Spatial Process Models for Social Network Analysis

by

Crystal D. Linkletter

B.Sc.H., Acadia University, 2000

M.Sc., Simon Fraser University, 2003

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
in the Department
of
Statistics and Actuarial Science

SIMON FRASER UNIVERSITY

Summer 2007

APPROVAL

Name: Crystal D. Linkletter

Degree: Doctor of Philosophy

Title of dissertation: Spatial Process Models for Social Network Analysis

Examining Committee: Dr. Richard Lockhart
Chair

Dr. Randy Sitter
Senior Supervisor
Simon Fraser University

Dr. Derek Bingham
Simon Fraser University

Dr. Steve Thompson
Simon Fraser University

Dr. Tim Swartz
Internal External Examiner
Simon Fraser University

Dr. Shane Reese
External Examiner
Brigham Young University

Date Approved: _____

Abstract

There has been a recent increase in the use of network models for representing interactions and structure in many complex systems. In this thesis we introduce the use of spatial process models for the statistical analysis of networks, emphasizing applications to social networks.

The first methodology we propose is the latent socio-spatial process model. In the spirit of a random effects model, pairwise connections are assumed to be conditionally independent given a latent spatial process evaluated at observed points in a covariate space. This smooths the relationship between connections and covariates in a sample network using relatively few parameters, so the probabilities of connection for a population can be inferred. The second model that is proposed is the meta-distance model. Here, a random function is used to represent the logistic relationship between covariates and binary relations. A spatial covariance structure is assumed for the random function, where the points in space are distances between attribute pairs. A Bayesian framework is used for estimation and prediction.

While spatial process models can be very flexible and provide reasonable fit and predictions in many contexts, interpretation of these models can be complicated. To aid in the identification of important covariates, we propose a reference distribution variable selection procedure. An inert variable is appended to the data for analysis, and the posterior distribution of an "activity" parameter associated with the covariate is used as a reference distribution against which the true variables can be assessed. The approach is Bayesian, but the variable selection has a frequentist flavor.

Finally, we illustrate one important application of the proposed methodology. Local network topology can have a significant impact on contact-based processes, such

as epidemics. This is demonstrated by looking at susceptible-infected-susceptible and susceptible-infected-removed epidemic models. We explore how using a predictive network model, such as the latent socio-spatial process model, can help in predicting how a disease might spread in a population.

Acknowledgements

I have been extremely fortunate to have had such amazing opportunities and mentors over the past few years. First, I have to thank my supervisor, Randy Sitter, for his endless support and encouragement. He has shown great faith in me, and I am grateful for the insights he has shared with me, and the freedom he has given me to explore new directions.

There are many others in the Department of Statistics and Actuarial Science at Simon Fraser University who have contributed to my wonderful years here. Special thanks to Derek Bingham, who has unselfishly involved me in many research projects. Also to my "office mates," Tom Loughin, Pritam and Chunfang, for many stimulating discussions.

I am indebted to the Statistical Sciences Group at Los Alamos National Laboratory. Without all that I have learned from Dave Higdon and Nicholas Hengartner, this research would not be possible. Everyone in the group has been so welcoming, and I am happy for the many good friends I have made there.

Funding support for this research has been provided by the Department of Statistics and Actuarial Science, Los Alamos National Laboratory, and the National Science and Engineering Research Council of Canada.

On a more personal note, I am honored to have the care and support of so many wonderful people. Thanks to all my friends who are so generous, too many to mention by name. Finally, a big thank-you to my parents. I appreciate all you have done to help me reach this goal.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

There has been a recent increase in the use of network models for representing structure in many complex systems. Of particular interest is modelling interactions and relations amongst a number of "units." Units can vary widely in terms of composition and function; they may be people, countries, computers, or proteins, to name a few. There is an equally diverse set of possible relations between units: international trade conflicts, the sharing of drug paraphernalia, paths of communication, corporate hierarchy, or simply friendships.

Despite the extremely broad range of applications, it is fascinating that all these problems (and many more) can be studied within one unified framework. Since Leonard Euler first used a network to solve the Seven Bridges of Königsberg problem in 1736 (for details, see Fowler (1988), e.g.), there have been many advancements in the field from a diverse number of sources – including graph theorists, physicists, computer scientists, and sociologists. Only in the last few decades, however, have statisticians made their own contributions with respect to modelling and inference for network data. The goal of this thesis is to suggest some new methodologies for the

statistical analysis of networks, with emphasis on social networks and their applications.

In social networks, the units of interest (often called "actors") are individual people or groups of people that have some connection or relational patterns of interest. Besides being interesting in their own right, social networks can also be used to study dynamic processes evolving through society. With recent concerns of bioterrorism, and the advent of new epidemics that spread with person-to-person contact, such as SARS, there is a great need for statistical models that emulate social networks in order to better understand the impact of the underlying social structure on the spread of infectious diseases and other processes.

In the remainder of this chapter, we will introduce social networks and corresponding data more formally, including ways to define and describe some important network features. We will also begin a discussion on stochastic network modelling, and why the ability to predict local network topology could improve understanding of epidemic progressions. After a brief overview of some Bayesian methodology we will use, an outline of the rest of this thesis is given at the end of the chapter.

1.1 Social Networks

Social network data typically consists of a set of n actors and a pairwise measurement $y_{i,j}$ made on all pairs $i, j = 1, \ldots, n$. The response could be the number of times two individuals come into contact with each other in a day, a measure of the strength of their acquaintance, or a count of conflicts between two countries. In many cases, $y_{i,j}$ is simply dichotomous, indicating only the presence or absence of a relation of interest, e.g. whether or not two children are friends, two families are united through

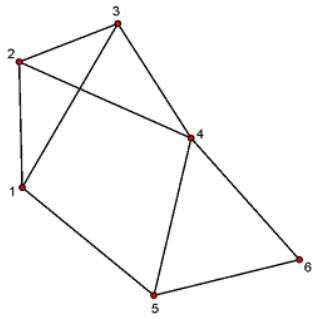


Figure 1.1: Example graph with $n = 6$ actors.

marriage, or two drug users share needles.

When the response $y_{i,j}$ is binary, the data can be represented as a *graph*, i.e. a collection of vertices (or nodes) and edges. In the context of *social* networks, each actor is a vertex in the graph and an edge exists between vertices i and j if $y_{i,j} = 1$. Figure 1.1 illustrates a graph for a small sample of $n = 6$ actors, created using the `R` package `network` (Butts, 2006).

A graph can be *undirected* or *directed*. An undirected graph is one in which $y_{i,j} = y_{j,i}$ for all pairs. Conversely, this symmetry does not hold in a directed graph. Direction in graphs can arise, for example, if Bob identifies Joe as a friend, but the favor is not returned. In this thesis, our primary focus will be on models for undirected data. This is a reasonable starting point if one is interested in how networks act as substrates for dynamic processes. In many applications, an edge acts as a channel for the flow of a disease or rumor, etc., where the transmission can occur between two adjoined vertices in either direction, thus implying an undirected underlying graph. As mentioned, there are many kinds of possible relations that one can measure, and in some cases they may be determined by a process of interest.

It is now a convenient time to introduce some language and terminology as it will be used in what follows. More algebraic definitions of graph topologies can be found

in West (2001), for example.

Definition 1.1. Two actors i and j are **connected**, denoted $i \sim j$, if there is an edge between vertex i and vertex j. Two connected actors are said to be **neighbours**, and the **neighbourhood** of vertex i, $\mathcal{N}_i$, is the set of indices corresponding to neighbours of vertex i:

$$\mathcal{N}_i = \{j : i \sim j\}.$$

Definition 1.2. A **sociomatrix** (or **adjacency matrix**), $\mathbf{Y}$, is an $n \times n$ matrix with elements $y_{i,j}$, where

$$y_{i,j} = \begin{cases} 1 & \text{if } i \sim j,\ i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

Note that for undirected networks $\mathbf{Y}$ is always a symmetric matrix since $y_{i,j} = y_{j,i}$ for all $i, j = 1, \ldots, n$. Thus, it is also convenient for notational purposes to introduce an $\binom{n}{2} \times 1$ vector, $\mathbf{y}$, with elements $\{y_{i,j}\}_{i<j}$. To illustrate, consider the graph in Figure 1.1. Here, for example, $\mathcal{N}_1 = \{2, 3, 5\}$. Also,

$$\mathbf{Y} = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

and $\mathbf{y} = (1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1)'$. In a slight abuse of terminology, we will sometimes refer to $\mathbf{Y}$ as a graph, by which we mean the graph corresponding to the

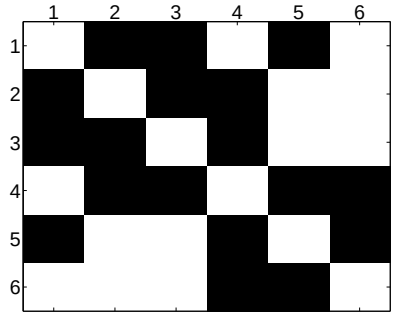


Figure 1.2: Example image representation of a graph for $n = 6$ actors.

sociomatrix $\mathbf{Y}$, as in the following definition.

Definition 1.3. The graph $\mathbf{Y}$ is a **subgraph** of $\mathbf{Y}'$, denoted $\mathbf{Y} \subset \mathbf{Y}'$, if every edge in $\mathbf{Y}$ is an edge in $\mathbf{Y}'$ and every vertex of $\mathbf{Y}$ is a vertex of $\mathbf{Y}'$.

The sociomatrix is a convenient representation for small network data sets, but may be cumbersome for large n. We propose an alternative way to visualize larger sociomatrices, as an image with $n \times n$ pixels. If each row and column of the image corresponds to the same row and column of the sociomatrix, we shade the $(i, j)^{th}$ pixel if $y_{i,j} = 1$, and leave it unshaded otherwise. This visualization technique will prove to be particularly useful later. Figure 1.2 is such a representation of the graph in Figure 1.1.

There are also a number of summary statistics available to reduce the $n \times n$ sociomatrix $\mathbf{Y}$ in ways to describe local and global network topologies. In general, we will use the notation $S(\mathbf{Y})$ for a statistic based on the sociomatrix. Many such statistics are explored in detail in Wasserman and Faust (1994) and Wasserman and Pattison (1996). There are some that are of particular interest to our discussion, however, and worth mentioning specifically here.

Definition 1.4. The **degree** of node i, d_i, is its number of neighbours. That is, $d_i = |\mathcal{N}_i|$. The **degree sequence** of a graph is $(D_0, \ldots, D_M)$, where D_k is the number of nodes in the graph with degree k, and M is the maximum observed degree in the graph. The **degree distribution** is $(D_0/n, \ldots, D_M/n)$, the *fraction* of nodes with degree k, $k = 0, \ldots, M$.

Definition 1.5. The **extended degree** of node i, e_i, is the number of its neighbour's neighbours, excluding itself. That is,

$$e_i = \sum_{j:i\sim j} |\mathcal{N}_j \setminus \{i\}|.$$

Returning to the graph in Figure 1.1, the degrees of all six actors are $d_1 = 3$, $d_2 = 3$, $d_3 = 3$, $d_4 = 4$, $d_5 = 3$, and $d_6 = 2$. Thus, the degree sequence for the graph is (0,0,1,4,1). The extended degree of actor 3, say, is $e_3 = \sum_{j\in\mathcal{N}_3} d_j - d_3 = 7$.

Definition 1.6. A **path** of length v between vertex i and vertex j is a sequence of indices $\{i_0, i_1, \ldots, i_{v+1}\}$ with $i_0 = i$ and $i_{v+1} = j$ such that

$$\prod_{t=0}^{v} y_{i_t,i_{t+1}} = 1.$$

A **cycle** is a path with $i_0 = i_{v+1}$, but no other repeated nodes.

Definition 1.7. The **minimum geodesic distance** between vertex i and j, $g_{i,j}$, is the length of the shortest path from i to j in $\mathbf{Y}$. If there is no path, then $g_{i,j} = \infty$. The **minimum geodesic sequence** is $(G_0, \ldots, G_M)$, where G_k is the number of node pairs that have minimum geodesic distance k, and M is the largest geodesic distance

that is not infinity.

Definition 1.8. A **triangle** is a cycle of length three in a graph. Three nodes i, j, and k form a triangle if $i \sim j$, $j \sim k$, and $k \sim i$.

To illustrate these last concepts, we return again to Figure 1.1. There exists many paths between actors 4 and 6, such as $\{4, 5, 6\}$, $\{4, 3, 2, 1, 5, 6\}$, or $\{4, 2, 1, 5, 6\}$. The shortest path, however, is $\{4, 6\}$, which has length 1. Note that the cycle $\{4, 5, 6, 4\}$ forms a triangle in the graph. Actor 4 is also a member of the triangle $\{4, 3, 2, 4\}$. There are a total of three triangles in the network.

While there exist a myriad of descriptive network statistics, considerably fewer network *models* are available. In the next section, we discuss some of the challenges with modelling network data.

1.2 Network Modelling

For the purpose of building statistical models, it is often assumed that the data $\{y_{i,j}\}_{i<j}$ are realizations of binary random variables. We will use the convention that the $y_{i,j}$ are called "random dyads," which is separate from the concept of an *edge*, the realization $y_{i,j} = 1$. In addition to collecting the pairwise responses, a key requirement of the models that we will be developing is the availability of covariate information for each actor in the sample. Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ denote measurements made on p characteristics of actor i. We assume this attribute information is easier to collect than the relational data $y_{i,j}$. A corresponding $n \times p$ matrix of covariates for the

sample, $\mathbf{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$, can be constructed by stacking the n covariate vectors.

In some applications, it is not obvious which of $\mathbf{Y}$ or $\mathbf{X}$ is the "response" variable or "explanatory" variables. For example, do friendships predict smoking habits or do smoking habits predict friendships? Either way, a relationship is implied between $\mathbf{Y}$ and $\mathbf{X}$, and under the assumption that the covariates are easier to collect, it is worth considering models that specify a probability distribution $P(\mathbf{Y}|\mathbf{X})$. In theory, such a model can be used to make inferences about network structure, to provide an explanation for the relationship between $\mathbf{Y}$ and $\mathbf{X}$, and to predict unobserved relations. It is this last point – prediction – that plays a central role in our exposition. In particular, we are interested in generating likely structures for $\mathbf{Y}'$ given $\mathbf{Y} \subset \mathbf{Y}'$ for a representative sample of n actors from the population.

As simple as this may sound – after all, the $\{y_{i,j}\}$ are just binary responses – there are some unique challenges associated with modelling this kind of relational data. One striking feature of social networks is that they tend to exhibit some inherent higher-order dependencies (e.g. Wasserman and Faust, 1994; Watts and Strogatz, 1998; Newman, 2003). Specifically, they usually contain a large amount of *clustering*, groups of nodes that have many within-group connections but fewer outer-group connections. This is related to the concept of *transitivity*, heuristically described as a "friends-of-my-friends-are-my-friends" phenomenon. In topological terms, this often manifests itself as a large number of triangles and "cliques" in the network (not in the formal graph-theoretical definition of *clique*). This propensity for clustering makes intuitive sense in many social contexts: If Bob and Joe are friends, and Joe and Kate are friends, then it is more likely that Bob and Kate are friends, forming a triangle with respect to the relation "friendship." One rationale for this is the notion of *homophily* by attributes (e.g. McPherson *et al.*, 2001; Handcock *et al.*, 2007), i.e. like attracts

like. Of course, this may not always be the case, but for some network data it is expected to apply.

The need for making predictive inferences has certainly been recognized, e.g. by Anderson *et al.* (1999), Hoff (2007a), and Handcock *et al.* (2007) amongst others; we will also provide our own motivation in the next section. To date, however, only modest progress has been made. The difficulty behind making predictions for social network data is in the specification of an estimable model $P(\mathbf{Y}|\mathbf{X})$ that accounts for the often-observed dependencies between the responses – such as clustering and transitivity. Currently, there are two main schools of thought when it comes to incorporating these dependencies, which we broadly classify as **exponential random graph models** and **latent factor models**. Our intention for this thesis is to propose a new, third class of models, **latent spatial process models.** Before discussing these models in more detail, we digress to present one motivation for building these models.

1.3 Social Networks and Disease Transmission

When this research began, our primary interest in social networks evolved around understanding the impact network structure has on disease propagation. In particular, we sought to provide an answer to the question: *what is the effect of local network topology on individual disease incidence?*

Historically, deterministic and stochastic models for the spread of infectious diseases assume homogeneous mixing, i.e. a susceptible individual can be infected by any infectious individual in the population (e.g. Bailey, 1975; Anderson and May, 1991; Andersson and Britton, 2000). Ball (1985), Andersson and Britton (1998) and

others have proposed generalizations of these models that allow for heterogeneity in contact and spread rates, but there has been an increasingly growing awareness that the underlying network structure plays an important role (Eubank *et al.*, 2006; Meyers, 2007). One recent example was the SARS outbreak which took place between November, 2002 and July, 2003. While mass-action models would have expected anywhere from 30,000 to 10 million cases in the first three to four months before the disease was recognized and interventions were made, the observed number of cases during this time – 782 – was significantly less (Meyers *et al.*, 2005). It was not that the infectiousness of the disease was overestimated, but rather that contact patterns restricted spread. Therefore, knowledge of the contact network structure can provide useful information about disease characteristics. This relationship between transmission models and network structure will be discussed in more detail in Chapter 6.

Unfortunately, the contact structure is rarely fully observed for a whole population of interest (e.g. for a city). Ideally, a predictive network model relating $\mathbf{Y}$ (for a sample of city residents) to demographic information $\mathbf{X}$ could, in principle, be used to generate likely social network structure, $\mathbf{Y}'$, for the entire city given urban census demographics. This in turn could contribute to understanding of how an infectious disease might spread throughout the city.

In the absence of such a model, continued growth in computing power has allowed researchers to develop complex computer programs that simulate social networks and disease paths. The EpiSims program at Los Alamos National Laboratory (Eubank *et al.*, 2004) is an example of such a simulator. EpiSims is a discrete event simulator that uses demographic information and other inputs, $\mathbf{X}$, to define rules for behavior and decision-making. Then, individuals move accordingly, coming into contact with each other, and outputting a corresponding network, $\mathbf{Y}$. A main application of EpiSims is

exploring the release of an infectious bioterrorist agent, such as smallpox, in an urban environment. Even with powerful computing, such simulators can be expensive, time-consuming and require large amounts of manual tuning. In the spirit of modelling computer experiments (e.g. Sacks *et al.*, 1989a and 1989b), a cheap statistical emulator $P(\mathbf{Y}|\mathbf{X})$ of the computer simulator could prove useful in this context as well. For a typical computer experiment, a number of input settings are selected (a design), and the computer code is run at those settings. By fitting a model to the observed input-output pairs, the output for the code can be predicted at untried inputs (see Chapter 5). Similarly, if an epidemic simulator is run using demographics from one city, say, then with an appropriate model one might be able to predict the outcome for a different city, without re-tuning and re-running the entire simulator.

Before moving on to discuss currently available network models, we present a brief overview of some important concepts in Bayesian estimation. A Bayesian framework will be used throughout this thesis to estimate parameters for network models, and introducing some details now will make presentation easier later (more details can be found in Gelman *et al.*, 2004).

1.4 An Overview of Bayesian Methodology

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_H) \in \Theta$ denote a vector of H unknown parameters in the sampling model for $\mathbf{y} = (y_1, \ldots, y_n)$ given covariates $\mathbf{X}$. For convenience, we drop the double subscript that is needed for network data. Let $[A]$ be the density of A and $[A|B]$ be the conditional density of A given B. In the Bayesian paradigm, parameters are assumed to be random variables with prior distribution $[\boldsymbol{\theta}]$, which is used to represent *a priori* beliefs in plausible parameter values before data is collected. Letting $[\mathbf{y}|\boldsymbol{\theta}]$

denote the likelihood model assumed for $\mathbf{y}$, then by Bayes' Rule

$$[\boldsymbol{\theta}|\mathbf{y}] \propto [\mathbf{y}|\boldsymbol{\theta}][\boldsymbol{\theta}]. \tag{1.1}$$

The distribution given by the left of equation (1.1) is termed the posterior distribution for $\boldsymbol{\theta}$, and it reflects updated beliefs about $\boldsymbol{\theta}$ given the data; it is this distribution that is used for inference about $\boldsymbol{\theta}$.

1.4.1 Estimation

Quite often, posterior quantities of interest involve integrating over the posterior distribution. For example, one point estimate of a function $g(\boldsymbol{\theta})$ is the posterior mean,

$$\mathbf{E}[g(\boldsymbol{\theta})|\mathbf{y}] = \int_{\Theta} g(\boldsymbol{\theta})[\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta}. \tag{1.2}$$

In practice, such integrals are usually solved numerically using Monte Carlo techniques. If $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}$ denote a large number of draws from $[\boldsymbol{\theta}|\mathbf{y}]$, then (1.2) can be approximated by

$$\mathbf{E}[g(\boldsymbol{\theta})|\mathbf{y}] \approx \frac{1}{T}\sum_{t=1}^{T} g(\boldsymbol{\theta}^{(t)}).$$

These draws from the posterior distribution can also be used to easily make other inferences about $\boldsymbol{\theta}$, e.g. for constructing credible intervals, appropriate percentiles of the posterior draws can be used.

Unfortunately, $[\boldsymbol{\theta}|\mathbf{y}]$ is usually a complex distribution with a nonstandard form. Thus, drawing realizations from the distribution requires the use of sophisticated sampling mechanisms, such as Markov Chain Monte Carlo (MCMC) algorithms (see

e.g. Gilks *et al.*, 1996). Gibbs sampling is one such tool. Given starting values $\boldsymbol{\theta}^{(0)}$, realizations from the joint posterior distribution can be generated by iteratively drawing each component of $\boldsymbol{\theta}$ from its full conditional distribution. Let $\boldsymbol{\theta}^{t'}_{-h} = (\theta^t_1, \ldots, \theta^t_{h-1}, \theta^{(t-1)}_{h+1}, \ldots, \theta^{(t-1)}_H)$. Then

- For $t = 1, \ldots, T$:
- For $h = 1, \ldots, H$:
  - Draw $\theta^{(t)}_h \sim [\theta_h | \boldsymbol{\theta}^{t'}_{-h}, \mathbf{y}]$.

If the full conditionals themselves are of a non-standard form, then a Metropolis-Hastings update can be used. Let $q(\theta^*_h | \theta^{(t-1)}_h)$ be a proposal distribution for generating proposed values of θ_h, θ^*_h, at iteration t given the preceding value $\theta^{(t-1)}_h$. For example, a $UNIF[\theta^{(t-1)}_h - c, \theta^{(t-1)}_h + c]$ for some chosen value of c might be typically used. The Metropolis-Hastings ratio for acceptance-rejection sampling is

$$R_{MH}(\theta_h) = \frac{[\theta^*_h | \boldsymbol{\theta}^{t'}_{-h}, \mathbf{y}] q(\theta^*_h | \theta^{(t-1)}_h)}{[\theta^{(t-1)}_h | \boldsymbol{\theta}^{t'}_{-h}, \mathbf{y}] q(\theta^{(t-1)}_h | \theta^*_h)}.$$

Specifically, if $[\theta_d | \cdot]$ is difficult to sample from directly, then sampling $\theta^{(t)}_h$ in step t of the Gibbs algorithm can then be done using the following algorithm:

- Draw $\theta^*_h \sim q(\theta^*_h | \theta^{(t-1)}_h)$
- Set

$$\theta^{(t)}_h = \begin{cases} \theta^*_h & \text{with probability } min\{1, R_{MH}(\theta_h)\} \\ \theta^{(t-1)}_h & \text{otherwise.} \end{cases}$$

1.4.2 Prediction

Predicting the response y^* at a new attribute vector $\mathbf{x}^*$ is quite straight-forward in the Bayesian context. Using standard techniques, the predictive distribution for y^* can be expressed as

$$[y^*|\mathbf{y}] = \int_\Theta [y^*|\mathbf{y}, \boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta}. \tag{1.3}$$

In practice, T draws from the predictive distribution can be obtained as follows:

- For $t = 1, \ldots, T$:
  - Draw $\boldsymbol{\theta}^{(t)} \sim [\boldsymbol{\theta}|\mathbf{y}]$
  - Draw $y^* \sim [y^*|\mathbf{y}, \boldsymbol{\theta}^{(t)}]$.

To simplify, realizations drawn from the posterior distribution generated for estimation can be used to generate predicted responses. Once samples from the predictive distribution are available, they can be used to get point predictions, such as $\mathbb{E}[y^*|\mathbf{y}]$, or prediction intervals.

1.4.3 Goodness-of-Fit

A natural approach to assessing the fit of a model is to generate "replicate" data and compare properties of the replicated data to the observed data. The predictive distribution evaluated at the original covariates $\mathbf{x}$ can be used to generate replicate data. That is,

$$[\mathbf{y}_{rep}|\mathbf{y}] = \int_\Theta [\mathbf{y}_{rep}|\mathbf{y}, \boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta}.$$

Given a large number of replicate draws, $\mathbf{y}_{rep}^{(1)}, \ldots, \mathbf{y}_{rep}^{(T)}$, statistical properties of the observed data, $S(\mathbf{y})$, can be compared to the distribution $S(\mathbf{y}_{rep}^{(1)}), \ldots, S(\mathbf{y}_{rep}^{(T)})$. If $S(\mathbf{y})$ is an *extreme value* in this distribution, then the model does not fit well with respect to that feature of the data. The intuition here is that if the same random mechanism that generated the observed $\mathbf{y}$ yesterday generated new observations $\mathbf{y}^{rep}$ tomorrow, then $\mathbf{y}^{rep}$ should "look like" $\mathbf{y}$ if the model is specified properly and fits the observed data well, as discussed in Gelman *et al.* (2004), for example.

1.5 Outline

Given these preliminaries, an outline of the remainder of this thesis is as follows. In Chapter 2, we will review the most well-studied models for social network data that are currently available, the exponential random graph models and the latent factor models. We will particularly highlight the point that if relating $\mathbf{Y}$ to $\mathbf{X}$ and making predictions is a goal of the network analysis, then there is room for the development of new models. Thus, a new class of models for social network analysis, latent spatial process models, which incorporate dependencies between network dyads through non-parametric function estimation, will be proposed in Chapter 3. The first model we will present in this class, the latent socio-spatial process model, will also be covered in this chapter. In Chapter 4, we will propose an alternative spatial process model, the meta-distance model. While extremely flexible, spatial process models can be difficult to interpret. So, in Chapter 5, we will develop a variable selection methodology, reference distribution variable selection, which can be used to identify important covariates in a spatial process model. We will revisit the relationship between disease transmission models and social networks in Chapter 6, demonstrating the potential

use of our proposed network models. Finally, we will conclude with a discussion of future research directions in Chapter 7.

Chapter 2

Review of Network Models

As mentioned in the introduction, developing probability models for social network data is a surprisingly challenging task. The greatest difficulty arises from trying to capture dependencies between the dyads, such as clustering and transitivity. Thus, the most obvious model for binary data – a logistic regression model (McCullough and Nelder, 1983) – is typically viewed as inappropriate, since the assumption of statistical independence between observations is violated. To date, there have been two well-studied classes of models proposed for allowing dyad dependencies: the exponential random graph models (ERGMs) and latent factor models.

Although tangential to the models we will be developing in Chapters 3 and 4, ERGMs have played a significant role in stochastic network modelling, particularly in the sociology and psychology literature. For this reason, we will devote the following section to this class. Due to some concerns that we will highlight, however, their use as predictive network models is limited. Latent factor models, on the other hand, are quite relevant to the models we propose, and thus will receive a more thorough treatment.

2.1 Exponential Random Graph Models

In spatial statistics, a result known as the Hammersley-Clifford (HC) Theorem (Besag, 1974) provides an important link between conditional distributions – that specify local stochastic dependencies – and joint probability distributions. ERGMs are a class of models that arise from extending this result to social network data. That is, the HC Theorem is a mechanism for assigning a probability to the random matrix $\mathbf{Y}$ by first specifying which network structural dependencies exist between the $y_{i,j}$.

Let $\mathcal{Y} \subset \{0,1\}^{n^2}$ denote the set of all permissible $n \times n$ sociomatrices. By permissible, we mean symmetric matrices with zeros on the diagonal. The class of ERGMs contains all models of the form

$$P(\mathbf{Y}) = c^{-1} exp\{\sum_{q=1}^{Q} \theta_q S_q(\mathbf{Y})\}, \quad \mathbf{Y} \in \mathcal{Y}, \tag{2.1}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_Q)$ is a vector of parameters and $\{S_q(\mathbf{Y})\}$ are a set of jointly sufficient statistics for $\mathbf{Y}$, specified by the user. The appropriate normalizing constant in (2.1) is

$$c = \sum_{\mathbf{Y} \in \mathcal{Y}} exp\{\sum_{q=1}^{Q} \theta_q S_q(\mathbf{Y})\}. \tag{2.2}$$

Any specification of sufficient statistics will yield a valid probability distribution for $\mathbf{Y}$, and in general, an ERGM is any network model in the form of their namesake, the canonical exponential family of distributions (Lehmann, 1983). For the alternative, i.e. for (2.1) to be motivated by the HC Theorem to correspond to specified conditional probability distributions, the $\{S_q(\mathbf{Y})\}$ are typically counts of subgraphs of $\mathbf{Y}$ (Frank and Strauss, 1986; Wasserman and Pattison, 1996). Subgraphs included in the model are referred to as *sufficient subgraphs*. In some cases, if the nodes in the graph are

divided into blocks by attributes $\mathbf{X}$, the sufficient subgraphs may also be functions of the blocking variables, that is $S(\mathbf{Y}, \mathbf{X})$ may be used in (2.1), as is done in Anderson *et al.* (1999). Other than this, however, incorporating attributes into the probability distribution for $\mathbf{Y}$ has not been a primary focus.

There are two main advantages to this class of models. First, models of the form (2.1) explicitly tie parameters to sufficient statistics, yielding an attractive interpretation. Second, courtesy of the HC Theorem there is a form of (2.1) consistent with given conditional distributions for the $y_{i,j}$. This implies an ERGM can be constructed to match beliefs on important network structures by first specifying an arbitrary set of dyads upon which each $y_{i,j}$ is dependent. For example, in the first application of the HC Theorem to social network data, Frank and Strauss (1986) assume a "Markov" dependency between the $y_{i,j}$. That is, letting $\vee$ denote the logical OR operator, they model $y_{i,j}$ as dependent on all $\{y_{r,s} : (r \vee s) = (i \vee j)\}$, i.e. on all incident dyads. This parallels a "nearest neighbour" dependency in spatial statistics (Besag, 1974). The special form of (2.1) corresponding to this assumption is known as the *Markov graph*.

There are three major criticisms of ERGMs, however, that highlight their limitations as useful inferential and predictive models. The first two points speak to the estimability of (2.1), while the last point touches on the problem of predictive inference:

i. The normalizing constant (2.2) is intractable in all but the simplest cases.

ii. Estimation of $\boldsymbol{\theta}$ is based on only *one* observation from $P(\mathbf{Y}|\mathbf{X})$.

iii. The support of $P(\mathbf{Y}|\mathbf{X})$ is sociomatrices of a fixed size, with n nodes.

To expand further, the biggest obstacle to estimating the parameters in ERGMs has been enumeration of the normalizing constant (2.2) as remarked in point (i). As

a note, models of the general form (2.1) have recently been popularized under the "p^*-model" moniker by Wasserman and Pattison (1996). This is in reference to the seminal "p_1-model" of Holland and Leinhardt (1980), which was the first log-linear model for network data, albeit one that assumed dyad independence. Though of exactly the same form as ERGMs, the term "p^*" has become synonymous with an estimation technique known as pseudo maximum likelihood estimation (PMLE; Besag, 1975). The wide-spread appeal of this approach is due in large part to its accessibility to non-statisticians; Strauss and Ikeda (1990) show PMLE is as easy to implement as standard logistic regression for binary network data – and most importantly, it does not require estimation of (2.2). The statistical properties of these estimators are unknown, however, and their use in this context has been criticized by Besag (2000) and Snijders (2002). Alternative MCMC maximum likelihood techniques have instead been recommended by these authors, and do show some promise (as well as highlight the sometimes poor performance of PMLE for network data). To improve estimability of ERGMs, there has also been a lot of recent consideration given to the choice of sufficient subgraphs (Snijders *et al.*, 2006), and extensions such as curved exponential families have been explored (Hunter and Handcock, 2006). Instabilities are to be expected, though, due to point (ii).

Even if estimation of these models becomes feasible, of additional concern to us is point (iii) above. ERGMs are a holistic modelling approach, and consciously model *conditional* probabilities instead of *marginal* probabilities. To illustrate a difficulty with this, consider for example trying to predict the probability that John is friends with Alice. This probably then, say, depends on knowing everyone that John is friends with (except for Alice), everyone that Alice is friends with (except for John), and the number of pairs of people in the population that have a mutual friend. If this

information is not known, the ERGM says nothing about the chances John and Alice are friends. Thus, the model implicitly assumes that the network is observed for the whole population of interest, and is not obviously useful for making predictions on unsampled pairs.

Given these concerns about ERGMs, other more local approaches to modelling dyad dependencies have been pursued.

2.2 Latent Factor Models

An alternative to modelling *conditional* probabilities, as ERGMs do, is to model *marginal* probabilities

$$\pi_{i,j} = P(y_{i,j} = 1).$$

This is the approach taken in a number of models that we categorize as latent factor models. Given the $\pi_{i,j}$, the random dyads are assumed to be conditionally independent Bernoulli random variables. All of the complexity of these models, then, lies in the specification of a model for $\pi_{i,j}$. Letting $\boldsymbol{\pi}$ denote the $\binom{n}{2} \times 1$ vector with elements $\{\pi_{i,j}\}_{i<j}$, the likelihood corresponding to a latent factor model is simply given by

$$\mathbf{L}(\mathbf{y}|\boldsymbol{\pi}) = \prod_{i<j} \pi_{i,j}^{y_{i,j}} (1 - \pi_{i,j})^{1-y_{i,j}}. \tag{2.3}$$

Clearly this is much more convenient to work with than (2.1).

In order to incorporate unconditional dependencies between the $y_{i,j}$, latent node-specific random effects are introduced in the model for $\pi_{i,j}$. Specifically, in this section we consider models of the form

$$\eta_{i,j} = logit(\pi_{i,j}) = \beta_0 + \mathbf{x}_{i,j}\boldsymbol{\beta}' + \zeta_{i,j}, \tag{2.4}$$

where $logit(a) = log(a/(1-a))$. In this class, statistical dependencies between the $\eta_{i,j}$ (and hence, the $y_{i,j}$) are captured by the term $\zeta_{i,j}$. This emphasis on modelling "noise," i.e. the lack-of-fit of the logistic-linear regression, is an interesting feature of latent factor models. The covariates $\mathbf{x}_{i,j} = (x_{i,j,1}, \ldots, x_{i,j,p})$ are presumed to be pair specific, and while there may be some attributes that are inherently pairwise, e.g. the distance between the capitals of two countries, typically $\mathbf{x}_{i,j}$ is constructed using individual covariates, $\mathbf{x}_i$ and $\mathbf{x}_j$. In keeping with the idea of homophily by attributes, we assume that $\mathbf{x}_{i,j} = (|x_{i,1} - x_{j,1}|, \ldots, |x_{i,p} - x_{j,p}|)$. The coefficients β_0 and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ are parameters to be estimated, as well as any other unknowns in $\zeta_{i,j}$.

Before discussing two particular cases of (2.4), we remark that other random effects models not of this form have been considered for modelling dependencies between network dyads. For example, Wong (1987) and Gill and Swartz (2004) use random effects to incorporate dependencies into Holland and Leinhardt's (1981) p_1-model. However, in what follows, we will restrict our attention to models of the form (2.4).

2.2.1 Latent Space Model

The first model we consider is the latent space model of Hoff, Raftery and Handcock (2002). For convenience, we will label this the HRH model. In order to capture network structure that is not accounted for by the logistic-linear function of observed attributes, this model posits the existence of a latent d-dimensional "social" space, $\mathbb{R}^d$, in which each actor has an unobserved position $\mathbf{z}_i$. We denote the collection of positions for all n individuals by $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)'$. The intuition behind this model is appealing. Actors that are close in the latent space are assumed to have a higher probability of connection than actors that are far apart, where closeness can

be measured, for example, using Euclidean distance; this implies a kind of homophily by unobserved characteristics. Specifically, the HRH model is given by

$$\eta_{i,j} = \beta_0 + \mathbf{x}_{i,j}\boldsymbol{\beta}' - \|\mathbf{z}_i - \mathbf{z}_j\|. \tag{2.5}$$

While d, the dimension of the social space, can be chosen to be any dimension, typically $d = 2$ is used, allowing the latent positions to be easily visualized. In the HRH model, the term $\zeta_{i,j} = -\|\mathbf{z}_i - \mathbf{z}_j\|$ is not interpretable as an "error" term since it is always negative. It does, however, aim to model structural features in the network that are not explained by the covariates. The use of distance measures (i.e. the absolute value on the marginal covariate differences and the Euclidean distance on the latent positions) induces a form of transitivity. If, for example, $\|\mathbf{z}_i - \mathbf{z}_k\|$ and $\|\mathbf{z}_j - \mathbf{z}_k\|$ are both small, then by the triangle inequality $\|\mathbf{z}_i - \mathbf{z}_j\|$ cannot be too large. It also suggests that actors close to each other in social space will relate similarly with other people, i.e. if $\mathbf{z}_i \approx \mathbf{z}_j$, then $\|\mathbf{z}_i - \mathbf{z}_k\| \approx \|\mathbf{z}_j - \mathbf{z}_k\|$ for all k (see Hoff, 2007b, for more on this point).

With $d = 2$, there are $2n + p + 1$ parameters to be estimated in the HRH model: β_0, $\boldsymbol{\beta}$, and $\mathbf{z}_i \in \mathbb{R}^2$, $i = 1, \ldots, n$. To proceed with Bayesian estimation, priors must be specified for each unknown. One choice of priors is

$$\begin{aligned} \beta_h &\sim N(0, \psi_\beta); \quad h = 0, \ldots, p \\ \mathbf{z}_i &\sim BVN(\mathbf{0}, \psi_z \mathbf{I}_2); \quad i = 1, \ldots, n, \end{aligned}$$

where $\mathbf{I}_2$ is the 2×2 identity matrix. In combination with the likelihood (2.3), the resulting full conditionals for each of the parameters are not of a standard form, but a Metropolis-Hastings update can be used.

Overall, implementation of this method is a little tricky, and we tried to stay close to the author's suggested approach. To select starting values for the latent

positions, for instance, the minimum geodesic distance is calculated between all pairs, and then a multidimensional scaling is used to reduce these distances to the desired two-dimensional positions. These are then used in an optimizer (we use `fminsearch` in MATLAB, for example) to yield a maximum likelihood estimate of $\mathbf{Z}$. Also, note that one inherent difficulty with estimating $\mathbf{Z}$ is that any rotation, reflection or translation of the positions yields the same likelihood value. Following Hoff *et al.* (2002), at each iteration t of the MCMC algorithm, the $\mathbf{Z}^{(t)}$ that is saved for inference is the Procrustean transformation of the currently accepted $\mathbf{Z}^*$ around a fixed set of positions $\mathbf{Z}_0$,

$$\mathbf{Z}^{(t)} = \mathbf{Z}_0\mathbf{Z}^{*'}(\mathbf{Z}^*\mathbf{Z}_0'\mathbf{Z}_0\mathbf{Z}^{*'})^{-1/2}\mathbf{Z}^*,$$

assuming all of the positions are centered at the origin. Hoff *et al.* (2002) take $\mathbf{Z}_0$ to be the maximum likelihood estimator of $\mathbf{Z}$ used as the starting value. To illustrate the performance of the methodology, we consider the following example.

Florentine Family (FF) Example. Padgett and Ansell (1993) collected data on marriages between 16 prominent Florentine families in the fifteenth century. The network is displayed in Figure 2.1, with $y_{i,j} = 1$ if there was a marriage between families i and j. The wealth of each family, x_i, $i = 1, \ldots, n$ is an available covariate. This network is analyzed without a covariate in Hoff *et al.* (2002); the following is our own implementation including x. For analysis, the x_i are scaled to $[0, 1]$, and pairwise variables $x_{i,j} = |x_i - x_j|$ are constructed (note $x_{i,j} \in [0, 1]$ as well). Normal priors with $\psi_\beta = \psi_z = 10$ are used for each of the parameters. Due to difficulties with visualizing the latent social positions for nodes with no connections inherent in the methodology, one family (labeled 12 in Figure 2.1) that has no marriages is removed

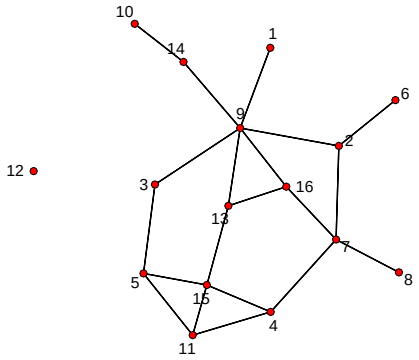


Figure 2.1: FF Example: Padgett and Ansell's (1993) Florentine Family network

from the network. Using the specified model and priors, an MCMC algorithm is run for posterior exploration. We find that it takes on the order of a million iterations for convergence. Draws from the posterior distribution of the latent social positions in $\mathbb{R}^2$ for the 15 families included in the analysis are plotted in Figure 2.2. The red points in the centers of the scattered points are the posterior mean position for each family.

Using results from the previous chapter, the posterior expectation of $\pi_{i,j}$ can be calculated as

$$\hat{\pi}_{i,j} = \mathbb{E}[\pi_{i,j}|\mathbf{y}, \mathbf{x}] = \frac{1}{T}\sum_{t=1}^{T} \frac{exp\{\eta_{i,j}^{(t)}\}}{1 + exp\{\eta_{i,j}^{(t)}\}},$$

where

$$\eta_{i,j}^{(t)} = \beta_0^{(t)} + |x_i - x_j|\beta_1^{(t)} - \|\mathbf{z}_i^{(t)} - \mathbf{z}_j^{(t)}\|,$$

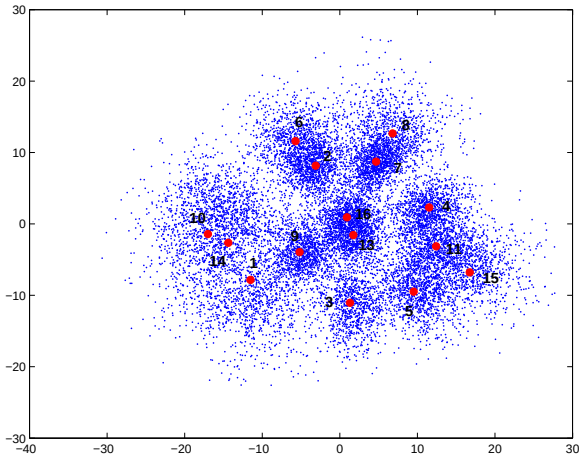
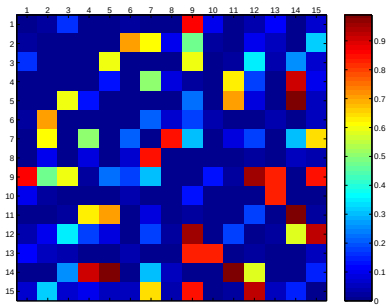


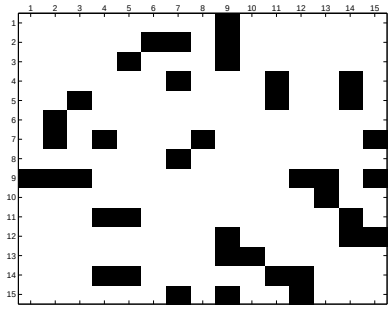
Figure 2.2: FF Example: Posterior draws and mean for latent positions

given draws $\beta_0^{(t)}$, $\beta_1^{(t)}$, and $\mathbf{Z}^{(t)}$, $t = 1, \ldots, T$ from the posterior distribution. Figure 2.3(a) is a plot of $\hat{\pi}_{i,j}$ for all pairs $i, j = 1, \ldots, 15$. We developed this plot to correspond to the image representation of a sociomatrix, but with pixels colored according to the posterior probability of connection for the $(i, j)^{th}$ pair. The original sociomatrix is plotted alongside for comparison. In the probability plot, "hotter" colors correspond to larger probabilities, while "cooler" colors are smaller. When the estimated probabilities are plotted alongside the sociomatrix in this way, a feature that stands out is how similar they are, i.e. high posterior probabilities are estimated where connections were observed. With $O(n)$ parameters, the HRH model is essentially able to reproduce the observed network $\mathbf{Y}$, and the primary interest seems to be on interpretation of relative social positions as depicted in Figure 2.2. This leads to an interesting observation on goodness-of-fit for network models.

As discussed in Chapter 1, to investigate goodness-of-fit we generate a large number, T, of replicate networks, $\mathbf{Y}_{rep}$. Then, graph statistics based on the sociomatrix, $S(\mathbf{Y})$, are compared to the replicate distribution $S(\mathbf{Y}_{rep}^{(1)}), \ldots, S(\mathbf{Y}_{rep}^{(T)})$. Figure 2.4(a) shows the degree sequence of $\mathbf{Y}$ for the Florentine Family network as a solid line,



(a) Posterior mean probabilities of connection



(b) Image representation of sociomatrix

Figure 2.3: FF Example: Sociomatrix and posterior mean probabilities

with pointwise 90% intervals of the replicate distribution in dashed lines. The box-plots show each element-wise distribution over all replicate draws. In Figure 2.4(b), the number of edges in $\mathbf{Y}$ is marked by a vertical line, and the smoothed histogram shows the distribution of the number of edges in the replicated sociomatrices. Recall that if the observed statistics are extreme values in the replicate distribution, then this suggests a lack-of-fit. This is clearly not the case here. However, the posterior mean plot (Figure 2.3(a)) suggests the fit almost reproduces $\mathbf{Y}$, so it is natural that it will do extremely well on these goodness-of-fit measures, i.e. the replicate sociomatrices essentially are $\mathbf{Y}$. When the variability of $\mathbf{Y}$ around the posterior mean probabilities is considered, there is more objective evidence of this over fitting.

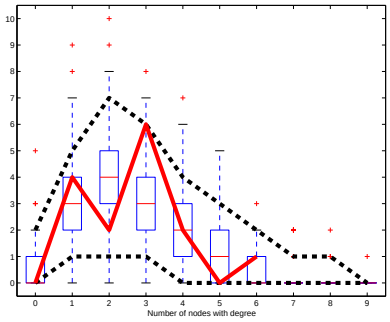
Consider a very basic χ^2-test:

$$H_0: \quad y_{i,j} \text{ independent } BER(\hat{\pi}_{i,j})$$

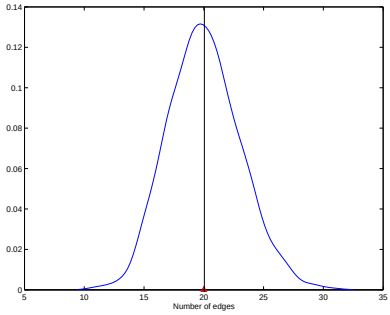
$$H_a: \quad y_{i,j} \text{ have a distribution other than } H_0$$

We compared the test statistic

$$X_0 = \sum_{i<j} \frac{(y_{i,j} - \hat{\pi}_{i,j})^2}{\hat{\pi}_{i,j}(1 - \hat{\pi}_{i,j})},$$



(a) Replicate degree sequence



(b) Replicate number of edges

Figure 2.4: FF Example: Goodness-of-fit statistics

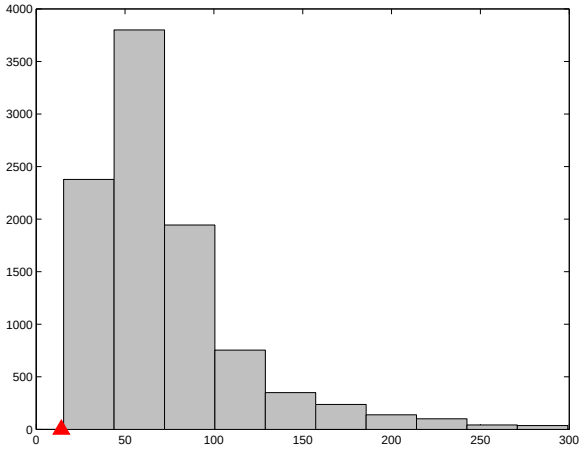


Figure 2.5: FF Example: χ^2 test statistic and null distribution

based on the observed sociomatrix $\mathbf{Y}$ to the same statistic calculated using a large number of sociomatrices drawn from the null distribution. The resulting p-value is 1, as can be seen in Figure 2.5. We note that in this plot the test statistic X_0 is marked by a triangle, and the histogram represents draws from the null distribution of the statistic. This suggests that the posterior estimates of the $\pi_{i,j}$ are indeed tailored to $\mathbf{Y}$; so much, in fact, that any other sociomatrix generated exhibits more variability around the posterior mean.

The Florentine Family example illustrates a fundamental difficulty with estimating models for binary network data. The best "fit" to the sociomatrix is the one with $\hat{\pi}_{i,j} = 1$ or 0 if $y_{i,j} = 1$ or 0, respectively. Models that achieve this fit may provide useful interpretations of the network structure – for example, parameters such as the latent positions serve as representations of unobserved characteristics and describe the network as a function of these. When it comes to prediction, however, the node-specific random effects do not model common underlying patterns. Consider predicting the probability that two individuals, i_0 and j_0, are connected, π_{i_0,j_0}. If there are no observed responses available for either of these actors, it is not possible to estimate $\mathbf{z}_{i_0}$ and $\mathbf{z}_{j_0}$. Thus, one would imagine that the best that could be done is to integrate over the prior distribution of positions. This could generate a lot of prediction error, particularly if the prior $[\mathbf{Z}]$ is taken to be non-informative. We anticipate that the scale of the parameters may also become uninterpretable in this case. For example, an alternative specification of the HRH model is the latent position cluster model, given in Handcock *et al.* (2007):

$$\eta_{i,j} = \beta_0 + \mathbf{x}_{i,j}\boldsymbol{\beta}' - \beta_z \|\mathbf{z}_i - \mathbf{z}_j\|.$$

Here it is assumed

$$\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{z}_i\|^2\right)} = 1$$

so the scales of all the coefficients are identifiable. If the model is used to make predictions, any proposed position for an actor outside the sample will alter this scaling.

2.2.2 Multiplicative Factor Models

The second latent factor model we consider is the multiplicative latent factor (MF) model proposed by Hoff (2007a). Here, the structure in the network is modeled using

$$\zeta_{i,j} = \mathbf{u}_i \mathbf{D} \mathbf{u}_j^{'} + \epsilon_{i,j}, \tag{2.6}$$

where the vectors $\mathbf{u}_i$, $i = 1, \ldots, n$ allow for differences in individual tendencies to connect. One rationale for this alternative model is that it reduces the confounding between homophily by attributes and stochastic equivalence (Hoff, 2007b).

Hoff (2007a) motivates model (2.6) by a matrix decomposition. Suppose $\mathbf{Z}$ is an $n \times n$ matrix that represents the lack-of-fit in $\eta_{i,j}$ from the logistic-linear regression. This matrix can be modeled as having a signal component and an error component, i.e. $\mathbf{Z} = \mathbf{M} + \mathbf{E}$. In the case of undirected networks, these are assumed to be symmetric matrices. Thus the decomposition

$$\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{U}'$$

can be found, where $\mathbf{U}$ is an orthonormal matrix and $\mathbf{D}$ is a diagonal matrix of real numbers. Putting all this together yields model (2.6). Typically one would use only the first k eigenvectors; we only consider the case $k = 2$. Hoff (2007a) emphasizes modelling directed graphs, and thus relies on a singular value decomposition of $\mathbf{M}$ instead of an eigenvalue decomposition. No results on analyzing undirected graphs using this model are currently available, so we derive the following algorithm. Note that unlike for the latent space (HRH) model, actors with all zero observations can be better accommodated here due to the structure of the orthonormal matrix. Actors with no observations, on the other hand, cannot be included, as we will discuss more below.

The full specification of the MF model is

$$\eta_{i,j} = \beta_0 + \mathbf{x}_{i,j}\boldsymbol{\beta}' + \mathbf{u}_i\mathbf{D}\mathbf{u}_j^{'} + \epsilon_{i,j}, \tag{2.7}$$

where given the $\eta_{i,j}$, it is again assumed that the $y_{i,j}$ are conditionally independent with likelihood (2.3). The quantities $\boldsymbol{\eta} = \{\eta_{i,j}\}_{i<j}$, β_0, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ in (2.7) are $\binom{n}{2} + p + 1$ parameters to be estimated, as well as the $n \times 2$ orthonormal columns $\mathbf{U}$, and the 2×2 diagonal matrix $\mathbf{D}$ with diagonal elements d_1 and d_2.

The inclusion of the error term in (2.7) in some sense simplifies model specification. In particular, if we begin with the assumption that $\epsilon_{i,j} \sim N(0, \phi)$ then

$$\eta_{i,j} \sim N(\beta_0 + \mathbf{x}_{i,j}\boldsymbol{\beta}' + \mathbf{u}_i\mathbf{D}\mathbf{u}_j^{'}, \phi).$$

Using vague $N(0, \psi)$ priors for β_h, $h = 0, \ldots, p$ and d_k, $k = 1, 2$, the full conditional distributions of these parameters are recognizable. Specifically,

$$\begin{aligned}
\beta_0|\cdot &\sim N\left[\frac{\psi\sum_{i<j} f_{i,j}(\beta_0)}{\psi\binom{n}{2} + \phi}, \frac{\psi\phi}{\psi\binom{n}{2} + \phi}\right] \\
\beta_h|\cdot &\sim N\left[\frac{\psi\sum_{i<j} f_{i,j}(\beta_h)}{\psi\sum_{i<j} x_{i,j,h}^2 + \phi}, \frac{\psi\phi}{\psi\sum_{i<j} x_{i,j,h}^2 + \phi}\right]; \quad h = 1, \ldots, p \\
d_k|\cdot &\sim N\left[\frac{\psi\sum_{i<j} f_{i,j}(d_k)}{\psi\sum_{i<j}(u_{ik}u_{jk})^2 + \phi}, \frac{\psi\phi}{\psi\sum_{i<j}(u_{ik}u_{jk})^2 + \phi}\right]; \quad k = 1, 2,
\end{aligned}$$

where

$$\begin{aligned}
f_{i,j}(\beta_0) &= \eta_{i,j} - \sum_{d=1}^{p} \beta_d x_{i,j,d} - \mathbf{u}_i\mathbf{D}\mathbf{u}_j^{'}), \\
f_{i,j}(\beta_h) &= (\eta_{i,j} - \beta_0 - \sum_{d\neq h} \beta_d x_{i,j,d} - \mathbf{u}_i\mathbf{D}\mathbf{u}_j^{'})x_{i,j,h}, \\
f_{i,j}(d_k) &= (\eta_{i,j} - \beta_0 - \sum_{d=1}^{p} \beta_d x_{i,j,d} - \sum_{r\neq k} u_{ir}u_{jr}d_r)u_{ik}u_{jk}d_k.
\end{aligned}$$

Sampling $\mathbf{U}$ from its full conditional requires a bit more thought. The prior we choose to assign to $\mathbf{U}$ is a uniform distribution (with respect to Haar measure) over

the Stiefel manifold, i.e. the space of $n \times 2$ orthonormal columns. Details on finding the appropriate normalizing constant for this distribution can be found in James (1954), for example, but this is unnecessary here since the constant cancels out in the Metropolis-Hastings (MH) step. To propose a new $\mathbf{U}^*$ at step t of the algorithm given $\mathbf{U}^{(t-1)}$, we use an idea from Villani and Larsson (2006) and apply a random Givens rotation (see e.g. Golub and Van Loan, 1996) to the current state. That is, we draw

$$\omega \sim U(-\frac{\pi}{2}, \frac{\pi}{2}),$$

and propose

$$\mathbf{U}^* = \mathbf{U}^{(t-1)} \begin{pmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{pmatrix}.$$

This results in a random counterclockwise rotation of the plane spanned by the columns of $\mathbf{U}$. Because the proposal for ω is symmetric around 0, the proposal distribution for $\mathbf{U}$ is also symmetric. Thus, the resulting MH ratio for this update at iteration t is

$$R_{MH}(\mathbf{U}) = \frac{[\boldsymbol{\eta}|\mathbf{U}^*, \beta_0^{(t)}, \boldsymbol{\beta}^{(t)}, \mathbf{D}^{(t)}]}{[\boldsymbol{\eta}|\mathbf{U}^{(t-1)}, \beta_0^{(t)}, \boldsymbol{\beta}^{(t)}, \mathbf{D}^{(t)}]}.$$

To finish the updates, new values are required for $\boldsymbol{\eta}$. Following Hoff (2007a) on this step, proposals are generated from

$$\eta_{i,j}^* \sim N(\beta_0^{(t)} + \mathbf{x}_{i,j}\boldsymbol{\beta}^{(t)'} + \mathbf{u}_i^{(t)}\mathbf{D}^{(t)}\mathbf{u}_j^{(t)'}, \phi)$$

and accepted according to the MH ratio

$$R_{MH}(\eta) = \frac{[\mathbf{y}|\boldsymbol{\eta}^*]}{[\mathbf{y}|\boldsymbol{\eta}]},$$

where $[\mathbf{y}|\boldsymbol{\eta}]$ is the likelihood (2.3).

Example: International Conflict (IC) Network. To illustrate the multiplicative factor methodology, we consider a network of international conflicts between 130 nations during the years 1990-2000. More details on the network can be found in Ward and Hoff (2007). The network can actually be analyzed as directed, since information on which country initiated each aggression is available; Hoff (2007a) focuses on modelling the directed network. To make this example applicable to the undirected case, we take $y_{i,j} = y_{j,i} = 1$ if either country i or country j initiated aggressive behavior against the other. The graph of this undirected network is shown in Figure 2.6. We consider two covariates,

1. The log population of country i, x_{i1}, and

2. The polity score of country i, x_{i2}.

The polity score is a measure of the political leanings of a country. Low scores correspond to authoritarian countries, while higher scores are given to more democratic countries (Ward and Hoff, 2007). As in the Florentine Family example, we scale the attributes to $[0, 1]$ and define $x_{i,j,1} = |x_{i1} - x_{j1}|$ and $x_{i,j,2} = |x_{i2} - x_{j2}|$.

Using the above algorithm, with $\phi = 1$ and $\psi = 1000$, we draw a large number of samples from the posterior distribution; we find that again on the order of a million iterations is needed. Figure 2.7(a) is a plot of the posterior mean probability of connection for each pair $i, j = 1, \ldots, n$,

$$\hat{\pi}_{i,j} = \frac{1}{T} \sum_{t=1}^{T} \frac{exp\{\eta_{i,j}^{(t)}\}}{1 + exp\{\eta_{i,j}^{(t)}\}},$$

where $\eta_{i,j}^{(t)}$, $t = 1, \ldots, T$, are draws from the posterior distribution. Recall that the log-odds themselves are treated as parameters in this model. The sociomatrix for the observed network is given alongside for comparison (Figure 2.7(b)). With probabilities shaded in reds and yellows being greater than the ones in blue, we see that the

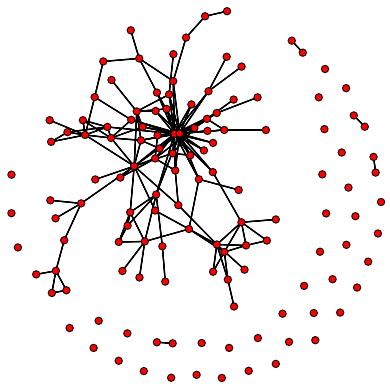
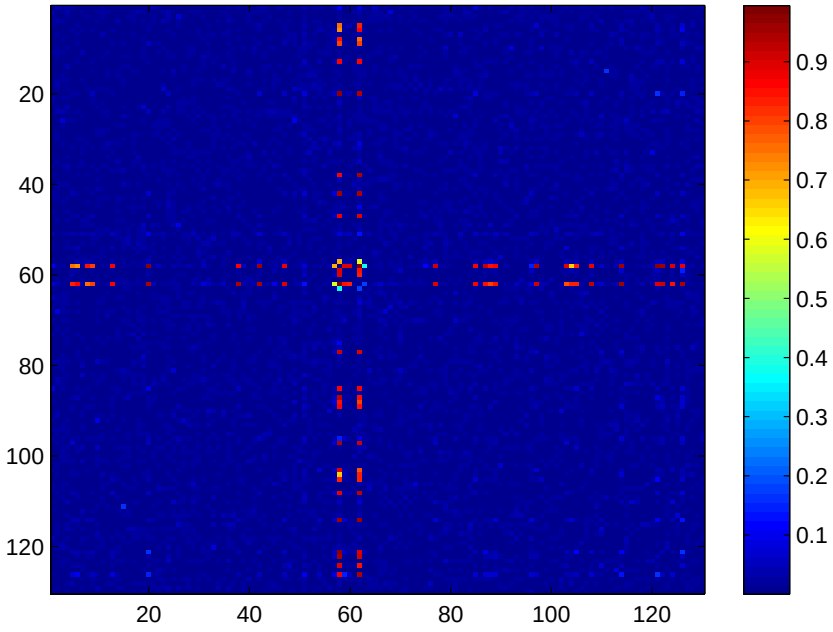


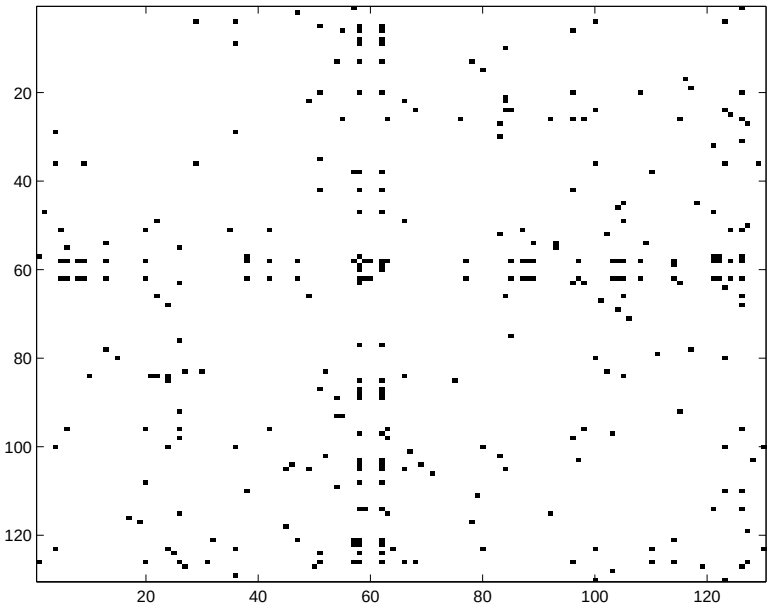
Figure 2.6: IC Example: Network of International conflicts.

posterior estimates of connection probabilities imitate the pattern of the observed network connections, as was the case with the HRH model. The χ^2 test shows additional evidence of this over fit (Figure 2.8).

As an alternative fit assessment, Hoff (2007a) proposes looking at the number of missing links that the model correctly predicts, i.e. consider deleting some links in the network and treating them as missing data. The model is then fit with these observations excluded, but note that every actor must have at least some connection information included to be able to fit the multiplicative factor model (Hoff, 2007a). Given the predictions $\hat{\pi}_{i,j}$ (estimated on the reduced network), the missing links are predicted by looking at a threshold probability. More explicitly, for some probability p_τ, predict $\hat{y}_{i,j} = 1$ if $\hat{\pi}_{i,j} > p_\tau$, and 0 otherwise, and then look at what proportion of missing links correctly predicted. With respect to this measure of fit, Hoff (2007a) shows the model does quite well. It is fairly clear from looking at Figure 2.7(a) that there should exist a threshold probability such that missing links will be recovered



(a) Posterior mean probabilities



(b) Observed sociomatrix

Figure 2.7: IC Example: Multiplicative latent factor model fit

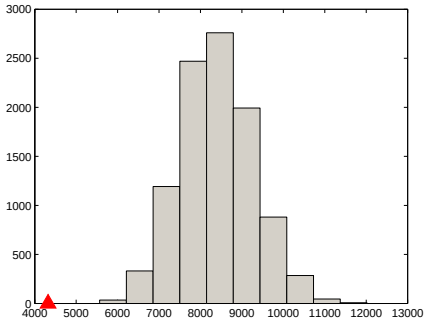


Figure 2.8: IC Example: χ^2 test for MF model

in this manner most of the time, i.e. imagine a plane slicing the three-dimensional histogram (with bar heights corresponding to probability color). Thus, while this suggests the model can be useful for predicting missing links between actors within a sample, it is not intended for predicting connections outside the sample.

We conclude this chapter with a few summary observations about these latent factor models. As best that we can tell, these models are not intended as predictive models, at least not for making predictions at a local, pair-specific level. It is possible that randomly generating random effects for a population and constructing corresponding networks may capture global properties, but we do not pursue this here. What these models are intended to do, however, they do very well: provide an intuitive interpretation of the structure in the network. The clustering version of the HRH model (Handcock *et al.*, 2007), for example, which places a mixture of multivariate normal priors on $\mathbf{Z}$, further improves the interpretation of network clustering. These models also provide estimates for the coefficients $\boldsymbol{\beta}$, so some inference on the importance of the observed covariates $\mathbf{X}$ is possible. Finally, we remark again on the large number of parameters for these models, $O(n)$ for the HRH model and $O(n^2)$

for the multiplicative factor model. While this may be feasible when fitting smaller networks, attempts at modelling larger networks may be problematic. The models we begin to develop in the next chapter attempt to address some of these issues.

Chapter 3

The Latent Socio-Spatial Process Model

The latent factor models discussed in the previous chapter have many attractive features. The Bernoulli likelihood is easy to work with, the concept of latent characteristics has an intuitive interpretation, and the random effects formulation provides a convenient mechanism for detecting and describing certain network structures. With their large number of node-specific or pair-specific parameters, however, they have the potential to over fit some data. While this makes them useful for predicting missing links within a sample network, they are not particularly geared toward predicting local network topologies or marginal probabilities of connection for actors outside the observed sample. In this chapter, we propose a new way of thinking about modelling network data. Rather than focusing on estimating the lack of fit in a logistic-linear regression, we instead turn attention to more flexible modelling of the relationship between $\mathbf{Y}$ and $\mathbf{X}$. This is motivated by a change in perspective from modelling the connections between "these particular actors" to modelling "actors like these."

When a goal of analysis is prediction, it is helpful to fit a model that links $\mathbf{Y}$ to $\mathbf{X}$, but does not simply reproduce $\mathbf{Y}$. Therefore, we look for a way to "smooth" the relationship between the probability of connections and attributes. How to incorporate covariates for a pair is a difficult question, but assuming homophily by attributes seems sensible for many kinds of network data. A literal interpretation, though, i.e. using marginal differences such as $\mathbf{x}_{i,j} = (|x_{i1} - x_{j1}|, \ldots, |x_{ip} - x_{jp}|)$, for example, may not be sufficient in all cases. It might be more reasonable to expect that the relationship between $\mathbf{Y}$ and $\mathbf{X}$ is quite complex, and may in fact change depending on the region of the covariate space. Let $\mathcal{X} \subseteq \mathbb{R}^p$ denote the p-dimensional covariate space. In contrast to the HRH model, which posits each actor has an unobserved position in a latent space, we focus on extracting as much information as possible from the *observed* positions $\mathbf{x}_i \in \mathcal{X}$.

We begin by assuming there exists a function $z : \mathcal{X} \mapsto \mathbb{R}$ that contains information on how a relative difference between covariates $\mathbf{x}_i$ and $\mathbf{x}_j$ affects $\eta_{i,j}$, the log-odds actors i and j are connected. That is, we propose to use

$$\eta_{i,j} = \mu - |z(\mathbf{x}_i) - z(\mathbf{x}_j)| \tag{3.1}$$

to link the expected value of $y_{i,j}$ to $\mathbf{x}_i$ and $\mathbf{x}_j$, the idea being that the relative position in social space depends on the covariates $\mathbf{x}$. To model the latent function (or surface, we will use the two words interchangeably), $z(\cdot)$, ideas from spatial process modelling will be used. Specifically, we model $z(\mathbf{x}_1), \ldots, z(\mathbf{x}_n)$ as a finite number of "observations" from a random process $\{z(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$. By estimating this random field, we are able to predict $z(\mathbf{x}_0)$ at any location $\mathbf{x}_0 \in \mathcal{X}$ using the height of the surface at that point. That is, we can make predictions for any actor given their attribute vector. In this present context, the $z(\mathbf{x}_i)$ are actually unobserved, and the "locations" $\mathbf{x}_i$ represent attributes measured to learn about social relations, so we

coin this the latent socio-spatial process (LSSP) model. The height of the surface at point $\mathbf{x}_i$, $z_i \equiv z(\mathbf{x}_i) \in \mathbb{R}$, is called the LSSP score for individual i. In this terminology, equation (3.1) specifies that actors with similar LSSP scores are more likely to be connected. The parameter μ is the average log-odds of connection for any two actors with the same LSSP score. Given the latent surface z, which captures the correlations between LSSP scores, the $y_{i,j}$ are modeled as conditionally independent binary random variables, so we maintain the same likelihood (2.3).

To help explain the intuition behind this model, we contrived the following example. Suppose one is interested in modelling the relationship between one covariate, age, and friendship. Consistent with the notion of homophily by attributes, it might be expected that two actors with similar ages are more likely to be friends than two with quite different ages. One might think, however, that the impact of age difference is *relative*, not *absolute*. If we have four actors with ages

$$x_1 = 5, \quad x_2 = 10, \quad x_3 = 50, \text{ and } x_4 = 55,$$

intuition suggests that actors 3 and 4 are more likely to be friends than actors 1 and 2, even though $|x_2 - x_1| = |x_4 - x_3| = 5$. In other words, the same absolute age difference is expected to have a different impact on the likeliness of friendship depending on the magnitude of x, or the particular region of the covariate space.

When distance is relative, sometimes a transformation can address this problem. The function $z(x)$ illustrated in Figure 3.1 suggests one possibility for this scenario. In this plot, age is represented by the horizontal axis. The LSSP score for an actor with a given age is the height of the function at that age, i.e. $z(50)$ in this plot is the LSSP score for any actor who is age 50. Looking at the projection of $z(x)$ onto the vertical axis, we see that $|z(10) - z(5)|$ is much greater than $|z(55) - z(50)|$. Thus, by comparing LSSP scores instead of the actual x values, the impact of age difference

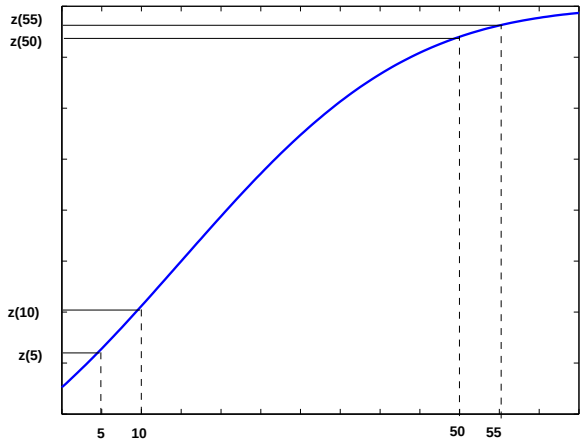


Figure 3.1: Absolute differences transformed to relative differences via the projection of a function.

more accurately matches intuition.

Now, consider a second covariate, annual gross income. Again, homophily by this attribute might be expected. Suppose we have the following information on four actors:

$$x_3 = (50, \$250\text{K}), \quad x_4 = (55, \$65\text{K}), \quad x_5 = (50, \$95\text{K}), \text{ and } x_6 = (55, \$100\text{K}).$$

Now, one might expect that actors 5 and 6 will be friends with higher probability than actors 3 and 4, even though the marginal age difference for each pair is the same. This suggests – at least intuitively – that attributes can interact in their effect on $\mathbf{Y}$.

As the number of covariates grows and the relationships between them become more complex, manually specifying transformations and interactions to find the most useful measure of relative distance is bound to become non-trivial. Thus, the function $z(\mathbf{x})$ in (3.1) is left unspecified, and during estimation the data suggests a simultaneous transformation of all covariates for which the corresponding projection onto $\mathbb{R}$ most accurately reflects homophily by attributes as it applies to the network under study.

Before moving on to discuss estimation of (3.1), we mention a few important

features of the model. First, with respect to the actual specification of the model, the use of absolute value, $|z(\mathbf{x}_i) - z(\mathbf{x}_j)|$, induces a certain amount of transitivity as in the HRH latent position model. The minus sign in front of this term implies that all covariates behave similarly in their impact on $\mathbf{Y}$, i.e. as currently expressed, increased distances in any covariate direction either have no effect or a decreasing effect on the likeliness of connection. Thus, (3.1) is only intended for network data where this is an appropriate assumption. We will revisit this point later. Also, no coefficient is included on this difference term because estimating its magnitude would be confounded with the estimation of z itself.

Given that actors with similar LSSP scores are more likely to be connected, the shape of the function $z(\mathbf{x})$ provides an interesting interpretation of clustering. In regions of $\mathcal{X}$ where the function is flat, differences in covariates have little impact on connection potential. Thus, these regions form groups of actors who are essentially equally likely to be connected with each other. Conversely, if the function is changing rapidly, this identifies regions in which small differences in $\mathbf{x}$ can have a big impact on the probability of connection. Thus, it can create a boundary that separates groups of actors. At a more global level, if the surface remains flat over the whole range of any particular covariate, then this covariate contains little information about $\mathbf{Y}$. This suggests that looking at function "activity" in each direction can be used as a means of identifying important variables. This is the topic of Chapter 5.

To conclude this section, we note that the proposed LSSP model in some ways closely resembles the HRH model. Although it might be arguable that (3.1) is only a special case of (2.5) – with the LSSP scores z_i just being latent positions in a one-dimensional social space – there are a few fundamental differences that we feel justify considering spatial process models a new class of models for network analysis. First,

whereas the HRH model assumes conditional independence of the $y_{i,j}$ given unobserved factors exogenous to $\mathbf{X}$, the LSSP model assumes conditional independence given an appropriate transformation of $\mathbf{X}$, and seeks to *average* the relationship between connections and attributes. As we will see, this change in perspective greatly reduces the required number of parameters and makes prediction feasible. Second, latent factor models emphasize fitting "noise" in order to fully explain a particular observed network. In contrast, the latent spatial process seeks to model common trends expected to hold at a population level, and the observed network is assumed to be a noisy realization of these underlying probabilities.

3.1 Estimation

To estimate model (3.1) within a Bayesian framework, priors must be specified for the parameter μ and the random field $\{z(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$. Choosing a prior for μ is a relatively standard procedure, so we begin our discussion by considering a prior for z, where more creativity is required. One of the primary advantages of the LSSP model is that $z(\mathbf{x})$ is left unspecified (to be determined by the data), so we wish to propose a very general prior class of functions.

The first approach we considered was to use a Gaussian Process (GP) prior for z. In a standard GP formulation (e.g. see Cressie, 1993), any finite number of observations $\mathbf{Z} = z(\mathbf{X}) = (z(\mathbf{x}_1), \ldots, z(\mathbf{x}_n))'$ are taken to have a multivariate normal distribution with a specified mean function – usually a constant or a polynomial – and a spatial covariance structure which determines

$$Cov[z(\mathbf{x}_i), z(\mathbf{x}_j)] \quad \forall \quad i, j = 1, \ldots, n.$$

This is a common tool for modelling spatial processes in many geostatistical and

environmental applications. And while its true that many of these processes take place in only $\mathbb{R}^2$ or $\mathbb{R}^3$, GP models have also been used in much higher dimensions. For example, Sacks *et al.* (1989a, 1989b) introduced their use as a form of nonparametric function estimation for the analysis of complex computer codes, which we will explore in Chapter 5.

Although GPs can represent a broad class of functions, there is a difficulty with using this modelling approach in the present context; namely, the function evaluations, i.e. the LSSP scores for the sample of n actors, are unobserved. One possible solution is to estimate $z(\mathbf{x}_i)$ for each individual $i = 1, \ldots, n$ as part of the analysis, but this is an extremely computationally intensive option and persists the issue of having $O(n)$ parameters.

In other situations where a traditional GP approach is prohibitive or too restrictive, an alternative process convolution representation (Barry and Ver Hoef, 1996; Higdon, 1998) has proved very efficient. Some applications can be found in Kern (2000), Higdon (2002) and Lee *et al.* (2005), for example. Particularly relevant, Higdon (2006) uses a process convolution for modelling a binary spatial process (though not pairwise, as in our situation) and Lee *et al.* (2007) use this approach for estimating an unobserved initial distribution in an inverse problem.

In general, instead of specifying a GP by its mean and covariance function, a GP can be constructed over the space $\mathcal{X}$ by convolving a Gaussian white noise process $\alpha(\mathbf{x})$ with a smoothing kernel $k(\mathbf{x})$,

$$z(\mathbf{x}) = \int_{\mathcal{X}} \alpha(\mathbf{x}) k(\mathbf{w} - \mathbf{x}) d\mathbf{w}, \quad \text{for } \mathbf{x} \in \mathcal{X}. \tag{3.2}$$

Of practical importance, Higdon (2002) suggests that if the white noise process is discretized, then the continuous spatial process $z(\mathbf{x})$ can be controlled by relatively few parameters. Specifically, if the support of α is restricted to a coarse grid $\mathcal{W} =$

$\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$, such that $\mathbf{w}_r \in \mathcal{X}$ for all $r = 1, \ldots, m$, then the discrete process convolution

$$z(\mathbf{x}) = \sum_{r=1}^{m} \alpha_r k(\mathbf{x} - \mathbf{w}_r) \tag{3.3}$$

is a good approximation to (3.2), where $\alpha_r = \alpha(\mathbf{w}_r)$ is the value of the noise process at site $\mathbf{w}_r$. In other words, the height of the surface at a point $\mathbf{x}_0 \in \mathcal{X}$, say, is the sum of an independent white noise process with support $\mathcal{W}$ weighted by a smoothing kernel centered at $\mathbf{x}_0$.

As noted in Kern (2000), for example, the choice of kernel can have a large impact on the resulting spatial process. In our case, since the process is latent, we make the innocuous assumption that $z(\mathbf{x})$ is a smooth surface over $\mathcal{X}$, and thus choose an independent p-dimensional multivariate Gaussian kernel for k. We let the width of the kernel vary in each covariate direction by introducing parameters $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)$, and use the unconventional parameterization

$$k_\rho(\mathbf{x}_i - \mathbf{w}_r) = \prod_{d=1}^{p} \rho_d^{(w_{rd} - x_{id})^2}, \tag{3.4}$$

where w_{rd} and x_{id} are the d^{th} element of $\mathbf{w}_r$ and $\mathbf{x}_i$ respectively. The ρ parameters measure the "correlation" between function evaluations in each direction. Consider the more standard notation

$$\rho_d = e^{-\frac{1}{2\sigma_d^2}},$$

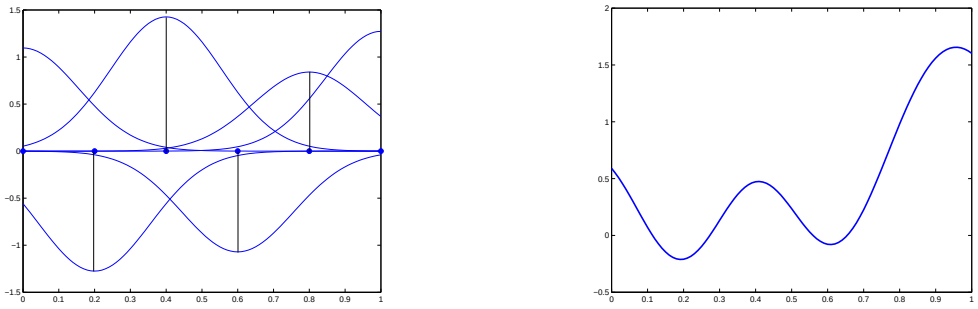
where σ_d is the standard deviation of the kernel in the d^{th} direction. As kernel widths increase in a particular direction, function evaluations become more correlated (see, e.g. Kern, 2000). Similarly, as $\rho_d \to 1$, the function becomes flatter in the d^{th} direction. As we will see in Chapter 5, this is related to the earlier comment that looking at the shape of the function can be used to identify important covariates.

Using (3.4) is also convenient because $0 \leq \rho_d \leq 1$, which simplifies prior specification and MCMC exploration.

This process convolution representation of a GP can be particularly useful in applications where a more flexible covariance function in the limiting GP is desired (as in Barry and ver Hoef, 1996 and Kern, 2000). For this purpose, the grid points $\mathcal{W}$ must be chosen sufficiently dense. For example, Higdon (2002) suggests a lattice with points no more than the kernel standard deviation apart. Now while using a lattice for $\mathcal{W}$ is feasible in $\mathbb{R}^2$ or $\mathbb{R}^3$, it will become extremely expensive in higher dimensions. For example, a lattice with, say, 10 points per dimension requires the estimation of upwards of $m = 10^p$ parameters. On the other hand, if the goal is simply function approximation (and not covariance estimation) as in our situation, then it might be possible to relax this requirement.

Given our particular choice of kernel, we see that (3.3) can also be viewed as a radial basis function approximation to $z(\mathbf{x})$. Hastie *et al.* (2001) is a useful reference on this approach to fitting data. From this perspective, (3.3) can be interpreted as a weighted sum of basis kernels, where $k_\rho(\mathbf{x} - \mathbf{w}_r)$ are the basis functions with centers $\mathbf{w}_r$, and the weights are α_r, $r = 1, \ldots, m$. This parallel is also recognized in Higdon (2006). Figure 3.2(a) illustrates how a smooth process is generated over $\mathcal{X} = [0, 1]$ using this approach. The circles on the horizontal axis mark the locations of 6 grid points, $w_1, \ldots, w_6$, and the heights of the kernels correspond to random weights $\alpha_1, \ldots, \alpha_6$. Figure 3.2(b) is the resulting sum of the weighted kernels.

For the purpose of function approximation, if the function is observed at n points $\mathbf{X} \in \mathcal{X}$, often these locations themselves are used as the centers for the kernels (again, we refer to Hastie *et al.*, 2001). As a justification for reducing the number of grid points, m, in our function representation, we think in terms of these radial basis



(a) Weighted basis functions (b) Resulting smooth function

Figure 3.2: Radial basis function approximation

functions rather than limiting GPs. In particular, we make the following argument. Suppose for the purpose of approximating a surface over $\mathcal{X}$, one has the opportunity to choose a set of locations in $\mathcal{X}$ at which the function will be observed. Given this chance, what is a good choice of locations?

When designing a computer experiment, for example, where input settings at which the computer code is to be run are selected, often a Latin hypercube design (LHD, McKay *et al.*, 1979) is a reasonable choice of input settings. Therefore, given that no evaluations of the function are ever observed in our case anyway, it seems reasonable to choose $\mathcal{W}$ for use in (3.3) as a LHD over $\mathcal{X}$, in the spirit that if we could observe the function, we would choose to do so at these points. In practice, we typically use a LHD with 10 points per dimension. This "$10p$" rule of thumb is fairly standard in choosing designs for high-dimensional computer experiments (Jones *et al.*, 1998). We also use a "space-filling" optimization to ensure the design points are evenly spread out over $\mathcal{X}$. This choice of $\mathcal{W}$ greatly reduces the required number of parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ as compared to a dense grid, and it appears to fill $\mathcal{X}$ enough for our needs. See Chapter 7 for more discussion on approximating radial basis functions in this way.

To summarize, the prior class of functions we choose for z is

$$z(\mathbf{x}) = \sum_{r=1}^{m} \alpha_r k_\rho(\mathbf{x} - \mathbf{w}_r),$$

with the kernel given by (3.4). The locations of the kernel centers, $\mathcal{W}$, we choose according to a Latin hypercube design on $\mathcal{X}$ with about 10 points per dimension. We scale $\mathcal{X}$ to $[0,1]^p$ to facilitate the selection of this design, which does imply there is a prior range of interest on each of the covariates. An advantage of this prior class of functions is that it is completely governed by only $m + p$ parameters, $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$. It remains to specify priors for these unknowns.

Because the function z only appears in (3.1) as a difference, we assume without loss of generality that it is a mean zero process – any mean parameter would be non-identifiable. In particular, the prior we use for $\boldsymbol{\alpha}$ is

$$\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{I}_m),$$

where $\mathbf{I}_m$ is the $m \times m$ identity matrix. Note that the resulting implied covariance structure for the latent surface is

$$cov[z(\mathbf{x}_i), z(\mathbf{x}_j)] = \sum_{r=1}^{m} k_\rho(\mathbf{x}_i - \mathbf{w}_r) k_\rho(\mathbf{x}_j - \mathbf{w}_r),$$

i.e. the covariance between two function evaluations depends on how far each attribute is from the fixed center points. For $\boldsymbol{\rho}$, we use non-informative priors

$$\rho_d \sim U[0,1], \quad d = 1, \ldots, p.$$

Though other choices can be used, these yield a broad enough class of prior functions for any applications we have considered to date. The remaining parameter in (3.1) is μ, to which we assign a

$$\mu \sim N(0, \psi_\mu)$$

prior. Taking ψ_μ to be fairly large makes this prior vague. Recall that this parameter represents the average log-odds of connection between two actors with the same LSSP score. This choice of prior centers this probability at 50%.

Given the Bernoulli likelihood (2.3) and the above priors for the $m+p+1$ parameters (where m is typically $10p$), the full posterior distribution for the LSSP model is simply

$$[\mu, \boldsymbol{\alpha}, \boldsymbol{\rho}|\mathbf{y}] \propto [\mathbf{y}|\mu, \boldsymbol{\alpha}, \boldsymbol{\rho}][\mu][\boldsymbol{\alpha}][\boldsymbol{\rho}].$$

Since the full conditionals for all parameters are nonstandard, an MCMC algorithm with a Metropolis-Hastings algorithm is used to generate realizations $(\mu^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\rho}^{(t)})$, $t = 1, \ldots, T$ from the posterior distribution. Using these draws, we are able to easily generate some posterior quantities that are of particular interest.

First, the height of the surface, i.e. the LSSP score, at any point $\mathbf{x}_0 \in \mathcal{X}$ can be predicted using the posterior expectation

$$\hat{z}(\mathbf{x}_0) = \mathbb{E}[z(\mathbf{x}_0)|\mathbf{y}] = \frac{1}{T}\sum_{t=1}^{T} z^{(t)}(\mathbf{x}_0) = \frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{m} \alpha_r^{(t)} k_{\rho^{(t)}}(\mathbf{x}_0 - \mathbf{w}_r).$$

Also, a prediction of the marginal probability that any two actors i_0 and j_0 are connected, given their attributes $\mathbf{x}_{i_0}$ and $\mathbf{x}_{j_0}$, is simply

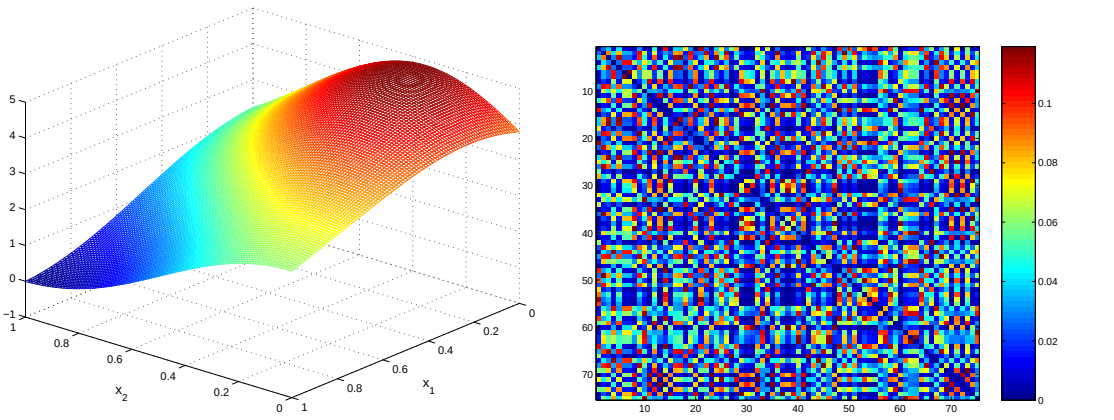
$$\hat{\pi}_{i_0,j_0} = \mathbb{E}\left[\frac{e^{\eta_{i_0,j_0}}}{1+e^{\eta_{i_0,j_0}}}|\mathbf{y}\right] = \frac{1}{T}\sum_{t=1}^{T} \frac{e^{\eta_{i_0,j_0}^{(t)}}}{1+e^{\eta_{i_0,j_0}^{(t)}}},$$

where

$$\eta_{i_0,j_0}^{(t)} = \mu^{(t)} - |z^{(t)}(\mathbf{x}_{i_0}) - z^{(t)}(\mathbf{x}_{j_0})|.$$

We refer back to Chapter 1 for these results.

Example: Synthetic Network (SN) of Actors with Two Covariates. To illustrate the proposed LSSP methodology, we demonstrate estimation and prediction



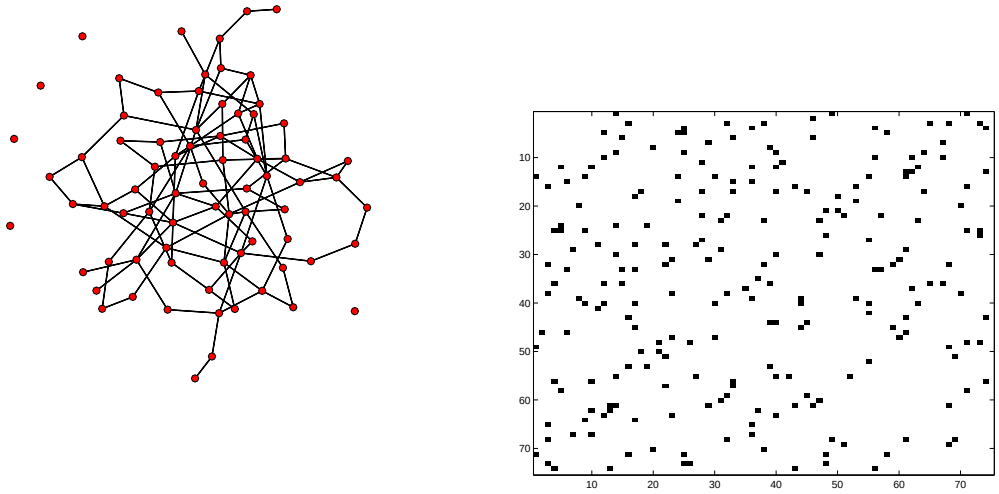
(a) An LSSP in $\mathbf{R}^2$ (b) Probability of connection for $n = 75$ individuals

Figure 3.3: SN Example: LSSP and connection probabilities.

using a network that is constructed under the pretense that the LSSP assumption is true. That is, suppose the surface over $\mathcal{X} = [0, 1]^2$ shown in Figure 3.3(a) is the LSSP that relates attributes to the log-odds of connection for a network. We generate a sample of $n = 75$ actors by uniformly choosing attribute vectors $\mathbf{x}_i \in \mathcal{X}$, $i = 1, \ldots, n$. LSSP scores are assigned to each actor according to the height of the surface in Figure 3.3(a) at locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Using (3.1) with $\mu = -2$, we calculate the probability that each pair of actors in this sample is connected. These probabilities are plotted in Figure 3.3(b). Again, in this type of plot, "hotter" colors represent larger probabilities, and "cooler" colors smaller. According to these probabilities, we independently generate binary random variables, which we then treat as the observed network for this sample – see Figure 3.4(a). The corresponding image representation is plotted alongside in Figure 3.4(b). Treating this network as observed, we implement the above estimation procedure to recover the underlying LSSP.

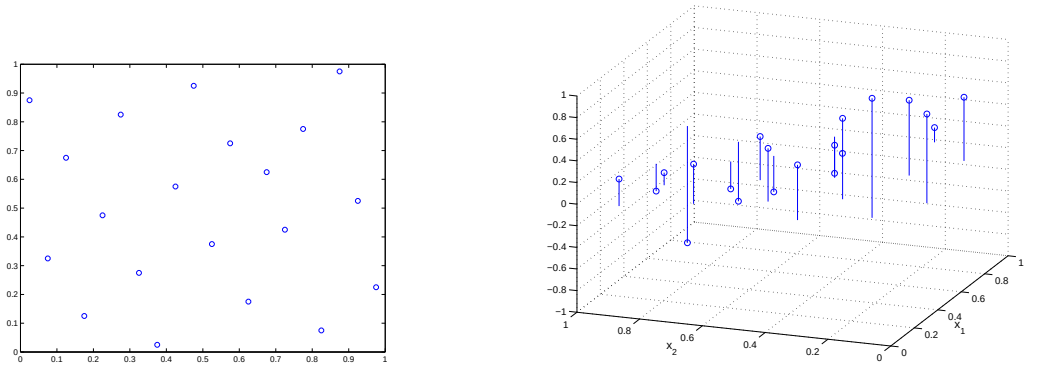
To begin, we choose $m = 20$ $\mathbf{w}_1, \ldots, \mathbf{w}_m \in \mathcal{X}$ as the centers for our basis kernels



(a) Graph representation (b) Image representation

Figure 3.4: SN Example: Alternative visual representations for a network with 75 people.

(3.4) points according to the "10p" rule. These locations are shown in Figure 3.5(a). In this example, there are a total of 23 parameters to be estimated: μ, $\boldsymbol{\rho} = (\rho_1, \rho_2)$, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{20})$. We use the priors given above, with the specific choice $\psi_\mu = 10$. An MCMC algorithm with a Metropolis-Hastings step is used to generate a large number of draws from the posterior distribution $[\mu, \boldsymbol{\alpha}, \boldsymbol{\rho}|\mathbf{y}]$. Figure 3.5(b) plots the posterior means of the parameters $\alpha_1, \ldots, \alpha_{20}$ at their respective locations $\mathbf{w}_1, \ldots, \mathbf{w}_{20} \in [0, 1]^2$. The stems represent the magnitude and direction of the parameters from zero. In particular, stems going up represent positive parameter estimates, and stems going down negative parameter estimates. The posterior mean of the latent surface over the covariate space is given in Figure 3.6(a). This is created by predicting the posterior mean of z over a fine grid in $[0, 1]^2$. The circles marked on this surface are the estimated LSSP scores for the original sample of $n = 75$ actors. By comparing Figures 3.3(a)



(a) Chosen basis kernel centers

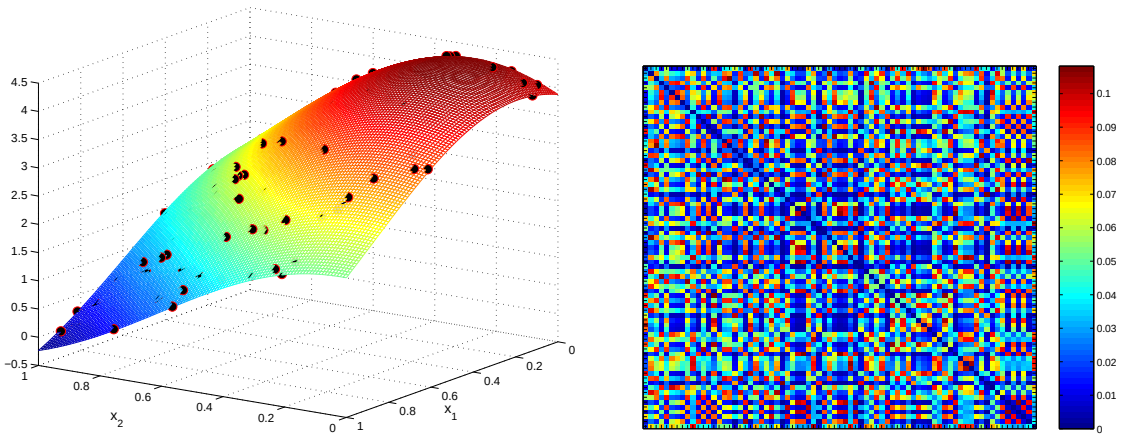
(b) Posterior mean estimates of basis weights

Figure 3.5: SN Example: Locations and posterior estimates of basis weights.

and 3.6(a), we see that the estimation procedure seems to have done a reasonable job of predicting the underlying function, i.e. the estimated surface and true surface appear very similar. The posterior mean probabilities of connection for this sample, evaluated for all pairs $i, j = 1, \ldots, n$, are plotted in Figure 3.6(b).

To illustrate how the LSSP model can reveal clusters of connections, we sort the actors with respect to their estimated LSSP scores, from lowest to highest. Figure 3.7(a) is a plot of the posterior probabilities of connection as in Figure 3.6(b), but with the rows and columns permuted in order of LSSP scores. In this case, since the true probabilities of connection are actually known, as shown in Figure 3.3(b), we can apply the same permutation to these as an *ad hoc* assessment of fit. This is done in Figure 3.7(b). Finally, we can permute the rows and columns of the sociomatrix, and see that there is evidence of this clustering pattern in the observed network. This is illustrated in Figure 3.8.

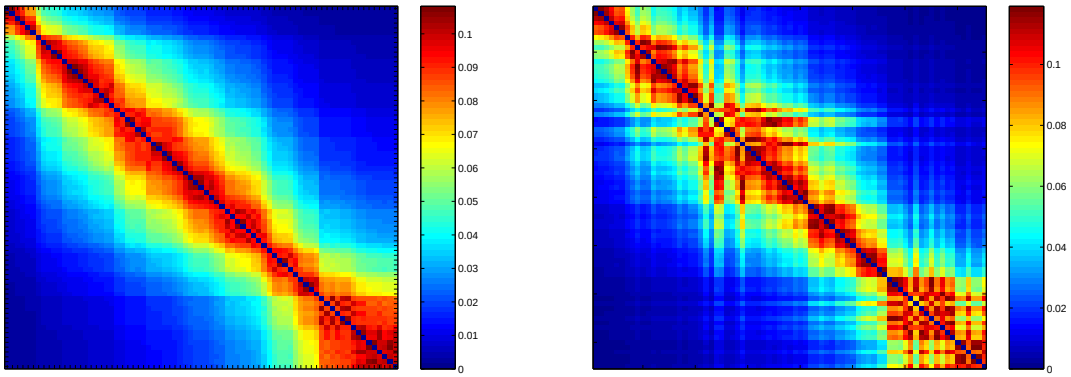
The above-mentioned plots give evidence of adequate fit. To more objectively assess fit, however, we generate a number of replicate sociomatrices (using the sample covariates) and compare a variety of network statistics from the observed sociomatrix



(a) LSSP posterior mean

(b) Posterior mean probabilities

Figure 3.6: SN Example: Posterior mean estimates.



(a) Sorted posterior mean probabilities

(b) Sorted true probabilities

Figure 3.7: SN Example: Probabilities of connection sorted by LSSP scores.

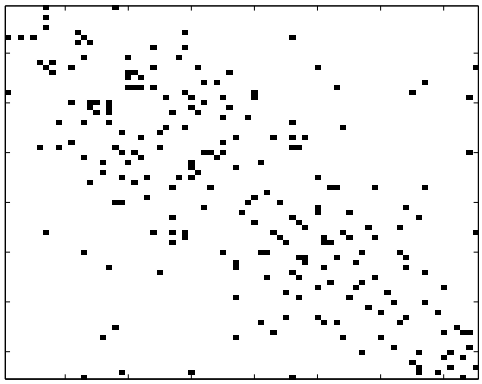
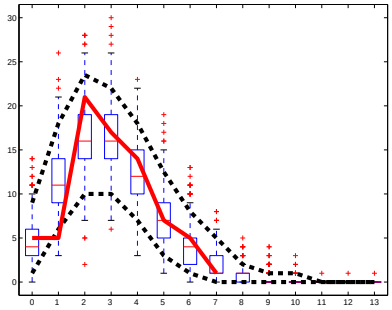


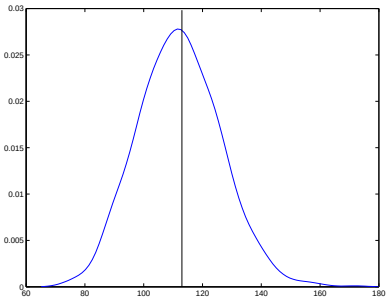
Figure 3.8: SN Example: Sociomatrix sorted by LSSP scores.

to the distribution of the same statistics calculated using the replicate sociomatrices. In Figure 3.9(a), the solid line is the degree sequence for the observed network, and the dashed lines are pointwise 90% credible intervals of degree sequences corresponding to replicate sociomatrices. The boxplots show the whole replicate distributions of each element in the degree sequence. We can also look at the number of edges in the graph as in Figure 3.9(b). This plot shows the observed number of edges as a vertical line and a smooth histogram of the replicate distribution of edges. The minimum geodesic sequence is shown in Figure 3.9(c), again with the observed sequence in a solid line and pointwise 90% intervals and boxplots from the replicate distribution. The last boxplot on the far right is for the number of pairs that have an infinite geodesic distance. To assess fit more locally, we plot the observed degrees for each actor in the network in Figure 3.9(d). Here, the observed degrees are plotted as circles, and 90% pointwise credible intervals for each actor are indicated by the dashed lines. The null distribution of a plug-in χ^2 goodness of fit test is drawn in Figure 3.10. The observed statistic (denoted by a triangle) is in the center of this distribution, suggesting a good fit.

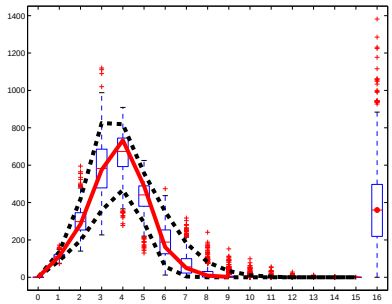
While all of these plots can be used to argue that the estimated model "fits" the



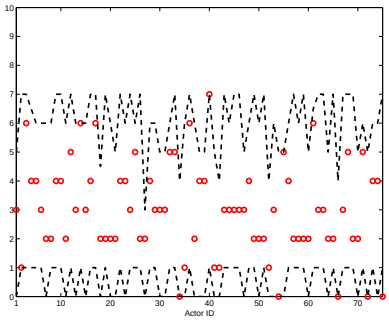
(a) Degree sequence



(b) Number of edges



(c) Minimum geodesic sequence



(d) Actor degrees

Figure 3.9: SN Example: Goodness-of-fit statistics

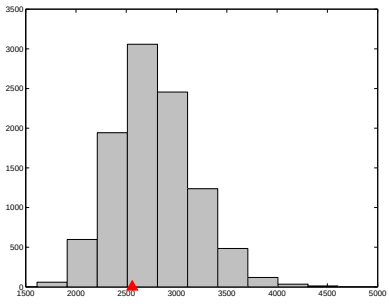


Figure 3.10: SN Example: χ^2 Test

data fairly well, they do not give any indication of its predictive ability. For this, we consider a Bayesian cross-validation (CV) procedure.

3.2 Bayesian Cross-Validation

Cross-validation is a useful procedure for testing the predictive ability of a model. Broadly speaking, the idea is to fit a model using only a portion of the available data, then predict responses for the withheld sample and make comparisons to the truth. Ideally, to minimize the impact of the choice of split, this will be repeated a number of times. A difficulty in the Bayesian context is that each fit of the model requires MCMC sampling of the posterior distribution. To make implementation feasible, we follow a Bayesian cross-validation technique proposed by Alqallaf and Gustafson (2001).

We begin by randomly splitting the n actors into two groups, a *training set* and a *validation set*. We fix the number of actors in the training set to be n_T, and thus the number in the validation set is taken to be $n_V = n - n_T$. Let $\mathcal{T}_s$ and $\mathcal{V}_s$ denote the set of indices of actors in the training set and validation set, respectively, given a particular split, s, of the actors. The training data to be used is the $\binom{n_T}{2} \times 1$ vectors with elements $\{y_{i,j} : i < j, (i,j) \in \mathcal{T}_s\}$, which we denote by $\mathbf{y}_{T(s)}$. The validation responses we wish to predict are $\mathbf{y}_{V(s)}$, which we define as all unique pairs of connections between each of the actors in the validation set and all other $n-1$ actors in the sample. Figure 3.2 is a conceptual drawing of these entities. For convenience, the actors are ordered by training set and then validation set in this drawing.

After dividing the actors, we sample $(\mu_s^{(t)}, \boldsymbol{\alpha}_s^{(t)}, \boldsymbol{\rho}_s^{(t)})$ from the posterior distribution

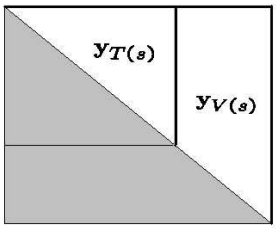


Figure 3.11: Training and validation data

given the training data,

$$[\mu_s, \boldsymbol{\alpha}_s, \boldsymbol{\rho}_s | \mathbf{y}_{T(s)}] \propto [\mathbf{y}_{T(s)} | \mu_s, \boldsymbol{\alpha}_s, \boldsymbol{\rho}_s][\mu_s][\boldsymbol{\alpha}_s][\boldsymbol{\rho}_s].$$

We use the subscript s here to emphasize that the training data corresponds to a particular split, s, of the actors. Then, a prediction $\hat{\mathbf{y}}^{(t)}_{V(s)}$ is drawn from

$$[\mathbf{y}_{V(s)} | \mu_s^{(t)}, \boldsymbol{\alpha}_s^{(t)}, \boldsymbol{\rho}_s^{(t)}, \mathbf{y}_{T(s)}].$$

Note that given the conditional independence assumption for the network connections, each element of $\hat{\mathbf{y}}^{(t)}_{V(s)}$ is predicted by the generation of an independent Bernoulli random variable with the log-odds specified by (3.1). This process is first repeated for many draws $t = 1, \ldots, T$ from the posterior distribution given the training data for a particular split. Finally, it is then repeated for many splits, $s = 1, \ldots, S$.

For each prediction $\hat{\mathbf{y}}^{(t)}_{V(s)}$, we wish to compare it to the observed responses $\mathbf{y}_{V(s)}$ using a specified "error" function

$$\epsilon(\hat{\mathbf{y}}^{(t)}_{V(s)}, \mathbf{y}_{V(s)}).$$

For example, one obvious choice might be

$$\epsilon(\hat{\mathbf{y}}^{(t)}_{V(s)}, \mathbf{y}_{V(s)}) = \|\hat{\mathbf{y}}^{(t)}_{V(s)} - \mathbf{y}_{V(s)}\|^2,$$

which for binary response is the total number of wrongly predicted responses. However, given the sparsity of most sociomatrices, i.e. the large number of zeros, we find

that this sum of squared prediction error gets "swamped" by correctly predicted zeros. That is, there are only two options for predicted responses, and any model that mostly predicts zeros will actually be right most of the time.

Since local topologies are one feature of networks that particularly interest us, we consider error functions that compare the degree of actors in the validation set to their true degrees, and similarly for extended degree. Note that our definition of $\mathbf{y}_{V(s)}$ includes connections between validation actors and all other $n-1$ actors, so we are able to calculate these predicted topological quantities. In particular, we consider two functions

$$\epsilon_1(\hat{\mathbf{y}}^{(t)}_{V(s)}, \mathbf{y}_{V(s)}) = \frac{1}{n_V} \sum_{i \in \mathcal{V}_s} \mathbb{I}[d_i^{(t)} \leq d_i], \tag{3.5}$$

and

$$\epsilon_2(\hat{\mathbf{y}}^{(t)}_{V(s)}, \mathbf{y}_{V(s)}) = \frac{1}{n_V} \sum_{i \in \mathcal{V}_s} \mathbb{I}[e_i^{(t)} \leq e_i], \tag{3.6}$$

where $\mathbb{I}$ is the indicator function. Here, $d_i^{(t)}$, and respectively $e_i^{(t)}$, are calculated using the predicted response $\hat{\mathbf{y}}^{(t)}_{V(s)}$, whereas d_i and e_i are the true degree and extended degree of each validation actor.

To average over the predictive distribution of the responses given the training data and the different splits, we calculate

$$\hat{\epsilon} = \mathbb{E}[\epsilon] \approx \frac{1}{S} \sum_{s=1}^{S} \frac{1}{T} \sum_{t=1}^{T} \epsilon(\hat{\mathbf{y}}^{(t)}_{V(s)}, \mathbf{y}_{V(s)})$$

for each of the error functions. From Alqallaf and Gustafson (2001), this is the so-called "silver estimator" for the cross-validation error, since we must sample the predictive distribution, i.e. we do not have an analytical solution. By using the discrepancy measure (3.5), we are finding on average what percentile of the predictive degree distribution is the true degree of a validation actor. A similar interpretation holds for (3.6).

As a final remark, one might recall that Hoff (2007a) uses a form of cross-validation to assess the ability of the multiplicative latent factor model to predict missing links. In that implementation, every actor in the sample must have at least one pairwise observation available in the training set. Here, on the other hand, validation actors are completely removed from the training data. This emphasizes the change in perspective toward making population-level predictive inference.

Example: The Synthetic Network (SN) Revisited. To implement the above cross-validation procedure for the synthetic network example, we divide the $n = 75$ actors into $n_T = 50$ training actors and $n_V = 25$ validation actors. For one particular split of the data, Figure 3.12 shows the location of the observed degree (the solid vertical line) in the predictive degree distribution for each of the n_V actors. Here, each plot corresponds to one actor in the validation set. If an actor has a degree of zero in the observed network, this is noted by a triangle. For this same split, we also consider the bivariate predictive distribution of degree and extended degree for each validation actor. This is shown in Figure 3.13 for the same 25 actors. In each of these plots, degree is on the horizontal axis and extended degree on the vertical. The true pair (d_i, e_i) for $i \in \mathcal{V}_s$ is indicated by a solid red dot. The other points are predicted pairs $(d_i^{(t)}, e_i^{(t)})$ for a number of draws of the predicted sociomatrix given the training data for this split. Averaged over $S = 50$ splits, we evaluate $\hat{\epsilon}_1 = 0.61$ and $\hat{\epsilon}_2 = 0.53$. As there are currently no other results against which to compare these, we can only comment that it seems positive that the observed degree (and extended degrees) for validation actors are on average not extreme values in the predictive distributions. We will revisit related issues in Chapter 6 when considering disease propagation.

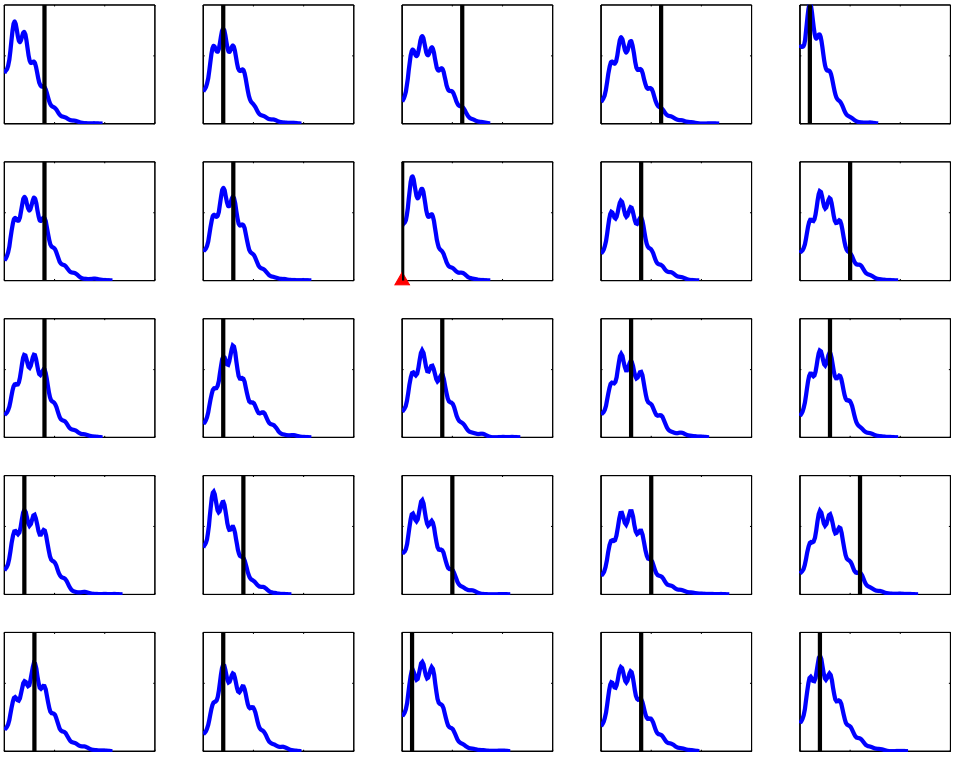


Figure 3.12: SN Example: Validation set degree distribution.

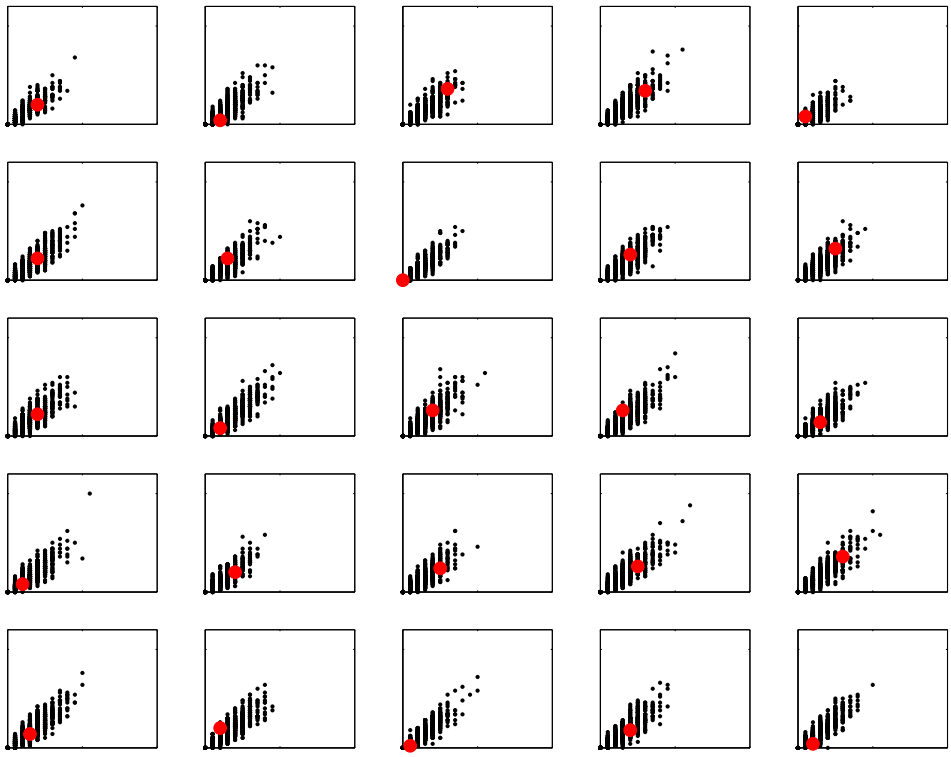


Figure 3.13: SN Example: Validation set degree by extended degree.

3.3 Examples

In this section, we use the LSSP model to analyze two networks that have been previously explored in the network literature (one being the International Conflict network which we have already seen). For each example, we estimate the parameters of the LSSP model, look at some goodness-of-fit plots, and run the cross-validation procedure.

Example: Adolescent Health (AH) Survey. The first network we consider is a friendship network compiled as part of the National Longitudinal Study of Adolescent Health (Resnick *et al.*, 1997). This is a very large survey of students in the USA conducted to monitor the long-term outcomes of health-related adolescent behaviours. One network from this study was considered by Handcock *et al.* (2007) in their exploration of the latent cluster model mentioned in Chapter 2. They also provide more details on the design of the study. The particular network we consider consists of 205 students from one school, which is available as part of the `R` package `statnet` (Handcock *et al.*, 2003). The graph for this network is shown in Figure 3.14. Here, $y_{i,j} = 1$ for a pair of students if either i or j reported the other as a friend on a questionnaire. We note that the network as given is not exactly one obtained from the study, but rather one that has been reconstructed for confidentiality reasons to be similar, as explained in the `statnet` manual. The covariate we consider is the grade of each student, which for this example are grades 7-12.

To estimate the LSSP model, the covariate *grade* is first scaled to $[0, 1]$, where 0 corresponds to Grade 7 and 1 corresponds to Grade 12. We use $m = 10$ kernel centers for this one-dimensional covariate, equally spaced between 0 and 1. The posterior mean of the function $z(x)$ is shown in Figure 3.15(a). The posterior mean probabilities

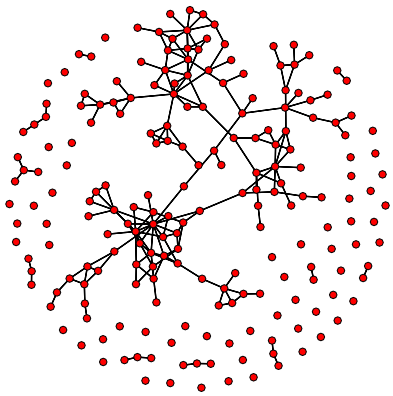
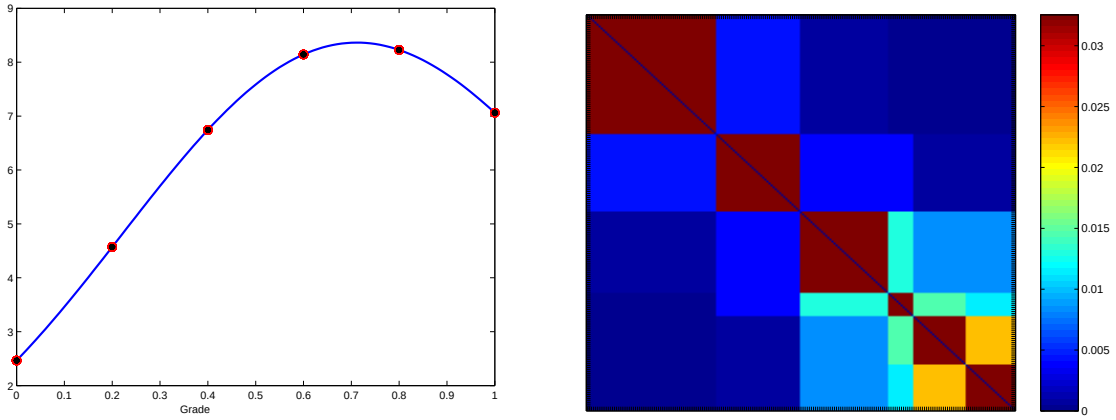


Figure 3.14: AH Example: Friendship network of 205 students

of connection for each pair of students, where the students are sorted by LSSP score (from lowest to highest), is shown in Figure 3.15(b). We note that by (3.1), students in the same grade will have the same LSSP score, and will be equally likely to be connected. This is clearly seen in the plot of the sorted posterior probabilities. This suggests that in future research it might be interesting to allow the parameter μ to change over the covariate space, but how to do this with covariates for each *pair* is still an open problem.

What is more interesting from Figure 3.15(a), however, is how LSSP scores change between grades. The estimated function changes rapidly between grades 7-9, so there is little probability of mixing between these groups. However, the rate of change slows between grades 10-12, implying more mixing in the higher grades. This is also seen in Figure 3.15(b), where the blocks of probability in the lower right corner (corresponding to the higher grades) show more blending than the disjoint blocks corresponding to grades 7, 8 and 9 (recall that the sorting in this plot is from lowest to highest LSSP



(a) LSSP posterior mean

(b) Posterior mean probabilities

Figure 3.15: AH Example: Posterior mean estimates.

scores). Figure 3.16 shows the observed sociomatrix with the same permutation of students with low to high LSSP scores, which shows these blocking patterns to appear in the observed connections. It might seem surprising that the posterior estimate of z *decreases* between Grades 11 and 12. This is due to the fact that there are very few Grade 12 students in this network, and it just so happens that of the few there are, many of them are connected to younger students, so the scoring for a Grade 12 student adjusts itself accordingly.

We remark that when Handcock *et al.* (2007) apply the latent position cluster model to a similar network (they use a connected network with 71 students, in part because the method cannot handle isolated individuals), they find clusters amongst the students that are very much associated with grade. In particular, they find little mixing between Grades 7, 8, and 9 students, and more mixing between the higher grades, similar to our interpretation. Fitting a latent position model, however, requires one parameter for each student plus some (i.e. approximately 230 parameters for the network we are exploring), whereas the LSSP is fit using only 12 parameters

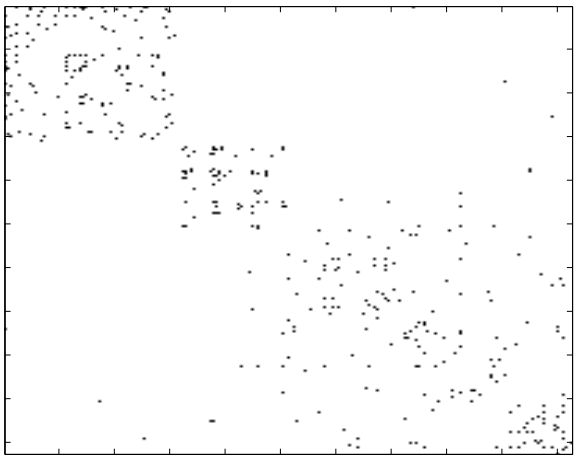
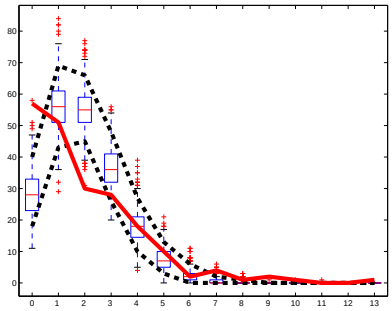


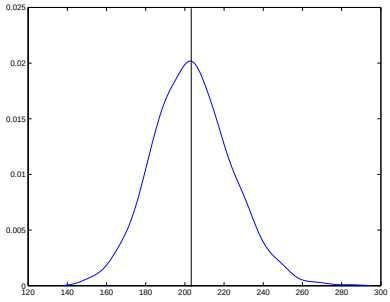
Figure 3.16: AH Example: Sorted sociomatrix.

$(\mu, \rho, \alpha_1, \ldots, \alpha_{10})$. We note, however, that this is not a totally fair comparison since it is possible the network we analyzed had emphasized clustering (due to its reconstruction). Also, we reiterate that the goals of the two analysis approaches are quite different.

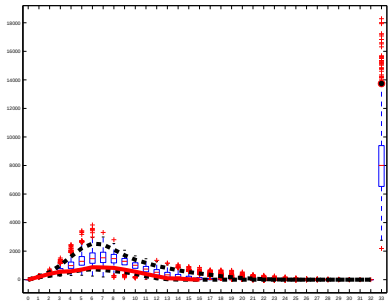
In any case, we move on to consider some goodness-of-fit plots. In particular, we consider the same five plots that we used for the SN Example. The first plot, Figure 3.17(a), shows the degree sequence for the observed friendship network in a bold, red line, and pointwise boxplots and 90% intervals corresponding to the replicate degree sequences. We can see here that the model expects many less isolates than were observed. Figure 3.17(b) plots a smoothed histogram of the edge count of a number replicated sociomatrices, as well as the observed count of edges as a vertical line. The minimum geodesic sequence is shown in Figure 3.17(c), with the far right boxplot representing pairs that have an infinite geodesic distance. Finally, the degrees of each of the 205 students are plotted in Figure 3.17(d) as red circles, with pointwise 90% intervals given by the dashed lines. The null distribution and test statistic for the χ^2 goodness-of-fit test is given in Figure 3.18. We see that this observed statistic



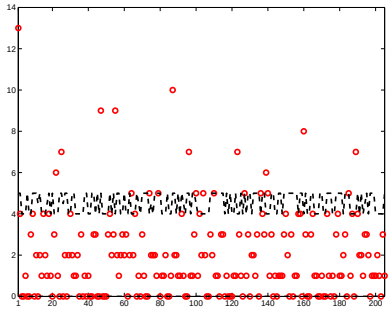
(a) Degree sequence



(b) Number of edges



(c) Minimum geodesic sequence



(d) Actor degrees

Figure 3.17: AH Example: Goodness-of-fit statistics

(identified by the triangle) is a bit extreme in the null distribution. This is partly due to the impact of the ordinal categorical covariate, grade. While such covariates can be included in the analysis, we find it is best if they have more categories than the 6 grades here. For example, though we exclude these results here, we found that substituting ages for the grades improved the fit but did not change the interpretation (we did this in an ad hoc way that assumed 2-3 different ages per grade). This suggests that the information in the grade covariate in this particular example may be just a bit too coarse. See the discussion in Chapter 7 for more on categorical covariates.

Using the cross-validation procedure described in the previous section, we divided the students into a training set with $n_T = 155$ students and a validation set with the

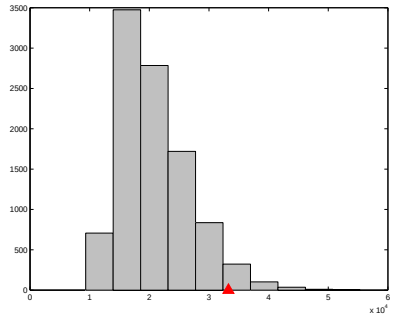


Figure 3.18: AH Example: χ^2 test.

remaining $n_V = 50$ students. For one particular split, Figure 3.19 shows the predictive degree distributions for each of the actors in the validation set given the training data, with the true degree of each actor represented by a vertical line. Figure 3.20 is the corresponding bivariate plot of predicted degree by extended degree. Again, the true pair for each actor is marked by a red circle. These plots suggest there are still some challenges to be resolved with handling outliers, i.e. students that have no connections or many connections. Averaging over $S = 50$ splits, we obtain $\hat{\epsilon}_1 = 0.54$ and $\hat{\epsilon}_2 = 0.56$ for the prediction of degrees and extended degrees, respectively.

Example: International Conflict (IC) Network. The second example we consider in this section is the network of international conflicts between 130 nations discussed in Chapter 2. Recall for this network (pictured again here in Figure 3.21) $y_{i,j} = 1$ if there was a conflict between countries i and j during the period 1990-2000. The covariates we consider are log population (x_1) and polity score (x_2).

After scaling the covariate space to $[0, 1]^2$, we use 20 basis kernel center $\mathbf{w}_1, \ldots, \mathbf{w}_{20}$ to anchor the LSSP. There are a total of 23 parameters to be estimated. The posterior mean of $z(\mathbf{x})$ is shown in Figure 3.22. The circles marked on the surface identify the LSSP scores for the 130 countries. This LSSP provides an interesting interpretation

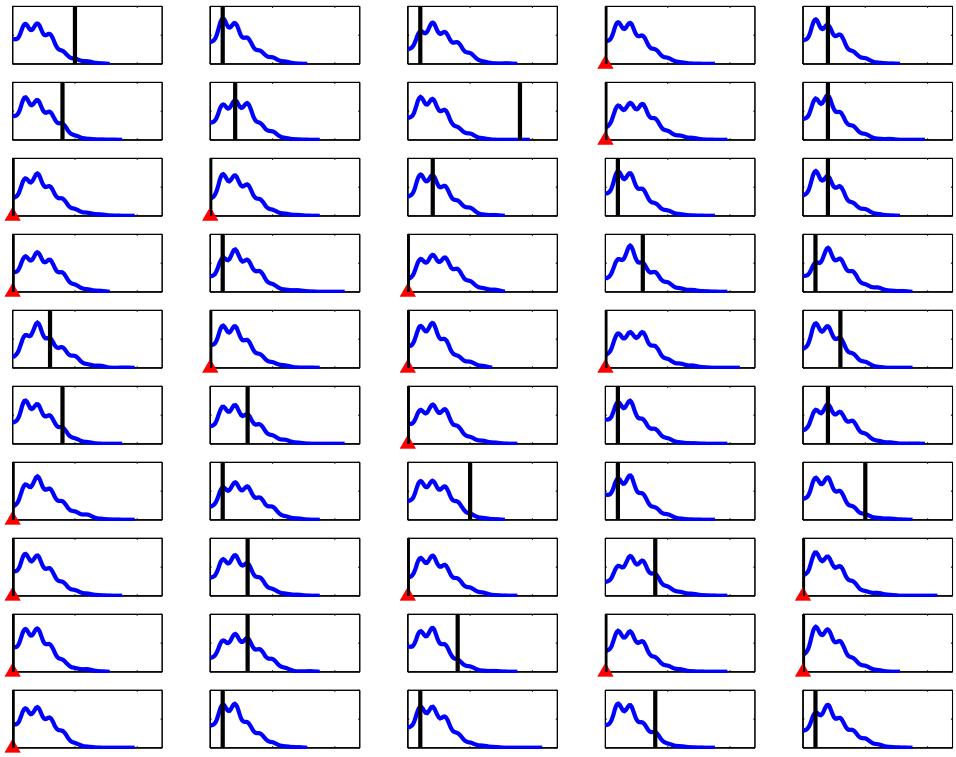


Figure 3.19: AH Example: Validation set degree distribution.

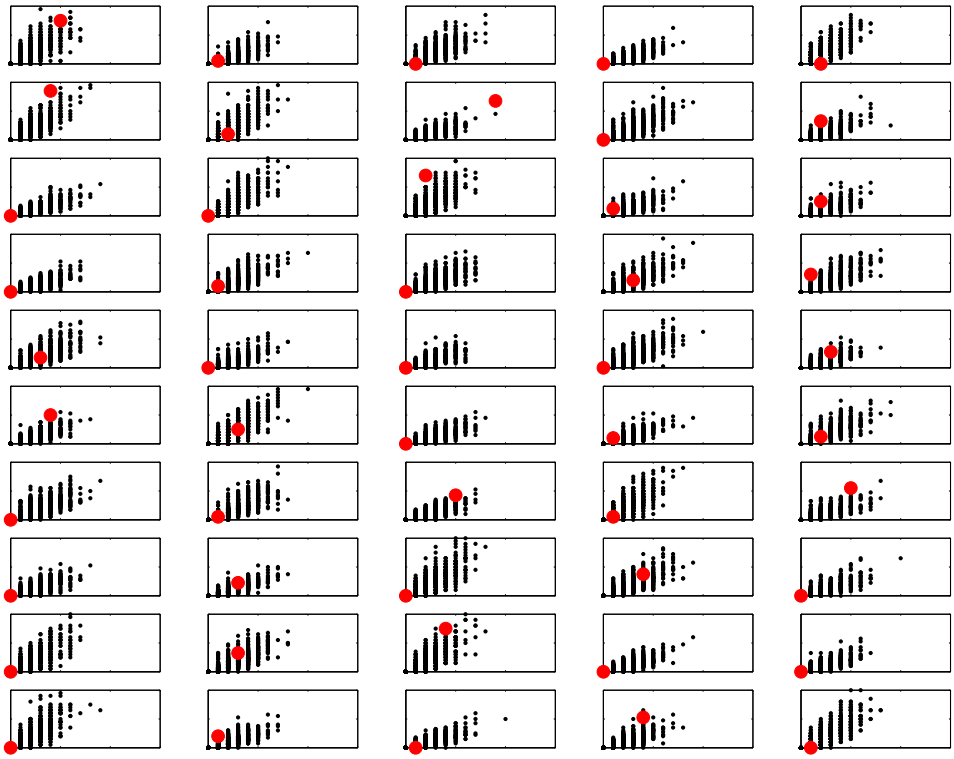


Figure 3.20: AH Example: Validation set degree by extended degree.

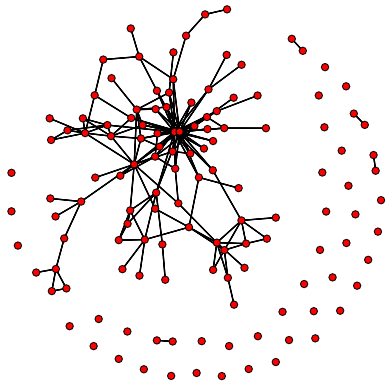


Figure 3.21: IC Example: Network of conflicts between 130 nations.

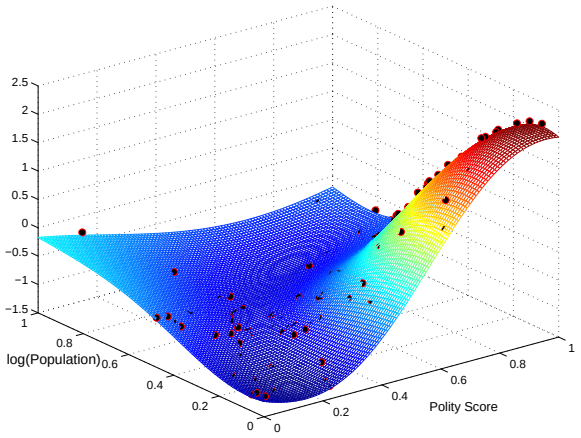
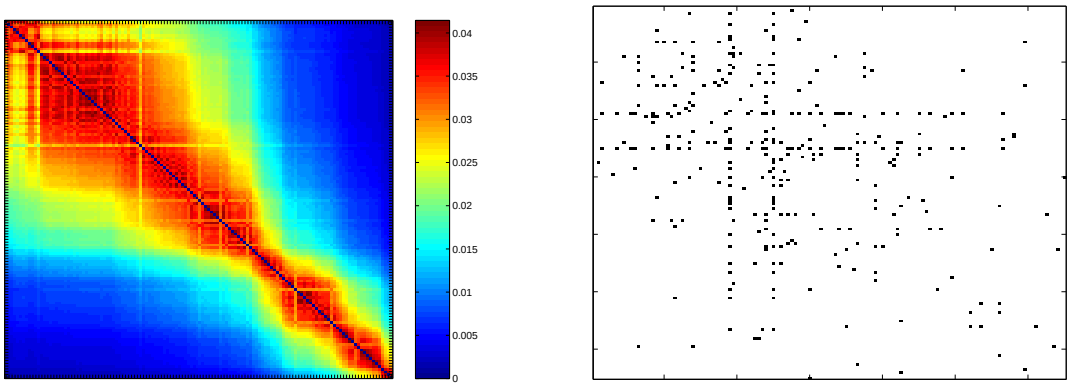


Figure 3.22: IC Example: Posterior mean LSSP

about the effect of these two covariates on the probability of a conflict. Notably, for democratic countries (high polity score), the surface changes rapidly as a function of population. Therefore, there appears to be a low probability of aggressive behaviour between large and small democratic countries (since they are at such different heights). This effect is less pronounced for more authoritarian countries, highlighting the interaction between these two attributes. Rather than having conflicts with small countries, large democratic countries are on a more "level set" with moderately-sized authoritarian countries (as seen in the LSSP plot). This may be in large part due to conflicts between, for example, the Unites States and Middle Eastern countries in the 1990s.

In Figure 3.23(a) we plot the posterior mean probability of connection between these 130 countries. As before, the rows and columns have been permuted from lowest to highest LSSP. This reveals the sizes of the clusters found by the function fitting. One might note that this yields quite a different image of connection probabilities than the multiplicative factor model discussed in Chapter 2. The sociomatrix for this network is plotted in Figure 3.23(b) with the same LSSP permutation. This reveals



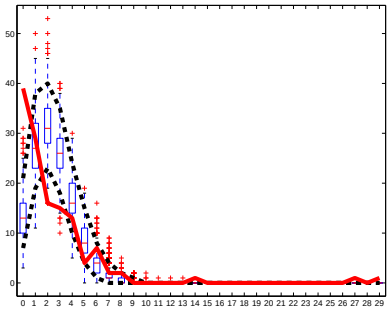
(a) Posterior mean probabilities (sorted)

(b) Sociomatrix (sorted)

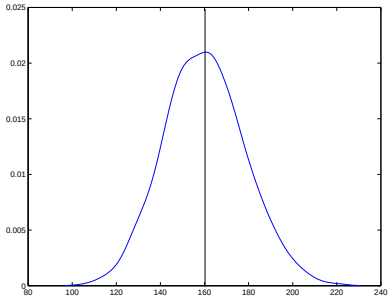
Figure 3.23: IC Example: Sorted posterior mean probabilities and sociomatrix

some evidence of this clustering, though the network is quite sparse.

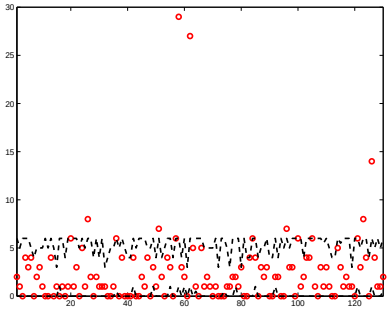
Looking at the more "formal" goodness-of-fit plots reveals some interesting observations about the impact of outliers. The plot of the degree sequence in Figure 3.24(a) shows again the difficulty with generating enough isolated nodes from a model fit to the average. Comparing the observed number of edges to those in the replicate distribution in Figure 3.24(b) shows good correspondence on this network property. More interesting is looking at local statistics such as degree and extended degree. Figure 3.24(c) plots the degree of each country as a red circle and gives pointwise 90% bounds. From this, it seems the model did a fairly good job of explaining individual degrees, except for two outlier countries which are much more connected than any of the others (Iraq and Jordan). Figure 3.24(d) shows the same local fit but of extended degree. This illustrates how inability to fit on some network properties (such as degree) can propagate through to higher-order topologies. By underfitting the highly connected countries Iraq and Jordan, there is much more error in the fit of extended degrees as a result, since their degrees contribute to the extended degrees of all of their neighbours. If we remove Iraq and Jordan from the calculation of degree and



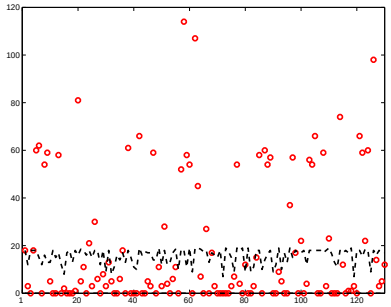
(a) Degree sequence



(b) Number of edges



(c) Actor degrees



(d) Actor extended degrees

Figure 3.24: IC Example: Goodness-of-fit statistics

extended degree, we see that it reduces the number of outliers in the extended degree plot (Figure 3.25). The null distribution of the χ^2 test shown in Figure 3.26 suggests a decent overall fit.

While fitting the LSSP model to this network yields some interesting interpretations, it is also useful to examine how it would predict conflicts if the network had not been collected for all 130 countries, but only a sample. We divide the countries into a random split of $n_T = 105$ training countries and $n_V = 25$ validation countries. For this one split, the observed degree compared to the predicted degree for each validation country is given in Figure 3.27. Figure 3.28 shows the corresponding bivariate distribution of degree (on the horizontal) and extended degree (on the vertical). Averaging

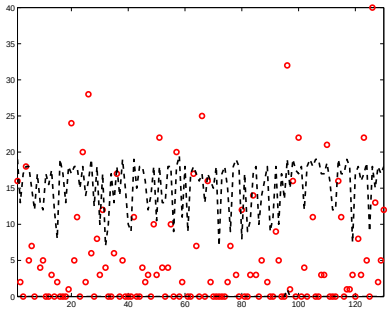


Figure 3.25: IC Example: Extended degree goodness-of-fit plot with highly connected countries removed.

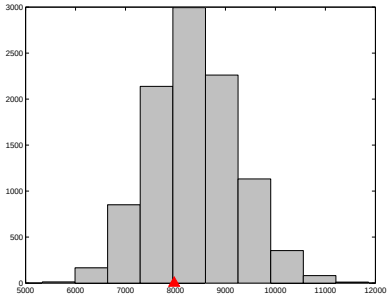


Figure 3.26: IC Example: χ^2 test.

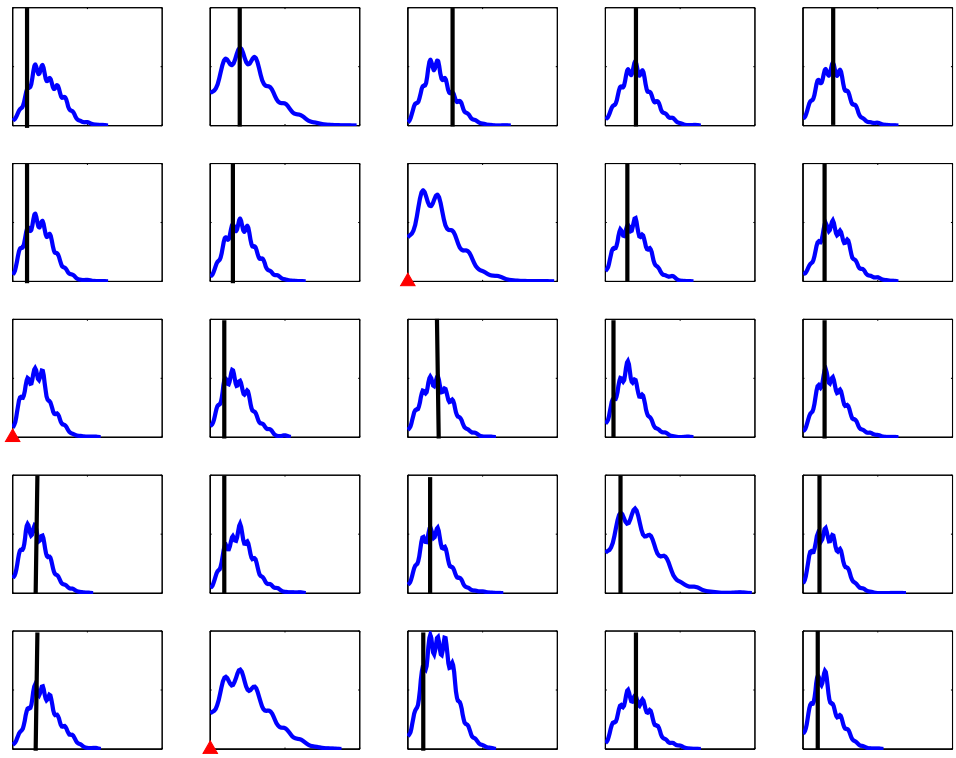


Figure 3.27: IC Example: Validation set degree distribution.

over $S = 50$ splits, we get $\hat{\epsilon}_1 = .49$ and $\hat{\epsilon}_2 = 0.58$.

3.4 Discussion

To conclude this chapter, we include a brief discussion on a couple of points. First, a few comments on computational requirements for the LSSP model. Second, we explore the LSSP assumption that increased differences between attributes always decrease the probability of connection.

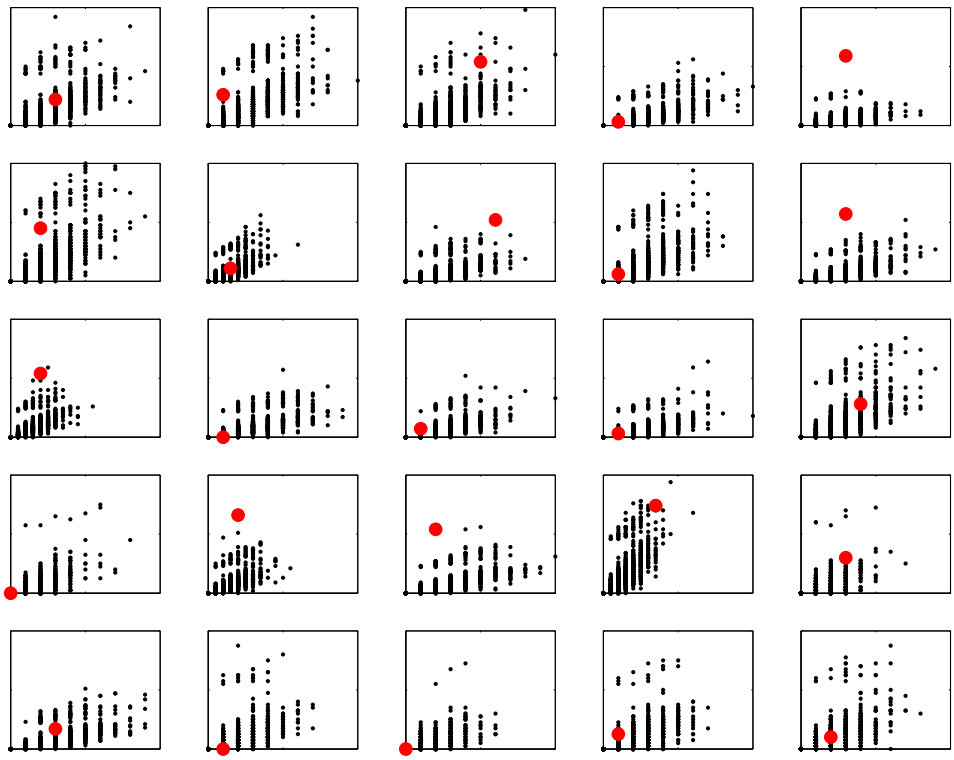
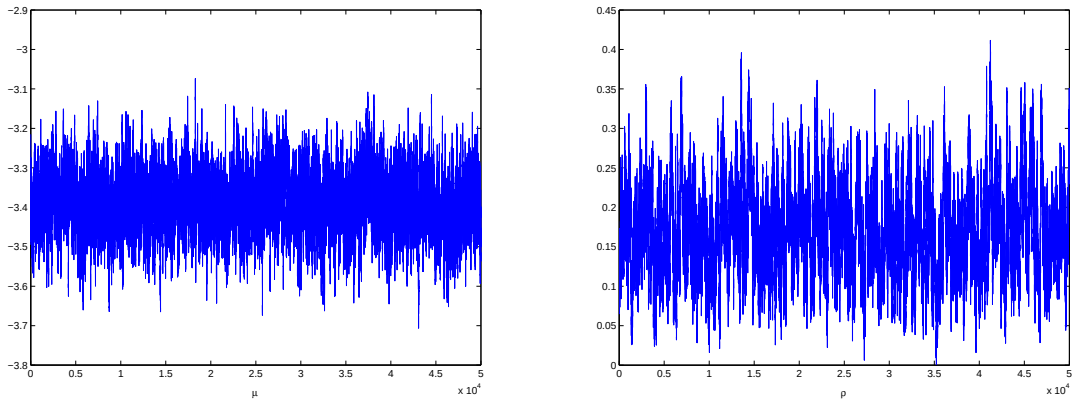


Figure 3.28: IC Example: Validation set degree by extended degree.



(a) Trace plot for μ

(b) Trace plot for ρ

Figure 3.29: AH Example: MCMC trace plots.

3.4.1 LSSP Computational Requirements

As mentioned in the last chapter, for latent factor models the MCMC algorithm must be run for quite a long time before it converges (in the ballpark of 1 million iterations). The LSSP examples we consider in this chapter typically require on the order of 100,000 iterations to converge. The largest example, the adolescent health example with 205 students, will run in about two hours on a laptop programmed in MATLAB. To illustrate what typical trace plots look like, we show trace plots for μ and ρ in the adolescent health example, which we ran for 100,000 iterations (Figures 3.29(a) and 3.29(b)). These plots show the last 50,000 scans which were kept after discarding the first half as burn-in. We find that these two parameters, especially ρ, are in general the slowest to show good mixing.

3.4.2 Revisiting the Homophily Assumption

Recall that the specification of the LSSP model in (3.1) forces the log-odds of connection to decrease as a function of increased differences in LSSP scores. While a

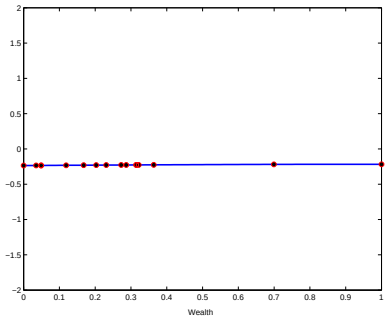


Figure 3.30: FF Example: Fitting an LSSP model

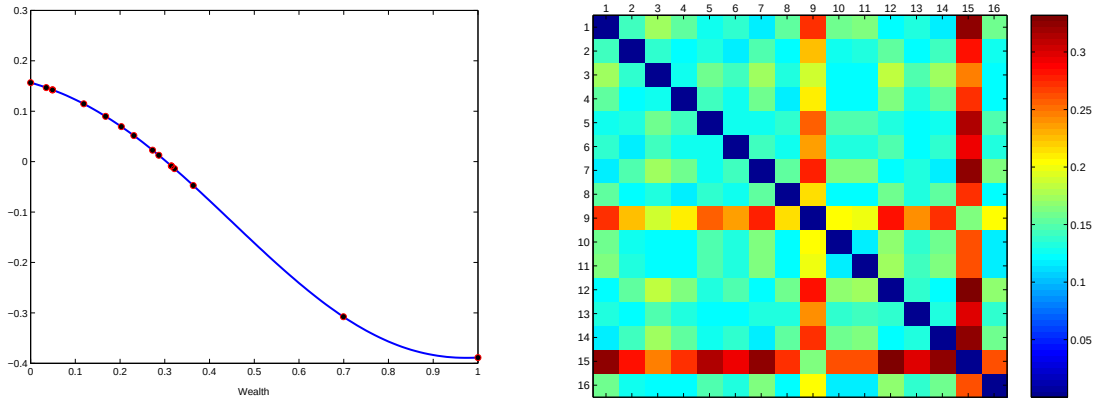
positive correlation between homophily by attributes and connections is a reasonable assumption for many social networks, it will not always be the case. We have actually already seen one network where this assumption is violated, the Florentine Family network. Here, the likeliness of a marriage increases with more discrepancy in wealth – perhaps some marriages were arranged for mercenary reasons.

If the LSSP model is fit to a network for which the assumptions do not hold, the result is what would be expected. Figure 3.30 shows the posterior mean of the LSSP over the range of x, wealth of family (scaled to $[0, 1]$) when the LSSP model is fit to the Florentine Family network. Seeing this result one would immediately question if the assumptions of the model hold.

Provided it is known *a priori* that all covariates have this opposite effect, the model can simply be specified as

$$\eta_{i,j} = \mu + |z(\mathbf{x}_i) - z(\mathbf{x}_j)|. \tag{3.7}$$

Fitting this alternative model to the Florentine Family network yields a more satisfactory fit. Figure 3.31(a) shows the new posterior mean of the LSSP over wealth (again scaled to $[0, 1]$). The posterior mean probabilities of connection are shown in Figure 3.31(b). Out of interest, the two families that are most likely to be involved



(a) Posterior mean LSSP (b) Posterior mean probabilities

Figure 3.31: FF Example: Fitting a positive LSSP model

in marriages are the wealthiest.

In general, suppose the covariate vector can be separated as $\mathbf{x}_i = (\mathbf{x}_i^-, \mathbf{x}_i^+)$, where $\mathbf{x}_i^-$ are the attributes for which (3.1) holds and $\mathbf{x}_i^+$ are better fit by the alternative (3.7). Then a more encompassing version of the LSSP model is potentially

$$\eta_{i,j} = \mu - |z_1(\mathbf{x}_i^-) - z_1(\mathbf{x}_j^-)| + |z_2(\mathbf{x}_i^+) - z_2(\mathbf{x}_j^+)|,$$

where $z_1(\mathbf{x})$ and $z_2(\mathbf{x})$ are two different surfaces modelled over the separate attribute spaces. We do not pursue this further here, but leave it as a potential avenue for future exploration.

In summary, we propose the LSSP model as a means to explain the tendency for actors to connect via a flexible measure of difference between attributes. Rather than describing how a particular set of actors connect, we seek to model how similar actors might connect. Of particular note, we are able to do this with significantly fewer parameters than the latent factor models previously discussed. There are some weaknesses of this methodology, however, such as the assumed direction of homophily

by attributes. In the next chapter, we consider the possibility of a more general model specification.

Chapter 4

The Meta-Distance Model

As we have been delineating, building models that average the relationship between attributes and pairwise binary observations is a challenging – but useful – exercise. In the last chapter we introduced the use of a spatial process model that assigns a score to each actor as one potential approach to this problem. In this chapter, we expand further upon the idea of using spatial process models for network analysis.

One weakness of the LSSP model is that it requires prior knowledge of whether differences in attributes increase or decrease the log-odds of connection. Spatial process models are so flexible it begs the question, why assume any kind of relationship at all? That is, we consider a more general model

$$\eta_{i,j} = \mu + z(\mathbf{x}_i, \mathbf{x}_j),$$

where z is some unspecified function of $\mathbf{x}_i$ and $\mathbf{x}_j$. Rather than yielding scores (which are then compared), here the latent function directly models the log-odds of connections up to a constant shift. As appealing as this sounds, we are somewhat restricted by the requirement that covariates for *two* actors have to be incorporated into the

model. In particular, in the case of undirected graphs, this has to happen in a symmetric way. Our intention in this chapter is to take some initial steps in the exploration of this kind of model. We emphasize that there are many kinds of network data, and this approach may be a reasonable alternative for some.

To incorporate pairs of covariates symmetrically, we turn again to the concept of homophily by attributes. Despite our argument in the last chapter that absolute differences between covariates may not always be appropriate, we rely on these for this alternative model. In particular, we consider using

$$\eta_{i,j} = \mu + z(|\mathbf{x}_i - \mathbf{x}_j|) = \mu + z(|x_{i1} - x_{j1}|, \ldots, |x_{ip} - x_{jp}|) \tag{4.1}$$

to directly model the log-odds of connection. Note that here we use a slight abuse of notation, $|\mathbf{x}_i - \mathbf{x}_j| = (|x_{i1} - x_{j1}|, \ldots, |x_{ip} - x_{jp}|)$ for convenience. We will use the same spatial process approach for modelling z as before, but here the correlation between $\eta_{i,j}$ and $\eta_{k,l}$ for any two pairs will depend on the distances (relative to fixed points $\mathcal{W}$) between their attribute dissimilarities, $|x_{id} - x_{jd}|$ and $|x_{kd} - x_{ld}|$, in each component $d = 1, \ldots, p$. That is, two pairs of actors that are equally dissimilar in all attributes will have the same log-odds of connection. This "distance of distances" interpretation prompted us to call this the meta-distance (MD) model.

We note that if comparing absolute differences is not ideal for a particular covariate, then if possible a marginal transformation should be made before using (4.1). For example, a log transformation may work well for some attributes. Whether the LSSP or this formulation (or neither) is preferable will certainly depend on the application. In general, leaving z unspecified in (4.1), however, does allow for interactions between the covariate dissimilarities as well as complex marginal effects.

Many of the ideas involved with the estimation of the LSSP model can be borrowed in this context. Assuming conditional independence of the $y_{i,j}$ given the $\eta_{i,j}$, i.e.

given μ and z, we have the same Bernoulli sampling model (2.3) for the pairwise observations. For the function z, which in this context we refer to as the MD surface, we use the prior class of functions

$$z(|\mathbf{x}_i - \mathbf{x}_j|) = \sum_{r=1}^{m} \alpha_r k_\rho(|\mathbf{x}_i - \mathbf{x}_j| - \mathbf{w}_r). \tag{4.2}$$

Assuming the covariate space is scaled to be $\mathcal{X} = [0,1]^p$, then $0 \leq |x_{id} - x_{jd}| \leq 1$ for $d = 1, \ldots, p$, as well. The MD surface (4.2) is therefore taken to be over the space of covariate differences, which we define as $\mathcal{X}_D = [0,1]^p$. This space ranges from differences of 0 to differences of 1 in each covariate. The centers $\mathcal{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_m : \mathbf{w}_r \in \mathcal{X}_D\}$ are now interpreted as fixed points in this space of differences. As before, we choose $\mathcal{W}$ as a LHD with typically $m = 10p$ points.

The unknowns in (4.2) are $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)$. The correlation parameters $\boldsymbol{\rho}$ for the MD model dictate how much a change in the dissimilarity of the d^{th} covariate impacts the log-odds of connection. For a prior for $\boldsymbol{\alpha}$, we use

$$\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{I}_m).$$

Note that this implies

$$cov[z(|\mathbf{x}_i - \mathbf{x}_j|), z(|\mathbf{x}_k - \mathbf{x}_l|)] = \sum_{r=1}^{m} k_\rho(|\mathbf{x}_i - \mathbf{x}_j| - \mathbf{w}_r) k_\rho(|\mathbf{x}_k - \mathbf{x}_l| - \mathbf{w}_r).$$

This reiterates the meta-distance interpretation. For the $\boldsymbol{\rho}$ parameters we assume

$$\rho_d \sim U[0,1]; \quad d = 1, \ldots, p.$$

While these priors specify the prior class of functions for z, we are left to determine a prior for μ in (4.1). Here, a vague $N(0, \psi_\mu)$ prior is used.

The posterior distribution for the MD model is simply

$$[\mu, \boldsymbol{\alpha}, \boldsymbol{\rho}|\mathbf{y}] \propto [\mathbf{y}|\mu, \boldsymbol{\alpha}, \boldsymbol{\rho}][\mu][\boldsymbol{\alpha}][\boldsymbol{\rho}],$$

where $[\mathbf{y}|\mu, \boldsymbol{\alpha}, \boldsymbol{\rho}]$ is the Bernoulli likelihood (2.3) with $\eta_{i,j}$ specified by (4.1). Sampling from the full conditionals requires a Metropolis-Hastings step. By taking a large number of draws $(\mu^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\rho}^{(t)})$, $t = 1, \ldots, T$, from the posterior distribution, we can easily predict the height of the MD surface for any pair of actors i_0 and j_0 with attributes $\mathbf{x}_{i_0}$ and $\mathbf{x}_{j_0}$ by

$$\hat{z}(|\mathbf{x}_{i_0} - \mathbf{x}_{j_0}|) = \mathbb{E}[z(|\mathbf{x}_{i_0} - \mathbf{x}_{j_0}|)|\mathbf{y}] \approx \frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{m} \alpha_r^{(t)} k_{\rho^{(t)}}(|\mathbf{x}_{i_0} - \mathbf{x}_{j_0}|).$$

Similarly, the predicted probability of connection for the pair is

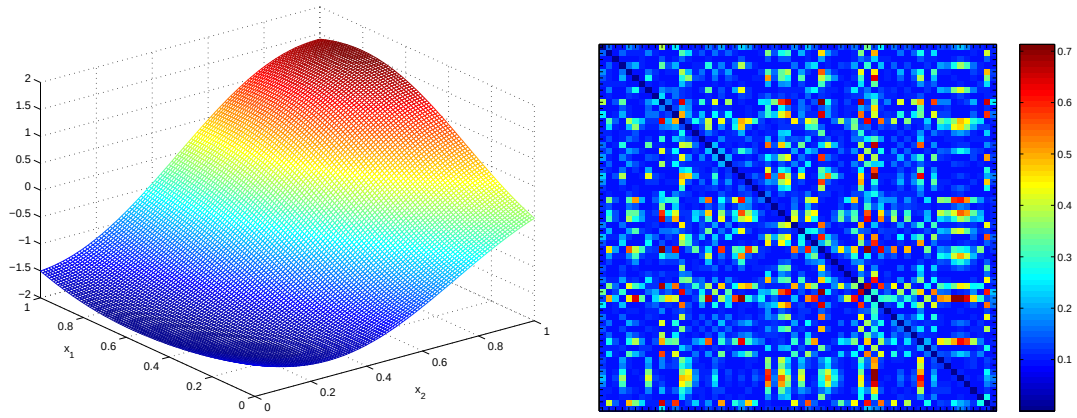
$$\hat{\pi}_{i_0,j_0} = \mathbb{E}\left[\frac{exp\{\eta_{i_0,j_0}\}}{1 + exp\{\eta_{i_0,j_0}\}}|\mathbf{y}\right] \approx \frac{1}{T}\sum_{t=1}^{T}\frac{exp\{\eta_{i_0,j_0}^{(t)}\}}{1 + exp\{\eta_{i_0,j_0}^{(t)}\}},$$

where

$$\eta_{i_0,j_0}^{(t)} = \mu^{(t)} + z^{(t)}(|\mathbf{x}_{i_0} - \mathbf{x}_{j_0}|).$$

We illustrate the interpretation and fitting of the MD model with two examples.

Example: Meta-distance (MD) Synthetic Network. We begin by constructing a network under the assumption that the MD model is true. Consider the covariate space $\mathcal{X} = [0, 1]^2$. We randomly generate $n = 60$ actors with attributes $\mathbf{x}_i \in \mathcal{X}$, $i = 1, \ldots, n$. The distances $|x_{id} - x_{jd}|$ are calculated for each pair $i < j$, $d = 1, \ldots, p$. These distances lie in the space $\mathcal{X}_D = [0, 1]^2$ representing distances between pairs. Figure 4.1(a) shows the MD surface over $\mathcal{X}_D$ that specifies the log-odds that each pair of actors is connected for this example. Here, we see that the log-odds of connection for a pair is greatest when they have a dissimilarity of one in both attributes. Conversely, it is lower when they are the same on both attributes. Note that the LSSP model would do poorly when this is an accurate description of the relationship between covariates and connections. We emphasize the difference in how the MD surface is



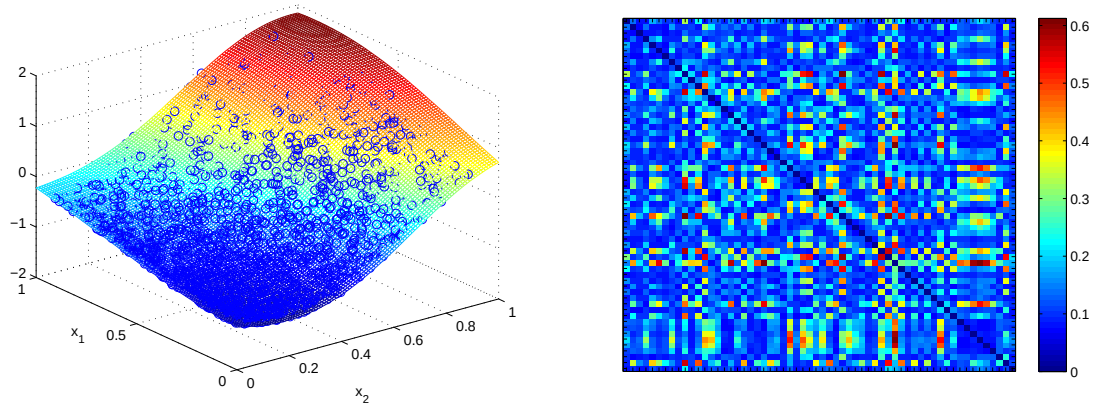
(a) The meta-distance surface

(b) Probabilities of connection

Figure 4.1: MD Example: Latent surface and connection probabilities.

interpreted; it is a literal representation of the log-odds of connection for each pair. Therefore, the MD surface is to be estimated for all $\binom{n}{2}$ *pairs*, rather than just on the n actors as for the LSSP model. We generate independent Bernoulli random variables according to the log-odds (4.1) with $\mu = -0.5$ to get our observed network. The probabilities of connection for this example data are shown in Figure 4.1(b).

We begin analysis by specifying 20 kernel basis centers according to a LHD in the difference space $\mathcal{X}_D$. Treating the generated network as the observed observations, we use the above priors (with $\psi_\mu = 10$) and draw realizations from the posterior distribution $[\mu, \boldsymbol{\alpha}, \boldsymbol{\rho}|\mathbf{y}]$. The posterior mean MD surface is shown in Figure 4.2(a). We see that it is a reasonable estimate of the true surface. The heights at locations corresponding to the $\binom{n}{2}$ pairs are marked by circles in this plot. Figure 4.2(b) shows the corresponding mean probabilities of connection (which are averages of draws of the MD surface adjusted by draws of μ). Again, we use the scale of dark blue to dark red to represent smaller to larger probabilities. We see that the fitted probabilities are slightly lower than the true probabilities, which appears to be a by-product of



(a) Posterior mean MD surface

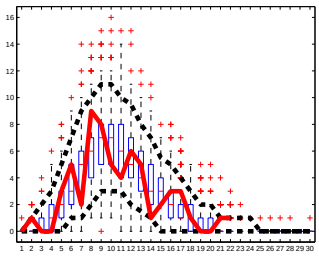
(b) Posterior mean MD probabilities

Figure 4.2: MD Example: Posterior means of latent surface and connection probabilities.

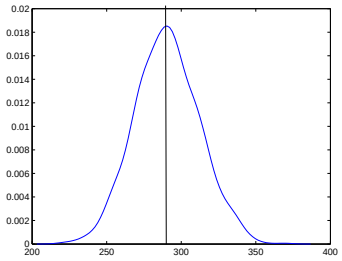
the posterior estimate of μ given this particular data set. With the estimates of the MD surface being pairwise instead of actor-oriented, there is no obvious sorting of the probability maps for the MD model as there was for the LSSP model. We can, however, look at goodness-of-fit plots as before.

For this synthetic example, we consider four goodness-of-fit plots. We begin by looking at the degree sequence and corresponding replicate bounds in Figure 4.3(a). The replicate distribution of the number of edges is shown in Figure 4.3(b). A more local assessment, the individual degrees, are plotted with pointwise 90% bounds in Figure 4.3(c). Finally, in Figure 4.3(d) we have the null distribution for the χ^2 goodness-of-fit test. The observed test statistic is marked by a red triangle. These plots suggest the model did fit the particular observed network well.

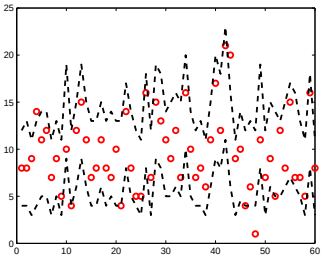
Example: Florentine Family Network. Recall that the LSSP model did not initially fit the Florentine Family marriage network until modifications were considered.



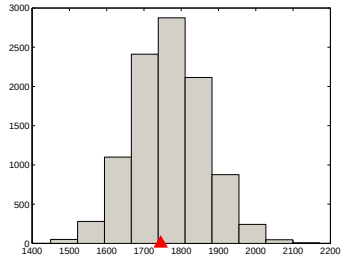
(a) Degree sequence



(b) Number of edges



(c) Degrees for actors



(d) χ^2 null distribution

Figure 4.3: MD Example: Goodness-of-fit plots.

The meta-distance model, on the other hand, will fit this network automatically. We scale the covariate wealth to $[0, 1]$, and then calculate the pairwise differences between the 16 families. Since there is one covariate, we use 10 kernel basis functions to model the latent MD surface. Figure 4.4 shows the resulting posterior mean of the MD function, with the locations of the $\binom{16}{2}$ pairs in the distance space marked by circles. As previously discussed for this network, we see that the log-odds for a pair increases as the difference in wealth growths. This is immediately detected by the MD model. We remark on the interpretation of the MD function. For example, pairs that are dissimilar by 0.6-0.7 in the scaled value of wealth have the highest posterior probability of being connected. On the other hand, the lowest probability of connection is for pairs that have the same wealth. Unlike for the LSSP model, the magnitude of the covariate does not play a role here, i.e. a difference of zero in wealth could imply two families are equally poor or equally wealthy. The resulting posterior mean probabilities of connection are plotted in Figure 4.5. Comparing these posterior mean probabilities to those calculated using the "positive" version of the LSSP model, we see similar results. Here, however, the meta-distance model detects this relationship on its own.

Goodness-of-fit plots can be generated as usual. In Figures 4.6(a) - 4.6(d) we include the degree sequence (and pointwise bounds), the replicate distribution of the number of edges with the observed, the minimum geodesic sequence and pointwise replicate bounds, and finally a χ^2 fit assessment. None of these show serious lack of fit.

We close this chapter with a brief discussion. As we have seen, implementation for the LSSP and MD models is essentially the same, though the interpretation is quite different. The MD model has the advantage that it makes no prior assumptions about

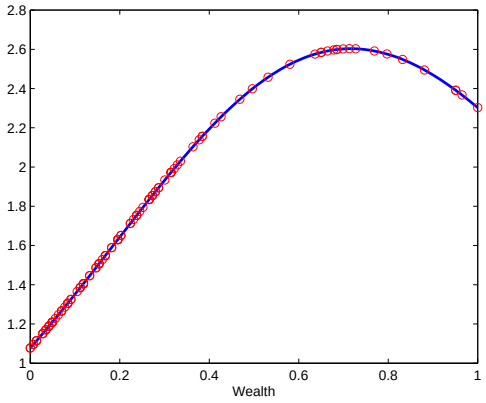


Figure 4.4: FF Example: Posterior mean of MD surface.

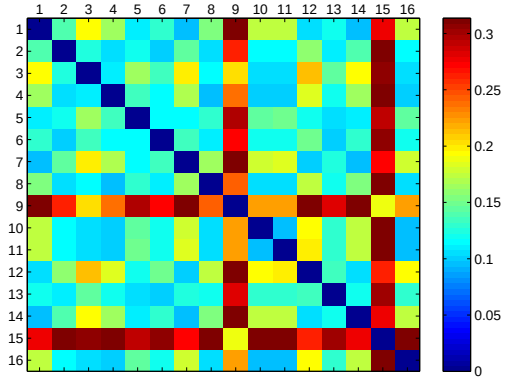
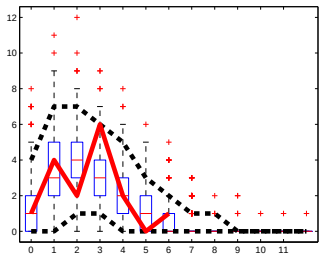
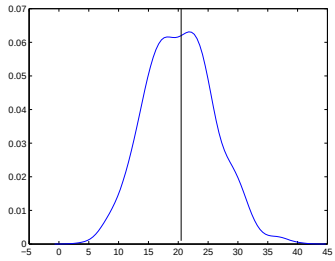


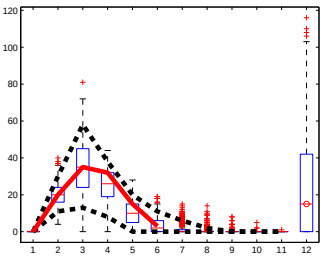
Figure 4.5: FF Example: Posterior mean probabilities of connection.



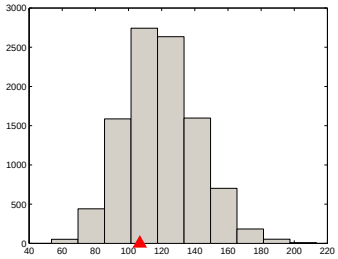
(a) Degree sequence



(b) Number of edges



(c) Minimum geodesic sequence



(d) χ^2 null distribution

Figure 4.6: FF Example: Goodness-of-fit plots.

the direction of the relationship between attributes differences and log-odds. Instead, the MD surface is a literal description of the log-odds of connection for every pair. The benefit of this can be seen in its application to the Florentine Family network, for example. It is also flexible enough to allow for a wide variety of interactions between covariate dissimilarities.

There are some aspects of the MD model of which to be aware, however. Using absolute differences to compare attributes may not be ideal, and this is an important assumption to consider. Possible remedies are covariate transformations, and in the future it may be worth trying to embed a transformation of the original covariate space within the model implementation. Finally, because the MD model operates on pairs, the computation time can be a bit longer. The number of parameters does not increase, but evaluating (4.2) at $\binom{n}{2}$ points is slower than evaluating the LSSP for the n actors. We also find that the MCMC algorithm has to run longer to get reasonable mixing in draws of μ for the meta-distance model. An example of the size we consider above (the two covariate example) takes approximately three hours to run in MATLAB on a laptop. Nonetheless, the MD model may be a reasonable alternative for some network data.

Chapter 5

Reference Distribution Variable Selection

We have now considered two ways that spatial process models can be used to relate attributes to pairwise binary responses. In the LSSP model, a spatial process model is used to assign scores to each actor, and these scores are then compared to predict connection probabilities. The meta-distance model, on the other hand, takes a more direct approach, using a spatial process model to represent a function that links attribute dissimilarities to the log-odds of connection.

A main advantage of spatial process models is that they are extremely flexible, but this can also make them difficult to interpret. Estimated surfaces cannot be easily visualized in more than two dimensions, and there are no coefficients with analytically derived null distributions that can be used to test the importance of covariates. As hinted earlier, the $\boldsymbol{\rho}$ parameters in the spatial process models we have been considering contain information on how much the surface is changing in each direction. In this chapter, we focus on developing a variable selection technique, which we call reference

distribution variable selection (RDVS), that can be used to assess the significance of each component of $\boldsymbol{\rho}$. We develop the methodology here for a completely different application we had in mind initially, screening for important factors in computer experiments. We will return again to our discussion of the LSSP model at the end of the chapter, and apply the ideas presented in what follows.

5.1 Computer Experiments

Rapid growth in computer power has made it possible to study complex physical phenomena that might otherwise be too time consuming or expensive to observe. Scientists are able to adjust inputs to computer simulators (or computer codes) in order to help understand their impact on a system. Many such computer simulators require the specification of a large number of input settings and are computationally demanding. As a result, only a limited number of simulation runs tend to be carried out. Scientists must therefore select the simulation trials judiciously and perform a designed computer experiment (or simply a computer experiment).

One main goal of experimentation, particularly in its early stages, is to determine the relative importance of each input variable in order to identify which have a significant impact on the process being studied. Since there can be many inputs into a computer code, an important problem is the identification of the most active factors.

Most computer experiments are unique in that the response has no random error component. That is, replicates of the same inputs to the computer code will yield the same response. To deal with this, Sacks *et al.* (1989a, 1989b) propose modelling the response from a computer experiment as a realization from a stochastic process. This allows for estimates of uncertainty in a deterministic computer simulation.

The input to the computer code is an $n \times p$ matrix $\mathbf{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$. At each input setting $\mathbf{x}_i$, an output from the computer $y_i = y(\mathbf{x}_i)$ is observed. For simplicity, we assume that the response is standardized to have mean 0 and standard deviation 1, and that the covariate space is scaled to $[0, 1]^p$. The responses $y(\mathbf{X})$ are modelled as the sum of a GP, $z(\mathbf{X})$, which depends on $\mathbf{X}$, and independent white noise. That is,

$$y(\mathbf{X}) = z(\mathbf{X}) + \epsilon, \tag{5.1}$$

where ϵ is a mean zero noise process with variance $1/\lambda_\epsilon$, independent of $z(\mathbf{X})$. Note that in this context, the responses y_i, $i = 1, \ldots, n$, are *observed* realizations of the spatial process z (with error). This is in contrast to the LSSP setting, where the function evaluations are latent and the responses $y_{i,j}$ are pairwise and binary. Recall that under the GP assumption, $(z(\mathbf{x}_1), \ldots, z(\mathbf{x}_n))$ have a multivariate normal distribution. We assume the process has mean zero (due to the standardization of the responses) and use the covariance function

$$cov[z(\mathbf{x}_i), z(\mathbf{x}_j)] = \frac{1}{\lambda_z} \prod_{d=1}^{p} \rho_d^{4(x_{id} - x_{jd})^2}. \tag{5.2}$$

Note that this is a bit different than the covariance structure implied by our radial basis function approximation in Chapter 3. This present formulation is an example of the more traditional approach to modelling Gaussian processes – through direct specification of the mean and covariance function. Here, $1/\lambda_z$ is the variance of the GP, and for this to be a valid covariance function, $0 \leq \rho_d \leq 1$. From (5.2), we see that if ρ_d is close to 1, the process does not depend on the d^{th} covariate. Therefore, estimation of the ρ_d's can help determine which of the input variables are having the most significant impact on the response.

The parameters in the model to be estimated are λ_z, λ_ϵ, and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)$. Here we are using the GP for fitting observed responses, so a little care is required in prior specifications. We assume that the GP $z(\mathbf{X})$ explains most of the variability in the (standardized) response $y(\mathbf{X})$, so for a prior we use

$$\lambda_z \sim GAM(a_z = 5, b_z = 5), \tag{5.3}$$

with the expectation that the variance of the GP should be close to 1, the variance of the response. For the noise, we expect it to be quite small, so we assume

$$\lambda_\epsilon \sim GAM(a_\epsilon = 2.5, b_\epsilon = .025). \tag{5.4}$$

In addition, we include the constraint that $\lambda_\epsilon > 5$. This prevents the standard deviation of the noise from being any more than about 45% of the response standard deviation, and puts its expectation at about 10%.

Since we are looking to use the parameters $\boldsymbol{\rho}$ for variable selection, we use a prior motivated by the variable selection priors from the regression context (e.g. George and McCulloch, 1993). Each component of $\boldsymbol{\rho}$ is given an independent mixture prior of a standard uniform and a point mass at 1,

$$[\rho_d] = \gamma \mathbb{I}[0 \leq \rho_d \leq 1] + (1 - \gamma)\delta_1(\rho_d). \tag{5.5}$$

Here, γ is the prior probability that input d is active and $\delta_1(\cdot)$ denotes a point mass at 1. This is a slight modification of the uniform priors we use in the LSSP setting, but this specification is particularly attractive in the variable selection context because the mixture probability can be chosen to reflect prior beliefs on the number of active factors, thereby incorporating a notion of effect sparsity. For our examples here, we specify $\gamma = 1/4$ to encode a prior belief that about one quarter of the variables will be important.

Let $\mathbf{R}(\rho)$ denote the matrix with entries

$$R_{i,j}(\rho) = \prod_{d=1}^{p} \rho_k^{4(x_{ik}-x_{jk})^2}.$$

Recall this is the correlation function from (5.2). Given the above priors for the parameters, the posterior distribution for the responses $y(\mathbf{X})$ is

$$[\lambda_z, \lambda_\epsilon, \boldsymbol{\rho}|\mathbf{y}] \propto [\mathbf{y}|\lambda_z, \lambda_\epsilon, \boldsymbol{\rho}][\lambda_z][\lambda_\epsilon][\boldsymbol{\rho}],$$

where now

$$\mathbf{y} \sim N\left(0, \frac{1}{\lambda_\epsilon}\mathbf{I}_n + \frac{1}{\lambda_z}\mathbf{R}(\rho)\right),$$

from (5.1). Here again, $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. Realizations $(\lambda_z^{(t)}, \lambda_\epsilon^{(t)}, \boldsymbol{\rho}^{(t)})$, $t = 1, \ldots, T$, can be drawn from the posterior distribution using an MCMC algorithm with a Metropolis-Hastings step. The posterior realizations $\boldsymbol{\rho}^{(t)}$, in particular, can be used to make variable selection decisions. Ideally, one can find a cutoff value for each component ρ_d that can be used to decide if a covariate is active or inert in the spirit of a frequentist hypothesis test. That is what RDVS aims to do.

5.2 Reference Distribution Variable Selection

The flexibility of a spatial process model makes detecting which variables are important a challenging task. When there are p variables, there are 2^p possible combinations of variables in the model. A good discussion on assigning model priors is given in Chipman, George, and McCullough (2001). A fully Bayesian implementation often requires one to specify a prior on all 2^p possible subsets, which is not always straightforward. In addition, variable selection decisions in this context are often subject and sensitive to prior specification.

In this section, a new, simple method for assessing the significance of factors in a GP is introduced. Our approach to identifying which individual estimates of ρ_k are small enough (far enough from 1) to be deemed as evidence of a significant variable parallels a frequentist's approach to variable selection. The central issue is to identify a reference distribution and selection criterion that can be used to assess the importance of each covariate.

Consider estimating the parameters $\boldsymbol{\rho}$ in the GP model. Because it is unknown which of them are important, directly gauging the relative magnitudes of the ρ_d's can be difficult. This is what RDVS seeks to address. To implement the method, an additional variable that is known to be inert – and thus has no impact on the response – is appended to $\mathbf{X}$. This provides information on how an inert variable behaves, and therefore can be used as a benchmark against which true covariates can be compared. We propose to use the distribution of the posterior median of the inert, or null, variable as a reference distribution to decide which of the real inputs are important.

The augmented covariate matrix is constructed by adjoining to $\mathbf{X}$ one additional column, $X^* = (x_{1(p+1)}, x_{2(p+1)}, \ldots, x_{n(p+1)})'$. This results in an $n \times (p+1)$ input matrix. To mimic the p real covariates, the elements of the additional column vector, X^*, range from 0 to 1 (since the covariate space is assumed to be $[0,1]^p$). Ideally, the column vector X^* is orthogonal to each set of columns in $\mathbf{X}$, but in practice this is unlikely to be the case. Instead, we randomly sample X^* from the space of the original matrix $\mathbf{X}$.

By construction, the augmented variable is not a true covariate, and thus has no impact on the response. The analysis proceeds as if there are $p+1$ inputs, but in this case it is known that the last variable is inert. Therefore, the posterior distribution

of ρ_{p+1} is the posterior distribution of the correlation parameter for an inert variable. Because the variable selection problem amounts to deciding which variables have an impact on the response that is distinguishable from noise, the posterior distribution of the true variables can be compared to that of the added variable to decide which can be claimed as active. That is, similar to frequentist hypothesis testing, the posterior distribution of the added variable is used as a reference distribution to assess the importance of the ρ_d's corresponding to the true inputs. The key feature of this approach is that it makes judging the actual magnitude of the ρ_d's unnecessary (i.e. there is no need to specify an arbitrary value below which ρ_d is considered to be sufficiently less than 1). This is beneficial because which ρ_d are "small" is often dependent on the particular data at hand. The only judgment that is necessary for RDVS is whether or not the posterior distribution of the true variables are distinguishable from the posterior distribution of the inert variable. Because the added column X^* could be correlated with some columns in $\mathbf{X}$, this procedure is repeated several times and the posterior distributions of the added inert variables from each iteration are combined to form one reference distribution corresponding to that of a null variable. This has the effect of averaging over all added columns.

There a number of ways one could imagine comparing the estimates of ρ_d for the true variables to this reference distribution, but we consider the following. Each time an inert factor is added to $\mathbf{X}$, the analysis is performed and we summarize the posterior distribution of ρ_{p+1} by its median, $\tilde{\rho}_{p+1}$. The process of inserting the inert variable, running the MCMC, and saving the posterior median of ρ_{p+1} is repeated many times. From this, an estimate of the distribution for the posterior median of a correlation parameter corresponding to an inert variable is obtained. In addition, every realization of ρ_d, $d = 1, \ldots, p$, is recorded at each step of the MCMC for the true

covariates. The posterior median over all the realizations for each ρ_d can be compared to the reference distribution of the inert factor median to assess the importance of factor d.

The disadvantage of this approach is the increase in computational time – though not the complexity – because the MCMC must be run many times in order to construct the reference distribution. However, the approach has many advantages. If $\tilde{\rho}_d$ is compared to, say, the 5^{th} percentile of the null distribution for posterior medians, a frequentist's interpretation of importance can be used, i.e. one would expect to falsely identify an inert factor as significant approximately 5% of the time. If one would prefer to err more on the conservative side, the $10th$ percentile could be used, for example. By using this approach, the posterior distributions of the ρ_d can be compared and assessed.

To summarize the RDVS procedure:

1. Augment $\mathbf{X}$ by creating a new column corresponding to a variable with no significant effect. The added column is selected at random from the covariate space of the original variables.

2. Find the posterior median $\tilde{\rho}_{p+1}$ of the added column.

3. Repeat steps 1 and 2 M times. Obtain a distribution for the posterior median of a null effect to be used as a reference distribution.

4. Compare the posterior medians $\tilde{\rho}_d$ of the true variables to the reference distribution to assess their importance. The percentile of the reference distribution used for comparison reflects the rate of falsely identifying the inert variable as important.

5.3 Simulated Examples

To illustrate the performance of RDVS, we have chosen three simulated examples of varying complexity. Using known functions allows us to evaluate the methodology. For all of the examples, the design matrix used is a 54-run Latin hypercube design (LHD) with $p = 10$ input variables. This parallels our decision to use a LHD for our radial basis function approximation of the LSSP. Latin hypercube designs are a popular choice (such designs were introduced by McKay et al. (1979) specifically for computer experiments) because they can be generated with minimal computational effort and fill the design space relatively well. In addition, when the sample inputs of such a design are projected into any one dimension, complete stratification is achieved. The particular design used in these examples has the additional property that the minimum pairwise distances in each two-dimensional projection is approximately maximized, yielding a space-filling design in each of the $p(p-1)/2$ two-dimensional projections of the design space.

Example 1. The first example is meant to demonstrate the performance of the RDVS methodology for a simple case. To begin, data are generated from the linear model

$$y(\mathbf{x}_i) = 0.2x_{i1} + 0.2x_{i2} + 0.2x_{i3} + 0.2x_{i4} + e_i, \tag{5.6}$$

where $e_i \sim N(0, \sigma^2)$ with $\sigma = 0.05$. After generation, the response is standardized to have mean zero and standard deviation one. For the simulation study, data are generated from the linear model given in (5.6) 1000 times and the important factors using RDVS are identified at each iteration of the simulation. For this example and each of the subsequent examples, $m = 100$ is used in step 3 of the algorithm.

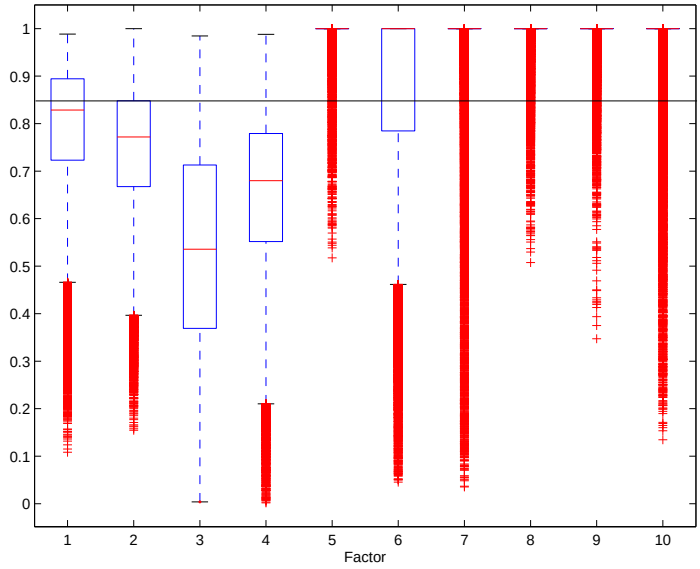
For illustration, consider in detail one iteration of the simulation study. First,

a response is generated as described above. To implement RDVS, an inert variable (i.e. an 11^{th} factor) is added to the design, with levels randomly selected from the design region of the original ten experimental inputs. The GP previously described in Section 5.1 is used to model the response surface. As mentioned, for all examples γ in (5.5) is taken to be 1/4 (in general, γ should be chosen to reflect the user's prior beliefs on effect sparsity).

Using the augmented design matrix, 600 iterations of the MCMC algorithm are run to generate posterior realizations of the ρ_k, $k = 1, \ldots, 11$, under the GP model, with the first 100 discarded as burn-in. The augmentation procedure and MCMC implementation is repeated $M = 100$ times. We find this is sufficient to obtain a reasonable estimate of the distribution for the posterior median of ρ_{11}. All 50,000 realizations of ρ_k for the ten experiment inputs are saved, and the posterior median of the correlation parameter for the inert variable is obtained. The combined 100 posterior medians $\tilde{\rho}_{11}$ form the reference distribution to be used for variable selection.

Figure 5.1 shows boxplots of the posterior realizations of ρ_k ($k = 1, \ldots, 10$) obtained from the MCMC corresponding to one iteration of the simulation study. The 10^{th} percentile of the reference null posterior median distribution is indicated by the solid horizontal line on the figure. There are some features of Figure 5.1 worth noting. As usual, the boxes of the boxplots denote the first, second and third quartiles of a distribution. One can see in this plot that for this data, the posterior distribution of an inert factor, such as factor 5, is pushed up against one. Indeed, for this factor, the upper three quartiles of the posterior distribution are all one. The "tail" on the distribution shows the range of the small fraction of posterior realizations that are less than one. This pattern is also observed for the other inert factors to varying degrees. Conversely, the posterior distribution of an active factor (e.g. factor 1) is centered far

Figure 5.1: Posterior distributions of ρ_k for one iteration of the simulation study in Example 1.



less than one.

By just inspecting these boxplots, an experimenter would likely correctly identify the first four inputs as having a significant impact on the response because the posterior medians are all much less than 1 relative to the other factors. Looking at Figure 5.1, one may be tempted to also declare factor 6 active. However, the posterior median for this factor is exactly 1. If the more formal rule of comparing the posterior distributions of ρ_k for the experimental variables to the 10^{th} percentile of the null median distribution is followed, the first four inputs are indeed correctly identified as being important. Thus, for this iteration of the simulation study, the decision is made to declare the first four inputs as active and the remaining factors as inert.

Table 5.1 summarizes the results for 1000 simulations. The performance of the approach is investigated using the 5^{th}, 10^{th}, and 15^{th} percentiles of the reference distribution as cut-off points. The results show that RDVS does well at correctly

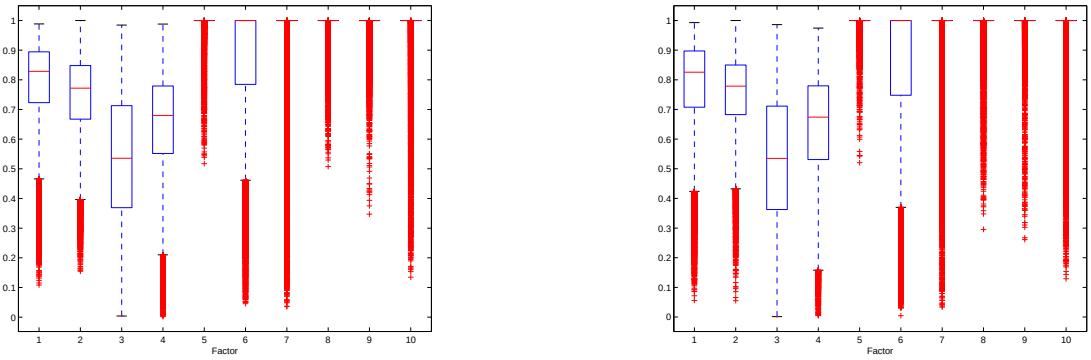
Table 5.1: Proportion of times each factor is identified as important in 1000 generations of the linear function given in (5.6).

	Factor									
Percentile	1	2	3	4	5	6	7	8	9	10
5^{th}	0.619	0.618	0.717	0.631	0.030	0.034	0.021	0.074	0.051	0.051
10^{th}	0.852	0.855	0.910	0.880	0.061	0.064	0.053	0.137	0.076	0.102
15^{th}	0.947	0.954	0.973	0.955	0.079	0.091	0.080	0.173	0.108	0.135

identifying the active factors in this simple example, as would be expected. It can also be seen from Table 5.1 that the false identification of inert inputs as active is at the expected level corresponding to the percentile used for decision making.

Before continuing, we make a brief digression back to the iteration of the simulation study explored in detail throughout this example. One might question if the addition of the extra variable for RDVS has an impact on the posterior distribution of the experimental variables. To explore this point, the MCMC analysis is repeated on this same response without adding the inert factor. Figure 5.2(a) shows the posterior distributions of the experimental variables when the extra factor is added, while Figure 5.2(b) shows the same distributions generated without using an augmented design matrix. The similarity of these plots suggests that there is no obvious altering of the experimental posterior distributions as a side effect of the methodology. Furthermore, inspection of the differences between posterior medians corresponding to the two approaches (with and without augmentation) showed an average difference of only 2.85×10^{-4}.

In order to explore the size of effects the RDVS selection method is able to detect, consider repeating this simulation study with a slightly more complex linear function. The response is now generated according to a linear function with decreasing



(a) With augmentation (b) Without augmentation

Figure 5.2: Posterior distributions of the experimental variables.

coefficients on the first eight inputs:

$$y(\mathbf{x}_i) = 0.2x_{i1} + \frac{0.2}{2}x_{i2} + \frac{0.2}{4}x_{i3} + \frac{0.2}{8}x_{i4} + \frac{0.2}{16}x_{i5} + \frac{0.2}{32}x_{i6} + \frac{0.2}{64}x_{i7} + \frac{0.2}{128}x_{i8} + e_i, \quad (5.7)$$

where again $e_i \sim N(0, \sigma^2)$ with $\sigma = 0.05$. After generation, the response is standardized to have a mean zero and standard deviation of one. Table 5.2 gives the results for 1000 simulations of this response. From these results, it can be seen that the first factor is still easily identified as active, which is consistent with the previous results. In addition, the second and third factors are detected as active more often than would be expected by chance, while the remaining inputs (which have relatively small coefficients) are determined to be inert for the most part.

Example 2. For our second example, we explore how well RDVS can correctly identify a complete lack of signal. Welch et al. (1992) observe it is difficult to distinguish between a model with no active factors and one with all active factors. Indeed, the sequential likelihood approach to screening they proposed does not distinguish between these two models. In this case, because RDVS decisions are made by making comparisons with an inert variable, it is anticipated the methodology will be able to

Table 5.2: Proportion of times each factor is identified as important in 1000 generations of the linear function given in (5.7).

	Factor									
Percentile	1	2	3	4	5	6	7	8	9	10
5^{th}	0.679	0.180	0.062	0.025	0.016	0.023	0.017	0.031	0.009	0.036
10^{th}	0.889	0.379	0.133	0.058	0.034	0.051	0.035	0.067	0.030	0.094
15^{th}	0.959	0.540	0.217	0.092	0.061	0.098	0.065	0.107	0.063	0.149

correctly detect a lack of activity amongst the experimental variables when none exists. For this example, the response is generated as random noise. That is, $y(\mathbf{x}_i) = e_i$, where $e_i \sim N(0, \sigma^2)$ with $\sigma = 0.05$, and analysis proceeds as in Example 1. Figure 5.3 shows boxplots corresponding to one iteration of this simulation study.

Note that in this plot it appears that all factors have correlations much less than one and seem to be significantly impacting the response. This is because the amount of variability that can be attributed to random noise is restricted in the model, and therefore the GP tries to interpolate a signal through most of the "jitter". In this case, based on a subjective examination of the boxplots, an experimenter would likely incorrectly declare all the ρ_k's to be less than one (and therefore important). When RDVS is used, however, the extra null factor added for the analysis looks and behaves like all the other inert factors, as indicated by the low value of the 10^{th} percentile of the reference distribution drawn as a solid horizontal line in the figure. As a result, when the RDVS decision rule is used, the correct variable selection decisions are made. This illustrates the point that RDVS is based on comparisons between the experimental factors and the inert factor, not on the actual values of the realized ρ_k's. The results from 1000 simulations are given in Table 5.3. It can be seen from these results that RDVS performs extremely well in this setting.

Figure 5.3: Posterior distributions of ρ_k for one iteration of the simulation study in Example 2.

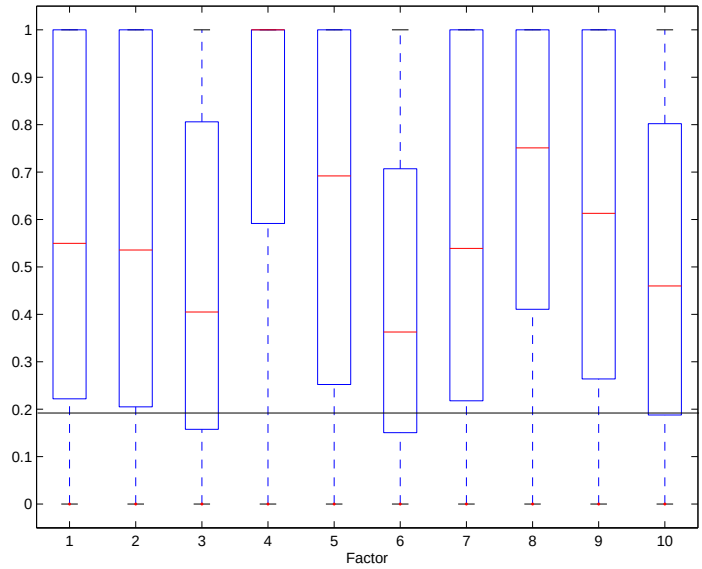


Table 5.3: Proportion of times each factor is identified as important in 1000 generations of random noise.

	Factor									
Percentile	1	2	3	4	5	6	7	8	9	10
5^{th}	0.003	0.013	0.004	0.007	0.006	0.008	0.001	0.006	0.003	0.005
10^{th}	0.012	0.039	0.009	0.013	0.016	0.022	0.010	0.017	0.011	0.013
15^{th}	0.033	0.064	0.032	0.039	0.029	0.041	0.023	0.027	0.027	0.031

Table 5.4: Proportion of times each factor is identified as important in 1000 generations of the response given by (5.8).

	Factor									
Percentile	1	2	3	4	5	6	7	8	9	10
5^{th}	1.000	1.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.001
10^{th}	1.000	1.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.001
15^{th}	1.000	1.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.001

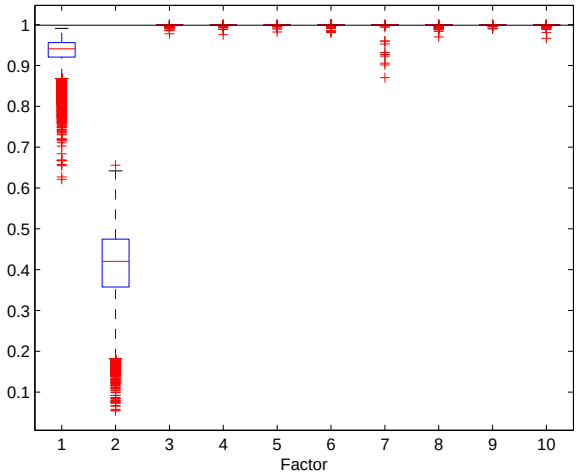
Example 3. For the third example, the data is generated according to

$$y(\mathbf{x}_i) = \sin(x_{i1}) + \sin(5x_{i2}) + e_i, \tag{5.8}$$

where again, $e_i \sim N(0, \sigma^2)$ with $\sigma = 0.05$ and the response is standardized. This function is slightly more complex than the others considered because factor 1 and factor 2 impact the response quite differently over their $[0, 1]$ ranges.

Figure 5.4 shows the posterior distribution of ρ_k, $k = 1, \ldots, 10$, for one iteration of this simulation. For this data, the posterior distributions corresponding to the inert variables are all pushed tightly against one. As it should, the added null variable mimics this behaviour, as can be seen by looking at the 10^{th} percentile of the distribution for posterior medians of inert variables drawn on the figure. As a result, RDVS correctly detects that the distributions for ρ_1 and ρ_2 look discernibly different than the distribution for ρ of an inert factor. Table 5.4 summarizes the results for 1000 simulations. For this example, RDVS does very well at identifying factors 1 and 2 as having a significant impact on the response.

Figure 5.4: Posterior distributions of ρ_k for one iteration of the simulation study in Example 3.



5.4 Sensitivity to Choice of Prior Distributions

To further understand the performance of RDVS, it would be beneficial to consider its robustness to the choice of prior distributions. Recall from Section 5.1 that priors are assigned for the GP parameters λ_z, λ_ϵ, and $\boldsymbol{\rho}$, and that in this case – where function evaluations are observed – these hyperparameters can have an impact on the model fit. The prior assigned to λ_z was a gamma distribution with parameters a_z and b_z chosen so that $E(\lambda_z) = 1$. This selection was made to reflect the prior belief that the GP $z(X)$ should account for essentially all of the variability in the standardized response. This is expected in this setting, so we do not explore alternative priors on λ_z.

A gamma prior was also used for the white noise variability λ_ϵ, governed by parameters a_ϵ and b_ϵ. The prior on λ_ϵ specifies the amount of variability in the response that can be attributed to random error. We chose a_ϵ and b_ϵ so that $E(\lambda_\epsilon) = 100$; that

is, so that it is expected only about 10% of the response standard deviation is explainable by random error. We also had the additional constraint that $\lambda_\epsilon < 5$, which prevented the white noise component from absorbing any more than about 45% of the response standard deviation at any realization of the MCMC analysis.

To investigate the robustness of RDVS to the choice of prior on λ_ϵ, we try varying the choice of b_ϵ. For fixed a_ϵ, changing b_ϵ allows for adjustments to the mean of this prior distribution. Consider again the linear response function given by (5.6) in Example 1 of the previous section. The simulation study on this response function is repeated with two alternative prior choices for λ_ϵ. First, a $\Gamma(a_\epsilon = 2.5, b_\epsilon = .0025)I_{[\lambda_\epsilon 5]}$ prior is used. Under this prior, $E(\lambda_\epsilon) = 1000$, which implies only about 3% of the response standard deviation is expected to be attributable to noise. The same lower bound constraint is kept. An example of the impact on the analysis due to making this particular change on b_ϵ can be seen in the boxplots of the ρ_k posterior distributions given in Figure 5.5(a). For this plot, the same linear response used for the detailed illustration of RDVS in Example 1 is used. This prior encourages the GP to account for more of the variability in the response, which manifests itself as an increased signal, or more values far from one in the boxplots. However, the added inert variable is given the same prior, and it self-calibrates itself to behave like the other inert factors. As a result, the 10^{th} percentile cut-off of the reference distribution is also farther from one, and the correct variable selections are still made. The results over 1000 simulations (with the response generated by (5.6)) are given in Table 5.5. This table shows a slight decrease in the frequency of the correct detection of the first four factors compared to Table 5.1 of Example 1.

Alternatively, we consider changing the prior on λ_ϵ to encourage more of the variability to be absorbed by the random error component. To do this, a $\Gamma(a_\epsilon =$

(a) $b_\epsilon = 0.0025$ (b) $b_\epsilon = 0.1$

Figure 5.5: Posterior distributions of the experimental variables corresponding to changes in the prior on λ_ϵ.

Table 5.5: Proportion of times each factor is identified as important in 1000 generations of the response when the prior on λ_ϵ has $b_\epsilon = 0.0025$.

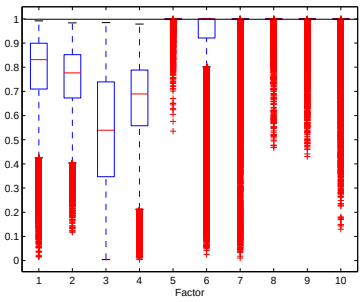
	Factor									
Percentile	1	2	3	4	5	6	7	8	9	10
5^{th}	0.568	0.578	0.680	0.566	0.043	0.041	0.028	0.079	0.050	0.067
10^{th}	0.777	0.809	0.877	0.801	0.078	0.081	0.058	0.135	0.101	0.130
15^{th}	0.902	0.909	0.944	0.915	0.108	0.107	0.092	0.194	0.133	0.180

Table 5.6: Proportion of times each factor is identified as important in 1000 generations of the response when the prior on λ_ϵ has $b_\epsilon = 0.1$.

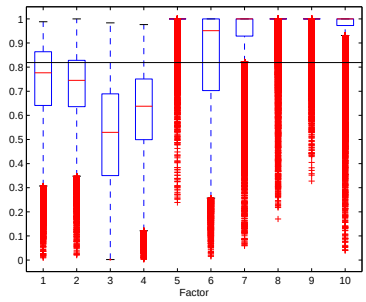
	Factor									
Percentile	1	2	3	4	5	6	7	8	9	10
5^{th}	0.854	0.862	0.895	0.871	0.028	0.027	0.028	0.078	0.040	0.039
10^{th}	0.969	0.979	0.988	0.976	0.043	0.044	0.041	0.107	0.058	0.058
15^{th}	0.995	0.997	0.998	0.995	0.047	0.055	0.046	0.121	0.068	0.063

$2.5, b_\epsilon = 0.1)I_{[\lambda_\epsilon > 5]}$ prior on λ_ϵ is used. Given this value of b_ϵ, $E(\lambda_\epsilon) = 25$, so about 20% of the response standard deviation is expected to be in the error. This has the opposite impact on the posterior distribution of the ρ_k as the previous change. In this case, the boxplots corresponding to inert factors are pushed against one. Mimicking this behaviour, the posterior distribution of the added inert factor is also pushed closer to one, as illustrated in Figure 5.5(b) (again, the same example response was used for this plot). The results from 1000 simulations are displayed in Table 5.6. Here, the first four factors are correctly determined to be active with a higher frequency than in Example 1. Overall, changing this prior does have some impact, but due to the self-calibration of the added inert variable, the performance of the RDVS methodology is still quite good.

We next explore the impact of changing the prior for ρ on the methodology. This mixture prior, given in (5.5), is specified by γ, the prior probability that a factor is active. In all of the previous examples, $\gamma = 1/4$ was taken to be a reasonable value. Consider now two alternative values of γ: $\gamma = 1/10$ and $\gamma = 1/2$. We believe these to be extremities in terms of prior beliefs on effect sparsity. Returning to the linear function given in (5.6), the simulations are repeated with these varying priors. Again, because the added factor has the same prior information as the other factors, its corresponding posterior distribution still mimics those of the other inert factors



(a) $\gamma = 0.1$



(b) $\gamma = 0.5$

Figure 5.6: Posterior distributions of the experimental variables corresponding to changes in the prior on ρ_k.

Table 5.7: Proportion of times each factor is identified as important in 1000 generations of the response when the prior on ρ has $\gamma = 0.1$.

	Factor									
Percentile	1	2	3	4	5	6	7	8	9	10
5^{th}	0.652	0.637	0.742	0.674	0.036	0.019	0.023	0.069	0.030	0.041
10^{th}	0.887	0.878	0.907	0.892	0.054	0.038	0.037	0.096	0.052	0.076
15^{th}	0.963	0.955	0.981	0.966	0.064	0.048	0.047	0.109	0.063	0.093

in the analysis. Figure 5.6 demonstrates this point for the same illustrative response used throughout. As can be seen in Tables 5.7 and 5.8, the performance of RDVS is quite robust to the prior choice of γ.

5.5 Cylinder Deformation Application

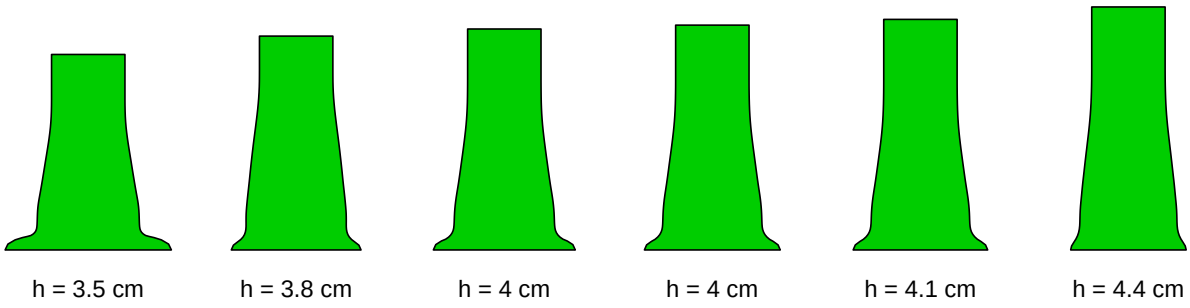
Detailed computer simulation of physical processes plays an important role in the development and understanding of physics-based mathematical models. One of the applications from Los Alamos National Laboratory (LANL) is a finite element code

Table 5.8: Proportion of times each factor is identified as important in 1000 generations of the response when the prior on ρ has $\gamma = 0.5$.

	Factor									
Percentile	1	2	3	4	5	6	7	8	9	10
5^{th}	0.659	0.675	0.759	0.661	0.037	0.035	0.035	0.114	0.043	0.072
10^{th}	0.844	0.875	0.921	0.868	0.068	0.070	0.074	0.172	0.093	0.120
15^{th}	0.939	0.945	0.974	0.948	0.099	0.110	0.113	0.235	0.148	0.176

that simulates a high velocity impact of a cylinder (hereafter referred to as the Taylor cylinder experiment). In this experiment, a copper cylinder (length = 5.08 cm, radius = 1 cm) is fired into a fixed barrier at a velocity of 177m/s. The resulting impact deforms the cylinder according to the elastic-plastic deformation model of Preston, Tonks, and Wallace (2003), the PTW model. This model is governed by 14 parameters (factors), which control the behaviour of the cylinder immediately after impact. Figure 5.7 shows a sample of cylinder deformations corresponding to a range of settings for these input parameters.

Figure 5.7: Collection of simulated cylinders ranging from most compressed to the least taken from the set of 118 simulations of the Taylor cylinder test.



The PTW model was developed to be applicable for a wide range of strain rates, and, in general, all of the factors play an important role in simulating the deformation. Indeed, this is why they were included as inputs to the computer code. However,

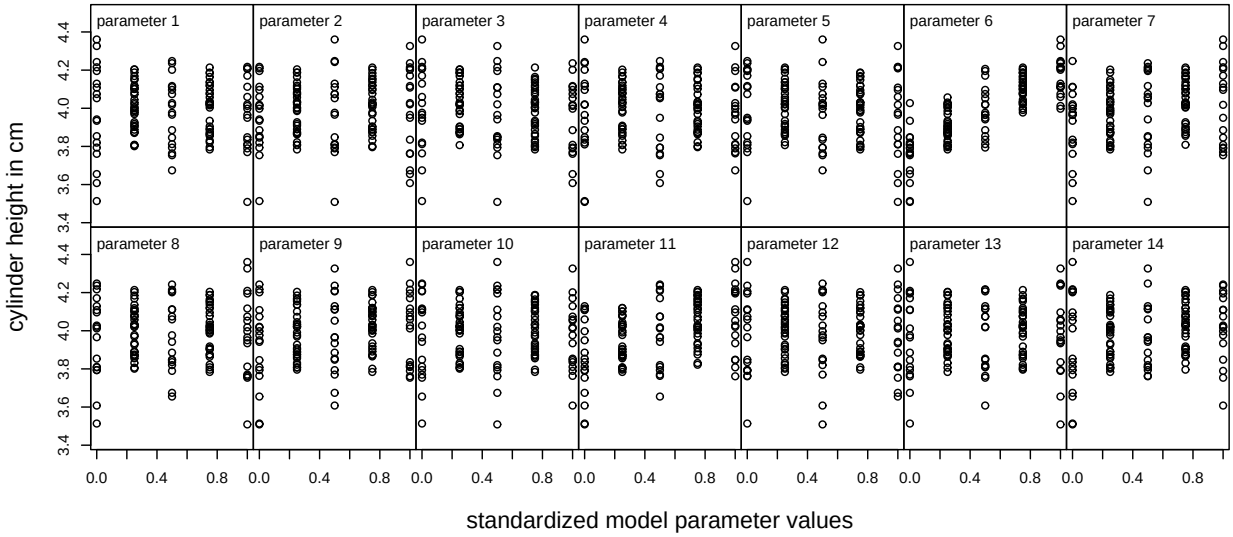
a computer experiment frequently exercises the simulator over a limited range of physical conditions (e.g., velocities or strain rates). Over this range, the simulator response is often dominated by a very limited number of input parameters.

At the input velocity of 177m/s used for this experiment, it is expected that deformation will only be affected by a subset of the 14 input parameters. Furthermore, the Taylor cylinder experiment is only a small component of broader experimentation, so reducing the number of factors to carry on to further experiments is beneficial. The goal of this study is to identify which factors most significantly impact the deformation (i.e., screening) over the reduced input space of the complex computer simulator.

A computer experiment was performed based on a five-level, nearly orthogonal array design (Wang and Wu, 1992), which prescribed 118 different input settings at which to carry out the simulation trials for the 14 factors. We look at the length of the cylinder after impact as our response. Figure 5.8 shows plots of the simulated cylinder length against the five standardized settings for each of the 14 input factors. From this rudimentary figure it appears that factor 6, which controls how temperature and density affect the plastic stress of the metal, is most important. It is difficult to otherwise distinguish between the factors, so RDVS is used to determine which of the factors are influencing the cylinder lengths after impact. To implement RDVS in this setting, the 118×14 design matrix, X, is (repeatedly) augmented with an additional column and the model outlined in Section 5.2 is fit. The idea is that this would be the analysis followed if this were a 15 factor experiment. In this case, however, it is known for certain that the 15^{th} factor is inert.

To be comparable, the added factor should be treated in the same manner as the experiment factors in both the design and analysis stages. Thus the added

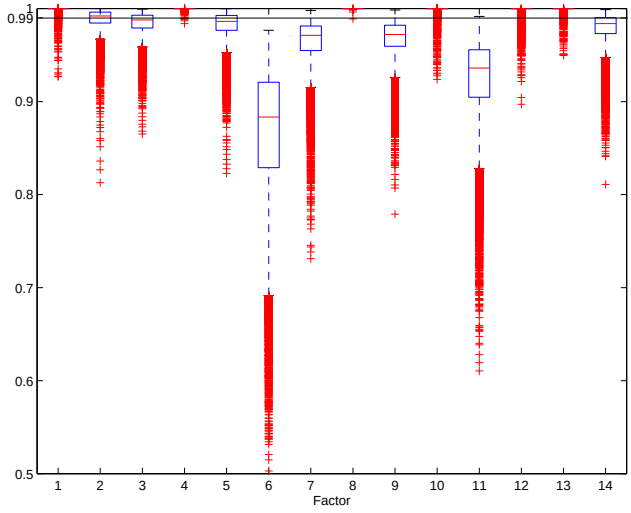
Figure 5.8: Plots of 118 simulated cylinder heights versus standardized input parameter settings for each of the 14 parameters governing the plastic-elastic flow model used to model the cylinder deformation.



factor should have five level settings, with 23 or 24 trials per setting, corresponding to the original 5-level, nearly orthogonal design matrix. To create the random added column, we begin with a vector which has five equally spaced level settings, $(0, 0.25, 0.50, 0.75, 1.00)$, with 23 copies of each level (i.e., a 115×1 vector). Next, three additional trials from the five level settings are randomly chosen, giving 118 trials for the added factor. The vector is then randomly permuted, resulting in the added column. This procedure is repeated for each of the $m = 100$ added columns.

A quick glance at Figure 5.9 reveals that our initial intuition is confirmed (i.e., factor 6, the impact of temperature and density on the stress rate, is an important factor). When the 10^{th} percentile of the posterior distribution of the median correlation parameter for the inert column is drawn (the solid horizontal line in Figure 5.9), seven factors are identified as active: factors 3, 5, 6, 7, 9, 11, and 14. Notice that

Figure 5.9: Posterior distributions of ρ_k for the experiment factors in the Taylor cylinder experiment.



factor 2 is deemed inert since the posterior median of ρ_2 (0.9921) is larger than the cut-off (0.9909) computed from the posterior distribution of the median correlation parameter for the added factor. It is likely, however, that an experimenter may consider carrying factor 2 forward to the next stage of investigation if the cost of doing so is not prohibitive.

After carrying out this analysis, a more in-depth investigation of the simulation output was made. It was found that the "overdriven" regime of the PTW model, which accounts for behaviour at very high strain rates, was never accessed in any of these 118 simulation runs. Hence factors 12 and 13, which govern the behaviour of the model in this regime, are truly inert for the purposes of this screening study and explained by known physics.

5.6 RDVS and the Latent Socio-Spatial Process Model

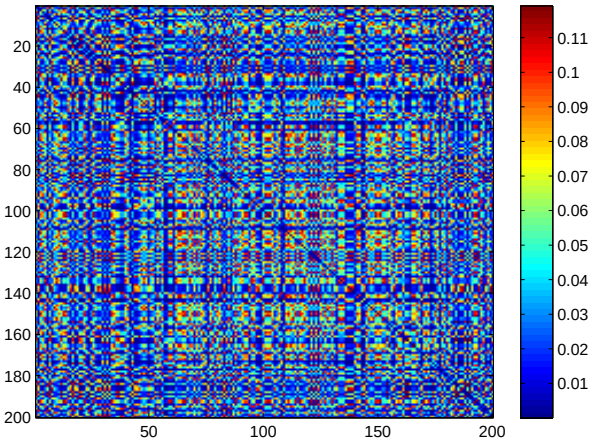
We have seen that RDVS can be useful for identifying important factors in computer experiments when a Gaussian spatial process model is used. We conclude this chapter by demonstrating how incorporating an inert covariate into an analysis can be used to identify important variables in the latent socio-spatial process (LSSP) model that we developed in Chapter 3. Thus far we have only considered using this model on examples with one or two covariates. This is partially due to the difficulties in interpreting the results in higher dimensions. Here, we consider an example with 10 covariates, of which only the first four are actually important.

To begin, using the covariate space $\mathcal{X} = [0,1]^{10}$, we generate covariate vectors corresponding to $n = 200$ actors. A random function that is active in four dimensions (corresponding to the first four covariates) is generated, and taken to be the LSSP for this example. The function cannot be visualized, but Figure 5.10(a) shows the resulting probabilities of connection for all $i, j = 1, \ldots, 200$. These probabilities were calculated using the LSSP model (3.1),

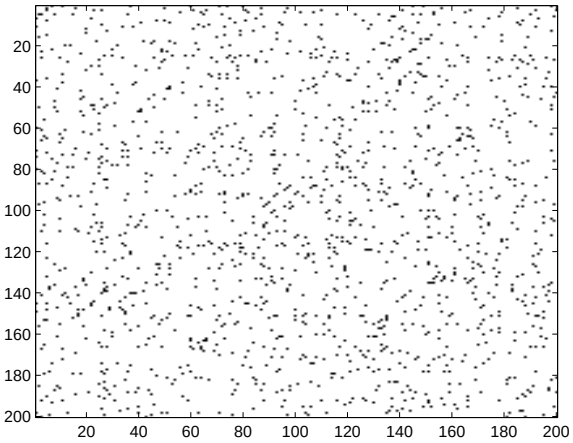
$$\eta_{i,j} = \mu - |z(\mathbf{x}_i) - z(\mathbf{x}_j)|,$$

with $\mu = -1.5$ and the "true" LSSP function. The sociomatrix we generate from these probabilities to be our observed network data is shown in Figure 5.10(b).

To implement the LSSP model we choose 100 basis centers in the 10-dimensional covariate space. Thus, there are 111 parameters to be estimated: μ, $\boldsymbol{\alpha}$, and $\boldsymbol{\rho}$. For these unknowns, we use the same priors previously specified in Chapter 3, i.e. $\mu \sim N(0, 10)$, $\rho_d \sim U[0,1]$, $d = 1, \ldots, p$ and $\alpha_r \sim N(0,1)$, $r = 1, \ldots, m$. Unlike in the designed experiment case, a variable selection prior on $\boldsymbol{\rho}$, such as the one



(a) Probabilities of connection



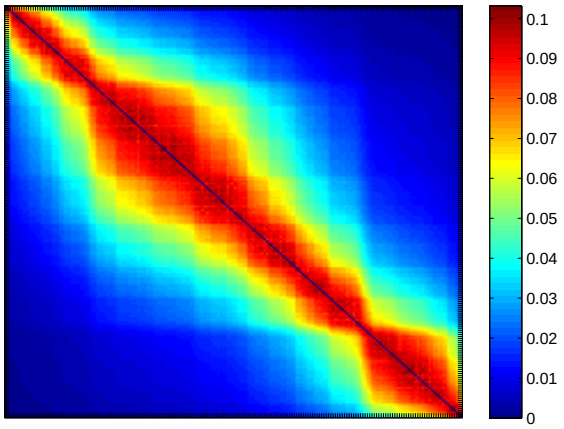
(b) Sociomatrix

Figure 5.10: Probabilities and sociomatrix for 10 covariate example.

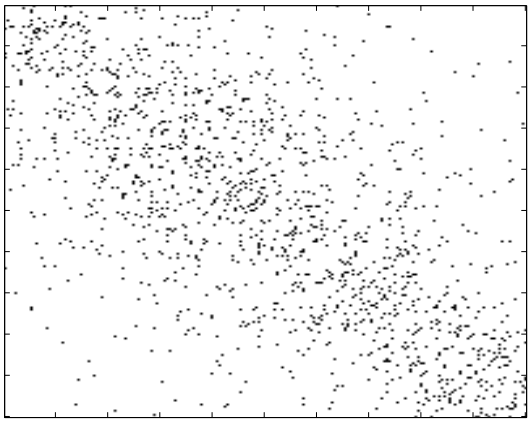
considered in (5.5), is not as intuitive here. Figure 5.11(a) is a plot of the posterior mean probabilities of connection for the sample, sorted by estimated LSSP score from lowest to highest. The sociomatrix sorted by the same permutation of actors is plotted in Figure 5.11(b). It seems from this plot that the overall pattern of probabilities was captured fairly well.

Using some more formal goodness-of-fit assessments, we see that the fit is satisfactory. Here we show the degree sequence of the true sociomatrix and the pointwise 90% bounds for a number of replicated sociomatrices in Figure 5.12(a). The replicate distribution of the number of edges is illustrated in Figure 5.12(b), with the observed value noted by a vertical line. For a more local fit, we look at the degree of each actor and the pointwise 90% replicate bounds in Figure 5.12(c). Finally, the χ^2 test we have been considering is plotted in Figure 5.12(d).

A difficulty with this analysis when there are multiple covariates is interpreting the relationship between attributes and network connections. Thus, we consider using the RDVS technique described in this chapter to see if it detects the active covariates in the example. To implement, we add an 11^{th} covariate to the attribute matrix, randomly selected from the values of the original 10 covariates. We fit the LSSP model, but use $m = 110$ basis functions, with the addition of the extra variable. Posterior draws are generated, and we save the posterior draws of $\boldsymbol{\rho}$ for the original 10 factors, and the posterior median $\tilde{\rho}_{11}$ for the inert variable. This is repeated 100 times, and the null distribution for the posterior median of ρ for an inactive variable is built up. Figure 5.13 shows the results of this analysis. The boxplots for the posterior distribution of ρ_d for $d = 1, \ldots, 10$ are shown, with the 10^{th} percentile of the distribution of the dummy variable marked by a horizontal line. We see that the first four factors are correctly identified as active, and the 6^{th} variable is incorrectly identified for this

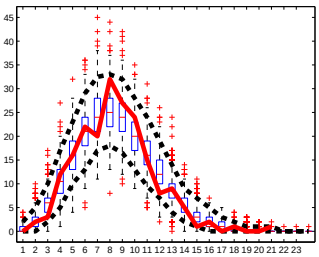


(a) Sorted posterior mean probabilities

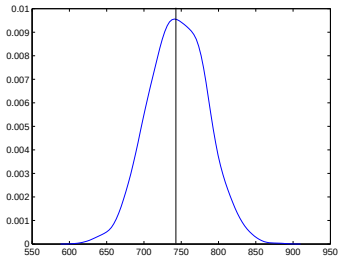


(b) Sorted sociomatrix

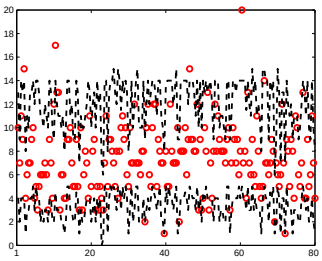
Figure 5.11: Posterior probabilities for 10 covariate example.



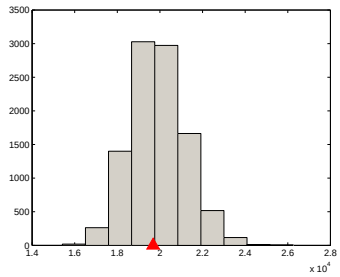
(a) Degree sequence



(b) Number of edges



(c) Degrees for actors



(d) χ^2 null distribution

Figure 5.12: Goodness-of-fit plots for 10 covariate example.

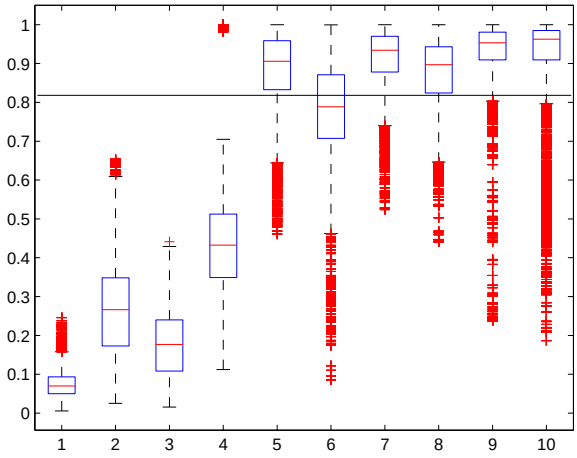


Figure 5.13: Posterior distributions of ρ_k for LSSP with 10 covariates.

particular example. To be thorough, it would be useful to do a simulation study on this to verify that the false identification rate is where it should be expected. We remark that computational time for this analysis is somewhat intensive. To run the LSSP model on the 10 covariate example (with $n = 200$ actors) requires between 2-3 hours, so conducting RDVS requires about 200 hours, at least conditional on using MATLAB and available computing. While this is feasible for analysis, it makes a simulation study something we will consider for the future.

Chapter 6

Application: Disease Transmission Modelling

To review, we have proposed a class of latent spatial process models for the analysis of social network data. Our approach is motivated by a new conceptualization of the link between individual attributes and pairwise relations. The framework we consider allows one to make inferences about likely connections between actors, given observed connections and attributes for a sample. The reference distribution variable selection technique discussed in the previous chapter – although initially developed for screening important factors in computer experiments – is one aid to interpretation of spatial process models. In this chapter, we consider an exciting potential application.

Recall that one of our primary motivations for developing network models was to gain insight into how an infectious disease might spread through a population. An infectious agent, released naturally or maliciously, has the potential to have a profound impact on society. In general, mathematical epidemiological models are

invaluable tools for allowing policy makers to experiment with different transmission scenarios and intervention strategies. Traditional compartmental transmission models (in which the population is divided up into disjoint "health" compartments) assume homogeneous mixing, i.e. an equal chance of contact between every person in the population. Anderson and May (1991) and Andersson and Britton (2000) are encompassing references on the discrete and stochastic versions of compartmental epidemiological models, respectively.

While such models can provide important information about the after-effects of many epidemics, some recent cases – such as SARS – highlight the weaknesses of the mass-action assumption. *Contact network epidemiology* (e.g. Newman, 2002; Meyers, 2007) has led to substantial refinements of the homogeneous mixing assumption. In this paradigm, transmission is assumed to be restricted to a network structure. That is, each infected individual is assumed to have a limited number of contacts. This can show improved predictions of final outbreak sizes, lengths of epidemics, threshold limits, and other global epidemic properties for contact-based infections. Despite the incorporation of more flexible contact structures, these models still treat every actor as completely exchangeable, making study of the effects of an infectious disease at a more local level infeasible.

6.1 Networks and Disease Incidence

To make this discussion more concrete, we introduce two elementary – but extensively used – disease transmission models. In particular, we aim to emphasize the impact of local network topologies on predicted individual disease incidence. In turn, we will revisit the latent socio-spatial process (LSSP) model and explore how generating

networks from the predictive distribution captures uncertainty about contact patterns in the predictions of local disease properties.

6.1.1 The SIS Transmission Model

The first transmission model we consider is a discrete-time contact dependent susceptible-infected-susceptible (SIS) model. This is in essence an individual-level compartmental model, where at any time t an actor can be in one of two disjoint categories: susceptible, meaning healthy at time t but susceptible to infection; or infective, ill at time t and capable of infecting contacts. Denote the health status of a population with N actors at time t by

$$\mathbf{H}^t = (H_1^t, \ldots, H_N^t) \in \{0,1\}^N$$

where

$$H_i^t = \begin{cases} 1 & \text{if actor } i \text{ is infected at time } t \\ 0 & \text{otherwise.} \end{cases}$$

To simulate an SIS epidemic, let ν be the probability that an infected person infects a susceptible neighbour between time t and $t+1$, and γ be the probability that an infected person at time t recovers at time $t+1$. In addition, we introduce a probability ϵ that a susceptible individual becomes infected between t and $t+1$ independently of the contact structure. We introduce ϵ partly to make exploration of the process stationary distribution possible, but we feel it is also reasonable from a practical point of view. In studying a disease on a network, even though every effort might be made to capture the whole population of interest, there is always a chance that someone in the network will be infected by someone outside. Anecdotal evidence

for this particular transmission model is the common cold. When it arrives in your circle of contacts, everyone seems to get it. Then it disappears for a while, only to make a random reappearance a few months later when it reenters your circle. It is this kind of persistent infection that this SIS model aims to capture.

The $\{\mathbf{H}^t\}$ form a Markov chain since the health status for the population at time t only depends on $\mathbf{H}^{t-1}$. The transition $H_i^{t-1} \to H_i^t$ for individual i, however, depends not only on H_i^{t-1} but all of $\mathbf{H}^{t-1}$, since the probability individual i moves from the susceptible state to the infected state depends on the health status of his/her neighbours. Having this global Markov property but dependencies at the individual level makes this SIS model an example of an *interacting particle system* (Liggett, 1985), with evolutionary equations

$$P(H_i^t = 1 | H_i^{t-1} = 0, \mathbf{H}^{t-1}) = 1 - (1-\epsilon)(1-\nu)^{\sum_{j \in \mathcal{N}_i} H_j^{t-1}} \tag{6.1}$$

$$P(H_i^t = 1 | H_i^{t-1} = 1, \mathbf{H}^{t-1}) = 1 - \gamma. \tag{6.2}$$

Under normal circumstances, $\mathbf{H} = \mathbf{0}$ would be an absorbing state for the SIS model. With our non-standard inclusion of $\epsilon > 0$, however, there exists a limiting probability

$$\lambda_i = \lim_{t \to \infty} P(H_i^t = 1).$$

We use simulations to begin looking at how λ_i depends on local network topologies. Suppose that a population of $N = 75$ individuals has the network structure shown in Figure 6.1. We randomly choose a "patient zero," i_0, and then run the SIS epidemic from this point for $T = 6,000$ time steps, with $\nu = 0.8$, $\gamma = 0.5$, and $\epsilon = 0.01$. Figure 6.2 is a plot of the cumulative estimates of λ_i,

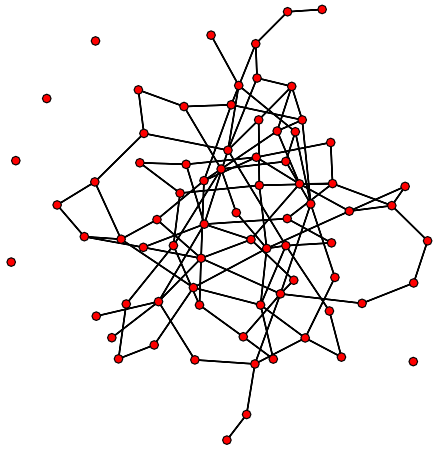


Figure 6.1: A network with $n = 75$ actors.

$$\hat{\lambda}_i(t) = \frac{1}{t}\sum_{x=1}^{t} H_i^x,$$

where each curve on the plot represents an individual actor. It can be seen from this simulation output that long-run probabilities of infection differ between actors. To be sure starting location is not having an impact, we repeat this simulation $M = 100$ times, randomly selecting i_0 and running the epidemic for $T = 6,000$ time steps at each iteration. Based on these runs, an estimate of the limiting probability (when $\epsilon > 0$)

$$\hat{\lambda}_i = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{T}\sum_{t=1}^{T} H_i^{t(m)}$$

is calculated for each individual, where $H_i^{t(m)}$ corresponds to the health status of actor i at time t in simulation m.

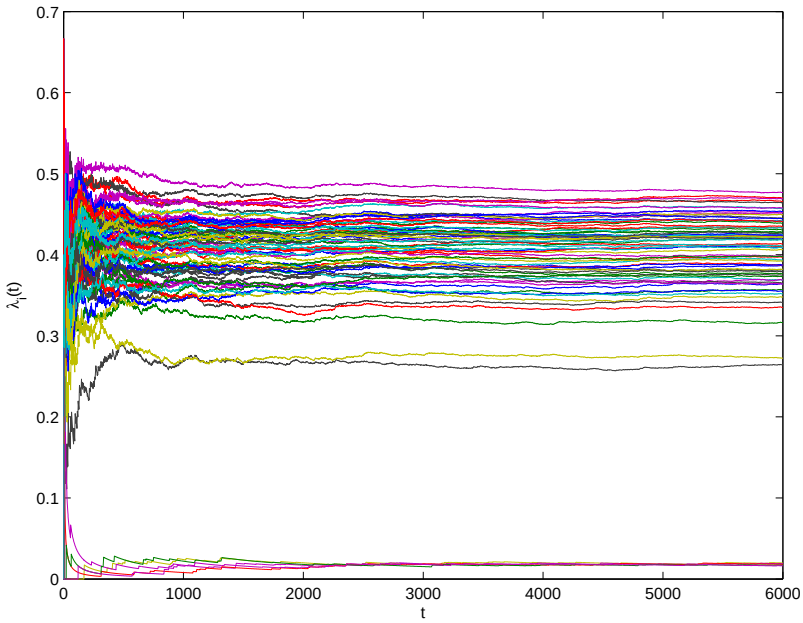


Figure 6.2: Cumulative probabilities of incidence for $n = 75$ actors

To illustrate how these probabilities relate to network topologies, we plot in Figure 6.3 the pairs $(d_i, \hat{\lambda}_i)$ for $i = 1, \ldots, N$. This reveals that if an SIS transmission is assumed, the degree of an actor, d_i, plays a significant part in his/her proneness to infection. Even for individuals with the same degree, however, there are some differences in limiting probabilities due to other local topologies, such as extended degree. Figure 6.4 is a plot of the triplets $(d_i, e_i, \hat{\lambda}_i)$, where e_i is the extended degree. Interestingly, differences between $\hat{\lambda}_i$ in this simulation are almost completely accounted for by these two properties. This can be seen by the very small differences in $\hat{\lambda}$ for individuals with the same (d, e) pair.

Some preliminary results suggest that for the SIS model (with spontaneous infection) this is likely true for any choices of ν, γ, and $\epsilon > 0$. To partially understand

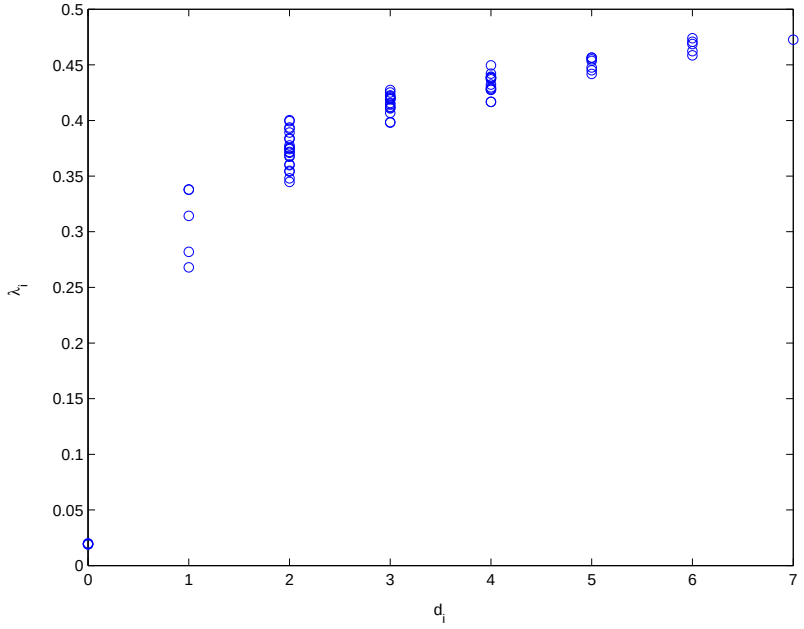


Figure 6.3: Probability of infection versus degree.

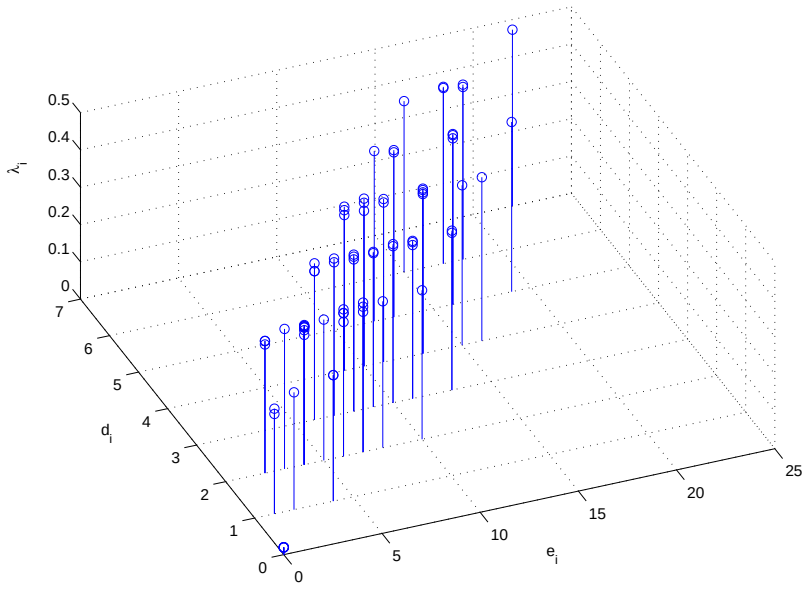


Figure 6.4: Probability of infection versus degree and extended degree.

this, we consider

$$\begin{aligned}
\lambda_i &= \lim_{t\to\infty} P(H_i^t = 1) \\
&= \lim_{t\to\infty} \left\{ P(H_i^t = 1 | H_i^{t-1} = 0, \mathbf{H}^{t-1}) P(H_i^{t-1} = 0) + P(H_i^t = 1 | H_i^{t-1} = 1) P(H_i^{t-1} = 1) \right\} \\
&= \lim_{t\to\infty} \mathbb{E}\left[1 - (1-\epsilon)(1-\nu)^{\sum_{j\in\mathcal{N}_i} H_j^{t-1}} | H_i^{t-1} = 0 \right] (1-\lambda_i) + \lambda_i(1-\gamma) \\
&= \frac{1-(1-\epsilon)E_i}{1-(1-\epsilon)E_i+\gamma},
\end{aligned}$$

where

$$E_i = \lim_{t\to\infty} \mathbb{E}\left[(1-\nu)^{\sum_{j\in\mathcal{N}_i} H_j^{t-1}} | H_i^{t-1} = 0 \right].$$

From this we see that λ_i depends on the network only through E_i. Intuitively, E_i represents the long-run probability that a susceptible individual escapes infection from his neighbours. Given the contact nature of transmission, this will most certainly depend on d_i, and similarly, how often the neighbours of i will be infected depends on e_i. These preliminary results suggest that it is important to accurately capture local topologies such as degree and extended degree in order to understand the long-run properties of even as simple a model as the SIS at a local level.

6.1.2 The SIR Transmission Model

A second, and arguably more widely used, transmission model is the susceptible-infected-recovered (SIR) model. Here, each person can be in one of three possible states at time t. The susceptible and infected states are defined as before. In this model, however, an infected person never returns to the susceptible state, but rather transitions to a recovered state. The recovered state describes actors who were infected during the course of the epidemic, but are now removed from the population (presumably through immunization or death) and can no longer be infected or infect

others. This transmission model is often a reasonable representation of some more severe infectious diseases, such as measles or smallpox.

Unlike our rendition of the SIS model – which reaches an endemic state – a disease spreading according to an SIR model will eventually cease. As a result, the typical epidemiological quantities of interest for the SIR model are usually the global properties mentioned previously, such as how long an epidemic lasts, the ultimate size of the epidemic, or the threshold limit, i.e. the critical point at which the epidemic will either take off and infect a large number of people or die off very quickly. Given that all actors are assumed exchangeable, local effects such as the probability that a particular actor (or type of actor) will be infected are not typically considered. A simulation of the SIR model on a network reveals that such local effects of the disease are again closely related to topological properties of the network.

As before, let ν be the probability that an infected individual infects a neighbour between time t and $t+1$. Thus, the evolutionary equation for the transition from the susceptible state to the infected state is the same as previously given in (6.1). For simplicity, we assume here than an infected individual at time t recovers with probability 1 at time $t+1$. This is consistent with a Reed-Frost "generation" interpretation of the SIR model (see, e.g. Andersson and Britton, 2000). Consider the graph in Figure 6.1. For the network SIR model, the starting location of the epidemic will have a significant effect on its predicted outcome. To demonstrate, we consider looking at the outbreak from two angles.

First, we begin the infection once on each of the $N = 75$ nodes. Over these 75 starts, we record the proportion of times each actor becomes infected during the course of the epidemic. Taking $\nu = 0.6$, Figure 6.5 shows the results of this simulation in blue. Here, the proportion of times each actor is infected is plotted as a function of his/her

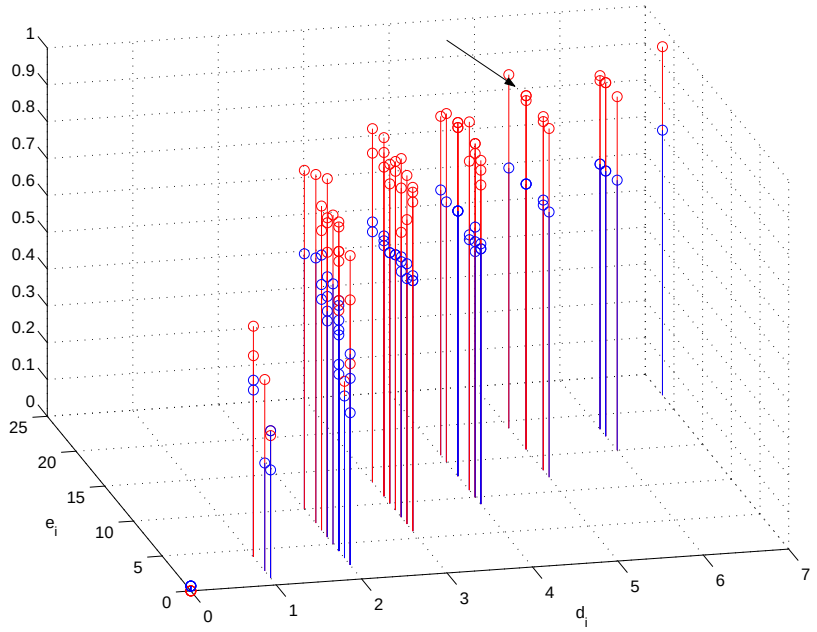


Figure 6.5: Probability of incidence by degree and extended degree. Blue: averaged over starting node; Red: conditional on a fixed starting node, identified by an arrow.

degree and extended degree. Alternatively, instead of averaging over all possible starting locations, we can condition on the epidemic starting at a particular person. An arrow points to this starting node in Figure 6.5. The probabilities marked in red in Figure 6.5 are the revised probabilities that each actor will be infected averaged over 75 simulations starting from the same point. In this simulation, given the fixed starting location, the predicted probability of infection for some individuals increases by as much as 20%. Clearly, the more information that can be incorporated about the structure of the population and the characteristics of patient zero, the better.

6.2 LSSP Model and SIR Incidence

The above explorations indicate that if the network structure of a population is known, it can provide extremely valuable information about how an infectious disease will impact each actor. This may, for example, help determine optimal intervention strategies – particularly if the characteristics of initially infected individuals are known. Unfortunately, while simulations can be conducted to learn about a disease spreading on one particular observed network, the results are not readily generalizable. For example, suppose a network is collected for one classroom, school or city, perhaps through observation, a questionnaire, or via a complex simulator such as EpiSims. What can be said about probable patterns of connection and the resulting disease incidence in other similar classrooms, schools, or cities?

Having predictive social network models suggests a new way to approach the exploration of disease transmission models. Likely network structures generated from the predictive posterior distribution (given an observed network) can be used to reflect uncertainty about contact patterns in similar populations. Advantageously, this leads to predictive posterior distributions of network topological features and disease properties. While ultimately we are interested in a variety of network and transmission models, we will restrict our attention here to the LSSP and SIR models, respectively. In particular, we conduct a small simulation study to look at the potential for using the LSSP predictive distribution to explore network topologies and predict local SIR disease incidence.

Recall that the premise behind the LSSP model is that it is possible to assign a score to each actor as a function of his/her attributes. Actors are modelled to have a

higher probability of connection if their scores are closer together,

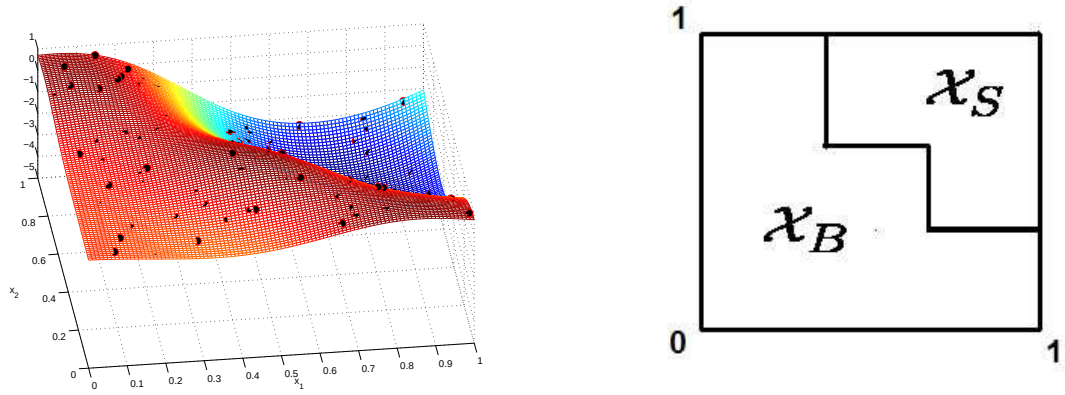
$$\eta_{i,j} = \mu - |z(\mathbf{x}_i) - z(\mathbf{x}_j)|. \tag{6.3}$$

The spatial process assumption for the scores $z(\mathbf{x})$ over the attribute space implies that actors with "similar" characteristics are likely to have "similar" scores, where similarity here is in essence governed by the correlation parameters. Therefore, the shape of the LSSP specifies what types of individuals are most likely to connect. In combination with the distribution of actors' attributes in the population, this provides information about expected network topologies.

As mentioned, in a typical contact epidemiology model, every actor is assumed to have the same degree distribution. Intuitively, if the attributes of the individuals in a population are known, the expected degree of any actor is going to be a complex combination of the kind of actors with whom he/she is most likely to connect, and how many actors with those characteristics are in the population. For example, if it is believed that homophily by age is explanatory for how people connect, then the number of relations a twenty-something is expected to form in a room filled with people his/her age versus in a room filled with teenagers is likely to be quite different. The LSSP model is a natural framework for incorporating this intuition.

To illustrate this point, consider the surface shown over $\mathcal{X} = [0, 1]^2$ in Figure 6.6(a). This could be an LSSP for a population in which actors are divided into two fairly clear groups, where the clusters are determined by a combination of two covariates. In principle, the groups could result from any number of attributes, but we use two for convenience. The sketch in Figure 6.6(b) shows roughly the partition of the covariate space that separates the groups. We let $\mathcal{X}_B$ denote the bigger region and $\mathcal{X}_S$ the smaller.

To emphasize the relationship between the LSSP and how the clusters are formed,



(a) An LSSP in $\mathbf{R}^2$

(b) Heuristic clustering of population

Figure 6.6: SIR Example: LSSP and implied clustering of population.

note that the surface in Figure 6.6(a) is flat over both $\mathcal{X}_B$ and $\mathcal{X}_S$, but changes quite drastically in height between them (by design). Under the assumptions of the LSSP model, for pairs $\{(i,j) : \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}_S\}$, $z(\mathbf{x}_i) \approx z(\mathbf{x}_j)$, so $\eta_{i,j} \approx \mu$. Similarly for two actors both in $\mathcal{X}_B$. In contrast, pairs $\{(i,j) : \mathbf{x}_i \in \mathcal{X}_S, \mathbf{x}_j \in \mathcal{X}_B\}$ are not likely to be connected, since $|z(\mathbf{x}_i) - z(\mathbf{x}_j)|$ is large, in log-odds terms. Our motivation for considering this exaggerated example is two-fold. First, it illustrates that the LSSP model can be quite flexible, representing as much or as little information between attributes and connections as is present. Second, it facilitates explanation of the interplay between the underlying trends modelled by the LSSP, the distribution of attributes in a population, and the resulting expected network topologies.

For example, we take a population of $N = 100$ actors with attributes uniformly distributed over $\mathcal{X}$. According to Figure 6.6(b), about 1/3 of the actors should be in $\mathcal{X}_S$ and the other 2/3 in $\mathcal{X}_B$ (due to the respective areas of the covariate space these regions cover). Using the LSSP in Figure 6.6(a), and taking $\mu = -1$ in (6.3), the probabilities of connection between these 100 actors can be found, and are shown in

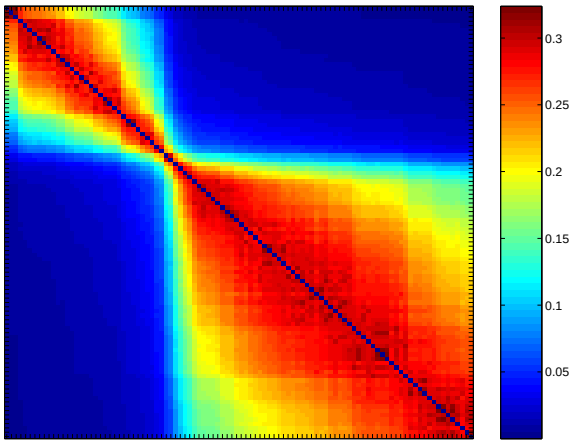


Figure 6.7: SIR Example: True sorted probabilities of connection for a uniformly distributed population of $N = 100$ actors.

Figure 6.7. In this plot, we have permuted the rows and columns according the the true LSSP scores of the actors (from lowest to highest). This clearly illustrates the two groups implied by the LSSP. We remark that in this example, clusters are formed by assuming a very specifically shaped LSSP and uniform distribution of attributes. Alternatively, a non-uniform distribution over the attribute space combined with a less extreme LSSP can represent similarly. Case in point is the adolescent health example in Chapter 3, which showed clustering of students, but in this case with a less extreme LSSP function and non-uniform distribution of attributes.

To summarize, our aim is to investigate the ability to recover what information an LSSP model contains in regards to local network topologies, and as a result, disease behaviour. From Figure 6.7, the expected degree of an actor in $\mathcal{X}_B$ is much greater than for an actor with attributes in $\mathcal{X}_S$, simply because there are more people in the first group with whom to potentially connect. Specifically, the expected degree of an actor in a specified population can be derived analytically with our assumption of conditional independence between dyads. Given probabilities $\boldsymbol{\pi}$ for the population,

the expected degree of actor i is

$$\mathbb{E}(d_i) = \sum_{j \neq i} \pi_{ij}.$$

We know from the previous section that this suggests the risks of infection for actors in the two groups is also expected to be different. A small simulation study shows that these expected differences can be captured through the posterior predictive distribution of the LSSP model.

For evaluating our simulations we need some measure of the true network topologies and disease incidence for each type of actor in our example. As mentioned, the expected degree can be found theoretically from the "true" probabilities in Figure 6.7. To get a measure of the expected extended degree, we generate 1000 networks from the true probabilities, and average each person's extended degree over these realized networks.

When looking at the expected SIR disease incidence, we consider three different scenarios. First, we average over all possible starting locations. To get the true expected incidence for each actor in this case, we generate 1000 networks from the true probabilities, and run many SIR epidemics on each one (here we use disease transmission probability $\nu = 0.1$). Specifically, we begin the epidemic ten times from each of the 100 nodes, and average the proportion of times each actor is infected. For the other two scenarios, we consider an SIR epidemic that spreads beginning from a particular person with attributes in $\mathcal{X}_B$, say i_B, and then one that begins from an actor in $\mathcal{X}_S$, i_S. To get the true incidence in each of these cases, on 1000 networks generated from the probabilities in Figure 6.7, we begin the epidemic 200 times starting at i_B and 200 times starting at i_S, averaging the results separately.

For our first simulation, we consider a somewhat optimal situation. Suppose a

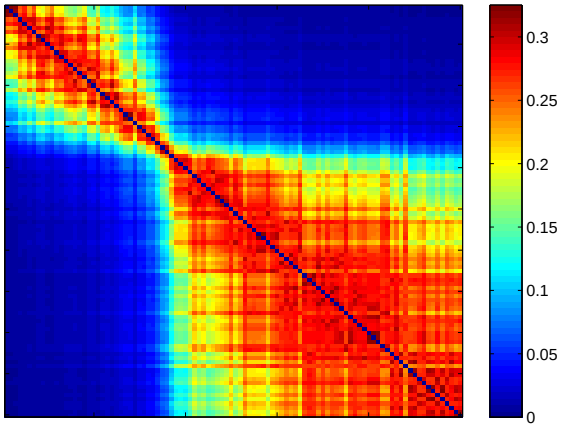


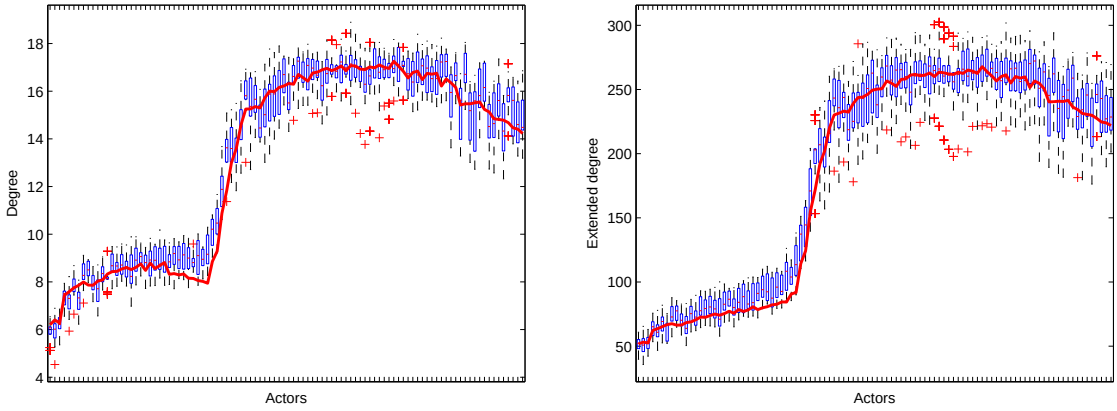
Figure 6.8: SIR Example: Sorted simulation probabilities of connection.

network is observed between $N = 100$ actors. For our purposes this network is generated as a single realization from the true connection probabilities in Figure 6.7. The LSSP model can be fit to this network, and posterior mean estimates of the probabilities of connection between any $N = 100$ actors with these same attributes can be found (as described in Chapter 3). From these fitted probabilities, we generate $R = 105$ replicate sociomatrices. Using these replicates, we calculate the expected degree and extended degree of each actor. In addition, we run SIR epidemics on each replicate (starting once on each node again with transmission probability $\nu = 0.1$), recording the proportion of times each actor is infected. Recall that when we were considering goodness-of-fit, we compared properties of the replicate sociomatrices to the particular network that was fit. Here we wish to assess performance against the "truth," where the truth is found as described above. To make our assessment, we repeat this entire process of observing, fitting, and predicting for $M = 100$ networks generated from the true connection probabilities.

In Figure 6.8, we show the average posterior mean connection probabilities for

the 100 actors, averaged over the R replicates for the M different fitted networks. Comparing Figures 6.7 and 6.8, we see that overall, given any observed network, we recover on average the connection probabilities that underly the observed networks quite well. The same sorting is applied in this plot as was previously done. In addition, we explore the expected degree and extended degree for each actor. For a particular observed network, we use the R replicates from the fitted model to get the posterior predictive estimate of the degree and extended degree for each actor. This is repeated for each of the M networks, and we plot the corresponding boxplots of the simulated expected degree and extended degrees in Figures 6.9(a) and 6.9(b), respectively. Again, we use the same sorting of actors to highlight the difference between the two groups. The red lines in these plots show the true expected degree and extended degree of each actor. We see that on average, over many different realizations, the expected degree and extended degree calculated for each actor from the fitted model tends to be a good estimate of the truth. In particular, the identification of the two groups is quite clear.

Finally, we look at what this suggests about potential SIR disease transmissions. There are multiple ways to look at performance, but as mentioned, we will focus on three scenarios in particular. First, we consider incidence for each actor averaged over starting location. For each of the R replicate networks generated from a fit of the LSSP model to an observed network, we run the SIR epidemic (starting on each node once) and save the proportion of times each actor is infected. Figure 6.10 shows the boxplots of these estimated probabilities over the $M = 100$ networks generated as "observed" networks. As expected, the predicted incidence for actors in each of the two groups is quite different. The red line shows the expected probability each actor is infected according to the true underlying model that we found previously.



(a) Expected degree

(b) Expected extended degree

Figure 6.9: SIR Example: Simulated distributions of network topologies.

Though there seems to be some over-estimation in this limited simulation, the general structure is readily apparent. We remark that there are two explanations for the over-estimation that seem plausible. One is the small size of the simulation. Many runs of the SIR model were performed to get the true expectation, and it is possible that the limited number of starts used in the simulation component reduce accuracy slightly. A second possibility is the challenging nature of the particular example function we use (Figure 6.6(a)). modelling the slope of this function is difficult – particularly when only a few binary responses are observed in the region of the change. To relate this to the over-estimation seen, note that if the slope is estimated to be more gradual than the truth, this will cause more mixing of the two groups than expected. This then increases the probability that the infection will spread from one group to the other, and thus the probability that any particular person will be infected. Further investigation will be needed to verify the cause.

Next, we consider the cases where we condition on the epidemic starting at particular locations. To do so, we look at the proportion of times each actor is infected

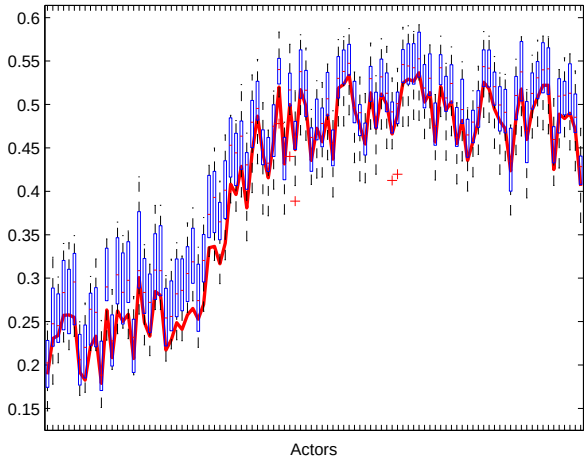
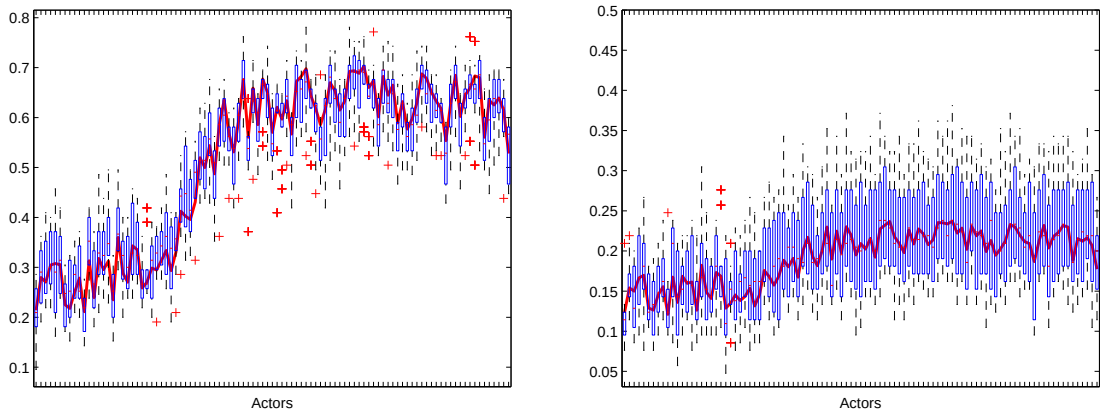


Figure 6.10: SIR Example: Simulated distribution of SIR on LSSP network

in our simulation conditional on the spread starting from i_B and i_S. Recall that i_B and i_S are chosen actors in the big and small groups, respectively. From the results of the previous section, the predicted probabilities of infection should depend strongly on which group the infection began in. In Figure 6.11(a), we show the simulated distribution of predicted disease incidence conditional on the epidemic beginning with a person in $\mathcal{X}_B$, as well as the true predicted path under this scenario in red. Similarly, in Figure 6.11(b), we consider the same, but starting from the person in $\mathcal{X}_S$. It is promising to see that over all fitted networks, the difference in disease incidence as a result of starting location is detected.

Admittedly, we are being somewhat generous in this simulation study, since we are only making predictions for the same number of actors (with the same attributes) as in the observed network. A key advantage of the LSSP model is that predictions can be made for actors outside the observed sample. To begin exploring the performance of the methodology in this context, we repeat the above simulation study, but assume the network is only observed for $n = 50$ actors. Specifically, for a network with



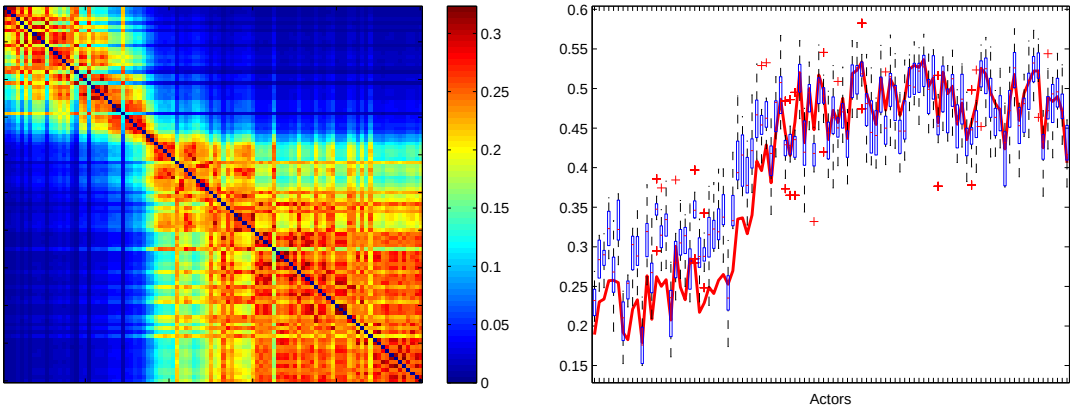
(a) Beginning on an actor in $\mathcal{X}_B$

(b) Beginning on an actor in $\mathcal{X}_S$

Figure 6.11: SIR Example: Simulation of SIR conditional on starting location.

$N = 100$ actors (generated from the truth, Figure 6.7, we take a random subset of $n = 50$ actors as the actors in the observed network. We repeat this for $T = 20$ randomly chosen training sets of size $n = 50$. Using only the training data, we predict the probabilities of connection for the entire $N = 100$ actors. In Figure 6.12(a), we show the average posterior mean connection probabilities for the 100 actors, averaged over the R replicates for the T training samples from M different fitted networks. The clusters are still detected quite well, though with more error than before. This is also reflected in the predictions of the SIR disease incidence for all 100 actors (averaged over all starting locations), shown in Figure 6.12(b).

These simulations suggest that it may be possible to detect differences in disease risks in a population of interest by first fitting a network model to an observed sample. In particular, using the posterior predictive connection probabilities to obtain a predictive distribution on network topologies and disease incidence is a natural framework for incorporating uncertainty about the population network structure. If it is believed that different types of people connect with different rates, this may impact



(a) Predicted posterior mean probabilities

(b) SIR on predicted network

Figure 6.12: SIR Example: Predicted contact probabilities and SIR incidence based on training samples.

intervention strategies – especially if it is known in what region of the attribute space early cases are detected. We note that this illustrates the advantage of considering network models that aim to capture trends, rather than over-fitting an observed network. If the fitted network model simply reproduces the observed network, then the ability to look at uncertainty over different possible networks is diminished. Certainly much more investigation is required here. Our simulation is very small, in terms of population size, training sizes, and sample sizes. We have also restricted our attention to only one particular example that was constructed to create a strong distinction between two clusters, albeit one that is not necessarily "easy" to fit. This is a research direction we will continue to develop in the future.

Chapter 7

Conclusions and Future Research

In this thesis we have considered a new class of models for analyzing social network data. This kind of data is challenging to model due to its pairwise and dependent nature. The class of models we proposed, latent spatial process models, use variations on the assumption of homophily by attributes to link pairs of covariates to the probability of connection. The first model we developed, the latent socio-spatial process (LSSP) model, assigns a score to each actor and then compares scores to determine the log-odds of connection. This approach has an intuitive interpretation of transitivity and clustering. Alternatively, the meta-distance (MD) model directly models the log-odds of connection for each pair as a function of the component-wise dissimilarities in each attribute. This model has the advantage that it requires no prior assumptions about how differences in attributes impact the likeliness of connection, i.e. whether it increases or decreases.

Using a radial basis function representation (similar to a process-convolution construction) for the latent surface in both models makes implementation feasible. Particularly important, by choosing the centers of the basis functions according to a Latin

hypercube design (LHD), we are able to estimate both models using only $m + p + 1$ parameters, where m is the chosen number of basis functions in the LHD and p is the dimension of the measured attribute vector. This can be a huge reduction from the $O(n)$ or $O(n^2)$ parameters that are typically used to model social network data.

While interpretation of these models can become difficult in higher covariate dimensions, the reference distribution variable selection technique proposed in Chapter 5 can aid in the detection of significant effects. Though originally developed for screening important factors in computer experiments, we find that the same philosophical approach appears promising for identifying active variables in the LSSP model.

One of the main motivations we had for building these network models was to find a way to easily predict unobserved relations. This plays an important role, for example, in understanding how an infectious disease might spread through a closed, structured population. We saw in Chapter 6 that if the LSSP model reasonably describes the structure in an observed sample, it can be used to predict likely structures for a larger population. A small simulation showed these predicted networks have similar topologies to networks generated according to an assumed true underlying trend. This shows promise for more accurately predicting how an infectious disease might spread in a population.

We close with a few remarks and an outline of some possible future research directions. One thing we note about using latent spatial process models as opposed to latent factor models, say, is that predicted probabilities of connection for each pair are very small. This is a by-product of the change in inferential perspective. From a population-inference perspective, the probability that any actor is connected to any other is very small (otherwise we would all know each other). Latent factor models, on the other hand, more accurately describe potentials of connection for a *particular*

set of actors. If this kind of description is desired, there is no reason why the latent spatial process and latent factor approaches could not be combined. Then, estimates of error found via the latent factor model can be used for descriptive purposes, and the latent spatial process component can be used for improved predictive ability.

This raises the issue of how to assess fit of network models. Models that maximize the likelihood by proposing fits close to 0 or 1 for unlinked pairs versus linked pairs, respectively, will fit the observed data very well. Indeed, by looking at even a simple χ^2 test, we see that sometimes latent factor models can fit too well. Bayesian cross-validation procedures seem promising for assessing predictive ability, but more needs to be done on choosing useful measures for comparison.

In our attempts at using cross-validation, we find that outliers in networks are a stumbling block to obtaining good fits. For network data, we consider outliers to be individuals that are highly connected or completely isolated. Completely removing isolates does not seem to us to be the best option, since they do actually contain information on the likeliness of connection, i.e. an individual with all observed zero connection is notably different than an individual with all missing observed connections. For the case of binary relations, it may be worth considering mixture distributions for the pairwise dyads. For example, we might assume that there are structural appearances of 0's and 1's as well as random realizations. In the spirit of modelling zero-inflated data, we could try to implement a 0-1 inflated model,

$$y_{i,j} = \begin{cases} 0 & \text{with probability } \lambda^0_{i,j} \\ BER(\pi_{i,j}) & \text{wp } 1 - \lambda^0_{i,j} - \lambda^1_{i,j} \\ 1 & \text{with probability } \lambda^1_{i,j}. \end{cases}$$

Here, $\lambda^0_{i,j}$ and $\lambda^1_{i,j}$ are the probabilities that a dyad will always be 0 or 1 for a particular pair. These might be assumed constant across pairs, or simultaneously modelled as

a function of pairwise covariates. A Bayesian approach to estimating finite mixture distributions is given in Diebolt and Robert (1994), for example.

We conclude by highlighting some more specific avenues for future research.

Handling Categorical Covariates

An obviously strong assumption we make in the development of our latent spatial process models is that the covariate space $\mathcal{X} \subseteq \mathbb{R}^p$. As we saw in the Adolescent Health and International Conflict examples, ordinal categorical variables (such as grade and polity score) can be directly included provided they have enough levels. Certainly handling non-ordinal categorical covariates is an important consideration – particularly for sociological network data. Incorporating categorical variables into spatial process models is a challenging task in general, and only a few approaches have been considered to date (e.g. McMillan *et al.*, 1999; Qian *et al.*, 2006).

There are a few potential directions we have been considering. Perhaps the most obvious choice is to take a block-model approach (e.g. Wang and Wong, 1987). Suppose, for example, that individuals are partitioned into blocks $B_1, \ldots, B_g$ by a number of categorical attributes. Following Wang and Wong (1987), a parameter $\lambda_{a,b}$ can be assigned to each $B_a \times B_b$ block. Let

$$\delta_{i,j,a,b} = \begin{cases} 1 & \text{if the pair } (i,j) \text{ is in } B_a \times B_b \\ 0 & \text{otherwise.} \end{cases}$$

Then, a possible extension of the LSSP model, say, is

$$\eta_{i,j} = \mu + \sum_{a,b} \lambda_{a,b}\delta_{i,j,a,b} - |z(\mathbf{x}_i) - z(\mathbf{x}_j)|,$$

where $\sum_a \lambda_{a,b} = 0$ and $\sum_b \lambda_{a,b} = 0$. This might be an improvement on the LSSP assumption of a constant mean, for example. It does not address any interactions

between the categorical and continuous covariates, however.

An alternative might be to take an optimal scoring approach. This is used to incorporate mixtures of continuous and categorical variables in discriminant analysis, for example (Huberty *et al.*, 1986; Tuv and Runger, 2004). Though we have not implemented the idea, it seems intuitive for network data. Generally speaking, one finds clusters of connections in the network data (using, for example, methods of Girvan and Newman, 2002). Let G denote a vector which indicates which group each actor belongs to as a result of this clustering. Let C contain the value of a categorical covariate for each actor, with say l levels.

The information contained in C can be represented by l indicator variables, $\delta_1, \ldots, \delta_l$. This is, for example, the trick used in linear regression when categorical variables are represented by dummy indicator variables. Replacing the levels of C with continuous scores $x_1, \ldots, x_l$ is equivalent to replacing $\delta_1, \ldots, \delta_l$ by a linear combination $x_1\mathbf{V}_1 + \ldots + x_l\mathbf{V}_l$. In particular, one chooses the combination $\mathbf{Vx}$ that maximizes the ability to discriminate between the groups G. More details on this approach can be found in Tuv and Runger (2004). Provided there are enough categories, l, for each categorical variable, the optimal scores can be assigned to each actor and used directly in the spatial process model for network analysis.

Directed Networks

Throughout this thesis we have focused on developing models for undirected links in networks. There are applications, however, where interest is in directed relations. In the case of a directed network, $y_{i,j}$ is observed to be 1 if a tie is sent from person i to person j, and $y_{j,i}$ does not have to be the same as $y_{i,j}$. As mentioned in Chapter 1, this can arise when person A claims person B as a contact, but not vice versa. Such

situations are likely to be expected if network data is collected via a questionnaire, for example. Alternatively, in the International Conflicts example we have considered, information is available on which country is the aggressor, so the directed network can be considered.

In addition to capturing tendencies such as homophily by attributes and clustering, directed models also look to quantify *reciprocity*, the tendency for links being sent to be returned. Directed versions of the latent factor models discussed in Chapter 2 have been developed. For example, in the latent space model, the social positions can be compared using $\mathbf{z}_i'\mathbf{z}_j$ (Hoff, 2005) instead of using the Euclidean distance. Similarly, the multiplicative factor model has an alternative singular value decomposition interpretation for handling directed connections.

In future work, we would like to extend the LSSP and MD models to handle directed networks. There are a couple of approaches we have considered. For the LSSP model, instead of assuming LSSP scores are realizations from one spatial process, we could introduce two correlated processes, a sender process, $z_s(\mathbf{x})$, and a receiver process. $z_r(\mathbf{x})$. These could be modelled, for example, as

$$z_s(\mathbf{x}) = \sum_{g=1}^{m} \alpha_g k(\mathbf{x} - \mathbf{w}_g; \boldsymbol{\rho}_s)$$

and

$$z_r(\mathbf{x}) = \sum_{g=1}^{m} \alpha_g k(\mathbf{x} - \mathbf{w}_g; \boldsymbol{\rho}_r).$$

Here, $\boldsymbol{\rho}_s$ and $\boldsymbol{\rho}_r$ are the correlation parameters that govern the sender and receiver functions, respectively. As before, they govern how much LSSP sender and receiver scores differ between actors as a function of the attributes. In addition, correlation is induced between the two processes via the use of the same grid $\mathcal{W}$ and weights $\boldsymbol{\alpha}$ (see, e.g. Higdon, 2002). This captures how similar an actor's LSSP sender score is

to his/her LSSP receiver score.

Using these two processes, the LSSP model can be reformulated as

$$\eta_{i,j} = \mu - |z_s(\mathbf{x}_i) - z_r(\mathbf{x}_j)|.$$

Here, the log-odds that actor i will send a tie to actor j, $\eta_{i,j}$, is modelled by comparing actor i's LSSP sender score to actors j's LSSP receiver score. Under this model, the probability for a tie being sent is highest if the scores are similar. Reciprocity is captured by how much correlation exists between the two processes. For example, if $z_s(\mathbf{x}\} \approx z_r(\mathbf{x})$, i.e. there is very high correlation between the two processes, then $\eta_{i,j} \approx \eta_{j,i}$ for all i and j.

As a second approach, instead of working within the LSSP framework (assigning scores and comparing them), we could think along the lines of the MD model which models $\eta_{i,j}$ directly. In the undirected case, we were restricted as to how attributes could be incorporated because we wanted to force symmetry. This led specifically to the MD model, which models differences between dissimilarities. When symmetry is not required, this yields more model choices. One attractive possibility is an additive model, i.e.

$$\eta_{i,j} = \mu + z_s(\mathbf{x}_i) + z_r(\mathbf{x}_j).$$

Here, z_s and z_r are correlated processes as specified above, where the correlation between them captures reciprocity. The function evaluations $z_s(\mathbf{x}_i)$ and $z_r(\mathbf{x}_i)$, however, are now interpreted more as random effects that capture actor i's contribution to the log-odds of connection as a sender or a receiver, respectively. What separates this from other random effect models for network data (such as Hoff, 2005, e.g.) is that the effects are assumed to be correlated according to actor attributes. This greatly reduces the number of required parameters, since only the $m + 2p$ parameters for the

processes z_s and z_r need to be estimated, rather than $2n$ individual-specific sender and receiver random effects.

Designed Radial Basis Functions

Our use of designed radial basis functions to represent a prior class of functions for the latent spatial process models raises some interesting questions about the general applicability of the method. This idea seems natural for predicting a latent surface, but it may be useful as a similar parameter reduction technique when fitting observed functions.

A typical approach to estimating radial basis functions is to center the kernels at the n data points. Since this can result in computational difficulties for large problems, a great deal of effort has been made on the selection of m basis kernels, where $m < n$. The number of basis functions that one should choose is going to be related to the perceived complexity of the function to be approximated. We suspect that this is similar to the issue of selecting input runs for the approximation of complex computer simulators. Heuristically, there may be an argument that a radial basis function centered at all data points can be well approximated by a designed radial basis function, with many less kernels. It would be interesting to see if there is any theoretical argument for whether or not a designed basis function approximates well a radial basis function with more kernels, and the rate at which the bias and error decrease.

Given the number of basis kernels, there has been much discussion about how best to place these points. For example, Yousef and el Hindi (2005) summarize a number of methods that have been tried. These include randomly selecting a subset of training data or clustering inputs and centering kernels at cluster centers. Genetic

algorithms and other optimizers have also been considered for choosing locations. Here we conduct a small simulation study to consider the performance of our idea to use a LHD to choose the centers.

We consider estimating a surface on $\mathcal{X} = [0,1]^2$ using

$$z(\mathbf{x}) = \sum_{r=1}^{m} \alpha_r k_\rho(\mathbf{x} - \mathbf{w}_r), \tag{7.1}$$

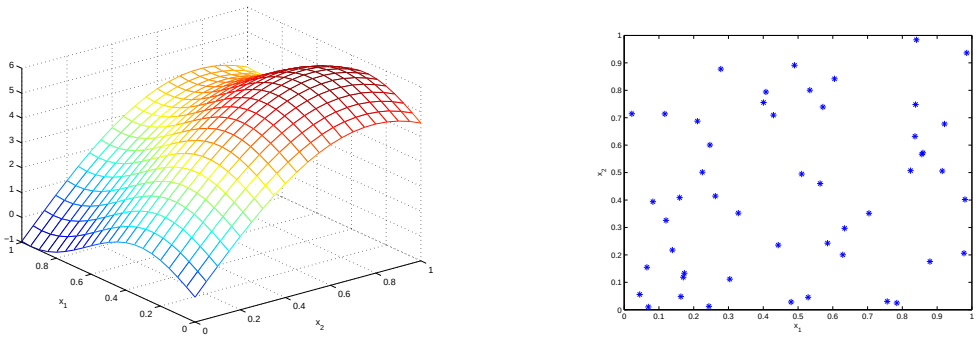
with six different choices of center points $\mathcal{W}$. Our goal is to compare prediction mean square error and estimates of the kernel widths.

The test function we use is

$$z(\mathbf{x}) = sin(5x_1) + 5sin(2x_2), \tag{7.2}$$

which is plotted over $[0,1]^2$ in Figure 7.1(a). We assume that the function is observed (with a small amount of error) at $n = 50$ points, $\mathbf{X}$. These points are selected at random, and we keep them fixed for all simulations. Figure 7.1(b) shows their locations. As mentioned, we consider six different choices for $\mathcal{W}$, each with a different number of points, m. The cases and a pointer to their illustrations are as follows.

1. A LHD with $m = 10$, Figure 7.2(a)
2. A LHD with $m = 20$, Figure 7.2(b)
3. A LHD with $m = 40$, Figure 7.2(c)
4. A LHD with $m = 50$, Figure 7.2(d)
5. The data points $\mathbf{X}$, $m = 50$, Figure 7.1(b)
6. An evenly spaced 11×11 grid, $m = 121$, Figure 7.2(e)



(a) True function to be approximated

(b) Data points, $n = 50$

Figure 7.1: A function and data points

One iteration of the simulation study proceeds as follows. Let

$$f(\mathbf{x}_i) = z(\mathbf{x}_i) + e_i$$

denote the observed response at point $\mathbf{x}_i$, $i = 1, \ldots, n$, where $e_i \sim N(0, \sigma^2)$ with $\sigma = 0.05$ known. Let $\mathbf{f} = f(\mathbf{X}) = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))'$. We standardize the response to have mean 0 and standard deviation 1 to simplify prior selection. We use (7.1) as a prior class of function (corresponding to the choice of $\mathcal{W}$). With the response standardized, the prior

$$\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{I}_m)$$

is taken for the weights. Non-informative $U[0, 1]$ priors are used for each component of $\boldsymbol{\rho}$. For convenience, let $\mathbf{K}$ denote the $n \times m$ matrix with elements

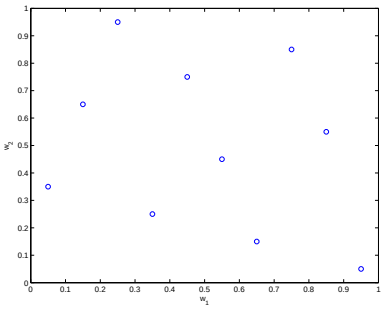
$$K_{i,j} = k_\rho(\mathbf{x}_i - \mathbf{w}_j).$$

The full posterior for this example is then

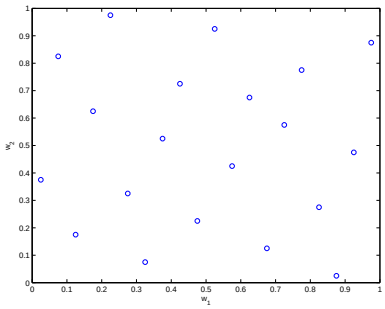
$$[\boldsymbol{\alpha}, \boldsymbol{\rho}] \propto [\mathbf{f}|\boldsymbol{\alpha}, \boldsymbol{\rho}][\boldsymbol{\alpha}][\boldsymbol{\rho}],$$

where

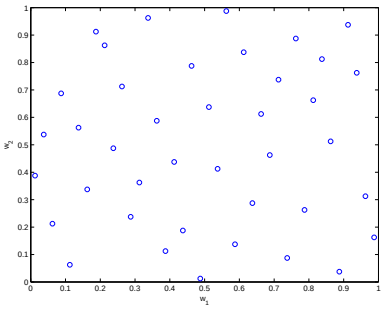
$$\mathbf{f}|\boldsymbol{\alpha}\boldsymbol{\rho} \sim N(\mathbf{K}\boldsymbol{\alpha}', \sigma^2)$$



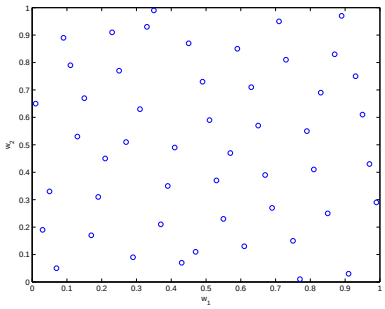
(a) Case 1



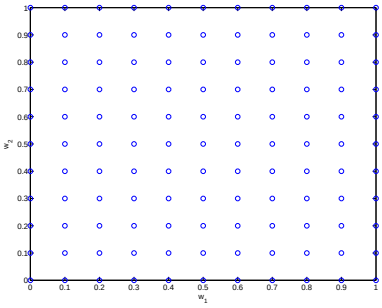
(b) Case 2



(c) Case 3



(d) Case 4



(e) Case 6

Figure 7.2: Different choices of kernel basis centers

is the sampling distribution for the observed function evaluations.

Given draws $(\boldsymbol{\alpha}^{(t)}, \ldots, \boldsymbol{\rho}^{(t)})$, $t = 1, \ldots, T$ from the posterior distribution, the function can be predicted at any point $\mathbf{x}_0$ by

$$\hat{z}(\mathbf{x}_0) = \frac{1}{T} \sum_{t=1}^{T} \sum_{r=1}^{m} \alpha_r^{(t)} k_{\rho^{(t)}}(\mathbf{x}_0 - \mathbf{w}_r).$$

Using this result, we make predictions for a find grid $\mathcal{G}$ over $[0,1]^2$ with $G = 101 \times 101$ points and calculate the predicted mean square error (PMSE) as

$$PMSE = \frac{1}{G} \sum_{g \in \mathcal{G}} [z(\mathbf{x}_g) - \hat{z}(\mathbf{x}_g)]^2,$$

where $z(\mathbf{x}_g)$ is the true value of the function from (7.2) at the location $\mathbf{x}_g$.

For each choice of $\mathcal{W}$, we repeat this process for 30 iterations. Figure 7.3 shows the average PMSE (over the 30 simulated responses) for each of the six cases. Unsurprisingly, using only $m = 10$ centers results in the highest PMSE. Increasing the number of centers to $m = 20$ results in a substantial gain in efficiency. The gains are much slower after this. Table 7.1 summarizes the PMSE and relative efficiency of each choice of centers, compared to using basis kernels centered at each data point. Finding a way to account for the number of basis kernels in the comparison of the performance would be useful in the future.

We also look at the impact of the choice of $\mathcal{W}$ on estimates of $\boldsymbol{\rho}$. There is expected to be an interplay between the distance between centers and the estimate of the kernel width. Figure 7.4(a) shows boxplots of 30 posterior median estimates of ρ_1 for each of the 5 cases. Figure 7.4(b) does the same for ρ_2. Interestingly, there is not much difference between estimates of ρ_1 over the last four cases. There is much more variability in the estimates of ρ_2. It is somewhat surprising that this variability does not decrease with an increase in the number of centers. Further exploration of this point is left for future study.

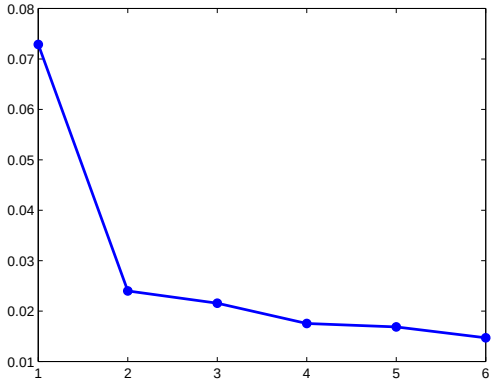
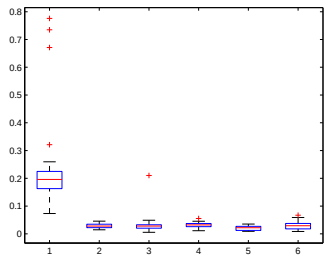


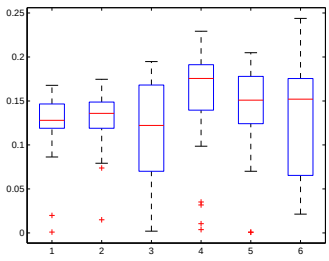
Figure 7.3: PMSE for six choices of grid

Table 7.1: PMSE and relative efficiency (compared to using data points as kernel centers) for 5 different choices of basis kernel centers.

Case	m	PMSE	Rel. Eff.
1	10	.0729	4.31
2	20	.0240	1.42
3	40	.0216	1.27
4	50	.0175	1.04
5	50	.0169	1.00
6	121	.0147	0.87



(a) Posterior medians for ρ_1



(b) Posterior medians for ρ_2

Figure 7.4: Posterior median estimates of $\boldsymbol{\rho}$ for 5 choices of grids.

These are some preliminary results on the impact of using an LHD for choosing basis centers. There is some indication that if prediction is the goal, using an LHD might be a reasonable first approximation. Certainly the computation saving would be enough (particularly in higher dimensions) to make this worth further exploration.

We remark that there may be other choices of design that are more optimal. For example, it may be more desirable to use something like a nearly-orthogonal LHD that has good projection properties onto subsets of two, three or more variables. Again, this choice will be related to how complicated the function is believed to be, and perhaps which directions have the most activity. Finally, Kern (2000) notes the importance of choosing basis centers outside the covariate space to reduce boundary effects on predictions near the edge of the covariate space. This is not an issue we have explored yet, but it is worth considering in the future.

Extensions to Other Data Structures

It will also be important to eventually consider other forms of network data. For example, in some cases non-binary responses may be measured – such as measurements made on the strength of a relationship, i.e. an ordinal scale of responses. Poisson distributed observations can arise when counts of events are made for each pair, such as the number of positive international relations between countries in Central Asia considered in Hoff (2005). The LSSP and MD models should be readily generalizable to non-binary (univariate) data in the spirit of a generalized linear model. To illustrate, if $y_{i,j}$ are counts, one might consider models of the general form

$$\eta_{i,j} = log(\lambda_{i,j}) = \mu + z(\mathbf{x}_i, \mathbf{x}_j),$$

with the sampling distribution of the responses being

$$[\mathbf{y}|\lambda] = \prod_{i<j} \frac{e^{-\lambda_{i,j}} \lambda_{i,j}^{y_{i,j}}}{y_{i,j}!}.$$

Being able to generalize to other response distributions is an advantage of the conditional independence assumption used in our models; exponential random graph models, for example, are only applicable to binary data.

Other more complex situations may also need to be considered. For instance, there may be more than one sociomatrix observed for a sample of actors (Fienberg *et al.*, 1985). This can arise if participants are asked multiple questions in a survey about different ways they relate. Developing models for networks that change over time, or networks that have a bipartite structure, are also possible future directions.

Estimation of Disease Transmission Parameters

In Chapter 6, we considered using social network models to help predict disease incidence for epidemics yet to happen. The other side of this coin is analyzing data collected on epidemics that have already occurred. Due to the dependent nature of infection data, inference on transmission rates and other disease properties is typically done by assuming a transmission model, such as the SIS or SIR model (see, e.g. Becker and Britton, 1999; Andersson and Britton, 2000). We expect, therefore, that many of our observations on the relationship between networks and transmission models will have counterpoints from an inferential perspective.

There are at least two aspects of epidemic inference that we would like to explore in the future. For one, we would like to investigate possible bias in the estimation of transmission parameters when the contact structure of the population is ignored or assumed to be homogeneous. Incorporating information from a partially observed

network may help reduce confounding between estimates of contact rates and transmission rates.

This leads to our second interest, which is in building a model that incorporates infection and network data simultaneously. We feel it will be beneficial to model the two components in unison, since both network data and infection data provide insight into the other. One reasonable approach may be to build a hierarchical model, with the transmission model being conditionally dependent on the network. For example, let $\mathbf{I}$ and $\mathbf{R}$ be data collected on infection and removal times, respectively. According to an assumed contact transmission model, these data have a likelihood $[\mathbf{I}, \mathbf{R}|\boldsymbol{\theta}, \mathbf{Y}]$, where $\boldsymbol{\theta}$ contains model parameters, such as transmission and recovery rates, and $\mathbf{Y}$ is the underlying network structure. A network model – such as the ones we have considered in this thesis – could be used as the prior for the network structure, $[\mathbf{Y}|\boldsymbol{\psi}]$. We use $\boldsymbol{\psi}$ to generically denote unknowns in the network model.

In principle, inference could be made about the network structure given the infection data,

$$[\mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{I}, \mathbf{R}] \propto [\mathbf{I}, \mathbf{R}|\boldsymbol{\theta}, \mathbf{Y}][\mathbf{Y}|\boldsymbol{\psi}].$$

This is an interesting point to consider from a sociological point of view, i.e. can one learn about social connections from observing a known process taking place on the network? Alternatively, with respect to estimation of transmission parameters, it would be interesting to explore if having partial network data can help improve inference. Britton and O'Neill (2002), for example, consider a hierarchical model similar to the one given above, but with the assumption of equal rates of contact between individuals. Some of their results, however, may suggest a means to estimate the more detailed model we propose.

Final Remarks

Overall, which assumptions and models will best describe a particular network is going to vary in each case. Fortunately, there are so many possible kinds of networks, it is hopeful that any reasonable model will be appropriate in at least some cases. To the best of our knowledge, there has been a limited collection of network data and attributes that can be analyzed with models like the LSSP or MD models. One of our hopes for this thesis is that it will encourage new directions in network sampling and modelling. For example, if a spatial model is to be used for analysis, ideally sample networks will be collected to "cover" the attribute space in a sufficient manner. We acknowledge that this is in no way a trivial matter, and we hope progress will continue to be made in the collection and analysis of network data.

Bibliography

[1] Alqallaf, F. and Gustafson, P. (2001). On cross-validation of Bayesian models. *The Canadian Journal of Statistics,* **29**, 333-340.

[2] Anderson, C., Wasserman, S. and Crouch, B. (1999). A p^* primer: logit models for social networks. *Social Networks*, **21**, 37-66.

[3] Anderson, R.M. and May, R.M. (1991). *Infectious diseases of humans; dynamics and control.* Oxford: Oxford Press.

[4] Andersson, H. and Britton, T. (1998). Heterogeneity in epidemic models and its effect on the spread of infection. *Journal of Applied Probability*, **35**, 662-670.

[5] Andersson, H. and Britton, T. (2000). *Stochastic epidemic models and their statistical analysis.* New York: Springer.

[6] Bailey, N.T.J. (1975). *The mathematical theory of infectious diseases and its applications.* London: Wiley.

[7] Ball, F. (1985). Deterministic and stochastic epidemics with several kinds of susceptibles. *Advanced Applied Probability*, **17**, 1-22.

[8] Barry, R.P. and Ver Hoef, J.M. (1996). Blackbox kriging: Spatial prediction without specifying variogram models. *Journal of Agricultural, Biological, and Environmental Statistics,* **1**, 297-322.

[9] Becker, N.G. and Britton, T. (1999). Statistical studies of infectious disease incidence. *Journal of the Royal Statistical Society B*, **61**, 287-307.

[10] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B*, **36**, 192-236.

[11] Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician,* **24**, 179-195.

[12] Besag, J. (2000). Markov Chain Monte Carlo for statistical inference. *Working Paper 9*. Center for Statistical and the Social Sciences, University of Washington, Seattle.

[13] Britton, T. and O'Neill, P.D. (2002). Bayesian inference for stochastic epidemics in populations with random social structure. *Scandinavian Journal of Statistics,* **29**, 375-390.

[14] Butts, C.T. (2006). network: Classes for relational data. `R package version 1.1-2.`

[15] Chipman, H.A., George, E.I. and McCulloch, R.E. (2001). The practical implementation of Bayesian model selection. *Institute of Mathematical Statistics Lecture Notes–Monograph Series,* **38**, 65-116.

[16] Cressie, N. (1993). *Statistics for spatial data*. New York: Wiley.

[17] Diebolt, J. and Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society B,* **56**, 363-375.

[18] Eubank, S., Guclu, H., Anil Kumar, V.S., Marathe, M.V., Srinivasan, A., Toroczkai, Z., Wang, N. (2004). Modeling disease outbreaks in realistic urban social networks. *Nature*, **429**, 180-184.

[19] Eubank, S., Anil Kumar, V.S., Marathe, M., Srinivasan, A., Wang, N. (2006). Structure of social contact networks and their impact on epidemics. In *Discrete methods in epidemiology*. Abello J., Cormode, G. (ed.). DIMACS Series in Discrete Mathematics and Theoretical Computer Science, **70**, 179-185.

[20] Fienberg, S.E., Meyer, M.M. and Wasserman, S. (1985). Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association,* **80**, 51-67.

[21] Fowler, P.A. (1988). The Konigsberg bridges–250 years later. *The American Mathematical Monthly,* **95**, 42-43.

[22] Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association,* **81**, 832-842.

[23] Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2004). *Bayesian data analysis.* London: Chapman and Hall.

[24] George, E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association,* **88**, 881-889.

[25] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in practice.* London: Chapman and Hall.

[26] Gill, P.S. and Swartz, T.B. (2004). Bayesian analysis of directed graphs data with applications to social networks. *Applied Statistics*, **53**, 249-260.

[27] Girvan, M. and Newman, M.E.J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Science USA*, **99**, 7821-7826.

[28] Golub, G.H. and Van Loan, C.F. (1996). *Matrix computations.* 3rd ed. London: John Hopkins University Press.

[29] Handcock, M.S., Hunter, D.R., Butts, C.T., Goodreau, S.M. and Morris, M. (2003). statnet: An R package for the Statistical modeling of Social Networks `http://www.csde.washington.edu/statnet`

[30] Handcock, M.S., Raftery, A.E. and Tantrum, J.M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society A*, **170**, 301-354.

[31] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The elements of statistical learning: data mining, inference, and prediction.* New York: Springer.

[32] Higdon, D.M. (1998). A process-convolution approach to modeling temperatures in the north Atlantic Ocean. *Journal of environmental and ecological statistics*, **5**, 173-190.

[33] Higdon, D.M. (2002). Space and space-time modeling using process convolutions. In *Quantitative methods for current environmental issues.* Andersons, C., Barnett, V., Chatwin, P.C. and El-Shaarawi, A.H. (ed.). London: Springer-Verlag, 37-56.

[34] Higdon, D. (2006). A primer on space-time modeling from a Bayesian perspective. In *Statistics of spatio-temporal systems.* Finkenstadt, B. and Held, L. (ed.). London: Chapman and Hall.

[35] Hoff, P.D., Raftery, A.E. and Handcock, M.S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, **97**, 1090-1098.

[36] Hoff, P.D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, **100**, 286-295.

[37] Hoff, P.D. (2007a). Multiplicative latent factor models for description and prediction of social networks. *Computational and Mathematical Organization Theory*, to appear.

[38] Hoff, P.D. (2007b). Discussion of Handcock, Raftery and Handcock 'Model-based clustering for social networks.' *Journal of the Royal Statistical Society A,* **170**, 339.

[39] Holland, P.W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, **76**, 33-65.

[40] Huberty, C.J., Wisenbaker, J.M., Smith, J.D. and Smith, J.C. (1986). Using categorical variables in discriminant analysis. *Multivariate Behavioral Research,* **21**, 479-496.

[41] Hunter, D.R. and Handcock, M.S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics,* **15**, 565-583.

[42] James, A.T. (1954). Normal multivariate analysis and the orthogonal group. *The Annals of Mathematical Statistics,* **25**, 40-75.

[43] Jones, D.R., Schonlau, M. and Welch, W.J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, **13**, 455-492.

[44] Kern, J.C. (2000). *Bayesian process-convolution approaches to specifying spatial dependence structure.* PhD. Thesis, Duke University, Institute of Statistics and Decision Sciences.

[45] Lee, H.K.H., Higdon, D.M., Calder, C.A. and Holloman, C.H. (2005). Efficient models for correlated data via convolutions of intrinsic processes. *Statistical Modeling*, **5**, 53-74.

[46] Lee, H.K.H., Sansó, B., Zhou, W. and Higdon, D. (2007). Inference for a proton accelerator using convolution models. *Journal of the American Statistical Association,* to appear.

[47] Lehmann, E.K. (1983). *Theory of point estimation.* New York: Wiley.

[48] Liggett, T.M. (1985). *Interacting particle systems.* New York: Springer.

[49] McCullough, P. and Nelder, J.A. (1983). *Generalized linear models.* London: Chapman and Hall.

[50] McKay, M.D., Beckman, R.J. and Conover, W.J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **21**, 239-245.

[51] McMillan, N.J., Sacks, J., Welch, W.J. and Gao, F. (1999). Analysis of protein activity data by Gaussian stochastic process models. *Journal of Biopharmaceutical Statistics,* **9**, 145-160.

[52] McPherson, M., Smith-Lovin, L. and Cook, J.M. (2001). Birds of a feather: homophily in social networks. *A. Rev. Sociol.*, **27**, 415-444.

[53] Meyers, L.A., Pourbohloul, B., Newman, M.E.J., Skrowronski, D.M. and Brunham, R.C. (2005). Network theory and SARS: predicting outbreak diversity. *Journal of theoretical biology*, **232**, 71-81.

[54] Meyers, L.A. (2007). Contact network epidemiology: bond percolation applied to infectious disease prediction and control. *Bulletin of the American Mathematical Society,* **44**, 63-86.

[55] Newman, M.E.J. (2002). Spread of epidemic disease on networks. *Physical Review E,* **66**, 016128.

[56] Newman, M.E.J. (2003). Random graphs as models of networks. In *Handbook of graphs and networks.* Bornholdt, S. and Schuster, H.G. (ed.). Berlin: Wiley-VCH, 35-68.

[57] Padgett, J.F. and Ansell, C.K. (1993). Robust action and the rise of the Medici. *American Journal of Sociology,* **98**, 1259-1319.

[58] Preston, D.L., Tonks, D.L. and Wallace, D.C. (2003). Model of plastic deformation for extreme loading conditions. *Journal of Applied Physics,* **93**, 211-220.

[59] Qian, Z., Wu, H. and Wu, C.F.J. (2006). Gaussian process models for computer experiments with qualitative and quantitative factors. Available at `http://www.stat.wisc.edu/ zhiguang/gpqq.pdf`

[60] Resnick M.D., Bearman, P.S., Blum R.W. et al. (1997). Protecting adolescents from harm. Findings from the National Longitudinal Study on Adolescent Health. *Journal of the American Medical Association*, **278**, 823-32.

[61] Sacks, J., Schiller, S.B. and Welch, W.J. (1989a). Designs for computer experiments. *Technometrics,* **31**, 41-47.

[62] Sacks, J., Welch, W.J., Mitchell, T.J. and Wynn, H.P. (1989b). Design and analysis of computer experiments. *Statistical Science*, **4**, 409-435.

[63] Snijders, T.A.B. (2002). Markov Chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, **3**. Available at `www.cmu.edu/joss/content/articles/volume3/Snijders.pdf`

[64] Snijders, T.A.B., Pattison, P.E., Robins, G.L. and Handcock, M.S. (2006). New specifications for exponential random graph models. *Sociological Methodology,* **36**, 99-153.

[65] Strauss, D. and Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association,* **85**, 204-212.

[66] Tuv, E. and Runger, G.C. (2004). Scoring levels of categorical variables with heterogeneous data. *IEEE Intelligent Systems,* **19**, 14-19.

[67] Villani, M. and Larsson, R. (2006). The multivariate split normal distribution and asymmetric principal components analysis. *Communications in Statistics–Theory and Methods,* **35**, 1123-1140.

[68] Wang, Y.J. and Wong, G.Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association,* **82**, 8-19.

[69] Wang, J.C. and Wu, C.F.J. (1992). Nearly orthogonal arrays with mixed levels and small runs. *Technometrics,* **34**, 450-456.

[70] Ward, M.D. and Hoff, P.D. (2007). Persistent patterns of international commerce. *Journal of Peace Research,* **44**, 157-175.

[71] Wasserman, S. and Faust, K. (1994). *Social network analysis: methods and applications*. Cambridge: Cambridge University Press.

[72] Wasserman, S. and Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^*. *Psychometrika*, **61**, 401-425.

[73] Watts, D.J. and Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. *Nature,* **393**, 409-410.

[74] Welch, W.J., Buck, R.J., Sacks, J., Wynn, H.P., Mitchell, T.J. and Morris, M.D. (1992). Screening, predicting and computer experiments. *Technometrics*, **34**, 15-25.

[75] West, D.B. (2001). *Introduction to graph theory.* 2nd ed. New Jersey: Prentice Hall.

[76] Wong, G.Y. (1987). Bayesian models for directed graphs. *Journal of the American Statistical Association,* **82**, 140-148.

[77] Yousef, R. and el Hindi, K. (2005). Training radial basis function networks using reduced sets as center points. *International Journal of Information Technology,* **2**, 21-35.