}, \mathbf{E})$.

We denote the subgraph induced by $\mathbf{Y}$ as $\mathcal{G}_{\mathbf{Y}} = (\mathbf{Y}, \mathbf{E}_{\mathbf{Y}} = \mathbf{E} \cap \{\mathbf{Y} \times \mathbf{Y}\})$, and similarly for $\mathbf{X}_j$, $\mathcal{G}_{\mathbf{X}_j}$. Site induced subgraphs are denoted as $\mathcal{G}_{s_i} = (\mathbf{Z}(s_i), \mathbf{E}(s_i) = \mathbf{E} \cap \{\mathbf{Z}(s_i) \times \mathbf{Z}(s_i)\})$.

The edges in $\mathbf{E}$ can explain both spatial dependency and within-site multivariate associations. For instance, the existence of edges between nodes from different sites, such as $Y(s_1) - Y(s_2) \in \mathbf{E}$, indicates that sites s_1, s_2 are spatially dependent, whereas edges such as $X_1(s_1) \rightarrow Y(s_1) \in \mathbf{E}$ denote bivariate associations between Y and X_1 at site s_1. For both types of associations directed or undirected edges are permitted.

Since directed edges usually denote "one-way" associations, and undirected edges denote "two-way" associations, they can naturally be adopted to represent regression effects or multivariate correlations between the nodes, respectively. This flexibility means that the graphical model framework can accommodate a variety of spatial models. One of the examples is the spatial regression model. If for each site there is one and only one node, namely $Y(s_i)$, such that $nb(Y(s_i)) \cap \mathbf{Z}(s_i) = \emptyset$, $de(Y(s_i)) = \emptyset$, *i.e.*, it does not have a child node and the other nodes from the same site are connected to it via directed edges, then we may name $\mathbf{Y}$ as the "response" of the model, while the other nodes at each site are p different "predictor" variables. In this scenario the graph would be considered as a spatial graphical model, On the other hand, if all edges in $\mathcal{G}$ are undirected, the graphical model can then be interpreted as a spatial multivariate system with its local and global structure visualized by the different end points of the edges (Thioulouse et al., 1995).

2.4.2 Isomorphic chain graphs

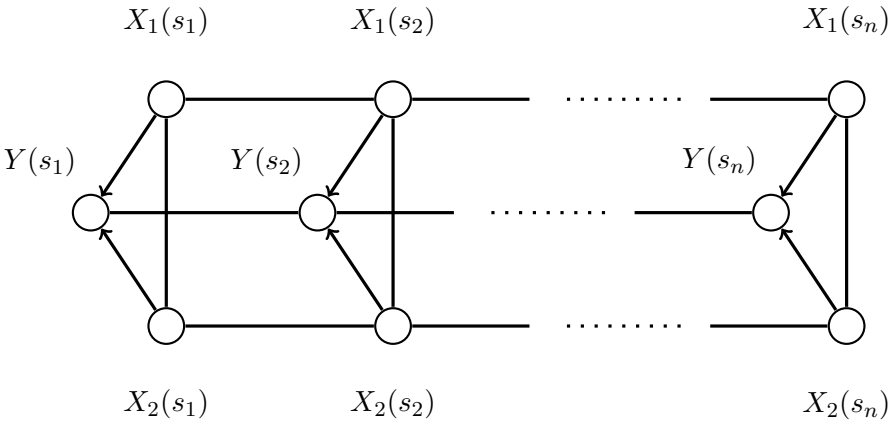
Looking at the site subsets of the chain graph, $\mathbf{Z}(s_i)$, individually, we see that the size of the subsets are equal ($p + 1$ in the previously defined graph $\mathcal{G}$). We might go one step further by assuming that their *structures* (*i.e.*, edge sets), are the same as well. For many problems this is an intuitive and natural solution, as it can be prohibitively demanding if each individual site has a unique graphical structure (it would greatly increase the number of unknown parameters to estimate.)

Formalizing this structural assumption, Gitelman and Herlihy (2007) call the graphical models that have the same within-site structure isomorphic chain graphs. We say a graph, $\mathcal{G}$, is *isomorphic* if every site induced subgraph, $\mathcal{G}_{s_i}$, is identical in the sense that they have identical edge sets and node sets.

Figure 2.4 gives an example of an isomorphic chain graph. In the graph, each triangular structure represents a site with three nodes, $\mathbf{Z}(s_i) = \{Y(s_i), X_1(s_i), , X_2(s_i)\}$. In structure, $\mathcal{G}_{s_1}$, $\mathcal{G}_{s_2}, \ldots, \mathcal{G}_{s_n}$ are equivalent. Notice that the ICG also requires that the type of edges among the same pairs of within-site variables be the same across the sites.

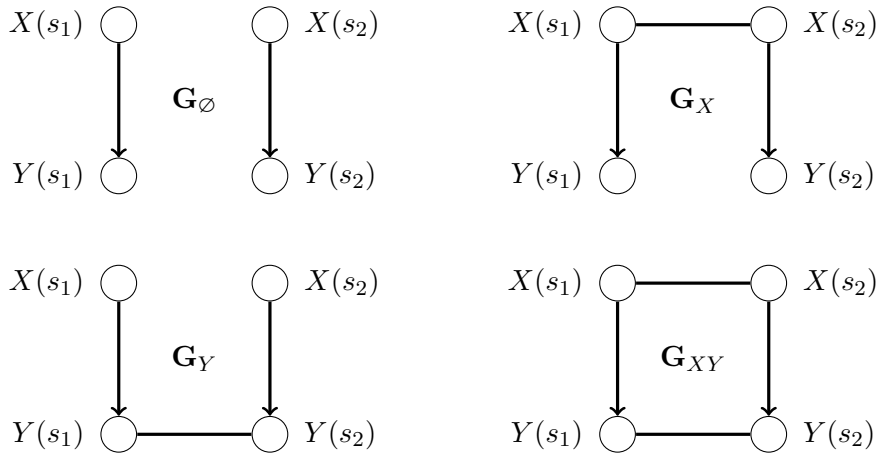
In spatial graphical models the between-site spatial dependence is represented by undirected edges connecting nodes from different sites. We may allow one or more variables to be spatially correlated across sites. In their paper discussing multivariate Gaussian spatial regression models, Irvine and Gitelman (2010) categorize the ICG into four groups according to their spatial associations. When neither the predictors nor the response variables are spatially correlated, in other words, when the data are spatially independent, the graph is denoted by $\mathbf{G}_{\varnothing}$. When some or all of the predictor variables,

Figure 2.4: Example of an isomorphic chain graph



$\mathbf{X}_j$'s, are spatially correlated while the responses are not, the graph is denoted by $\mathbf{G}_X$. Conversely, the $\mathbf{G}_Y$ model is the graph with spatially correlated responses $\mathbf{Y}$ but not $\mathbf{X}_j$'s. Finally, the $\mathbf{G}_{XY}$ model denotes the graphs where both the responses and some or all of the predictor variables are spatially correlated.

Figure 2.5: Isomorphic chain graphs $\mathbf{G}_\varnothing$, $\mathbf{G}_X$, $\mathbf{G}_Y$ and $\mathbf{G}_{XY}$



In Figure 2.5, the four types of isomorphic chain graph are shown in elementary

examples, where each graph has two sites, s_1 and s_2, and each site has only two variables, X and Y. The within-site structures are the same for all sites ($X \to Y$), whereby each graph is isomorphic.

Of the four ICG models presented in Figure 2.5, $\mathbf{G}_{XY}$ attracts our attention the most. The other three ICG, $\mathbf{G}_{\varnothing}$, $\mathbf{G}_X$ and $\mathbf{G}_Y$, can be considered special cases of it. In some spatial data sets more than one variable may be taken to be stochastic. In the context of ICG models, this means that there are multiple nodes in a single site that are connected to the isomorphic nodes at other sites. If we want to estimate multivariate relationships while accounting for spatial dependence of multiple random variables, then models that accommodate multiple spatial dependences are an appropriate way to do it. The demand for multivariate dependence analyses has led to the development of multivariate spatial models in recent years (Anselin, 1988, 2002; Banerjee et al., 2004; Fischer and Getis, 2010). These models should not only allow multiple spatial dependences for its multiple variables, but also be capable of assigning different dependence structures among them because the scales and units of the set of variables recorded at one site may be greatly different (temperature, precipitation, *etc.* for example). Hence it may be reasonable to assume that, in a multivariate spatial regression model, the responses and predictors exhibit different correlation patterns.

The ICG serves as the foundation for us as we build an estimable multivariate spatial model. We must first transform the graphical structures of the ICG into a probability model. The crucial link between this transition from the graph to the model are the conditional independence properties. Using LWF or AMP Markov properties, we may summarize conditional independences among the nodes based on their graphical structures. Some spatial regression models then take advantage of these conditional

independences to formalize statistical models such that we may make inference on the unknown parameters. Such models and the necessary transition process will be discussed in Section 2.5.

2.5 Connecting graphs with spatial models

2.5.1 A graph is not the whole picture

In the previous sections of this chapter, we have established graphical models as an effective and straightforward method to represent associations between multivariate variables across different spatial locations. Thanks to the definitions of nodes-and-edges diagrams, spatial graphical models can denote complicated multivariate spatial relationships using relatively simple and intuitive graphs. It is easier to draw graphs than to develop statistical models to fit them, but the graph alone only gives qualitative description of the data. Just as every spatial statistical model has assumptions, every graph possesses a dependence structure determined by its edges. Careful examination is needed to match a graph to a probability model to ensure that the model appropriately represents the spatial structure of the graph.

Although an ample collection of spatial statistical models have been developed for many types of spatial data, the models are still relatively limited compared to the virtually unlimited possibilities of graphs. Because there is no limit to the scope of the graph other than a finite domain, it can always grow in size and complexity. While it might be possible to write a model equation based on a graph representation, whether the equation can be realistically calculated and estimated is another story. What we

desire is a spatial model with a manageable variance-covariance rendition that remains true to the edges and neighborhood structure of the graph.

In our work we focus on lattice data (Cressie, 1993) particularly, which are defined as a fixed set of locations on a finite and discrete spatial domain. In Section 2.4 we introduced the isomorphic chain graph as a modeling tool to visualize the relationships among multivariate, spatially correlated random variables. When applied to lattices, we want the model to accommodate multiple variables, to handle large sets of nodes and sites, as well as large cliques of neighbors. We also desire a model that allows both Gaussian and non-Gaussian variables at the same time.

The underlying variable associations represented by the edges in the isomorphic chain graph can be summarized from the graphical structure using the Markov properties. In Gaussian cases, bivariate conditional independences correspond to zero entries in a variance-covariance matrix, which can lead us to a specific model. In non-Gaussian cases, although conditional independences do not directly suggest zero entries, they still provide insight to the choice of models.

2.5.2 Conditional independence of ICG models

The conditional independences on a chain graph can be summarized using either LWF or AMP Markov properties, though as we have noted before, each Markov property may induce different sets of conditional independences. Table 2.1 lists the resulting conditional independences from the four simple two-variable, two-site ICG, $\mathbf{G}_{\varnothing}$, $\mathbf{G}_X$, $\mathbf{G}_Y$ and $\mathbf{G}_{XY}$, from Figure 2.5. Irvine (2007) worked out the ICG AMP conditional independences and we added those corresponding to the LWF Markov property. Of

the four ICG models, the conditional independences from $\mathbf{G}_\varnothing$ and $\mathbf{G}_X$ models are identical under both LWF and AMP properties; whereas for the $\mathbf{G}_Y$ and $\mathbf{G}_{XY}$ models they are somewhat different. This disparity is due to the difference of requirement in (B3) and (B3*) in Section 2.3, which stated the conditioning set to be $bd_{\mathcal{G}}(v)$ for LWF property and $pa_{\mathcal{G}}(v)$ for AMP property. Because the parent set $pa_{\mathcal{G}}(v)$ is a subset of the boundary set $pa_{\mathcal{G}}(v)$ in $\mathbf{G}_Y$ and $\mathbf{G}_{XY}$, the AMP property conditioning set (for instance, the first independence of $\mathbf{G}_Y$ is conditioned on $X(s_2)$ only) is included in that of the LWF property (on $X(s_2)$ and $Y(s_1)$).

Table 2.1: Conditional independences of the four ICG in Figure 2.5

ICG Model	**LWF *cond. indep.***	**AMP *cond. indep.***
$\mathbf{G}_\varnothing$	$\{X(s_1), Y(s_1)\} \perp \{X(s_2), Y(s_2)\}$	$\{X(s_1), Y(s_1)\} \perp \{X(s_2), Y(s_2)\}$
$\mathbf{G}_X$	$Y(s_1) \perp\!\!\!\perp \{X(s_2), Y(s_2)\} \mid X(s_1)$ $Y(s_2) \perp\!\!\!\perp \{X(s_1), Y(s_1)\} \mid X(s_2)$	$Y(s_1) \perp\!\!\!\perp \{X(s_2), Y(s_2)\} \mid X(s_1)$ $Y(s_2) \perp\!\!\!\perp \{X(s_1), Y(s_1)\} \mid X(s_2)$
$\mathbf{G}_Y$	$X(s_1) \perp\!\!\!\perp Y(s_2) \mid \{X(s_2), Y(s_1)\}$ $X(s_2) \perp\!\!\!\perp Y(s_1) \mid \{X(s_1), Y(s_2)\}$ $X(s_1) \perp X(s_2)$	$X(s_1) \perp\!\!\!\perp Y(s_2) \mid X(s_2)$ $X(s_2) \perp\!\!\!\perp Y(s_1) \mid X(s_1)$ $X(s_1) \perp X(s_2)$
$\mathbf{G}_{XY}$	$X(s_1) \perp\!\!\!\perp Y(s_2) \mid \{X(s_2), Y(s_1)\}$ $X(s_2) \perp\!\!\!\perp Y(s_1) \mid \{X(s_1), Y(s_2)\}$	$X(s_1) \perp\!\!\!\perp Y(s_2) \mid X(s_2)$ $X(s_2) \perp\!\!\!\perp Y(s_1) \mid X(s_1)$

The differences between the two sets of independences increase when more sites and variables are added to the model. This has provided a challenge yet an opportunity for us. It is a challenge in the sense that when they differ we interpret the same chain graph's structure using different spatial models. At the same time, it is also an opportunity since different conditional independences may suggest additional modeling options.

2.5.3 Connecting multivariate spatial models: achievements and problems

The different types of ICG models correspond to different spatial models under an assumption of normality. Irvine and Gitelman (2010) have shown that under the AMP Markov property, $\mathbf{G}_{XY}$ ICG models can be matched with the separable model (Banerjee et al., 2004) and the linear model of coregionalization (LMC) (Schmidt and Gelfand, 2003); and that $\mathbf{G}_X$ ICG can be matched with separable models under additional conditions. Irvine (2007) has also pointed out that under the AMP property these ICG cannot be connected with the simultaneous autoregressive (SAR) (Whittle, 1954) nor the conditional autoregressive (CAR) models (Besag, 1974). The key difference between the models is that the spatial regression models (SAR and CAR) allow only one variable, usually the response, $\mathbf{Y}$, to be stochastic. Their explanatory variables are considered to be fixed. The separable model and LMC model permit both $\mathbf{Y}$ and $\mathbf{X}$ to be stochastic and spatially dependent.

In this thesis, we establish connections between Gaussian $\mathbf{G}_{XY}$ graphs and multivariate conditional autoregressive (MCAR) (Mardia, 1988) models under the LWF Markov property. The MCAR model allows multiple spatially dependent variables and fully utilizes the conditional distributions induced by Markov properties. For the other two types of ICG, the $\mathbf{G}_Y$ graphs may be viewed as univariate spatial regression models with fixed covariates, while for $\mathbf{G}_\varnothing$ the spatial structure is irrelevant.

To show the connection between $\mathbf{G}_{XY}$ and MCAR, we now give a brief introduction to the various spatial models mentioned above, starting with univariate ones.

Simultaneous autoregressive models (SAR) are sometimes also referred to as *spatial*

error models (Fischer and Wang, 2011, SEM). They were first developed by Whittle (1954). The SAR model is frequently used in ecological studies (Cliff and Ord, 1981; Cressie, 1993; Legendre and Fortin, 1989; Haining, 2003; Kissling and Carl, 2008). The simultaneous autoregressive model can be written as

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \boldsymbol{\epsilon}, \quad \text{where} \\ \boldsymbol{\epsilon} &= \lambda \mathbf{W}\boldsymbol{\epsilon} + \mathbf{u}. \end{aligned} \tag{2.13}$$

In this model $\mathbf{X}$ is considered fixed, and $\mathbf{Y}$ stochastic. The spatial structure is specified in its error term $\boldsymbol{\epsilon}$. λ is the autoregressive parameter, and $\mathbf{W}$ is an $n \times n$ adjacency matrix, with $W_{ij} \neq 0$ if and only if $Y(s_i)$ and $Y(s_j)$ are neighbors. $\mathbf{u}$ is the spatially independent errors. Assume $\text{var}(\mathbf{u}) = \boldsymbol{\Sigma}_{\mathbf{u}}$, the joint distribution of $\mathbf{Y}$ is then

$$\begin{aligned} \mathbf{Y} &\sim \text{MVN}(\mathbf{X}\beta, \boldsymbol{\Sigma}_{\mathbf{YY}}^{(SAR)}), \\ \boldsymbol{\Sigma}_{\mathbf{YY}}^{(SAR)} &= (\mathbf{I}_n - \lambda \mathbf{W})^{-1} \boldsymbol{\Sigma}_{\mathbf{u}} (\mathbf{I}_n - \lambda \mathbf{W}')^{-1}. \end{aligned}$$

In contrast to the SAR, which models the spatial dependence of $\mathbf{Y}$ via its error terms, the *spatial lag model* (Anselin, 2002) deals with the same types of situations (where only the responses are correlated), but models its dependence in its mean structure. The model can be specified as

$$\mathbf{Y} = \lambda \mathbf{W}\mathbf{Y} + \mathbf{X}\beta + \epsilon, \tag{2.14}$$

where $\epsilon_i \sim i.i.d.\ \mathrm{N}(0, \sigma_\epsilon^2)$. Its joint distribution is

$$\begin{aligned} \mathbf{Y} &\sim \mathrm{MVN}([\mathbf{I}_n - \lambda\mathbf{W}]^{-1}\mathbf{X}\beta, \boldsymbol{\Sigma}_{\mathbf{YY}}^{(SLM)}), \\ \boldsymbol{\Sigma}_{\mathbf{YY}}^{(SLM)} &= \sigma_\epsilon^2[(\mathbf{I}_n - \lambda\mathbf{W})'(\mathbf{I}_n - \lambda\mathbf{W})]^{-1}. \end{aligned}$$

Both the simultaneous autoregressive model and spatial lag model are by definition univariate models, and, as noted by Irvine and Gitelman (2010), cannot be matched with any $\mathbf{G}_X$, $\mathbf{G}_Y$, or $\mathbf{G}_{XY}$ under the AMP Markov property.

The conditional autoregressive model (CAR) was first proposed by Besag (1974) and arose from the study of Markov properties based on lattice data. Unlike the simultaneous autoregressive model and the spatial lag model, the CAR model specifies the joint distribution of the lattice not through its unconditional likelihood function, but rather its conditional distributions. This makes it very appealing for the graphs which summarize conditional independences. For Gaussian distributed random variables $Y(s_i)$, the CAR model defines its conditional distribution given all the other Y variables, $\mathbf{Y}(s_{-i})$, as multivariate Gaussian:

$$Y(s_i) \mid \mathbf{Y}(s_{-i}) \sim \mathrm{N}\left[\mathbf{X}(s_i)'\beta + \sum_{j \neq i} \psi\omega_{ij}\left(\mathbf{Y}(s_j) - \mathbf{X}(s_j)'\beta\right), \sigma_i^2\right], \quad i = 1, 2, \ldots, n \tag{2.15}$$

where ψ is the dependent parameter, and ω_{ij}'s are the corresponding adjacency elements from matrix $\mathbf{W}$. The dependent parameter measures the magnitude of spatial effects between neighboring $Y(s_i)$'s, and $\mathbf{W}$ determines where they occur. Besag (1974)

showed that the conditional model is equivalent to its unconditional form,

$$\begin{aligned} \mathbf{Y} &\sim \text{MVN}\left(\mathbf{X}'\beta, \boldsymbol{\Sigma}_{\mathbf{YY}}^{(CAR)}\right), \quad \text{with} \\ \boldsymbol{\Sigma}_{\mathbf{YY}}^{(CAR)} &= (\mathbf{I}_n - \psi\mathbf{W})^{-1}\mathbf{D}, \quad \mathbf{D} = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2). \end{aligned}$$

Results based on CAR models are sensitive to the choice of adjacency matrix $\mathbf{W}$ (Earnest et al., 2007). The model has a close connection to the SAR model. Haining (1993) and Wall (2004) explored similarities and differences between SAR and CAR models. For applications with multivariately correlated random variables, Kittler and Föglein (1984); Mardia (1988) generalized the univariate CAR models to the *multivariate conditional autoregressive* (MCAR) models. These are further expanded in hierarchical data scenarios to *generalized multivariate CAR* (GMCAR) (Jin et al., 2005) and *coregionalized CAR* (Sang and Gelfand, 2009).

We have discovered that, on bivariate lattices, the $\mathbf{G}_{XY}$ graphs can be connected with Gaussian MCAR models using LWF Markov property. To see this, we remember the form of conditional independence on chain graphs under LWF Markov property. In the two site example $\mathbf{G}_{XY}$ in Table 1, we have $X(s_1) \perp\!\!\!\perp Y(s_2) \mid \{X(s_2), Y(s_1)\}$ holds. Generalizing this to n sites using Markov property (CL), we have

$$X(s_i) \perp\!\!\!\perp Y(s_j) \mid \{X(s_j), nb[Y(s_j)]\}, \tag{2.16}$$

where s_i and s_j are adjacent sites and $nb[Y(s_i)]$ is the neighboring set of nodes of $Y(s_j)$ (we say two sites are adjacent, meaning they share an undirected edge, so that it is not to be confused with neighbor nodes). Observing Equation (2.16), we can see that it holds true for both $\mathbf{G}_{XY}$ graphs with directed edges $X(s_i) \to Y(s_i)$, and

undirected edges $X(s_i) - Y(s_i)$ within sites (if the undirected graph is defined such that $X(s_j) \in nb[Y(s_j)]$). This conditional independence equivalence is no coincidence, and will be formalized by the moralization process in Section 3.1.

From Equation (2.16) and the conditional independence axioms in Proposition 2.3.1, we can further write the conditional independence among sites (*i.e.*, among pairs of nodes $\{X(s_i), Y(s_i)\}$ in the graph). Specifically, $\{X(s_i), Y(s_i)\}$ is conditionally independent from $\{X(s_j), Y(s_j)\}$ given the adjacent sites (for non-adjacent $s_i \neq s_j$). Mathematically, if $\mathbf{V}(s_i) = \{X(s_i), Y(s_i)\}$ and define the adjacent sites $nb[\mathbf{V}(s_i)] \equiv \{nb[X(s_i)] \cup nb[Y(s_i)]\}$, then

$$\mathbf{V}(s_i) \perp\!\!\!\perp \mathbf{V}(s_j) \mid nb[\mathbf{V}(s_i)], \quad \text{when } s_i \text{ and } s_j \text{ are non-adjacent.} \tag{2.17}$$

Equation (2.17) suggests that we may parameterize spatially correlated Gaussian random variables to match with $\mathbf{G}_{XY}$ under the LWF Markov property because they have the same conditional independences specified in MCAR models. The MCAR is defined by the conditional distributions of each random vector (site) given its adjacent sites. For a finite lattice of n sites with zero centered, Gaussian distributed random site $\mathbf{V}(s_i)$,

$$\mathbf{V}(s_i) \mid \mathbf{V}(s_{-i}) \sim \mathrm{N}(\sum_{j=1}^{n} \mathbf{C}_{ij}\mathbf{V}(s_j), \mathbf{\Gamma}_i). \tag{2.18}$$

s_{-i} denotes all the other sites but s_i. Both $\mathbf{C}_{ij}$ and $\mathbf{\Gamma}_i$ are $p \times p$ matrices, with p equal to the length of the random vector. In our case, $p = 2$ because we only have two random variables per site. $\mathbf{C}_{ij}$ measures the spatial dependence between sites s_i

and s_j and $\mathbf{C}_{ii} = \mathbf{0}$. The LWF Markov property in Equation (2.17) commands that $\mathbf{C}_{ij} = \mathbf{0}$ when s_i and s_j are nonadjacent and $\mathbf{C}_{ij} \neq \mathbf{0}$ when they are adjacent. $\mathbf{\Gamma}_i$ is the conditional variance-covariance matrix of $\mathbf{V}(s_i)$ and since the graph is isomorphic, we may write

$$\mathbf{\Gamma}_i \equiv \mathbf{\Gamma}_0 = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}.$$

For $\mathbf{G}_{XY}$ ICG models, Irvine (2007) showed that

$$\mathbf{\Gamma}_0 = \mathbf{\Gamma}_{XY} = \begin{pmatrix} \sigma_x^2 & \beta\sigma_x^2 \\ \beta\sigma_x^2 & \sigma_y^2 \end{pmatrix}.$$

Similarly, $\mathbf{C}_{ij}$ is also constant when s_i and s_j are adjacent and we may define $\mathbf{C}_{ij} \equiv \mathbf{C}^{adj}$. Mardia (1988) proved that the joint distribution of the above model is uniquely determined by the stated conditional distributions and can be written as

$$\mathbf{V} \sim \mathrm{N}\left[\mathbf{0}, (\mathbf{I}_n - \mathbf{C})^{-1}\mathbf{\Gamma}\right], \tag{2.19}$$

where $\mathbf{V}' = \{\mathbf{V}(s_1)', \mathbf{V}(s_2)', \ldots, \mathbf{V}(s_n)'\}$. $\mathbf{C}$ is a $2n \times 2n$ matrix with the ij-th block $\mathbf{C}_{ij}$, and $\mathbf{\Gamma}$ is the block diagonal matrix $\mathbf{\Gamma} = \mathrm{diag}(\mathbf{\Gamma}_{XY}, \mathbf{\Gamma}_{XY}, \ldots, \mathbf{\Gamma}_{XY})$. The diagonal blocks in $\mathbf{C}$ are all zero, and on the off-diagonal side only adjacent s_i, s_j will yield a $\mathbf{C}_{ij} \neq \mathbf{0}$. We can write

$$\mathbf{C} = \mathbf{W} \otimes \mathbf{C}^{adj}, \quad \mathbf{\Gamma} = \mathbf{I}_n \otimes \mathbf{\Gamma}_{XY},$$

where $\mathbf{W}$ is the adjacency matrix as defined in Equation (2.13). The "$\otimes$" sign in the equation stands for the matrix Kronecker product. This MCAR model is a valid

match for Gaussian $\mathbf{G}_{XY}$ as long as the $\mathbf{C}_{ij}$'s are determined based on the conditional independences in Equation (2.17) and $(\mathbf{I}_n - \mathbf{C})^{-1}\mathbf{\Gamma}$ is positive definite. For three or more spatially dependent random variables, the matching between $\mathbf{G}_{XY}$ and MCAR models also holds and the proof is analogous.

Besides the Gaussian ICG, the matching to the spatial models is not straightforward when graphs have discrete nodes. We propose an entirely new approach for such chain graphs, by first transforming them to the easier-to-handle undirected graphs, and then partitioning them into smaller components to be evaluated and modeled separately. These partitions heavily rely on the local Markov properties, cliques, and the Hammersley-Clifford Theorem. This is the topic of Chapter 3.

2.6 Discussion

Existing literature primarily utilize prediction methods such as the co-kriging model for multivariate spatial system. On lattices graphical model permits spatial error terms on more than one variable and provides useful alternative to co-kriging. Based on the Markov property on undirected graphs or the LWF and AMP Markov properties on chain graphs, the conditional independences may be summarized to construct the joint distribution function. When the joint distribution matches with some spatial regression models, their estimable distribution function may be used to help obtain estimates on the graphs.

This chapter showed that the MCAR, separable or LMC models can transform some of the ICG model dependence structures to the estimable probability models under Gaussian cases. Besides these discussed models there is still a very wide spectrum

of models worth exploring. Examples include the Spatial Durbin model, an augmented spatial lag model (LeSage and Pace, 2009; Fischer and Wang, 2011), local indicators for spatial association model (LISA) developed by Anselin (1995), and Ising, auto-binomial, and auto-Poisson models for binomial and discrete spatial data (Anselin, 1988; Florax and Folmer, 1992; Cressie, 1993; Fischer and Getis, 2010), among others.

There are a few problems we would like to point out that may hinder the general applications of these models. The first and foremost problem is, none of the models have fully addressed the correspondence in non-Gaussian or discrete cases. In Gaussian situations, conditional independence among two nodes in the graph may be effortlessly translated as a zero correlation coefficient element in the variance-covariance matrix of the regression or multivariate model. There is, however, no equivalence between non-Gaussian conditional independence and the variance-covariance element and such absence denies the non-Gaussian spatial data with many of the most useful spatial models.

The second problem is, the model matching is always graph-dependent. In every situation a spatial model may only correspond to one or a few types of isomorphic graphs. There is no "universal" spatial model that is adapted to all four types of isomorphic chain graphs, even under Gaussian cases. Part of the reason for it is because sometimes the graph assumes that all its nodes are stochastic, while the model is only univariate. The other reason for it includes the dependence of the variance-covariance matrix on the graph structure under either LWF or AMP Markov properties. A change in the pattern of the graph, for example from regular to irregular lattice sites, may render a new variance-covariance matrix and demand a new spatial model.

Moreover, computational issues also affect the practicability of the multivariate

and regression models. Dealing with the whole spatial domain and data set all at once requires long computing time and powerful resources, for the sake of tasks such as the inversion of very big matrices like $[\mathbf{I}_n - \lambda\mathbf{W}]^{-1}$.

This connection to a spatial model is crucial for the graphs. An important generalization would be expanding it to the non-Gaussian models, a topic which will be heavily covered in Chapter 3 and 4. An equally consequential expansion is to establish similar links for graphical structures other than ICG. This may include geostatistical and point patterns data, as well as lattice with distinct site structures. In practice each variable may not necessarily be recorded at every site, creating a mismatched set of sites. Caetano et al. (2006) suggested an approach on the point pattern data, although they have used a jitter to connect the graph. On geostatistical studies, it is yet unclear whether a similar approach exists due to the continuous domains they possess. On lattices, the question is equally intriguing. For instance, it would be interesting to see whether an estimable joint distribution function may be established with observations on $X(s_{ij}), Y(s_{kl})$ only, but with missing $X(s_{kl})$ and $Y(s_{ij})$.

Chapter 3

Parametric Estimates for Discrete Spatial Graphical Models

3.1 Introduction

In this chapter we develop estimable joint distribution functions for multivariate spatial data that have discrete components. These methods employ a series of graphical model tools, including generalized tree networks, moralization, the Hammersley-Clifford theorem, junction trees, as well as copula models. Collectively they have the ability to transform isomorphic chain graphs (ICG) on regular or irregular lattices into conditional Markov random fields, which can then be estimated using Markov chain Monte Carlo algorithms.

The goal of this chapter is to extend the work commenced in Irvine and Gitelman (2010) to connect ICG representing multivariate spatial data in continuous and discrete cases to estimable models. To achieve this goal we first turn to undirected graphs, or, more specifically, to the moralized undirected graphs corresponding to an ICG.

The reason we choose this path is twofold: a conversion procedure known as moralization allows us to convert most chain graphs into undirected graphs; and the Hammersley-Clifford theorem applied to those moralized graphs allows us to partition the probability distribution on the graph into small components that each involves only a few nodes. In Section 3.2 we introduce the moralization method in detail and give the definition of generalized tree networks. Once converted to undirected graphs, the Hammersley-Clifford theorem ensures that the joint distribution of the whole graph does not depend on any graphical feature other than the maximal cliques. In Section 3.3 we present this pivotal theorem, and in Section 3.4 we give methods for writing the partition functions needed for the final estimable joint distribution function. This can be done seamlessly in Gaussian cases, but non-Gaussian cases require more effort. We

solve this by creating junction trees and conditional Markov random fields, borrowing influences from Jensen et al. (1990) and Besag (1974). Finally, in Section 3.5 we discuss how to model the joint distribution in each clique using copula models.

3.2 Converting chain graphs

3.2.1 Moralization

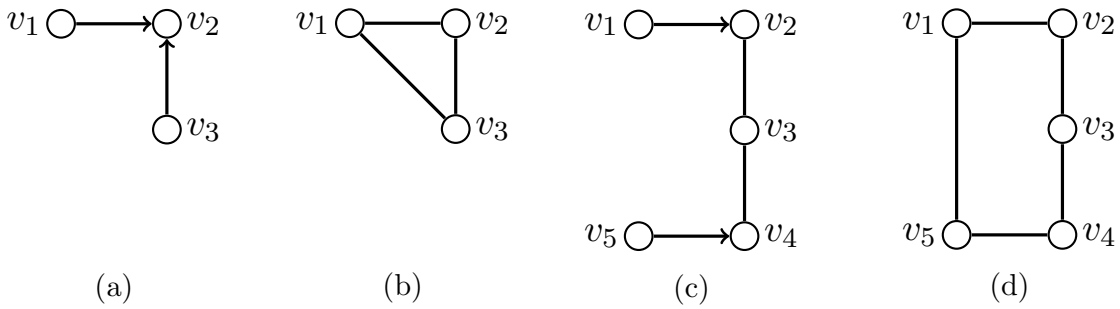
A directed graph or chain graph can be transformed into an undirected graph by dropping all the directed edges from the graph, replacing them with undirected edges, and connecting all parental nodes that share the same child node with undirected edges as well. This procedure is known as moralization. If the original directed graph or chain graph is $\mathcal{G}$, the moralized, undirected graph will be denoted as $\mathcal{G}^M$. The procedure is named as such because by connecting all the parental nodes that share the same children, we are essentially "marrying" them and their children are no longer "illegitimate."

The moralization of chain graphs is conducted at the "complex level" of the graph. We first encountered the concept of a complex in Chapter 2, where a 1-complex was defined as a triplet of nodes $v_1 \rightarrow v_2 \leftarrow v_3$. This triplet is also known as an *immorality*, since the parents, v_1 and v_3, are not directly connected. In the moralizing procedure, such immorality will be converted into a 3-clique, by first replacing $v_1 \rightarrow v_2$ and $v_3 \rightarrow v_2$ with $v_1 - v_2, v_3 - v_2$, and then adding a new undirected edge $v_1 - v_3$. Figure 3.1 (a) and (b) show this immorality before and after moralization.

A three-node subgraph, $v_1 \rightarrow v_2 - v_3$, with one directed and one undirected edge is

known as a *flag* (Frydenberg, 1990). It is moralized by converting $v_1 \rightarrow v_2$ into $v_1 - v_2$, without the connecting edge between v_1 and v_3. The moralization for a k-complex, $v_0 \rightarrow v_1 - \ldots - v_k \leftarrow v_{k+1}$, with $k \geqslant 2$ is analogous to that of the immorality, with only one new edge added between the parents. It is moralized by first replacing $v_0 \rightarrow v_1$, $v_k \leftarrow v_{k+1}$ with undirected edges, and then connecting v_0 with v_{k+1} with an undirected edge. An example for the moralization of a 3-complex is shown in Figure 3.1 (c) and (d). We provide in Figure 3.2 a more complete example of moralization. The parent nodes in graph (a) are first "married," denoted by dashed lines in (b), and then all the directed edges in graph (a) and the dashed "marriages" in (b) are replaced by undirected edges in graph (c). We say graph (c) is the moralized graph of (a) and (a) is the immoral graph of (c).

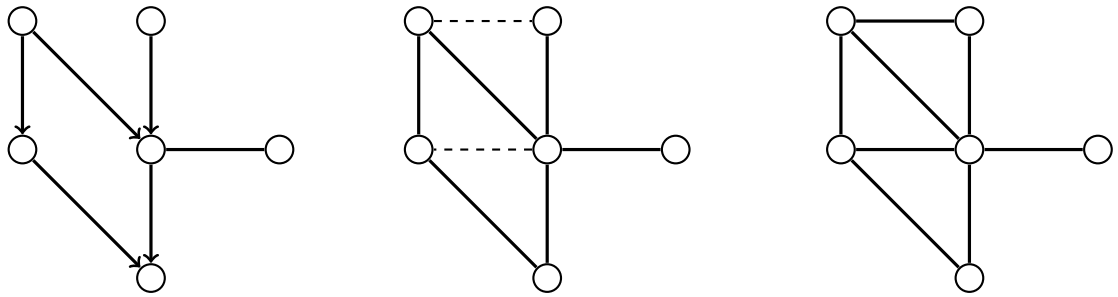
Figure 3.1: Immorality (a) and its moralization (b), 3-complex (c) and its moralization (d)



All chain graphs can be broken into a connected combination of immoralities and k-complexes. No matter how big and complicated the chain graphs might be, they can always be subdivided into these basic elements alongside with undirected subgraphs, and the moralization can then be performed accordingly.

An important consequence of the moralizing procedure, and the reason for us to employ it, is that the conditional independences represented in a moralized undirected

Figure 3.2: Moralization of a chain graph



(a) The immoral CG (b) "Marrying" parents (c) The moralized UG (MRF)

graph are also represented under the LWF Markov property of the immoral chain graph (Lauritzen, 1996; Paskin, 2009). This is not true, however, for the AMP Markov property (Andersson et al., 2001) that summarizes conditional independences of the chain graph differently from the LWF Markov property. Take Figure 3.1 (c) and (d) again for example. The LWF Markov property leads to $v_1 \perp\!\!\!\perp v_4 \mid \{v_2, v_3, v_5\}$ in graph (c), which is the same as in undirected graph (d), whereas at the same time the AMP property in (c) implies $v_1 \perp\!\!\!\perp v_4 \mid v_5$, which is not true in (d).

This property allows the conversion of chain graphs into undirected graphs without further restricting their underlying conditional independence structures. The reverse of the property, however, cannot be stated. There are conditional independences in the immoral graph that are not preserved by its moralized version (Cowell et al., 1999; Paskin, 2009). The loss of some of the conditional independences in the chain graph during moralizing may be considered as a simplification of the dependence structure. We demonstrate in Section 3.4 that by working with the undirected graph, Markov properties provide simple partition of the joint distribution.

We devise the following proposition regarding the structural change caused by the

additional edges of a moralized graph. This result assures that the form of the joint distribution on the original chain graph is a special case of the joint distribution on the moralized undirected graph.

Proposition 3.2.1. *For a triplet of subsets* $\mathbf{A}, \mathbf{B}$ *and* $\mathbf{S}$ *in chain graph* $\mathcal{G}$*, if* $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{S}$ *with respect to* $\mathcal{G}^M_{An(\mathbf{A},\mathbf{B},\mathbf{S})}$*, the moralized undirected graph of the smallest ancestral set containing* $\mathbf{A} \cup \mathbf{B} \cup \mathbf{S}$*, then the conditional independence also holds with respect to* $\mathcal{G}$*.*

Proof. Under LWF Markov property, $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{S}$ implies the global Markov property; *i.e.*, $\mathbf{S}$ separates $\mathbf{A}$ from $\mathbf{B}$ in $\mathcal{G}^M_{An(\mathbf{A},\mathbf{B},\mathbf{S})}$.

Consider any path from $\mathbf{A}$ to $\mathbf{B}$. Because $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{S}$ we can write any path that originates from $\mathbf{A}$ and ends in $\mathbf{B}$ as $\alpha \rightleftharpoons \beta \equiv \alpha \rightleftharpoons s \rightleftharpoons \beta \in \mathcal{G}^M_{An(\mathbf{A},\mathbf{B},\mathbf{S})}$, where $\alpha \in \mathbf{A}$, $\beta \in \mathbf{B}$, $s \in \mathbf{S}$, and $\rightleftharpoons$ denotes the connected components. Moralization only adds edges to the chain graph (either add a new undirected edge, or a $v_1 \to v_2$ to the already existing $v_1 \leftarrow v_2$) but never deletes one. The edge set of the immoral chain graph is included in the edge set of its moralization, or $\mathbf{E}_{\mathcal{G}} \subseteq \mathbf{E}_{\mathcal{G}^M_{An(\mathbf{A},\mathbf{B},\mathbf{S})}}$, which means that there is no path that is only in $\mathcal{G}$ but not in $\mathcal{G}^M_{An(\mathbf{A},\mathbf{B},\mathbf{S})}$.

When $\alpha \rightleftharpoons s \rightleftharpoons \beta \in \mathcal{G}$ the separation rule and global Markov property hold. $\alpha \rightleftharpoons s \rightleftharpoons \beta \notin \mathcal{G}$ implies either path $\alpha \rightleftharpoons s \notin \mathcal{G}$ or $s \rightleftharpoons \beta \notin \mathcal{G}$, which also supports the global Markov property of $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{S}$ in $\mathcal{G}$. Therefore the global Markov property in $\mathcal{G}^M_{An(\mathbf{A},\mathbf{B},\mathbf{S})} \Rightarrow$ the global Markov property in $\mathcal{G}$. □

Proposition 3.2.1 confirms that a moralized graph is loyal to the original graph in that it does not further restrict the conditional independences. This is important for preserving the original correlation structure in the chain graph. The conditional

independences in the moralized graph form a subset of those from the immoral chain graph.

For a majority of chain graphs, moralization provides a direct conversion to Markov random fields. We may encounter some problems, however, with chordless directed cycles in a chain graph. Remember that a *chord* is an edge that connects two non-consecutive nodes in a cycle. Since no two parent nodes in a chordless directed cycle share the same child, a moralized chordless directed cycle will not include any new edges. This can create large, chordless undirected cycles that complicate the partition of the graph, especially for cycles with length of more than 4 or 5. Although directed cycles are not unusual in some spatial-temporal related studies, for instance the predator-prey oscillation patterns (Rosenzweig and MacArthur, 1963) and feedback networks (Spirtes, 1995), we do not consider them here. The between-site dependence of an ICG lattice is instead denoted by undirected edges (Besag, 1975; Cressie and Verzelen, 2008) with directed edges representing only acyclic within-site relationships.

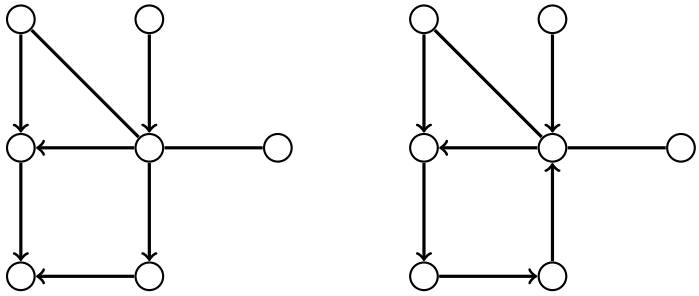
In the following narrative we consider only those chain graphs that do not have chordless directed cycles. This class of chain graph is important to us. We summarize the class under the term generalized tree networks, noting its natural connection to tree networks.

3.2.2 From tree networks to generalized tree networks

We define a *generalized tree network* (GTN) as a chain graph with no chordless directed cycles. In other words, it is a chain graph that permits undirected and semi-directed cycles. Figure 3.3 shows two chain graphs that are identical except for two edges. The

graph in Figure 3.3 (a) is a generalized tree network, whereas Figure 3.3 (b) is not because it possesses a directed 4-cycle.

Figure 3.3: Example (a) and counter-example (b) of generalized tree networks



(a) GTN (b) Not a GTN

The generalized tree network, as its name suggests, essentially generalizes a *tree network* (or sometimes simply a *tree*, TN), which is an undirected graph without cycles (Meilă and Jordan, 2001). In a tree network there is only one path between one node to any other node. We may arbitrarily designate a node in a tree network as the *root* of the graph, and all the other nodes can be subsequently ordered by their distance to the root. The nodes that are connected to only one other node are known as *leaves* of the tree. A *binary tree* is a tree network that each non-leaf node has no more than three neighboring nodes: one on the higher level and two on the lower level (Meilă and Jaakkola, 2006).

The tree network presents one of the easiest structures to work with among graphical models because the hierarchy of the tree translates well into the conditional distributions (Kirshner, 2007). The joint distribution of a tree network can always be partitioned along its edges. This revelation is important to us, since it leads to the Hammersley-Clifford theorem partition. When the connectivity in a tree network is

low, there are some nice results for its partition (see Appendix of Chapter 4). Kirshner (2007) showed that the joint distribution, $f_{\mathcal{T}}(\mathbf{V})$, for a tree network $\mathcal{T} = (\mathbf{V}, \mathbf{E})$ can be written as

$$f_{\mathcal{T}}(\mathbf{V}) = \prod_{v_i - v_j \in \mathbf{E}} g_{ij}(v_i, v_j) \tag{3.1}$$

as long as $g()$'s are bivariate functions of $\{v_i, v_j\}$, $i, j = 1 \ldots n$, the neighboring pairs of nodes in the tree. The product is over all the undirected edges in $\mathcal{T}$. In other words, the joint distribution can be decomposed as the product of pairwise distributions based on the tree. As later generalized by the Hammersley-Clifford theorem, this partition is possible because the pairs $v_i - v_j$ are the maximal cliques (defined in the next section) of the graph.

One of the downsides of the tree networks, however, is the structural constraint of no cycles and binary hierarchy. These are often too restrictive for many practical applications. Efforts have been made to relax these restraints and generalize the findings from tree networks. Chow et al. (1968) developed techniques that expand the scope of tree networks to adapt them to a broader set of problems. Bedford and Cooke (2002) introduced the concept of vines, and Meilă and Jaakkola (2006) used averaged weighted trees over all possible tree structures in a graph. Our introduction of the GTN represents the latest of these exercises.

The focus of our interest is the spatial structure on chain graphs. The GTN accommodates a boarder scope of chain graphs with varied dependency structures. With the exclusion of chordless directed cycles, the GTN can be moralized and then partitioned as a Markov random field. Similar to tree networks, the joint distribution of moralized

GTN can be determined by a product involving conditional distributions.

The connectivity in moralized GTN increases to the point that it is usually impossible to partition it by pairs of nodes. There are often clusters of unbreakable sets of nodes (*i.e.*, cliques) that need to be considered as a whole. Accordingly, instead of writing the joint distribution as a product of marginal distributions of neighboring pairs, $g_{ij}(v_i, v_j)$, we use products of marginal distributions of the maximal cliques in the moralized GTN. This partition follows from the Hammersley-Clifford theorem (Hammersley and Clifford, 1971).

3.3 Undirected graph partition

3.3.1 Cliques

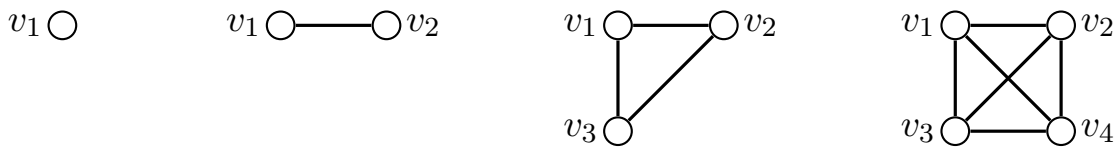
Both the partition of moralized chain graphs and the identification of generalized tree network neighborhood structures rely heavily upon the cliques of the graph. We may also say that it is the cliques that impose the "tree network" on generalized tree networks since they serve the same function in a moralized GTN as pairs of nodes serve in tree networks. As later sections will demonstrate, if we know the distribution of each maximal clique, it is essentially the same to say that we known the joint distribution of the whole graph (conditionally).

Cliques help to summarize the neighborhood structure of an undirected graph. For a node subset C of an undirected graph/Markov random field $\mathcal{G}$, if every node is a neighbor to all the other nodes, the subset is known as a *clique*. Mathematically, $C \subseteq \mathbf{V}$ is a clique iff $C \subseteq \{v, nb(v)\}, \forall v \in C$. If $v \in C$, we say C is a clique on v, and its

size d_C equals the number of nodes in the clique. When a clique has size d we called it a d-clique. Any node, v, can be considered as a clique itself, with $d_v = 1$, and any pair of neighbors $\{u - v\}$ is a clique with $d_{\{u,v\}} = 2$. If C is the largest clique that contains all its nodes, then it is called a maximal clique (Lauritzen, 1996). If C is a maximal clique, then for $\forall v_j \notin C, \exists v_i \in C, v_i - v_j \notin \mathbf{E}$. There is at least one non-neighbor node in C to any other node outside of C. In other words, if you add one more node from $\{\mathbf{V} \backslash C\}$ to the maximal clique C, it will no longer be a clique (if C is a maximal clique and $C \subset C^*$, C^* is not a clique).

Figure 3.4 lists examples of cliques up to size four. These cliques can also be denoted as $C_1 = \{v_1\}$, $c_{12} = \{v_1, v_2\}$, $C_{123} = \{v_1, v_2, v_3\}$, and $C_{1234} = \{v_1, v_2, v_3, v_4\}$.

Figure 3.4: Cliques with $d = 1 - 4$ in lattice neighborhood structures



(a) C_1 (b) C_{12} (c) C_{123} (d) C_{1234}

It is an important task to identify all the cliques and maximal cliques so that we can partition the graph. This is relatively easy to do when the numbers of nodes are few. Also, a regular scheme, such as a regular lattice, makes the clique layout systematic. Identifying cliques becomes more difficult, however, when the number of nodes increases, more edges are connecting to the nodes, and/or when the lattice space does not have a regular neighborhood structure (Tjelmeland and Besag, 1998).

With regular lattice structures, the cliques on the graph can form clusters whose

dependences have intuitive spatial interpretations. Two commonly used structures are first and second order nearest neighbors (Brook, 1964; Besag, 1974). If, in a Markov random field with regular lattice scheme, each node is a neighbor to the nodes directly above, below, and to the left and right of it, the graph is said to have *first order nearest neighbor* structure. If, on top of the first order nearest neighbors, the node is also a neighbor to its four diagonally adjacent nodes, it is called *second order nearest neighbor* structure.

Figure 3.5: First (a) and second order (b) nearest neighbor structures

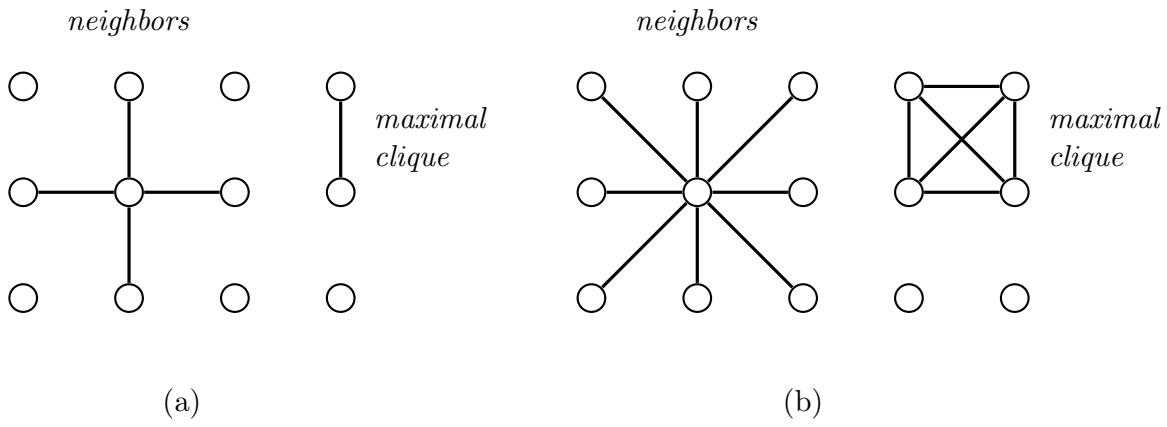


Figure 3.5 shows examples of first and second order nearest neighbor lattices. In the first order nearest neighbor lattice, Figure 3.5 (a), each non-border node has four neighbors, whereas in second order nearest neighbor lattice, Figure 3.5 (b), each one has eight neighbors. For the first order nearest neighbor structure, cliques may have a size of one or a maximal size of two. This can be interpreted in the sense that strong spatial dependence is only observed between nodes with one unit distance (left to right, or up and down). For the second order nearest neighbor structure, cliques may have sizes of one, two, three, or a maximal size of four. We may interpret this structure as

a combination of two by two node blocks, with each pair of nodes within the block having the same level of spatial dependence.

Not only can the cliques be identified in regular lattices, they can also be found in other schemes as well. A notable example is when the nodes are laid out triangularly or following the Delaunay triangulation (Schabenberger and Gotway, 2004). In these graphs cliques have sizes of one and two, with a maximal size of three. This is simply because the graph with a Delaunay triangulation structure is formed by the union of 3-cliques like the one in Figure 3.4 (c).

Cliques play an important part in the partitioning of joint distributions of generalized tree networks. When considering the Markov random field obtained from a moralized generalized tree network, its maximal cliques form a new graph known as the *clique tree* (Barber, 2003). The nodes of the clique tree are the maximal cliques of the original generalized tree network, and thus we call them *clique nodes*. The (undirected) edges among these clique nodes are determined by whether two maximal cliques are adjacent (*i.e.,* whether they share a non-empty intersection in the generalized tree network). These intersections are called the *separators* of the cliques. Mathematically, we can write a clique tree $\mathcal{T}_{\mathcal{C}(\mathcal{G})} = (\mathcal{C}, \mathcal{E}_{\mathcal{C}})$, where $\mathcal{C}$ is the set of all maximal cliques in $\mathcal{G}$, and $\mathcal{E}_{\mathcal{C}}$ is the edge set between cliques: $C_i - C_j \in \mathcal{E}_{\mathcal{C}} \iff C_i \cap C_j \neq \varnothing, C_i, C_j \in \mathcal{C}$. Figure 3.7 and 3.8 in Section 3.4 show some practical examples of clique trees.

Not all moralized GTN can be treated as clique trees, but for those that can, when considering the cliques as nodes, these clique trees are actually tree networks (Barber, 2003). In these graphs we can therefore use the method for tree networks (Silva and Gramacy, 2009) to calculate the graph's joint distribution by partitioning it on its cliques and separators. For those GTN whose moralized Markov random fields are not

clique trees, more work is required to obtain the joint distribution partition, and we address this in Section 3.4.

The adoption of the moralized GTN and the identification of maximal cliques means we have shifted the focus of model building from individual nodes and edges in the chain graph to cliques and their connectivity and adjacency in the moralized GTN. This process is a simplification, since the graphical structure is far less complicated on the clique level than on the whole graph level. The next step is to take advantage of this simplification. The maximal cliques partition the moralized GTN, and the dependence structure of the graph is represented by the adjacency of the cliques and separators. In other words, the Markov random field structure is realized among its cliques through global Markov property and separation rules. Because of the equivalence between global and local Markov properties on undirected graphs, this provides a way to combine marginal clique distributions on local levels.

3.3.2 Hammersley-Clifford theorem

The Hammersley-Clifford theorem, first introduced by Hammersley and Clifford (1971), states that the joint distribution of a Markov random field equals the product of its conditional distributions. This factorization in turn equals the product of "potential functions" (defined below) measured along its maximal cliques. The theorem frees us from building the joint distribution of the graph directly, instead, we revert to using the maximal cliques.

The Hammersley-Clifford theorem was never published by its discoverers. Numerous restatements and proofs have been given in the ensuing years, including Besag

(1972, 1974) on lattice neighboring structures, Ripley and Kelly (1977) on Markov point processes, and Robins et al. (1999) on non-lattice general graphs. The beauty of the theorem is, even when spatial correlation structure is complicated for the undirected graph, it is still relatively easy to determine the joint probability distribution function.

Two requirements precede the theorem. First, for the theorem to apply, the graph has to be a Markov random field, which a moralized generalized tree network is by definition. Second, the probability distribution on the graph has to be strictly positive. This requirement can be checked by the positivity and zero assumptions (Besag, 1974).

The *positivity assumption* says that if the values $x_1, x_2, \ldots, x_n$ can individually appear on sites $v_1, v_2, \ldots, v_n$, then they can appear together in all sites at the same time. Precisely, if $\mathbf{V} = \{v_1, v_2, \ldots, v_n\}$, $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ and $p(v_i = x_i) > 0, \forall i = 1, \ldots, n$, then we have $p(\mathbf{V} = \mathbf{X}) > 0$. Besag (1975) has noted that the positivity assumption is usually satisfied by most real-world data sets.

The *zero assumption* requires that the value 0 is an acceptable realization on any site, so $p(v_i = 0) > 0, \forall i = 1, \ldots, n$. Combine this with positivity, it is clear to see that $p(\mathbf{V} = \mathbf{0}) > 0$. In their proofs of the theorem, Grimmett (1973) and Preston (1973) have shown that this is a technical assumption; *i.e.,* any data set which violates this assumption can be easily brought to validity by re-indexing the values.

Both requirements are almost always met in moralized generalized tree networks. We now state the theorem here.

Theorem 3.3.1. *(Hammersley-Clifford theorem): Assume a strictly positive joint probability distribution, $p_{\mathcal{G}}(\mathbf{V})$, on Markov random field, $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, exists. Denote*

the set of unknown parameters on $\mathcal{G}$ as Θ_G. The following properties are equivalent.

1. Global Markov property: $p_{\mathcal{G}}(\mathbf{V})$ can be written as the product of conditional densities obtained by recognizing the conditional independencies from the graph separation of $\mathcal{G}$. That is, for any non-empty triple subsets $\mathbf{A}, \mathbf{B}, \mathbf{S} \in \mathbf{V}$, if $\mathbf{S}$ graph separates $\mathbf{A}$ and $\mathbf{B}$,

$$p_{\mathcal{G}}(\mathbf{A}, \mathbf{B}|\mathbf{S}) = p_{\mathcal{G}}(\mathbf{A}|\mathbf{S}) p_{\mathcal{G}}(\mathbf{B}|\mathbf{S}) \tag{3.2}$$

2. Factorization property: $p_{\mathcal{G}}(\mathbf{V})$ can be factorized by

$$p_{\mathcal{G}}(\mathbf{V}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \phi_C(\mathbf{V}_C|\Theta_C),\ \Theta_C \in \Theta_G. \tag{3.3}$$

where $\mathcal{C}$ is the set of all the maximal cliques on $\mathcal{G}$, $\mathbf{V}_C$ is the subset of nodes on clique C, and Θ_C is the parameters associated with $\mathbf{V}_C$. $\phi_C(\mathbf{V}_C|\Theta_C)$ is called a potential function or simply a potential. It is a non-negative function that is proportional to some density function $f_C(\mathbf{V}_C)$. $Z = \sum_v \prod_{C \in \mathcal{C}} \phi_C(\mathbf{V}_C|\Theta_C)$ is a normalizing constant called the partition function. It ensures that the normalized product of the potentials is a density that summarizes to 1. The summation $\sum_v$ is taken over all the possible realizations of the variables $\mathbf{V} = v$. Therefore $Z = Z(\Theta_{\mathcal{G}})$ is a function of the unknown parameters to be estimated from the graph rather than a function of the nodes.

The theorem states that the joint distribution of a Markov random field equals to its *Gibbs distribution*, the partition on the right hand side of Equation (3.3). There are several proofs to the theorem, including those from Grimmett (1973), Preston (1973), Sherman (1973) and Besag (1974). The proof for the global Markov property can be

found in Theorem 1 of Andersson et al. (2001), while the second part, the factorization property, was detailed in Cheung (2008).

The global Markov property captures the cohesiveness and connectivity of the graph, but considering the number of nodes it involves, it can be difficult to work with directly. Together with the moralization procedure, the Hammersley-Clifford theorem connects the global Markov property, and the joint distribution of a chain graph, to the local neighborhood structure and partition on its moralized undirected graph using maximal cliques. This connection is expressed in Equation (3.4) below, where $\mathbf{E}^M$ is the set of moralized edges of a generalized tree network, $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, and $\mathcal{G}^M = (\mathbf{V}, \mathbf{E}^M)$ is its moralized undirected graph:

$$p_{\mathcal{G}}(\mathbf{V}) = p_{\mathcal{G}}^{CI}(\mathbf{V}) \Rightarrow p_{\mathcal{G}^M}^{CI}(\mathbf{V}) = p_{\mathcal{G}^M}(\mathbf{V}) = \frac{1}{Z}\prod_{C\in\mathcal{C}} \phi_C(\mathbf{V}_C|\Theta_C). \tag{3.4}$$

Here $p_{\mathcal{G}}$ stands for the joint distribution of chain graph $\mathcal{G}$, and $p_{\mathcal{G}^M}$ is the joint distribution for undirected graph $\mathcal{G}^M$. $P_{(.)}^{CI}$ is the joint distribution written out in its conditional independence form. The first equal sign in Equation (3.4) is true because of the LWF Markov property on the chain graph; the implication is because of the partial equivalence the moralization maintains the conditional independencies in the chain graph; and finally, the third and fourth equal signs are true due to the first and second properties of Hammersley-Clifford theorem. One should note that the expression in (3.4) is not unique because the ϕ_C's can be arbitrarily selected.

Because on lattices it is usually easy to identify maximal cliques and their separators, the partition of the Gibbs distributions is relatively easy to obtain. Equation (3.3) is also mathematically straightforward and readily interpretable. It is based on

factorization, without attention to the details on graph variance-covariance matrix. The theorem does not impose strong restrictions on the neighborhood structure, the separation rule, nor the marginal distributions of the nodes other than the positivity assumption [which is satisfied by most of the practical problems, also see the discussion between Besag and Hammersley in Besag (1974)]. The theorem makes a great candidate when the interests are on those chain graphs with discrete components, as long as the clique distributions involved in the partition can be found. As demonstrated later, we employ a copula model to write distributions of the maximal cliques when the nodes in the cliques are discrete and correlated.

On a clique C, the only requirement for the potential function $\phi_C(\mathbf{V}_C|\Theta_C)$ to be valid is that it is a non-negative function that is proportional to some density function, $f_C(\mathbf{V}_C)$. In other words, for nodes $\mathbf{V}_C$ of clique C and two of its possible realizations, $\mathbf{v}_{c_1}$ and $\mathbf{v}_{c_2}$, the potential function only needs to satisfy

$$\frac{\phi_C(\mathbf{V}_C = \mathbf{v}_{c_1}|\Theta_C)}{\phi_C(\mathbf{V}_C = \mathbf{v}_{c_2}|\Theta_C)} = \frac{p_C(\mathbf{V}_C = \mathbf{v}_{c_1})}{p_C(\mathbf{V}_C = \mathbf{v}_{c_2})}. \tag{3.5}$$

As we will see later this is actually sufficient when developing the joint distribution estimate for a large and highly connected chain graph. (see Appendix A).

In sum, the theorem allows us to use an explicit and general factorization function to model most chain graphs, with flexibility on selecting the individual factors of the partition. These properties made it very attractive when dealing with large dimension problems, and with non-Gaussian graphs, just like the cases of our ICG models.

As much as all of this sounds too good to be true, there are indeed a few drawbacks on the theorem that may limit its application. One of the issues worth mentioning is

that, since the theorem and the partition rely almost entirely on the maximal cliques of the graph, their characteristics such as size and correlation structure can affect the modeling and computational effort. For instance, if the clique sizes are big, calculating $\phi_C(\mathbf{V}_C|\Theta_C)$ can be quite complicated and time consuming.

Another issue is determining the partition function $Z(\Theta_{\mathcal{G}})$ in Equation (3.3). Since it is a normalization over all the potentials, it is a function with respect to the unknown parameters $\Theta_{\mathcal{G}}$. However, it is often that the form of $Z(\Theta_{\mathcal{G}})$ is unknown. Since $Z(\Theta_{\mathcal{G}})$ goes hand-in-hand with the potential functions, an arbitrary definition of the potentials can lead to complicated and unknown form of $Z(\Theta_{\mathcal{G}})$. The only exceptions may be the Gaussian cases (see Appendix B). For general cases, there is usually no closed algebraic form for $Z(\Theta_{\mathcal{G}})$ (Lauritzen, 1996; Bishop, 2006). The function can only be computed using numeric methods for a majority of the graphs. By normalizing over all the realization of the graph, the calculation of $Z(\Theta_{\mathcal{G}})$ demands a huge computational effort. For a graph with discrete nodes, the summation of $Z(\Theta_{\mathcal{G}})$ involves n^p members when the graph has n nodes and each node has p different states. For graphs with continuous nodes, $Z(\Theta_{\mathcal{G}})$ requires a n-fold integration. None of these is easy to accomplish, and the direct calculation often meets its end without success.

Nevertheless, to obtain estimates of parameters associated with the graph mandates the knowledge of $Z(\Theta_{\mathcal{G}})$. This is problematic and poses a challenge for the applications of the Hammersley-Clifford theorem. In the next section, we focus on developing a method that is able to overcome this challenge. It involves identifying a certain type of graphical structure known as a junction trees, as well as creating conditional Markov random fields on the moralized undirected graphs, to make progress in the partition function.

3.4 Finding the joint distribution

This section deals directly with the problem of how to explicitly write out the partition function $Z = Z(\Theta_G)$ in a moralized Markov random field. If $Z(\Theta_G)$ were known, we may skip this section completely and go directly into the next one on modeling the clique distributions.

Two approaches can help us to deal with $Z(\Theta_G)$. The first approach is distribution-based (Cressie and Lele, 1992) and applies to a limited set of chain graphs. In this approach, we need to pre-determine the potentials according to some known form, such that their product, $\prod_{\mathcal{C}} \phi_C$, can be recognized as part of a known distribution. In these cases, we don't need to calculate $Z(\Theta_G)$ to know $\frac{1}{Z}\prod_{\mathcal{C}} \phi_C$ because $\prod_{\mathcal{C}} \phi_C$ is already known. These potentials are usually from exponential families, and the approach is permitted for some exponential family-based graphs only (Hamze and Freitas, 2006; Moura and Balram, 1992). By learning the family of distribution of the potential product before fitting the graphical model, the normalizing procedure of $Z(\Theta_G)$ can be effectively avoided. These graphs are known as *Exponential family Markov random fields* and *Gaussian Markov random fields* (Rue and Knorr-Held, 2005), and they provide effective alternatives to model the joint densities on some of these undirected graphs. This approach is detailed in Appendix B.

We propose a second, more general approach that is structural-based. The idea of *decomposable* graphs will be precisely defined later, but generally speaking it can be viewed as a graph whose joint distribution can be written as the product of some known functions of its maximal cliques, including the normalizing part. By studying the neighborhood structure of the moralized undirected graph, we can identify whether

the graph is decomposable. If the graph is indeed decomposable, the junction tree algorithm provides a suitable solution to write down $Z(\Theta_G)$ in the form of the product of clique separators; if the graph is non-decomposable, we will have the options to convert it into decomposable graph by either adding some edges or eliminating some nodes. We here introduce the approach and show how $Z(\Theta_G)$ can be obtained in a decomposable graph.

3.4.1 Junction tree algorithm

A certain class of undirected graphs is said to have the decomposable property if their maximal cliques are connected in the graph the same way as the nodes in a tree network are connected. Their potential functions and the normalizing partition can be written as functions within the cliques. These decomposable graphs are known as junction trees (Jensen et al., 1990). They have been discussed in detail by Jensen and Jensen (1994); Cowell et al. (1999); Barber (2003); Cevher (2008); and Wainwright and Jordan (2008).

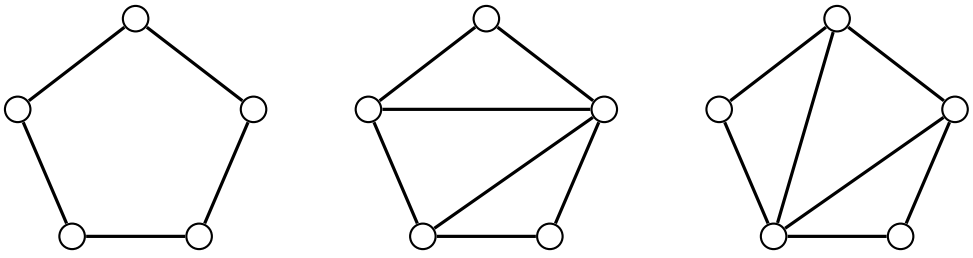
Any undirected graph that satisfies the defining decomposability is a junction tree. If a generalized tree network can be moralized into a junction tree, the partition function can be factored or sometime even reduced to 1. On the other hand, if the moralized undirected graph is not a junction tree, we may still be able to either delete some nodes or add some edges in the undirected graph to convert it to a junction tree without satisfying too much of the original graph.

Apart from the general lexicon of graphical model stated in the previous chapter, we need to give a few additional definitions that facilitate the definition of a junction

tree. As before, many authors have proposed different definitions, assigned different names for the junction tree, such as the *join tree* by Maier (1983) and *k-tree* by Tidén and Arnborg (1987). Our version follows Cowell et al. (1999).

Recall the definition of a cycle, where a sequence of nodes $\{v_1, v_2, \ldots, v_n\}$ is a n-cycle when every $v_i - v_{i+1} \in \epsilon$ and $v_1 = v_n$. The length of the cycle is n. A *chord* is an edge that connects two non-consecutive nodes in a cycle. They can be denoted as $v_i - v_j$, with $j \neq i+1$ or $i-1$. Though chords could be either directed or undirected, we are exclusively discussing issues in the context of moralized graphs. Therefore, when we mention a chord in this chapter, we are referring to an undirected chord. A chord may only appear in cycles with length $n \geqslant 4$. An undirected graph is *chordal* or *triangulated* when every cycle with length greater than 4 in the graph has at least one chord. The process of adding chords to a cycle to make it chordal is called the *triangulation* (Jordan et al., 1999). It is important to point out that triangulation is not unique for almost all undirected graphs. Figure 3.6 provides an example of a simple undirected graph with two of the many possible triangulations.

Figure 3.6: A 5-cycle (a) and two of its triangulations (b, c)



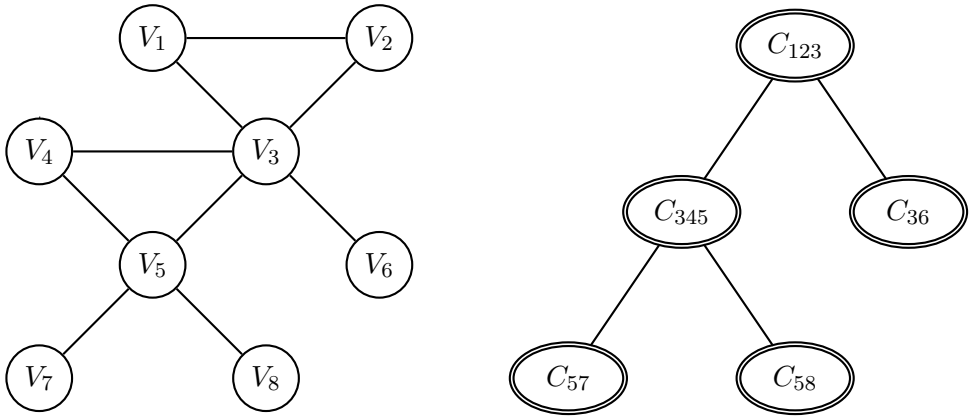
(a) (b) (c)

A *triplet* is three subsets, $\mathbf{A}, \mathbf{B}, \mathbf{S}$, of the vertex set $\mathbf{V}$ satisfying $\mathbf{A} \cup \mathbf{B} \cup \mathbf{S} = \mathbf{V}$ and $\mathbf{A} \cap \mathbf{B} = \mathbf{B} \cap \mathbf{S} = \mathbf{A} \cap \mathbf{S} = \varnothing$. We say $\mathbf{S}$ separates $\mathbf{A}$ and $\mathbf{B}$ if any path between

a node in $\mathbf{A}$ and another node in $\mathbf{B}$ will pass through $\mathbf{S}$. When $\mathbf{S}$ is also a maximal clique, the triplet $\mathbf{A}, \mathbf{B}, \mathbf{S}$ is said to have *decomposed* $\mathcal{G} = G(\mathbf{V}, \mathbf{E})$ (Bishop, 2006). A decomposition is *proper* if both $\mathbf{A}$ and $\mathbf{B}$ are non-empty. An undirected graph is *decomposable* when it is a maximal clique, or when it can be decomposed properly by a triplet into decomposable subgraphs $\mathcal{G}_{\mathbf{A}\cup\mathbf{S}}$ and $\mathcal{G}_{\mathbf{B}\cup\mathbf{S}}$. This is a recursive definition that is only permissible when each step of decomposition may reduce the size of the graph, hence the proper requirement. We can continuously apply the decomposition step to an undirected graph until all the subgraphs are maximal cliques. Because there is no triplet that separates a maximal clique, it can not be properly decomposed.

The decomposable undirected graphs share an uniformity that is important for our model development: the maximal cliques of a decomposable undirected graph can be ordered and connected in a tree network. In other words, the decomposable undirected graph is a clique tree. It preserves the dependence structure of the original graph, for instance Figure 3.7 (a), and effectively simplifies each maximal clique into a single node in Figure 3.7 (b). The maximal clique set of the left undirected graph is $\mathcal{C} = \{C_{123} = \{V_1, V_2, V_3\}, C_{345} = \{V_3, V_4, V_5\}, C_{36} = \{V_3, V_6\}, C_{57} = \{V_5, V_7\}, C_{58} = \{V_5, V_8\}\}$. These cliques form the corresponding clique tree on the right. Notice that there are no edges between maximal clique nodes C_{345} and C_{36}, C_{57} and C_{58} because otherwise the graph will contain cycles and no longer be a tree. The definitions above pave the way for the identification of junction trees. An undirected graph $\mathcal{G}$ with maximal clique set $\mathcal{C} = \{C_i, C_j, \ldots\}$ is a *junction tree* if, for any pair of its clique nodes C_i, C_j, all nodes on the path between C_i and C_j contain the intersection $C_i \cap C_j$. This is also known as the *running intersection property*. The intersection $C_i \cap C_j$ is called the *separator* of cliques C_i and C_j.

Figure 3.7: Example of a MRF (a) its clique tree (b)



(a) (b)

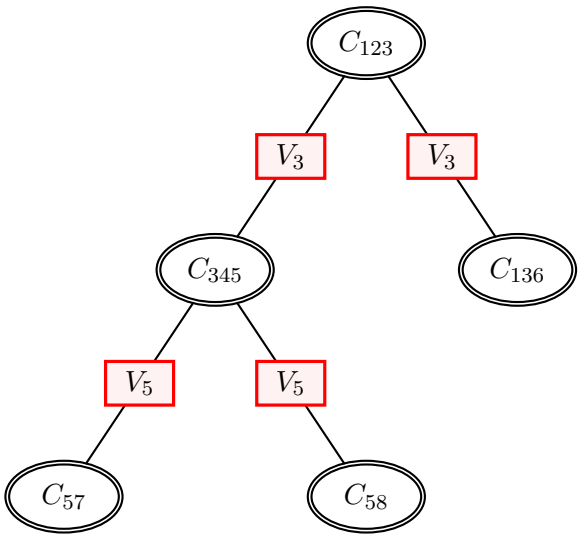
Pictorially, this property states that there is no "shortcut" between two clique neighbors. No active path exists through a third clique such as $C_i - C_k - C_j$ for the neighbors: any connection between the two neighbors has to pass through the "front door," the intersection $C_i \cap C_j$. A junction tree eliminates the possibility of clique cycles, because otherwise $C_i - C_k - C_j - C_i$ would have violated the running intersection property.

Figure 3.7 (a) yields the junction tree Figure 3.8. The clique tree structure remains the same in the junction tree, with the addition of the rectangles denoting separators. For instance, $\{V_1, V_2, V_3\} \cap \{V_3, V_4, V_5\} = V_3$, therefore V_3 is the separator of C_{123} and C_{345}, etc. Both V_3 and V_5 serve as separators twice because each one is connected with three maximal cliques.

The ideas of chordal, decomposition and junction tree have very profound underlying connections. Cowell et al. (1999) has proved the following key theorem.

Theorem 3.4.1. *The following properties are equivalent for an undirected graph $\mathcal{G}$:*

Figure 3.8: Junction tree based on Figure 3.7 (a).



1. $\mathcal{G}$ is chordal.

2. $\mathcal{G}$ is decomposable.

3. The clique tree of $\mathcal{G}$ is a junction tree.

These properties in Theorem 3.4.1 are essentially stating the same underlying characteristic of junction trees in three different ways: maximal cliques cannot be connected in a cycle. From Figure 3.7 we see that a junction tree is a tree network with respect to the cliques. By forbidding clique cycles, a junction tree permits only one "path" of connection among any pair of maximal cliques.

Junction trees allow a roadmap where not only joint distributions are partitioned, but the normalizing constant, $Z(\Theta_{\mathcal{G}})$, is explicit as well. Its ability to partition the graph can be described by the message passing algorithm (Wainwright and Jordan, 2008) that is inversely analogous to the process of nutrition transported in a real tree

from its root to the leaves. If we consider the marginal distribution of C as a *message* stored in the clique, then when two maximal cliques C_1 and C_2 are non-independent, we may say that the finding of joint density $p_{\mathcal{G}}(C_1, C_2)$ via the conditional marginal density $p_{\mathcal{G}}(C_1|C_2)$ using equation $p_{\mathcal{G}}(C_1, C_2) = p_{\mathcal{G}}(C_1|C_2)p_{\mathcal{G}}(C_2)$ is a *passing of the message* from C_2 into C_1. After passing, $p_{\mathcal{G}}(C_1, C_2)$ possesses the message (distribution) from both clique $C1$ and $C2$. A message can only be passed between adjacent clique nodes. Graphically in Figure 3.8, every message passing step involves transporting the information of a clique node through a rectangular separator to an adjacent clique node. Because there is only one "path" between any two arbitrary maximal cliques in a junction tree, there is only one way which the message can be passed from one maximal clique to another. Since there is a predetermined root in the tree network, we may order the maximal cliques in the junction tree by its distance from the root. By repeatedly passing the message from the maximal clique that is farther away from the root to the one that is closer to the root, all the message from all maximal cliques can be passed to the root. This total message is the joint distribution of the graph.

The message passing algorithm should work with any junction tree, unaffected by whether the nodes and cliques are marginally discrete or continuous. It applies to the moralized ICG, as long as the moralized graphs are junction trees. Other than the descriptive definition above, the algorithm can also be written in the form of a factorization of the graph, with every factor explicitly stated and readily modeled. This is an important step in our model development. We propose to formalize the algorithm into the following theorem.

Theorem 3.4.2. *The joint distribution of a finite junction tree $\mathcal{G} = G(\mathbf{V}, \epsilon)$ with maximal clique set $\mathcal{C}$ and separator set $\mathcal{S}$ can be factorized by the marginal densities*

of its maximal cliques and separators:

$$p_{\mathcal{G}}(\mathbf{V}) = \frac{\prod_{C \in \mathcal{C}} p_C(\mathbf{V}_C)}{\prod_{S \in \mathcal{S}} [p_S(\mathbf{V}_S)^{d_S - 1}]} \tag{3.6}$$

where $\mathbf{V}_C$ *and* $\mathbf{V}_S$ *are the node sets of* C *and* S*, and* d_S *is the number of maximal cliques joining together at separator* S*.*

Proof. Because there is no factorization when the junction tree $\mathcal{G}$ is a maximal clique itself, we start the proof by showing the theorem is true when $\mathcal{G}$ has two maximal cliques. Suppose $\mathcal{G} = C_1 \cup C_2$, $\mathcal{C} = \{C_1, C_2\}$, and the separator between the two cliques is $S = C_1 \cap C_2$.

$$p_{C_i}(\mathbf{V}_{C_i}) = p(\mathbf{V}_{C_i \backslash S} | \mathbf{V}_S) p(\mathbf{V}_S), i \in 1, 2. \tag{3.7}$$

Because $\mathcal{G}$ is a junction tree, every path from a node in $C_1 \backslash S$ to another node in $C_2 \backslash S$ passes S, therefore S graph separates $C_1 \backslash S$ and $C_2 \backslash S$. By the global Markov property we have $\{C_1 \backslash S\} \perp\!\!\!\perp \{C_2 \backslash S\} \mid S$ and

$$\begin{aligned}
p_{C_1}(\mathbf{V}_{C_1}) p_{C_2}(\mathbf{V}_{C_2}) &= p(\mathbf{V}_{C_1 \backslash S} | \mathbf{V}_S) p(\mathbf{V}_S) p(\mathbf{V}_{C_2 \backslash S} | \mathbf{V}_S) p(\mathbf{V}_S) \\
&= p(\mathbf{V}_{C_1 \backslash S}, \mathbf{V}_{C_2 \backslash S} | \mathbf{V}_S) p(\mathbf{V}_S)^2 \\
&= p(\mathbf{V}_{C_1 \backslash S}, \mathbf{V}_{C_2 \backslash S}, \mathbf{V}_S) p(\mathbf{V}_S) \\
&= p_{\mathcal{G}}(\mathbf{V}) p(\mathbf{V}_S), \qquad (3.8) \\
\text{hence we have } \; p_{\mathcal{G}}(\mathbf{V}) &= \frac{p_{C_1}(\mathbf{V}_{C_1}) p_{C_2}(\mathbf{V}_{C_2})}{p(\mathbf{V}_S)}. \qquad (3.9)
\end{aligned}$$

Here $d_S = 2$ because S is separating two maximal cliques, so Equation (3.6) is true.

Now assume the theorem is true for junction trees with n maximal cliques, and

consider the trees with $n+1$ cliques. Because a graph $\mathcal{G}_{n+1} = G(\mathbf{V}, \epsilon)$ with $n+1$ maximal cliques is also a clique tree, there is at least one maximal clique node, denoted by C_{n+1}, is adjacent to one and only one neighboring clique C_n. For the sub-graph $\mathcal{G}_n \subset \mathcal{G}_{n+1}$ that is induced by node set $\{C_1 \cup C_2 \cup \ldots \cup C_n\}$, let $\mathcal{S}^*$ be its separator set, and d_S^* be the number of maximal cliques joining at a given separator S on the sub-graph $\mathcal{G}_n$. Equation (3.6) holds for $\mathcal{G}_n$ and its joint density can be factored as

$$p(\mathbf{V}_{C_1 \cup C_2 \cup \ldots \cup C_n}) = \frac{\prod_{i=1}^{n} p_{C_i}(\mathbf{V}_{C_i})}{\prod_{S \in \mathcal{S}^*} [p_S(\mathbf{V}_S)^{d_S^* - 1}]} \tag{3.10}$$

Now add C_{n+1} to the graph. Since it only has one neighboring clique, C_{n+1} can be arbitrarily named as the root of the clique tree. In this way, all the other maximal cliques are farther away from the root than C_n, therefore $p_{\mathcal{G}}(\mathbf{V}_{C_1 \cup C_2 \cup \ldots \cup C_n})$ contains the messages that were passed from all $\{C_1, C_2, \ldots, C_{n-1}\}$ to C_n. The last step of message passing will occur from C_n to C_{n+1}, which forms the joint density of $\mathcal{G}_{n+1}$. Denote $S_{n+1} = C_n \cap C_{n+1}$ to be the separator between C_n and the root clique. We call $\mathbf{A} = \mathbf{V}_{C_1 \cup C_2 \cup \ldots \cup C_n \setminus S_{n+1}}$, $\mathbf{B} = \mathbf{V}_{C_{n+1} \setminus S_{n+1}}$. By global Markov property the triplet $\{\mathbf{A}, \mathbf{B}, \mathbf{V}_{S_{n+1}}\}$ decomposes $\mathcal{G}_{n+1}$ and $\mathbf{A} \perp\!\!\!\perp \mathbf{B} | \mathbf{V}_{S_{n+1}}$, therefore

$$\begin{aligned} p(\mathbf{V}_{C_1 \cup C_2 \cup \ldots \cup C_n}) p(\mathbf{V}_{C_{n+1}}) &= p(\mathbf{A}|\mathbf{V}_{S_{n+1}}) p(\mathbf{V}_{S_{n+1}}) p(\mathbf{B}|\mathbf{V}_{S_{n+1}}) p(\mathbf{V}_{S_{n+1}}) \\ &= p(\mathbf{A}, \mathbf{B}, \mathbf{V}_{S_{n+1}}) p(\mathbf{V}_{S_{n+1}}) \\ &= p_{\mathcal{G}_{n+1}}(\mathbf{V}) p(\mathbf{V}_{S_{n+1}}). \end{aligned}$$

And we have

$$\begin{aligned} p_{\mathcal{G}_{n+1}}(\mathbf{V}) &= \frac{p(\mathbf{V}_{C_1 \cup C_2 \cup \ldots \cup C_n}) p(\mathbf{V}_{C_{n+1}})}{p(\mathbf{V}_{S_{n+1}})} \\ &= \frac{\prod_{i=1}^{n} p_{C_i}(\mathbf{V}_{C_i})}{\prod_{S \in \mathcal{S}^*} [p_S(\mathbf{V}_S)^{d_S^* - 1}]} \times \frac{p_{C_{n+1}}(\mathbf{V}_{C_{n+1}})}{p(\mathbf{V}_{S_{n+1}})} \\ &= \frac{\prod_{C \in \mathcal{C}} p_C(\mathbf{V}_C)}{\prod_{S \in \mathcal{S}} [p_S(\mathbf{V}_S)^{d_S - 1}]}. \end{aligned} \tag{3.11}$$

When S_{n+1} has already been counted as a separator between C_n and some other maximal clique $C_k \neq C_{n+1}$, $\mathcal{S} = \mathcal{S}^*$ and $d_{S_{n+1}} = d^*_{S_{n+1}} + 1$; when it is a unique separator, $\mathcal{S} = \mathcal{S}^* \cap S_{n+1}$ and $d_S = d^*_S$ for all the $S \neq S_{n+1}$ and $d_{S_{n+1}} = 2$. Either way, the last equal sign for Equation (3.11) always holds and Theorem 3.4.1 is true for any clique tree size n. □

Before leaving the proof of the theorem we would also like to point out the fact that Equation (3.11) is always *invariant under absorption*, or *globally consistent* (Lauritzen, 1996) in a junction tree. This means that the distribution of the separator is the same when marginalized from either side. When S separates cliques A and B, we have $\sum_{A \backslash S} p(A) = \sum_{B \backslash S} p(B)$ for discrete graphs and $\int_{A \backslash S} p(A) = \int_{B \backslash S} p(B)$ for continuous graphs. Barber (2003) has shown that an information update in $\{A \backslash S\}$ will trigger the update of $\{B \backslash S\}$ as part of the message passing process, which keeps Equation (3.11) invariant.

Comparing Equation (3.11) with the Hammersley-Clifford partition, one can see that in Equation (3.11) the denominator coincides with the partition function $Z(\Theta_{\mathcal{G}})$. With $Z(\Theta_{\mathcal{G}}) = \prod_{S \in \mathcal{S}} [p_S(\mathbf{V}_S)^{d_S - 1}]$, we can calculate this partition function using the marginal distributions of the separators on the graph. If we know the distributions of

the maximal cliques, and consider the fact that separators are subsets of the cliques, we can obtain each of the $p_S(\mathbf{V}_S)$ as well. Finally, in junction trees we can solve for $Z(\Theta_{\mathcal{G}})$ directly. It completes the final step writing down the joint distribution partition.

The key to our approach is now apparent. When the moralized generalized tree network is a junction tree, we can use Equation (3.11) for the joint distribution. Everything depends on this prerequisite. When the moralized graph is not a junction tree, the move towards the joint distribution comes to a halt. It turns out, however, that the moralized generalized tree networks can be converted into junction trees using some conversion algorithms. In doing so we create conditional Markov fields, which when conditioned on a certain set of nodes, are junction trees themselves.

3.4.2 Building junction trees

Since a junction tree facilitates a simple distribution partition with explicitly written normalizing constants, it is natural that we want to apply the junction tree framework to as many moralized generalized tree networks as possible. When the moralized UG is not a junction tree, Algorithms 3.4.3 and 3.4.4 provide guidance on how a moralized generalized tree network may be converted into a junction tree.

Algorithm 3.4.3. *Building a junction tree.*

- **Step 1.** *Read in a generalized tree network $\mathcal{G}$;*
- **Step 2.** *Is $\mathcal{G}$ an undirected graph?*
  - ◇ **Yes** → *Make $\mathcal{G}^M = \mathcal{G}$, go to* **Step 4**.
  - ◇ **No** → *Go to* **Step 3**.

- **Step 3. Moralize** $\mathcal{G}$*, return the moralized Markov random field* $\mathcal{G}^M$*.*

- **Step 4.** *Is* $\mathcal{G}^M$ *chordal?*

  - ◇ **Yes** → *Make* $\mathcal{G}^{M\Delta} = \mathcal{G}^M$*, go to* **Step 6***.*

  - ◇ **No** → *Go to* **Step 5***.*

- **Step 5. Triangulate** $\mathcal{G}^M$ *into* $\mathcal{G}^{M\Delta}$ *by adding chords to non-chordal cycles in* $\mathcal{G}^M$*.*

- **Step 6.** *Identify maximal clique nodes* $\mathcal{C}$ *and separators* $\mathcal{S}$ *of* $\mathcal{G}^{M\Delta}$*. Form the* **junction tree***.*

- **Step 7.** *Assign an arbitrary* **root** *clique.*

- **Step 8. Pass messages** *from the farthest leaves to the nearest. Update the joint distribution.* **Repeat** *until reaching root.*

- **Step 9.** *Return the* **joint distribution** *factorization of* $\mathcal{G}^{M\Delta}$*.*

- *EXIT.*

Algorithm 3.4.3 guarantees that the resulting graph is a junction tree. Just like in the moralization process, adding new edges during triangulation does not contradict the original GTN, since it does not impose new conditional independence properties to the graph. However, there is no restriction on how big its maximal cliques would be. A junction tree with cliques too large in size is unsuitable for partition, because the marginal distributions of the large cliques might still be intractable. In step 5, when the graph is triangulated, we have no control on how many edges we need to add into

the graph, because we are bound to triangulate the nodes into maximal cliques. It is purely determined by the connectivity and structure of the original graph. Since there is no cap for the numbers of edges we might need to add into the graph, there is no way to control what the largest maximal clique size is in the triangulated junction tree. In the worst case scenario, the graph can be triangulated into one giant single clique, which makes the Hammersley-Clifford partition a moot point.

To solve this problem, we must constrain adding an excessive number of edges during triangulation. To satisfy the chordal requirement and maintain the final graph as a junction tree, some edges and nodes may need to be removed from the graph. By removing a few strategically selected nodes and the edges they are connected to the graph, the sub-graph induced by the remaining nodes can be triangulated with a controlled size of maximal clique. This process will create the graphs known as *conditional Markov random field* (CMRF), introduced in Besag (1974). The removed nodes are the *conditioned-on nodes*, while the remaining ones the *conditioning nodes*.

Although being an incomplete graph, Besag (1974, 1975) and our simulated example suggested that, if selected wisely, the conditional Markov random field would still retain much of the structural and dependent information of the original graph. The choice of the conditioned-on nodes is also in most instances non-unique, which gives us the opportunity to obtain multiple estimates on the same set of parameters for one graph. Each estimate corresponds to one choice of the conditioning, and they can be combined together to form a final estimate. The following algorithm is a revision of Algorithm 3.4.3 and allows for the node conditioning to create a conditional Markov random field.

Algorithm 3.4.4. *Building a junction tree (CMRF).*

- **Step 1.** *Read in a generalized tree network* $\mathcal{G}$*;*

- **Step 2.** *Is* $\mathcal{G}$ *an undirected graph?*

  - ◇ **Yes** → *Make* $\mathcal{G}^M = \mathcal{G}$*, go to* **Step 4**.

  - ◇ **No** → *Go to* **Step 3**.

- **Step 3. Moralize** $\mathcal{G}$*, return the moralized Markov random field* $\mathcal{G}^M$*.*

- **Step 4.** *Is* $\mathcal{G}^M$ *chordal?*

  - ◇ **Yes** → *Make* $\mathcal{G}^{M\Delta} = \mathcal{G}^M$*, go to* **Step 6**.

  - ◇ **No** → *Go to* **Step** $4\frac{1}{2}$.

- **Step** $4\frac{1}{2}$**.** *Set aside a node* D_i *on one of the non-chordal cycles into* $\mathcal{D}$ *(the conditional node set). Delete all edges linking* D_i *with the graph. Go to* **Step 5**.

- **Step 5. Triangulate** $\mathcal{G}^M$ *into* $\mathcal{G}^{M\Delta}$ *by adding chords to non-chordal cycles in* $\mathcal{G}^M$*. Is the maximal clique sizes in* $\mathcal{G}^{M\Delta}$ *too big?*

  - ◇ **Yes** → *Go back to* **Step** $4\frac{1}{2}$**.**

  - ◇ **No** → *Go to* **Step 6**.

- **Step 6.** *Identify maximal clique nodes* $\mathcal{C}$ *and separators* $\mathcal{S}$ *of* $\mathcal{G}^{M\Delta}$*. Form the* **junction tree**.

- **Step 7.** *Assign an arbitrary* **root** *clique.*

- **Step 8. Pass messages** *from the farthest leaves to the nearest. Update the* **conditional** *joint distribution of* $\mathcal{G}^{M\Delta}$ *given* $\mathcal{D}$*.* **Repeat** *until reaching root.*

- **Step 9.** *Return the* **conditional joint distribution** *factorization of* $\mathcal{G}^{M\Delta}$.

- *EXIT.*

The main difference between Algorithm 3.4.3 and Algorithm 3.4.4 is the addition of Step 4½, which allows the node conditioning. As demonstrated in Chapter 4, the choice of conditional node set, D_i, is influential for the estimates, especially on irregular lattices. It may depend on many factors including the size, shape and background of the data set. After the junction tree is produced based on Algorithm 3.4.4, Equation (3.6) can be applied to obtain the conditional likelihood and estimate for each conditional Markov random field.

By means of the junction tree or conditional Markov random field, the modeling and estimation of the undirected graphical model $\mathcal{G}^M$ have all come down to its maximal cliques $C \in \mathcal{C}$. We need to know the value of the marginal distributions $p_C(\mathbf{V}_C)$ and $p_S(\mathbf{V}_S)$ in Equation (3.6) to calculate the graph density, $p_{\mathcal{G}}(\mathbf{V})$. Because the cliques and its separators are on a size much smaller than the full graph, we should be able to calculate their marginal distributions directly. Since in a clique, or separator, every node is dependent upon all the other nodes, copula models provide a great tool to untangle these dependencies.

3.5 Copula model

It can be difficult to estimate the joint distribution with a complex dependency pattern. Various models have been proposed to measure dependency structure among variables; however, they usually require each variable to follow the same family of distribution.

Examples of these models include the Gaussian MRF model for normal data, and the Ising model for binomial data. They are powerful analytical tools when the data are distributed approximately as a known distribution, but few can be generalized to include a mix of different families of distributions for the different variables involved in a data set.

The Copula has provided an ideal candidate for this situation. It separates the dependency among variables from their marginal distributions, and allows flexible definitions of both terms. It has the minimum restriction on variables' marginal distributions, and is a perfect candidate for building a complex correlation structure. In this section we will show that it is possible to express the joint density of a maximal clique using a multivariate copula, even when the variables are distributed differently and some discretely. We first give the definition of a copula, and then provide a multivariate copula designed to model multivariate spatially correlated data.

Fundamental work on the copula model can be found in Hoeffding (1940) and Nelsen (1999). Fisher (1997) introduces the copula in his paper as a function that "joins or couples multivariate distribution functions to their one-dimesional marginal distribution functions." Nelsen (1999) gives a more precise definition of the *bivariate copula* or *2-copula* $C(u, v)$. It is a function with domain $\mathbf{I}^2 = [0, 1]^2$ and satisfies $C(u, 1) = u$, $C(1, v) = v$.

The bivariate copula, also known as just the *copula*, is the simplest copula that involves only two random variables. Its concept can be generalized to d-dimensional cases. A copula with more than two variables is known as a *multivariate copula* or *d-copula*. The generalized case of d-copula is defined as follows. Let $\mathbf{U} = (u_1, u_2, \ldots, u_d)$

be a realization of vector of random variables with each u_i defined on $[0, 1]$ and joint CDF $F(\mathbf{U})$ on $\mathbb{R}^d$. Let each of the marginal univariate distribution functions, $F_i(u_i)$, and each of the marginal probability density function, $f_i(u_i)$, be well defined for $i = 1, 2, \ldots d$, meaning that the range of $F_i(u_i)$ is $[0, 1]$ and $\int f_i(u_i)du_i = 1$ for all i. The multivariate copula $C_{\mathbf{U}}$ is then defined as the function with domain $\mathbf{I}^d$ and range $\mathbf{I}$ that satisfies

$$C_{\mathbf{U}}(F_1(u_1), F_2(u_2), \ldots, F_d(u_d)) = F(\mathbf{U}), \mathbf{U} \in \mathbb{R}^d \tag{3.12}$$

The definitions suggest that copulas separate the modeling of marginal distributions of its random variables from the dependency among variables. In Equation (3.12), the marginal distributions depend only on $F_i(u_i)$, whereas the variance-covariance terms are involved only in $C_{\mathbf{U}}$. These functions (F_i and $C_{\mathbf{U}}$) can be selected individually and independently from each other. For example, we may define a multivariate copula $C_{\mathbf{U}}$ with Gaussian marginal CDF F_i's. The copula can even adapt to the cases where random variables are from different families of distribution, for instance, one continuous and the other discrete. For another example, if we had determined that in the copula specified above, one of the variables u_i is from the binomial distribution rather than Gaussian, then we may update the involved marginals F_i^*'s and the newly update copula will still be a valid joint distribution function. It is also ensured that such copula always exists, no matter what marginal or correlation structure there might be. Moreover, it is also unique under continuous cases. This result is knowns as Sklar's Theorem (Nelsen, 1999). We first state here the theorem in its bivariate form, and then move to its multivariate form as well.

Theorem 3.5.1. *(Sklar's Theorem in bivariate form): Let H be a joint distribution*

function with marginal CDFs F and G. Then there exists a copula, C, such that for all x, y in $\mathbb{R}$,

$$H(x, y) = C[F(x), G(y)]. \tag{3.13}$$

If F and G are continuous, then C is unique; otherwise, C is uniquely determined on the range $Ran(F) \times Ran(G)$. Conversely, if C is a copula and F and G are distribution functions, then the function H defined by (3.13) is a joint distribution function with marginals F and G.

The proof can be found in Nelsen (1999). For bivariate cases, Theorem 3.5.1 links the joint density function with copula functions, and guarantees the copula's existence, and its uniqueness under continuity. It offers a versatile way to define the bivariate joint density given any two marginal density functions without parametric restrictions.

There is, however, a set of bounds that is imposed upon the copulas. The range of the copula does not always project onto $\mathbf{I} = [0, 1]$, but in most of times onto a subset of it. This subset is the *Fréchet-Hoeffding bounds.*

Corollary 3.5.2. *(Fréchet-Hoeffding bounds): For any bivariate copula $C : \mathbf{I}^2 \to \mathbf{I}$ and any magical CDFs F and G, the following bounds hold:*

$$W(F, G) \leqslant C(F, G) \leqslant M(F, G), \tag{3.14}$$

$$\textit{where} \quad W(F, G) = max\,(F + G - 1, 0), \tag{3.15}$$

$$\textit{and} \quad M(F, G) = min\,(F, G). \tag{3.16}$$

$W(F, G)$ is called the *Fréchet-Hoeffding lower bound*, and $M(F, G)$ the *Fréchet-*

Hoeffding upper bound. They are both copulas themselves. As one would expect, these bounds have very close connections with the monotonicity of the distributions of random variables x and y. They measure the minimum and maximum Pearson correlation or other types of monotone dependence measurements the two variables are allowed to have under the copula jargon (Madsen and Birkes, 2011; Genest and Nešlehová, 2007). This property of the copula is further revealed in Theorem 3.5.3.

Theorem 3.5.3. *A subset $S \subset \mathbb{R}^2$ is nondecreasing if for any (z,t) and (u,v) on S, $z < u$ implies $t \leqslant v$. Similarly, S is nonincreasing if $z < u$ implies $t \geqslant v$. Let X and Y be two random variables with joint distribution function H. Then H equals the Fréchet-Hoeffding lower bound if and only if the support of $H(x,y)$ is non-increasing, and it equals the Fréchet-Hoeffding upper bound if and only if its support is nondecreasing.*

The proof can be found in Nelsen (1999). Theorem 3.5.3 gives the minimum and maximum dependence allowed by bivariate random variables, where Y is almost surely an increasing or decreasing function of X. It is also noted in Genest and Nešlehová (2007) that independence is induced by another special copula $\Pi(x,y) = xy$. When X, Y are continuous, they are independent if and only if $C(x,y) = \Pi(x,y)$. When they are discrete, $C(x,y) = \Pi(x,y)$ still implies independence, but this condition is not necessary anymore, i.e., there could be some other $C' \neq \Pi$ which induces independence of X, Y as well.

To adapt the clique distribution of a graphical model to a copula, we need to first make sure that the spatial autocorrelation satisfies these dependence requirements. Various types of formula exist for various types of marginal distributions. Prentice (1988) suggested the equation for maximal Pearson correlation for binary random variables, while Madsen (2009) had given the upper limit for general discrete cases.

For more realistic multivariate cases, copulas can also provide some insight into the complicated correlation structure through the generalized form of Sklar's Theorem and Fréchet-Hoeffding bounds.

Theorem 3.5.4. *(Sklar's Theorem in multivariate form): Let H be a joint distribution function for $\mathbf{U} = \{u_1, u_2, \ldots, u_d\} \in \mathbb{R}^d$ with marginal CDFs $F_1, F_2, \ldots, F_d$. There exists a d-copula $C_{\mathbf{U}}$ such that for all $u_1, u_2, \ldots, u_d$,*

$$H(u_1, u_2, \ldots, u_d) = C_{\mathbf{U}}[F_1(u_1), F_2(u_2), \ldots, F_d(u_d)]. \tag{3.17}$$

Moreover, if F_i's are all continuous, then $C_{\mathbf{U}}$ is unique; otherwise, $C_{\mathbf{U}}$ is uniquely determined on the range $\prod_i Ran(F_i)$. Conversely, if $C_{\mathbf{U}}$ is a copula and F_i's are distribution functions, then the function H defined by (3.17) is a joint distribution function with marginal F_i's.

When $F_i(u_i)$ is continuous for all i in Equation (3.17), according to Sklar's Theorem, $C_{\mathbf{U}}$ is the unique d-copula for $\mathbf{U}$. With $m_i = F_i(u_i)$ and $\mathbf{m} = (m_1, m_2, \ldots, m_d)$, the d-copula from Equation (3.17) can be written as

$$C_{\mathbf{U}}(m_1, m_2, \ldots, m_d) = F(F_1^{-1}(m_1), F_2^{-1}(m_2), \ldots, F_d^{-1}(m_d)). \tag{3.18}$$

The multivariate Fréchet-Hoeffding bounds are:

$$W_{\mathbf{U}}^d \leqslant C_{\mathbf{U}} \leqslant M_{\mathbf{U}}^d, \tag{3.19}$$

$$W_{\mathbf{U}}^d = \max\left(\sum_i u_i - n + 1, 0\right), \tag{3.20}$$

$$M_{\mathbf{U}}^d = \min\,(u_1, u_2, \ldots, u_d). \tag{3.21}$$

In this case, $M_{\mathbf{U}}^d$ is always a d-copula, but $W_{\mathbf{U}}^d$ is never a d-copula for any $d > 2$ (Nelsen, 1999).

Equation (3.18) is the copula form for which we will be utilizing to express the clique marginal distributions. When the d-th derivatives exist for H, as most continuous cases satisfy, one can differentiate $C_{\mathbf{U}}$ to obtain the joint PDF:

$$\begin{aligned} h(\mathbf{U}) &= \frac{\partial^d H(\mathbf{U})}{\partial u_1 \partial u_2 \dots \partial u_d} = \frac{\partial^d C_{\mathbf{U}}(m_1, m_2, \dots, m_d)}{\partial u_1 \partial u_2 \dots \partial u_d} \\ &= \frac{\partial^d C_{\mathbf{U}}}{\partial m_1 \partial m_2 \dots \partial m_d} \prod_{i=1}^{d} \frac{\partial m_i}{\partial u_i} \\ &= c(\mathbf{m}) \prod_{i=1}^{d} f_i(u_i). \end{aligned} \tag{3.22}$$

The derivative, $c(\mathbf{m}) = \frac{\partial^d C_{\mathbf{U}}}{\partial m_1 \partial m_2 \dots \partial m_d}$, is called the *copula density*.

For discrete random variables, $\mathbf{U} = \{u_1, \dots u_d\}$, Song (2000) has shown that the joint probability mass function, $p(\mathbf{U})$, can be written in the format of the summation of Gaussian copulas:

$$c(\mathbf{U}) \equiv p(\mathbf{U}) = \sum_{j_1=1}^{2} \dots \sum_{j_d=1}^{2} (-1)^{j_1 + \dots + j_d} \Phi_{\Sigma} \left[\Phi^{-1}(t_{1j_1}), \dots, \Phi^{-1}(t_{dj_d}) \right], \tag{3.23}$$

where $t_{i1} = F_i(u_i)$, and $t_{i2} = F_i(u_i-)$ denotes the limit from the left of F_i at $u_i, i = 1, \dots, d$. $\Phi_{\Sigma}(\cdot)$ is the d-order multivariate normal CDF with mean vector zero and variance-covariance matrix Σ, and $\Phi^{-1}(\cdot)$ is the inverse standard normal CDF. Finding the probability mass function (3.23) written in the copula summation format is computationally demanding, since it involves an addition of 2^d Gaussian d-copulas. One must use special care when d increases. Generally speaking, this calculation is

not efficient for $d > 4$. For discrete copulas with orders greater than 4, some types of approximation techniques, such as the jittering step proposed by Denuit and Lambert (2005), might be needed. It is also worthy exploring alternative copulas for higher dimensional discrete maximal cliques. Plackett copula (Nelsen, 1999) is one of the potential candidates. Mayor and Torrens (2005); Ma and Sun (2008) and Erdely et al. (2008) have also proposed methods connecting various types of copulas with discrete distribution functions. On the other hand, for calculations of Gaussian copulas with size $d = 4$, such as the cases of the distributions of 4-cliques in a second order nearest neighbor lattice, we have used function `pmvnorm` in `R` package and noticed substantial fluctuation of the result due to the numerical method the function incorporated. Such calculations require close examination and careful monitoring.

To denote the clique distribution, $p_C(\mathbf{V}_C)$, in terms of copulas, we need to specify the marginals, dependence parameters, and derive the CDF, $C_{\mathbf{U}}$, to obtain $c(\mathbf{m})$. As a consequence, there is no universal form for the copula distributions. All these steps need to be executed on a model-by-model basis, based on factors such as whether the variables are continuous or discrete and the size of the clique.

3.6 Discussion

In this chapter we described tools for modeling spatially dependent multivariate data denoted by ICG. We first introduced the moralization as a method to convert chain graphs into undirected graphs. We call those graphs that do not have chordless directed cycles generalized tree networks, and their moralized Markov random fields may be partitioned along their maximal cliques. The Hammersley-Clifford theorem provides a

straightforward form of the partition, with the joint distribution of the whole graph written as the product of potential functions along the maximal cliques, multiplied by a normalizing partition function. The theorem gives us great flexibility to assign arbitrary potential function for each maximal clique, but usually comes with the limitation of non-closed form partition function. Since the partition function usually involves unknown estimating parameters, we could not ignore it but have to find alternative ways to solve it. One way to do it is to assign exponential family potential function to each clique, in the hope that the their product form, and the form of the partition function, are identifiable. This method is most effective for Gaussian random fields. For non-Gaussian Markov random fields, we may convert them to junction trees by adding chords to the graph, following either one of the two junction tree algorithms. We may also produce the junction tree by selecting a subset of nodes to be conditioned on in the calculation of the joint distribution, thus creating the conditional Markov random fields. In these instances, the partition function can be written as the product of the separator distributions.

We believe that the moralizing procedure is an acceptable tool for graph conversion, because the conditional independences summarized by the undirected graph Markov properties in the moralized Markov random fields are also preserved in the chain graphs under LWF Markov properties. It is yet unclear, however, on what is the link between the AMP Markov property and the undirected graph Markov property.

When the moralized generalized tree network is not a junction tree, or when its partition function is unknown, the conversion process is influenced by which set of edges we select to add to the graph to create chordal cycles, or which set of nodes is conditioned on. This process also depends on the nature of the lattice: *i.e.*, which

kind of spatial scheme it has, whether it is regular or irregular. Different choices of the edges and nodes will almost certainly result in different junction trees, and introduce different estimates for the unknown parameters. A natural question we would like to answer is: how do the choices of conditioning nodes or chord edges affect the estimates? In the next chapter, we will explore this aspect of modeling.

We introduce copulas as a tool for modeling joint distributions of correlated and discrete random variables. Copulas deal with marginal distributions of each node and their correlation structure separately, and Sklar's Theorem allows us to construct correlated joint distributions using the marginal distributions even when each node is from a different family of distribution. Discrete multivariate copula functions are presented, calculated according to an inclusion-exclusion algorithm for each individual Gaussian copula. This algorithm is effective up to the fourth or fifth order, but beyond this point the computational time and precision will decrease. In this study we have not studied graphs with spatial schemes more connected than the regular lattice design, therefore the largest clique size is limited to four. However, for larger cliques and higher order copulas, we may need to explore alternative algorithm besides Equation (3.23) to model the clique distribution functions.

In the next chapter, we present examples on how to deal with different types of spatial data, with a focus on the lattice schemes. In each example we first denote the data using isomorphic chain graphs, and then proceed to demonstrate how these ICG can be converted to generalized tree networks, and subsequently junction trees or conditional Markov random fields. We will present both simulated data with Gaussian and bivariate variables, as well as real world data set obtain from the 2008 U.S. presidential election.

Chapter 4

Data Analysis and MCMC Simulation

4.1 Introduction

In this chapter we evaluate the methods from Chapter 3 for partitioning generalized tree networks (GTN) with continuous and discrete nodes using both simulated and real world data sets. The first data set is a simulated second order nearest neighbor regular lattice, from which inferences are made for parameters of a Gaussian and a Bernoulli random variable, both spatially correlated. We generate a regular lattice space with 15×15 sites and the second order nearest neighbor structure using simulation methods proposed by Rue (2000), with each non-border site having eight neighbors. Treating the graph as conditional Markov random fields, the unknown parameters for the correlation structures and the regression-like effects are estimated.

We then move on to investigate the GTN on irregular lattices. The data we use are county level election results during the 2008 United States presidential election in the states of Oregon and Washington. The geographical adjacency among the countries in the two states provides a good example of irregular lattice space. Each county has different numbers of "neighbors" (adjacent counties). Within each county we use the election result as a Bernoulli response and the median household income as a covariate. Assuming an isomorphic relationship between the variables in each county, the GTN can be used to model the among-county spatial correlation as well as within-county regression-like associations. Many lattice-type spatial data are aggregated over administrative units, study blocks, or other types of regions, both regular and irregular, and this example showcases how generalized tree networks can be used on these lattices.

For both the simulated and election data, the graphs' joint distribution functions are written according to their clique partitions. Unknown parameters are estimated

using a Bayesian approach with Markov chain Monte Carlo (MCMC) algorithms. We compare the generalized tree network model with Bayesian Gaussian kriging model (Diggle et al., 1998), conditional autoregressive (CAR) model (Besag, 1974), and simple logistic regression model without considering any spatial dependence, and notice an improvement of the generalized tree networks over the other models on the parameters' posterior estimates and predictions.

4.2 ICG on second order nearest neighbor lattices

In this section we define isomorphic chain graphs on second order nearest neighbor regular lattices and propose a coding method to the conditional Markov random fields. These lattices and coding method can be viewed as a progression from the first order nearest neighbor lattices examined in Appendix C. We show how to partition the graphs using cliques to obtain estimable joint distributions. The models for the cliques use multivariate copulas for discrete and continuous nodes.

The discussion in this section revolves around a somewhat general case, a $\mathbf{G}_{XY}$ ICG defined on a finite lattice. The example shows how a typical ICG, considered as a generalized tree network, may be applied on its second order nearest neighbor lattice. With modifications for the different between-site and within-site association patterns, we can create many different types of ICG that span over all four categories that were listed in Chapter 2, namely, $\mathbf{G}_{\varnothing}$, $\mathbf{G}_{Y}$, $\mathbf{G}_{X}$, and $\mathbf{G}_{XY}$, all of which may be partitioned by their induced generalized tree networks and junction trees.

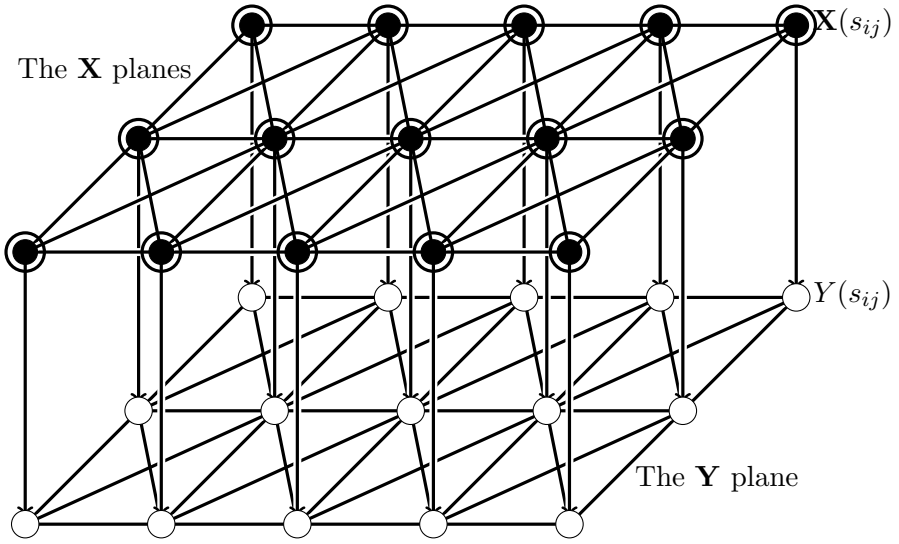
4.2.1 Graphical representation

In the accompanying example of this section, we assume there are $n \times m$ sites on the lattice and $p + 1$ random variables observed at each site with $p \geqslant 2$. We assume that at each site the $p + 1$ variables have the same association structure (*i.e.,* the graph is isomorphic). One of the random variables, $\mathbf{Y}$, is assumed to be Bernoulli distributed, while the rest of them (all $\mathbf{X}$'s) are assumed to be Gaussian. Each $\mathbf{X}$ node is assumed to connect with its within-site counterpart $\mathbf{Y}$ node by a directed edge, and the $\mathbf{X}$ nodes themselves are assumed to be independent from each other. Put into a regression context, we may consider $\mathbf{X}$'s at each site to be the explanatory variables, and $\mathbf{Y}$ to be the responses. We also assume all variables to be spatially correlated. Since at each site there are more than one node that is correlated with nodes at other sites, the graph can be categorized as a $\mathbf{G}_{XY}$ ICG.

We focus primarily on the $\mathbf{G}_{XY}$ graphs, because the other three types of ICG, $\mathbf{G}_X$, $\mathbf{G}_Y$, and $\mathbf{G}_\varnothing$, can all be considered as special cases of this type. For the sake of simplicity, we consider only finite lattices. Assume that on a lattice, $\mathcal{G}$, with $n \times m$ sites, there is one "response" and p "explanatory variables" per site. Each of the p explanatory variables is denoted as $\mathbf{X}_1$, $\mathbf{X}_2, \ldots, \mathbf{X}_p$, and together with response $\mathbf{Y}$ they can be perceived graphically as $p + 1$ separate planes of lattice nodes.

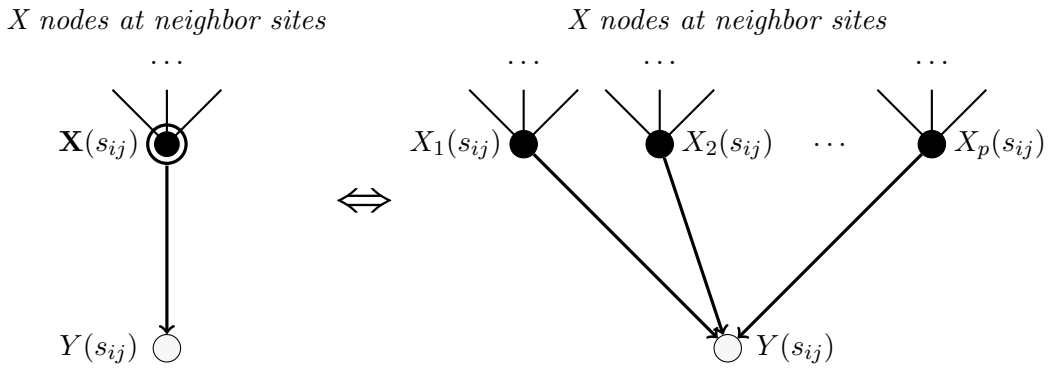
Graphically, the isomorphic chain graph $\mathbf{G}_{XY}$ can be represented by Figure 4.1 under a second order nearest neighbor scheme. The response neighborhood structure is illustrated by the lower, white nodes $\mathbf{Y}$ plane, $\mathcal{G}_{\mathbf{Y}}$, while the explanatory variable spatial correlations are shown by the upper, double black nodes $\mathbf{X}$ planes, $\mathcal{G}_{\mathbf{X}}$. The directed arrows represent the logit regression, $Y(s_{ij})|\mathbf{X}(s_{ij}) \sim$ Bernoulli (π_{ij}), where

Figure 4.1: Second order nearest neighbor lattice: ICG $\mathbf{G}_{XY}$



π_{ij} is modeled with the logit link. For visual clarification, the p different predictors $\{\mathbf{X}(s_{ij}) = X_1(s_{ij}), \ldots, X_p(s_{ij})\}$ at each site are confined into one double black node. If we take one site s_{ij} only, then the subgraph induced on the site can be properly expanded into Figure 4.2. These induced subgraphs are all the same for each site, hence keeping $\mathbf{G}_{XY}$ an ICG.

Figure 4.2: Expanding Figure 4.1.



4.2.2 Parameterization of ICG $\mathbf{G}_{XY}$

Denote the ICG in Figure 4.1 as $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V} = (\mathbf{Y}, \mathbf{X})$. We introduce the following notation as formal definitions of the nodes:

$$\begin{aligned}
\mathbf{Y} &= \{Y(s_{11}), \ldots, Y(s_{nm})\}^T, \\
\mathbf{X}(s_{ij}) &= \{X_1(s_{ij}), \ldots, X_p(s_{ij})\}, \\
\mathbf{X}_d &= \{X_d(s_{11}), \ldots, X_d(s_{nm})\}_{1\times nm}, \\
\mathbf{X} &= \{\mathbf{X}_1^T \ldots, \mathbf{X}_p^T\}_{nm\times p}, \text{ and} \\
\mathbf{V}_{ij} &= \{Y(s_{ij}), X(s_{ij})\},
\end{aligned}$$

with $i = 1, \ldots, n, j = 1, \ldots, m, d = 1, \ldots, p$. The node set of the graph has dimensions $nm \times (p + 1)$, and the node set at a particular site s_{ij} is $\mathbf{V}_{ij}$. Besides being the explanatory variable, we also refer to $\mathbf{X}_d$ as the d-th predictor in the graph.

The edges of the graph are determined by the spatial and regressional associations in the ICG. As mentioned above, we consider the model with Bernoulli responses and Gaussian explanatory variables. We assume that each predictor is independent from the other ones at the same site, or $X_{d_1}(s_{ij}) \perp\!\!\!\perp X_{d_2}(s_{ij})$. For the $\mathbf{X}_d$ predictor, its spatial dependence structure is modeled by partial correlation coefficients, $\xi^{(d)}$,

between neighboring sites.

$$
\begin{aligned}
X_d(s_{ij}) &\sim N(\mu_d, \sigma_d^2). && (4.1)\\
\text{Corr } (X_{d_1}(s_{ij}), X_{d_2}(s_{ij})) &= 0 \text{ , when } d_1 \neq d_2. && (4.2)\\
\text{Corr } (X_d(s_{ij}), X_d(s_{kl}) \mid \mathbf{X}_{d-ij,kl}) &= \begin{cases} 0, & \text{when } s_{ij}, s_{kl} \text{ are non-neighbors,} \\ \xi_{ij\_kl}^{(d)}, & \text{when } s_{ij}, s_{kl} \text{ are neighbors,} \end{cases} \\
\xi_{ij\_kl}^{(d)} &= \exp\left[\frac{-3\delta(ij,kl)^2}{r_d^2}\right], && (4.3)
\end{aligned}
$$

where $\mathbf{X}_{d-ij,kl}$ denotes all the nodes from $\mathbf{X}_d$ except the indexed two exclusions, $X_d(s_{ij})$ and $X_d(s_{kl})$, and $\xi_{ij\_kl}^{(d)}$ is the partial correlation coefficient between two neighboring sites $X_d(s_{ij})$ and $X_d(s_{kl})$. $\delta(ij,kl)$ measures the Euclidean distance between sites s_{ij} and s_{kl}. In a second order nearest neighbor regular lattice there are two distinct $d(i,j)$'s: 1 and $\sqrt{2}$, associated with first order and diagonal neighbors respectively. r_d is known as the *effective range of correlation* or *correlation length* of the d-th predictor. It measures the greatest distance at which the correlation would be more than 0.05. The random field, $\mathbf{X}_d$, specified by Equation (4.1) - (4.3) is therefore second order stationary and isotropic (Gaetan and Guyon, 2010, page 8), which means that its distribution is invariant by site location and by direction. For instance, $\xi_{ij\_kl}^{(d)}$ is only dictated by the distance $\delta(ij,kl)$ between the two sites s_{ij} and s_{jk}, but not by the relative location of these sites in terms of the lattice nor by the direction of the distance.

We have chosen this conditional approach over the unconditional one, because in this approach each predictor plane ($\mathcal{G}_{\mathbf{X}_d}$) may be treated as a standalone Gaussian Markov random field. Besag and Kooperberg (1995) introduced the class of Gaussian Markov random fields and noted that the partial correlation is always zero between two

non-neighboring nodes. When the nodes are neighbors, we assume that the non-zero pairwise correlation (4.3) follows a Gaussian correlation model as in Schabenberger and Gotway (2004). The partial correlation also has a very close tie to the inverse of the unconditional variance-covariance matrix of $\mathbf{X}_d$, given that it is invertible (Rue, 1999). If the precision matrix is denoted as $\mathbf{Q}^{(d)}$ and its elements $Q^{(d)}_{ij,sl}$, then

$$\text{Corr } (X_d(s_{ij}), X_d(s_{kl}) \mid \mathbf{X}_{d-ij,kl}) = -\frac{Q^{(d)}_{ij,kl}}{\sqrt{Q^d_{ij,ij} Q^{(d)}_{kl,kl}}}.$$

The $\mathbf{Q}^{(d)}$ matrix may be inverted for the unconditional correlations. To ensure a positive definite $\mathbf{Q}^{(d)}$, the Gaussian Markov random field does not allow negative pairwise correlations.

The design of the response variable structure follows a similar logic to the predictors. The correlation between two neighboring nodes, $Y(s_{ij})$ and $Y(s_{kl})$, is measured by a conditional correlation coefficient ρ_{ij_kl}. Between response and predictors, the link function to associate them is set to be the logit link:

$$\begin{aligned}
\text{logit } (\pi_{ij}) &= b_0 + \mathbf{X}(s_{ij})\mathbf{B_1}, && (4.4)\\
Y(s_{ij}) \mid \mathbf{X}(s_{ij}) &\sim \text{Bernoulli } (\pi_{ij}), && (4.5)\\
\text{Corr } (Y(s_{ij}), Y(s_{kl}) | \mathbf{X}(s_{ij}), \mathbf{X}(s_{kl}), \mathbf{Y}_{-ij,kl}) &= \begin{cases} 0, \\ \text{when } s_{ij}, s_{kl} \text{ are non-neighbors;} \\ \rho_{ij_kl}, \\ \text{when } s_{ij}, s_{kl} \text{ are neighbors;} \end{cases} \\
\rho_{ij_kl} &= \exp\left[\frac{-3\delta(ij,kl)^2}{r_y^2}\right]. && (4.6)
\end{aligned}$$

In our notation $\mathbf{Y}_{-ij,kl}$ denotes the set of response nodes except the two with indices ij and kl. ρ_{ij_kl} is the partial correlation coefficient, $\delta(ij, kl)$ the Euclidean distance, and r_y is the effective range parameter. π_{ij} determines the mean level of $Y(s_{ij})$ since it is a Bernoulli random variable, as $\mathbf{Y}$ is also second order stationary and isotropic. b_0 and $\mathbf{B_1}$ are the regression intercept and parameter vector in the link function, with $\mathbf{B_1} = \{b_{11}, \ldots, b_{1p}\}^T$.

Equations (4.3) and (4.6) state the local conditional correlation structures for variables $\mathbf{X}_d$'s and $\mathbf{Y}$. With the help of Figure 4.1, the conditional independence of the ICG can be identified under the LWF Markov property (Lauritzen, 1996):

$$\mathbf{X}_{d_1} \perp\!\!\!\perp \mathbf{X}_{d_2} \qquad \forall d_1 \neq d_2, \tag{4.7}$$

$$Y(s_{ij}) \perp\!\!\!\perp \mathbf{X}(s_{kl}) \mid \{\mathbf{X}(s_{ij}), nb(Y(s_{ij}))\} \qquad \forall ij \neq kl, \tag{4.8}$$

where $nb(Y(s_{ij}))$ denotes the neighbor set of $Y(s_{ij})$. Equation (4.7) confirms the conditional independences specified by (4.2), while Equation (4.8) likewise corroborates the local Markov properties. The conditional dependences among $\mathbf{V}$ are:

$$Y(s_{ij}) \not\perp\!\!\!\perp \mathbf{X}(s_{ij}) \qquad \forall\ ij, \tag{4.9}$$

$$X_d(s_{ij}) \not\perp\!\!\!\perp X_d(s_{kl}) \qquad \forall\ d, ij \neq kl, \tag{4.10}$$

$$Y(s_{ij}) \not\perp\!\!\!\perp Y(s_{kl}) \mid \{\mathbf{X}(s_{ij}), \mathbf{X}(s_{kl})\} \qquad \text{when } s_{ij}, s_{kl} \text{ are neighbors.} \tag{4.11}$$

Equation (4.9) accounts for the within-site regression from (4.4) and (4.5). Equation (4.10) and (4.11) aligns with the spatial correlation among the explanatory and response variables, respectively.

Although our purpose in graphical model development is to accommodate multivariate spatial dependences in the graph, it is often the estimates of regression parameters that are of the most interest. In practice we want to account for, rather than estimate, the spatial dependence structure, and therefore the partial correlation coefficients, $\xi^{(d)}_{ij\_kl}$ and ρ_{ij_kl}, as well as the range parameters, r_d and r_y, are considered to be nuisance parameters in the model. Our main focus is to estimate b_0 and $\mathbf{B_1}$, the intercept and slope vectors.

In a completely full model (*i.e.*, if each pair of neighbors was allowed to have its own unique dependence strength, measured by ρ_{ij_kl}) there would effectively be $nm(nm-1)/2$ unknown correlation coefficients to be estimated in the model on a regular $n \times m$ plane. By assuming stationarity and isotropy, however, we reduce the number of parameters to one for each variable only. Because there is a one-to-one correspondence between r_y and ρ_{ij_kl} and between r_d and $\xi^{(d)}_{ij\_kl}$, we may calculate different correlation coefficients given the different pairings of the neighbors in the lattice. When $\delta(ij,kl) = 1$, that is, when s_{ij} and s_{kl} are neighbors on the first order, we have $\rho_{ij\_kl} = \rho$, $\xi^{(d)}_{ij\_kl} = \xi_d$. When the pairs are diagonal neighbors, we have $\rho_{ij\_kl} = \rho^2, \xi^{(d)}_{ij\_kl} = \xi^2_d$, and $\delta(ij,kl) = \sqrt{2}$.

Given all the previous definitions and assumptions, the parameter space of $\mathbf{V} = (\mathbf{X}, \mathbf{Y})$ is denoted as

$$\Theta_{\mathcal{G}} = \{\rho, \xi_1, \ldots, \xi_p, \mu_1, \ldots, \mu_p, \sigma^2_1, \ldots, \sigma^2_p, b_0, \mathbf{B}_1\}, \quad \text{or equivalently} \tag{4.12}$$

$$\Theta^*_{\mathcal{G}} = \{r_y, r_1, \ldots, r_p, \mu_1, \ldots, \mu_p, \sigma^2_1, \ldots, \sigma^2_p, b_0, \mathbf{B}_1\}, \tag{4.13}$$

that is, there are a total of $4p + 2$ parameters to be estimated in $\mathbf{G}_{XY}$.

4.2.3 Cliques and the partition of $\mathbf{G}_{XY}$

Since there are no directed cycles in Figure 4.1 and 4.2, the $\mathbf{G}_{XY}$ graph $\mathcal{G}$ can be moralized. When moralizing $\mathcal{G}$, directed edges $X_d(s_{ij}) \rightarrow Y(s_{ij})$ are replaced by undirected edges $X_d(s_{ij}) - Y(s_{ij})$. Since $X_1(s_{ij}), X_2(s_{ij}), \ldots, X_p(s_{ij})$ are all common parents of $Y(s_{ij})$, every pair of them will be connected by undirected edges. Between any two nodes of $\{X_1(s_{ij}), X_2(s_{ij}), \ldots, X_p(s_{ij})\}$ there is an undirected edge after moralization. This subset forms a clique, and with the inclusion of the response node, $\mathbf{V}_{ij} = \{Y(s_{ij}), X_1(s_{ij}), X_2(s_{ij}), \ldots, X_p(s_{ij})\}$ is a maximal clique. These cliques characterize the within-site structure of the moralized Markov random field. Each maximal clique has a size of $p+1$. To emphasize its within-site effect and the fact that they are summarized over the vertical direction, these $(p+1)$-cliques will be denoted as Cl^Z cliques (this should not be confused with the partition function $Z(\Theta_{\mathcal{G}})$).

Apart from the Cl^Z cliques, the other maximal cliques can be identified on the $\mathbf{Y}$ and $\mathbf{X}_d$ planes and will be known as the Cl^Y and Cl^{X_d} cliques. Under the second order nearest neighbor structure, each set of four square-positioned nodes on the $\mathbf{Y}$ or $\mathbf{X}_d$ planes consists of a maximal 4-cliques, for example $Cl^Y_{11} \equiv \{Y(s_{11}), Y(s_{12}), Y(s_{21}), Y(s_{22})\}$. Formally we may denote all the maximal cliques of the moralized graph of Figure 4.1 by

$$\mathbf{V}_{ij} = Cl^Z_{ij} \equiv \{Y(s_{ij}), X_1(s_{ij}), \ldots, X_p(s_{ij})\}, \tag{4.14}$$

$$Cl^Y_{ij} \equiv \{Y(s_{ij}), Y(s_{i,j+1}), Y(s_{i+1,j}), Y(s_{i+1,j+1})\}, \tag{4.15}$$

$$Cl^{X_d}_{ij} \equiv \{X_d(s_{ij}), X_d(s_{i,j+1}), X_d(s_{i+1,j}), X_d(s_{i+1,j+1})\}. \tag{4.16}$$

The superscripts of the cliques denote which plane it is on, and the subscripts ij mark

the top-left site of the clique in the cases of the Cl^Y and Cl^{X_d} cliques. At any given set of four square-positioned sites there are $p+1$ overlapping Cl^{X_d} and Cl^Y cliques because there are p predictor and one response nodes. There is no clique spanning across different predictors.

A key advantage of this moralized generalized tree network is that the distribution $f(\mathbf{X})$ depends only on the Cl^{X_d} cliques, while $f(\mathbf{Y}|\mathbf{X})$ depends only on Cl^Y and Cl^Z cliques. We may therefore, based on Hammersley-Clifford Theorem and Equation (3.3), partition the joint distribution of the graph, $p_{\mathcal{G}}$, using distributions on the cliques separately. To be more specific, we can write

$$p_{\mathcal{G}^M}(\mathbf{X}) = \prod_d p_{\mathcal{G}^M}(\mathbf{X}_d) \propto \prod_d \prod_{Cl_{ij}^{X_d}} \phi_{Cl_{ij}^{X_d}}(\mathbf{X}), \tag{4.17}$$

$$p_{\mathcal{G}^M}(\mathbf{Y} \mid \mathbf{X}) \propto \prod_{Cl_{ij}^{Y}} \phi_{Cl_{ij}^{Y}}(\mathbf{Y}) \times \prod_{\mathbf{V}_{ij}} \phi_{\mathbf{V}_{ij}}(\mathbf{V}), \tag{4.18}$$

$$\begin{aligned} p_{\mathcal{G}^M}(\mathbf{V}) &= p_{\mathcal{G}^M}(\mathbf{Y} \mid \mathbf{X}) p_{\mathcal{G}^M}(\mathbf{X}) \\ &= \frac{1}{Z(\Theta_{\mathcal{G}^M})} \prod_{Cl_{ij}^{Y}} \phi_{Cl_{ij}^{Y}}(\mathbf{Y}) \times \prod_{\mathbf{V}_{ij}} \phi_{\mathbf{V}_{ij}}(\mathbf{V}) \times \prod_d \prod_{Cl_{ij}^{X_d}} \phi_{Cl_{ij}^{X_d}}(\mathbf{X}). \end{aligned} \tag{4.19}$$

Because the moralized GTN $\mathcal{G}^M$ is not a junction tree, there is no closed form for $Z(\Theta_{\mathcal{G}^M})$. When we consider $\mathbf{Y}$ and $\mathbf{X}_d$ planes from Figure 4.1 we may see that many 4-cliques join together to form large chordless cycles. Since the size of these chordless cycles are already fairly large, adding more edges to make them cliques is not a very effective remedy. Instead, we use coding methods to create conditional Markov random fields based on these moralized GTNs.

4.2.4 Coding the conditional Markov random field

For second order nearest neighbor regular lattices we propose a coding scheme similar to the snake coding in the first order lattices (see Appendix C). We choose the conditioning nodes to be every third rows or columns, in order to separate the junction trees in the rest of the graph. Figure 4.3 provides an illustration to this coding method. In the figure we are considering only one plane of the graph, for instance the $\mathbf{Y}$ plane. We may do this without loss of generality since all the $\mathbf{X}_d$ and $\mathbf{Y}$ planes in a ICG are identical in structure.

In Figure 4.3, we consider every third row of nodes and mark it white. We call these rows the *conditioning rows*. We may also call the first and second rows of black nodes in the figure the first *belt* of the graph, and the fourth and fifth rows the second belt, *etc*, because they represent the first and second rows of maximal cliques to be included in the junction tree on which the conditional likelihood function is calculated.

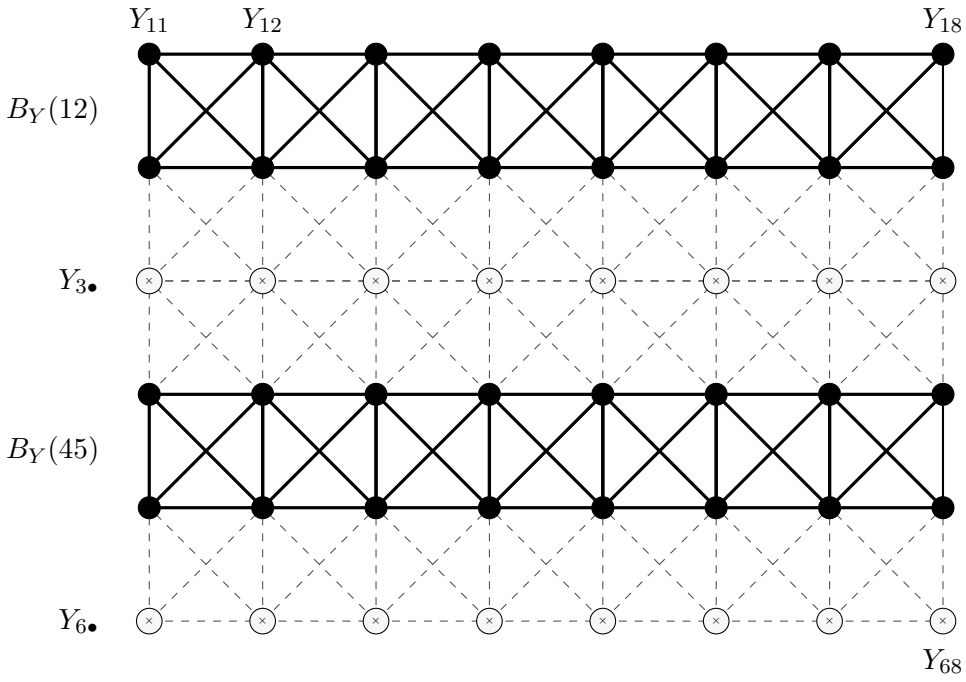
For simplicity, we write $Y_{ij} \equiv Y(s_{ij})$, and propose to use $Y_{i\bullet}$ to denote the i-th row of nodes, and $Y_{\bullet j}$ as the j-th column of nodes in Figure 4.3 (that is, $Y_{i\bullet} \equiv \{Y_{i1}, Y_{i2}, \ldots, Y_{im}\}, Y_{\bullet j} \equiv \{Y_{1j}, Y_{2j}, \ldots, Y_{nj}\}$, $i = 1, \ldots, n; j = 1, \ldots, m$). A *belt* $B_Y(i, i+1)$ or $B_Y(j, j+1)^T$ on the plane is defined as two adjacent rows or columns of black nodes, or, $B_Y(i, i+1) = \{Y_{i\bullet}, Y_{i+1\bullet}\} = \cup_j Cl_{ij}^Y$, and $B_Y(j, j+1)^T = \{Y_{\bullet j}, Y_{\bullet j+1}\} = \cup_i Cl_{ij}^Y$. For instance, in Figure 4.3 we have shown that $B_Y(12) = \{Y_{1\bullet}, Y_{2\bullet}\}$ and $B_Y(45) = \{Y_{4\bullet}, Y_{5\bullet}\}$, as they represent two of the row belts. Between any pair of belts there is at least one row of conditioning white nodes, such as $Y_{3\bullet}$, $Y_{6\bullet}$, *etc*. Based

on the separation rule of the global Markov property on undirected graphs, we have

$$\begin{aligned} B_Y(12) \perp\!\!\!\perp B_Y(45) \perp\!\!\!\perp \ldots \quad | \quad \{Y_{3\bullet}, Y_{6\bullet}, \ldots\}, \quad \text{and} \\ B_Y(12)^T \perp\!\!\!\perp B_Y(45)^T \perp\!\!\!\perp \ldots \quad | \quad \{Y_{\bullet 3}, Y_{\bullet 6}, \ldots\}, \end{aligned} \tag{4.20}$$

because graphically $Y_{3\bullet}$ separates $B_Y(12)$ and $B_Y(45)$, and $Y_{6\bullet}$ separates $B_Y(45)$ and $B_Y(78)$, and so forth. In other words, in the $\mathbf{Y}$ plane the belts are conditionally independent from each other given the conditioning rows. A similar result holds in $\mathbf{X}$ planes.

Figure 4.3: Belt coding for the second order nearest neighbor lattice



After conditioning, the remaining nodes form a conditional Markov random field, and belts $B_Y(12)$, $B_Y(45)$ are junction trees because there is no chordless cycle in them anymore. Any cycle with four or more nodes in the belts has at least one chord.

$\mathbf{X}_1, \ldots \mathbf{X}_p$ planes may be coded in the exact same manner.

Similar to the snake coding scheme in Appendix C, there could be more than one possible way of specifying the conditional rows or columns. Subsequently we may create multiple conditional Markov random fields from the same undirected graph whose parameter estimates do not necessarily have to agree. If we moved all the conditioning rows in Figure 4.3 down one row, so that the conditioning rows include $Y_{1\bullet}, Y_{4\bullet}$, and $Y_{7\bullet}$, and the first belt is $B_Y(23)$, and second belt is $B_Y(56)$, then together they form a new conditional Markov random field, different from the first run shown in the figure. Move down the conditioning rows yet again to $Y_{2\bullet}, Y_{5\bullet}$, and $Y_{8\bullet}$ *etc*, and we would have the third conditional Markov random field. Each run is expected to produce a different set of parameter estimates. Repeat this process to the columns, and three more runs and three more sets of estimates can be acquired. Together we would have six different sets of estimates for a second order nearest neighbor regular lattice, all based on different conditional Markov random fields. Based on these six estimates a single combined estimate may be obtained. These can be either simple or weighted averages of the estimates when the conditioning sets are relatively comparable (Besag, 1974, 1975).

Another benefit of this coding scheme is the increase of utility rate. The *utility rate* measures the percentage of nodes and edges included in the conditional Markov random fields. For each run of the lattice this rate varies according to its size and shape and could be different, but generally speaking we can expect to utilize about two third of the nodes, and slightly less than half of the edges for a single run. These utility rates are much larger and more efficient compared with the chess board coding schemes, which are only 50% for the nodes and 0% for the edges (see Figure C.4).

4.2.5 The joint distribution function

With each conditional Markov random field consisting of several row or column belts, its joint distribution may be first partitioned along the belts. They are subsequently further partitioned on a finer level of maximal cliques nested within the belts. Consider the conditioning rows as a whole, and the belts are conditionally independent to each other and their distribution product is the joint distribution. For instance, if we name the conditioning set of nodes from Figure 4.3 as $\mathbf{C}_3 = \{Y_{3\bullet}, Y_{6\bullet}, \ldots\}$, then the conditional Markov random field obtained may be defined as $\tilde{\mathbf{Y}}_{\mathbf{C}_3} \equiv \{B_Y(12) \cup B_Y(45) \cup \ldots\} = \mathbf{Y} \backslash \{Y_{3\bullet}, Y_{6\bullet}, \ldots\}$. According to the global Markov property on undirected graphs, we may write the joint distribution as

$$p_{\mathcal{G}}(\tilde{\mathbf{Y}}_{\mathbf{C}_3}) = p_{\mathcal{G}} \left\{ \bigcup_{k=12,45,\ldots} B_Y(k) \right\} = \prod_{k=12,45,\ldots} p_{\mathcal{G}}[B_Y(k)]. \tag{4.21}$$

The subscript for the conditioning set indicates the first row of nodes that was conditioned on, in this case the third row, hence $\mathbf{C}_3$. For the other two row runs, the conditioning sets are $\mathbf{C}_1$ and $\mathbf{C}_2$, and the conditional Markov random fields change accordingly. The question now reduces to finding $p_{\mathcal{G}}[B_Y(k)]$. Since $B_Y(k)$ is a junction tree, $p_{\mathcal{G}}[B_Y(k)]$ in turn may be factored by the maximal cliques and separators. Consider the first two 4-cliques from the left side of $B_Y(12)$, namely, C_{11}^Y and C_{12}^Y. Following Kirshner (2007), we have

$$\begin{aligned} p_{\mathcal{G}}(C_{11}^Y) p_{\mathcal{G}}(C_{12}^Y) &= p_{\mathcal{G}}(Y_{11}, Y_{21}, Y_{12}, Y_{22}) p_{\mathcal{G}}(Y_{12}, Y_{22}, Y_{13}, Y_{23}) \\ &= p_{\mathcal{G}}(Y_{11}, Y_{21} | Y_{12}, Y_{22}) p_{\mathcal{G}}(Y_{12}, Y_{22}) p_{\mathcal{G}}(Y_{13}, Y_{23} | Y_{12}, Y_{22}) p_{\mathcal{G}}(Y_{12}, Y_{22}) \\ &= p_{\mathcal{G}}(Y_{11}, Y_{21}, Y_{12}, Y_{22}, Y_{13}, Y_{23}) p_{\mathcal{G}}(Y_{12}, Y_{22}) \end{aligned} \tag{4.22}$$

Since $S^Y_{12} \equiv \{Y_{12}, Y_{22}\}$ is the separator between C^Y_{11} and C^Y_{12}, Equation (4.22) can be rewritten as

$$p_{\mathcal{G}}(C^Y_{11} \cup C^Y_{12}) = \frac{p_{\mathcal{G}}(C^Y_{11}) p_{\mathcal{G}}(C^Y_{12})}{p_{\mathcal{G}}(S^Y_{12})}. \tag{4.23}$$

The joint distribution of two adjacent cliques in a belt may be partitioned as the product of the distributions for each individual clique divided by the distribution of their separator. Repeatedly applying (4.23) from one end (left or top) of the belt to another (right or bottom), the distribution on the belt can be expressed as

$$p_{\mathcal{G}}(B_Y(12)) = \frac{\prod_{i=1}^{m-1} p_{\mathcal{G}}(C^Y_{1i})}{\prod_{i=2}^{m-1} p_{\mathcal{G}}(S^Y_{1i})}. \tag{4.24}$$

Here, m is the length of the rows in the lattice. For each belt, a partition like Equation (4.24) may be computed. Combining these using Equation (4.21), and we may obtain a distribution for $\tilde{\mathbf{Y}}_{\mathbf{C}_3}$ based on the cliques. In the Hammersley-Clifford Theorem notation, $p(\tilde{\mathbf{Y}}_{\mathbf{C}})$ can be written as

$$p_{\mathcal{G}}(\tilde{\mathbf{Y}}_{\mathbf{C}_3}) = \frac{1}{Z(\Theta_{\mathcal{G}})} \prod_{Cl^Y_{ij} \in \tilde{\mathbf{Y}}_{\mathbf{C}_3}} \phi_{Cl^Y_{ij}}(\mathbf{Y}), \quad \text{where} \tag{4.25}$$

$$Z(\Theta_{\mathcal{G}}) = \prod_{k=1,4,7,\ldots} \prod_{j=2}^{m-1} p_{\mathcal{G}}(S^Y_{kj}). \tag{4.26}$$

The joint distributions of the 4-cliques on $\mathbf{Y}$ can be treated as the potential functions, *i.e.*, $\phi_{Cl^Y_{ij}}(\mathbf{Y}) = p_{\mathcal{G}}(Cl^Y_{ij})$. Equation (4.25) also applies to the all six different runs on the graph, including the other two row runs $\tilde{\mathbf{Y}}_{\mathbf{C}_1}$ and $\tilde{\mathbf{Y}}_{\mathbf{C}_2}$, as well as column runs, $\tilde{\mathbf{Y}}^T_{\mathbf{C}_1}$ to $\tilde{\mathbf{Y}}^T_{\mathbf{C}_3}$, with a simple change of conditioning and separator sets. Because $\mathbf{Y}$ is

Bernoulli distributed, the numerator and denominator of Equation (4.25) are products of discrete 4-copula and bivariate copula. Following the multivariate discrete Gaussian copula function (Madsen, 2009) we may write

$$\begin{aligned}\phi_{Cl_{ij}^{Y}}(\mathbf{Y}) &= p_{\mathcal{G}}(Cl_{ij}^{Y}) = \sum_{j_{kl}=1}^{2} \cdots \sum_{j_{k+1,l+1}=1}^{2} \{(-1)^{j_{kl}+\ldots+j_{k+1,l+1}} \\ &\quad \Phi_{\Sigma}\left[\Phi^{-1}(u_{kl j_{kl}}), \ldots, \Phi^{-1}(u_{k+1,l+1 j_{k+1,l+1}})\right]\}; \qquad (4.27)\\ \text{with } u_{kl1} &= F_{kl}[Y(s_{kl})], \quad u_{kl2} = F_{kl}[Y(s_{kl})-]. \qquad (4.28)\end{aligned}$$

Separator distributions $p_{\mathcal{G}}(S_{kj}^{Y})$ in Equation (4.26) can be modeled in the same manner.

The joint distributions of the predictor variables $\mathbf{X}$ can be partitioned in a way very much like on the $\mathcal{G}_{\mathbf{Y}}$ sub-graph. The potentials for the cliques on the $\mathbf{X}_d$ planes can also be written as the product of $p_{\mathcal{G}}(Cl_{ij}^{X_d})$, with the partition function being the product of distributions of separators $S_{kj}^{X_d}$ on $\mathbf{X}_d$ planes. Because the predictors are Gaussian, the maximal cliques and separators $Cl_{ij}^{X_d}$ and $S_{ij}^{X_d}$ follow multivariate Gaussian distributions. Lastly, the distribution functions for Cl_{ij}^{Z}, the connecting edges between $\mathbf{X}$ and $\mathbf{Y}$ planes, are

$$\begin{aligned}\phi_{Cl_{ij}^{Z}}(\mathbf{V}) &= p_{\mathcal{G}}(\mathbf{V}_{ij}) \\ &= p_{\mathcal{G}}(Y_{ij} \mid X_1(s_{ij}), \ldots X_p(s_{ij})) p_{\mathcal{G}}(X_1(s_{ij}), \ldots X_p(s_{ij})) \\ &= p_{\mathcal{G}}(Y_{ij}) \mid X_1(s_{ij}), \ldots X_p(s_{ij})) \prod_d p_{\mathcal{G}}(X_d(s_{ij})) \\ &= \pi_{ij}^{Y_{ij}}(1-\pi_{ij})^{(1-Y_{ij})}(\prod_d \sigma_d)^{-1}(2\pi)^{-p/2} \exp\left[-\sum_d \frac{(X_d(s_{ij})-\mu_d)^2}{2\sigma_d^2}\right] \qquad (4.29)\end{aligned}$$

when Y_{ij}'s are Bernoulli responses and $X_p(s_{ij})$'s are Gaussian. Together $p_{\mathcal{G}}(\mathbf{V}_{ij})$ and $\phi_{Cl_{ij}^{Y}}(\mathbf{Y})$ will determine $p_{\mathcal{G}}(\mathbf{Y}|\mathbf{X})$.

With all maximal cliques on the junction tree written out in closed forms, we can combine them together using the Hammersley-Clifford theorem. The partitions such as Equation (4.19) means that to write an estimable joint distribution for the junction tree, we need to first identify the conditioning set of nodes and the belts, evaluate the maximal cliques on the belts, then combine the belts, and finally multiply together all the variable planes and sites. Take Figure 4.3 for example again. If we use $\mathbf{C}_{d3} \equiv \{X_{d3\bullet}, X_{d6\bullet}, \ldots\}$ to denote the set of conditioning rows on $\mathbf{X}_d$ plane that contains the third, sixth, and ninth rows, then $\tilde{\mathbf{X}}_{d\mathbf{C}_3}$ is the conditional Markov random field on $\mathbf{X}_d$ with $\tilde{\mathbf{X}}_{d\mathbf{C}_3} \equiv \mathbf{X}_d \backslash \mathbf{C}_{d3}$. Combine the predictor conditioning sets $\mathbf{C}_{d3}$'s with the response set $\mathbf{C}_3$, and $\tilde{\mathbf{X}}_{d\mathbf{C}_3}$'s with $\tilde{\mathbf{Y}}_{\mathbf{C}_3}$, we have

$$\mathbf{C}_3^V \equiv \bigcup_{d=1\ldots p} \mathbf{C}_{d3} \bigcup \mathbf{C}_3 \tag{4.30}$$

$$\tilde{\mathbf{V}}_{\mathbf{C}_3} \equiv \bigcup_{d=1\ldots p} \tilde{\mathbf{X}}_{d\mathbf{C}_3} \bigcup \tilde{\mathbf{Y}}_{\mathbf{C}_3}. \tag{4.31}$$

$\tilde{\mathbf{V}}_{\mathbf{C}_3}$ is the junction tree after the nodes from $\mathbf{C}_3^V$ are conditioned on. Writing the joint distribution on $\tilde{\mathbf{V}}_{\mathbf{C}_3}$ by the conditional distribution function, we have

$$p_{\mathcal{G}}(\tilde{\mathbf{V}}_{\mathbf{C}_3}) = p_{\mathcal{G}}(\mathbf{V} \backslash \mathbf{C}_3^V) = p_{\mathcal{G}}(\tilde{\mathbf{Y}}_{\mathbf{C}_3} \mid \bigcup_{d=1\ldots p} \tilde{\mathbf{X}}_{d\mathbf{C}_3}) \times p_{\mathcal{G}}(\bigcup_{d=1\ldots p} \tilde{\mathbf{X}}_{d\mathbf{C}_3}). \tag{4.32}$$

Equation (4.32) rewrites the conditional distribution $f(\mathbf{Y}, \mathbf{X}) = f(\mathbf{Y}|\mathbf{X})f(\mathbf{X})$ on the belt conditional Markov random fields. It means that the joint distribution of the whole graph is contributed by each of the $\mathbf{Y}$, $\mathbf{X}_d$ planes and sites $\mathbf{V}_{ij}$. The joint

distribution for junction tree $p_{\mathcal{G}}(\tilde{\mathbf{V}}_{\mathbf{C}_3})$ with respect to its maximal cliques is:

$$
\begin{aligned}
p_{\mathcal{G}}(\tilde{\mathbf{V}}_{\mathbf{C}_3}, \Theta_{\mathcal{G}}) &= \frac{1}{Z(\Theta_{\mathcal{G}})} \prod_{Cl_{ij}^{Y} \in \tilde{\mathbf{Y}}_{\mathbf{C}_3}} \phi_{Cl_{ij}^{Y}}(\mathbf{Y}) \times \prod_{\mathbf{V}_{ij} \in \tilde{\mathbf{V}}_{\mathbf{C}_3}} \phi_{\mathbf{V}_{ij}}(\mathbf{V}) \times \prod_{d} \prod_{Cl_{ij}^{X_d} \in \tilde{\mathbf{X}}_{d\mathbf{C}_3}} \phi_{Cl_{ij}^{X_d}}(\mathbf{X}) \\
= & \prod_{k=1,4,7,\ldots} \prod_{j=2}^{m-1} \left\{ p_{\mathcal{G}}(Y_{kj}, Y_{k+1,j}) \prod_{d} p_{\mathcal{G}}[X_d(s_{kj}), X_d(s_{k+1,j})] \right\}^{-1} \prod_{k=1,2,4,5,7,8,\ldots} \prod_{j=2}^{m-1} p_{\mathcal{G}}(Y_{kj})^{-p} \\
& \times \prod_{Cl_{ij}^{Y} \in \tilde{\mathbf{Y}}_{\mathbf{C}_3}} \left\{ \sum_{j_{kl}=1}^{2} \cdots \sum_{j_{k+1,l+1}=1}^{2} [(-1)^{j_{kl}+\ldots+j_{k+1,l+1}} \Phi_{\Sigma}(\Phi^{-1}(u_{klj_{kl}}) \ldots \Phi^{-1}(u_{k+1,l+1j_{k+1,l+1}}))] \right\} \\
& \times \prod_{\mathbf{V}_{ij} \in \tilde{\mathbf{V}}_{\mathbf{C}_3}} \left\{ \pi_{ij}^{Y_{ij}} (1-\pi_{ij})^{(1-Y_{ij})} (\prod_{d} \sigma_d)^{-1} (2\pi)^{-p/2} \exp\left[-\sum_{d=1}^{p} \frac{(X_d(s_{ij}) - \mu_d)^2}{2\sigma_d^2} \right] \right\} \\
& \times \prod_{d} \prod_{Cl_{ij}^{X_d} \in \tilde{\mathbf{X}}_{d\mathbf{C}_3}} p_{\mathcal{G}}[X_d(s_{kj}), X_d(s_{k,j+1}), X_d(s_{k+1,j}), X_d(s_{k+1,j+1})]. \qquad (4.33)
\end{aligned}
$$

Equation (4.33) may seem daunting at first, nevertheless it is not at all impossible to process. When we look at each individual factor of the equation, the most complicated ones only involve four correlated variables. The most complex factor, $\phi_{Cl_{ij}^{Y}}(\mathbf{Y})$, involves an inversion of 4×4 matrices, and a summation over 16 Gaussian copulas.

4.3 Simulating spatial correlations

4.3.1 Gaussian Markov random field

In this section we show how to generate spatially correlated normal predictors and Bernoulli responses on a lattice according to Equation (4.1) - (4.6). For simplicity and resource reasons, we limit the lattice to one predictor only, *i.e.* $p = 1$ and $\mathbf{X} = \mathbf{X}_1 = \{X_1(s_{ij}), i = 1 \ldots n, j = 1 \ldots m\}$.

The first step towards simulating the data we need on a second order lattice is to generate a stationary, autocorrelated multivariate Gaussian variable $\mathbf{X} = \{x_1, x_2, \ldots x_{n\times m}\}$. We assume that the autocorrelations between x_i's are governed by a precision matrix $\mathbf{Q}$. On the graph these autocorrelations are represented by undirected edges. The subgraph induced by $\mathbf{X}$ is a Markov random field. According to its local Markov property, we know that $x_i \perp\!\!\!\perp x_k \mid \{\mathbf{X}\backslash x_i, x_k\}$ when x_i and x_k are non-neighbors.

It is difficult to generate the autocorrelated Gaussian variable based on a zero mean vector and a given unconditional correlation or variance-covariance matrix, because the Markov properties do not clearly state the unconditional correlation $corr(x_i, x_k)$ for non-neighboring x_i and x_k. The LWF Markov properties only regulate the neighboring node correlations $corr(x_i, x_j)$ conditionally, and do not specify a clear result when x_i and x_k are on a path. Hence, the explanatory variable's correlation matrix is unknown to us. We cannot generate these variables directly using multivariate Gaussian distributions and their variance-covariance matrices. Instead, we use the algorithm proposed by Rue (1999, 2000) to generate these variables conditionally.

In this algorithm, instead of focusing the Gaussian variables based on the unconditional correlation matrix, we generate them based on the conditional variances and partial autocorrelation. The algorithm is based on the theory that a multivariate distribution can be defined by its conditional distributions. For instance, a Gaussian Markov Random Field (Gaussian MRF as defined in Chapter 4) $\mathbf{X}$ with respect to graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ can be defined as

$$\mathbf{X} \sim N_n(\mu, \Sigma), \qquad \text{and} \tag{4.34}$$

$$x_i \perp\!\!\!\perp x_j \mid \mathbf{X}_{-i,j}, \qquad \text{when } x_i, x_j \text{ are not neighbors.} \tag{4.35}$$

$\mathbf{X}_{-i,j}$ here denotes $\{\mathbf{X} \backslash x_i, x_k\}$, and $\mathbf{X}_{-i} \equiv \mathbf{X} \backslash x_i$. Alternatively, we may also specify the exact same Gaussian MRF through its conditional distribution functions:

$$x_i \mid \mathbf{X}_{-i} \sim N(\mu_{-i}, \Sigma_{-i}),\ i = 1 \ldots n. \tag{4.36}$$

When the variance-covariance matrix Σ of a Gaussian MRF is unknown, as in our case, its precision matrix, $\mathbf{Q}$, the inverse of its variance-covariance matrix, $\mathbf{Q} = \Sigma^{-1}$, is usually not difficult to specify. For a zero mean Gaussian MRF $\mathbf{X} \sim N_n(0, \mathbf{Q}^{-1})$, its joint density can be written as

$$p_{\mathcal{G}}(\mathbf{X}) = (2\pi)^{n/2} |\mathbf{Q}|^{1/2} \exp\left(-\frac{1}{2}\mathbf{X}^T \mathbf{Q} \mathbf{X}\right). \tag{4.37}$$

The precision matrix tells us everything that the variance-covariance matrix informs us, plus a little bit more. Supported by Brook's Lemma, Theorem 4.3.1 (Rue and Tjelmeland, 1999) holds true on Gaussian MRF, which greatly help us facilitate the simulation of autocorrelated Gaussian variables.

Theorem 4.3.1. *Let* $\mathbf{X}$ *be a Gaussian Markov random field with respect to* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$. *The precision matrix of* $\mathbf{X}$ *is* $\mathbf{Q}$. *Denote the* ij*-th entry of* $\mathbf{Q}$ *as* $Q_{i,j}$, *then*

1. $x_i \perp\!\!\!\perp x_j \mid \mathbf{X}_{-i,j} \Longleftrightarrow Q_{i,j} = 0$.

2. *Var* $(x_i \mid \mathbf{X}_{-i}) = Q_{i,i}^{-1}$.

3. *Corr* $(x_i, x_j \mid \mathbf{X}_{-i,j}) = -(Q_{i,i} Q_{j,j})^{-1/2} Q_{i,j}$.

The first part of the theorem says that when two nodes are non-neighbors, their corresponding precision matrix entry is zero; and only neighbors have non-zero entries.

The second part means the conditional variance of each node conditioned on the rest of the graph equals the inverse of the diagonal entry. The third part says that the partial correlation is a ratio between diagonal and off diagonal entries.

There are two major benefits of Theorem 4.3.1. Firstly, by discriminating the neighboring pairs of nodes from non-neighbors pairs, the theorem makes clear use of the precision matrix. Secondly, the predominant number of zero entries in $\mathbf{Q}$ ensure that it is a sparse matrix which, combined with the proper optimal indexing of the sites, may produce fast simulation (Rue, 1999).

To populate the non-zero elements of the precision matrix, we need to know the first and second order conditional moments of the Gaussian variables. All the elements in $\mathbf{Q}$ can then be solved using the result from Theorem 4.3.1. The conditional mean and variance of a single node x_i can be written as

$$\begin{aligned} \mathrm{E}\left(x_i \mid \mathbf{X}_{-i}\right) &= -Q_{i,i}^{-1} \sum_{i-j \in \mathbf{E}} Q_{i,j} x_j, & (4.38) \\ \mathrm{Var}\left(x_i \mid \mathbf{X}_{-i}\right) &= Q_{i,i}^{-1}. & (4.39) \end{aligned}$$

Per Gaussian correlation functions (Schabenberger and Gotway, 2004), in a second order nearest neighbor lattice we have

$$\begin{aligned} \mathrm{Var}\left(x_i \mid \mathbf{X}_{-i}\right) &= \sigma_x^2, \quad \text{and} & (4.40) \\ \mathrm{Corr}\left(x_i, x_j \mid \mathbf{X}_{-i,j}\right) &= \xi_x = \exp\left[\frac{-3d(i,j)^2}{r_x^2}\right], & (4.41) \end{aligned}$$

where $d(i,j)$ is the Euclidean distance between two neighboring sites x_i and x_j. In the lattice there are two distinct $d(i,j)$'s: 1 and $\sqrt{2}$, associated with first order and

diagonal (second order) neighbors respectively. r_x is the effective range of correlation or correlation length. Under this parameterization, entries in $\mathbf{Q}$ can be written as:

$$Q_{i,j} = \begin{cases} 1/\sigma_x^2 & \text{if } i = j, \\ -\xi_x/\sigma_x^2 & \text{if } i, j \text{ are neighbors,} \\ 0 & \text{otherwise.} \end{cases} \tag{4.42}$$

A Gaussian MRF with these known parametric terms can be simulated from its mean vector μ and precision matrix $\mathbf{Q}$. To ensure the existence and uniqueness of the full conditional densities, there is no extra constraint imposed upon $\mathbf{Q}$ besides it being a positive definite matrix. Without loss of generality, we can always assume that $\mathrm{E}(\mathbf{X}) = \mathbf{0}$, since if $\mathbf{X} \sim N_n(\mathbf{0}, \mathbf{Q}^{-1})$, then $\mathbf{Z} = \mathbf{X} + \mu \sim N_n(\mu, \mathbf{Q}^{-1})$. Algorithm 4.3.2 due to Rue (1999) summarizes the steps needed to simulate a Gaussian MRF with known mean vector μ and precision matrix $\mathbf{Q}$.

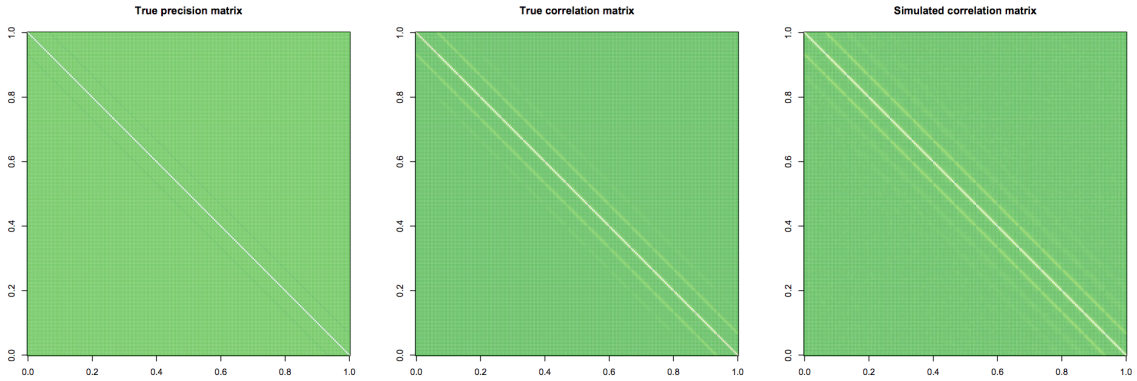
Algorithm 4.3.2. *Simulating Gaussian MRF* $\mathbf{Z} \sim N_n(\mu, \mathbf{Q}^{-1})$

- **Step 1.** *Find the Cholesky decomposition* $\mathbf{Q} = \mathbf{L}\mathbf{L}^T$*;*
- **Step 2.** *Generate* $\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{I})$*;*
- **Step 3.** *Solve* $\mathbf{L}^T\mathbf{X} = \mathbf{Y}$*;*
- **Step 4.** *Return* $\mathbf{Z} = \mathbf{X} + \mu$*.*

Figure 4.4 shows the theoretical precision matrix (a), correlation matrix (b), and the empirical correlation matrix (c) based on 10,000 simulations. The simulation is done on a 15×15 lattice so each of the three square matrices contains 225×225 elements. Darker shades denotes zero entries whereas lighter denotes non-zeros. One

can see the sparseness in all three matrices outside of their diagonal bands.

Figure 4.4: Simulated Gaussian MRF explanatory variable



4.3.2 Correlated Bernoulli responses

It is difficult to simulate dependent discrete random variables directly from their probability mass functions, partly because of the lack of knowledge on how to meet the restrictions on the limits of the dependency measurements (Madsen and Birkes, 2011). Instead, the simulation of correlated Bernoulli response vector $\mathbf{Y}$ with variance-covariance matrix Σ_Y is based on the Gaussian MRF $\mathbf{X}$ with the same length and variance-covariance terms. A copula approach was proposed in Madsen and Birkes (2011) to generate dependent discrete variables from the Gaussian MRF using either their Pearson or Spearman correlations.

We have discussed how to create a Gaussian random vector, $\mathbf{X}$, which has pairwise correlations governed by its precision matrix $\mathbf{Q}$. Now we relate $\mathbf{X}$ to an arbitrary distribution function, F, with respect to the same Markov random field. Take each element $x_i \in \mathbf{X}$ and transform it so that $U_i = \Phi(x_i)$, where Φ is the univariate standard

normal distribution function. Since Φ is a CDF, U_i's are Uniform(0, 1). Define $y_i \equiv F_i^{-1}(U_i)$, where

$$F_i^{-1}(u) = \inf\{y : F_i(y) > u\} \tag{4.43}$$

and the function produces vector $\mathbf{Y} = \{y_1, \ldots, y_n\}$ with desired, arbitrary distribution functions $F_1, \ldots, F_n$ under the transformations

$$y_i = F_i^{-1}[\Phi_i(x_i)]. \tag{4.44}$$

When F_i's are binomial CDFs, this is equivalent to transforming the Gaussian random variable x_i according to some threshold p_i:

$$y_i = \begin{cases} 1 & \text{when } x_i \geqslant p_i, \\ 0 & \text{when } x_i < p_i, \end{cases} \tag{4.45}$$

But more importantly, we have

$$\text{Corr}\ (x_i, x_j) = \text{Corr}\ (U_i, U_j) = \text{Corr}\ (y_i, y_j), \quad \text{and} \tag{4.46}$$

$$\Sigma_{\mathbf{X}} = \Sigma_{\mathbf{Y}}, \tag{4.47}$$

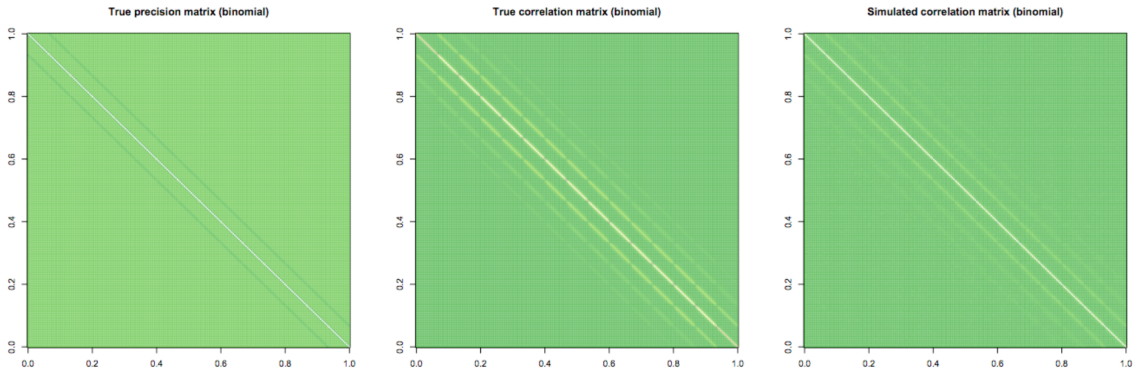
for both continuous and discrete variables. From (4.44) we have $x_i = \Phi_i^{-1}[F_i(y_i)]$, and by Sklar's theorem the joint distribution function of $\mathbf{Y}$, $C_{\mathbf{Y}} = C(y_1, \ldots y_n)$, can be expressed by a multivariate Gaussian n-copula Φ_Σ with respect to $\mathbf{X}$ and correlation

matrix $\mathbf{\Sigma} = \mathbf{Q}^{-1}$, given marginal functions $F_1, \ldots, F_n$.

$$C_{\mathbf{Y}} = \Phi_{\mathbf{\Sigma}}\{x_1, \ldots, x_n\} = \Phi_{\mathbf{\Sigma}}\{\Phi_1^{-1}[F_1(y_1)], \ldots, \Phi_n^{-1}[F_n(y_n)]\}. \tag{4.48}$$

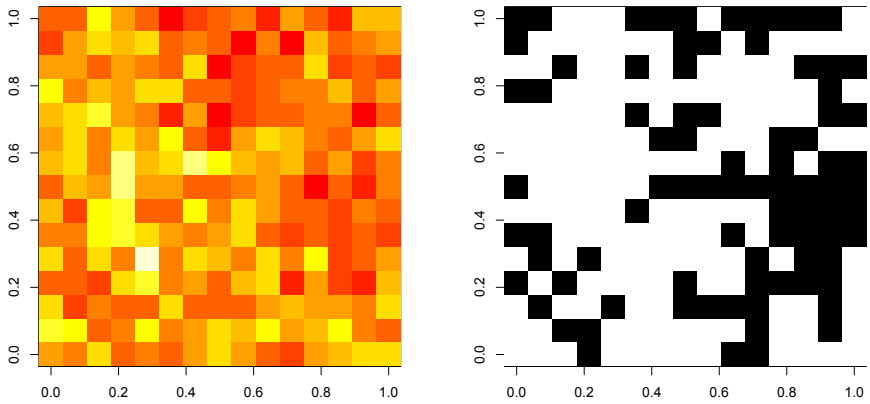
Figure 4.5 shows the theoretical precision matrix (left), correlation matrix (center), and empirical correlation matrix (right) for Bernoulli responses under second order nearest neighbor scheme. The dependency is determined by Gaussian correlation functions. The only unknown parameter needed to be estimated from the correlation is the effective range, r_y. Figure 4.6 is an example of both simulated Gaussian MRF $\mathbf{X}$ (left) and Bernoulli MRF $\mathbf{Y}$ (right) with means of $\mathbf{Y}$ determined by the logit link of $\mathbf{X}$. Both plots exhibit similar spatial dependency patterns.

Figure 4.5: Simulated correlated Bernoulli response



From Figure 4.4, 4.5, and 4.6 we can see that the supposedly non-zero entries of the simulated correlation matrices are well identifiable from zero entries, and the matrices exhibit the dependence structures one would expect from the second order regular lattices, with two non-zero bands on each side of the diagonal line.

Figure 4.6: Simulated explanatory variable and Bernoulli response



The analysis on the second order nearest neighbor lattice follows in Section 4.5, with the implementation of MCMC algorithms based on Equation (4.33) to estimate the unknown parameter set $\Theta_{\mathcal{G}}$ of the lattice. Before doing so, we will introduce another example, a data set with irregular lattice spatial structure.

4.4 U.S. presidential election data: Oregon and Washington

There are ample situations where data are collected at finite numbers of sites over a spatial domain, but the sites are not separated by a constant distance. This type of spatial data may still be considered as lattice data. However, since there are no regular intervals between sites, these are usually referred to as irregular lattices (Schabenberger and Gotway, 2004). This type of lattice structure is very common for data collected over regions, administrative or census units, or any other type of data that is aggregated over an area to a single location. The irregular lattice is prominent in social and human

sciences. We take this chance to demonstrate that generalized tree networks can be applied to irregular lattices as well, when the inference tools are not so different from the first and second order regular lattices.

To analyze the spatial and multivariate relationships on the irregular lattices, we need to first define their neighborhood structures. This is slightly more complicated than on the regular lattices, since we cannot define the nearest neighbors of a site based on its relative position to the other sites. The only requirement for the distance between a pair of sites in an irregular lattices is that it is a positive rational number. Luckily for data collected over regions, there is a substitute for the nearest neighbor scheme, because neighboring regions are usually perceived as the regions sharing a same border line. Those regions are also known as *adjacent* (for example we can say that the U.S. is adjacent to Canada and Mexico).

When the irregular lattice is defined by regions or geographical units on a spatial domain, we define adjacent sites as those that share a border. As in Section 4.3, we assume that there are multiple random variables at each site, and these variables are spatially autocorrelated. The neighboring nodes, $V(s_{ij})$ and $V(s_{kl})$, are instances of the same random variable observed at adjacent sites s_{ij} and s_{kl}. We continue to assume that the graphical structure at every site s_{ij} to be the same.

To demonstrate how generalized tree networks can be used to model irregular lattices, we use data from the 2008 presidential election. We are interested is accounting for spatial correlation on the county level in the United States while establishing the relationship between election results and income levels. The two variables included in the example are popular vote outcome by county during the 2008 U.S. presidential election (data from *USA Today*), and the county level median household income (MHI,

from the U.S. Census Bureau) in 2008. For computational reasons we look at two states only - Oregon and Washington. The total sample size, which equals the numbers of counties in these two states, is 75.

The full lattice may be denoted as $\mathbf{V} = \{V(s_i), i = 1 \dots 75\}$, where each $V(s_i)$ represents a county. Notice that we index the counties by a single subscript because there are no longer rows and columns to identify. In each county the election result is taken as a Bernoulli variable, with 0 indicating that the county favored the Republican Party candidate (John McCain), and 1 that the county favored the Democrats' candidate (Barack Obama). The only demographic variable, county median household income, is recorded in dollar values and after transformation and centering may be considered Gaussian. In a way, we may consider the election outcome as the $\mathbf{Y}$ response variable and median household income as the $\mathbf{X}$ variable, so that at each county, or site, we observe the familiar isomorphic structure $X(s_i) \to Y(s_i), V(s_i) = \{X(s_i), Y(s_i)\}$. We specify the following definitions for the irregular lattice $\mathring{\mathcal{G}} = (\{V(s_i)\}, \mathbf{E})$.

$$X(s_i) \sim N(\mu_x, \sigma_x^2), \tag{4.49}$$

$$\text{logit } (\pi_i) = b_0 + b_1 X(s_i), \tag{4.50}$$

$$Y(s_i) \mid X(s_i) \sim \text{Bernoulli } (\pi_i). \tag{4.51}$$

$$\text{Corr}(X(s_i), X(s_j) \mid \mathbf{X}_{-ij}) = \begin{cases} 0, & \text{when } s_i, s_j \text{ are non-neighbors;} \\ \xi_{ij}, & \text{when } s_i, s_j \text{ are neighbors.} \end{cases} \tag{4.52}$$

$$\text{Corr}(Y(s_i), Y(s_j) \mid X(s_i), X(s_j), \mathbf{Y}_{-ij}) = \begin{cases} 0, & \text{when } s_i, s_j \text{ are non-neighbors;} \\ \rho_{ij}, & \text{when } s_i, s_j \text{ are neighbors; and} \end{cases} \tag{4.53}$$

$$\xi_{ij} = \exp\left[\frac{-3\delta_{ij}^2}{r_x^2}\right], \quad \rho_{ij} = \exp\left[\frac{-3\delta_{ij}^2}{r_y^2}\right].$$

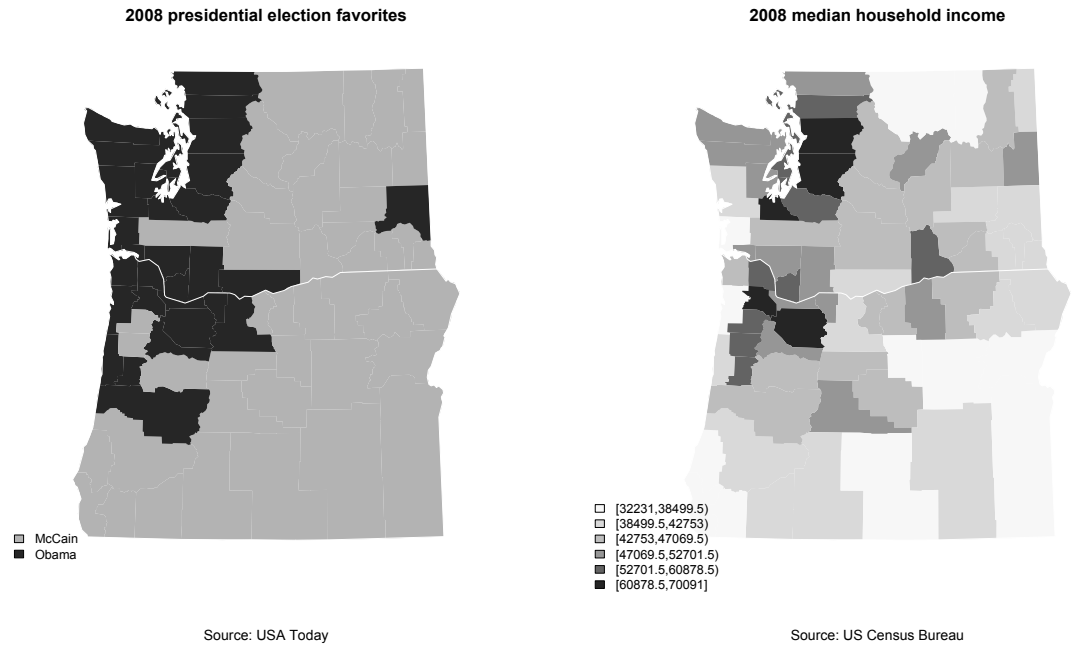
In these equations $\mathbf{X}_{-ij} = \{\mathbf{X}\backslash X(s_i), X(s_j)\}, \mathbf{Y}_{-ij} = \{\mathbf{Y}\backslash Y(s_i), Y(s_j)\}$. r_x, r_y are the range parameters governing the scale of the spatial autocorrelation, and d_{ij} is the Euclidean distance between s_i and s_j county seats. $\Theta_{\mathring{\mathcal{G}}} = \{r_x, r_y, \mu_x, \sigma_x^2, b_0, b_1\}$ is the parameter space on $\mathring{\mathcal{G}}$.

The sites' neighborhood structure, as defined by the shared borders between counties and Equations (4.52) and (4.53), needs to be determined from the map. Figure 4.7 shows side-by-side maps of the election result and MHI for each county in Oregon and Washington. On the left hand map lighter counties indicate where the Republican Party won, while darker indicates a Democratic victory. On the right hand map, darker shades means higher MHI levels, and lighter shades means lower MHI of the county.

From these maps we may summarize the data into a graph represented by Figure 4.8. Shades still denote the election result, while the size of the nodes indicates the magnitude of MHI. An undirected edge between two nodes means that these two county are adjacent geographically (with the exception of one island county, San Juan Island County, which is considered to be isolated and not adjacent to any other county). Each node is plotted at its county seat, so that the length of the edge suggests the Euclidean distance between two county seats.

The major difference between the regular and irregular lattice resides in how do we code their conditional Markov random fields. Because Figure 4.8 is not a junction tree, we need to select a set of nodes as the conditioning nodes, such that the rest of the

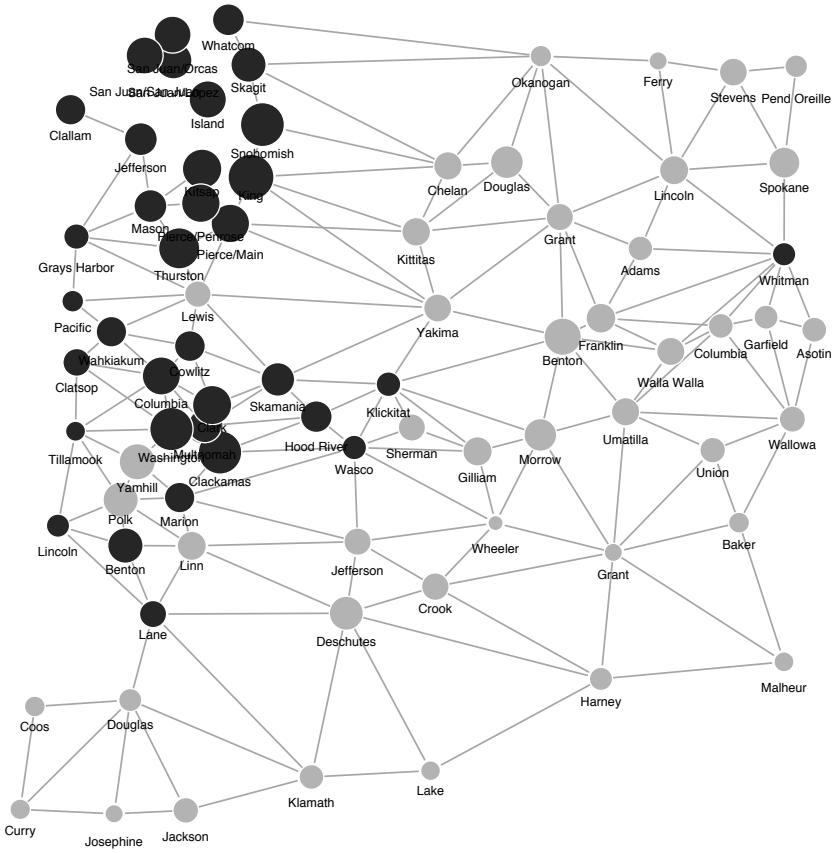
Figure 4.7: Maps of the election outcome (left) and MHI (right)



graph is a junction tree and partition-able after conditioning. In regular lattices, the coding can be done by selecting either every third rows or columns of nodes, which is a nice, systematic method. In irregular lattices, however, there are no pre-determined rows and columns, and we cannot use this method.

The coding scheme we propose strives to find a junction tree based on the undirected graph that contains as many maximal cliques as possible. We would like to have only one junction tree per run, which would preferably include as many nodes and edges as possible (*i.e.*, with the largest possible utility rate). We call it the *maximum tree* coding scheme. The maximum tree on an irregular lattice is not unique. Figure 4.9 illustrates three different maximum trees, each developed from a different "starting cliques", *i.e.*, a different corner of the graph. Not surprisingly, combined estimates from the maximal clique coding differs from each other. We will elaborate on

Figure 4.8: Graphical model of the election data

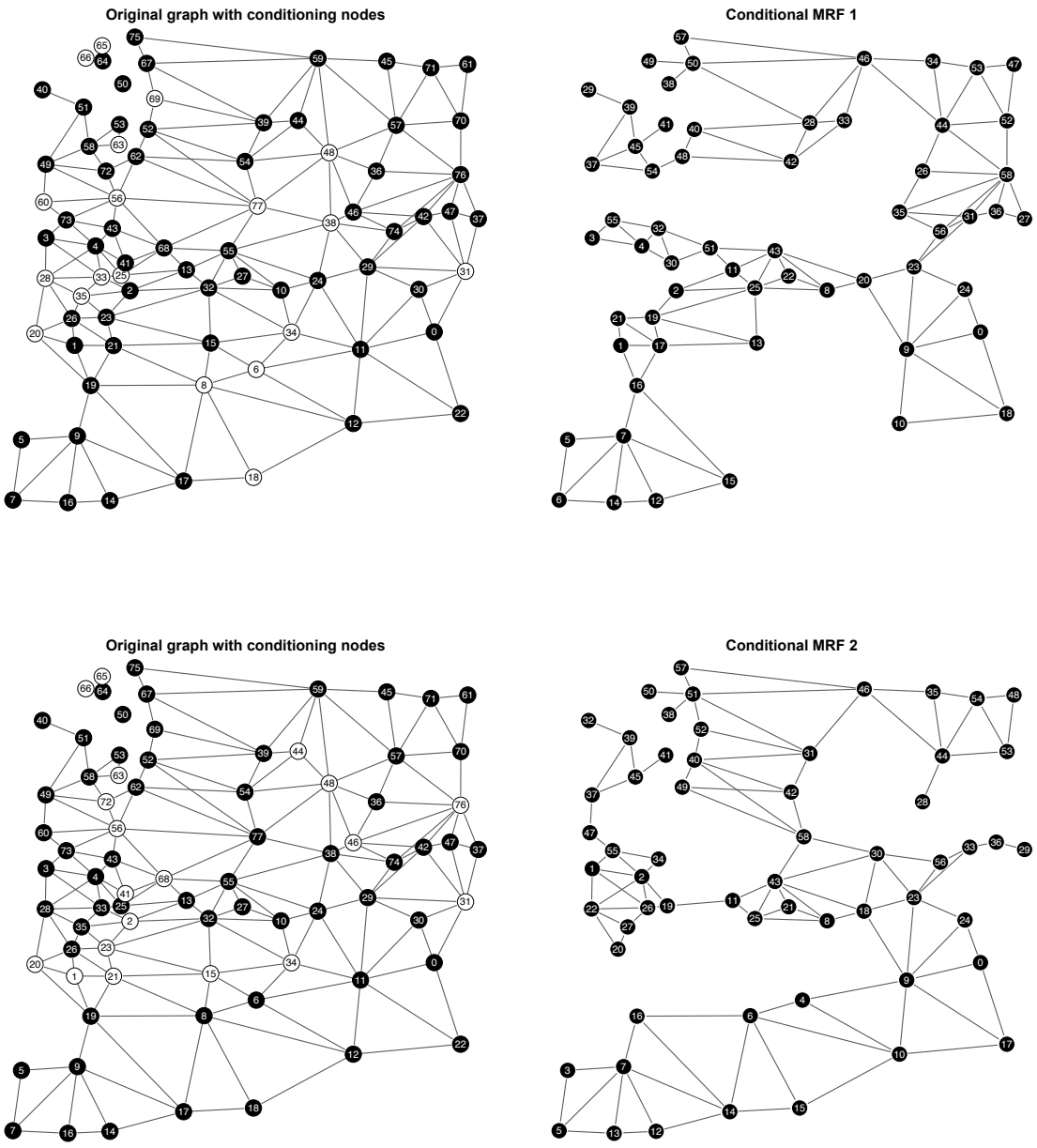


these findings and results in Section 4.6.

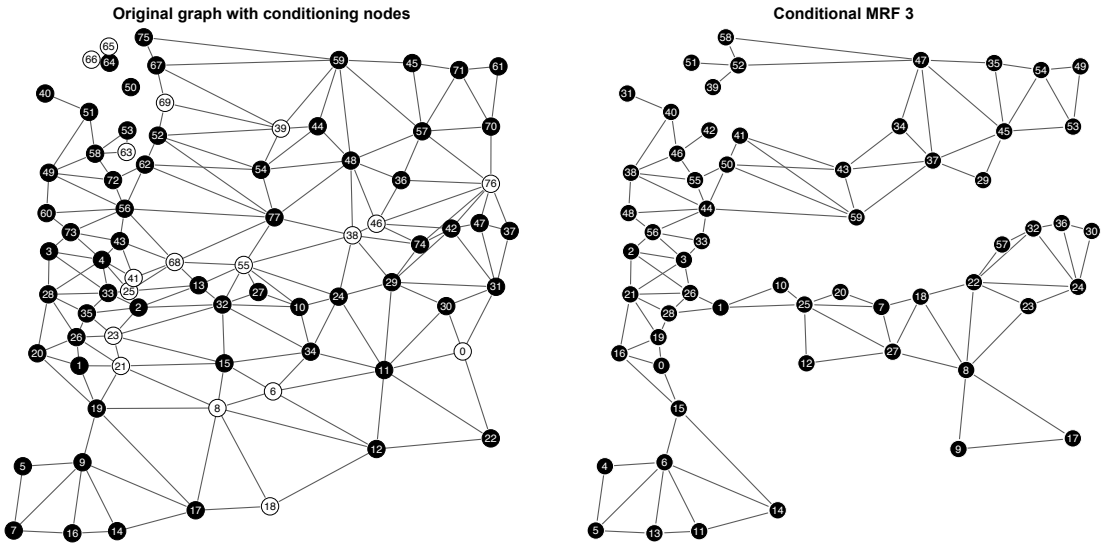
4.5 Markov Chain Monte Carlo simulation

Although MCMC algorithms (Gelman et al., 1995; Gilks et al., 1996; Andrieu et al., 2003; Robert and Casella, 2010) can be complicated and computationally demanding, in many cases they are easier to implement than numerical methods for maximum likelihood estimates. Bayesian inferences on high dimensional data, for example predictions from Equation (4.33), typically involve normalization, marginalization or

Figure 4.9: Maximum tree coding scheme



finding expectation based on intractable integrations that the MCMC algorithms are usually the most and only reliable processes we may count on. Recent advancement of the computer sciences has also lent great power to the MCMC methods for chain



convergence and parameter inference, which narrows the gap when comparing these simulation methods with classic estimation.

To make posterior inference using the distribution function (4.33), we specify appropriate priors and generate a series of random walks for each unknown parameter using the Metropolis-Hasting algorithm. This algorithm was first developed by physicists (Metropolis et al., 1953) for Boltzmann distributions, inspired by the calculation of a partition function not so different from our own $Z(\Theta_{\mathcal{G}})$. It was later generalized by Hastings (1970) to universal cases.

The objective for our application of Metropolis-Hasting algorithm to the ICG models is to obtain parameter estimates for all unknown parameters. For the $\mathbf{G}_{XY}$ ICG regular lattices with one predictor only, this refers to $\Theta_{\mathcal{G}} = \{\rho, \xi, \mu_x, \sigma_x^2, b_0, b_1\}$. For the irregular lattice of the presidential election data, this refers to $\Theta_{\hat{\mathcal{G}}} = \{r_x, r_y, \mu_x, \sigma_x^2, b_0, b_1\}$. Since r_y is one-to-one function of ρ in the regular lattice, as is r_x of ξ, these two parameter spaces are essentially equivalent.

The joint distribution for junction trees on the regular ICG which involves $\Theta_{\mathcal{G}}$ is given in Equation (4.33) in Section 4.2. Similarly, a joint distribution is found for each of the irregular lattice conditional MRF. With only one predictor and one response variable at each site, and the conditioning set of nodes denoted by $\mathbf{N}$, the separators on $\mathbf{Y}$ and $\mathbf{X}$ planes denoted by S_i^Y and S_i^X, its joint distribution may be written as

$$
\begin{aligned}
& p_{\mathring{\mathcal{G}}}(\mathbf{V}\backslash\mathbf{N} \mid \Theta_{\mathring{\mathcal{G}}}) \\
= \quad & \frac{1}{Z(\Theta_{\mathring{\mathcal{G}}})} \prod_{Cl_i^Y \in \mathbf{Y}\backslash\mathbf{N}} \phi_{Cl_i^Y}(\mathbf{Y} \mid \Theta_{\mathring{\mathcal{G}}}) \times \prod_{\mathbf{V}_i \in \mathbf{V}\backslash\mathbf{N}} \phi_{\mathbf{V}_i}(\mathbf{V} \mid \Theta_{\mathring{\mathcal{G}}}) \times \prod_{Cl_i^X \in \mathbf{X}\backslash\mathbf{N}} \phi_{Cl_i^X}(\mathbf{X} \mid \Theta_{\mathring{\mathcal{G}}}) \\
= \quad & \prod_{Cl_i^Y \in \mathbf{Y}\backslash\mathbf{N}} p(Cl_i^Y \mid \Theta_{\mathring{\mathcal{G}}}) \times \prod_{\{X(s_i),Y(s_i)\} \in \mathbf{V}\backslash\mathbf{N}} p[Y(s_i), X(s_i) \mid \Theta_{\mathring{\mathcal{G}}}] \times \\
& \prod_{Cl_i^X \in \mathbf{X}\backslash\mathbf{N}} p(Cl_i^X \mid \Theta_{\mathring{\mathcal{G}}}) \times \prod_{S_i \in \mathbf{V}\backslash\mathbf{N}} p(S_i^Y \mid \Theta_{\mathring{\mathcal{G}}})^{-d_{S_i}} p(S_i^X \mid \Theta_{\mathring{\mathcal{G}}})^{-d_{S_i}} \times \\
& \prod_{Y(s_i) \in \mathbf{Y}\backslash\mathbf{N}} p[Y(s_i) \mid \Theta_{\mathring{\mathcal{G}}}]^{-1}. \qquad (4.54)
\end{aligned}
$$

d_{S_i} is the number of maximal cliques joining at separator S_i. We select noninformative priors for most of the parameters to remain as "objective" as possible. Because regular and irregular lattice codings result in varying conditional MRF, we also set the priors differently in the two cases. For the range parameters, r_y, r_x, and mean μ_x we assume uniform distributions. The variance parameters σ_x^2 were set to have inverse-gamma distribution, and for the regression parameters, b_0 and b_1, priors were set as normal $N(0, 1000)$ for the regular lattices. On the irregular lattices, their priors are calibrated to center at the non-spatial estimates from logistic regressions (around $b_0 = -8.25, b_1 = 2$).

Using Equation (4.54) and the priors, we implement a Metropolis-Hastings algorithm to draw parameter samples. Algorithm 4.5.1 summarizes the detail of such

implementation, following notation of Gelman et al. (1995).

Algorithm 4.5.1. *Metropolis-Hastings algorithm*

- **Step 1.** *Initialize a parameter, θ, from the parameter set $\Theta_{\mathcal{G}}$ or $\Theta_{\mathring{\mathcal{G}}}$, at θ_0. The initial value has to satisfy that $p(\theta_0|\mathbf{V}) > 0$ and the first jump $J_1(\theta|\theta_0)$ is well defined. Set $t = 0$;*

- **Step 2.** *Set $t = t+1$. Draw a new candidate iteration θ_t^* from jumping distribution $J_t(\theta_t^*|\theta_{t-1})$, which is allowed to be asymmetric.*

- **Step 3.** *Calculate*

$$\lambda_t = \frac{p(\theta_t^* \mid \mathbf{V})}{p(\theta_{t-1} \mid \mathbf{V})} \cdot \frac{J_t(\theta_{t-1} \mid \theta_t^*)}{J_t(\theta_t^* \mid \theta_{t-1})} \tag{4.55}$$

- **Step 4.** *Assign the value for the next iteration to be:*

$$\theta_t = \begin{cases} \theta_t^*, & \textit{with probability } \min(1, \lambda_t), \\ \theta_{t-1}, & \textit{with probability } 1 - \min(1, \lambda_t); \end{cases}$$

- **Step 5.** *Go to* Step 2, *unless $t = N$;*

- *EXIT.*

To estimate all six parameters in the posterior distribution would require simultaneous updates of all these parameters. It means for each iteration of the chain we would have to update the whole equation six times and calculate the ratios λ_t six times, once for each parameter being estimated. This is resource-demanding and time consuming, especially considering the large numbers of discrete Gaussian copulas needed to be

calculated at each update.

To speed up the calculation, Gelman et al. (1995) recommended a sectional approach to the algorithm. They noticed the fact that a good portion of Equation (4.54) remains the same between $p(\theta_t^*|\mathbf{V})$ and $p(\theta_{t-1}|\mathbf{V})$ when there is only one parameter updated. For each parameter θ, because the jumping distribution $J_t(\theta_t|\theta_{t-1})$ depends only on θ but not the other parameters nor the data, and the accepting rule λ_t is a ratio of conditional distribution of θ given data, all the factorial components of the target distribution which does not involve θ will be canceled out in the calculation of r_t. In other words, when we iterate on parameter θ in the Metropolis-Hastings algorithm, we may isolate the factorial components from the target distribution which involve θ and update them only, while holding all other parameters fixed. In this way, we do not have to worry about the remaining, invariant component of the distribution with respect to θ.

This approach fits well with our example of the ICG models. By the nature of the junction tree, its joint distribution $p_{\mathcal{G}}$ is written as the product of the cliques' marginal distributions based on the planes of random variables, and each plane's distribution is governed by a subset of parameters from $\Theta_{\hat{\mathcal{G}}} = \{r_x, r_y, \mu_x, \sigma_x^2, b_0, b_1\}$ only. Borrowing this idea from Gelman et al. (1995), we can consolidate each parameter in Equation (4.54) into the components it involves only. This will greatly reduce the amount of calculation needed, and therefore, cut time for the simulation.

Table 4.1 lists the components in the posterior distribution that need to be updated for each parameter. Take r_y for instance: assuming it was sampled at step $t-1$ as $r_{y,t-1}$,

Table 4.1: Updated components for Metropolis-Hasting algorithm

Estimand	Updated components in the posterior function
r_y	$p(Cl_i^Y \mid r_y, b_0, b_1), \quad p(S_i^Y \mid r_y, b_0, b_1), \quad p(r_y)$
r_x	$p(Cl_i^X \mid r_x, \mu_x, \sigma_x^2), \quad p(S_i^X \mid r_x, \mu_x, \sigma_x^2), \quad p(r_x)$
μ_x, σ_x^2	$p[Y(s_i), X(s_i) \mid \mu_x, \sigma_x^2, b_0, b_1], \quad p(Cl_i^X \mid r_x, \mu_x, \sigma_x^2), \quad p(S_i^X \mid r_x, \mu_x, \sigma_x^2)$ $p(\mu_x)$ or $p(\sigma_x^2)$
b_0, b_1	$p(Cl_i^Y \mid r_y, b_0, b_1), \quad p[Y(s_i) \mid b_0, b_1], \quad p[Y(s_i), X(s_i) \mid \mu_x, \sigma_x^2, b_0, b_1],$ $p(S_i^Y \mid r_y, b_0, b_1), \quad p(b_0)$ or $p(b_1)$

to find its value at step t of the Metropolis-Hasting algorithm we only need to calculate

$$\lambda_{r_y,t} = \frac{\prod p(Cl_i^Y \mid r^*_{y,t}, b_{0,t-1}, b_{1,t-1}) \prod p(S_i^Y \mid r^*_{y,t}, b_{0,t-1}, b_{1,t-1})}{\prod p(Cl_i^Y \mid r_{y,t-1}, b_{0,t-1}, b_{1,t-1}) \prod p(S_i^Y \mid r_{y,t-1}, b_{0,t-1}, b_{1,t-1})} \times \frac{\prod p(r^*_{y,t}) J_t(r_{y,t-1} \mid r^*_{y,t})}{\prod p(r_{y,t-1}) J_t(r^*_{y,t} \mid r_{y,t-1})},$$

$$\text{instead of} \quad \lambda_{r_y,t} = \frac{p(r^*_{y,t} \mid \mathbf{V}, r_{x,t-1}, \mu_{x,t-1}, \sigma^2_{x,t-1}, b_{0,t-1}, b_{1,t-1})}{p(r_{y,t-1} \mid \mathbf{V}, r_{x,t-1}, \mu_{x,t-1}, \sigma^2_{x,t-1}, b_{0,t-1}, b_{1,t-1})} \times \frac{J_t(r_{y,t-1} \mid r^*_{y,t})}{J_t(r^*_{y,t} \mid r_{y,t-1})}.$$

Apart from the choices of priors and acceleration by the selected components, the proper utilization and efficiency of the Metropolis-Hastings algorithm also depends on the selection of the jumping distribution J. Extensive discussion was drawn on the practicality of suitable candidate jumping distributions (Gilks et al., 1996; Gelman et al., 1996; Tierney and Mira, 1999), and the debate is far from concluded.

There are a few common, well established and tested jumping distributions that many researchers use and we may loosely group them into five categories. Metropolis et al. (1953) and Müller (1991) have given in their respective papers the first category of jumping distribution $J_t(\theta_t|\theta_{t-1}) = q_1(\theta_t - \theta_{t-1})$, where $q_1(\cdot)$ is some multivariate density function. This category is known as the *random walk* family. The second category,

given by Hastings (1970), generates new states of the chain by $J_t(\theta_t|\theta_{t-1}) = q_2(\theta_t)$. Independent from the current state, this category is termed as the *independence chain* in Tierney (1994). The third category involves exploiting the target distribution $p(\theta|\mathbf{V})$ and trying to mimic it as closely as possible. If the form of the target distribution is known, we may use some other distributions with simpler form but similar shape as the jumping distribution. For another alternative, if we know that the target distribution can be written as $p(\theta|\mathbf{V}) \propto q_3(\theta|\mathbf{Y})h(\theta|\mathbf{V})$, where $q_3(\theta|\mathbf{V})$ is known and easier to sample from, whereas $h(\theta|\mathbf{V})$ is difficult to sample but bounded, then $J_t(\theta_t|\theta_{t-1}) = q_3(\theta_t|\mathbf{V})$ is usually an efficient jumping distribution choice as well. A fourth category, developed by Tierney (1994), uses the accept-rejection method with a pseudo dominating function to generate the iterations. And finally, the fifth category is the *autoregressive chain* and utilizes an $AR(1)$ process for the jumping distribution, $J_t(\theta_t|\theta_{t-1}) = q_5[(\theta_t - a) - b(\theta_{t-1} - a)]$.

In the simulated data set, since we know the true value for each parameter and the assumed distributions $p(\theta|\mathbf{V})$, we will use the third category of jumping distributions to choose the best candidates resembling them. In real world data sets we may practice similar techniques by plotting and imitating the distribution curves for each parameter. By generating the Markov chains from some J's similar to these true distributions, these choices will hopefully provide efficient sampling and fast convergence time. For the irregular lattice of the presidential election data, we validate the third category choice by first plotting them and then mimic the curves with the optimal candidate distributions. The choices of the parameters' jumping distributions are summarized in Table 4.2.

We discuss the MCMC simulation results for both studies in Section 4.6.

Table 4.2: Jumping distributions for the ICG Metropolis-Hasting chains

Simulated study		Presidential election data	
Estimand	Jumping distribution	Estimand	Jumping distribution
r_y	Uniform	r_y	Uniform
r_x	Uniform	r_x	Gaussian (bounded)
μ_x	Gaussian	μ_x	Gaussian
σ_x^2	Gamma	σ_x^2	Gamma
b_0	Gaussian	b_0	Bivariate Gaussian (with b_1)
b_1	Gaussian	b_1	Bivariate Gaussian (with b_0)

4.6 Results

4.6.1 Simulation study results

Simulated data sets were generated on second order nearest neighbor regular lattices as described in Section 4.3. This task was carried out in `R` (R Core Team, 2012, http://cran.r-project.org), in particularly relying on the library `mtvnorm`. Each lattice is laid out as 15×15 sites, and at each site one Bernoulli variable and one Gaussian variable are assumed to have a directed relationship. We randomly simulated 50 different lattices, all based on an identical set of parameters. The true values are noted in Table 4.3. The values of the effective range parameters, r_y and r_x, were carefully chosen such that the spatial autocorrelation was strong enough to be detected between neighboring sites (for both first order and diagonal ones), but declined substantially for the non-neighbors.

Estimation of the corresponding generalized tree network using the Metropolis-Hasting algorithm is performed using `R`. For each of the 50 lattices three runs of

row belt coding are conducted. Each parameter Markov chain is updated 5,000 iterations, with the first 500 iterations considered as the burn-in period. We see from the trace plots that all chains achieved convergence fairly quickly. Partial autocorrelation function (PACF) curves of the Markov chains have reduced to close to zero after a reasonable lag. Posterior means and intervals were calculated for the last 4,500 iterations per chain. Each lattice's three posterior estimates of the same parameter were then combined to form a "lattice estimate," and together they were summarized again over all 50 lattices into Table 4.3, forming the "overall estimate." Collectively there were 150 junction trees evaluated for each parameter, with each tree equivalent to four or five 15×2 belts on $\mathbf{Y}$ and $\mathbf{X}$ planes, depending on where the belt started.

For the spatial dependence measurement, we use the effective ranges, rather than the correlation coefficients ρ and ξ, as the spatial parameters so that these results correspond to similar results reported in Irvine (2007) for multivariate Gaussian lattices. Table 4.3 lists the overall posterior means, standard deviations, posterior intervals, and for the six parameters comprising $\Theta_{\mathcal{G}}$.

All six posterior means are close to the true values used to generate the data. In other words, they are theoretical values of the parameters, rather than the empirical values based on the 50 realized simulated lattices. Because of the uncertainty during simulation, a deviation between the empirical and theoretical true values of the parameters is expected. The estimate of r_x has a smaller posterior standard deviation than that of r_y. This is expected since r_x only involves calculations on $\mathcal{G}_{\mathbf{X}}$, while the calculation for r_y depends on the whole graph, $\mathcal{G}$. Overestimation of the effective range parameters means an underestimation of the autocorrelation. We suspect this may be due to the relatively small sample size not fully capturing the spatial structure. All

posterior intervals include the true values of their respective parameters, except for r_y, whose lower bound sits almost exactly on the true value.

Table 4.3: Metropolis-Hasting estimates based on second order ICG GTN

Estimand	True	Mean	SD	2.5%	Median	97.5%
r_y	1.30	1.581	0.153	1.301	1.576	1.924
r_x	1.35	1.404	0.084	1.262	1.398	1.573
μ_x	0.50	0.435	0.191	0.065	0.429	0.782
σ_x^2	2.00	2.019	0.252	1.554	2.013	2.513
b_0	0.00	-0.063	0.367	-0.864	-0.045	0.568
b_1	2.00	1.794	0.326	1.128	1.788	2.423

The generalized tree network method gives us the flexibility to specify a factorization model when some of the nodes are non-Gaussian. The other types of spatial models either lack the ability to explain simultaneous spatial dependence or cannot be applied to discrete random variables. For instance, the separable model and linear model of coregionalization both require that both $\mathbf{X}$ and $\mathbf{Y}$ are Gaussian. The simultaneous and conditional autoregressive models, on the other hand, allow only spatially-correlated error terms in $\mathbf{Y}$, but not in $\mathbf{X}$, and as a consequence they are limited to a graph with only the $\mathbf{Y}$ plane with no $\mathbf{X}$ dependence and also, without r_x, σ_x^2 parameters.

We now compare our GTN results with the spatial regression models that consider no spatial effect, univariate spatial dependent and bivariate dependency. For the simplest, "reference" model we use a non-spatial logistic regression model. It can be considered as a model with both $\mathbf{Y}$ and $\mathbf{X}$ spatial correlations completely removed, with only the $X(s_{ij}) \rightarrow Y(s_{ij})$ effect being estimated from the model. The conditional autoregressive (CAR) model allows a spatially correlated $\mathbf{Y}$ with second order nearest neighbors, but assumes $\mathbf{X}$ to be fixed. Structurally it resembles more with a $\mathbf{G}_Y$ ICG

rather than a $\mathbf{G}_{XY}$. Following the notations in Chapter 2, this model may be specified by

$$\text{logit } (\mathbf{\Pi}) \sim \text{MVN}\left[\mathbf{X}'\beta, (\mathbf{I}_n - \psi\mathbf{W_Y})^{-1}\mathbf{D_Y}\right], \quad \mathbf{\Pi} = \{\pi_{11}, \ldots, \pi_{nm}\}. \tag{4.56}$$

We also fit the data with a generalized linear mixed model (Schabenberger and Gotway, 2004, GLMM), where both $\mathbf{Y}$ and $\mathbf{X}$ are consider to be stochastic and spatially dependent. The dependency structure of $\mathbf{Y}$ is again evaluated by Equation (4.56), while the structure of $\mathbf{X}$ is modeled by Bayesian Gaussian kriging model (Diggle et al., 1998):

$$\mathbf{X} = \mu_{\mathbf{X}}\mathbf{1} + \mathbf{e}, \quad \mathbf{e} \sim \text{N}(\mathbf{0}, \mathbf{\Sigma_X}). \tag{4.57}$$

Both the CAR and GLMM models are realized using `GeoBUGS`, a package specialized in spatial models and maps in `WinBUGS` (Thomas et al., 2004). The univariate CAR model uses `car.normal` function in `GeoBUGS`, while the GLMM uses both `car.normal` and `spatial.exp` functions, specifying a variance-covariance matrix $\mathbf{\Sigma_X}$ for $\mathbf{X}$ with pairwise entries determined by the Euclidean distances between the county seats, a precision parameter τ, and a spatial range parameter ϕ (Thomas et al., 2004).

We obtain estimates on the same 50 lattices under these three models, and compare them with the GTN model. Summary of the comparison is presented in Table 4.4.

All mean estimates from GTN, GLMM, CAR and logistic models are close to the true value, and agree among each other. The major differences between the models are at the posterior standard deviation values and the corresponding widths of the posterior intervals. By looking at the first three lines of the table we can see that the

Table 4.4: GTN, GLMM, CAR, and logistic regression estimates of b_0 and b_1

Estimand, true value	Model	Mean	SD	2.5%	Median	97.5%
b_0=0	**GTN**	-0.063	0.367	-0.864	-0.045	0.568
	CAR+spatial.exp	-0.052	0.321	-0.705	-0.023	0.493
	CAR only	-0.057	0.321	-0.703	-0.047	0.504
	Logistic	-0.051	0.314	-0.682	-0.039	0.487
b_1=2	**GTN**	1.794	0.326	1.128	1.788	2.423
	CAR+spatial.exp	1.808	0.247	1.320	1.812	2.197
	CAR only	1.794	0.256	1.297	1.800	2.182
	Logistic	1.756	0.248	1.271	1.765	2.136

GTN model has the largest standard deviation and widest posterior interval, while the logistic regression possesses the smallest standard deviation and narrowest interval; the GLMM and CAR models reside between the two. The same pattern occurs again for the b_1 estimates.

We believe that the cause of smaller standard deviations in CAR and logistic regression models is that they have likely falsely exaggerated the effective sample size on the graph. The logistic regression model overlooks both variables' spatial autocorrelation structure, and considers all sites to be spatially independent from each other. Essentially, it assumes an effective sample size larger than what is supported by data and therefore has the smallest standard deviation and narrowest interval. The CAR spatial effect model also suffers from the same mistake by specifying only $\mathbf{Y}$ dependence while assuming fixed, spatially invariable $\mathbf{X}$ nodes.

The GLMM (CAR + spatial.exp) has provided very similar estimation results to the CAR model and logistic regressions. It did not improve the standard deviations with its consideration of additional $\mathbf{X}$ spatial structure. Since both the $\mathbf{Y}$ and $\mathbf{X}$ spatial terms are fitted before the logit link transformation, GLMM finds it difficult to distinguish the two types of spatial dependence. Cressie (1993, p. 127-129) had also

noted the difficulty and importance on identifying the formation of Bayesian kriging process components, which he referred to as "measurement error" and "microscale" variances. The GTN method, on the other hand, models the two spatial components individually, resulting in larger posterior standard deviations and wider posterior intervals. Among the four models, we believe that GTN is the one that provides the most reliable posterior estimates.

4.6.2 Election data results

Since there are no recurrent spatial patterns to form the maximal cliques in an irregular lattice, the junction trees that we built appear to be highly influenced by the coding scheme selected, and subsequently the selection of conditioning nodes has non-neglectable impacts on the posterior estimations of the trees.

For the election data, we use similar techniques as in the regular lattice, and the chains for each junction tree are updated for 10,000 iterations. We build junction trees from the generalized tree networks using the maximum tree coding schemes. Table 4.5 lists the result of three maximum trees and their combination on the election data. The maximum trees 1, 2, and 3 correspond to the conditional MRF plots 1-3 in Figure 4.9 respectively. For each maximum tree, the posterior mean, standard deviation, and posterior intervals were reported for each parameter of the lattice, and the pooled overall estimates and posterior intervals of the three trees are also calculated.

From Table 4.5 we can see that some of the estimates for the same parameter, such that those for b_0, b_1, and r_y, are quite different, while those for μ_x and σ_x^2 are fairly similar across runs. This indicates the difficulty for the chains when evaluating the

Table 4.5: Metropolis-Hasting estimates, maximum tree coding scheme

Estimand	Tree	Mean	SD	2.5%	Median	97.5%
r_y	#1	0.948	0.455	0.120	0.931	1.904
	#2	1.460	0.295	0.847	1.485	1.941
	#3	1.180	0.620	0.114	1.140	2.531
r_x	#1	1.060	0.140	0.778	1.061	1.339
	#2	1.214	0.176	0.881	1.211	1.569
	#3	1.475	0.213	1.031	1.480	1.881
μ_x	#1	4.641	0.088	4.470	4.643	4.815
	#2	4.570	0.091	4.390	4.571	4.745
	#3	4.671	0.091	4.494	4.669	4.847
σ_x^2	#1	0.720	0.099	0.554	0.714	0.936
	#2	0.704	0.106	0.526	0.694	0.938
	#3	0.639	0.100	0.465	0.629	0.859
b_0	#1	-7.416	1.596	-10.098	-7.661	-3.869
	#2	-7.796	1.418	-11.454	-7.712	-5.109
	#3	-9.909	2.545	-14.223	-9.870	-5.514
b_1	#1	1.530	0.339	0.777	1.583	2.108
	#2	1.631	0.308	1.053	1.614	2.410
	#3	2.098	0.544	1.149	2.091	3.032
Combined		Mean	SD	2.5%	Median	97.5%
r_y		1.196	0.520	0.169	1.213	2.195
r_x		1.250	0.248	0.838	1.217	1.786
μ_x		4.627	0.100	4.428	4.628	4.817
σ_x^2		0.688	0.108	0.502	0.679	0.919
b_0		-8.374	2.209	-13.551	-8.031	-4.585
b_1		1.753	0.480	0.923	1.677	2.873

whole irregular lattice and approximating the Gaussian copulas: the estimates that involve only $\mathbf{X}$ are the ones that are more consistent across the trees.

We also compare the GTN methods with the logistic regression model and the multivariate spatial model. Estimate results of the regression parameters from the GLMM (CAR + spatial.exp), logistic regression, and the GTN method are presented in Table 4.6.

The first thing we notice from Table 4.6 is that the posterior means for both b_0 and

Table 4.6: GTN, GLMM, and logistic regression estimates of b_0 and b_1

Estimand	Model	Mean	SD	2.5%	Median	97.5%
b_0	**GTN (max tree)**	-8.374	2.209	-13.551	-8.031	-4.585
	CAR + spatial.exp	-8.129	1.477	-11.143	-8.082	-5.229
	Logistic	-8.203	1.393	-10.961	-8.195	-5.437
b_1	**GTN (max tree)**	1.753	0.480	0.923	1.677	2.873
	CAR + spatial.exp	1.694	0.317	1.074	1.684	2.335
	Logistic	1.708	0.299	1.117	1.703	2.302

b_1 are similar between GTN model, the GLMM and the logistic regression. In terms of the standard deviations and posterior intervals, the GTN model with maximum tree coding tends to have the largest standard deviations and widest posterior interval, followed by GLMM, and the logistic regression.

While we are still not persuaded by the standard deviation from the logistic regression, it is also debatable how much confidence we should place on those estimates obtained from the Markov chains on the generalized tree networks. The expansion of standard deviations for the GTN parameters on irregular lattices appears to exceed the level which is validated by the extra correlation they allow. This is especially intriguing since both the GTN and the GLMM models consider both $\mathbf{Y}$ and $\mathbf{X}$ to be spatially dependent.

We suspect that a major, if not the only, reason for the underperformance of the GTN method and conditional MRF is, the junction tree depends too much on the choice of coding scheme, especially on the ways maximal cliques are selected or dropped from the final graph. The coding scheme, the starting clique, directions of the newly added cliques, and the final forms of the junction trees all contribute to the problem and complicated the interpretation of the results. This is especially questionable when the sites are distributed unevenly geographically, or when the variables on sites show

evidence of anisotropy and clustering. When the conditional Markov random field shows its insufficiency with respect to the original chain graph on an anisotropic, heteroscedastic irregular lattice, it is reflected to the the estimates that cause bias and loss of precision.

We believe that results from this section have demonstrated that the method we proposed works effectively on the regular lattice. On the second order nearest neighbor regular lattices, by first moralizing the types of chain graphs known as generalized tree networks into undirected graphs, applying necessary coding technique to obtain conditional Markov random fields and junction trees, and then conducting MCMC algorithm calculation to these conditional Markov random fields, we achieve satisfactory and improved Bayesian inference for both the spatial dependence and regression parameters. Though we might consider the GTNs on regular lattices with the advantages over the other spatial models, the irregular lattice draws a somewhat different picture. It is yet unclear if an effective method exists for the irregular lattices. While the maximum tree coding scheme has provided some promising results, they have obvious drawbacks. We might need to explore alternative ways of creating the conditional Markov random field, or even maybe other ways to partition the graph, so that the estimate in irregular lattices can 1) increase its precision, 2) have smaller or no bias, and 3) be straightforward to interpret.

4.7 Discussion

This chapter demonstrates how to estimate parameters of ICG with discrete and Gaussian variables on both regular and irregular lattices. In the first example, overlaying

Gaussian and Bernoulli Markov random fields are generated on a regular lattice, with each site having second order nearest neighbors. Gaussian variables are generated using multivariate normal function. For non-Gaussian distributed variables, we first simulated multivariate Gaussian variables with the same first and second moments, and then transformed them into the desired distribution using Cholesky decomposition. This approach has proven to be quite efficient (Rue, 1999).

Because the lattice has a second order nearest neighbor structure, we created conditional MRF from its moralized GTN using the belt coding scheme. When conditioning on every third row or column, the remaining nodes form conditional MRF which are also junction trees. Bayesian estimation was made from these junction trees using Metropolis-Hastings algorithm. Among all four models (GTN, GLMM, CAR only and logistic) that were considered on this lattice, the GTN estimates have the largest posterior standard deviations, and we believe it indicates that it follows the intrinsic spatial structure of the graph most closely.

On irregular lattices, we investigate the chain graph representing spatial and regressional associations between county median household income and presidential election results. The irregular lattice CG is considered isomorphic, and we created similar parametric model as appeared in the regular lattice. We attempted to use a similar belt coding scheme to create the conditional MRF on the irregular lattice, that is, by starting from a corner of the graph, and move towards one direction and select the maximal cliques along the way that do not create chordless cycles to be included in the belt. The estimate, however, is somewhat unsatisfactory. It shows big deviations from the other model estimates when compared with GLMM and logistic regression models. Overall, the belt coding suggests inconsistency among the estimates, and is

strongly influenced by the shapes and sizes of the coded junction trees, making it a sub-optimal approach.

We experienced another coding scheme, this time the maximum tree coding, in order to achieve a more consistent estimate. Although the influence of coding preference is still observable, generally speaking the estimates are much more invariant against the choice of the junction trees. A smaller sample size and irregular graph structure may have contributed to the large variation of the regression parameters. With a possible increase of sample size and a shift towards new coding methods, we might expect the GTN performance on irregular lattices to improve concretely.

Chapter 5

Discussion and Conclusion

The main focus of this thesis is multivariate spatial associations between continuous and discrete random variables. Variable dependence and autocorrelation are likely to be present in spatially collected data. It occurs not only in univariate and Gaussian cases, but also in multivariate and discrete data as well. No two spatial data sets are exactly the same. The great profusion of spatial structures exceeds the types of spatial models that have been proposed to appraise them, and researchers commonly find themselves struggled at pinpointing an available model that may adapt to a small variety of data sharing similar spatial structures.

Graphical models have seen advancement and popularity in recent year as a modeling tool assessing multivariate dependent systems in areas such as path analysis, image restoration, and machine learning. Comparably, it seems to have fewer applications in spatial analysis, besides on simple graphical structures (Haas et al., 1994; Rue, 2000; Adriaenssens et al., 2004). One of the main reasons for graphical models' absence in spatial studies is the difficulty of translating dependence structures from the data to a spatial model with known joint distribution. Per definitions by the neighborhood structure, each site on the domain usually has many neighbors. The result is a node-and-edge diagram with high connectivity. For more complicated cases such as chain graphs, graphical models made it simple to visualize and write down the data dependency based on the diagram, but the following step which requires matching the dependency with an estimable probability distribution is not necessarily simple. Another difficulty occurs when we introduce discrete components into the graphs. Even when the global Markov property can be used explicitly, it is nevertheless unclear whether a probabilistic model may be directly associated with the discrete random variable conditional independences.

The main contribution of this thesis is that we use isomorphic chain graphs (ICG) to denote highly connected graphical structure, and introduce generalized tree network and junction trees as means to parameterize and estimate ICG with discrete and/or continuous nodes as Markov random fields. Chapter 2 of the thesis mainly deals with ICG. When there are both predictor and response random variables in the graphs, each one of the four ICG type graphs ($\mathbf{G}_{\varnothing}$, $\mathbf{G}_X$, $\mathbf{G}_Y$, and $\mathbf{G}_{XY}$) may be identified using either the LWF or AMG Markov property. Since the two properties induce different sets of conditional independence, the spatial regression models parameterized by both properties are different as well. For example, Irvine (2007) showed that using AMP Markov property, ICG $\mathbf{G}_{XY}$ under Gaussian distributions may be associated with separable and LMC models, while we have showed that under the same assumptions, the graph associates with the MCAR model using LWF Markov property. We suspect that an "equivalence" class might exist between the corresponding probabilistic models (*i.e.* LMC, separable, and MCAR models) since they all trace back to the same graph. A closer inspection and comparative study between these models may formalize the equivalence class, as they can provide insights on the similar underlying characteristics among the models and the bond between them.

ICG apply to a variety of spatial data, but for data with discrete components they do not always associate with a known model. Chapter 3 highlights our effort to distinguish joint distributions of discrete ICG using Gibbs distributions (Hammersley and Clifford, 1971). Since the Gibbs distribution is defined on Markov random fields (MRF), we need to convert the chain graph first. We select those chain graphs that are generalized tree networks, and moralize them into Markov random fields. The requirement of no chordless directed cycles in a GTN ensures that its moralized graph

does not have large chordless undirected cycles unfavorable to the graph separation. Using global Markov property graph separation rules, we obtain the partition of the Gibbs distribution on the moralized MRF using the Hammersley-Clifford theorem. Subsequently we employ a combination of exponential family MRF, junction tree, and copula model to account for multivariate spatial dependence in the presence of discrete nodes.

A noteworthy point for this chapter is the conditional independence reduction in the moralization process. In particular it might raise someone's question when two Markov equivalent CG (Lauritzen et al., 1990; Verma and Pearl, 1992) might moralize into two MRF with different conditional independence properties. In our defense, we believe that the conditional dependence comes as a trade-off for straightforward joint distribution partition made possible on the MRF. More importantly, the process made sure that the conditional independences specified in the moralized MRF are loyal to the immoral chain graph under the LWF Markov property. In essence, the moralization process preserves a subset of conditional independence of the chain graph, and by connecting parent nodes moralization reveals conditional dependences that have not been directly stated in chain graphs. The chain graph conditional dependences may alternatively be summarized by the Bayes ball rules (Shachter, 1998). A similar practice on CG also "augment" them to UG. Further investigations into these alternative options of the conversion process under both the AMP and LWF Markov property have the potential to provide more insight and penetration about the relationships between CG, UG and their partitioning functions.

The successful application of the Gibbs distribution in practice depends on immediate and efficient evaluations of its partition factors, *i.e.*, the maximal clique potentials.

In our approach the potential functions are model using copulas. The magnitude of spatial dependency is limited by the maximum correlation allowed by the copula. As demonstrated in Chapter 3, multivariate Gaussian copula is ideal for maximal cliques with size smaller or equals to four. Larger cliques, on the other hand, requires more careful designation and more complicated copulas. We did not study maximal cliques with size greater than four in the two examples, but one should not be surprised by a 5- or larger clique in spatial data. For discrete nodes, Madsen (2009) suggested adding jitters to the nodes in order to conduct continuous approximation of their joint distribution. It is also worthy exploring other types of copulas, such as the Plackett copula (Nelsen, 1999), or even other models for multivariate dependence, that is more suitable for large clique sizes and more discretely distributed nodes.

Chapter 4 describes the analyses of two lattice datasets, one of which is regular and the other one irregular. It also includes simulation methods for generating both Gaussian and discrete lattice Markov random fields, and harnessed Metropolis-Hastings algorithm attaining Bayesian inference from the data. For both regular and irregular lattices, we compare our GTN estimates to those obtained from a multivariate conditional autoregressive model (MCAR) and a logistic regression model which does not take into consideration the spatial effects. On the regular lattice it is revealed that the three models provide similar posterior mean estimates, while the GTN always have the largest standard deviations due to the extra spatial stochasticity it factors in. It suggests that while any of the three models is able to estimate the parameter posterior mean, only the GTN model provides a correct variation measurement by employing all spatial structures resides within the data. The high degrees of representativeness of the conditional MRF, combined with a complete permutation of all maximal cliques by

the three row or column conditioning runs supported the stability of their estimates. The lack of stability thereof might also explain the contrary trend observed in irregular lattices, when maximum tree coding leads to estimates of the GTN quite different from those of MCAR, co-kriging and logistic regression models. For maximum trees as well as belt codings, it is difficult to obtain representative junction trees from an irregular lattice because of its anisotropic variation and unorthodox clique tree structure. The fact that the coding methods are graph initiated, as they are influenced by the initializing clique and expansion direction of the junction tree, is not justifiable to the idea of model generality, which is one of the crucial reasons why we developed this modeling approach in the first place. We believe that a better devised coding method, or a more exhaustive combining process agreeing with the irregular structure of the lattice may improve the precision and consistency of the estimates.

The spatial effect is a natural circumstance occurring in many data sets and is difficult to comprehend. Graphical model is proved to be helpful identifying the spatial structures using variable conditional independences. The thesis shows that, besides using the AMP Markov property, the separable model, and the linear model of coregionalization, there are more estimable models we may connect to the ICG in Gaussian cases. Under the LWF Markov property and normality, $\mathbf{G}_{XY}$ can be connected with MCAR models, and we believe that there should be other spatial models suitable for the various types of chain graphs. For discrete cases, we have also found estimable models to which the graph may be associated. We show that directed effects on the graph may be converted undirectedly while partly preserving their conditional independences. We devise algorithms to convert the undirected graphs into junction trees, with a large subset of nodes and edges of the original graph factored into the estimation

function. Either fully or conditionally, the joint distribution of the undirected graph may be partitioned into an estimable function with respect to the maximal clique potential functions, with the partition function $Z(\Theta_{\mathcal{G}})$ explicitly written as the product of separator distributions. Using numerical methods, such as MCMC algorithms, this approach estimates the unknown parameters directly instead of approximately. We believe that this thesis improves our ability to address both issues of multivariate association and spatial dependence.

Appendix A

Notes on Hammersley-Clifford theorem

A.1 Well-defined conditional independence

There are a few notes worth mentioning for the Hammersley-Clifford theorem. The first one is one of the most desirable aspects of the Hammersley-Clifford theorem. By partitioning the Gibbs distribution based on potential functions on the graph, rather than distribution functions, the theorem allows next to no restriction whatsoever on the specification of the conditional independence structures across the cliques. The potential function can take on an arbitrary form of the researchers' choice, as long as it involves only the variables from its own clique, and can be normalized to 1 by the partition function Z.

Not only does the theorem allow great flexibility on defining the potential functions, it also ensures the existence of the conditional dependences upon which they

are based. This is a great advancement from the general Markov property. Although it is acknowledged that any conditional dependence structure induced by local Markov properties is easier to work with than the joint density, there is a hidden limitation that researchers need to keep in mind. The flexible allowance of the conditional independences does not always imply its existence. In other words, the Markov property will equate the global Markov property and the conditional independence property only when the true conditional independence property exists: It does not guarantee a set of arbitrary conditional independences to be well defined.

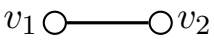
The main reasons for a conditional independence to be ill-defined include it has either taken an improper prior, or had degrees of freedom (in terms of numbers of estimating unknown parameters) from the conditional independence less than the degree of freedom from the joint distribution. In Bayes analyses it is common to specify an improper prior for the unknown probability, which does not integrate to 1, due to simplicity reasons or the lack of knowledge to the parameters. But in the cases of graphical models, we need to move with caution when specifying an improper prior since it might lead to non-existence conditional independences. An example is not at all difficult to construct.

Example A.1.1. *Non-existence conditional independence of a very simple MRF.*

Consider one of the easiest non-trivial Markov random fields, with only two nodes v_1 *and* v_2 *connected by an undirected edge. Let us suppose that there are one Bernoulli distributed variable at each node, namely,* V_1 *and* V_2*.*

When the undirected edge suggests that the joint density $p(V_1, V_2)$ *may be specified*

Figure A.1: A very simple MRF



by

$$p(V_1, V_2) = \begin{cases} \alpha_{11}, & \textit{when } V_1 = 1, V_2 = 1, \\ \alpha_{10}, & \textit{when } V_1 = 1, V_2 = 0, \\ \alpha_{01}, & \textit{when } V_1 = 0, V_2 = 1, \\ 1 - \alpha_{11} - \alpha_{10} - \alpha_{01}, & \textit{when } V_1 = 0, V_2 = 0. \end{cases} \tag{A.1}$$

When constructing a conditional distribution of V_1 and V_2, it is perfectly realistic to assume a form of V_1 given V_2 and V_2 given V_1 as

$$p(V_1 = 1|V_2) = \begin{cases} \beta_1, & \textit{when } V_2 = 1, \\ \beta_2, & \textit{when } V_2 = 0. \end{cases} \tag{A.2}$$

$$p(V_1 = 2|V_1) = \begin{cases} \beta_3, & \textit{when } V_1 = 1, \\ \beta_4, & \textit{when } V_1 = 0. \end{cases} \tag{A.3}$$

However, comparing Equation (A.1) with Equation (A.2) and (A.3), one can immediately see the dilemma: There are more unknown parameters in the conditional model than in the joint model. Hence, although the conditional distributions of Equation (A.2) and (A.3) are simple and intuitive enough, they actually contradict each other and cannot be combined to form the joint density (A.1) without additional assumptions.

The Hammersley-Clifford theorem prevents improperly defined conditional independences. The potential functions over the maximal cliques limit the numbers of unknown parameters to be estimated, and the rule of defining potentials following the clique structures rather than a freelance conditional independence ensures the propriety of the local Markov property.

A.2 The partition function of Hammersley-Clifford theorem

When the distribution of a Markov random field is defined through Hammersley-Clifford theorem, it is written in the form of its Gibbs distribution and is determined by the potential functions, ϕ, and the normalizing partition function $Z(\Theta_{\mathcal{G}})$. The potentials do not need to be probability density functions; rather, they only need to be proportional to some density function on the clique. This proportionality does not involve $Z(\Theta_{\mathcal{G}})$, the normalizing partition function. It means that without knowing the partition function, we can execute the inferences on the graph that do not require the calculation of full joint density $p_{\mathcal{G}}(\mathbf{V})$ (Bishop, 2006). Examples of these types of inferences include evaluating the local conditional, marginal distributions, or the most likely observed state of a graph, $\mathbf{V} = \mathbf{v}_{max}$, such that $p_{\mathcal{G}}(\mathbf{V} = \mathbf{v}_{max}) = max(p_{\mathcal{G}}(\mathbf{V}))$. In all these cases, we are more interested in the relative odds for a state of the graph compared to other states, rather than the absolute values or the parameter estimates of the graph.

The local conditional distribution of node, v_i, can always be written as the ratio of two joint distributions $p(v_i \cup bd(v_i))/p(bd(v_i))$, where $bd(v_i)$ is the boundary set of v_i, whereby Z is canceled out. Alternatively, we can work on the unnormalized

functions of v_i up to a proportion of its marginal distribution, and normalize it after the summation or integration process is taken over all the nodes.

The proportional equality (3.5) also makes it easy to compare different realizations of the same graph. Such problem is known as the the *most likely state* problem (Barber, 2003). It is the equivalence of finding maximum likelihood estimates of the nodes on a graphical model. To be more specific, for any two realizations $\mathbf{V} = \mathbf{v}_a$, $\mathbf{V} = \mathbf{v}_b$ of the graph $\mathcal{G} = G(\mathbf{V}, \mathbf{E})$, a most likely state problem concerns whether $p_{\mathcal{G}}(\mathbf{V} = \mathbf{v}_a) \geqslant p_{\mathcal{G}}(\mathbf{V} = \mathbf{v}_b)$ or $p_{\mathcal{G}}(\mathbf{V} = \mathbf{v}_a) \leqslant p_{\mathcal{G}}(\mathbf{V} = \mathbf{v}_b)$. It is also interested in finding the $\mathbf{v}_{max}$. From Equation (3.3) and (3.5) we can see that:

$$\frac{p_{\mathcal{G}}(\mathbf{V} = \mathbf{v}_a)}{p_{\mathcal{G}}(\mathbf{V} = \mathbf{v}_b)} = \frac{Z^{-1} \prod_{\mathcal{C}} \phi_C(\mathbf{V}_{aC}|\Theta_C)}{Z^{-1} \prod_{\mathcal{C}} \phi_C(\mathbf{V}_{bC}|\Theta_C)} = \frac{\prod_{\mathcal{C}} p_C(\mathbf{V}_{aC})}{\prod_{\mathcal{C}} p_C(\mathbf{V}_{bC})}. \tag{A.4}$$

And the ordering of the realizations from the "most likely state" to the "least likely state" can be simply identified by comparing the proportions of the potential products with 1 without knowing anything about Z. In practical studies this, algorithm is usually enhanced by evaluating $p^k_{\mathcal{G}}(\mathbf{V} = \mathbf{v}_a)/p^k_{\mathcal{G}}(\mathbf{V} = \mathbf{v}_b)$ with a large power k rather than Equation (A.4) to make the differences stand out even more.

Appendix B

Exponential family and Gaussian Markov random fields

The exponential family generalized tree networks are utilized by an alternative statement of the potential functions: the Hammersley-Clifford factorization (3.3) can usually be enhanced by the exponential specification of the potential functions. Whenever suitable, the potentials, ϕ_C, are chosen to be exponential family functions (they do not necessarily have to be exponential family densities). It was under this formulation that Besag (1974) presented the Hammersley-Clifford theorem. There are several benefits behind this statement. Firstly, because the Gibbs distribution is presented as the product of the potentials, the joint distribution can be modeled using the summation of log-potentials:

$$p_{\mathcal{G}}(\mathbf{V}) = \frac{1}{Z}\prod_{C\in\mathcal{C}}\phi_C(\mathbf{V}_C|\Theta_C) \tag{B.1}$$

$$\propto \exp\{\sum_{C\in\mathcal{C}}\log(\phi_C(\mathbf{V}_C)|\Theta_C)\} \tag{B.2}$$

The negative log-potential $e_C(\mathbf{V}_C) \equiv -\log(\phi_C(\mathbf{V}_C)|\Theta_C)$ is known as the *energy*. We have $-\log p_{\mathcal{G}}(\mathbf{V}) \propto -\sum_{C\in\mathcal{C}} e_C(\mathbf{V}_C)$. The second, and more important reason, to choose exponential family potentials is to avoid the summarization of the partition function Z. If $\prod_{\mathcal{C}} \phi_C(\mathbf{V}_C|\Theta_C)$ is a recognizable form, then there is no need to calculate Z anymore: this usually only happens when ϕ_C's are exponential family functions (Li, 2001).

One of the most widely used exponential family distribution is the so-called *Gaussian Markov random field* (Gaussian MRF), where the nodes of a MRF are marginally normal and pairwise correlated. Its popularity is not only due to the fact that Gaussian dependent data are common in various types of statistical studies, but also because it can serve as a "base" model on which other graphs can be built. More on this will be discussed in Chapter 5. For now, we will explore how to obtain its joint density.

To get the joint density without knowing Z, we need to be able to recognize the product of the potentials $\prod_{\mathcal{C}} \phi_C(\mathbf{V}_C)$ or the sum of the energies $\sum_{\mathcal{C}} e_C(\mathbf{V}_C)$. To achieve this, we may arbitrarily define the clique energy function in a partition form. In the Gaussian MRF case where $\mathcal{G} = G(\mathbf{V}, \epsilon), \mathbf{V} = \{v_1, \ldots, v_n\}$, Rue and Knorr-Held (2005) have shown that

$$\begin{aligned} e_C(\mathbf{V}_C) &= -\log(\phi_C(\mathbf{V}_C)) \\ &= \sum_{v_i\in C} \frac{n_{Ci}}{n_i}\xi_i(v_i) + \sum_{v_i - v_j \in \epsilon_C} \xi_{ij}(v_i, v_j) \end{aligned} \tag{B.3}$$

$$\text{where} \qquad \xi_i(v_i) = \frac{1}{2}Q_{ii}v_i^2 - \gamma_i v_i, \tag{B.4}$$

$$\xi_{ij}(v_i, v_j) = \begin{cases} \frac{1}{2}v_i Q_{ij} v_j & \text{when } v_i, v_j \text{ are neighbors,} \\ 0 & \text{otherwise.} \end{cases} \tag{B.5}$$

n_{Ci} is the numbers of neighbors of v_i within clique C, while n_i is the total numbers of neighbors of v_i in the graph. Q_{ij}'s and γ_i's associate with the unknown parameters to be estimated. Denote $\mathbf{Q} = \{Q_{ij}\}_{n\times n}$ and $\mathbf{\Gamma} = \{\gamma_1, \ldots, \gamma_n\}$. Rewriting the Hammersley-Clifford factorization using the energies defined by (B.3) through (B.5), we have

$$\begin{aligned}
\log p_{\mathcal{G}}(\mathbf{V}|\Theta_{\mathcal{G}}) &= \frac{1}{Z}\prod_{C\in\mathcal{C}}\phi_C(\mathbf{V}_C|\Theta_C) = -\log(Z) - \sum_{C\in\mathcal{C}} e_C(\mathbf{V}_C) \qquad \text{(B.6)}\\
&= -\log(Z) - \sum_{C\in\mathcal{C}}\sum_{v_i\in C}\frac{n_{Ci}}{n_i}\xi_i(v_i) - \sum_{v_i - v_j\in\epsilon}\xi_{ij}(v_i, v_j)\\
&= -\log(Z) + \sum_i[\gamma_i v_i(\sum_{C\in\mathcal{C}}\frac{n_{Ci}}{n_i})] - \frac{1}{2}\sum_i[Q_{ii}v_i^2(\sum_{C\in\mathcal{C}}\frac{n_{Ci}}{n_i})] - \frac{1}{2}\sum_{i\neq j} v_i Q_{ij} v_j\\
&= -\log(Z) + \sum_i \gamma_i v_i - \frac{1}{2}\sum_{i,j} v_i Q_{ij} v_j\\
&= -\log(Z) + \mathbf{\Gamma}^T\mathbf{V} - \frac{1}{2}\mathbf{V}^T\mathbf{Q}\mathbf{V}. \qquad \text{(B.7)}
\end{aligned}$$

We can immediately see that without the calculation of Z, this joint density is already explicit: this representation forms a multivariate Gaussian distribution $\mathbf{V} \sim \text{MVN}(\mu, \mathbf{\Sigma})$ with $\mu = \mathbf{Q^{-1}\Gamma}$ and $\mathbf{\Sigma} = \mathbf{Q^{-1}}$. Being the inverse of the variance-covariance matrix $\mathbf{\Sigma}$, $\mathbf{Q}$ is known as the precision matrix.

A very good result of the Gaussian MRF is the representativeness of its Markov properties by its precision matrix $\mathbf{Q}$ (Rue and Knorr-Held, 2005). To be precise, in a Gaussian MRF,

$$v_i \perp\!\!\!\perp v_j|\mathbf{V_{-ij}} \Longleftrightarrow Q_{ij} = 0. \qquad \text{(B.8)}$$

This can be shown using our constructed energies. By the local Markov property, the conditional independence $v_i \perp\!\!\!\perp v_j|\mathbf{V_{-ij}}$ implies that v_i, v_j are not neighbors. If v_i, v_j

are not neighbors, then by (B.5) we have $\xi_{ij}(v_i, v_j) = 0$, which means that $Q_{ij} = 0$. The proof from the other direction is analogous. Equation (B.8) means that there is an one-to-one correspondence between the non zero elements of $\mathbf{Q}$ and the pairs of neighbors, and it is very useful when it comes to fast, efficient simulations and estimations on a random field.

As good as it may get for the Gaussian MRF, our ability to come to an recognizable joint density similar to (B.7) is distribution-dependent. A different specification of the neighborhood structure, or a different marginal distributions, especially a non-exponential one, may render the energy sum unrecognizable. This method only applies to continuous MRF, but not discrete ones.

Appendix C

Modeling tree network and chain graphs

C.1 Tree network

A tree network is an undirected graph with simple structure. Since there are no loops in a tree network, all the maximal cliques in a tree network $\mathcal{T} = (\mathbf{V}, \mathbf{E})$ have size two. Although in discrete cases the clique factorization $\prod_C \phi_C(v_i, v_j)$ is generally not from an identifiable family, we would still be able to write down the joint distribution $p(\mathbf{V})$ explicitly based on its conditional independencies. In a tree network two maximal cliques C_1, C_2 are either disjoint or have one common node. For two adjacent cliques C_1 and C_2, let us assume that $C_1 = \{v_1, w\}, C_2 = \{v_2, w\}$, so that $C_1 \cap C_2 = w$. w is the separator between the cliques. Since v_1 and v_2 are separated by w, they are conditionally independent given w, or $v_1 \perp\!\!\!\perp v_2 \mid w$. This benefits us on the partitioning,

since Kirshner (2007) has pointed out that

$$\begin{aligned} f_{\mathcal{T}}(v_1, v_2, w) &= f(w)f(v_1, v_2|w) = f(w)f(v_1|w)f(v_2|w), \\ &= \frac{f(w)f(v_1|w)f(w)f(v_2|w)}{f(w)} = \frac{f(v_1, w)f(v_2, w)}{f(w)}. \end{aligned} \tag{C.1}$$

Immediately, we may recognize resemblances between Equation (C.1) and the Hammersley-Clifford Theorem. If the graph is only consisted of three nodes $\{v_1, v_2, w\}$, then (C.1) means the partition function of the Hammersley-Clifford Theorem is $Z(\Theta_{\mathcal{T}}) = f(w)$. This process of simplifying the joint distribution from the left side of Equation (C.1) to the right is known as the *node elimination process* (Barber, 2003), in the sense that the three nodes may be eliminated from the graph in place of a new hyper-node $\tilde{v} = \{v_1, v_2, w\}$ representing their joint density.

Applying the node elimination process through the main trunk of the tree network $\mathcal{T}$ from its lowest to highest hierarchy branches, the joint distribution of the whole tree can subsequently be partitioned into the product of bivariate distributions normalized by the product of separator nodes along the tree. Precisely for tree network $\mathcal{T}$ with maximal clique set $\mathcal{C}, C_i \in \mathcal{C}$, we may write that

$$f_{\mathcal{T}}(\mathbf{V}) = \frac{\prod_{v_i - v_j \in \mathbf{E}} f_{ij}(v_i, v_j)}{\prod_{w_{ij} = C_i \cap C_j} f(w_{ij})} \tag{C.2}$$

$$= \prod_{i=1}^{n} f_i(v_i) \prod_{v_i - v_j \in \mathbf{E}} \frac{f_{ij}(v_i, v_j)}{f_i(v_i) f_j(v_j)} \tag{C.3}$$

This partition rule applies here because the highest order of the multivariate distribution that needs to be calculated was reduced from n to 2. Remember that for copula models, Equation (3.22) partitions the multivariate joint density $h(\mathbf{U})$ as the product

of the marginal distributions $f_i(u_i)$ times the copula density $c(\mathbf{m})$. From (3.22) we have

$$c(\mathbf{m}) = \frac{h(\mathbf{U})}{\prod_{i=1}^{d} f_i(u_i)}, \quad \text{and we may write} \tag{C.4}$$

$$c_{ij}(v_i, v_j) = \frac{f_{ij}(v_i, v_j)}{f_i(v_i) f_j(v_j)}. \tag{C.5}$$

This reveals that the copula density is in fact the ratio between the joint distribution and the product of the marginal distributions. It is especially useful in discrete cases, when the joint distribution can be conveniently modeled by the already constructed copula densities $c_{ij}(v_i, v_j)$.

Figure C.1 illustrates an example of a tree network with nine nodes. In this example, $\mathcal{T} = (\mathbf{V}, \mathbf{E})$ where $\mathbf{V} = \{v_1, \ldots, v_9\}$ and $\mathbf{E} = \{v_1 - v_3, v_2 - v_3, v_3 - v_4, v_3 - v_5, v_4 - v_6, v_4 - v_7, v_5 - v_8, v_5 - v_9\}$. The maximal clique set $\mathcal{C}$ coincides with the edge set, hence $\mathcal{C} = \mathbf{E}$. The separator set consists of three nodes, $\mathcal{S} = \{v_3, v_4, v_5\}$ because they are the only nodes shared by more than one clique. These nodes are denoted by darker shades in the figure. We may arbitrarily select a node, say v_1, as the root of the tree, and $v_1 - v_3$ would subsequently be the highest hierarchy branch, followed by $v_3 - v_2$, $v_3 - v_4$, and so forth. Partition along this hierarchy, and the joint distribution of the tree network $f_{\mathcal{T}}(\mathbf{V})$ can be written as

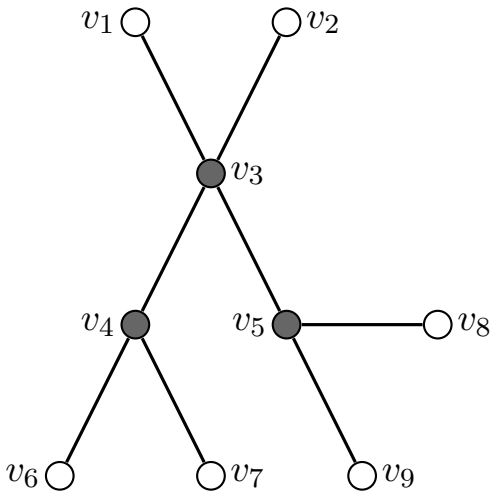
$$\begin{aligned} f_{\mathcal{T}}(\mathbf{V}) &= \prod_{i=1}^{9} f_i(v_i) \prod_{v_i - v_j \in \mathbf{E}} \frac{f_{ij}(v_i, v_j)}{f_i(v_i) f_j(v_j)}, \\ &= \frac{f(v_1, v_3) f(v_2, v_3) f(v_3, v_4) f(v_3, v_5) f(v_4, v_6) f(v_4, v_7) f(v_5, v_8) f(v_5, v_9)}{f(v_3)^3 f(v_4)^2 f(v_5)^2} \end{aligned} \tag{C.6}$$

The powers of the denominator marginals equal to the numbers of edges connected

to the joint nodes minus 1. In copula notations, Equation (C.6) can be refashioned as

$$f_{\mathcal{T}}(\mathbf{V}) \;=\; \prod_{i=1}^{9} f_i(v_i) \prod_{v_i - v_j \in \mathbf{E}} c_{ij}(v_i, v_j). \tag{C.7}$$

Figure C.1: A tree network with nine nodes

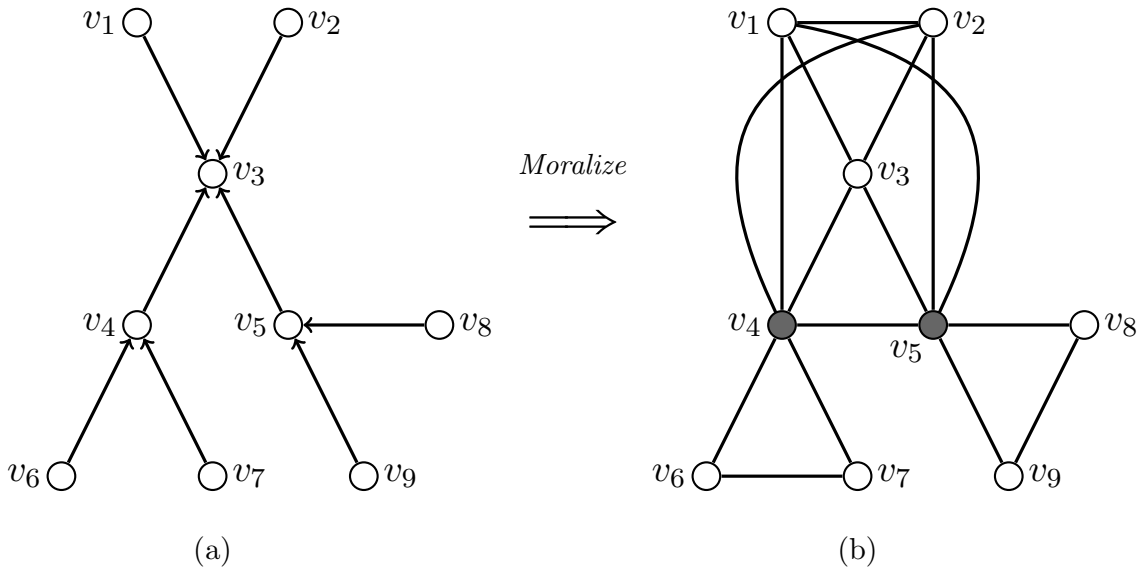


C.2 Bayes network

For a Bayes network, the final form of partition varies graph by graph, and is dependent upon their neighborhood structures. Let us look at an example similar to that of Figure C.1. The only difference between Figure C.1 and Figure C.2 (a) is, the undirected edges in the tree network are replaced by directed edges in the Bayes network. However, we cannot work with the Bayes network directly in terms of the partition. In order to do that it has to be moralized first. Only after the moralization could we identify the maximal cliques, separators, and the junction tree induced by the Bayes network. Figure C.2 (a) plots the Bayes network $\mathcal{B} = (\mathbf{V}, \mathbf{E})$ with nine nodes.

Recognizing that $v_1 \rightarrow v_3 \leftarrow v_2, v_1 \rightarrow v_3 \leftarrow v_4, v_2 \rightarrow v_3 \leftarrow v_5, v_4 \rightarrow v_3 \leftarrow v_5, v_1 \rightarrow v_3 \leftarrow v_5, v_2 \rightarrow v_3 \leftarrow v_4, v_6 \rightarrow v_4 \leftarrow v_7$ and $v_8 \rightarrow v_5 \leftarrow v_9$ are all immoralities, we are able to moralize it into the Marko random field illustrated in Figure C.2 (b).

Figure C.2: An example of Bayes network $\mathcal{B}$ (a) and its moralization (b)



Immediately we can see that the moralized network is more complicated than the tree network in Figure C.1, as it possesses bigger maximal cliques. The added undirected edges establish new connections between nodes, and only three maximal cliques remain in Figure C.2 (b): $C_{12345}, C_{467}, C_{589} \in \mathcal{C}$, with $C_{12345} = \{v_1, v_2, v_3, v_4, v_5\}$, $C_{467} = \{v_4, v_6, v_7\}$, and $C_{589} = \{v_5, v_8, v_9\}$. There are two separators in the moralized graph now: v_4 and v_5.

From Figure C.2 (b) we may see that $\{C_{12345}, C_{467}, C_{589}\}$ forms a clique tree ($C_{467} - C_{12345} - C_{589}$) and therefore, it is also a junction tree. Remember that based on the junction tree partition function Equation (3.11) in Chapter 3 and the copula density function, we would be able to write the joint density of the junction tree as the product

of the maximal clique densities divided by the separator product. For this particular example, we can write

$$f_{\mathcal{B}}(\mathbf{V}) = \frac{f_{C_{12345}}(v_1, v_2, v_3, v_4, v_5) f_{C_{467}}(v_4, v_6, v_7) f_{C_{589}}(v_5, v_8, v_9)}{f(v_4) f(v_5)}, \quad \text{(C.8)}$$

$$= \prod_{i=1}^{9} f_i(v_i) c_{12345}(v_1, v_2, v_3, v_4, v_5) c_{467}(v_4, v_6, v_7) c_{589}(v_5, v_8, v_9), \quad \text{(C.9)}$$

where c_{12345}, c_{467}, and c_{589} are the copula densities on cliques C_{12345}, C_{467}, and C_{589}, respectively. Moreover, this partition applies to more than just the Gaussian cases. Assuming variables at each site are discrete, we can exchange the densities $f(\cdot)$'s at the nodes for the probability mass functions $p(\cdot)$'s. The partition (C.8) holds again and we still have

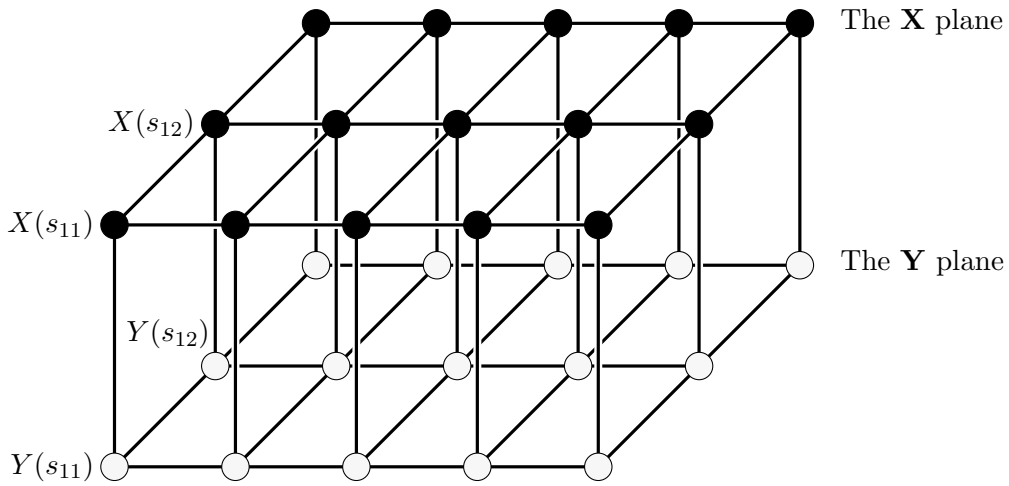
$$p_{\mathcal{B}}(\mathbf{V}) = \frac{p_{C_{12345}}(v_1, v_2, v_3, v_4, v_5) p_{C_{467}}(v_4, v_6, v_7) p_{C_{589}}(v_5, v_8, v_9)}{p(v_4) p(v_5)}.$$

For discrete cases, the calculation of the partition for the Bayes network is more demanding, because after moralization the undirected graph contains maximal cliques up to size of five and the orders of the multivariate copulas increase accordingly as well. The calculation for a 5-clique discrete copula is pushing the limit for the approximate copula density function because it involves an inclusion-exclusion algorithm over $2^5 = 32$ Gaussian copula terms, challenging the upper cap for the size of maximal cliques. We have discovered that, even for 4-cliques, the numerical Gaussian multivariate copula calculation is already not very stable in both `R` and `Matlab` packages. It would certainly raise a flag for 5-clique calculations. In those cases it might be helpful to explore alternative methods, such as the Plakett copula (Nelsen, 1999), to model the clique mass functions on higher order cliques.

C.3 ICG on first order nearest neighbor lattices

Regular lattices are typically easier to understand and to analyze than the irregular ones. For regular lattices two of the most commonly specified structures are the first and second order nearest neighbor schemes, first introduced by Besag (1972, there are higher order schemes that have been proposed as well, but they are uncommon in use). For the first order schemes, each non-bordering site has four neighboring or adjacent sites. If we consider the case where there is two variables, *i.e.* two nodes, $X(s_{ij})$ and $Y(s_{ij})$, per site s_{ij}, and the isomorphic property of the graph ($X(s_{ij}) - Y(s_{ij})$ for every site), then allowing both $\mathbf{X}$ and $\mathbf{Y}$ to be spatially dependent, an ICG $\mathbf{G}_{XY}$ data can be represented by a graph similar to Figure C.3.

Figure C.3: First order nearest neighbor regular lattice with two variables per site



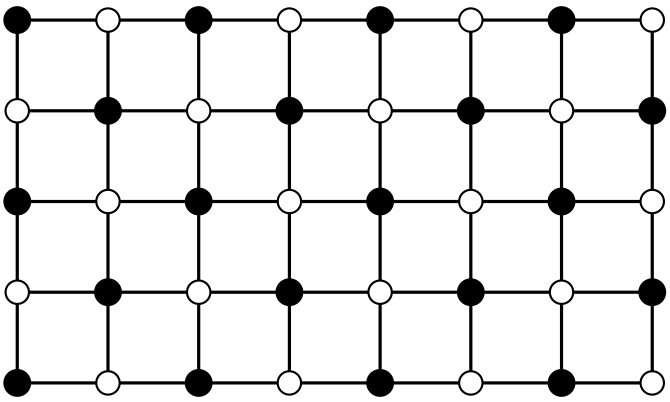
In Figure C.3 all the $\mathbf{X}$ variables among sites are denoted by black nodes, while all the $\mathbf{Y}$ variables are white nodes. The vertically aligned nodes, such as $X(s_{11})$ and $Y(s_{11})$, belong to the same site. One can clearly see that each site has four neighboring sites, one in each compass direction. The graph also contains many undirected cycles,

such as $X(s_{11}) - X(s_{12}) - Y(s_{12}) - Y(s_{11}) - X(s_{11})$.

Because of the presence of chordless cycles (like the one we have mentioned above), Figure C.3 is not a junction tree. What we may do is to use the idea mentioned in the conditional Markov random field part in Chapter 3 to convert this Markov random field into a junction tree, after conditioning on a selected set of nodes, also known as "*codes*" (Besag, 1974). His first coding method proposal is illustrated in Figure C.4. Let us consider one plane of the graph only, say the $\mathbf{X}$ plane. In his paper Besag suggests that the node set of this plane be halved alternately in a chess board pattern. In Figure C.4 any black node is conditionally independent from any other black nodes given the white node, and vice versa. Due to the conditional independence axioms, all the black nodes can therefore be treated as conditionally independently distributed given all the white nodes. This conditional independence leads to simple estimates. Denote $\mathbf{X}_b = \{X(s_{ij}) \text{ are black nodes}\}$ and $\mathbf{X}_w = \{X(s_{ij}) \text{ are white nodes}\}$, we then have $\mathbf{X}_b|\mathbf{X}_w, \mathbf{X}_w|\mathbf{X}_b \sim i.i.d.$ and the conditional maximum likelihoods $l(\mathbf{X}_b|\mathbf{X}_w)$ and $l(\mathbf{X}_w|\mathbf{X}_b)$ can be calculated. Based on the conditioning of black, or white nodes, the two runs of likelihood maximization of $l(\mathbf{X}_b|\mathbf{X}_w)$ and $l(\mathbf{X}_w|\mathbf{X}_b)$ will most likely produce two different estimates for each parameter, and these two estimates can eventually be combined together to form a final answer.

Although the conditional maximum likelihood estimate for this chess board coding method is algebraically intuitive, it does have a weakness. By conditioning on every other node, the junction tree takes into consideration only half of the nodes at a time. More importantly, it essentially ignores all the edges in the graph. The edges are important parts of the graph, and by removing them from the junction tree we overlooked local dependence structures. In the original, unconditional Markov random field, the

Figure C.4: First order nearest neighbor lattice, chess board coding

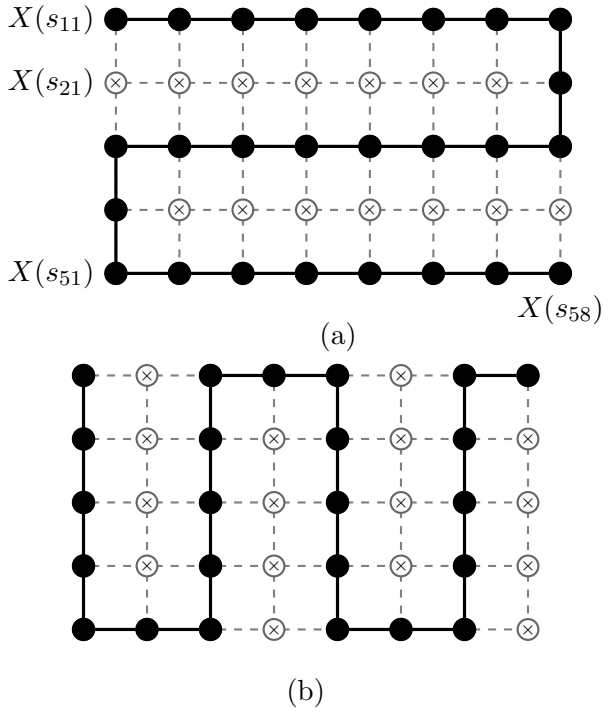


maximal cliques are 2-cliques. In the conditional Markov random field, however, the cliques are merely single standalone nodes. This means that none of the original cliques are evaluated as a whole in the junction tree. As we have pointed out before, cliques inherit the local dependence structure of the graph and are responsible for the joint distribution partition, the loss of this information during the chess board method is too much to bear.

We believe that the chess board coding method can be improved by incorporating more nodes and edges to the conditional Markov random field. One way to achieve this is by using what we call the "snake coding" method, as appears in Figure C.5. For both plots, when conditioned on the white nodes, the black nodes form a conditional Markov random field and a junction tree, which happens to be a tree network as well. The dashed lines indicate the edges removed from the original graph due to conditioning, while the black edges denote the ones preserved in the junction tree.

Compared to the chess board coding, there are more nodes included in these junction trees (always more than half), and they also retain some of the maximal cliques in the original graph. The calculation for the conditional maximal likelihood $l(\mathbf{X}_w|\mathbf{X}_b)$

Figure C.5: First order nearest neighbor lattice, two snake codings



should still be simple since $\mathbf{X}_w|\mathbf{X}_b$ is a tree network. Similar to the chess board coding, we end up with multiple estimates for the same graph. By shifting the starting node of the path, for instance from $X(s_{11})$ to $X(s_{21})$, we can create a new junction tree and new set of estimates. The black path, or "snake", may also travel horizontally [Figure C.5 (a)] or vertically [Figure C.5 (b)]. Altogether we may create four junction trees (from $X(s_{11})$ and $X(s_{21})$ vertically, and from $X(s_{11})$ and $X(s_{12})$ horizontally) to include all the edges from the original graph in the analysis.

Bibliography

Adriaenssens, V., Goethals, P., Charles, J., and De Pauw, N. (2004). Application of Bayesian belief networks for the prediction of macroinvertebrate taxa in rivers. *Annales de limnologie*, 40(3):181–192.

Andersson, S. and Perlman, M. (1998). Normal linear regression models with recursive graphical Markov structure. *Journal of Multivariate Analysis*, 66(2):133–187.

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

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997b). On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs. *Scandinavian Journal of Statistics*, 81:81–102.

Andersson, S. A., Madigan, D., and Perlman, M. D. (2001). Alternative Markov properties for chain graphs. *Scandinavian Journal of Statistics*, 28:33–85.

Andersson, S. A., Madigan, D., Perlman, M. D., and Triggs, C. M. (1997c). A graphical characterization of lattice conditional independence models. *Annals of Mathematics and Artificial Intelligence*, 21:27–50.

Andersson, S. A. and Perlman, M. D. (2006). Characterizing Markov equivalence classes for AMP chain graph models. *Annals of Statistics*, 34:939–972.

Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to

MCMC for machine learning. *Machine Learning*, 50:5–43.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models (Studies in Operational Regional Science)*. Springer, New York, New York.

Anselin, L. (1995). Local indicators of spatial association lisa. *Geographical Analysis*, 27:93–115.

Anselin, L. (2002). Under the hood: Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, 27(3):247267.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Heirarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, London, UK.

Barber, D. (2003). Probabilistic modelling and reasoning: the junction tree algorithm. Department of Computer Sciences, University College London, UK.

Bedford, T. and Cooke, R. M. (2002). Vines - a new graphical model for dependent random variables. *Annals of Statistics*, 30(4):1031–1068.

Besag, J. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34:75–83.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York, New York.

Brook, D. (1964). On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbor systems. *Biometrika*,

51:481–483.

Caetano, T., Caelli, T., Schuurmans, D., and Barone, D. (2006). Graphical models and point pattern matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1646–1663.

Carlin, B. P., Banerjee, S., and Gelfand, A. E. (2003). Hierarchical multivariate car models for spatio-temporally correlated survival data. In *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, pages 45–64. Oxford University Press, Oxford, UK.

Cevher, V. (2008). Junction tree algorithm. Electrical and Computer Engineering Department, Rice University, Houston, Texas. Unpublished.

Cheung, S. (2008). Proof of Hammersley-Clifford theorem. Department of Electrical and Computer Engineering, University of Kentucky, Lexington, Kentucky. Unpublished.

Chow, C. K., Member, S., and Liu, C. N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14:462–467.

Cliff, A. D. and Ord, J. K. (1981). *Spatial Processes: Models and Applications*. Pion Ltd, London, UK.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer, New York, New York.

Cox, D. R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs. *Statistical Science*, 8(3):204–128.

Cressie, N. and Lele, S. (1992). New models for Markov random fields. *Journal of Applied Probability*, 29(4):877–884.

Cressie, N. and Verzelen, N. (2008). Conditional-mean least-squares fitting of Gaussian

Markov random fields to Gaussian fields. *Computational Statistics & Data Analysis*, 52(5):2794–2807.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley-Interscience, Hoboken, New Jersey.

Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *Annals of Statistics*, 8(3):522–539.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31.

Dawid, A. P. (1980). Conditional independence for statistical operations. *Annals of Statistics*, 8(3):598–617.

Denuit, M. and Lambert, P. (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis*, 93(1):40–57.

Diggle, P., Tawn, J., and Moyeed, R. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 47(3):299–350.

Drton, M. and Eichler, M. (2006). Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property. *Scandinavian Journal of Statistics*, 33:247–257.

Earnest, A., Morgan, G., Mengersen, K., Ryan, L., Summerhayes, R., and Beard, J. (2007). Evaluating the effect of neighbourhood weight matrices on smoothing properties of Conditional Autoregressive (CAR) models. *International Journal of Health Geographics*, 6(1):54.

Erdely, A., Gonzalez-Barrios, J., and Nelsen, R. (2008). Symmetries of random discrete copulas. *Kybernetika*, 44(6):846.

Fedorov, V. (1996). Design of spatial experiments: model fitting and prediction. *Handbook of Statistics*, 13:515–553.

Fischer, M. M. and Getis, A. (2010). *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*. Springer, New York, New York.

Fischer, M. M. and Wang, J. (2011). *Spatial Data Analysis: Models, Methods and Techniques*. Springer, New York, New York.

Fisher, M. and Burford, R. (1967). Theory of critical-point scattering and correlations. I. The Ising model. *Physical Review*, 156(2):583.

Fisher, N. (1997). *Copulas, in Encyclopedia of Statistical Sciences, Volume 1*. Wiley, New York, New York.

Florax, R. and Folmer, H. (1992). Specification and estimation of spatial linear regression models: Monte Carlo evaluation of pre-test estimators. *Regional Science and Urban Economics*, 22(3):405432.

Frydenberg, M. (1990). The chain graph Markov property. *Scandinavian Journal of Statistics*, 17(4):333–353.

Gaetan, C. and Guyon, X. (2010). *Spatial Statistics and Modeling*. Springer, New York, New York.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman & Hall/CRC, London, UK.

Gelman, A., Roberts, G. O., and Gilks, W. R. (1996). Efficient Metropolis jumping rules. *Bayesian Statistics*, 5:599–608.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 721–741.

Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data. *Astin Bulletin*, 37(2):475–515.

Gibbs, W. (1902). *Elementary principles of statistical mechanics*. Yale University

Press, New Haven, Connecticut.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC, London, UK.

Gillespie, T. W. and Agnew, J. A. (2009). Finding Osama bin Laden: An application of biogeographic theories and satellite imagery. Online Working Paper Series, California Center for Population Research, Los Angeles, California.

Gitelman, A. I. and Herlihy, A. (2007). Isomorphic chain graphs for modeling spatial dependence in ecological data. *Environmental and Ecological Statistics*, 14(1):27–40.

Grimmett, G. R. (1973). A theorem about random fields. *Bulletin of the London Mathematical Society*, 5(1):81–84.

Haas, T., Mowrer, H., and Shepperd, W. (1994). Modeling aspen stand growth with a temporal Bayes network. *AI applications*, 8(1):15–28.

Haining, R. (1993). *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press, Cambridge, UK.

Haining, R. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge University Press, Cambridge, UK.

Hammersley, J. M. and Clifford, P. (1971). Markov fields on finite graphs and lattices. Unpublished.

Hamze, F. and Freitas, N. D. (2006). Hot coupling: a particle approach to inference and normalization on pairwise undirected graphs. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems 18*, pages 491–498. MIT Press, Cambridge, Massachusetts.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Hoeffding, W. (1940). Masstabinvariante Korrelationstheorie. *Schriften des Mathe-*

matischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin, 5:179–233.

Huang, F. and Ogata, Y. (2002). Generalized pseudo-likelihood estimates for Markov random fields on lattice. *Annals of the Institute of Statistical Mathematics*, 54(1):1–18.

Irvine, K. (2007). *Graphical Models for Multivariate Spatial Data*. PhD thesis, Oregon State University, Corvallis, Oregon.

Irvine, K. and Gitelman, A. I. (2010). Graphical spatial models: a new view on interpreting spatial pattern. *Environmental and Ecological Statistics*, 18(3):447–469.

Jensen, F. V. and Jensen, F. (1994). Optimal junction trees. In *Proceeding of the Tenth Conference on Uncertainty in Artificial Intelligence*. Seattle, Washington.

Jensen, F. V., Lauritzen, S. L., and Olesen, K. G. (1990). Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quarterly*, 4:269–282.

Jensen, J. and Møller, J. (1991). Pseudolikelihood for exponential family models of spatial point processes. *The Annals of Applied Probability*, 1(3):445–461.

Jin, X., Carlin, B. P., and Banerjee, S. (2005). Generalized hierarchical multivariate CAR models for areal data. *Biometrics*, 61:950961.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, 19(1):140–155.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

Kato, Z., Berthod, M., and Zerubia, J. (1996). A hierarchical Markov random field model and multitemperature annealing for parallel image classification. *Graphical Models and Image Processing*, 58(1):18–37.

Kiiveri, H., Speed, T. P., and Carlin, J. B. (1984). Recursive causal models. *Journal*

of the Australian Mathematical Society. Series A, 36:30–52.

Kirshner, S. (2007). *Learning with tree-averaged densities and distributions*. Conference on Neural Information Processing Systems (NIPS). Vancouver, British Columbia.

Kissling, W. D. and Carl, G. (2008). Spatial autocorrelation and the selection of simultaneous autoregressive models. *Plant Ecology*, 17(1):59–71.

Kittler, J. and Föglein, J. (1984). Contextual classification of multispectral pixel data. *Image and Vision Computing*, 2(1):13–29.

Knorr-Held, L. and Rue, H. (2002). On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29(4):597–614.

Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press, Oxford, UK.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H.-G. (1990). Independence properties of directed Markov fields. *Networks*, 20:491–505.

Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of statistics*, 17(1):31–57.

Legendre, P. and Fortin, M. J. (1989). Spatial pattern and ecological analysis. *Plant Ecology*, 2:107–138.

LeSage, J. and Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Chapman & Hall/CRC, London, UK.

Levitz, M., Perlman, M. D., and Madigan, D. (2001). Separation and completeness properties for AMP chain graph Markov models. *Annals of Statistics*, 29(6):1751–1784.

Li, S. Z. (2001). *Markov Random Field Modeling in Image Analysis*. Computer Science Workbench. Springer, New York, New York.

Ma, J. and Sun, Z. (2008). Dependence structure estimation via copula.

http://arxiv.org/abs/0804.4451.

Madigan, D., Andersson, S., Perlman, M., and Volinsky, C. (1996). Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Communications in Statistics: Theory and Methods*, 25(11):2493–2519.

Madsen, L. (2009). Maximum likelihood estimation of regression parameters with spatially dependent discrete data. *Journal of Agricultural, Biological and Environmental Statistics*, 14(4):375–391.

Madsen, L. and Birkes, D. (2011). Simulating dependent discrete data. *Journal of Statistical Computation and Simulation*. In press.

Maier, D. (1983). *The Theory of Relational Databases*. Computer Science Press.

Mardia, K. V. (1988). Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. *Journal of Multivariate Analysis*, 24(2):265–284.

Matúš, F. (1992). On equivalence of Markov properties over undirected graphs. *Journal of Applied Probability*, pages 745–749.

Mayor, G. and Torrens, J. (2005). *Logical, Algebraic, Analytic, and Probabilistic Aspects of Triangular Norms*, chapter 7, pages 189–230. Elsevier B.V., Amsterdam, the Netherlands.

Meilă, M. and Jaakkola, T. (2006). Tractable Bayesian learning of tree belief networks. *Statistics and Computing*, 16(1):77–92.

Meilă, M. and Jordan, M. (2001). Learning with mixtures of trees. *The Journal of Machine Learning Research*, 1:1–48.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092.

Moura, J. M. F. and Balram, N. (1992). Recursive structure of noncausal Gauss-Markov random fields. *IEEE Transactions on Information Theory*, 38(2):334 –354.

Müller, P. (1991). A generic approach to posterior integration and Gibbs sampling. Technical report, Purdue University, West Lafayette, Indiana.

Nelsen, R. (1999). *An Introduction to Copulas*. Springer-Verlag, New York, New York.

Ogata, Y. and Tanemura, M. (1984). Likelihood analysis of spatial point patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(3):496–518.

Paskin, M. (2009). A short course on graphical models. Stanford University, Palo Alto, California. Unpublished.

Pearl, J. and Paz, A. (1987). Graphoids: A graph based logic for reasoning about relevance relations. *Advances in Artificial Intelligence*, 2:357–363.

Pearl, J. and Verma, T. (1987). The logic of representing dependencies by directed graphs. In *Proceedings of the Sixth Conference of American Association of Artificial Intelligence*, pages 374–379. Seattle, Washington.

Pitcher, K. W., Burkanov, V. N., Calkins, D. G., Boeuf, B. J. L., Mamaev, E. G., Merrick, R. L., and Pendleton, G. W. (2001). Spatial and temporal variation in the timing of births of Steller sea lions. *Journal of Mammalogy*, 82(4):1047–1053.

Polastre, J., Szewczyk, R., Mainwaring, A., Culler, D., and Anderson, J. (2004). Analysis of wireless sensor networks for habitat monitoring. *Wireless sensor networks*, 6:399–423.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44(4):1033–1048.

Preston, C. J. (1973). Generalized Gibbs states and Markov random fields. *Advances in Applied Probability*, 5(2):242–261.

Quinn, G. and Keough, M. (2002). *Experimental Design and Data Analysis for Biol-*

ogists. Cambridge University Press, Cambridge, UK.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ripley, B. and Kelly, F. (1977). Markov point processes. *Journal of the London Mathematical Society*, 2(1):188.

Robert, C. P. and Casella, G. (2010). *Monte Carlo Statistical Methods*. Springer-Verlag, New York, New York.

Robins, G., Pattison, P., and Wasserman, S. (1999). Logit models and logistic regressions for social networks: Iii. valued relations. *Psychometrika*, 64:371–394.

Rosenzweig, M. and MacArthur, R. (1963). Graphical representation and stability conditions of predator-prey interactions. *American Naturalist*, pages 209–223.

Roverato, A. (2006). A graphical representation of equivalence classes of AMP chain graphs. *Journal of Machine Learning Research*, 7:1045–1078.

Rue, H. (1999). A fast and exact simulation algorithm for general Gaussian Markov random fields. Technical report, Norwegian University of Science and Technology, Trondheim, Norway.

Rue, H. (2000). Fast sampling of Gaussian Markov random fields with applications. *Journal of the Royal Statistical Society. Series B (Methodological)*, 63:325–338.

Rue, H. and Knorr-Held, L. (2005). *Gaussian Markov random fields: theory and applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, UK.

Rue, H. and Tjelmeland, H. (1999). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, 29:31–49.

Sang, H. and Gelfand, A. E. (2009). Hierarchical modeling for extreme values observed over space and time. *Environmental and Ecological Statistics*, 16(3):407–426.

Schabenberger, O. and Gotway, C. A. (2004). *Statistical Methods for Spatial Data Analysis*. Chapman & Hall/CRC, London, UK.

Schmidt, A. M. and Gelfand, A. E. (2003). A Bayesian coregionalization approach for multivariate pollutant data. *Journal of Geophysical Research*, 108(24):8783–8790.

Shachter, R. (1998). Bayes-ball: Rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). In *Proceedings of the fourteenth conference on Uncertainty in Artificial Intelligence*, pages 480–487. Morgan Kaufmann Publishers, Burlington, Massachusetts.

Sherman, S. (1973). Markov random fields and Gibbs random fields. *Israel Journal of Mathematics*, 14(1):92–103.

Silva, R. and Gramacy, R. B. (2009). MCMC methods for Bayesian mixtures of copulas. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, pages 512–519. JMLR W & CP, Clearwater Beach, Florida.

Smith, J. Q. (1989). Influnence diagrams for statistical modelling. *Annals of Statistics*, 17:654–672.

Song, P. X.-K. (2000). Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics*, 27:305–320.

Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. McGill University, Montreal, Quebec.

Strauss, D. and Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85(409):204–212.

Studený, M. (2001). *On Mathematical Description of Probabilistic Conditional Independence Structures*. PhD thesis, Academy of Sciences of the Czech Republic, Prague, Czech Republic.

Thioulouse, J., Chessel, D., and Champely, S. (1995). Multivariate analysis of spatial patterns: a unified approach to local and global structures. *Environmental and Ecological Statistics*, 2:1–14.

Thomas, A., Best, N., Lunn, D., Arnold, R., and Spiegelhalter, D. (2004). GeoBUGS User Manual. *http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/geobugs.shtml*.

Tidén, E. and Arnborg, S. (1987). Unification problems with one-sided distributivity. *Journal of Symbolic Computation*, 3(1/2):183–202.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, 22:1701–1762.

Tierney, L. and Mira, A. (1999). Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine*, 18(1718):2507–2515.

Tjelmeland, H. and Besag, J. (1998). Markov random fields with higher-order interactions. *Scandinavian Journal of Statistics*, 25(3):415–433.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2):234–240.

Vanhatalo, J. and Vehtari, A. (2007). Sparse log Gaussian processes via MCMC for spatial epidemiology. In *JMLR: Workshop and Conference Proceedings, Gaussian Processes in Practice*, volume 1, pages 73–89. Bletchley, UK.

Verma, T. and Pearl, J. (1992). An algorithm for deciding if a set of observed independencies has a causal explanation. In *Proceedings of the Eighth international conference on uncertainty in artificial intelligence*, pages 323–330. Morgan Kaufmann Publishers, Burlington, Massachusetts.

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.

Wall, M. M. (2004). A close look at the spatial correlation structure implied by the car and sar models. *Journal of Statistical Planning and Inference*, 121(2):311–324.

Wang, Y.-G. (1994). Statistical issues in mapping and monitoring using remote sensing data. Technical report, CSIRO Mathematical and Information Sciences, Clayton South, Australia.

Wartenberg, D. (1985). Multivariate spatial correlation: a method for exploratory geographical analysis. *Geographical Analysis*, 17(4):263–283.

Wermuth, N. (1991). *On Block-Recursive Linear Regression Equations*. Johannes Gutenberg-Universität, Mainz, Germany.

Wermuth, N., Cox, D. R., and Pearl, J. (1994). Explanations for multivariate structures derived from univariate recursive regressions. Technical report, Johannes Gutenberg-Universität, Mainz, Germany.

Wermuth, N. and Lauritzen, S. L. (1989). Graphical and recursive models for contingency tables. *Biometrika*, 70:537–552.

Wermuth, N. and Lauritzen, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(1):21–50.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 41(3/4):434–449.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20(7):557–585.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215.

Zimmerman, D. and Harville, D. (1991). A random field approach to the analysis of field-plot experiments and other spatial experiments. *Biometrics*, 47:223–239.