# A Geometric Approach for Inference on Graphical Models

by

Simón Lunagómez

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Sayan Mukherjee, Advisor

_____
Robert L. Wolpert, Advisor

_____
Mike West

_____
John Harer

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2009

# Abstract

(Statistics)

A Geometric Approach for Inference on Graphical Models

by

Simón Lunagómez

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Sayan Mukherjee, Advisor

_____
Robert L. Wolpert, Advisor

_____
Mike West

_____
John Harer

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University

2009

# Abstract

We formulate a novel approach to infer conditional independence models or Markov structure of a multivariate distribution. Specifically, our objective is to place informative prior distributions over graphs (decomposable and unrestricted) and sample efficiently from the induced posterior distribution. We also explore the idea of factorizing according to complete sets of a graph; which implies working with a hypergraph that cannot be retrieved from the graph alone. The key idea we develop in this thesis is a parametrization of hypergraphs using the geometry of points in $\mathbb{R}^m$. This induces informative priors on graphs from specified priors on finite sets of points. Constructing hypergraphs from finite point sets has been well studied in the fields of computational topology and random geometric graphs. We develop the framework underlying this idea and illustrate its efficacy using simulations.

To My Parents

# Contents

# List of Tables

# List of Figures

# Acknowledgements

How did a class project became a dissertation? The answer is: Because my advisors believed in me and supported me even when it was uncertain where this project was going. So let me thank first Professors Sayan Mukherjee and Robert Wolpert for all their support, guidance, and encouragement during my grad school years.

I want to thank Professors John Harer and Herbert Edelsbrunner for their support, time, and curiosity. They and their students, Paul Bendich, Dimitriy Morozov, and Bei Wang (specially Bei) were invaluable for this research.

I thank Professor Mike West for his suggestions and interest in the project. His curiosity and comments were always a source of encouragement. I also want to thank Professors Jonathan Mattingly, Mauro Maggioni, and David Banks for their advice and encouragement. I will remember their generosity.

During this four years I was blessed with the gift of friendship. I specially want to thank Gavino Puggioni, Francesca Petralia, Veronica Berrocal, Kristian Lum and Paul Richard Hahn for their friendship. I also want to thank David and Lenka Siroky for their generosity, they always supported me when I needed it.

Let me finish by thanking my parents, without their unconditional love none of this would have happened. This work is a product of their love and efforts.

# 1

# Introduction

A wide variety of problems in the natural and social sciences reduce to investigating which elements in a system influence other elements. One way to approach this problem is to think of each element as an entry of a random vector; then specific questions about elements influencing others translate to statements about the dependence structure. The graphical models framework offers algorithms and mathematical results that make this approach appealing for practical use; this perspective provides a way for constructing high dimensional probability models in terms of lower dimensional ones. These models encode a series of conditional independence statements (or Markov structure) in a graph. Graphical models are modular, both in terms of computation (of marginal, conditional distributions) and interpretation (because of the Markov structure). Inferences on this models can be performed according to either the frequentist or the Bayesian approach. Although in the last 10 years they have become increasingly popular within the Bayesian community since they encompass

hierarchical models, hidden Markov models and other constructions commonly used by Bayesians.

The origins of these models can be traced back to several applied areas; one of them is statistical physics. Gibbs [46] was concerned with the interaction of particles in a system (possibly atoms of a gas or solid). Each particle was associated to a state and it was assumed that neighbor particles (this is, physically close) were likely to influence each other. Another area that motivated this modeling approach was genetics (Wright [107],[108], [109] ); here assumptions regarding heritable properties were encoded in a directed graph.

In their seminal work Cowell, Dawid, *et al.* [22] showed a way of encoding conditional probability statements in a directed graph to solve a concrete expert system problem. A formal treatment for combining graph theory and probability was developed by Pearl [83]; this was done in the context of artificial intelligence. Lauritzen and Wermuth [70] discussed how to define a graphical model when some of the marginals are discrete and some are continuous. The monographs by Whittaker [105] and Lauritzen [67] provide a nice overview of the mathematical and statistical foundations of this area.

Dawid and Lauritzen [26] defined the idea of hyper Markov law, this is, a probability distribution define on the set of probability measures that are Markov with respect to a graph; this concept is crucial for Bayesian inference and model selection. Simultaneous inference of a decomposable graph and marginals in a fully Bayesian framework was developed in Green and Giudici [49] using local proposals to sam-

ple graph space. A promising extension of this approach called Shotgun Stochastic Search (SSS) takes advantage of parallel computing to select from a batch of local moves Jones, *et al.* [61]. A stochastic search method that incorporates local moves as well as more aggressive moves in graph space has been developed by Scott and Carvalho [96]. These stochastic search methods are intended to identify regions with high posterior probability, but their convergence properties are still not well understood. Bayesian models for non-decomposable graphs have been proposed by Roverato [93] and Wong *et al.* [106]. These two approaches focus on Monte Carlo sampling of the posterior distribution from appropriate hyper Markov prior laws.

Erdös-Rényi random graphs (those in which each of the $\binom{d}{2}$ possible edges $(i, j)$ is included in $\mathcal{E}$ independently with some specified probability $p \in [0, 1]$), and variations where the edge inclusion probabilities $p_{ij}$ are allowed to be edge-specific, have been used to place informative priors on decomposable graphs Heckerman *et al.* [54, 72]. The number of parameters in this prior specification can be enormous if the inclusion probabilities are allowed to vary, and some interesting features of graphs may be hard to express solely by specifying edge probabilities. Mukherjee and Speed [75] developed methodology for placing informative distributions on directed graphs by using what they call concordance functions; this is a function that is increasing as the graph agrees more with a given feature.

In this thesis we propose parametrizations of graph spaces based on intersection patterns of convex sets. These convex sets are parametrized in terms of a configuration of points in $\mathbb{R}^m$ and a size parameter $r \geq 0$. To perform inferences on the graph

structure according to the Bayesian approach, it is necessary to propose priors and MCMC algorithms suitable to these parametrizations. To give an idea of the type of priors we use, just observe that by placing a probability model on the configuration of points we obtain a distribution for a space of graphs. An advantage of this way of inducing probability distributions on graphs is that it allows for simultaneous control of two or more graph features, something that cannot be achieved by an Erdös-Rényi model with single inclusion probability. In the same way, MCMC procedures designed for sampling distributions that have a region of $\mathbb{R}^m$ as support can be used in this setting. An implication of this is that one can propose algorithms that allow local and global moves in a space of graphs within the MCMC framework. Another interesting feature of our methodology is that it permits the parametrization and inference of hypergraphs, which can prove useful when dealing with non-Gaussian models, since there are aspects of the dependence structure that go beyond pairwise relationships.

In Chapter 2 we provide examples to motivate the use of geometric graphs and patterns of convex sets in graphical models. Chapter 3 gives an overview of graph theory and graphical models. In Chapter 4 we introduce concepts from computational topology and random geometric graphs; from the former we obtain a way for parametrizing spaces of graphs and hypergraphs, the later gives a framework for proposing priors for graphs based on distributions on point patterns. We present concrete probability models and propose MCMC procedures in Chapter 5. Empirical results are discussed in Chapter 6.

# 2

# Why a Geometric Approach?

In this chapter we motivate a geometric representation of graphs. In Section 2.1 we discuss two approaches for defining a graph. The focus of Section 2.2.3 is to discuss features that make this perspective attractive for statistical modelling

## 2.1 Two ways to Understand a Graph

The most common way to define a graph $\mathcal{G}$ is to regard it as a set of abstract objects (called vertices) and a collection of pairs of such objects (the edges of the graph), let us call this the algebraic approach. Most graph properties used in applications are phrased according to this perspective. For example: If all pairs are regarded as unordered (ordered) then the graph is said to be undirected (directed). If no more than one edge involves the same pair of vertices and no edge consists of the same vertex repeated, the graph is said to be simple.

A different way to represent a graph is to understand the vertices as a set of points

in $\mathbb{R}^m$ and the edges as curves joining those points; let us refer to this perspective as the geometric approach. This is in fact the usual way to illustrate simple undirected graphs (Figures 2.1 and 2.2). We adopt the convention that crossings between the curves are allowed unless stated otherwise.

The fields of Computational Topology and Topological Graph Theory deal with the interplay between both representations. In Computational Topology the goal is to represent a topological space in a way that permits efficient computations [34]. Often this implies obtaining a combinatorial structure from a geometric or topological object (think of the process of automating the triangulation a manifold that is embedded in $\mathbb{R}^m$). Topological Graph Theory studies how graphs can be represented as topological spaces. It is intuitively clear that each edge of a graph can be represented as a homeomorphism between the interval $[0, 1]$ and a subspace of $\mathbb{R}^m$. This area comprehends results like Kuratowski's Theorem, which fully characterizes which graphs can be drawn in $\mathbb{R}^2$ without crossings.

## 2.2 Why a Geometric Perspective Can Be Useful

Most methodologies for making inferences on graphical models use parametrizations and priors (in the Bayesian setting) implied by the algebraic approach; the work of [74] is a remarkable exception. We believe that adopting a geometric perspective can be helpful for the statistician. The following discussions will motivate the key ideas.

### *2.2.1 If You Can Draw It, It is Sparse*

Nowadays there is an emphasis on developing statistical methodologies for graphical models that incorporate the assumption that the underlying graph is sparse. There are good reasons for this: prior knowledge about the application may suggest that each component in the model can only interact with few other components, this is the case in gene expression data and most image analysis problems. Another reason is computational efficiency, for some operations (e.g. obtaining marginal distributions) it is hard or even prohibitive to work with graphs with many edges.

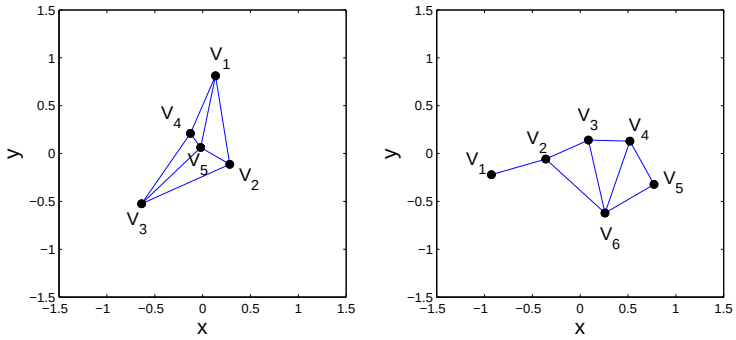


FIGURE 2.1: Two examples of planar graphs. For (A) we have $n = 5$, $m = 8$ and $k = 5$, while for (B) $n = 6$, $m = 8$, $k = 4$. In both cases $n - m + k = 2$.

A graph is called planar if it can be represented as a set of points (vertices) and curves (edges) with no crossings. How does this restriction affect features like the number of edges and maximal clique size of the graph? To answer this question we refer to two basic results in Topological Graph Theory. We first look at

**Theorem 2.2.1** (Euler Relation For Planar Graphs)**.** *Every embedding of a connected graph in the plane satisfies* $n - m + k = 2$*, where* $n$ *denotes the number of*

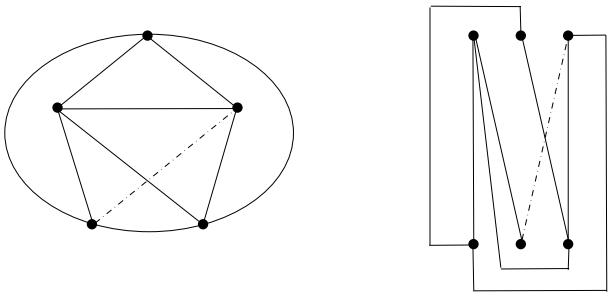

FIGURE 2.2: $K_5$ and $K_{3,3}$. These are the graphs referred in Kuratowski's Theorem, they cannot be drawn without crossings (dotted lines).

*vertices, $m$ the number of edges and $k$ the number of connected components of the complement of the graph with respect to the plane.*

A consequence of this result is that $m \leq 3n - 6$ and $k \leq 2n - 4$ for any $n \geq 3$ (see Figure 2.1). Therefore planar graphs are forced to be sparse. Let us investigate if something can be said on the maximal clique size of the graph. In Figure 2.2 we show representations of the graphs $K_5$ and $K_{3,3}$; neither of them can be drawn without crossings. Now we are ready to state:

**Theorem 2.2.2** (Kuratowski's Theorem)**.** *A simple graph is planar if and only if it has no subgraph homeomorphic to $K_5$ or to $K_{3,3}$.*

This result implies that for a planar graph the maximum clique size cannot be greater than 4. Therefore we have that planarity controls at least two key features of the graph. In contrast, any graph can be embedded in $\mathbb{R}^3$, however one can still use the dimension of the space where the graph is embedded to control specific attributes. This can be done by working only with graphs produced by a geometric

construction contained in $\mathbb{R}^m$, therefore by manipulating the parameters in the construction, different types of restrictions on graph space can be imposed. Specifically, we will work with algorithms that generate simplicial complexes to achieve this. For the moment think of a simplicial complex as a topological subspace of $\mathbb{R}^m$ with a graph structure associated to it. This idea is explored in Sections 4.1 and 4.2 .

### *2.2.2 The Affordable Hypergraph*

Let $(X, Y, Z)$ be a random vector and assume that

- $X \perp\!\!\!\perp Y \mid Z$ and
- $X \perp\!\!\!\perp Z \mid Y$

Therefore, if one encodes this information in a graph following the convention that the absence of an edge implies that variables incident to that edge are conditional independent given all other variables (see Section 3.3), we end up with the graph displayed in 2.3 (A).

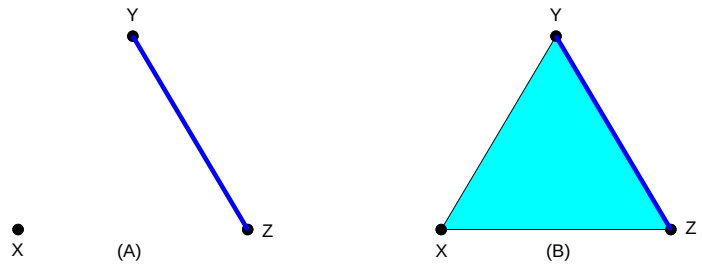


FIGURE 2.3: (A) is a graphical model that encodes the assumptions $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$. (B) is an hypergraph that distinguishes the case when $X$ is not independent from $X \perp\!\!\!\perp (Y, Z)$.

What can be said about the dependence structure between $X$ and $(Y, Z)$? It is tempting to claim that $X \perp\!\!\!\perp (Y, Z)$, however this is true only if the density of all variables is continuous and positive with respect to a product measure (Proposition 3.1 in [67]). This would be the case if $(X, Y, Z)$ followed a multivariate normal distribution. In contrast if we set $\Pr\{\alpha = 1\} = \Pr\{\alpha = -1\} = 1/2$ and

$$\begin{aligned} \eta_i &\sim \mathrm{N}(0, 1) \\ \nu_i &= \alpha \times |\eta_i| \end{aligned}$$

for $i = 1, 2, 3$ with $\{\alpha, \eta_1, \eta_2, \eta_3\}$ independent, then we have that any pair of the $\nu_i$'s is conditionally independent given the third one, but since the three variables are forced to have the same sign $\{\nu_1, \nu_2\}$ cannot be independent of $\nu_3$. Note that the condition of Proposition 3.1 in [67] is violated since the joint density of $\{\nu_1, \nu_2, \nu_3\}$ is positive for only two of the octants in $\mathbb{R}^3$. This example suggests that there can be relevant features in the dependence structure of a random vector that are beyond pairwise relationships (in this case the relationship is given by the absence of an edge in the graph). A natural way to proceed is to encode all conditional independence statements via a hypergraph. For instance one could add a hyperedge $\{X, Y, Z\}$ to distinguish the case when $X$ and $(Y, Z)$ are not independent; see Figure 2.3 (B).

It is been widely acknowledged that making inference on graphical models is a challenging problem, mainly because the model space grows superexponentially with respect to the number of nodes in the graph. Moving from graphs to hypergraphs may seem to turn this problem into an even worse one. However if one assumes that the hypergraph is in some sense sparse, a different representation of it may be

feasible and in fact manageable for Bayesian computations.

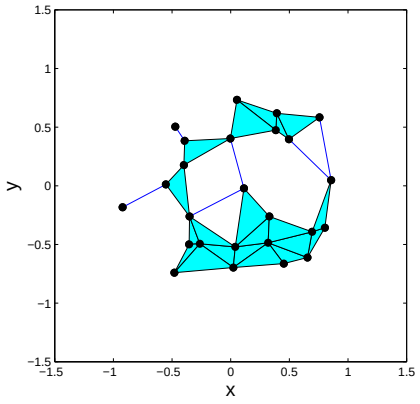


FIGURE 2.4: Alpha complex in $\mathbb{R}^2$ with 25 vertices and $r = 0.274$. Note that not all complete sets of size 3 are associated to 3−dimensional hyperedge (filled triangle).

A specific example of such representation is the Alpha complex in $\mathbb{R}^2$ (p. 81 in [34]). This construction can be understood as a hypergraph that is computed from the coordinates of a finite set of points in $\mathbb{R}^2$ or $\mathbb{R}^3$ and a tuning parameter $r > 0$ (Figure 2.4). The idea of using geometric and topological constructions to obtain low dimensional representations of complex combinatorial objects is central in this work. In Chapter 5 and Section 6.3 we develop methodology based on this idea.

### *2.2.3 From Joining the Dots to Getting the Prior*

How to design probability distributions on graph spaces that give higher probability to graphs with certain features? This question is relevant when designing prior distributions over graph spaces, and some answers can be found in the Random Graph literature. Still there is the need for models that can be both rich and simple. By richness we mean that one should be able to control the distribution of several

graph features simultaneously, and by simplicity it is meant that this should be achieved without introducing a large number of parameters in the model.

While proposing a probability distribution on a space of graphs tailored to a specific application is not an easy task, designing distributions on $\mathbb{R}^m$ is a well understood problem for which there are many tools available. In Figure 2.5 we show realizations from 4 different distributions of point patterns in the unit square. Now, for each of these point patterns join each pair of points $\{v_i, v_j\}$ for which $\|v_i - v_j\| \leq 0.01$. The graphs produced by this procedure are shown in Figure 4.1. It is clear that these graphs are qualitatively different in terms of number of edges, maximum clique size, maximum edge degree and number of connected components. This phenomenon goes beyond a specific realization of these point patterns, in fact this is a way of inducing probability distributions on graph spaces. The the theory of Random Geometric Graphs is a branch of Stochastic Geometry that studies these relationships [84].

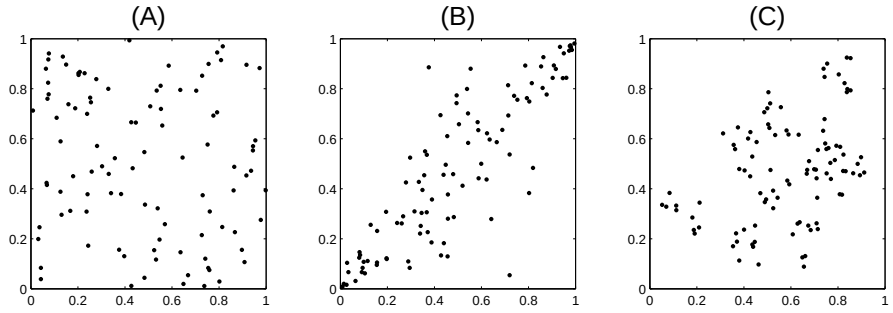


FIGURE 2.5: Random point patterns of size 100 on $[0, 1]^2$ sampled from (A) uniform distribution, (B) t-copula with $\rho = 0.85$ and 3 degrees of freedom, and (C) cluster Poisson process.

In Section 4.3 we investigate how the Random Geometric Graph Theory tools

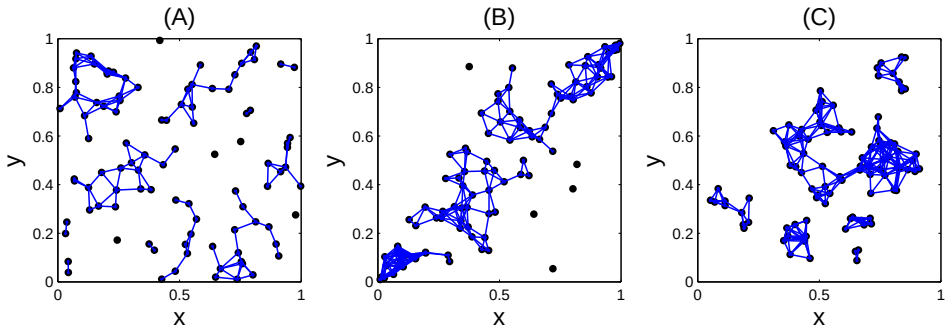


FIGURE 2.6: Proximity graphs implied by the point patterns displayed in Figure 2.5.

can be used as an alternative to standard Random Graph approaches for proposing priors on graphs.

# 3

# A Primer on Graphical Models

In this chapter we present the concepts and results from graph theory and graphical models that will be used in this thesis. In Section 3.1 the basic concepts of graph theory are defined. Section 3.2 focuses on the idea of decomposability. In Section 3.3 we present the basic definitions of graphical models. Section 3.4 provides background on Bayesian inference on graphical models. In Section 3.5 we discuss some technical aspects of Gaussian graphical models. In section 3.6 we discuss some differences between decomposable and non decomposable models.

## 3.1 Graph Theory

A *graph* $\mathcal{G}$ is a pair composed by a set $\mathcal{V}$ and a collection $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$; this is usually written as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$; $\mathcal{V}$ and $\mathcal{E}$ are called the vertex set and the edge set, respectively. The elements of $\mathcal{V}$ are called *vertices* and the elements of $\mathcal{E}$ are called the *edges* of the graph. If all the elements $(v, w)$ of $\mathcal{E}$ are ordered (unordered) pairs, the graph

is called *directed* (*undirected*). $v$ and $w \in \mathcal{V}$ are said to be *adjacent* if $(v, w) \in \mathcal{E}$. For the rest of this chapter, we will regard all graphs as undirected unless stated otherwise.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph; the quantities $|\mathcal{V}|$ and $|\mathcal{E}|$ are called, respectively the *size* and the *order* of $\mathcal{G}$. If $(v_i, v_j) \in \mathcal{E}$ we say that the vertices $v_i$ and $v_j$ are *adjacent*. A *walk* between two vertices $v_i$ and $v_j$ is a sequence of vertices and edges $\left\{v_{m_1}, e_1, v_{m_2}, e_2, \ldots, v_{m_{k-1}}, e_{k-1}, v_{m_k}\right\}$ such that $e_i = (v_{m_i}, v_{m_{i-1}})$, with $v_i = v_{m_i}$, and $v_j = v_{m_k}$. A walk for which all edges are distinct is called a *trail*. If a walk involves $k \geq 3$ vertices, all of them distinct but $v_{m_1} = v_{m_k}$, then it is said to be a *cycle*. A *triangulated graph* is a graph with the property that for every cycle of lenght $\geq 4$, there are two non-consecutive vertices that are adjacent. Let $A$, $B$ and $S$ be subsets of $\mathcal{V}$, then $S$ separates $A$ from $B$ if every walk starting in $A$ and finishing in $B$ contains a vertex in $S$ (equivalently, $S$ is a separator for $A$ and $B$).

A *connected component* of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a maximal subset $\mathcal{V}_0$ of $\mathcal{V}$ such that there is no pair of elements of $\mathcal{V}_0$ that cannot be connected by a walk; let $v \in \mathcal{V}$, the connected component that contains $v$ is denoted by $[v]_{\mathcal{G}}$. A graph with a single connected component is said to be *connected*. If a connected graph has no cycles, then it is called a *tree*; in general, a graph with no cycles is said to be a *forest*.

A *subgraph* $\mathcal{G}_0$ of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a graph $(\mathcal{V}_0, \mathcal{E}_0)$ with $\mathcal{V}_0 \subseteq \mathcal{V}$ and $\mathcal{E}_0 \subseteq \mathcal{E}$; let $\mathcal{V}_0 \subseteq \mathcal{V}$, the *induced subgraph* $\mathcal{G}_{\mathcal{V}_0}$ is the graph with vertex set $\mathcal{V}_0$ and edge set $\{(v, w) \in \mathcal{E} : \ v, w \in \mathcal{V}_0\}$. A *complete graph* is a graph where every pair of vertices is adjacent, clearly a complete undirected graph of size $k$ has order $\binom{k}{2}$. A *clique* of a

graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a maximal complete subgraph (with respect to inclusion). The *neighborhood* $\Gamma(v)$ of a vertex $v \in \mathcal{V}$ is the set of vertices that are adjacent to it. The *degree* of a vertex $v$ is defined as $|\Gamma(v)|$.

Two graphs $\mathcal{G}_0 = (\mathcal{V}_0, \mathcal{E}_0)$ and $\mathcal{G}_1 = (V_1, \mathcal{E}_1)$ are said to be *isomorphic* if there exists a bijection $\phi : \mathcal{V}_0 \to \mathcal{V}_1$ such that all adjacency relations in $\mathcal{G}_0$ are preserved in $\mathcal{G}_1$ (see Figure 3.1 ). This is denoted by $\mathcal{G}_0 \cong \mathcal{G}_1$.

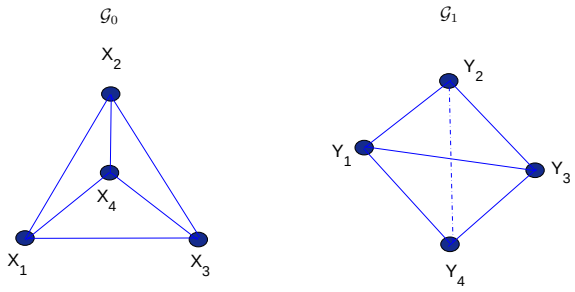


FIGURE 3.1: $\mathcal{G}_0$ is isomorphic to $\mathcal{G}_1$, because the mapping $\varphi(X_i) = Y_i$, $1 \leq i \leq 4$ is bijective and preserves the adjacency relations.

A natural extension to the concept of graph is the one of an hypergraph. Again it is formed by a pair $(\mathcal{V}, \mathcal{H})$ where $\mathcal{V}$ is the vertex set and $\mathcal{H}$ is a collection of subsets of $\mathcal{V}$. The elements of $\mathcal{H}$ are called hyperedges (For an example see Figure 3.2).

## 3.2 Decomposability

In this section we introduce the ideas related to the concept of decomposability. In later sections we will introduce probabilistic concepts that will lead to certain computations on graphs. From a conceptual as well as practical point of view it is

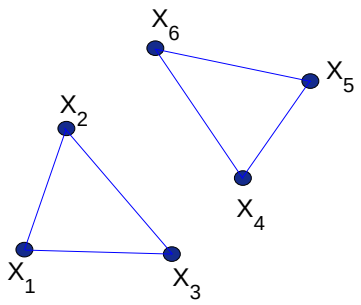


FIGURE 3.2: The set of all cliques of the this graph $\{1, 2, 3\}$ and $\{4, 5, 6\}$ constitutes an hypergraph. Another hypergraph that defined from this graph is formed by its complete sets, this is, $\{1, 2, 3\} \cup \{4, 5, 6\} \cup \mathcal{E} \cup \mathcal{V}$.

convenient to perform such operations locally. The notions of decomposability and prime component determine what is meant by local. Complete sets of the graph will play a special role here.

Before discussing decomposability, we need to define one more concept: A marked graph is a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, such that $\mathcal{V} = \Delta \cup \Psi$ and $\Delta \cap \Psi = \varnothing$, this is $(\Delta, \Psi)$ form a partition of the vertex set (one of them can be empty). For the moment $\Delta$ and $\Psi$ will be labels without specific meaning. We will provide an interpretation to them in the next section.

**Definition 3.2.1** (Decomposition)**.** Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected marked graph with vertex set $\mathcal{V}$ and edge set $E$, and $(A, B, S)$ a partition of $V$ into nonempty subsets such that:

1. $S$ separates $A$ from $B$;

2. $S$ is a complete subset of $\mathcal{V}$;

3. if either $S \subseteq \Delta$ or $B \subseteq \Psi$

Then it is said that $(A, B, S)$ is a decomposition of $\mathcal{G}$ into the components $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ ($\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ are the induced subgraphs for $A \cup S$ and $B \cup S$), respectively. If only conditions 1 and 2 hold, then $(A, B, S)$ is said to be a weak decomposition of $\mathcal{G}$.

The concept of decomposable graph derives from the idea of graph decomposition and has a iterative flavor.

**Definition 3.2.2** (Decomposable Graph)**.** An undirected marked graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is said to be decomposable if either

1. $\mathcal{G}$ is complete, or

2. there exists a decomposition $(A, B, S)$ of $\mathcal{V}$ into decomposable subgraphs $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$.

The notion of weakly decomposable graph is defined in a completely analogous manner (just replace decomposition by weak decomposition in 2). The following result (Proposition 2.5 in [67]) relates the concepts of weak decomposability and chordless graph.

**Proposition 3.2.1.** *Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be an undirected marked graph. The following conditions are equivalent:*

- *$\mathcal{G}$ is weakly decomposable;*

- *$\mathcal{G}$ is triangulated;*

- *every minimal* $(\alpha, \beta)-$*separator is complete.*

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph. Now, compute decompositions on the graph in an iterative manner. This is: if at least one component resulting from the previous decomposition can be decomposed, then the decomposition is performed. Components that cannot be further decomposed are left static. This procedure ultimately results in what are called the prime components of $\mathcal{G}$. Clearly, the set of vertices shared between two prime components forms a complete subgraph (from the definition of decomposition). See Figure 3.3 for an example. If all the prime components of a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ are complete, then by Proposition 3.2.1 the graph is weakly decomposable.

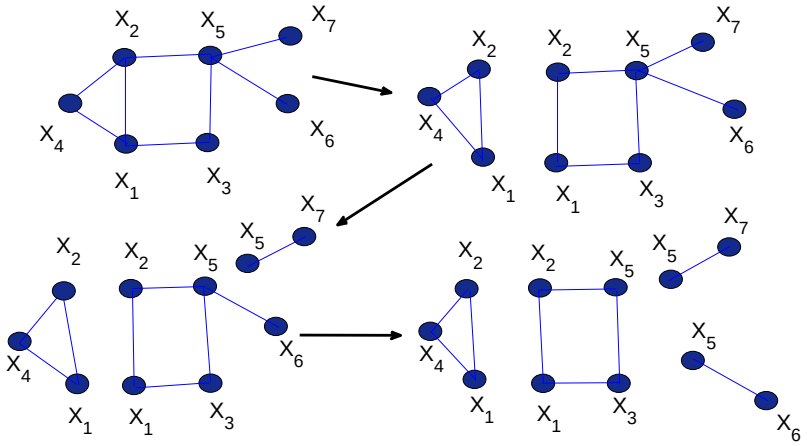


FIGURE 3.3: Here we illustrate the decomposition of a graph into its prime components: First decomposition has $\{1, 2\}$ as separator, the second and third decompositions have $\{5\}$ as separator. Note that all prime components are complete graphs but $\{1, 2, 3, 5\}$.

It is well known that a connected graph can be represented as a tree of its prime components. This construction is called junction tree and it is key for the Graphical

Models framework. The nodes of the tree are the prime components of the graph. The nodes of the tree fulfill what is called the junction property, this is, for any $a$, $b$ nodes in the tree and $c$ in the unique path connecting $a$ and $b$ we have that $a \cap b \subseteq c$. Clearly, if $\mathcal{G}$ is decomposable then the vertices of the tree will be the cliques of the graph. The idea of junction tree is the most useful way to extend the idea of decomposability to hypergraphs. The characterization given in Theorem 2.25 in [67] gives a clear intuitive idea of what is a decomposable hypergraph. It states that a hypergraph is decomposable if and only if it is the clique hypergraph of a weakly decomposable graph.

**Theorem 3.2.1.** *Given that a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is weakly decomposable, it is of interest to know if adding an edge to the graph will maintain weak decomposability. This is if $\mathcal{G}' = (\mathcal{V}, \mathcal{E} \cup \{u, v\})$ is weakly decomposable, with $\{u, v\} \in \mathcal{V} \times V$, $\{u, v\} \notin \mathcal{E}$ and $u \neq v$. Guidici and Green [48] proved that $\mathcal{G}'$ is weakly decomposable iif.*

- $[u]_{\mathcal{G}} \neq [v]_{\mathcal{G}}$.
- $[u]_{\mathcal{G}} = [v]_{\mathcal{G}}$ *and there exists* $R, T \subset V$ *such that* $u \cup R$ *and* $v \cup T$ *are cliques in* $\mathcal{G}$, *and* $S = R \cap T$ *is a separator on the path between* $u \cup R$ *and* $v \cup T$ *in a junction forest representation of* $\mathcal{G}$.

## 3.3 Elements of Graphical Models

The graphical models framework can be described as the use of graph theory tools to encode the conditional independence structure of a random vector. This connection has proven to be useful for computing quantities of interest relative to multivariate

distributions, such as marginal probabilities. Let $\{X_1, X_2, \ldots, X_d\}$ be a random vector and let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph of size $d$. We adopt the convention that every univariate marginal is associated to a node of the graph; this is $v_i \in V$ is associated to $X_i$. The edges of the graph determine how the distribution of $X$ is factorized. The set of vertices associated to discrete variables is denoted by $\Delta$, while the set of vertices associated to continuous marginals is called $\Psi$. Denote by $F$ the distribution on $\{X_1, X_2, \ldots, X_d\}$. There are several conditional independence properties (also called Markov properties) that a distribution $F$ may have with respect to a given graph.

- the pairwise Markov property with respect to $\mathcal{G}$. If for any given pair $(\alpha, \beta)$ of non-adjacent vertices

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

- the global Markov property relative to $\mathcal{G}$, if for any triple $(A, B, S)$ of disjoint subsets of $V$ such that $S$ separates $A$ from $B$ in $\mathcal{G}$.

$$A \perp\!\!\!\perp B \mid S$$

It can be proved (Proposition 3.4 in [67]) that the global Markov property implies the pairwise Markov property.

The following definitions will help to make the idea of correspondence between factorizations and graphs precise. It is said that a distribution $F$ factorizes according to a graph $\mathcal{G}$ if there exist non-negative functions $\psi_a$ such that

$$f(x) = \prod_{a \in \mathcal{C}(\mathcal{G})} \psi_a(x_a), \tag{3.1}$$

where $\mathcal{C}(\mathcal{G})$ denotes the complete sets of the graph. Clearly, the factorization given by Expression 3.1 could be written in several different ways; it depends on how the factors (the $\psi_a$'s) are associated. A common practice is to factorize according to the cliques of the graph, this is:

$$f(x) = \prod_{a \in \mathcal{Q}(\mathcal{G})} \psi_a(x). \tag{3.2}$$

A different approach (which has been increasingly exploited in the Gaussian graphical models literature) is to factorize according to the prime components:

$$f(x) = \prod_{a \in \mathcal{P}(\mathcal{G})} \psi_a(x). \tag{3.3}$$

In this work we will focus on making inferences on models that factorize according to either the prime components (Expression 3.3) or the complete sets of the graph (Expression 3.1); This will be discussed in Chapter 5.

The following result will be used extensively in the following chapters:

**Theorem 3.3.1** (Hammersley and Clifford)**.** *A probability distribution $F$ with positive and continuous density $f$ with respect to a product measure $\mu$ satisfies the pairwise Markov property with respect to an undirected graph $\mathcal{G}$ if and only if factorizes according to $\mathcal{G}$.*

In other words: if a distribution $F$ on $\{X_1, \ldots, X_d\}$ (such that fulfills the assumptions of the theorem) factorizes according to $\mathcal{G}$, then

$$X_i \perp\!\!\!\perp X_j \mid X_{\{1,\ldots,d\}\setminus\{i,j\}} \quad \text{iif } \{i,j\} \notin \mathcal{V} \tag{3.4}$$

The Hammersley and Clifford Theorem is discussed in [51]. In Section 3.2 we defined the concept of a junction tree. One reason why this construction is widely used is that it ensures certain type of coherency when making probability statements involving different marginals. Let us make this notion precise:

**Definition 3.3.1** (Consistency)**.** It is said that distributions $q_A$ over $A$ and $q_B$ over $B$ are consistent if they yield the same distribution over $A \cap B$.

Theorem 2.6 in [26] states that if $\mathcal{G}$ is decomposable, then the junction tree is the unique Markov distribution over $\mathcal{G}$ having the given consistent distributions as its clique marginals.

Assume that a distribution $f$ factorizes according to a graph $\mathcal{G}$. This means that $f$ is restricted by a set of conditional independence statements. Clearly, knowing $\mathcal{G}$ does not determine $f$. In this thesis we assume that the functional form of the terms in the factorization (the $\psi_a(x_a)'$s) is known, but we allow those terms to depend on unspecified parameters (therefore, instead of $\psi_a(x_a)$, we should write $\psi_a(x_a \mid \theta_a)$). This implies that, to estimate a model of the form 3.1 or 3.3, inference on the graph and on the parameters has to be performed simultaneously. We will call this the *estimation problem*. There are applications when knowing the graph structure suffices, therefore the inference procedure identifies each graph with a model. Therefore, inferring the graph becomes a *model selection* problem.

## 3.4 Bayesian inference for graphical models

To perform inferences on a model of the form 3.1 or 3.3 according to the Bayesian approach it is necessary to specify:

1. $p(x \mid \mathcal{G}, \Theta)$ a probability model given the Markov structure and parameters involved in each term of the factorization.

2. $p(\Theta \mid \mathcal{G})$ a prior for the parameters given the graph.

3. $p(\mathcal{G})$ a prior on the space of graphs.

Here $\Theta \mid \mathcal{G} = \{\theta_a : \ a \in \mathcal{C}(\mathcal{G}\}$ (or $\{\theta_a : \ a \in \mathcal{P}(\mathcal{G}\}$, depending on the context). The term $p(\Theta \mid \mathcal{G})$ corresponds to what is called a hyper Markov law. The term was coined by Dawid and Lauritzen [26]. A hyper Markov law can be seen as a probability measure that is concentrated on the probability models that are Markov with respect to $\mathcal{G}$ and such that inherits the conditional independence structure of the sample distribution. Dawid and Lauritzen make this idea precise via the concept of hyperconsistency (Definition 3.2 in [26])

**Definition 3.4.1** (Hyperconsistency)**.** It is said that laws $p_A$ over $A$ and $p_B$ over $B$ are hyperconsistent if they both induce the same law over $A \cap B$.

Note that hyperconsistency is a property of the prior, so statements derived from it are at the parameter level, in contrast with consistency, which is defined at the sampling distribution level. Let $M(\mathcal{G})$ be the set of all probability models that are Markov with respect to $\mathcal{G}$. A law $p(\Theta \mid \mathcal{G})$ on $M(\mathcal{G})$ is said to be hyper Markov over

$\mathcal{G}$ if for any decomposition $(A, B)$ of $\mathcal{G}$

$$\theta_A \perp\!\!\!\perp \theta_B \mid \theta_{A \cap B} \tag{3.5}$$

Dawid and Lauritzen use the junction tree decomposition to prove that for decomposable models there is a unique hyper Markov law that is hyper consistent over clique marginals (Theorem 3.9 in [26]). For the case where the sampling distribution is Gaussian, the corresponding hyper Markov law is the hyper Inverse Wishart.

The usual choices for $p(\mathcal{G})$ are: the uniform distribution on the space of decomposable graphs with $d$ vertices, or an Erdös-Rényi random graph with single edge inclusion probability. The later is a graph $(\mathcal{V}, \mathcal{E})$ where each of the $\binom{d}{2}$ possible edges $(i, j)$ is included in $\mathcal{E}$ independently with some specified probability $p \in [0, 1]$. Therefore, once the distribution of edge counts has been fixed, the distribution of any other graph feature (cliques of size 3, for example) is already determined. In order to obtain more flexible priors, variations where the edge inclusion probabilities $p_{ij}$ are allowed to be edge-specific, have been proposed [54],[72].

The strategy proposed at the beginning of this section (this is, to use priors of the form $p(\Theta \mid \mathcal{G}) \times p(\mathcal{G})$) was conceived to deal with the problem of estimating the model; this is, to infer the graph and the parameters involved in the terms of the factorization. The work of Giudici and Green [48] is an elegant example of this. If the inference of interest is model selection (Chapter 7 of [90]), then the problem turns into sampling directly from the posterior of $\mathcal{G}$. This implies working with

$$\Pr\{\mathcal{G} \mid x\} \propto \int_{M(\mathcal{G})} p(x \mid \Theta, \mathcal{G}) p(\Theta \mid \mathcal{G}) p(\mathcal{G}) d\Theta, \tag{3.6}$$

which is known as the marginal likelihood (see pg 348 of [90]).

One of the most studied problems in the graphical models literature is to infer the Markov structure of a multivariate normal distribution. This implies defining a suitable hyper Markov law and computing the corresponding marginal likelihood. We review those results in the following section.

## 3.5 Gaussian Graphical Models

Assume that $X \sim \text{MVN}(0, \Sigma)$, and let $K = \Sigma^{-1}$, that is $K$ is the precision matrix. It holds that $x_i \perp\!\!\!\perp x_j \mid x_{\{1,\ldots,d\}\setminus(i,j)}$ if and only if $k_{i,j} = 0$, where $k_{i,j}$ is the $(i,j)-$entry of $K$ (See Section 5.1.3 of [67]). This implies that the distribution of $X$ is Markov (in the sense of the Hammersley and Clifford Theorem) with respect to a graph such that edges are omitted only for those pairs where the corresponding entry of the precision matrix is zero. Therefore, inferring the Markov structure for a multivariate normal is equivalent to making inferences about the pattern of zeros in the precision matrix. This setting is called covariance selection. The term was coined by Dempster [29].

Now we will present some key facts about the hyper Markov law for the problem of covariance selection. We follow closely the presentation by [3]. Let $x$ be a single observation from a multivariate normal with mean 0 and precision matrix $K$:

$$p(x \mid K, \mathcal{G}) = \frac{(K)^{\frac{n}{2}}}{(2\pi)^{\frac{nd}{2}}} \exp\left(-\frac{1}{2}\langle K, x^t x\rangle\right). \tag{3.7}$$

Note that we are conditioning with respect to $\mathcal{G}$, this means that some entries of $K$

may equal to zero. The conjugate prior for $K$ is of the one proposed by [30]:

$$p(K \mid \mathcal{G}) = \frac{1}{I_{\mathcal{G}}(\delta, D)} |K|^{\frac{\delta-2}{2}} \exp\left(-\frac{1}{2}\langle K, D\rangle\right). \tag{3.8}$$

This is called a $\mathcal{G}$-Wishart with shape parameter $\delta$ (a scalar greater than zero) and inverse scale matrix $D$. Note that the support of this distribution is the set of all positive definite matrices that respect the restriction (pattern of zeros) imposed by $\mathcal{G}$, this set is denoted by $M^+(\mathcal{G})$. The normalizing constant of this distribution is given by:

$$I_{\mathcal{G}}(\delta, D) = \int_{M^+(\mathcal{G})} |K|^{\frac{\delta-2}{2}} \exp\left(-\frac{1}{2}\langle K, D\rangle\right) dK. \tag{3.9}$$

Clearly $\delta$ and $D$ must be such that the integral in Expression 3.9 is finite. The normalizing constant of the $\mathcal{G}$-Wishart is crucial for Bayesian inference, since the marginal likelihood for $\mathcal{G}$ is

$$p(x \mid \mathcal{G}) = \frac{1}{(2\pi)^{\frac{nd}{2}}} \frac{I_{\mathcal{G}}(\delta + n, D + x^t x)}{I_{\mathcal{G}}(\delta, D)}. \tag{3.10}$$

Here $n$ is the number of observations. In the case that $\mathcal{G}$ is complete, the normalizing constant can be expressed in close form:

$$I_{\mathcal{G}}(\delta, D) = \frac{2^{\frac{nd}{2}} \Gamma_d\left(\frac{\delta-d+1}{2}\right)}{|D|^{\frac{\delta+d-1}{2}}}, \tag{3.11}$$

where $\Gamma_d(\cdot)$ is the multivariate Gamma function:

$$\Gamma_d(a) = \pi^{\frac{d(d-1)}{4}} \prod_{i=0}^{d-1} \Gamma(a - \frac{i}{2}). \tag{3.12}$$

The probability model for $\Sigma$ implied by expression 3.7 is the Inverse Wishart:

$$\mathrm{IW}(\Sigma \mid \delta, D) = \frac{1}{I_{\mathcal{G}}(\delta, D)} |\Sigma|^{-\frac{\delta-2}{2}} \exp -\frac{1}{2}\langle \Sigma^{-1}, D\rangle. \tag{3.13}$$

Let $\mathcal{G}$ be a graph and consider a junction tree decomposition in terms of its prime components (this is, we are not assuming decomposability). The Hyper Inverse Wishart for $\Sigma$, where $\text{MVN}(0, \Sigma)$ is Markov with respect to $\mathcal{G}$ is given by:

$$\text{HIW}_{\mathcal{G}}(\Sigma \mid \delta, D) = \frac{\prod_{j=1}^{k} \text{IW}(\Sigma_{P_j} \mid \delta, D_{P_j})}{\prod_{j=2}^{k} \text{IW}(\Sigma_{S_j} \mid \delta, D_{S_j})}, \tag{3.14}$$

Where the $P_j$'s are the prime components and $S_j'$s are the separators in the junction tree representation. The normalizing constant of this distribution is:

$$I_{\mathcal{G}}(\delta, D) = \frac{\prod_{j=1}^{k} I_{\mathcal{G}_{P_j}}(\delta, D_{P_j})}{\prod_{j=2}^{k} I_{\mathcal{G}_{S_j}}(\delta, D_{S_j})} \tag{3.15}$$

If $\mathcal{G}$ is decomposable 3.15 can be computed in close form. If one of the prime components is not complete, the corresponding normalizing constant can be obtained via simulation;[27], [93] and [3] have proposed methods for doing this. We adopted the method in [3], which takes advantage of the Choleski decomposition and it is straightforward to implement (Section 4.2 of [3]).

The hyper inverse Wishart is an example of an hyperconsistent law; this is because for any pair of neighbors in the prime component decomposition, the corresponding marginals of the HIW agree on the marginal distribution for the entries of the covariance matrix associated to the separator.

## 3.6 Decomposable *vs.* Non Decomposable Models

Most of the literature on undirected graphical models focuses on the decomposable case, there are good reasons for this:

- In a decomposable model the $\psi(x_a)'s$ correspond to marginal distributions. In a non decomposable models these functions are potentials (See [63]).

- Close form of maximum likelihood estimates are available for decomposable models (Sections 5.3.2 and 6.3.2 of [67]). Currently there are no results of that sort for the non-decomposable case.

- Junction trees for decomposable models are consistent (Theorem 2.6 in [26]). There is no guarantee of consistency for the non-decomposable case.

Still non-decomposable models are worth of consideration. For instance [93] and [3] proposed methodology for non-decomposable Gaussian models, they tested it for the Fisher's iris data and observed that the model with highest posterior probability was the four-cycle, which is non-decomposable.

The methodology proposed in this thesis is able to make inferences on either decomposable or non decomposable models. We make the distinction since each case poses different computational challenges, also because all the references on graphical models cited in this work specialize in either one case or the other. Although we provide a good amout of background on Gaussian Graphical models, we are also interested in more general examples; these are discussed on Chapter 6.

# 4

# Geometric Graphs and Nerves

The objective of this chapter is to provide the concepts needed for defining parametrizations on graph spaces and priors suitable to such parametrizations. In Section 4.1 we define the concept of nerve and use it to propose parametrizations for spaces of graphs. We relate these ideas with the notion of weak decomposability in Section 4.2. Basic ideas on Random Geometric Graphs are developed in Section 4.3.

## 4.1 Nerves

In the previous chapter, a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ was defined as an ordered pair consisting of a set $\mathcal{V}$ of objects called vertices and a collection $\mathcal{E}$ of unordered pairs of those objects, the edges of the graph. Note that this definition does not consider any topology or metric on the elements of $\mathcal{V}$. The core idea of this chapter is that, by regarding $\mathcal{V}$ as a subset of a metric space, efficient parametrizations of graph spaces can be obtained and flexible probability measures on such spaces can be defined. By

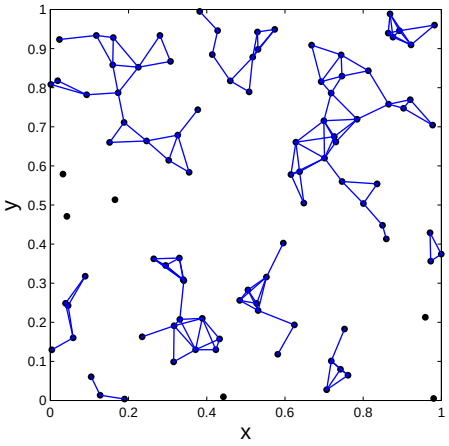


FIGURE 4.1: Proximity graph with 100 vertices and $r = 0.05$.

flexibility we mean having simultaneous control over 2 graph features (*e.g.* subgraph counts).

We first consider this idea in its simplest form: Let $A$ be a region in $\mathbb{R}^m$ and choose $\mathcal{V}$ a finite set of points in such region, now join every unordered pair of elements in $\mathcal{V}$ by an edge if the distance between them is less or equal than some pre-specified value $2r$ (see Figure 4.1); this construction is known as the proximity graph. Observe that:

- To compute the graph it is only necessary to specify $\mathcal{V}$ and $r$ and that the metric of $\mathbb{R}^m$ is explicitly used. This gives a low-dimensional representation of a space of graphs, which has dimension $m|\mathcal{V}| + 1$.

- Not all graphs with $|\mathcal{V}|$ vertices can be represented as a proximity graph. Think of a star graph with one vertex of degree $s$ and the other $s-1$ vertices of degree 1. For $m = 2$ this graph can be represented as a proximity graph for $s \leq 6$ but

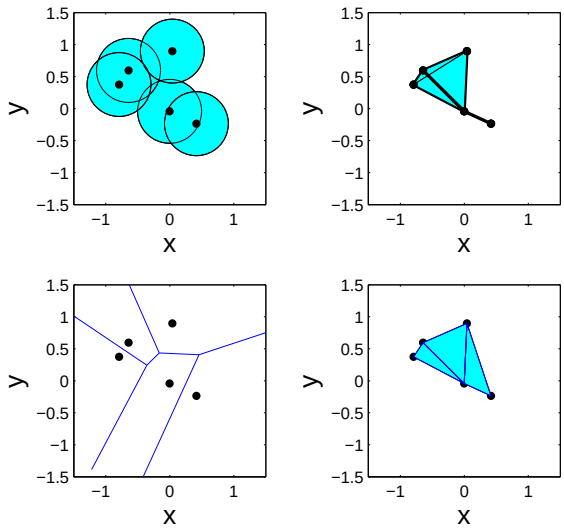


FIGURE 4.2: Given a set of vertices in $\mathbb{R}^2$ and a radius ($r = 0.5$) a family of disks is generated (top left) and its nerve (top right) can be computed. This is an example of a Čech complex. For the same vertex set, the Voronoi diagram is computed (bottom left) and the nerve of the Voronoi cells is obtained (bottom right). This is an example of the Delaunay triangulation. Note that the maximum clique size of the Delaunay is bounded the dimension of the space where the vertex set lies plus one; such restriction does not apply to the Čech complex.

not for $s \geq 7$.

The proximity graph is a particular case of a more general construction which is illustrated in Figure 4.2; the main point is that a graph (or a hypergraph) can be computed from the intersection pattern of a family of subsets in Euclidean space. The methodology developed in this thesis exploits several aspects of this idea. The key concept is the one of nerve:

**Definition 4.1.1** (Nerve)**.** Let $F = \{A_j,\ j \in I\}$ be a finite collection of nonempty

convex sets. The *nerve* of $F$ is given by

$$\mathrm{Nrv}F = \{\sigma \subseteq I \mid \cap_{j\in\sigma} A_j \neq \varnothing\},$$

this is, the sets of indices for which the intersections of $A_j$'s are non empty. Since the objectives of this work concern statistical modeling and Bayesian simulation, we will focus our efforts to families of subsets that are simple to parametrize and efficient to compute. In particular we will work with sets that can be indexed by a center $v$ and a radius $r$; the generic set of this type will be denoted by $A(v, r)$. The nerves that will be used in the subsequent sections of the chapter are now defined:

**Definition 4.1.2** (Čech Complex)**.** Let $\mathcal{V}$ be a finite set of points in $\mathbb{R}^m$ and $r > 0$. The *Čech complex* corresponding to $\mathcal{V}$ and $r$ is the nerve of the sets $B_v = v + r\mathbb{B}_k$, $v \in \mathcal{V}$. This is denoted by $\check{\mathrm{C}}\mathrm{ech}(\mathcal{V}, r)$.

**Definition 4.1.3** (Delaunay Triangulation)**.** Let $\mathcal{V}$ be a finite set of points in $\mathbb{R}^m$. The *Delaunay triangulation* corresponding to $\mathcal{V}$ is the nerve of the sets

$$C_v = \{x \in \mathbb{R}^m \mid \|x - v\| \leq \|x - u\|,\ u \in \mathcal{V}\}, \quad v \in \mathcal{V}.$$

This is denoted by $\mathrm{Delaunay}(\mathcal{V})$. The sets $C_v$ are called *Voronoi cells*.

**Definition 4.1.4** (Alpha Complex)**.** Let $\mathcal{V}$ be a finite set of points in $\mathbb{R}^m$ and $r > 0$. The *Alpha complex* corresponding to $\mathcal{V}$ and $r$ is the nerve of the sets $B_v \cap C_v$, $v \in \mathcal{V}$. This is denoted by $\mathrm{Alpha}(\mathcal{V}, r)$.

The nerve of a family of sets is a hypergraph, more specifically it is a particular case of a class of hypergraphs known as simplicial complexes:

**Definition 4.1.5** (Simplicial Complex)**.** Let $\mathcal{V}$ be a finite set, a *Simplicial Complex* with base set $\mathcal{V}$ is a family of subsets $\mathcal{K}$ of $\mathcal{V}$ such that $\tau \in \mathcal{K}$ and $\zeta \subseteq \tau$ implies $\zeta \in \mathcal{K}$. The elements of $\mathcal{K}$ are called simplices.

The notion of simplicial complex is a fundamental one in the field of Computational Topology. The idea of a filtration (which is the object of the next section) is based upon this concept. Most of the applications in this work concern graphs; There is a graph associated to each nerve, to make this precise we introduce the following concept:

**Definition 4.1.6** ($p$-Skeleton)**.** Let $\mathcal{K}$ be a simplicial complex. The $p-$skeleton of $\mathcal{K}$ is the collection of all $\tau \in \mathcal{K}$ such that $|\tau| \leq p+1$. The elements of the $p-$skeleton are called $p-$simplices.

In particular the $1-$skeleton is a graph. Clearly the $1-$skeleton of a nerve is the graph that is obtained by keeping track of nonempty pairwise intersections of convex sets. The process of obtaining the nerve and the $1-$skeleton from a family of sets is illustrated in Figure 4.3. Observe that different families of convex sets induce different restrictions in graph space: in the case of the Delaunay triangulation and the Alpha complex there is a cap for the clique size equal to $m+1$. For the Čech complex that restriction does not apply.

The Čech and the alpha complex can be thought as hypergraphs that are indexed in terms of a finite set of points $\{V_1, \ldots, V_d\}$ and a size parameter $r$. This automatically defines an indexing on the $1-$skeleton of the nerves. In other words: this construction induces a parametrization of the space of graphs $(\mathcal{V}, r) \to \mathcal{G}(\mathcal{V}, r)$.

Clearly, one can also parametrize spaces of hypergraphs $(\mathcal{V}, r) \to \mathcal{H}(\mathcal{V}, r)$, by working with the nerve directly. Since the graph depends also on the class of sets used to compute the nerve, which we denote by $\mathcal{A}$, one could write $\mathcal{G}(\mathcal{V}, r, \mathcal{A})$. $\mathcal{A} = \check{\text{C}}\text{ech}$ means that the sets used to compute the nerve are $A(v, r) = v + r\mathbb{B}_m$ while $\mathcal{A} = \text{Alpha}$ means that the sets $A(v, r) = (v + r\mathbb{B}_m) \cap C_v$ will be used instead. To keep the notation simple, $\mathcal{A}$ will be omitted whenever obvious by the context.

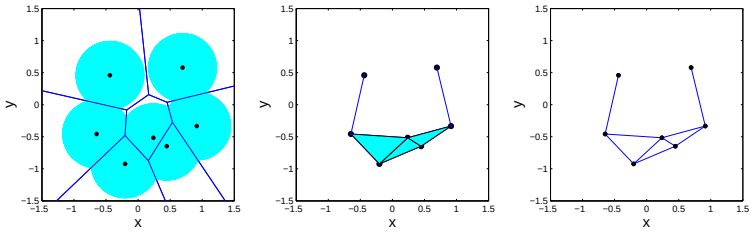


FIGURE 4.3: Given a set of vertices and a radius ($r = 0.5$) one can compute $A_i = C_i \cap B_i$, where $C_i$ is the Voronoi cell for vertex $i$ and $B_i$ is the ball of radius $r$ centered at vertex $i$ (left). The Alpha complex is the nerve of the $A_i$'s (center). Often the main interest will be the $1-$skeleton of the complex, which is the subset of the nerve that corresponds to (nonempty) pairwise intersections (right).

## 4.2 Filtrations

The intuitive idea behind a filtration is the following: Given a vertex set $\mathcal{V}$ and a radius $r$ a simplicial complex (more specifically a nerve) can be obtained by computing the nerve for $\mathcal{V}$ and $s$, where $s$ increases from 0 to $r$. By doing this, sets are added to the complex and never removed; which is obvious since $\cap_{j=1}^{m} A(V_j, s_1) = \varnothing$ implies $\cap_{j=1}^{m} A(V_j, s_2) = \varnothing$ for all $s_2 > s_1 \geq 0$. A filtration can be seen as a structure that keeps track of this process.

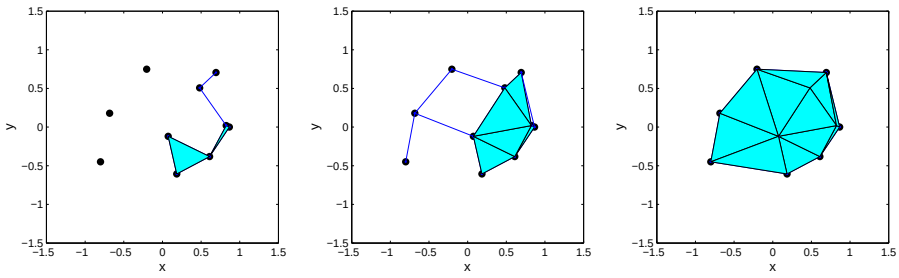

FIGURE 4.4: Filtration of Alpha complexes, here $r = 0.31$ (left), $r = 0.45$ (center) and $r = 0.86$ (right).

**Definition 4.2.1** (Filtration)**.** A *filtration* for a simplicial complex $\mathcal{K}$ is a collection $\mathcal{L} = \left\{\mathcal{K}^0, \mathcal{K}^1, \ldots, \mathcal{K}^l\right\}$ of simplicial complexes such that

$$\varnothing = \mathcal{K}^0 \subset \mathcal{K}^1 \subset \ldots \subset \mathcal{K}^l = \mathcal{K}.$$

The inclusions are proper.

A particular case of a filtration that we will use in this and the following chapters is the following:

**Definition 4.2.2** (Alpha Shape)**.** A filtration for the Alpha Complex with parameters $\mathcal{V}$ and $r$ is called an Alpha Shape.

One way filtrations will be useful for the applications discussed in this thesis is that they provide Algorithms for computing specific nerves (that is the case for the Alpha complex [36]). A second way to take advantage of this construction is to relate the concept of simplicial complex to the one of decomposability. As illustrated in Figures 4.1 and 4.4, the $1-$Skeleton of a Čech or an Alpha complex does not have to be a weakly decomposable graph. Since a good amount of literature on Graphical

Models focuses on Markov structures derived from weakly decomposable graphs, it is desirable to have a procedure that guarantees that the graphs produced by a nerve are weakly decomposable. That is the objective of Algorithm 1.

The input to the procedure is a filtration $\mathcal{L} = \left\{\mathcal{K}^0, \mathcal{K}^1, \ldots, \mathcal{K}^l\right\}$. This filtration can be constructed using either $\mathcal{A} =$ Čech or $\mathcal{A} =$ Alpha. The output of the procedure will be a weakly decomposable graph $\mathcal{G}$. Since the simplices are added in a hierarchical manner ($\tau$ cannot be added before $\kappa$ if $\kappa \subset \tau$), the $1-$skeleton of every simplex $\tau$ is a added before $\tau$. Because of this observation and the fact that weak decomposability is a property of the $1-$skeleton, we only have to consider only the elements of the filtration where a $1-$simplex is added. The initial value of the procedure is $\mathcal{K}^0$ the empty graph, which is weakly decomposable, then the question is: given that a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is weakly decomposable, how to ensure that the graph obtained by adding an edge $(v_i, v_j)$ is also weakly decomposable? The answer is given by the criterion proposed by Giudici and Green (Theorem 2 in [48]). We first check if adding $\tau_j$ reduces the number of connected components, if this is the case, then the simplicial complex $\mathcal{K}^{j+1}$ is a weakly decomposable graph and $\tau_j$ is added to the graph $\mathcal{G}$. If the addition of $\tau_j$ does not reduce the number of connected components we need to make sure that there are no chordless cycles in the new simplex $\mathcal{K}^{j+1}$. This question is answered by examining the cliques and separators that include the vertexes of $\tau_j$. The $1-$ simplexes $\tau_j$ are added until $\mathcal{K}^l$ is examined. $\mathcal{G}$ is weakly decomposable by construction. Algorithm 1 outlines the above steps.

**Proposition 4.2.1.** *The graph $\mathcal{G}$ produced by the above algorithm is weakly decom-*

**Algorithm 1**: Weakly decomposable graph from filtration with $\sharp(\mathcal{B})$ defined as the number of connected components of the simplicial complex $\mathcal{B}$.

**input** : a filtration $\mathcal{L} = \{\mathcal{K}^0, \mathcal{K}^1, \ldots, \mathcal{K}^l\}$
**return**: a weakly decomposable graph $\mathcal{G}$
$\mathcal{G} = \varnothing$;
**for** $j = 1$ **to** $m$ **do** $\mathcal{A}^j = \mathcal{K}^j$; `// initialize the elements of the filtration`
$i = 0$;
**while** $\mathcal{A}^{i+1}$ *is nonempty and* $i < l$ **do**
  $\tau_i = \{\kappa \in \mathcal{A}^{i+1} - \mathcal{A}^i \mid \dim(\kappa) = 1\}$; `// the 1-simplex added`
  **if** $\sharp(\mathcal{A}^{i+1}) < \sharp(\mathcal{A}^{i+1})$ *or* $\tau_i = [v_1, v_2]$ *is such that* $v_k \in \gamma_k$ *(*$k \in \{1, 2\}$*) cliques and* $\gamma_1 \cap \gamma_2$ *is a separator of on the path between* $\gamma_1$ *and* $\gamma_2$*.* **then**
    add $\tau_i$ to $\mathcal{G}$;
  `// no chordless cycle is created`
  i=i+1;

*posable.*

*Proof.* The algorithm is initialized with the empty graph and weak decomposability is tested every time one tries to add an edge (*i.e.* a $1-$simplex in $\mathcal{L}$) to the graph. Since there are only finitely many edges that could be added, $\mathcal{G}$ is weakly decomposable by construction. □

Figure 4.5 illustrates the output of Algorithm 1 for a Čech filtration corresponding to 100 points sampled uniformly from the unit square and $r = 0.05$. The $1-$skeleton of the Čech complex for $r = 0.05$ is also shown. One observes that few edges were deleted from the original graph. This is because geometric graphs tend to be triangulated, in the sense that if $\{v_1, v_2\}$ and $\{v_2, v_3\}$ belong to a geometric graph, then very likely $\{v_1, v_3\}$ will also be in the graph.

Consider the vertex set displayed in Table 4.1 and apply Algorithm 1 for $\mathcal{A} =$ Čech

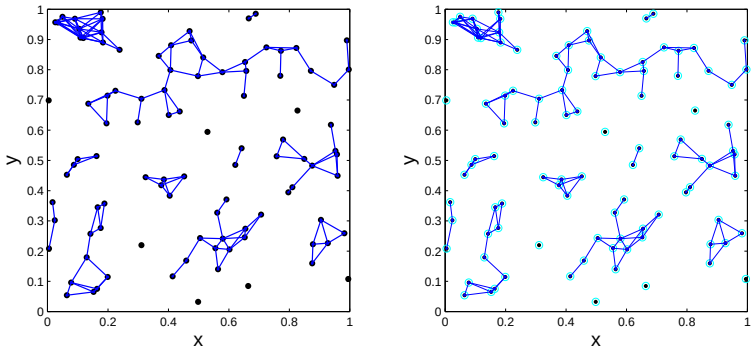


FIGURE 4.5: Proximity graph and decomposable graph for same vertex set and $r = 0.05$.

| Coordinate | $V_1$ | $V_2$ | $V_3$ | $V_4$ | $V_5$ |
|---|---|---|---|---|---|
| $x$ | 0.686 | 0.214 | 0.846 | 0.411 | 0.089 |
| $y$ | 0.151 | 0.194 | 0.420 | 0.567 | 0.553 |

Table 4.1: Vertex set used to illustrate Algorithm 1. See Table 4.2.

and $r = 0.5$. It follows that (Table 4.2) the first 3 edges in the filtration will be added, because by doing so, the number of connected components is reduced. The edge $\{2, 4\}$ is added to $\mathcal{G}$ since the intersection of $\{2, 5\}$ and $\{4, 5\}$ is $\{5\}$, a separator in the junction tree representation of $\mathcal{G}$. $\{3, 4\}$ reduces the number of connected components, then it is added to the edge set of $\mathcal{G}$. The next edge in the filtration is $\{1, 2\}$, which cannot be added, since the intersection of $\{1, 3\}$ and $\{2, 4, 5\}$ is empty, and therefore not a separator. Finally $\{1, 4\}$ is added since $\{3\}$ is a separator (See Figure 4.6).

| Cliques | Separators | $r$ | Update |
|---|---|---|---|
| $\{1\}\,\{2\}\,\{3\}\,\{4\}\,\{5\}$ | - | 0 | - |
| $\{1,3\}\,\{2\}\,\{4\}\,\{5\}$ | - | 0.313 | $\{1,3\}$ |
| $\{1,3\}\,\{2\}\,\{4,5\}$ | - | 0.321 | $\{4,5\}$ |
| $\{1,3\}\,\{2,5\}\,\{4,5\}$ | $\{5\}$ | 0.379 | $\{2,5\}$ |
| $\{1,3\}\,\{2,4,5\}$ | - | 0.421 | $\{2,4\}$ |
| $\{1,3\}\,\{3,4\}\,\{2,4,5\}$ | $\{3\}\,\{4\}$ | 0.459 | $\{3,4\}$ |
| $\{1,3\}\,\{3,4\}\,\{2,4,5\}$ | $\{3\}\,\{4\}$ | 0.474 | $\sim\{1,2\}$ |
| $\{1,3,4\}\,\{2,4,5\}$ | $\{4\}$ | 0.498 | $\{1,4\}$ |

Table 4.2: Evolution of cliques and separators in the junction tree representation of $\mathcal{G}$ as edges are added according to Algorithm 1. The edge $\{1,2\}$ is left out of the graph.

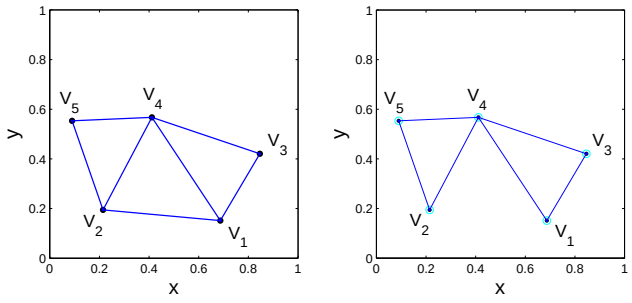


FIGURE 4.6: Proximity graph and decomposable graph computed from the vertex set given in Table 4.1. The edge $\{1,2\}$ is not included in the graph.

## 4.3 Random Geometric Graphs

### *4.3.1 General Setting*

In Section 4.1 we defined a parametrization $(\mathcal{V}, r) \to \mathcal{G}(\mathcal{V}, r)$ for a space of graphs. One implication of such parametrization is that, given the family of sets to compute the nerve $\mathcal{A}$, a probability distribution on $(\mathcal{V}, r)$ automatically induces a probabil-

ity distribution on $\mathcal{G}(\mathcal{V}, r, \mathcal{A})$. For this reason, this parametrization is attractive for Bayesian inference since placing prior distributions on $(\mathcal{V}, r)$ should be, in principle, more straightforward than placing priors on graphs. This leads to the following question: which distributions on $(\mathcal{V}, r)$ would induce meaningful distributions on $\mathcal{G}(\mathcal{V}, r, \mathcal{A})$? From a statistical modeling perspective it is desirable to have insight about the distribution of key features of the graph while allowing for certain flexibility. To address these issues we rely on the following construction.

**Definition 4.3.1** (Random Geometric Graph)**.** Given a class $\mathcal{A}$ of convex sets in $\mathbb{R}^m$ of the form $A(v, r)$ (see Section 4.1) and $r > 0$, let $\mathbf{V} = (V_1, \ldots, V_d)$ be sampled according to a $(m \times d)-$dimensional distribution $\mathcal{Q}$ (there are $d$ vertices in $\mathbb{R}^m$). The graph $\mathcal{G}(\mathbf{V}, r, \mathcal{A})$ is said to be a *Random Geometric Graph* (RGG).

It is clear that different choices for $\mathcal{A}$, $\mathcal{Q}$ and $r$ will have an impact on the support of the implied RGG distribution. To make this notion precise we define:

**Definition 4.3.2** (Feasible Graph)**.** Given a class $\mathcal{A}$ of convex sets in $\mathbb{R}^m$ and $r > 0$, and a distribution $\mathcal{Q}$. A graph $\Gamma$ is said to be feasible if

$$\Pr\{\mathcal{G}(\mathbf{V}, r, \mathcal{A}) \cong \Gamma\} > 0, \quad \mathbf{V} \sim \mathcal{Q}.$$

While Definition 4.3.1 is more general than the one used in [84], it still does not comprehend all the random graphs discussed in [85], just the ones that are of use in our methodology. Unless stated otherwise, it will be assumed that $\mathbf{V} = \{V_1, \ldots, V_d\}$ is sampled from a distribution $\mathcal{Q}$ that has density with respect to Lebesgue measure; such density will be denoted by $q$.

In contrast to Erdös-Rényi models where the inclusion of two different edges in the graph are independent events. The RGG models imply a dependence structure for the edge inclusions mainly due to the metric in $\mathbb{R}^m$ (if the vertex $x$ is adjacent to $y$ and the later to $z$, with high probability $x$ and $z$ will be adjacent) and the type of sets used to compute the nerve (for the Delaunay triangulation the existence of an edge between two vertices does not depend on the position of those vertices only, but on the position of all other vertices). While exact computation of the distribution of a specific graph feature for a given $d$ may be cumbersome, there exists a good (for certain distributions $\mathcal{Q}$) amount of asymptotic results for most quantities of interest. Such quantities include: subgraph counts (this comprehends edge counts, which encode sparsity), typical vertex degree, order of the largest clique, and maximum vertex degree among others. The monograph [84] is an encyclopedic work of results developed explicitly for the $1-$skeleton of the Čech complex. Several results for the Delaunay triangulation are available [85]. Results specific to the Alpha complex have not been developed in the literature, but the $1-$skeleton of the Alpha complex fits in the general framework presented in [85].

### *4.3.2 Subgraph Counts*

Results regarding subgraph counts are of special interest because they give insight about the distribution of the number of edges in the graph, which encodes sparsity. In addition they provide information about the distribution of the number of cliques of a given size. In this section we discuss some key results regarding subgraph counts and present simulations in order to gain understanding about the regimes for which

the approximations can be useful. Most results focus on the proximity graph (i.e., the 1−skeleton of the Čech complex).

Unless stated otherwise, we will work on the following setting: $\mathbf{V} = (V_1, \ldots, V_d)$ is sampled from $\mathcal{Q}$, which has a density, denoted by $q$. It will be assumed that the components of $\mathbf{V}$ are independent and with same marginal distribution, which we denote by $q_{\text{marg}}$. Remember that each $V_i$ is an element of $\mathbb{R}^m$. The radius parameter $r$ is regarded as fixed and $\mathcal{G}(\mathbf{V}, r)$ refers to the proximity graph.

Let $\Gamma$ be a graph of order $s$, define

$$h_\Gamma(\mathbf{V}) = \mathbf{1}_{\{\mathcal{G}(\mathbf{V},1)\cong\Gamma\}}. \tag{4.1}$$

This r.v. keeps track of the event when the proximity graph with $r = 1$ is isomorphic to $\Gamma$. Let $D$ be an open set that contains the origin; it will be assumed that $\text{Leb}(\partial D) = 0$. Now set

$$\mu_{\Gamma,D} = s!^{-1} \int_D q_{\text{marg}}(x)^s dx \int_{(\mathbb{R}^m)^{s-1}} h_\Gamma(\{\mathbf{0}, x_1, \ldots, x_{s-1}\}) d(x_1, \ldots, x_{s-1}) \tag{4.2}$$

The second integral in Expression 4.2 is the volume of the $m(k-1)$−dimensional set (we are talking about the coordinates of $(k-1)$ vertices) such that the proximity graph of radius 1 is isomorphic to $\Gamma$. Here one of the vertices is placed at the origin. Now think that the vertex at the origin could be translated to any point in $D$, and that point is picked by sampling according to $q_{\text{marg}}$; that is the meaning of the first integral. The factor $s!^{-1}$ takes into account the fact that any of the $k$ vertices can be set at $\mathbf{0}$.

**Example 4.3.1.** *Let $q_{marg}$ be the uniform distribution on the disk of radius 3, set $D$ as the disk of radius 2 and $\Gamma = K_2$ (the complete graph with 2 vertices). This is, we are looking at the edge count of the graph (Figure 4.7). Clearly $h_\Gamma(\mathbf{0}, x_1) = 1$ if and only if $\|x_1\| \leq 1$, therefore the value of the second integral in Expression 4.2 is $\pi$; the area of the disk of radius 1. The value of the first integral is*

$$\int_D q_{marg}(x)^s dx = (2^2\pi) \times \left(\frac{1}{3^2\pi}\right)^2.$$

*Therefore $\mu_{\Gamma,D} = \frac{1}{2!} \times \frac{4}{81\pi} \times \pi = \frac{2}{81}$.*

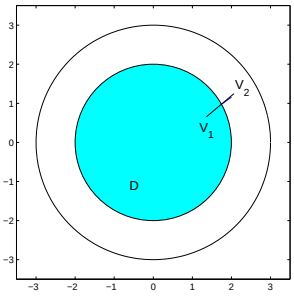


FIGURE 4.7: In Example 4.3.1 we set the support of $q_{\text{marg}}$ as the disk of radius 3 and $D$ as the disk of radius 2. One of the vertices of $\Gamma$ ($K_2$ in this example) is sampled according to $q_{\text{marg}}$ restricted to $D$.

We follow the notation of [84] and define $G_{n,D}(\Gamma)$ as the number of subgraphs in $\mathcal{G}(\mathbf{V}, r)$ that are isomorphic to $\Gamma$ and such that its left most point (according to lexicographic ordering in $\mathbb{R}^m$) is in $D$. Now we can state:

**Proposition 4.3.1.** *Assume that $\Gamma$ is a feasible connected graph $(\mathcal{V}(\Gamma), \mathcal{E}(\Gamma))$ with $|\mathcal{V}(\Gamma)| = k \geq 2$ and let $D \subseteq \mathbb{R}^m$ be open with $Leb(\partial D) = 0$. Consider a sequence of*

*positive numbers* $\{r_n\}_{n\geq 1}$ *such that* $\lim_{n\to\infty}(r_n) = 0$. *Then*

$$\lim_{n\to\infty} r_n^{-m(k-1)} n^{-k} E[G_{n,D}(\Gamma)] = \mu_{\Gamma,D}. \tag{4.3}$$

Result 4.3.1 is useful because once we know for which $n_0$ the asymptotics kick in (this can be investigated via simulation) a very simple computation will provide an estimate for $E[G_{n,D}(\Gamma)]$ for $n \geq n_0$.

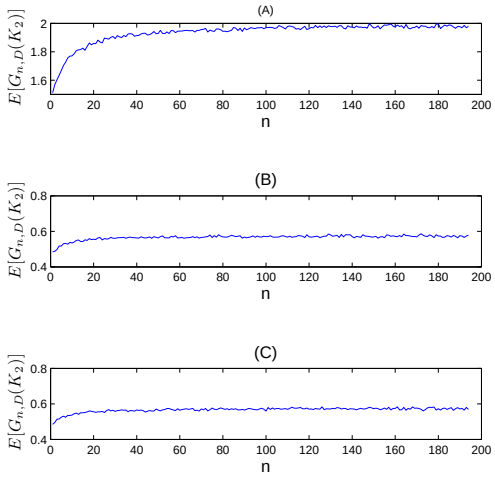


FIGURE 4.8: $E[G_{n,D}(K_2)]$ as a function of $n$. Each vertex of $\mathcal{G}(\mathbf{V}, r_n)$ is sampled according to: (A) uniform on $[0,1]^2$, (B) a multivariate normal $Y$ with mean $\mathbf{0}$ and $\sigma_1^2 = 1$, $\sigma_2^2 = 3$, $\sigma_{1,2}^2 = 1.5$, (C) a mixture of multivariate normals, distributed as $Y$, $Y + (2,0)$, $CY + (-2,2)$, where $C$ is the rotation matrix corresponding to $\pi$ radians; all elements in the mixture are sampled with equal probability. $E[G_{n,D}(K_2)]$ was estimated using $25,000$ simulations for each $n$.

In Figure 4.8 we illustrate the asymptotic behavior of $E[G_{n,D}(\Gamma)]$ for $\Gamma = K_2$; this is, we are looking at the expected number of edges as a function of the number of vertices in the graph. We considered 3 options for $q_{\text{marg}}$: Uniform in $\mathbb{B}_2$, Multivariate normal in $\mathbb{R}^2$ and a mixture of 3 Multivariate normals in $\mathbb{R}^2$. This was

done for the sequence $r_n = \frac{1}{n}$, $n \geq 1$. For these choices $r_n^{-m(k-1)} n^{-k} = 1$, therefore $\lim_{n\to\infty} E[G_{n,D}(\Gamma)] = \mu_{\Gamma,D}$. Observe that the expected number of edges stabilizes fairly quickly for the three regimes; but the speed of convergence is sensitive to the distribution for sampling the vertices. In this setting it is clear how to control $E[G_{n,D}(\Gamma)]$ for large $n$ so it is approximately equal to some desired value $a$: it is enough to multiply the sequence $\{r_n\}_{n\geq 1}$ by a constant (namely $a^{\frac{-1}{m}}$).

While having an estimate of the expected number of subgraphs of certain type may give some insight about how to calibrate a prior, it is far more useful to have an approximation of the subgraph count distribution. The following result provides this information. The following theorem is discussed in pg. 52 of [84].

**Theorem 4.3.1** (Penrose)**.** *Assume that* $\Gamma$ *is a feasible connected graph* $(\mathcal{V}(\Gamma), \mathcal{E}(\Gamma))$ *with* $|\mathcal{V}(\Gamma)| = s \geq 2$ *and let* $G_n(\Gamma) = G_{n,\mathbb{R}^m}$*. Suppose* $\{nr_n^m\}_{n\geq 1}$ *is a bounded sequence. Let* $Z_n$ *be Poisson with mean* $E[G_n]$*. Then there is a constant* $c$ *such that for all* $n$*,*

$$d_{TV}(G_n, Z_n) \leq cnr_n^m$$

*If* $n^s r_n^{m(s-1)}$ *converges to* $\alpha \in (0, \infty)$*, then* $G_n$ *converges in distribution to* $Po(\lambda)$ *with* $\lambda = \alpha\mu_\Gamma$*. If* $n^s r_n^{m(s-1)}$ *tends to* $\infty$ *and* $nr_n^m \to 0$*, then* $(n^s r_n^{m(s-1)}\mu_\Gamma)^{-1/2}(G_n - E[G_n])$ *converges to* $N(0,1)$ *in distribution.*

Observe that $q_{\text{marg}}$, the distribution of each $V_i$ does not appear directly in the statement of the theorem, but it has an effect on the speed of convergence ($c$ depends on $q_{\text{marg}}$). Note that the theorem is based the relationship between the number of edges of the graph and the radius for computing the proximity graph. One way to

use this result is to fix $n$ and then look for a sequence that will lead to a sparse regime (Poisson) or to a regime where the subgraph count would converge to normal.

Consider a proximity graph on $\mathbb{R}^2$ (this is, $m = 2$) and let us focus on the edge counts ($k = 2$). In this setting, the sequence $\left\{r_n = \frac{1}{n}\right\}_{n\geq 1}$ corresponds to the first regime (Poisson) with $\alpha = 1$; in contrast $\left\{r_n = \frac{1}{n^{3/4}}\right\}_{n\geq 1}$ corresponds to the second regime (normal). In Figures 4.9 and 4.10 we show how the empirical quantiles of the edge count distribution evolve as a function of $n$ for two different choices for $q_{\text{marg}}$.

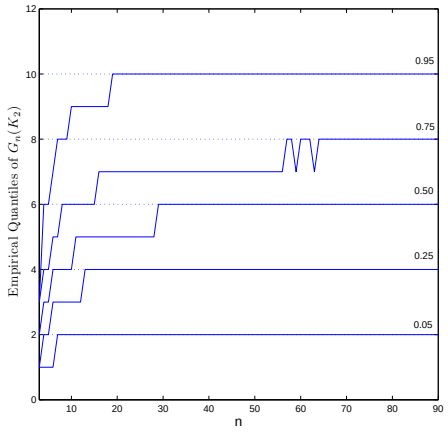


FIGURE 4.9: Empirical quantiles of $G_n(K_2)$ as a function of $n$. Here the sequence is $\left\{r_n = \frac{1}{n}\right\}_{n\geq 1}$ (Poisson regime) and $q_{\text{marg}}$ was set as uniform on $[0, 1]^2$. The dotted lines correspond to quantiles from the Poisson limit.

Because of properties of the Lebesgue measure, it is clear that for fixed $r$, increasing $m$ will lead to sparser graphs. Theorem 4.3.1 confirms this statement and provides a guideline of how large $r$ has to be so that the distribution of certain subgraph count falls in the Gaussian regime.

In Section 4.2 we discussed how to construct weakly decomposable graphs based on a filtration. It is natural to ask if asymptotic approximations are still reasonable if

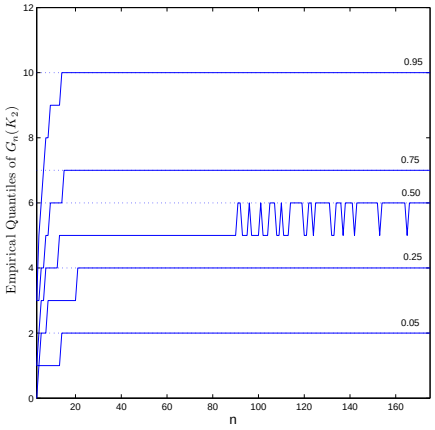


FIGURE 4.10: Empirical quantiles of $G_n(K_2)$ as a function of $n$. Here the sequence is $\left\{r_n = \frac{10}{n}\right\}_{n\geq 1}$ (Poisson regime) and $q_{\text{marg}}$ was set as multivariate normal in $\mathbb{R}^2$ with mean $\mathbf{0}$ and $\sigma_1^2 = 1$, $\sigma_2^2 = 3$, $\sigma_{1,2}^2 = 1.5$. The dotted lines correspond to quantiles from the Poisson limit.

graphs are produced by sampling a vertex set according to a distribution $\mathcal{Q}$ and then applying Algorithm 1. In Figure 4.11 we show a comparison between the empirical distribution of edge counts for a proximity graph and a weakly decomposable graph derived from Algorithm 1. The empirical distributions are very similar, this is not surprising, since as suggested by Figure 4.6, our procedure tends to remove few edges from the graph. In Figure 4.12 we repeat the simulation experiment summarized in Figure 4.9, now forcing the graphs to be decomposable.

We will now discuss a multivariate extension of Theorem 4.3.1. This will allow us to talk about asymptotics that involve the joint distribution of several subgraph counts. The first step is to generalize the indicator function defined in 4.1

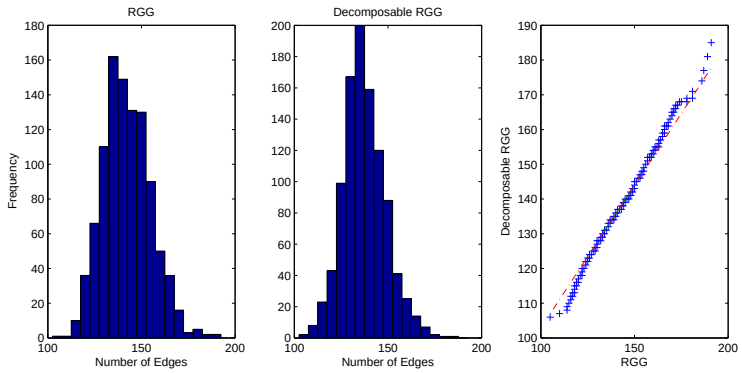


FIGURE 4.11: Distribution of edge counts for both unrestricted and decomposable graphs. Graphs were computed via a filtration of Čech complexes and setting $V_i \sim \text{Unif}([0,1]^2)$; here $|\mathcal{V}| = 100$.

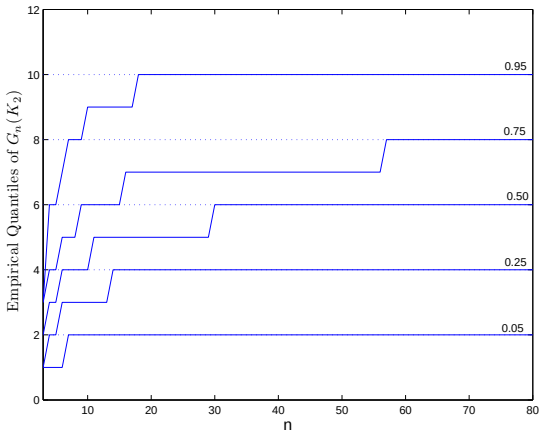


FIGURE 4.12: Empirical quantiles of $G_n(K_2)$ as a function of $n$. Here the sequence is $\left\{r_n = \frac{1}{n}\right\}_{n\geq 1}$ (Poisson regime) and $q_{\text{marg}}$ was set as uniform on $[0,1]^2$. The dotted lines correspond to quantiles from the Poisson limit. Graphs were forced to be decomposable by applying Algorithm 1.

$$h^j_{\Gamma,\Gamma'}((x_1,\cdots,x_{k+k'-j})) = h_\Gamma(\{x_1,\ldots,x_k\}) \tag{4.4}$$

$$\times h_{\Gamma'}(\{x_1,\ldots,x_j,x_{k+1},\cdots,x_{k+k'-j}\}) \tag{4.5}$$

This is, Expression 4.4 keeps track of the event when the subgraph induced by $k+k'-j$ vertices contains $\Gamma$ and $\Gamma'$ as subgraphs. In the same way we can define

$$h^{j}_{\Gamma,\Gamma',n,D}((x_1,\cdots,x_{k+k'-j})) = h_{\Gamma,n,D}(\{x_1,\ldots,x_k\}) \tag{4.6}$$

$$\times h_{\Gamma',n,D}(\{x_1,\ldots,x_j,x_{k+1},\cdots,x_{k+k'-j}\}) \tag{4.7}$$

where $n$ is the index for the sequence $\{r_n\}_{n\geq 1}$. Again $D$ is an open set that contains $\mathbf{0}$ and $\mathrm{Leb}(\partial D)=0$. Let us give a more general version of Expression 4.2

$$\Phi_{j,A}(\Gamma,\Gamma') = \left(\frac{\int_A q_{\mathrm{marg}}(x)^{k+k'-j}dx}{j!(k-j)!(k'-j)!}\right) \tag{4.8}$$

$$\times \int_{\mathbb{R}^d} dx_2 \cdots \int_{\mathbb{R}^d} dx_{k+k'-j} h^{j}_{\Gamma,\Gamma'}(\mathbf{0},x_2,\ldots,x_{k+k'-j}) \tag{4.9}$$

This quantity can be seen as the probability of observing $\Gamma$ and $\Gamma'$ as subgraphs of a geometric graph that has $k+k'-j$ vertices; those vertices are restricted to $D$ and one of them is placed at $\mathbf{0}$. Now we are able to state the following result (see pg. 67 of [84]):

**Theorem 4.3.2** (Penrose)**.** *Let $\Gamma_1,...,\Gamma_s$ be feasible non-isomorphic connected subgraphs. Here $\Gamma_j$ is of order $k_j\in[2,\infty)$. Assume that $\lim_{n\to\infty}(nr_n^d)=\rho\in(0,\infty)$, then the joint distribution of the variables $n^{-1/2}(G'_{n,A}(\Gamma_j)-E[G'_{n,A}(\Gamma_j)])$, $1\leq m$ converges as $n\to\infty$, to a centered multivariate normal with covariance matrix whose $(i,l)$th entry is*

$$\left(\sum_{j=1}^{\min(k_i,k_l)} \rho^{k_i+k_l-j-1}\Phi_{j,A}(\Gamma_i,\Gamma_l)\right) - k_i k_l \rho^{k_i+k_l-2}\mu_{\Gamma_i}\mu_{\Gamma_i}$$

While the computation of the parameters involved in Theorem 4.3.2 can be quite demanding, the relevant information from this result is that the distribution of properly scaled subgraph counts is multivariate normal. One approach for using this result to calibrate a prior would be:

1. Pick $d$ and $q_{\text{marg}}$.
2. Propose a sequence of positive numbers $\{r_n\}_{n\geq 1}$, such that $\lim_{n\to\infty}(nr_n^d) = \rho \in (0, \infty)$.
3. Verify that the joint distribution of subgraph counts is multivariate normal for a given $n$.
4. Modify $\{r_n\}_{n\geq 1}$ to tune the parameters of the prior, while respecting the conditions of the theorem.

Clearly the order of these steps is not strict; returning to a previous step may be necessary.

Assume that each $V_i$ is sampled from an uniform distribution in the $d-$dimensional unit cube. Set $d = 2$ and $\left\{r_n = \frac{1}{\sqrt{n}}\right\}_{n\geq 1}$. According to the power function of Royston's test (which is derived from the Shapiro-Wilk test), the multivariate normal approximation is reasonable for $n \geq 200$ (See Figure 4.14). The joint distribution of 2 and 3−cliques for $n = 75$ is illustrated in Figure 4.13.

### *4.3.3 Vertex Degree*

Another interesting feature of a prior on a space of graphs is the distribution of vertex degree. One way to understand this is to look at the distribution of the number of

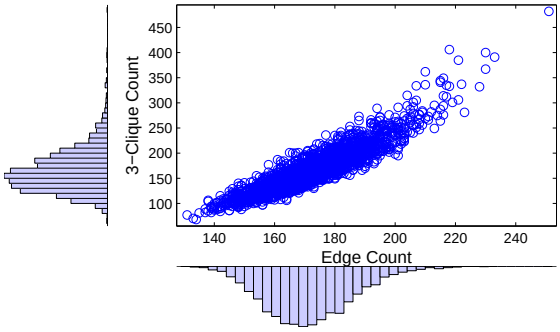


FIGURE 4.13: Edge counts and 3−Clique counts from $2,500$ simulated samples of $\mathcal{G}(\mathbf{V}, 1/\sqrt{2 \cdot 75}, \check{\text{C}}\text{ech})$ where $|\mathbf{V}| = 75$ and $V_i \sim \text{Unif}([0,1]^2)$, $1 \leq i \leq 75$. The multivariate normal appears as a reasonable approximation for the joint distribution; as suggested by Theorem 3.10 in [84]. For this particular case $r_n = 1/\sqrt{2n}$.

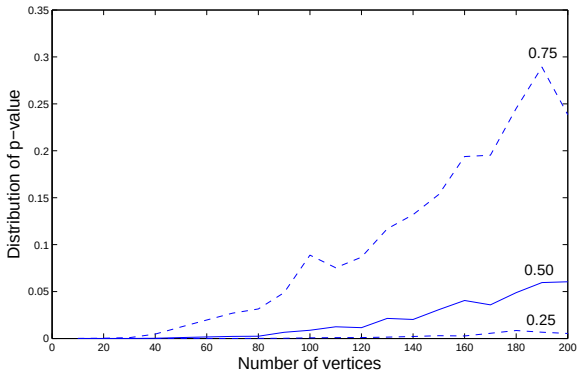


FIGURE 4.14: Empirical quantiles of the $p - value$ from the Royston's test. The quantiles are seen as a function of the size of the graph. The figure suggests that $|\mathcal{V}|$ has to be greater or equal to 200 so half of the tests are rejected at the 0.05. The null hypothesis of the test is that the data were sampled from a multivariate normal.

vertices that have degree greater than or equal to $k$. If the geometric graph of interest is $\mathcal{G}(\mathcal{V}, r, \check{\text{C}}\text{ech})$ (the 1−skeleton of the Čech complex) in $\mathbb{R}^m$, then the statements

1. $v_i \in \mathcal{V}$ has degree $\geq k$ in $\mathcal{G}(\mathcal{V}, r)$.

2. The intersection of $v_i + r\mathbb{B}_m$ and $\mathcal{V} - \{v_i\}$ has at least $k$ elements.

are equivalent. Another way to phrase the second formulation of the problem is that the $k-$nearest neighbor of $v_i$ is at a distance less or equal to $r$.

We follow closely the presentation of [84]: Denote by $\mathcal{V}_n$ a generic vertex set with $n$ elements and let

$$\left\{ r_n(t) = \left(\frac{t}{n}\right)^{\frac{1}{m}} \right\}_{n\geq 1}, \quad t > 0.$$

Now, define $Z_n(t)$ as the number of vertices in $\mathcal{G}(\mathcal{V}_n, r_n(t))$ with degree at least $k$. Let

$$B(n, x, t) = x + r_n(t)\mathbb{B}_m, \tag{4.10}$$

this is, a ball with center $x$ and radius $r$, then

$$Z_n(t) = \sum_{i=1}^{n} \mathbf{1}_{\{|B(n,v_i,t)\cap\mathcal{V}|\geq k+1\}}. \tag{4.11}$$

Clearly, for $n$ fixed $Z_n(t)$ is a non-decreasing discrete-valued stochastic process. In this section we discuss a result that describes the behavior of $\{Z_n(t)\}_{t\geq 0}$ as $n$ increases.

Let $A$ be a subset of $\{0, 1, 2, \ldots\}$. Denote by $\pi_\lambda(A)$ the probability that a r.v. with distribution $\mathrm{Po}(\lambda)$ takes a value in $A$. Let $\nu_m$ be the volume of the $m-$dimensional unit ball. Define

$$Z_n(t, A) = \sum_{i=1}^{n} \mathbf{1}_{\{|B(n,v_i,t)\cap\mathcal{V}|\geq k+1\}\cap\{v_i\in A\}}. \tag{4.12}$$

Now we are able to state (see pg 76 of [84] ):

**Theorem 4.3.3** (Penrose). *Suppose $A \subseteq \mathbb{R}^m$ is a Borel set. Then*

$$\lim_{n\to\infty}(n^{-1}E[Z_n(t,A)]) = \int_A \pi_{\nu_m t q_{marg}(x)}([k,\infty))q_{marg}(x)dx \tag{4.13}$$

This result is useful because it allows the use of efficient Monte Carlo estimates for $E[Z_n(t,A)])$. Let us outline a procedure for approximating the right side of Expression 4.13:

1. Sample $X_1, X_2, \ldots, X_l$ from $q_{\text{marg}}$ restricted to $A$.
2. For each $X_i$ compute $\lambda_i = \nu_m t q_{\text{marg}}(X_i)$, $i \in \{1, 2, \ldots, l\}$.
3. For each $i$ compute $y_i = \pi_{\lambda_i}([k,\infty))$, $i \in \{1, 2, \ldots, l\}$.
4. Let $\widehat{y} = \frac{1}{l}\sum_{i=1}^{l} y_i$

example here.

### *4.3.4 Repulsive Point Processes*

While asymptotic results like Theorem 3.4 in [84] give insight about the implications of choosing a particular $\mathcal{Q}$ that assumes independence among the components of $\mathbf{V}$, allowing more general choices for $\mathcal{Q}$ may lead to better control over the joint distribution of certain graph features. For example, one may be interested in imposing sparsity by penalizing the presence of high dimensional cliques. Using an Alpha complex with $\mathbf{V} \sim \text{Unif}(\mathbb{B}_k)$ gives a way to encode such information, another possibility would be to use a Čech complex and a repulsive point process as prior for $\mathbf{V}$.

A repulsive point process is a probability model on point patterns in $\mathbb{R}^m$ such that penalizes configurations of points where 2 or more points are closer than certain

threshold value $t$. Since we are interested in using a point process as a prior, it would be convenient that the joint density of any finite set of points could be computed efficiently. It is also desirable that the process could be simulated exactly, or that there exist efficient MCMC algorithms to obtain samples from it. Two processes that fulfill these requirements are the Strauss and the Matérn III.

The Strauss process is a repulsive model that is based on the idea of surrounding each point with a disc of radius $t$ and for each pair of points whose discs overlap is penalized in the density. The parameters of this process are: $\lambda \in [0, \infty)$ which plays a similar role as the intensity of a homogeneous Poisson process, and $\gamma \in [0, 1]$ which controls the penalty for each pair of points being close, and $t$ is the radius of the disc. If the point patterns are subsets of $\mathbb{R}^2$, the joint density of a finite set of points is proportional to:

$$q(v) = \lambda^{|v|}\gamma^{s(v)}, \quad s(x) = |\{\{i, j\} : \text{dist}(v_i, v_j) < 2t\}| \tag{4.14}$$

This process can be simulated efficiently via reversible jump MCMC or by a continuous birth and death chain (See [58])

While the Strauss process is easily defined by giving an expression for its density, the Matérn III process is better described in terms of the algorithm that generates it

Input: t, $\lambda(\cdot)$. Output:$V$

1. Simulate $A \leftarrow$ from a Poisson process $(\lambda(\cdot))$

2. For all points $p \in A$

3. Simulate a label $t_p \leftarrow \text{Unif}([0,1])$

4. Let $L \leftarrow \{p_1, p_2, \ldots, p_{|A|}\}$ where $t_{p_1} \leq t_{p_2} \leq \cdots t_{p_{|A|}}$

5. While $l \neq \varnothing$

6. Let $p$ be the element of $L$ with smallest label value

7. Let $A \leftarrow A \ \{q : t_p < t_q\}$

**Example 4.3.2.** *By performing a similar Monte Carlo experiment as the one used to produce Figure 4.13, setting $\mathcal{Q}$ as a Matérn III with parameter $s = 0.35$, we obtain Figure 4.15. Observe that the distribution of the number of edges does not change much (the sparsity of the graph), but the presence of larger cliques is being penalized (having 200 or more $3-$cliques in the uniform case has probability higher than 0.25, for the Matérn III that probability is lower than 0.05). Clearly this could not be achieved by an Erdös-Rényi model with a single edge inclusion probability, and it is not obvious how to encode this type of information in the approach proposed by [75].*

This example illustrates an advantage of the proposed methodology over Erdös-Rényi models. For those models designing a flexible distribution involving more than one graph feature implies specifying a large number of parameters; for our approach such flexibility can be achieved by tuning $\mathcal{Q}$ or by making choices over $\mathcal{A}$ and $r$.

Repulsive processes are by no means the only alternative for generalizing the priors derived from the REG setting; In principle one can induce different dependence structures on $\mathbf{V}$ via $\mathcal{Q}$ to obtain richer models on graph space.

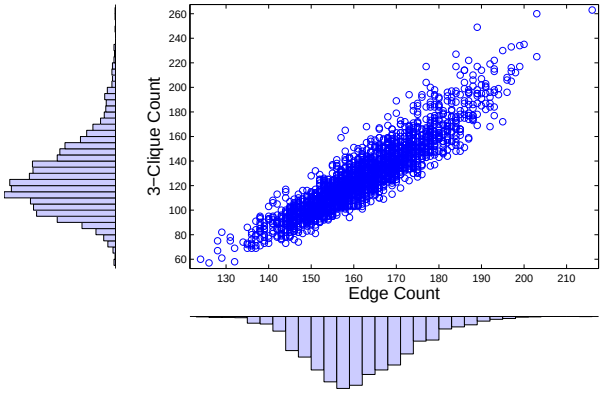


FIGURE 4.15: Edge counts and 3−Clique counts from $2,500$ simulated samples of $\mathcal{G}(\mathbf{V}, 1/\sqrt{2 \cdot 75}, \check{\text{C}}\text{ech})$ where $|\mathbf{V}| = 75$ and $V$ sampled from a Mattérn III with parameter 0.35.

# 5

# Bayesian Inference

In this Chapter we specify all the elements required for performing inferences on graphical models according to the parametrizations proposed in Chapter 4. In Section 5.1 we review the types of factorizations we are interested in. In Section 5.2 we define the priors that will be used. The corresponding MCMC algorithms are proposed in Section 5.3.

## 5.1 General Setting

As discussed in Section 3.3 we are interested in making inferences on a model that can be represented as

$$f(x) = \prod_{a \in \mathcal{P}(\mathcal{G})} \psi_a(x_a \mid \theta_a), \tag{5.1}$$

this is, a junction tree where the underlying graph is not necessarily decomposable (remember that $\mathcal{P}(\mathcal{G})$ represents the prime components of $\mathcal{G}$). We will also work

with the more general form

$$f(x) = \prod_{a \in \mathcal{C}(\mathcal{G})} \psi_a(x_a \mid \theta_a), \tag{5.2}$$

where $\mathcal{C}(\mathcal{G})$ denotes the complete sets of $\mathcal{G}$. Inferences will be performed according to the Bayesian approach. In general, we adopt prior specification that can be written as

$$p(\Theta \mid \mathcal{G}) \times p(\mathcal{G}), \tag{5.3}$$

where $\Theta = \{\theta_a\}_{a \in \mathcal{P}(\mathcal{G})}$ (or $\{\theta_a\}_{a \in \mathcal{C}(\mathcal{G})}$, depending on the context). The first term in Expression 5.3 corresponds to a hyper Markov law suited to the application at hand. To specify that part of the model we rely on standard methods (e.g. a hyper Inverse Wishart if the sampling distribution is Multivariate normal). Our work is focused on defining $p(\mathcal{G})$ when $\mathcal{G}$ is understood as a geometric graph, and on designing suitable MCMC algorithms.

## 5.2 Prior Specification

All the graphs used in our statistical models are computed using nerves based on the position of the vertex set elements $\mathcal{V}$ and a scale parameter $r$. This results in a parametrization $(\mathcal{V}, r)$ where $\mathcal{V} = (V_i)_{i=1}^d \in \mathbb{R}^m$. The graph $\mathcal{G}$ is completely determined by the location of each vertex, the value of $r$, and the class of sets $\mathcal{A}$ used to compute the nerve. Thus, the mapping $(\mathcal{V}, r) \to \mathcal{G}(\mathcal{V}, r)$ is deterministic. Observe that $\mathcal{G}(\mathcal{V}, r) = \mathcal{G}(h(\mathcal{V}), r)$ for any rigid transformation $h$, similarly $\mathcal{G}(\mathcal{V}, r) = \mathcal{G}(a\mathcal{V}, ar)$ for any $a > 0$. Therefore, due to invariance in the map $(\mathcal{V}, r) \to \mathcal{G}(\mathcal{V}, r)$

$$\mathcal{G}(\mathcal{V}, r) = \mathcal{G}(ah_0(\mathcal{V}), ar),$$

where $h_0$ is a composition of rigid transformations that maps the mean of the vertex set to the origin and $a \times \max\{|h_0(V_i)| : V_i \in \mathcal{V}\} < 1$. Restricting the support of the prior to $\mathbb{B}_m$, the closed unit ball in $\mathbb{R}^m$ does not reduce the model space as long as the scale parameter is restricted to $0 < r < \frac{2}{d}$. We will use this fact heavily, so from now on we will write $\mathcal{G}(\mathcal{V})$ instead of $\mathcal{G}(\mathcal{V}, r)$.

The rationale of using the unit ball in instead of other subset of $\mathbb{R}^m$ is the following: First, it is desirable that the vertices are restricted to a bounded set, since that greatly simplifies arguments regarding the ergodicity of MCMC algorithms (Section 5.3.4). The set that contains $\mathcal{V}$ should be simply connected; it is clear that a disconnected set may prevent the Markov chain from being irreducible and holes in the set can make some computations cumbersome without any clear advantage. Convexity is also desirable, since we would like to avoid bottlenecks in the space (wich would prevent the chain from mixing). In the rest of the chapter we will assume $\mathcal{V} \subset \mathbb{B}_m$; We acknowledge that this choice is still somewhat arbitrary and other simply connected bounded set (like $[0, 1]^n$) could be used instead.

The most relevant implication of the construction we use is that we can induce priors on a space of graphs by sampling configurations of $d$ points on the unit ball in $\mathbb{R}^m$. One way to achieve this is to sample $\mathbf{V} = (V_1, \ldots, V_d)$ according to a distribution $\mathcal{Q}$. If the components of $\mathbf{V}$ are set to be independent, then results for Random Geometric Graphs can be used to understand the implications of a particular $\mathcal{Q}$ and calibrate the prior accordingly (see [84] and [85]). If the components of $\mathbf{V}$ are not independent, the calibration of the prior can be performed via simulation. We will

always assume that $\mathcal{Q}$ has a density, which we denote by $q$.

It is worth mentioning that calibrating a prior for a space of graphs is a different task from working with a space of hypergraphs. One reason for this is that a specific clique (which is computed directly from the graph, equivalently from pairwise intersections of convex sets) can have relatively high prior probability, but this subgraph is computed most of the time from the same simplicial complex (computed from all the intersections of a set of convex sets); some simplicial complexes may even be excluded from the support of the prior. Think of a complete graph with 3 vertices: For $m = 2$, $\mathcal{A} = \check{\text{C}}\text{ech}$, $r = \sqrt{0.3}$ and $V_i \sim \text{Unif}(\mathbb{B}_2)$ with the $V_i$'s independent ($1 \leq i \leq 3$), the complete graph with 3 vertices has a prior probability of 0.2466. Under these conditions the nerve with maximal simplices $\{V_1, V_2, V_3\}$ has probability 0.2376, while the nerve with maximal simplices $\{V_1, V_2\}$, $\{V_2, V_3\}$, and $\{V_1, V_3\}$ has prior probability 0.009 (see Table 5.1). This is, one of the hypergraphs has the highest prior probability while the other has the lowest, despite having the same 1−skeleton.

## 5.3 Posterior Sampling

### *5.3.1 Local and Global Moves in Graph Space*

In Sections 4.1, 4.2 and 4.3 we discussed how to use the parametrization $(\mathcal{V}, r) \to \mathcal{G}(\mathcal{V}, r)$ to propose priors on graph space. In the previous section we adopted the convention of restricting $\mathcal{V}$ to be a subset of $\mathbb{B}_m$, where $m \in \{2, 3\}$. Now we proceed to design algorithms to sample from the posterior that are suited to this type of parametrizations. One of the main reasons that lead us to develop the approach

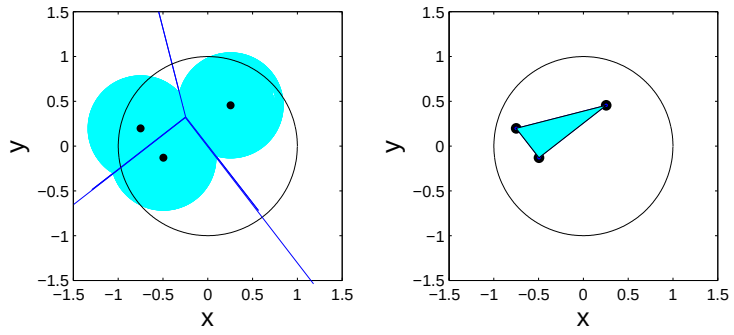


FIGURE 5.1: Pattern of convex sets and corresponding nerve. Here $\mathcal{A}$ = Alpha and $r = \sqrt{0.3}$.

| Graph Topology | Prior Probability |
|---|---|
| $\{V_1, V_2, V_4\}$ | 0.2376 |
| $\{V_1, V_2\}\,\{V_2, V_3\}$ | 0.1344 |
| $\{V_2, V_3\}\,\{V_1, V_3\}$ | 0.1316 |
| $\{V_1, V_2\}\,\{V_1, V_3\}$ | 0.1314 |
| $\{V_1, V_2\}\,\{V_3\}$ | 0.1126 |
| $\{V_1, V_3\}\,\{V_2\}$ | 0.1098 |
| $\{V_2, V_3\}\,\{V_1\}$ | 0.1016 |
| $\{V_1\}\,\{V_2\}\,\{V_3\}$ | 0.0320 |
| $\{V_1, V_2\}\,\{V_2, V_3\}\,\{V_1, V_3\}$ | 0.0090 |

Table 5.1: Estimated prior probabilities for the 9 possible nerves based on 3 vertices. Here $\mathcal{A}$ = Alpha, $r = \sqrt{0.3}$ and $V_i \sim \text{Unif}(\mathbb{B}_2)$, $1 \leq i \leq 3$.

presented in this thesis is that spaces of graphs do not have an obvious topology or metric (as in $\mathbb{R}^m$), therefore research in this area has developed in a way that MCMC algorithms tend to use proposals based on local moves (add or delete one edge at a time); this is especially true for methodology dealing with decomposable

graphs [48]. A move that is not local is called global. The algorithms discussed in this section do not distinguish between local and global moves. In our approach graphs are computed from nerves, which characterize the intersection pattern of a collection of convex sets $A(v_1, r), \ldots, A(v_d, r)$. Intuitively a small perturbation on the vertex set would lead to adding or deleting few edges from the graph. This intuition comes from the fact that if two compact sets $A_1$ and $A_2$ in $\mathbb{R}^m$ do not overlap, there exist $B_1$ and $B_2$ open such that $A_i \subset B_i$, $i \in \{1, 2\}$ and $B_1 \cap B_2 = \varnothing$. A similar argument can be formulated when the interior of two compact sets overlap. However, if the magnitude of the perturbation is given beforehand (this is regardless of the positions of the sets, like in a random walk proposal ) there is no guarantee that the move in graph space will be local (See Figure 5.2). Because of invariance to rigid transformations, it does not hold that a drastic change in the vertex set would lead to big changes in the graph structure (Figure 5.2). Still, a proposal based on adding small perturbations on the elements of $\mathcal{V}$ is the most natural way to obtain a procedure that performs local moves (in graph space) with high probability. If one wants to explore the a space of graphs in an efficient ways and deal with multiple modes in the posterior, it is necessary to use proposals that generate local as well as global moves; we will discuss that in the following section.

### *5.3.2 Theoretical Justification of Random Walk*

One way to induce local moves in graph space with high probability is by perturbing each element of the vertex set independently using a random walk. We will use a random walk in the half plane that approximates Brownian motion with reflecting

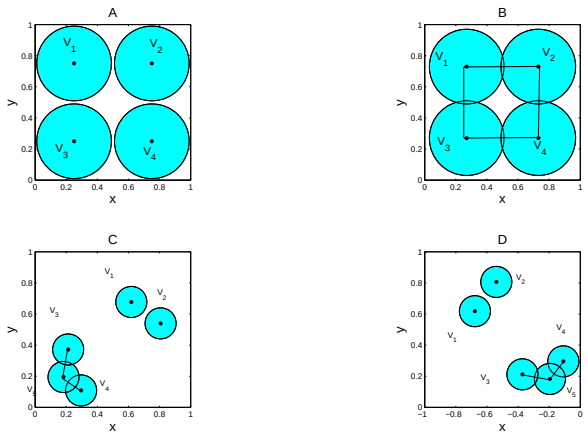


FIGURE 5.2: Here we illustrate how a small perturbation on the vertex set (A) may lead to a global move on the 1−skeleton of the nerve (B). In contrast rotating the vertex set (C) by $\pi/2$ radians does not produce any change in the nerve (D).

boundary condition, then we apply a conformal mapping to transform it into a random walk that approximates Brownian motion on the unit disc. Brownian motion on the plane maps into Brownian motion on the disc as a consequence of Lévy's Theorem [71]. The conformal map we use is:

$$F(z) = i\left(\frac{1+z}{1-z}\right), \quad |z| \leq 1, \tag{5.4}$$

By assuming that the angle of reflection is always the same when hitting the disc, we can apply the result from [52]. They proved that the stationary distribution of a Brownian motion restricted to a simply connected region $S$ and with reflection on the boundary has the form:

$$p(z) = c\ \Re(\exp(L(z))), \quad z \in S^{\circ}, \tag{5.5}$$

where $S^{\circ}$ denotes the interior of $S$, $\Re$ represents the real part of a complex number

and

$$L(z) = \frac{1}{\pi} \int_{\partial S} \left[ \frac{\vartheta(\xi) - \vartheta(\xi_0)}{F(z) - F(\xi)} \right] dF(\xi) - i\vartheta(\xi_0), \quad z \in S. \tag{5.6}$$

$F$ is a conformal mapping such that the image of $S$ is the upper half-plane (Expression 5.4) and $\vartheta(\xi)$ is the angle of reflection at the boundary point $\xi$. When $S = \mathbb{B}_2$, $\partial S$ is $C$, therefore

$$L(z) = \frac{1}{\pi} \int_C [\vartheta(\xi) - \vartheta(1)] \frac{(1-z)d\xi}{(z-\xi)(1-\xi)} - i\vartheta(1). \tag{5.7}$$

If $\xi$ is set to be constant, Equation 5.7 simplifies to:

$$L(z) = -i\vartheta(1), \tag{5.8}$$

This is $p(z) \propto \cos(\vartheta(1))$; the stationary distribution is uniform. We also have that the stationary distribution of the random walk matches the stationary distribution of Brownian motion.

### *5.3.3 MCMC Algorithms*

Given the prior discussed Section 5.2 we will use a random walk that approximates Brownian motion as a proposal distribution and the Metropolis-Hastings algorithm to sample from the posterior. We first specify the procedure for $m = 2$ and then for $m = 3$.

For the case where $m = 2$ the random walk is more conveniently parametrized in polar coordinates $(\rho_i, \varphi_i)$

$$V_i = (x_i, y_i), \quad \rho_i = \sqrt{x_i^2 + y_i^2}, \quad \varphi_i = \arctan\left(\frac{y_i}{x_i}\right).$$

For each point $V_i^{(t)}$ the proposed move $V_i^{(t+1)}$ is

$$\begin{aligned} V_i^{(t+1)} &= \left(\rho_i^{(t)}\cos(\varphi_i^{(t)}), \rho_i^{(t)}\sin(\varphi_i^{(t)})\right), \\ \rho_i^{(t+1)} &= \left(f_b\left(\left(\rho_i^{(t)}\right)^2 + \varepsilon\right)\right)^{\frac{1}{2}}, \quad \varepsilon \sim \text{No}\left(0, \eta^2\right) \\ \varphi_i^{(t+1)} &= \varphi_i^{(t)} + \delta, \quad \delta \sim \text{No}\left(0, \left(\rho_i^{(t)}\right)^{-2}\eta^2\right). \end{aligned}$$

The angles are computed modulo $2\pi$ and $f_b$ reflects the walk when it hits the boundary.

For $m = 3$ it is convenient to express the random walk in terms of spherical coordinates $(\rho, \varphi, \nu)$; where $\rho$ denotes the radial distance with respect to the origin, $\varphi$ the polar angle from the $z-$axes and $\nu$ is the azimuthal angle in the $xy-$plane from the $x-$axis. The proposal move for each $V_i$ is

$$\begin{aligned} V_i^{(t+1)} &= \left(\rho_i^{(t)}\cos(\varphi_i^{(t)})\sin(\nu_i^{(t)}), \rho_i^{(t)}\sin(\varphi_i^{(t)})\sin(\nu_i^{(t)}), \rho_i^{(t)}\cos(\nu_i^{(t)})\right), \\ \rho_i^{(t+1)} &= \left(f_b\left(\left(\rho_i^{(t)}\right)^3 + \varepsilon_1\right)\right)^{\frac{1}{3}}, \quad \varepsilon_1 \sim \text{No}\left(0, \eta^2\right), \\ \varphi_i^{(t+1)} &= \varphi_i^{(t)} + \varepsilon_2, \quad \varepsilon_2 \sim \text{No}\left(0, \left(\rho_i^{(t)}\right)^{-1}\eta\right), \\ \nu_i^{(t+1)} &= \begin{cases} \nu_i^{(t)} + \varepsilon_3 & \text{if } \nu_i, \varphi_i \text{ are both } > 0 \text{ or } < 0 \\ (\nu_i^t + \pi) + \varepsilon_3 & \text{otherwise,} \\ \text{here } \varepsilon_3 \sim \text{No}\left(0, \left(\rho_i^{(t)}\right)^{-2}\eta^2\sin^2(\varphi_i^{(t)})\right). \end{cases} \end{aligned}$$

Note that it is necessary to add $\pi$ radians to $\nu_i^{(t+1)}$ when $\nu_i^{(t)}$ and $\varphi_i^{(t)}$, otherwise the random walk would reflect every time it hits the $z-$axes.

As explained in Section 5.3.1, the reason to use a random walk as proposal distribution is that it can be tuned so it produces local moves with high probability. For

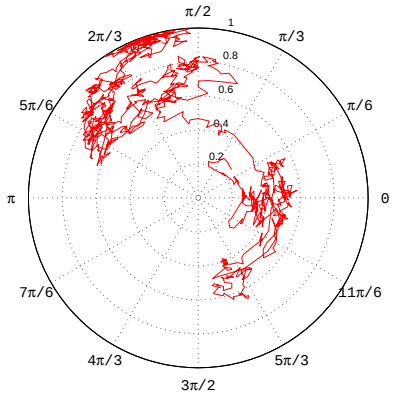


FIGURE 5.3: Trajectory of random walk proposal for 500 iterations and $\eta = \frac{1}{50}$. Here $m = 2$.

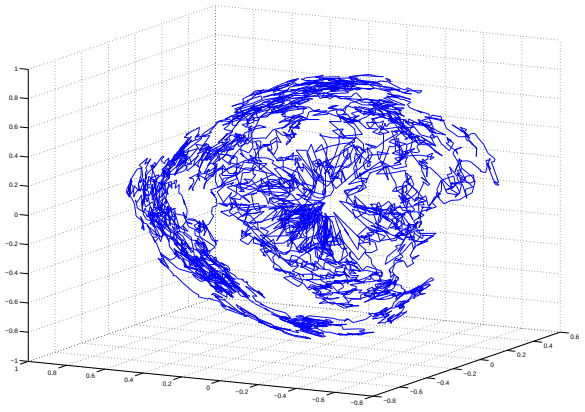

FIGURE 5.4: Trajectory of random walk proposal for $5,000$ iterations and $\eta = \frac{1}{50}$. Here $m = 3$.

problems involving a moderate number of variables, it is desirable to also propose global moves, so the MCMC is robust to local modes of the posterior. To obtain an algorithm able to combine local and global moves we can use an hybrid kernel. For example, let $0 < \kappa < 1$ and set

1. A random walk on $\mathbb{B}_m$ for each $V_i$, $1 \leq i \leq d$, for a given value of $\eta$. This proposal is picked with probability $1 - \kappa$.

2. An element of $\{1, 2, \ldots, d\}$ is picked at random, with equal probability. The vertex corresponding to that index is sampled from Unif($\mathbb{B}_m$). This proposal is picked with probability $\kappa$.

A slight generalization of this hybrid kernel is to set $0 \leq \kappa_j$, $1 \leq j \leq d$ such that $\sum_{j=1}^{d} \kappa_j < 1$, then propose moves according to

1. A random walk on $\mathbb{B}_m$ for each $V_i$, $1 \leq i \leq d$, for a given value of $\eta$. This proposal is picked with probability $1 - \sum_{j=1}^{d} \kappa_j$.

2. A subset of $\{1, 2, \ldots, d\}$ of size $j$ ($1 \leq j \leq d$) is sampled uniformly. The vertices corresponding to those indices are sampled independently from Unif($\mathbb{B}_m$). This proposal is picked with probability $\kappa_j$.

Another strategy that we found to be useful is to use a hybrid kernel where some of the proposals are random walks on $\mathbb{B}_m$; what is varying across those proposals is the tuning parameter $\eta$, which determines the 'speed' of the random walk. The objective of such proposals is to provide some tempering effect.

Now we present the expression of the Metropolis ratio for different settings and inference problems. If the interest is on estimating the graph structure only, then the Metropolis ratio is given by

$$R_{(i,i+1)} = \frac{\mathcal{M}(\mathcal{V}^{(i+1)})q(\mathcal{V}^{(i+1)})}{\mathcal{M}(\mathcal{V}^{(i)})q(\mathcal{V}^{(i)})}, \tag{5.9}$$

where

$$\mathcal{M}(\mathcal{V}) = \int_{S(\mathcal{G})} p(x \mid \Theta, \mathcal{G}(\mathcal{V}))p(\Theta \mid \mathcal{G}(\mathcal{V}))d\Theta. \tag{5.10}$$

This is, Expression 5.10 is proportional to the marginal likelihood for the graph $\mathcal{G}(\mathcal{V})$ (see pg. 348 of [90]). In particular, if the sampling distribution is multivariate normal then the hyper Markov law to use is the hyper inverse Wishart. More precisely, assume that the multivariate normal has mean zero and that the prior for the covariance conditioning on $\mathcal{G}$ is $\Sigma \sim \text{HIW}_{\mathcal{G}}(\delta, D)$. Under these assumptions

$$\mathcal{M}(\mathcal{V}) = \frac{1}{(2\pi)^{\frac{nd}{2}}} \frac{I_{\mathcal{G}(\mathcal{V})}(\delta + n, D + X^T X)}{I_{\mathcal{G}(\mathcal{V})}(\delta, D)}, \tag{5.11}$$

where $I_{\mathcal{G}}(\delta, D)$ is the normalizing constant of $HIW_{\mathcal{G}}(\delta, D)$ and $n$ is the number of observations. If we assume $V_i \sim \text{Unif}(\mathbb{B}_2)$ a priori with $V_1, V_2, \ldots, V_d$ independent, then Equation 5.12 becomes

$$R_{(i,i+1)} = \frac{I_{\mathcal{G}(\mathcal{V}^{(i+1)})}(\delta + n, D + X^T X)}{I_{\mathcal{G}(\mathcal{V}^{(i)})}(\delta + n, D + X^T X)} \frac{I_{\mathcal{G}(\mathcal{V}^{(i)})}(\delta, D)}{I_{\mathcal{G}(\mathcal{V}^{(i+1)})}(\delta, D)}. \tag{5.12}$$

If $\mathcal{G}(\mathcal{V}^{(j)})$ is weakly decomposable for either $j = i$ or $j = i + 1$, then $I_{\mathcal{G}(\mathcal{V}^{(j)})}(\delta, D)$ can be expressed analytically (Section 3.5) and Equation 5.12 can be computed more efficiently. If one of the graphs is not weakly decomposable, the corresponding normalizing constants most be estimated via simulation; we used the method developed by Atay-Kayis and Massam to perform this task [3] for low-dimensional examples. For higher dimensional problems we recommend the algorithm proposed by [19].

For the problem of estimating the model in Expression 5.2, it is necessary to design proposals for the parameters involved in the terms of the factorization given

the graph; the generic notation $p(\theta^{(i+1)} \mid \theta^{(i)})$ will be used for such proposals. We first consider the simplest case possible, this is, when the parameter space for the terms in the factorization does not depend on the graph (concrete examples of this will be discussed on Sections 6.1, 6.3 and 6.4). For the proposals discussed in this section, the Metropolis ratio is given by

$$R_{(i,i+1)} = \frac{p(\mathbf{x}|\mathcal{V}^{(i+1)}, \theta^{(i+1)})p(\theta^{(i+1)} \mid \mathcal{V}^{(i+1)})q(\mathcal{V}^{(i+1)})p(\theta^{(i)} \mid \theta^{(i+1)})}{p(\mathbf{x}|\mathcal{V}^{(i)}, \theta^{(i)})p(\theta^{(i)} \mid \mathcal{V}^{(i)})q(\mathcal{V}^{(i)})p(\theta^{(i+1)} \mid \theta^{(i)})}, \tag{5.13}$$

The term $p(\mathbf{x}|\mathcal{V}^{(i)}, \theta^{(i)})$ is very general: It is a composition of the maps $\mathcal{V}^{(i)} \to \mathcal{G}(\mathcal{V}^{(i)})$ (or $\mathcal{V}^{i} \to \mathcal{H}(\mathcal{V}^{(i)})$, depending on the context), and the likelihood of the data given $\mathcal{G}(\mathcal{V}^{(i)})$ ($\mathcal{H}(\mathcal{V}^{(i)})$) and the parameters involved in the terms of the factorization. It is also the only term that would suffer changes if one opted for a different class of sets for computing the nerve, or if one decided to factorize according to the complete sets of a graph instead of the cliques. This observations are valid for all Metropolis algorithms discussed in this chapter. Despite the proposals and metropolis ratios have a simple and general form, the behavior of the MCMC algorithms is sensible to the class of sets used to compute the nerve. This will be discussed in Section 5.3.4.

Typically the model 5.2 will be of variable dimension, by this we mean that $\Theta \mid \mathcal{G}(\mathcal{V})$, the parameter space for all marginals given the factorization can change according to $\mathcal{G}(\mathcal{V})$. Remember that the proposal distribution for each $V_i$ is a random walk in $\mathbb{B}_m$ and $r$ is regarded as fixed. In this setting not all accepted moves for $\mathcal{V}$ will lead to a change in $\mathcal{G}(\mathcal{V})$; for accepting such moves an ordinary Metropolis ratio would suffice (Equation 5.13). For the case where accepting the move would change $\mathcal{G}(\mathcal{V})$ (and therefore $\Theta \mid \mathcal{G}(\mathcal{V}, r)$), the computation of the Metropolis ratio is

slightly more involved. Let $(\mathcal{V}^{(i)}, \theta^{(i)})$ be the current state of the chain, $\theta^{(i)} \in \Theta_u$. We think that the saturated approach proposed by [18] is a sensible choice; it is based on sampling $W_{\dim(\Theta^{(i)})+1}, \ldots, W_D$ iid auxiliary variables where $D$ is the maximum possible dimension for $\Theta$; we denote by $\zeta$ the univariate density of the auxiliary variables. If we want to propose a move from $\Theta_u$ to $\Theta_v$ where $n_v = \dim(\Theta_v) > n_u = \dim(\Theta_u)$, then this approach is similar to the one proposed by [49], in the sense that the instrumental variables $W_{n_u+1}, \ldots, W_{n_v}$ are paired with an injective mapping $f_{u,v} : \Theta_u \times \mathbb{R}^{n_v - n_u} \to \Theta_v$ to generate the point $\theta^{(i+1)} \in \Theta_v$ to be used as proposed move. In this setting the Metropolis ratio has the form

$$R_{(i,i+1)} = \frac{p(\mathbf{x}|\mathcal{V}^{(i+1)}, \theta^{(i+1)}) p(\theta^{(i+1)} \mid \mathcal{V}^{(i+1)}) q(\mathcal{V}^{(i+1)})}{p(\mathbf{x}|\mathcal{V}^{(i)}, \theta^{(i)}) p(\theta^{(i)} \mid \mathcal{V}^{(i)}) q(\mathcal{V}^{(i)}) \prod_{k=n_u+1}^{n_v} \zeta(u_i, k)} |J_{i,i+1}| \tag{5.14}$$

The last term in the Expression 5.14 represents the Jacobian of the transformation. In the case where $n_u > n_v$ the Metropolis ratio is the reciprocal of 5.14. The advantage of using [18] instead of [49] is that it can also handle non-nested models, which is a situation that will happen given our approach for moving on graph space. Observe that the term corresponding to the probability of moving from $\Theta_u$ to $\Theta_v$ (and the one from moving $\Theta_v$ to $\Theta_u$) in any particular iteration is absent in 5.14 (see [49]). This is because the search in the graph space is driven by the moves in $\mathcal{V}$, which are produced by a proposal that cancels out in the Metropolis ratio as in 5.12.

### *5.3.4 Convergence of the Markov chain*

To guarantee that summaries computed from samples of the algorithms described above can be used for Bayesian inference, conditions regarding the convergence of the chain must be verified. A sufficient condition for applying the Law of Large Numbers for Markov chains (see Section 6.7.1 [91]) is Harris recurrence.

We first introduce some notation: Denote by $\mathcal{G}_d$ the set of all graphs with $d$ vertices. Let $\dot{\mathcal{G}}(d, m, r)$ be the family of the graphs than can be produced by computing the $1-$skeleton of the complex isomorphic to the nerve of $d$ open balls with radius $r$ in $\mathbb{R}^m$. In an analogous way define $\bar{\mathcal{G}}(d, m, r)$ as the family of graphs corresponding to the $1-$skeleton of a Čech complex $\mathbb{R}^m$.

**Proposition 5.3.1.** *The MCMC procedures described in Section 5.3 produce samples from a Harris recurrent chain in* $\mathcal{G}_d \cap \dot{\mathcal{G}}(d, m, r)$.

*Proof.* Let $\mathcal{G} \in \dot{\mathcal{G}}(d, m, r)$; which means that there exists a vertex set $\mathcal{V}$ such that $\mathcal{G} = \dot{\mathcal{G}}(\mathcal{V}, r)$. Set

$$\varepsilon_{\mathcal{G}} = \frac{1}{2} \min_{i<j} |2r - |V_i - V_j||$$

and $Y_k = V_k + h_k$, where $V_k \in \mathcal{V}$ and $|h_k| \leq \varepsilon_{\mathcal{G}}$. Clearly $\dot{\mathcal{G}}(\mathcal{V}, r) = \dot{\mathcal{G}}(\mathcal{Y}, r)$; this implies that there exists a set $\Delta$ of positive Lebesgue measure in $(\mathbb{B}_m)^d$ such that $\dot{\mathcal{G}}(\mathcal{V}, r) = \dot{\mathcal{G}}(\mathcal{W}, r)$ for any $\mathcal{W} \in \Delta$. For this setting, an Independent Metropolis Hastings with proposal Unif $(\mathbb{B}_m)$ for each vertex will be irreducible for the space $\mathcal{G}_d \cap \dot{\mathcal{G}}(d, m, r)$. Since such space is finite the chain will be positive recurrent. Let $A \subset (\mathbb{B}_m)^d$ with positive Lebesgue measure; since the number of possible sample

paths that visit $A$ only a finite number of times is countable, the chain will also be Harris recurrent. The same argument follows if instead we use a hybrid kernel such that, with probability $\pi_i > 0$ a move for $V_i$ is proposed from $\text{Unif}\,(\mathbb{B}_m)$, $1 \leq i \leq d$. For the priors on the vertex set that arise from strictly positive density functions on $(\mathbb{B}_m)^d$, the set $\bar{\mathcal{G}}(d, m, r) - \dot{\mathcal{G}}(d, m, r)$ has probability zero; this is because those priors are equivalent to Lebesgue measure in the sense of absolute continuity. It follows that such set is not relevant for arguments concerning the ergodicity of the MCMC. □

One could have phrased this argument in terms of the random walk proposed in Section 5.3.3. This is because the stationary distribution for each vertex is $\text{Unif}(\mathbb{B}_m)$. However, some effort is necessary to ensure that a chain based only on a random walk is mixing; mainly because the subset of $(\mathbb{B}_m)^d$ constituted of the point configurations corresponding to any specific model is not necessarily connected. Therefore a random walk can be trapped in one component of the region for a large number of iterations. Such phenomenon depends on the class of sets used to compute the nerve and on $r$.

# 6

# Simulation Experiments

We use four examples to illustrate different aspects of our methodology. The first example illustrates that our method works when the graph encoding the Markov structure is contained in the space of graphs spanned by the class of convex sets assumed when fitting the model. In the second example we apply our method to Gaussian Graphical Models. The third example shows that our methodology can be used to infer hypergraphs, and therefore a way to encode dependence features that go beyond pairwise relationships. In the fourth example we compare results obtained by using different classes of convex sets to make inferences for the same underlying graph.

## 6.1 Example 1: The Graph is in the Space Generated by $\mathcal{A}$

Let $(X_1, \ldots, X_{10})$ be a random vector and assume its distribution has the following factorization:

$$f_\theta(\mathbf{x}) = \frac{f_\theta(x_1, x_3, x_{10}) f_\theta(x_1, x_8, x_{10}) f_\theta(x_2, x_7) f_\theta(x_2, x_4, x_6) f_\theta(x_5, x_9)}{f_\theta(x_1, x_{10}) f_\theta(x_2)}. \tag{6.1}$$

The Markov structure implied by Expression 6.1 can be encoded by the geometric graph displayed in Figure 6.1. More precisely: the factorization 6.1 corresponds to a junction tree of the 1−skeleton of the complex illustrated in Figure 6.1. We specified the clique marginals as a Clayton copula, a multivariate model with density:

$$C_\theta^n(\mathbf{u})^{1+n\theta} \prod_{j=1}^{n} u_j^{-(1+\theta)} (1 + (j-1)\theta), \tag{6.2}$$

where

$$C_\theta^n(\mathbf{u}) = \left(u_1^{-\theta} + u_2^{-\theta} + \cdots + u_n^{-\theta} - n + 1\right)^{-1/\theta}. \tag{6.3}$$

Here $\mathbf{u} \in [0,1]^n$ and $\theta > 0$ and $n \geq 2$ (in this example $n = 10$); $C_\theta^n$ is the cdf. For $n = 1$ we assume Unif$(0,1)$. Properties of the Clayton copula can be reviewed in [21] and page 152 in [78]. This choice implies that all the univariate distributions are Unif$(0,1)$, in addition all clique marginals become exchangeable. For the sake of simplicity all copula parameters are specified with the same value ($\theta = 4$), this way we avoid issues regarding compatibility among the clique marginals and we can focus on the conditional independence structure.

To estimate $\theta$ we need to specify a prior and a proposal distribution for the Metropolis algorithm. We use $\theta \sim$ Exp(1) as the prior and the following random

walk proposal:

$$\theta^{(t+1)} = \begin{cases} \theta^{(t)} + \varepsilon & \text{if } \theta^{(t)} + \varepsilon \geq 0 \\ -\theta^{(t)} - \varepsilon & \text{if } \theta^{(t)} + \varepsilon < 0 \end{cases} \tag{6.4}$$

with $\varepsilon_t \sim \text{Unif}(-\beta, \beta)$, $\beta > 0$.

For the graph $\mathcal{G}(\mathcal{V})$ we used a uniform on $\mathbb{B}_2$ as prior for each element of $\mathcal{V}$ and the random walk described in Section 5.3.3.

We drew 300 samples from the distribution with Markov structure given by Expression 6.1 and clique marginals specified as above, then we fit a model with factors specified by a Clayton copula and the graph structure implied by setting $\mathcal{A} = \text{Alpha}$ in $\mathbb{R}^2$ with $r^2 = 0.15$, Algorithm 1 was applied to enforce decomposability. Posterior samples were obtained via a Metropolis Hastings algorithm. The proposals we used for the vertex set are:

- A random walk on the unit disk for each $V_i$, $1 \leq i \leq 10$ as described in Section 5.3. Here $\eta = \frac{1}{50}$. This proposal is picked with probability 0.95
- An element $k$ of $\{1, 2, \ldots, 10\}$ is chosen at random; the proposed move for $V_k$ is sampled form $\text{Unif}(\mathbb{B}_2)$. This proposal is picked with probability 0.025.
- All vertices are sampled independently from $\text{Unif}(\mathbb{B}_2)$. This proposal is picked with probability 0.025.

We obtained 1 000 samples after a burn-in period of 75 000 draws. The three models with the highest posterior probabilities are displayed in Table 6.2. The geometric graphs computed from nine posterior samples, specifically from every one hundred draws, are shown in Figure 6.2. Note that Figure 6.2 illustrates that the nerve

computed from the vertex set stabilizes after certain number of iterations while the actual position of the vertex set is allowed to vary. Inferences regarding the copula parameter are summarized in Figure 6.3.

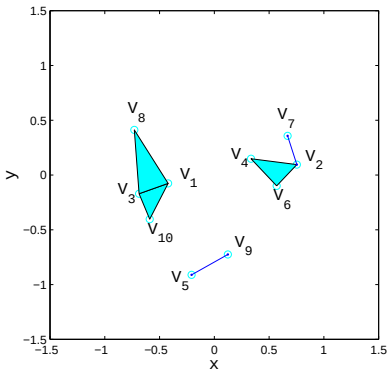


FIGURE 6.1: Geometric graph corresponding to the true model.

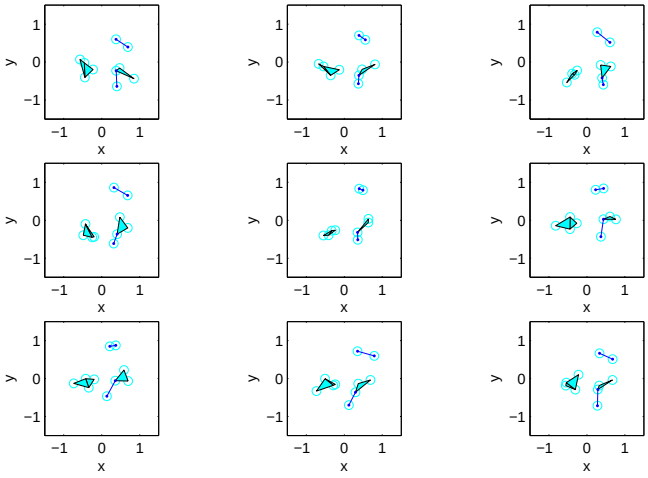


FIGURE 6.2: Geometric graphs corresponding to snapshots of posterior samples. For most samples the graphs obtained coincide with the true model.

| Graph Topology | Posterior Probability |
|---|---|
| $[1,3,10][1,3,8][2,4,6][2,7][5,9]$ | 0.964 |
| $[1,3,10][1,3,8][2,4,6][2,4,7][5,9]$ | 0.017 |
| $[1,3,10][1,3,8][2,4,6][2,7][5][9]$ | 0.015 |

Table 6.1: The 3 models with highest estimated posterior probability. In this case the true model is $[1,3,10][1,3,8][2,4,6][2,7][5,9]$ (see Figure 6.1). Here $\theta = 4$.

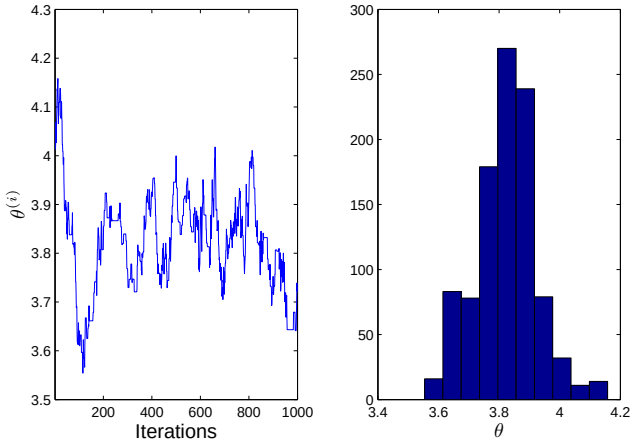


FIGURE 6.3: Traceplot for the sequence of $\theta^{(i)}$'s and histogram of posterior samples. The 0.95 credible interval for $\theta$ is $[3.62, 4.04]$. The true value of $\theta$ is 4.

Now consider a model that factorizes in the following way:

$$f_\theta(\mathbf{x}) = \frac{f_\theta(x_1,x_2,x_3,x_4)f_\theta(x_1,x_2,x_5)f_\theta(x_2,x_3,x_6)}{f_\theta(x_1,x_2)f_\theta(x_2,x_3)} \cdot \times \frac{f_\theta(x_2,x_6,x_7)f_\theta(x_6,x_8,x_{10})f_\theta(x_6,x_8,x_9)}{f_\theta(x_2,x_6)f_\theta(x_6)f_\theta(x_6,x_8)} \tag{6.5}$$

Like Expression 6.1 this factorization can be expressed as the junction tree of a weakly decomposable graph. Note that the corresponding clique hypergraph cannot be represented as an Alpha complex in $\mathbb{R}^2$ (there is a clique of size 4), but it can be computed as an Alpha complex in $\mathbb{R}^3$ (Figure 6.4). Again, we assume a Clayton

copula for the clique marginals and $\theta$ is specified as 4. We drew 500 samples from this distribution.

To fit the model, we specified the functional form of the factors as a Clayton copula and the graph structure implied by setting $\mathcal{A} = \text{Alpha}$ in $\mathbb{R}^3$ with $r^2 = 0.15$, once more, we applied Algorithm 1 to ensure decomposability. To sample from the posterior we implemented a Metropolis Hastings algorithm with random walk proposal on $\mathbb{B}_3$ as described in Section 5.3.3. We obtained 1 000 samples from the posterior after a burn-in period of 135 000 draws. Results are summarized in Figure 6.5 and Table 6.2.

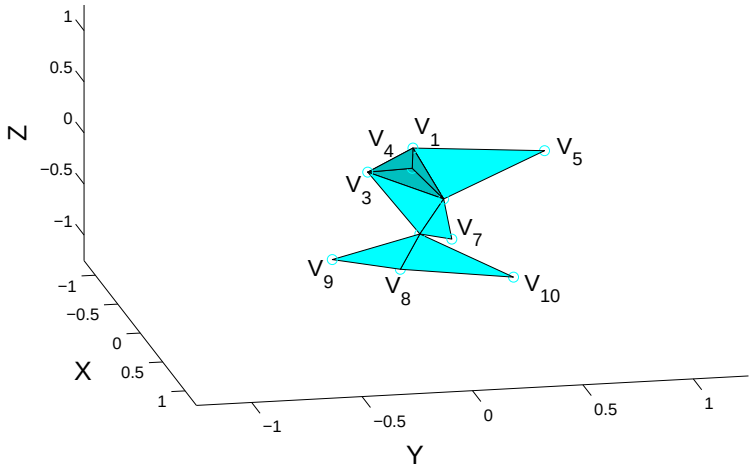


FIGURE 6.4: Geometric graph corresponding to the true model. The 4−clique is associated to a tetrahedron (darker color).

## 6.2 Example 2: Gaussian Graphical Model

We used our procedure to perform model selection for Gaussian Graphical Models, this is $X \sim \text{MN}(0, \Sigma_{\mathcal{G}})$, where $\mathcal{G}$ encodes the zeros in $\Sigma^{-1}$. We adopted a Hyper

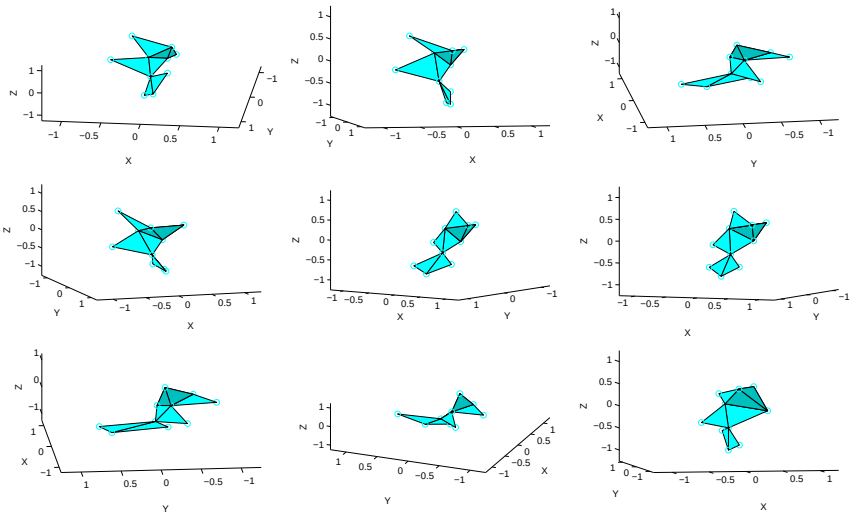

FIGURE 6.5: Geometric graphs corresponding to snapshots of posterior samples. For most samples the graphs obtained coincide with the true model. Axis were rotated to show the graphs clearly. We used a darker color to indicate the tetrahedron.

Inverse Wishart (HIW) as prior for $\Sigma \mid \mathcal{G}$. In this setting the marginal likelihood is given by

$$\mathcal{M}(\mathcal{V}) = \frac{1}{(2\pi)^{\frac{nd}{2}}} \frac{I_{\mathcal{G}(\mathcal{V})}(\delta + n, D + X^T X)}{I_{\mathcal{G}(\mathcal{V})}(\delta, D)}, \tag{6.6}$$

where $I_{\mathcal{G}(\mathcal{V})}(\delta, D)$ denotes the normalizing constant of $\text{HIW}_{\mathcal{G}(\mathcal{V})}(\delta, D)$ (Section 3.5); we are using the parametrization used by [3]. Expression 6.6 can be computed in close form when $\mathcal{G}(\mathcal{V})$ is weakly decomposable. We will work on the case when $\mathcal{G}(\mathcal{V})$ is unrestricted, which implies that the right side of Expression 6.6 has to be approximated via simulation. Since we will work with small examples, we can apply the method proposed by [3]; for dealing with a larger number of variables we recommend the method by [19]. We set $\delta = 3$ and $D = 0.4I_6 + 0.6J_6$; here $I_6$ and $J_6$ denote the identity matrix and the matrix with all elements equal to 1, respectively. We

| Graph Topology | Posterior Probability |
|---|---|
| $[1,2,3,4][1,2,5][2,3,6][2,6,7][6,8,9][6,8,10]$ | 0.946 |
| $[1,2,3,4][1,2,5][2,3,6][2,7][6,8,9][6,8,10]$ | 0.011 |
| $[1,2,3,4][1,2,5][2,3,6][2,7][6,9][6,8,10]$ | 0.011 |

Table 6.2: The 3 models with highest estimated posterior probability. In this case the true model is $[1,2,3,4][1,2,5][2,3,6][2,6,7][6,8,9][6,8,10]$ (see Figure 6.1). Here $\theta = 4$.

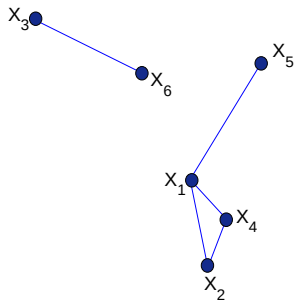


FIGURE 6.6: This graph encodes the Markov structure of the true model.

sampled 300 observations from a Multivariate Normal with conditional independence structure given by the graph shown in Figure 6.6 and the precision matrix:

$$\begin{pmatrix} 18.18 & -6.55 & 0 & 2.26 & -6.27 & 0 \\ -6.55 & 14.21 & 0 & -4.90 & 0 & 0 \\ 0 & 0 & 10.47 & 0 & 0 & -3.65 \\ 2.26 & -4.90 & 0 & 10.69 & 0 & 0 \\ -6.27 & 0 & 0 & 0 & 27.26 & 0 \\ 0 & 0 & -3.65 & 0 & 0 & 7.41 \end{pmatrix}.$$

We fit the model described in 5.3.3 using a uniform prior for each $V_i$ and $r = 0.25$ and the following proposal distributions for the Metropolis algorithm:

- A random walk on the unit disk for each $V_i$, $1 \leq i \leq 6$ as described in Section

5.3. Here $\eta = \frac{1}{50}$. This proposal is picked with probability 0.85

- An element $k$ of $\{1, 2, \ldots, 6\}$ is chosen at random; the proposed move for $V_k$ is sampled form Unif($\mathbb{B}_2$). This proposal is picked with probability 0.15.

We sampled $1,000$ observations from the posterior after a burn in of $750,000$. Results are summarized in Table 6.3

| Graph Topology | Posterior Probability |
|---|---|
| $[X_1, X_2, X_4][X_1, X_5][X_3, X_6]$ | 0.152 |
| $[X_1, X_5][X_2, X_3, X_4][X_2, X_3, X_6]$ | 0.072 |
| $[X_1, X_2, X_3, X_4, X_6][X_1, X_5]$ | 0.069 |
| $[X_1, X_4][X_2, X_4][X_2, X_3, X_6]$ | 0.055 |
| $[X_1, X_2, X_4][X_2, X_3, X_4][X_1, X_5][X_3, X_6]$ | 0.052 |

Table 6.3: The 5 models with highest estimated posterior probability. In this case the true model is $[X_1, X_2, X_4][X_1, X_5][X_3, X_6]$.

## 6.3 Example 3: Inferences on Hypergraphs

Let us consider a model such that its density factorizes in the following way:

$$f(x) = f(x_2, x_6) f(x_1 \mid x_2, x_6) f(x_3, x_4, x_5), \tag{6.7}$$

where

$$f(x_1 \mid x_2, x_6) \propto f(x_1, x_2) f(x_1, x_6). \tag{6.8}$$

Assume that the joint density is positive and continuous (see Proposition 3.1 in [67]) therefore the graph shown in Figure 6.7 encodes the Markov structure. This is an example of a distribution that is factorized according to the complete sets of the

graph. Here the factor associated to $(X_3, X_4, X_5)$ may have a different functional form from the factor corresponding to $(X_1, X_2, X_6)$. For example, assume that the potential function associated to each factor is a Clayton copula (see [21] and page 152 in [78]). For the sake of simplicity we assume that all potentials share the same value for the association parameter $\theta$, then:

$$\begin{aligned} f(x) \quad \propto \quad & C^3(x_3, x_4, x_5)^{1+3\theta} \left(C^2(x_2, x_6)C^2(x_1, x_2)C^2(x_1, x_6)\right)^{1+2\theta} \\ & \times \left(x_1, x_2, x_5, x_1^2, x_2^2, x_6^2\right)^{-(1+\theta)} (1+\theta)^4(1+2\theta). \end{aligned} \tag{6.9}$$

The question now is how to infer a factorization over complete sets of a graph in the case where grouping the factors differently may lead to a different functional form for the density. In previous examples we factorized according to the $1-$skeleton of a nerve. Since each element of the nerve is a complete set of its $1-$skeleton, it is natural to determine the factors according to the maximal simplices. For example: the Alpha complex computed from the vertex set displayed in Table 6.4 and $r = \sqrt{0.075}$ has $\{3, 4, 5\}$, $\{1, 2\}$, $\{1, 6\}$, and $\{2, 6\}$ as its maximal simplices (Figure 6.8); by associating a Clayton copula to each of these hyperedges we recover the model shown

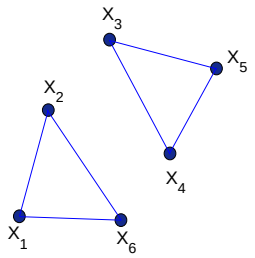


FIGURE 6.7: Graph encoding the Markov structure of the model given in Expression 6.7.

in Expression 6.9.

| Coordinate | $V_1$ | $V_2$ | $V_3$ | $V_4$ | $V_5$ | $V_6$ |
|---|---|---|---|---|---|---|
| $x$ | $-0.0936$ | $-0.4817$ | 0.0019 | 0.0930 | 0.2605 | $-0.5028$ |
| $y$ | 0.6340 | 0.7876 | 0.0055 | 0.0351 | $-0.0702$ | 0.2839 |

Table 6.4: Vertex set used for generating a factorization based on nerves.

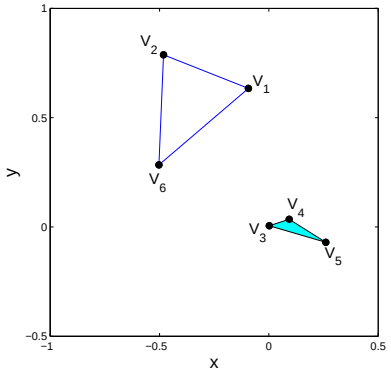


FIGURE 6.8: Alpha complex corresponding to the vertex set in Table 6.4 and $r = \sqrt{0.075}$.

Note that to fit this model one can use the same priors an proposals as in Section 5. What has changed is the way the nerve is being used: before it was an intermediate step to obtain a graph, now it is understood as a hypergraph whose maximal hyperedges represent factors.

We sampled 650 observations from model 6.9 with $\theta = 4$. We specified $\mathcal{A} =$ Alpha, $r = \sqrt{0.075}$ and $V_1, V_2, \ldots, V_6$ were assumed independent a priory with $V_i \sim$ Unif($\mathbb{B}_2$). The proposals for the vertex set are given by:

- A random walk for each $V_i$, $1 \leq i \leq 6$ as described in Section 5.3. Here $\eta = \frac{1}{50}$. This proposal is picked with probability 0.89

- A random walk for each $V_i$, $1 \leq i \leq 6$ as described in Section 5.3. Here $\eta = \frac{1}{40}$. This proposal is picked with probability 0.05

- A subset of size $k$, $1 \leq k \leq 6$ is sampled uniformly from $\{1, 2, 3, 4, 5, 6\}$; the vertices corresponding to those indices are sampled independently from Unif($\mathbb{B}_2$). This proposal is picked with probability 0.01 for each $k$.

For $\theta$ we used the same prior and proposal as in Example 6.1.

We obtained $5,000$ samples from the posterior after a burn-in period of $95,000$ iterations. The models with highest posterior probability are summarized in Table 6.5.

| Maximal Simplices | Posterior Probability |
|---|---|
| $\{3,4,5\}\,\{1,2\}\,\{2,6\}\,\{1,6\}$ | 0.374 |
| $\{3,4,5\}\,\{1,2,6\}$ | 0.259 |
| $\{3,4,5\}\,\{1,6\}\,\{1,2\}$ | 0.068 |
| $\{1,2,6\}\,\{4,5\}\,\{3,5\}\,\{3,4\}$ | 0.042 |

Table 6.5: The 4 models with highest estimated posterior probability. In this case the true model is $\{3,4,5\}\,\{1,2\}\,\{2,6\}\,\{1,6\}$.

There are two interesting differences between this example and the examples in Section 6.1. First: because in this example we are factorizing according to the complete sets of a graph, there is no need to ensure decomposability; here we are using the Alpha complexes directly to obtain the hypergraphs. Second: the models in Section 6.1 could be simulated exactly, in contrast, to sample from the true model in this section we had to apply an accept-reject algorithm (we used Unif($[0, 1]$) as

proposal). This implies that the normalizing constant had to be estimated; we used the same accept-reject algorithm to achieve this.

## 6.4 Example 4: The Graph is not Necessarily Contained in the Space Generated by $\mathcal{A}$

The simulation studies discussed in the previous sections were performed under the assumption that $\mathcal{A}$, the class of sets used to compute the nerve, was known. In this example we investigate the behavior of our methodology when the class of convex sets used when fitting the model is different from the one corresponding to the true graph. We consider 3 possibilities for the class of sets: $\mathcal{A}$ = Alpha in $\mathbb{R}^m$, with $m \in \{2, 3\}$ and $\mathcal{A} =$ Čech in $\mathbb{R}^2$. We performed 2 experiments: one when the graph is contained in each of the spaces of graphs spanned by the 3 classes, and another example where the graph could be generated by only 2 of the classes.

First consider a model that has the factorizes in the following way.

$$f_\theta(\mathbf{x}) = \frac{f_\theta(x_2, x_3, x_4) f_\theta(x_1, x_3) f_\theta(x_5)}{f_\theta(x_3)} \tag{6.10}$$

The conditional independence structure of this model corresponds to a junction tree of the $1-$skeleton of the complex displayed in Figure 6.9. Again, the clique marginals are specified as a Clayton copula with association parameter $\theta = 4$. We simulated 300 samples from this distribution.

We fitted the model assuming each of the 3 classes of convex sets using the Metropolis Hastings algorithm discussed in Section 5.3.3 with the random walk proposal on $\mathbb{B}_m$ (were $m = 2$ or 3, depending on $\mathcal{A}$). We specified $r^2 = 0.15$ and $\eta = \frac{1}{50}$;

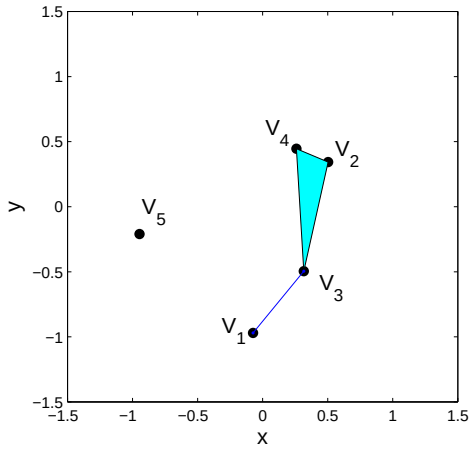


FIGURE 6.9: Graph encoding the Markov structure of the model given in Expression 6.10.

Algorithm 1 was applied to enforce decomposability. For $\theta$ we used the same prior and proposal as in Example 6.1. We obtained 1 000 samples after a burn-in period of 50 000 draws. Results are summarized in Table 6.6. Not surprisingly, the posterior mode coincided with the true model in the 3 cases.

The second model we considered has the following factorization:

$$f_\theta(\mathbf{x}) = \frac{f_\theta(x_1, x_2, x_4) f_\theta(x_1, x_3, x_4) f_\theta(x_1, x_4, x_5)}{\left(f_\theta(x_1, x_4)\right)^2} \tag{6.11}$$

The corresponding graph cannot be obtained from an Alpha complex in $\mathbb{R}^2$, but it can be computed from an Alpha complex in $\mathbb{R}^3$ (Figure 6.10) or a Čech complex in $\mathbb{R}^2$. We made the same assumptions as in the previous model regarding the clique marginals, sampled 300 observations from this distribution and fitted the model using the 3 classes of convex sets. We obtained 1 000 samples after a burn-in period of 75 000 draws; results are summarized in Table 6.7. We observed that when the graphs were obtained from Alpha complexes in $\mathbb{R}^2$, there was no clear posterior mode

| Nerve | HPP Models | Posterior |
|---|---|---|
| $\alpha$ in $\mathbb{R}^2$ | $[2,3,4][1,3][5]$ | 0.972 |
| | $[2,3,4][1,2,3][5]$ | 0.016 |
| | $[2,3,4][1,2,3][3,5]$ | 0.009 |
| $\alpha$ in $\mathbb{R}^3$ | $[2,3,4][1,3][5]$ | 0.490 |
| | $[1,3][2,3][3,4][5]$ | 0.103 |
| | $[1,3][2,3][2,4][5]$ | 0.068 |
| Čech in $\mathbb{R}^2$ | $[2,3,4][1,3][5]$ | 0.607 |
| | $[1,2,3,4][5]$ | 0.098 |
| | $[1,2,3][2,3,4][5]$ | 0.061 |

Table 6.6: Models with highest posterior probability. The table is divided according to the class of convex sets used when fitting the model. The true model has $[2,3,4]$, $[1,3]$ and $[5]$ as cliques.

(unlike the previous example, or Sections 6.1 and 6.3). The posterior mode for the Čech complex coincided with the true model. For the Alpha complex in $\mathbb{R}^3$ the second model with highest posterior probability matched the true model.

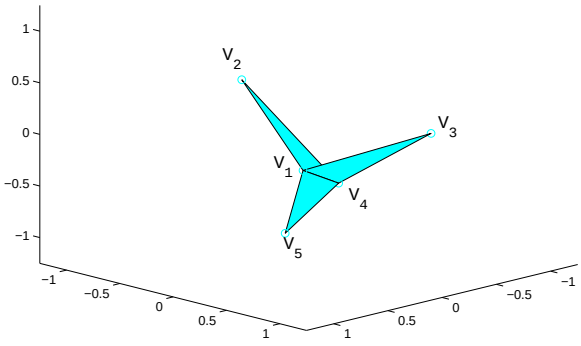


FIGURE 6.10: Graph encoding the Markov structure of the model given in Expression 6.11.

| Nerve | HPP Models | Posterior |
|---|---|---|
| $\alpha$ in $\mathbb{R}^2$ | $[1,2,3][1,3,4][1,4,5]$ | 0.383 |
| | $[1,3,4][2,3,4][1,4,5]$ | 0.136 |
| | $[1,2,3][1,2,4][1,4,5]$ | 0.104 |
| $\alpha$ in $\mathbb{R}^3$ | $[1,2,3][2,3,4][1,2,5]$ | 0.490 |
| | $[1,2,4][1,3,4][1,4,5]$ | 0.103 |
| | $[1,3,5][3,4,5][2,3,4]$ | 0.068 |
| Čech in $\mathbb{R}^2$ | $[1,2,4][1,3,4][1,4,5]$ | 0.607 |
| | $[1,2,4][1,3,4][1,3,5]$ | 0.098 |
| | $[1,2,4][1,3,4][4,5]$ | 0.061 |

Table 6.7: Models with highest posterior probability. The table is divided according to the class of convex sets used when fitting the model. The true model has $[1,2,4]$, $[1,3,4]$ and $[1,4,5]$ as cliques.

# Bibliography

[1]

[2] Kjersti Aas, Claudia Czado, Arnoldo Frigessi, and Henrik Bakken. Pair-copula constructions of multiple dependence. Technical Report 487, Technische Universität München Zentrum Mathematik, 2006. On-line at http://epub.ub.uni-muenchen.de/1855/1/paper_487.pdf.

[3] Aliye Atay-Kayis and H. Massam. A Monte Carlo method to compute the marginal likelihood in non decomposable graphical Gaussian models. *Biometrika*, 92:317–335, 2005.

[4] Thomas F. Banchoff and Clint McCrory. Combinatorial formula for normal Stiefel-Whitney classes. *Proc. Amer. Math. Soc.*, 76:171–177, 1979.

[5] R. Beach, A.O. Chan, Wu T.T, J.A. White, J.S. Morris, S. Lunagomez, R.R. Broaddus, J.P. Issa, S.R. Hamilton, and A. Rashid. Braf mutations in aberrant crypt foci and hyperplastic polyposis. *The American Journal of Pathology*, 166(4):1069–1075, 2005.

[6] Anne Berry, Jean R. S. Blair, Pinar Heggernes, and Barry W. Peyton. Maximum cardinality search for computing minimal triangulations. In Gerhard Goos, Julias V. Hartmanis, and Jan van Leeuwen, editors, *Graph-Theoretic Concepts in Computer Science*, volume 2573 of *Lecture Notes in Computer Science*, pages 1–12. Springer-Verlag, 2002.

[7] Julian Besag and Peter Clifford. Generalized Monte Carlo significance tests. *Biometrika*, 76:633–642, 1989.

[8] Julian Besag and Peter Clifford. Sequential Monte Carlo $p$-values. *Biometrika*, 78:301–304, 1991.

[9] Julian Besag, Peter J. Green, Dave Higdon, and Kerrie Mengersen. Bayesian computation and stochastic systems (with discussion). *Statist. Sci.*, 10:3–66, 1995.

[10] Julian Besag and Charles L. Kooperberg. On conditional and intrinsic autoregressions. *Biometrika*, 82:733–746, 1995.

[11] Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Statist.*, 43:1–59, 1991.

[12] Julian E. Besag. Nearest-neighbour systems and the auto-logistic model for binary data. *J. Roy.Statist. Soc. Ser. B*, 34:75–83, 1972.

[13] Julian E. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy.Statist. Soc. Ser. B*, 36(2):192–236, 1974.

[14] Julian E. Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society Series D*, 24(3):179–195, 1975.

[15] Patrick Billingsley. *Convergence of Probability Measures*. Wiley, 1999.

[16] Béla Bollobás. *Modern Graph Theory*. Springer-Verlag, 1998.

[17] Wolfgang Breymann, Alexandra Dias, and Paul Embrechts. Dependence structures for multivariate high-frequency data in finance. *Quantitative Finance*, 3(1):1–14, 2003.

[18] Stephen P. Brooks, Paolo Giudici, and Gareth O. Roberts. Efficient construction of reversible jump markov chain Monte Carlo proposal distributions. *J. Roy.Statist. Soc. Ser. B*, 65(1):3–55, 2003.

[19] Carlos M. Carvalho, H. Massam, and Mike West. Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika*, 94(3):719–733, 2007.

[20] Fan R. K. Chung. *Spectral Graph Theory*, volume 92 of *Regional Conference Series in Mathematics*. CBMS-AMS, Providence, RI, 1997.

[21] David G. Clayton. A model for association in bivariate life tables and its applications in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978.

[22] Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Statistics for Engineering and Information Science. Springer-Verlag, 1999.

[23] A. Phillip Dawid. Conditional independence in statistical theory (with discussion). *J. Roy.Statist. Soc. Ser. B*, 41(1):1–31, 1979.

[24] A. Phillip Dawid. Some misleading arguments involving conditional independence. *J. Roy.Statist. Soc. Ser. B*, 41(2):249–252, 1979.

[25] A. Phillip Dawid. Conditional independence for statistical operations. *Ann. Statist.*, 8(3):598–617, 1980.

[26] A. Phillip Dawid and Steffen L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.*, 21(3):1272–1317, 1993.

[27] Petros Dellaportas, Paolo Giudici, and Gareth Roberts. Bayesian inference for nondecomposable graphical Gaussian models. *Sankhya: The Indian Journal of Statistics*, 65(1):43–55, 2003.

[28] Stefano Demarta and Alexander J. McNeil. The $t$ copula and related copulas. *Internat. Statist. Rev.*, 73(1):111–129, 2005.

[29] A. P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

[30] Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families. *Ann. Statist.*, 7(2):269–281, 1979.

[31] Mathias Drton, Bernd Sturmfels, and Seth Sullivant. Algebraic factor analysis: tetrads, pentads and beyond. *Probability Theory and Related Fields*, 138:463–493, 2007.

[32] Rick Durret. *Random Graph Dynamics*. Cambridge Univ. Press, 2007.

[33] Herbert Edelsbrunner and John Harer. Persistent homology— a survey. In Jacob E. Goodman, Janos Pach, and Richard Pollack, editors, *Surveys on Discrete and Computational Geometry: Twenty Years Later*, volume 453 of *Contemporary Mathematics*, pages 257–282. American Mathematical Society, 2008.

[34] Herbert Edelsbrunner and John Harer. Lecture notes from the course 'Computational Topology'. Available on-line at `http://www.cs.duke.edu/courses/fall06/cps296.1/`, 2009.

[35] Herbert Edelsbrunner, John Harer, and Afra Zomorodian. Hierarchical Morse-Smale complexes for piecewise linear 2-manifolds. *Discrete and Computational Geometry*, 30:87–107, 2003.

[36] Herbert Edelsbrunner and Ernst P. Mücke. Three-dimensional alpha shapes. *ACM Transactions on Graphics*, 13:43–72, 1994.

[37] Paul Embrechts. Copulas: A personal view. *Journal of Risk and Insurance*, page Forthcoming, 2009.

[38] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, 1997.

[39] Paul Embrechts and Giovanni Puccetti. Bounds for functions of multivariate risks. *Journal of Multivariate Analysis*, 97(2):526–547, 2006.

[40] Paul D. Feigin and Sidney I. Resnick. Pitfalls of fitting autoregressive models for heavy-tailed time series. *Extremes*, 1(4):391–422, 1999.

[41] Jason P. Fine and Hongyu Jiang. On association in a copula with time transformations. *Biometrika*, 87(3):559–571, 2000.

[42] Robin Forman. A user's guide to discrete Morse theory. In Hélène Barcelo and Volkmar Welker, editors, *FPSAC'01: Proceedings of the 13$^{th}$ International Conference on Formal Power Series and Algebraic Combinatorics, A special volume of Advances in Applied Mathematics*. Advances in Applied Mathematics, 2001.

[43] Morten Frydenberg and Steffen L. Lauritzen. Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika*, 76(3):539–555, 1989.

[44] Dan Geiger, Thomas Verma, and Judea Pearl. Identifying independence in Bayesian networks. *Networks*, 20(5):507–534, 1990.

[45] Christian Genest, Kilani Ghoudi, and Louis-Paul Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552, 1995.

[46] W. Gibbs. *Elementary Principles of Statistical Mechanics*. Yale University Press, New Haven, Connecticut, 1902.

[47] Joachim Giesen and Matthias John. The flow complex: A data structure for geometric modeling. *Computational Geometry*, 39:178–190, 2008.

[48] Paolo Giudici and Peter J. Green. Decomposable graphical Gaussian model determination. *Biometrika*, 86(4):785–801, 1999.

[49] Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, December 1995.

[50] Stephen Halpering and Domingo Toledo. Stiefel-Whitney homology classes. *Annals of Mathematics*, 96(3):511–525, 1972.

[51] John M. Hammersley and Peter Clifford. Markov fields on finite graphs and lattices. (unpublished), 1971.

[52] J.M. Harrison, H.J. Landau, and L.A. Shepp. The stationary distribution of reflected Brownian motion in a planar region. *The Annals of Probability*, 13(3):744–757, 1985.

[53] Jean-Claude Hausmann. On the Vietoris-Rips complexes and a cohomology theory for metric spaces. In Frank Quinn, editor, *Prospects in Topology: Proceedings of a conference in honour of William Browder*, volume 138 of *Annals of Mathematics Studies*, pages 175–188. Princeton University Press, Princeton, NJ, 1995.

[54] David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

[55] Janet E. Heffernan and Jonathan A. Tawn. A conditional approach for multivariate extreme values. *J. Roy.Statist. Soc. Ser. B*, 66(3):497–530, 2004.

[56] Peter D. Hoff. Extending the rank likelihood for semiparametric copula estimation. *Annals of Applied Statistics*, 1(1):265–283, 2007.

[57] D. Hong, S. Lunagomez, E.E. Kim, J.H. Lee, R.S. Bresalier, S.G. Swisher, T.T. Wu, J.S. Morris, Z. Liao, R. Komaki, and J.A. Ajani. Value of baseline positron emission tomography for predicting overall survival in patient with nonmetastatic asophageal or gastroesophageal junction carcinoma. *Cancer*, 104(8):1620–1626, 2005.

[58] Mark Huber. *Spatial point processes*. Chapman & Hall/CRC Press. To appear.

[59] Rustam Ibragimov. Copula-based characterizations for higher-order Markov processes. *Econometric Theory*, page In press, 2008. Harvard Discussion Paper 2094.

[60] Harry Joe. *Multivariate Models and Multivariate Dependence Concepts*, volume 73 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, 1997.

[61] Beatrix Jones, Carlos Carvalho, Adrian Dobra, Christopher Hans, Christopher K. Carter, and Mike West. Experiments in stochastic computation for high-dimensional graphical models. *Statist. Sci.*, 20(4):388–400, 2005.

[62] Michael I. Jordan. An introduction to graphical models. Technical report, MIT Center for Biological and Computational Learning, December 1997.

[63] Michael I. Jordan. Graphical models. *Statist. Sci.*, 19(1):140–155, 2004.

[64] Ross Kindermann and James Laurie Snell. *Markov random fields and their applications*, volume 1 of *Contemporary Mathematics*. Amer. Math. Soc., 1980.

[65] Henry King, Kevin Knudson, and Neza Mramor. Generating discrete Morse functions for point data. *Experimental Math*, 14:435–444, 2005.

[66] Samual Kotz and Saralees Nadarajah. *Multivariate t Distributions and Their Applications*. Cambridge Univ. Press, 2004.

[67] Steffen L. Lauritzen. *Graphical Models*, volume 17 of *Oxford Statistical Science Series*. Oxford Univ. Press, 1996.

[68] Steffen L. Lauritzen, A. Philip Dawid, B. N. Larsen, and Hanns-Georg Leimer. Independence properties of directed Markov fields. *Networks*, 20(5):491–505, 1990.

[69] Steffen L. Lauritzen and David J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. Roy.Statist. Soc. Ser. B*, 50(2):157–224, 1988.

[70] Steffen L. Lauritzen and N. Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.*, 17(1):31–57, 1989.

[71] Paul Lévy. *Processus Stochastiques et Mouvement Brownien*. Gauthier-Villars, Paris, 1948.

[72] Vikash K. Mansinghka, Charles Kemp, Joshua B. Tenenbaum, and Thomas L. Griffiths. Structured priors for structure learning. In Leopoldo Bertossi, Anthony Hunter, and Torsten Schaub, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Twenty-second Annual Conference (UAI 2006)*, 2006.

[73] Yukio Matsumoto. *An Introduction to Morse Theory*, volume 208 of *Translations of Mathematical Monographs*. Amer. Math. Soc., 2002. Translated from Japanese by Kiki Hudson and Masahico Saito.

[74] Jason Morton and Gunnar Carlsson. Topological regularization on large graphical models. In preparation, 2009.

[75] Sach Mukherjee and Terence P. Speed. Network inference using informative priors. *PNAS*, 105(38):809–830, 2008.

[76] James R. Munkres. *Elements of Algebraic Topology*. Perseus, 1984.

[77] Daniel Q. Naiman and Henry P. Wynn. Abstract tubes, improved inclusion exclusion identities and inequalities and importance sampling. *Ann. Statist.*, 25(5):1954–1983, 1997.

[78] Roger B. Nelsen. *An Introduction to Copulas*. Springer-Verlag, 1999.

[79] Mark E. J. Newman, Steven H. Strogatz, and Duncan J. Watts. Random graph with arbitrary degree distribution and their applications. *Physical Review E*, 64:026118, 2001.

[80] Mark E. J. Newman, Steven H. Strogatz, and Duncan J. Watts. Random graph models of social networks. 99:2566–2572, 2002.

[81] Dimitris Nicoloutsopoulos. *Parametric and Bayesian non-parametric estimation of copulas*. PhD thesis, University College, 2005.

[82] David Oakes. Bivariate survival models induced by frailties. *J. Amer. Statist. Assoc.*, 84(406):487–493, 1989.

[83] Judea Pearl. *Probabilistic Reasoning in intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo,California, 1988.

[84] Mathew D. Penrose. *Random Geometric Graphs*. Oxford Univ. Press, 2003.

[85] Mathew D. Penrose and Joseph E. Yukich. Central limit theorems for some graphs in computational geometry. 11(4):1005–1041, 2001.

[86] Les A. Piegl and Wayne Tiller. *The NURBS Book*. Springer-Verlag, second edition, 1996.

[87] G. Pistone, Henry Wynn, G. Sáenz de Cabezón, and J.Q. Smith. Junction tubes and improved factorisations for bayes nets. 2009.

[88] Sidney I. Resnick. Heavy tail modeling and teletraffic data. *Ann. Statist.*, 25(5):1805–1869, 1997.

[89] Sidney I. Resnick and Holger Rootzén. Self-similar communication models and very heavy tails. *Annals of Applied Probability*, 10(3):753–778, 2000.

[90] Christian P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer-Verlag, second edition, 2001.

[91] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, second edition, 2004.

[92] G.L. Rosner, P. Mueller, S. Lunagomez, and P.A. Thompson. *Parmackinetics in clinical oncology:statistical issues*. Chapman & Hall/CRC Press, 2006.

[93] Alberto Roverato. Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Statist.*, 29(3):341–411, 2002.

[94] Paul D. Sampson and Peter Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc.*, 87(417):108–119, 1992.

[95] Huiyan Sang and Alan E. Gelfand. Hierarchical modeling for extreme values observed over space and time. *EES*, In press, 2009.

[96] James G. Scott and Carlos M. Carvalho. Feature-inclusion stochastic search for Gaussian graphical models. *J. Comput. Graph. Statist.*, 17(4):790–808, 2008.

[97] Terrence J. Sejnowski. Higher-order Boltzmann machines. In John S. Denker, editor, *AIP Conference Proceedings on Neural Networks for Computing*, volume 151, pages 398–403, 1987.

[98] Adam Silberstein, Gavino Puggioni, Alan E. Gelfand, Kamesh Munagala, and Jun Yang. Making sense of suppressions and failures in sensor networks: A Bayesian approach. In Kristoph Koch, editor, *VLDB '07: Proceedings of The* $33^{rd}$ *International Conference on Very Large Data Bases (VLDB '07)*, Vienna, AT, 2007.

[99] Terence P. Speed. A note on nearest-heighbour Gibbs and Markov probabilities. *Sankhya: The Indian Journal of Statistics*, 41(3/4):184–197, 1979.

[100] David J. Spiegelhalter and Steffen L. Lauritzen. Sequential updating of conditional probabilities on directed gaphical structures. *Networks*, 20(5):579–605, 1990.

[101] Daniel A. Spielman. Lecture notes from the course 'Spectral Graph Theory and its Applications'. Available on-line at `http://www.cs.yale.edu/homes/spielman/eigs/`, 2004.

[102] P.A. Thompson, D.J. Murray, G.L. Rosner, S. Lunagomez, S.M. Blaney, S.L. Berg, B.M. Camitta, Z.E. Dreyer, and L.R. Bomgaars. Metrotrexate pharmacokinetics in infants with acute lymphoblastic leukemia. *Cancer Chemotheraphy and Pharmacology*, 59(6):847–853, 2006.

[103] Leopold Vietoris. Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen. *Mathematische Annalen*, 97(1):454–472, 1927.

[104] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, Department of Statistics, U. C. Berkeley, 2003.

[105] Joe Whittaker. *Graphical Models in Applied Multivariate Statistics*. John Wiley and Sons, Chichester, 1990.

[106] Kevin K. F. Wong, Christopher K. Carter, and Robert Kohn. Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Biometrika*, 90(4):809–830, 2003.

[107] S. Wright. Correlation and causation. *Journal of Agricultural Research*, 1(20):557–585, 1921.

[108] S. Wright. Theory of path coefficients: a reply to Nile's criticism. *Genetics*, 1(8):239–255, 1923.

[109] S. Wright. The method of path coefficients. *Ann. Statist.*, 1(5):161–215, 1934.

# Biography

Simón Lunagómez was born in Xalapa, Mexico, on 29 December 1976. In 2001 he obtained a Bachelor degree in Actuarial Science from the National Autonomous University of Mexico (UNAM), the next year he was awarded with a Master's degree in Statistics from CIMAT \ University of Guanajuato under the supervision of Professor José Andrés Christen Gracia.

From 2003 to 2005 he worked as a statistical analyst at the Department of Biostatistics and Applied Mathematics, University of Texas MD Anderson Cancer Center; he worked under the direction of Professors Gary Rosner and Jeffrey Morris. He coauthored several publications ([102], [5] and [57]) and a book chapter on Population Pharmacokinetics [92].

In 2005 he started his doctoral studies. He coauthored a publication on computer experiments [1] and soon will submit the publication that conveys the research in this dissertation; which is work joint with his advisors, Professors Sayan Mukherjee and Robert L. Wolpert. He passed his doctoral examination on July 2009. Soon he will start a position as research officer at the London School of Economics, under the supervision of Professor Henry Wynn.