STATISTICAL METHODS FOR STUDYING

THE EVOLUTION OF NETWORKS AND BEHAVIOR
RIJKSUNIVERSITEIT GRONINGEN

STATISTICAL METHODS FOR STUDYING
THE EVOLUTION OF NETWORKS AND BEHAVIOR

Proefschrift

ter verkrijging van het doctoraat in de
Gedrags- en Maatschappijwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op
maandag 21 mei 2007
om 14.45 uur

door

Michael Schweinberger
geboren op 5 oktober 1976
te Wangen im Allgäu, Duitsland
Promotor: Prof. dr. T.A.B. Snijders

Beoordelingscommissie: Prof. dr. O. Frank
Prof. dr. M.S. Handcock
Prof. dr. H.J.A. Hoijtink
## Contents

1 **Introduction**  
   1.1 Longitudinal network models .......................... 2  
   1.2 Longitudinal network and behavior models ............... 5  
   1.3 Structure of the thesis ................................ 6  
   1.4 Computer programs and documentation ...................... 10  

2 **Estimating functions: derivative estimation**  
   2.1 Introduction .............................................. 11  
   2.2 Probabilistic framework .................................... 12  
   2.3 Derivative estimation ...................................... 15  
       2.3.1 Finite differences method ........................... 16  
       2.3.2 Likelihood ratio / score function method .......... 18  
   2.4 Applications .............................................. 23  
       2.4.1 Application: Jacobian matrix known ................. 23  
       2.4.2 Application: Jacobian matrix unknown .............. 28  
   2.5 Discussion ................................................ 31  

Appendix A: Complete-data efficient score  
   A.1 Conventional models .......................... 34  
   A.2 More general models .......................... 34  
   A.3 Independent Arcs model ......................... 36  

Appendix B: Independent Arcs model: Jacobian matrix ........... 37  

3 **Tests of goodness-of-fit**  
   3.1 Introduction .............................................. 39  
   3.2 Model specification and estimation .................... 41  
       3.2.1 Model specification ............................. 41  
       3.2.2 Model estimation ............................. 43  
   3.3 Test statistic ............................................ 43
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3.1</td>
<td>Literature review</td>
<td>44</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Goodness-of-fit test statistic</td>
<td>44</td>
</tr>
<tr>
<td>3.4</td>
<td>Model misspecifications</td>
<td>49</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Misspecifications I</td>
<td>49</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Misspecifications II: homogeneous periods</td>
<td>50</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Misspecifications II: homogeneous nodes</td>
<td>51</td>
</tr>
<tr>
<td>3.5</td>
<td>Monte Carlo study</td>
<td>51</td>
</tr>
<tr>
<td>3.6</td>
<td>Application</td>
<td>57</td>
</tr>
<tr>
<td>3.7</td>
<td>Discussion</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>Random effects models</td>
<td>61</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>61</td>
</tr>
<tr>
<td>4.2</td>
<td>Model</td>
<td>63</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Fixed effects models</td>
<td>63</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Random effects models</td>
<td>65</td>
</tr>
<tr>
<td>4.3</td>
<td>Maximum likelihood estimation</td>
<td>66</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Implementation: data augmentation</td>
<td>68</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Implementation: closed-form expressions</td>
<td>69</td>
</tr>
<tr>
<td>4.4</td>
<td>Bayesian estimation</td>
<td>71</td>
</tr>
<tr>
<td>4.5</td>
<td>Application</td>
<td>72</td>
</tr>
<tr>
<td>4.6</td>
<td>Discussion</td>
<td>77</td>
</tr>
<tr>
<td>Appendix: Gradients and Hessian matrices</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Bayesian modeling and estimation</td>
<td>81</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>81</td>
</tr>
<tr>
<td>5.2</td>
<td>Data structure</td>
<td>83</td>
</tr>
<tr>
<td>5.3</td>
<td>Model</td>
<td>83</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Timing of changes</td>
<td>84</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Changes: conditional probability law</td>
<td>84</td>
</tr>
<tr>
<td>5.4</td>
<td>Bayesian estimation</td>
<td>86</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Notation</td>
<td>86</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Motivation</td>
<td>86</td>
</tr>
<tr>
<td>5.4.3</td>
<td>Bayesian methods</td>
<td>90</td>
</tr>
<tr>
<td>5.5</td>
<td>Applications</td>
<td>93</td>
</tr>
<tr>
<td>5.5.1</td>
<td>Application I</td>
<td>94</td>
</tr>
<tr>
<td>5.5.2</td>
<td>Application II</td>
<td>101</td>
</tr>
</tbody>
</table>
CONTENTS

5.6 Discussion .................................................. 103

6 Discussion ............................................ 107
  6.1 Summary ................................................. 108
  6.2 Future research ......................................... 110

Appendix A: Guide to R functions ............................. 113

Summary in Dutch ............................................. 115

Bibliography ................................................ 119

ICS dissertation series ....................................... 125
Chapter 1

Introduction

Human beings are interdependent, not the least due to the need to survive, i.e., to produce food, clothing, shelter, etc. Human interdependence is evident in the mere existence of human language, reflecting the need to coordinate common tasks, and in social relationships between human beings.

Social relationships between actors (e.g., human beings, enterprises), called “social networks”, have been studied for decades (see Freeman, 2004); examples include friendships and sexual relationships among human beings, collaboration among employees, and ownership relations among enterprises. It is common that social relationships have an impact on actor-level outcome variables, here called “behavior”: e.g., friends may influence adolescents to initiate health-risky or criminal behavior such as smoking, drinking alcohol, or taking drugs; friends or collaborators at the workplace may influence the commitment of employees. The behavior in turn may have an impact on social relationships: e.g., adolescents may select others as friends who are similar in terms of alcohol consumption; employees may select others as friends who show similar levels of commitment. In other words, there may be a mutual dependence between social relationships on one hand and behavior on the other hand.

Empirical data generated by such processes show that relationships among actors are associated with similar behavior; one example is the finding that friendships among adolescents tend to be associated with smoking initiation (see, e.g., Fisher and Bauman, 1988). An important question is whether similar behavior among related actors is due to (1) the influence of related actors (influence processes); (2) the fact that actors select similar others as relation partners (selection processes); (3) both influence and selection processes, which may, or may not, operate at the same time;
CHAPTER 1. INTRODUCTION

or (4) other processes. To answer such questions, it is important to have longitudinal network and behavior data, because non-longitudinal data make it hard to tell whether (1), (2), (3), or (4) holds, and in case of (3), whether the processes operated at the same time or not, and what process preceded the other.

To extract the information contained in longitudinal network and behavior data about influence and selection processes, one needs statistical inference. However, statistical inference proved to be hard: that human beings are interdependent is what makes the social sciences interesting on one hand and statistical modeling hard on the other hand, because statistical modeling is typically based on (conditional) independence assumptions. As a consequence, scientific issues involving influence and selection processes could not be settled in important areas of scientific and societal interest, such as in the mentioned example of smoking initiation among adolescents (Fisher and Bauman, 1988). However, recent advances in the statistical modeling of longitudinal network data have paved the way for the statistical modeling of longitudinal network and behavior data.

Section 1.1 reviews pioneering work on longitudinal network models where the behavior was treated as exogenous, while Section 1.2 reviews recent developments concerning longitudinal network and behavior models. Section 1.3 sketches the advances made by the present thesis in terms of the statistical modeling of longitudinal network and behavior data, while Section 1.4 reviews computer programs (including documentation) which implement the proposed models and methods.

1.1 Longitudinal network models

Let \( N = \{1, \ldots, n\} \) be a finite set of actors. Consider a network on \( N \) and an actor-level outcome variable (“behavior”). It is assumed that the network is directed, i.e., the relationship from actor \( i \) to actor \( j \) is not constrained to be equal to the relationship from \( j \) to \( i \); the directed network is assumed to be binary, so that it can be represented by a binary \( n \times n \) matrix \( x \equiv x(t) \), where element \( x_{ij} \equiv x_{ij}(t) \) is 1 if actor \( i \) is related to actor \( j \) at time \( t \) and 0 otherwise; the convention is adopted that \( x_{ii} \equiv 0 \), and the terms (directed) network and (di)graph are used synonymously.

The behavior is assumed to be binary or integer-valued and is represented by a \( n \times 1 \) vector \( y \equiv y(t) \), where \( y_i \equiv y_i(t) \) refers to the behavior of actor \( i \) at time \( t \). Extensions to integer-valued \( x_{ij} \) are possible; note that the assumption that \( x_{ij} \) and \( y_i \) are binary or integer-valued is not too restrictive, because such data are common; in addition, extensions to more than one outcome variable are possible.
1.1. **LONGITUDINAL NETWORK MODELS**

The basic idea underlying the longitudinal network models pioneered by Holland and Leinhardt (1977) was stated as follows:

“We see two distinct ideas in the statement: “Choices tend to be reciprocated.” First, that the digraph can change; and second, that this change will depend on the digraph’s current structure. That is, current structure influences the probability of change into future structure. More generally, all notions of structural tendencies have this quality.” (Holland and Leinhardt, 1977, pp. 7—8).

Motivated by these considerations, Holland and Leinhardt (1977) postulated that changes of the current network $x$ into a future network $x^*$ are governed by a continuous-time Markov process. The rate of changing $x$ into $x^* \neq x$ was assumed to be of the form

$$q(x, x^*) = \begin{cases} 
\pi_{x,ij}(x) & \text{if } x \text{ and } x^* \text{ differ only in the } (i,j)\text{-th element,} \\
0 & \text{if } x \text{ and } x^* \text{ differ in more than one element.} 
\end{cases} \quad (1.1)$$

Thus, the Markov process proceeds in mini-steps, changing one relationship $x_{ij}$ at a time, and the rate of changing $x_{ij}$ is given by $\pi_{x,ij} \equiv \pi_{x,ij}(x)$. A number of specifications of $\pi_{x,ij}$ was discussed by Holland and Leinhardt (1977), including second-order dependence (reciprocity), third- and higher-order dependencies (such as transitivity), and covariates (including exogenous behavior $y$). To estimate such models, Holland and Leinhardt (1977) proposed the method of moments (MM), but the estimation framework remained incomplete. Wasserman (1979, 1980), Leenders (1995a,b), and Snijders (1999) elaborated on Holland and Leinhardt (1977), however, the proposed estimation methods were not applicable to models with third- and higher-order dependencies.

Significant progress was made by Snijders (1996, 2001), who combined the mini-step Markov models of Holland and Leinhardt (1977) with random utility models (see, e.g., McFadden, 1974) and, more importantly, showed how to estimate models with third- and higher-order dependencies. Snijders (1996, 2001) assumed that the rates of change are of the form (1.1), but modeled the mini-steps as driven by actors $i$ by decomposing $\pi_{x,ij}$ according to

$$\pi_{x,ij}(x) = \lambda_{x,i}(x) \ p_{X,i}(j \mid x),$$

where $\lambda_{x,i}$ is the rate at which $i$ changes one of the relationships $x_{ih}$ ($h \neq i$), while $p_{X,i} \equiv p_{X,i}(j \mid x)$ is the conditional probability that $i$ selects the actor $j \neq i$ and...
changes the relationship $x_{ij}$ into $x^*_{ij} = 1 - x_{ij}$. The conditional probability $p_{X,i}$ was modeled as having multinomial logit form, which (cf. McFadden, 1974) corresponds to assuming that $i$ selects the actor $j \neq i$ which maximizes

$$f_{X,i}(j \mid x) + U_{X,ij}(t),$$

where $f_{X,i}$ is called the network objective function of $i$ and $U_{X,ij}(t)$ is a Gumbel(0, 1) distributed random variable, representing the unknown determinants that influence the choice of $i$ and randomness. The network objective function is a weighted sum of statistics,

$$f_{X,i}(j \mid x) = \eta^X X s_{X,i}(j, x),$$

where $\eta^X$ is a vector-valued weight (parameter) and $s_{X,i}$ is a vector-valued statistic. Such models can capture second-, third-, and higher-order dependencies as well as the impact of covariates (including exogenous behavior $y$) by specifying suitable statistics $s_{X,i}$. Some aspects of model specification were discussed in Snijders (1995, 1996), Snijders and Van Duijn (1997), Snijders (2001, 2003), Huisman and Snijders (2003), and Snijders (2005). To estimate such models from discrete observations of the network (network panel data, the common form of longitudinal network data), Snijders (1996, 2001) proposed MM estimation, implemented by stochastic approximation (Robbins and Monro, 1951) and Monte Carlo (MC) methods. Advances in terms of estimation were made by Koskinen (2004), who proposed Bayesian estimation using Markov chain Monte Carlo (MCMC) methods, and Snijders, Koskinen, and Schweinberger (2006), who proposed maximum likelihood (ML) estimation using stochastic approximation and MCMC methods.

Another significant step forward was the basic idea of Snijders to extend the longitudinal network models by assuming that the Markov process governs the change in both networks and behavior. The idea led to the research programme “The Dynamics of Networks and Behavior” (DNB), funded by the Netherlands Organization for Scientific Research (NWO). Its objective was formulated as follows:

“In the first place, a statistical methodology will be elaborated for analysing longitudinal data on social networks and individual characteristics... Second, these questions will be studied theoretically and empirically...” (NWO dossier no. 401-01-550).

The present thesis is the result of research carried out by the author within the research programme DNB and is concerned with its first objective, the elaboration of
1.2. Longitudinal network and behavior models

The basic idea of Snijders was first given form by Snijders, Steglich, and Schweinberger (2006) and applied to empirical data by Steglich, Snijders, and Pearson (2006), and may be sketched as follows. At stochastic times, either the network $x$ or the behavior $y$ is allowed to change. Conditional on the event that $x$ is allowed to change, an actor $i$ is sampled and assumed to select the actor $j \in N$ which maximizes (1.2), and if $i = j$, then $i$ is assumed to change nothing, otherwise $i$ transforms $x_{ij}$ into $x_{ij}^* = 1 - x_{ij}$. Conditional on the event that $y$ is allowed to change, an actor $i$ is sampled. Let $\mathcal{I} = \{-1, 0, 1\}$ be the set of possible increments of $y_i$, and $\mathcal{M} \subseteq \mathcal{I}$ be the set of increments which are admissible given the current behavior $y_i$ (if $\{a, a+1, \ldots, b\}$, for two integers $a, b$ such that $a < b$, is the outcome space of variable $Y_i$, then $\mathcal{M} = \mathcal{I}$ unless $y_i = a$ or $y_i = b$). Actor $i$ is assumed to transform $y_i$ into $y_i^* = y_i + j$ by selecting the increment $j \in \mathcal{M}$ which maximizes

$$f_{Y,i}(j \mid y) + U_{Y,ij}(t),$$

where $f_{Y,i}$ is called the behavior objective function of $i$ and $U_{Y,ij}(t)$ is a Gumbel$(0, 1)$ distributed random variable. The behavior objective function is a weighted sum of statistics,

$$f_{Y,i}(j \mid y) = \eta_Y \cdot s_{Y,i}(j, y),$$

where $\eta_Y$ is a vector-valued weight (parameter) and $s_{Y,i}$ is a vector-valued statistic.

The model can be used to study influence and selection by including, e.g., terms of the form $\beta_X \sum_{h=1}^n x_{ih}^* d(y_i, y_h)$ in $f_{X,i}$ (selection) and terms of the form $\beta_Y \sum_{h=1}^n x_{ih} d(y_i^*, y_h)$ in $f_{Y,i}$ (influence), where $\beta_X$ and $\beta_Y$ are parameters and $d(y_i, y_h)$
is a similarity (distance) function which expresses the similarity of \( i \) and \( h \) with respect
to \( y_i \) and \( y_h \); countless other specifications of influence and selection are conceivable,
depending on both network statistics (involving, e.g., popular actors, subsets of equivalent actors, cohesive subsets of actors) and behavior statistics (e.g., \( y_h, d(y_i, y_h) \)).

To estimate such models from network and behavior panel data, Snijders, Steglich, and Schweinberger (2006) proposed MM estimation.

1.3 Structure of the thesis

The point of departure is the longitudinal network and behavior modeling framework of Section 1.2. Chapters 2—5 advance the framework both in terms of modeling and statistical inference. Note that the chapters are articles, tailored to statistical journals with diverse aims and readerships, and thus are heterogeneous in terms of notation and (conciseness of) presentation.

Chapter 2: Estimating functions: derivative estimation. The chapter considers network and behavior panel data. In the MM framework, the derivatives of the estimating function with respect to the parameters are required for (1) the standard errors of the parameter estimates (see Snijders, Steglich, and Schweinberger, 2006); (2) goodness-of-fit test statistics (see Chapter 3); (3) sensitivity analyses: identifying parameters for which small changes in parameter value are associated with huge changes in expected values of statistics (which could cast doubt on the model specification); (4) linear extrapolation: computing, if parameters (interpretable as preferences of actors) changed, how much the expected values of statistics would change.

Since the derivatives of the estimating function are not available in closed form, the derivatives are estimated by MC methods. The conventional finite differences (FD) estimator of the derivatives, proposed by Snijders (1996, 2001), is studied and found to be biased and inconsistent. What is more, the FD estimator has the drawback that a perturbation entity \( \epsilon \) has to be specified, which is associated with a bias-variance dilemma, because under regularity conditions the bias of the FD estimator is of order \( \epsilon \) while its MC variance is at least of order \( \epsilon^{-1} \) (see Lemma 1), so that decreasing \( \epsilon \) tends to decrease the bias but increase the MC variance, while increasing \( \epsilon \) tends to decrease the MC variance but increase the bias. In practice, the bias-variance dilemma implies that specifying FD algorithms is hard. In addition, the number of iterations required to evaluate the FD estimator is linear in the number of parameters, and its evaluation is hence very time-consuming when the number of parameters is
moderate or large.

Three alternative estimators of the derivatives are proposed based on the likelihood ratio / score function method of derivative estimation; two of these estimators exploit variance-reduction methods based on control variates. By Lemmas 2 and 3, the estimators are unbiased and consistent, and one has minimum variance in a large class of estimators. Besides, for moderate or large numbers of parameters, the two control variate estimators tend to be more efficient than the FD estimator in the sense that, to obtain the same MC variance, the two estimators tend to require less computation time than the FD estimator. Therefore, the proposed estimators have great theoretical and practical advantages compared to the FD estimator.

The finite-sample behavior of the estimators is studied by presenting theoretical insight and, in addition, two Monte Carlo simulation studies, which compare the FD estimator with the proposed estimators.

Chapter 3: Tests of goodness-of-fit. The chapter considers network and behavior panel data (though the behavior remains implicit for ease of presentation). In the MM framework, statistical inference has been limited to parameter estimates, standard errors, and one-parameter $t$-tests, while the issue of goodness-of-fit—i.e., whether the observed data support a given model specification—has received little attention.

A test statistic is proposed, which can be regarded as a generalized score test statistic based on regular estimating functions, and which (1) has many applications; (2) does not require to estimate the parameters to be tested, which (a) saves computation time and (b) allows to test hard-to-estimate parameters; (3) admits one- and two-sided tests as well as (4) multi-parameter tests; and (5) has an appealing interpretation in terms of goodness-of-fit, in the sense that it compares the expected value of some function of the data to the observed value of the function; that is, it uses the observed data as an external benchmark to which model predictions are compared.

When the goodness-of-fit of the model restricted by the null hypothesis is found to be unacceptable, it is typically desired to compute the unrestricted estimates of all parameters of interest, which requires additional, time-consuming MC simulations. To cut down the computational burden, one-step estimators are derived, which are crude approximations of the unrestricted estimators, but have the advantage that no additional MC simulations are required to evaluate them.

The finite-sample behavior of one- and multi-parameter goodness-of-fit tests is explored by a Monte Carlo simulation study, and compared to the one-parameter $t$-tests. In addition, the test statistic is applied to a large empirical data set of more
than 400 actors, and its usefulness in forward model selection procedures is demonstrated.

Chapter 4: Random effects models. The chapter considers network panel data (though most arguments can be extended to network and behavior panel data: see Chapter 5). An important assumption of Snijders (1996, 2001) and Snijders, Steglich, and Schweinberger (2006) is that the weights $\eta_X$ are constant across actors. The constant-weights assumption facilitates statistical estimation, but the goodness-of-fit of corresponding models may be doubtful, because the implicit assumption is that all relevant actor-level covariates are (a) observed and (b) correctly incorporated in the model. In applications, both (a) and (b) may not be satisfied: (a) in the social sciences, it is common that not all relevant covariates are observed, due to data collection constraints (e.g., limited time and willingness of respondents to respond to countless stimuli) and limited knowledge of the researcher as to what covariates are relevant; (b) the model admits to incorporate covariates in the rate function $\lambda_{X,i}$ and the objective function $f_{X,i}$ in many ways, and even when all relevant covariates are collected and incorporated in the model, some covariate-related effects of the data-generating model may be excluded from the specified model.

A model is proposed which replaces the constant-weights assumption by the assumption that the weights are unobserved (latent) outcomes of actor-dependent random variables, governed by a probability law that is common to all actors. If $x$ is allowed to change, an actor $i$ is sampled and supposed to select the actor $j \in N$ which maximizes

$$f_{X,i}^*(j \mid x) + T_{X,j} + U_{X,ij}(t), \quad (1.3)$$

where

$$f_{X,i}^*(j \mid x) = \eta_{X,i}^sX,i(j, x).$$

The model is more general than (the network part of) the model of Snijders (1996, 2001) and Snijders, Steglich, and Schweinberger (2006) in two respects. First, and foremost, the weight $\eta_{X,i}$ is allowed to depend on actor $i$ by modeling $\eta_{X,i}$ as

$$\eta_{X,i} = \beta_X + A_XV_{X,i},$$

where $\beta_X$ is a parameter vector, $A_X$ is a design matrix, and $V_{X,i} \sim N(0, \Sigma_X)$ is a Gaussian distributed random vector; observe that $V_{X,i}$ can be conceived as the sum of
covariate-related effects which are part of the data-generating model but are excluded from the specified model. Second, scalar $T_{X,j} \sim N(0, \tau_X)$ is a Gaussian distributed random variable, which expresses the latent popularity of actor $j$ and is an alternative to conventional terms capturing the effect of the popularity of $j$ in $f_{X,i}$. The terms $V_{X,i}$ and $T_{X,j}$ are assumed to be constant over time.

Estimating such models is complicated by the fact that (a) the Markov process is not observed in continuous time but at discrete time points, and (b) the estimation of variance-covariance matrix $\Sigma_X$ is subject to the constraint that estimates of $\Sigma_X$ must be symmetric and positive definite. Both ML and Bayesian estimation methods are proposed, using an unconstrained parametrization of $\Sigma_X$; for ML estimation, the non-redundant elements of $\Sigma_X$ are reparametrized so that estimates of $\Sigma_X$ are by construction symmetric and positive definite. In addition, the reparametrization facilitates the estimation of very small variances, which is an important practical advantage. Both ML and Bayesian estimation exploit MCMC-based data augmentation.

The model is applied to an empirical data set, and ML and Bayesian estimation methods are compared.

Chapter 5: Bayesian modeling and estimation. The chapter considers network and behavior panel data. The model of Chapter 4 and the model of Snijders, Steglich, and Schweinberger (2006) are generalized as follows.

Conditional on the event that $x$ is allowed to change, an actor $i$ is sampled and supposed to select the actor $j \in N$ which maximizes (1.3). Conditional on the event that $y$ is allowed to change, an actor $i$ is sampled and supposed to select the increment $j \in M$ which maximizes

$$f^\star_{Y,i}(j \mid y) + U_{Y,ij}(t),$$

where

$$f^\star_{Y,i}(j \mid y) = \eta_{Y,i} s_{Y,i}(j, y).$$

The model is more general than the model of Chapter 4 because the behavior $y$ is endogenous, and more general than the model of Snijders, Steglich, and Schweinberger (2006) because the weight $\eta_{Y,i}$ is allowed to depend on actor $i$ by modeling $\eta_{Y,i}$ as

$$\eta_{Y,i} = \beta_Y + A_Y V_{Y,i},$$

where $\beta_Y$ is a parameter vector, $A_Y$ is a design matrix, and $V_{Y,i} \sim N(0, \Sigma_Y)$ is a Gaussian distributed random vector.
While ML estimation of network and behavior models with constant or actor-dependent weights may proceed along similar lines as in Chapter 4, it is argued that Bayesian estimation is a viable alternative and complement to ML estimation, because (1) social processes are unique and non-repeatable, and Bayesian methods are well-suited to studying such data, and (2) when ML algorithms do not converge, then the incorporation of prior knowledge can enable Bayesian estimation, and when ML algorithms do converge, then Bayesian estimation can complement ML estimation by giving—under vague prior distributions—insight into the shape of the likelihood function and non-linear dependencies between parameters.

A MCMC algorithm for sampling from the posterior distribution is proposed, which can handle (1) networks and behavior and (2) actor-dependent weights; (3) is based on the embedded Markov process, avoiding needless complications encountered by Koskinen (2004); and (4) allows to sample $\beta_X$ (and $\beta_Y$) from the posterior distribution more efficiently than Koskinen (2004).

The Bayesian methods are applied to two small, empirical data sets, and compared to MM and ML estimation.

1.4 Computer programs and documentation

All models and methods proposed in the present thesis were implemented by the author in the computer program Siena. Siena is written in the programming language Delphi, and the source code as well as Windows-executables of Siena can be downloaded from the website http://stat.gamma.rug.nl/stocnet. In practice, users tend to find it more convenient to use Siena in the Windows-based program collection StOCNET, which can be downloaded from the same website. The author programmed in addition the supplementing R functions siena.mle and siena.bayes for ML and Bayesian estimation, respectively, so that the (graphical) facilities of R can be exploited to examine the MCMC output of Siena; these functions can be downloaded from the website mentioned above.

Comprehensive guides to Siena and StOCNET are given by Snijders, Steglich, Schweinberger, and Huisman (2006) and Boer, Huisman, Snijders, Steglich, Wichers, and Zeggelink (2006), respectively. A how-to-use guide for siena.mle and siena.bayes can be found in Appendix A.
Chapter 2

Estimating functions: derivative estimation†

A parametric, continuous-time Markov model for digraph panel data is considered. The parameter is estimated by the method of moments. A convenient method for estimating the variance-covariance matrix of the moment estimator relies on the delta method, requiring the Jacobian matrix—that is, the matrix of partial derivatives—of the estimating function. The Jacobian matrix was estimated hitherto by Monte Carlo methods based on finite differences. Three new Monte Carlo estimators of the Jacobian matrix are proposed, which are related to the likelihood ratio / score function method of derivative estimation and have theoretical and practical advantages compared to the finite differences method. Some light is shed on the practical performance of the methods by applying them in a situation where the true Jacobian matrix is known and in a situation where the true Jacobian matrix is unknown.

Keywords: digraphs, continuous-time Markov process, gradient estimation, likelihood ratio / score function method, variance reduction, control variates.

2.1 Introduction

The present paper considers digraph panel data, that is, data that can be represented by a directed relation (or digraph) on a set of nodes observed at two or more discrete time points. Social scientists use digraphs to study, among other things, informal relations (e.g., friendship) among individuals embedded in formal organizations (business

†Published as Schweinberger and Snijders (2007). Markov models for digraph panel data: Monte Carlo-based derivative estimation. Computational Statistics and Data Analysis. In press.
firms, schools, etc.). Such data tend to display second- and third-order dependencies among observed arcs, so that modeling digraph panel data requires to take into account such dependencies.

A flexible approach to model such dependent data is based on the assumption that the observed digraphs are outcomes at some discrete time points of a Markov process evolving in continuous time, indexed by a parameter $\theta$. The basic, underlying idea dates back to Holland and Leinhardt (1977) and Wasserman (1979, 1980), and was expanded by Snijders (2001) to model third- and higher-order dependencies.

Since the continuous-time Markov process is not observed in continuous time but at discrete time points, the likelihood function cannot be written in closed form and thus likelihood-based inference is hard. The method of moments was proposed by Snijders (2001) to estimate the parameter $\theta$. A convenient method to estimate the variance-covariance matrix of the moment estimator is based on the delta method, which requires the Jacobian matrix of the estimating function, that is, the matrix of first-order partial derivatives. Snijders (2001) estimated the Jacobian matrix by Monte Carlo methods based on finite differences with common random numbers.

The present paper proposes three new estimators of the Jacobian matrix, all of them related to the likelihood ratio / score function method of derivative estimation (Aleksandrov, Sysoyev, and Shemeneva, 1968), and two of them utilizing variance reduction methods based on control variates. The three estimators have theoretical advantages compared to the finite differences method, but an important practical motivation for the estimators is that they roughly cut down the computational burden by a factor $L + 1$, where $L$ is the dimension of $\theta$. The achieved reduction in computation time is of great practical value, since in practice computation time is an issue, and it is not unusual for $L$ to be between 10 and 30.

The paper is structured as follows. The probabilistic framework is outlined in Section 2.2. The central argument is presented in Section 2.3. Section 2.4 compares the estimators of the Jacobian matrix in a situation where the true Jacobian matrix is known and in a situation where the true Jacobian matrix is unknown.

## 2.2 Probabilistic framework

It is assumed that a binary, directed relation $\rightarrow$ on a finite set of nodes $N = \{1, 2, \ldots, n\}$ has been observed at discrete, ordered time points $t_0 < t_1 < \cdots < t_G$. These observations may be represented by digraphs and stored as binary matrices.
2.2. PROBABILISTIC FRAMEWORK

\[ x(t_0), x(t_1), \ldots, x(t_G), \] where element \( x_{ij}(t_g) \) of \( n \times n \) matrix \( x(t_g) \) is defined as

\[
x_{ij}(t_g) = \begin{cases} 
1 & \text{if } i \rightarrow j \text{ at time point } t_g, \\
0 & \text{otherwise}, 
\end{cases} \tag{2.1}
\]

where \( i \rightarrow j \) means that node \( i \) is related to node \( j \); the fact that the relation is directed means that \( x_{ij}(t_g) \) may be different from \( x_{ji}(t_g) \); the diagonal elements \( x_{ii}(t_g) \) are regarded as structural zeros.

It is postulated that the observed digraphs are generated by an unobserved, continuous-time stochastic process. The discrete time points \( t_0 < \cdots < t_G \) are embedded in the time interval \([t_0, t_G]\). The digraph \( x(t_0) \) observed at time point \( t_0 \) is not modeled, that is, the statistical modeling is done conditional on \( x(t_0) \). Consider the case \( G = 1 \).

A simple process can be constructed by assuming that the process is a Markov process. Then the model is specified by the generator of the Markov process, which corresponds to a \( W \times W \) matrix \( Q_{\theta} \) indexed by a parameter \( \theta \), where \( W = 2^{n(n-1)} \) is the number of digraphs on \( N \). The elements \( q_{\theta}(x^*, x) \) of generator \( Q_{\theta} \) are the rates of moving from digraph \( x^* \) to digraph \( x \).

Let \( x^* \) and \( x \) be two arbitrary digraphs on \( N \). If \( x \) deviates from \( x^* \) in more than one arc variable \( x^*_{ij} \), then Snijders (2001) assumes that \( q_{\theta}(x^*, x) = 0 \); in other words, the process moves forward by changing not more than one arc variable \( x^*_{ij} \) at the time.

Let \( x^* \) be an arbitrary digraph on \( N \), and let \( x \) be the digraph that is obtained from \( x^* \) by changing one and only one specified arc variable, say \( x^*_{ij} \). Since the transition from \( x^* \) to \( x \) involves only the ordered pair of nodes \((i, j)\), one can rewrite \( q_{\theta}(x^*, x) \) as \( q_{\theta}(x^*, i, j) \) and decompose \( q_{\theta}(x^*, i, j) \) as follows:

\[
q_{\theta}(x^*, i, j) = \lambda_i(\theta, x^*) \ r_i(\theta, x, j), \tag{2.2}
\]

where

\[
\lambda_i(\theta, x^*) = \sum_{h \neq i}^n q_{\theta}(x^*, i, h) \tag{2.3}
\]

and

\[
r_i(\theta, x, j) = \frac{q_{\theta}(x^*, i, j)}{\lambda_i(\theta, x^*)}. \tag{2.4}
\]

The interpretation is that \( \lambda_i \)—called the rate function—is the rate at which the set \( \{x^*_{ij} : j \neq i \in N\} \) of arcs emanating from node \( i \) is changed, while \( r_i \) gives the conditional probabilities of such changes.
A simple specification of the rate function is
\[
\lambda_i(\theta, x^*) = \rho,
\]
that is, as constant across nodes \(i\) and digraphs \(x^*\), equal to some rate parameter \(\rho\). A more general specification of the rate function is given by
\[
\lambda_i(\theta, x^*) = \rho \exp \left[ \alpha' a_i(x^*, c_i) \right],
\]
where \(\alpha = (\alpha_k)\) is a vector valued parameter and \(a_i = (a_{ik})\) is a vector valued function of covariates \(c_i\) depending on node \(i\) and graph-dependent statistics involving the arcs of node \(i\). When there is more than one time interval, that is, \(G > 1\), then parameter \(\rho\) can be made dependent on time interval \([t_{g-1}, t_g]\).

A convenient, multinomial logit parametrization of \(r_i(\theta, x, j)\) is given by
\[
r_i(\theta, x, j) = \frac{\exp \left[ f_i(\beta, x, j) \right]}{\sum_{h=1}^{n} \exp \left[ f_i(\beta, x, h) \right]},
\]
where
\[
f_i(\beta, x, j) = \beta' s_i(x, j),
\]
while \(\beta = (\beta_k)\) is a vector valued parameter and \(s_i = (s_{ik})\) is a vector valued statistic. The function \(f_i\) is called the objective function. Examples of statistics \(s_{ik}\) are the number of arcs \(\sum_{h=1}^{n} x_{ih}\), the number of transitive triplets \(\sum_{h,l=1}^{n} x_{ih} x_{hl} x_{il}\), or other statistics involving the arcs of node \(i\) and covariates; such statistics can be used to define third- and higher-order dependencies.

Remark 2.2.1: simulation of the Markov process The methods proposed in Section 2.3 rely on Monte Carlo simulation of the Markov process. Therefore, it is worthwhile to consider how the Markov process can be simulated. Let \(\text{Exp}(\psi)\) be the negative exponential distribution with parameter \(\psi\). The Markov process can be simulated in time interval \([t_0, t_1]\) conditional on digraph \(x(t_0)\) observed at time point \(t_0\) by starting at time point \(t = t_0\) at digraph \(x^{(0)} = x(t_0)\) and iterating the following steps (with initial value \(M = 0\)):

1. Increment \(M\).

(1) Increment \(t\):
2.3. DERIVATIVE ESTIMATION

— Sample \( h_M \sim \text{Exp} \left( \sum_{k=1}^{n} \lambda_k \left( \theta, x^{(M-1)} \right) \right) \).
— Set \( t = t + h_M \).

(2) If \( t < t_1 \) then:

— Sample node \( i \) with probability \( \lambda_i \left( \theta, x^{(M-1)} \right) / \sum_{k=1}^{n} \lambda_k \left( \theta, x^{(M-1)} \right) \).
— Given \( i \), sample node \( j \) with probability \( r_i \left( \theta, x^{(M)}, j \right) \) and set \( x^{(M)}_{ij} = 1 - x^{(M-1)}_{ij} \) and \( x^{(M)}_{kl} = x^{(M-1)}_{kl} \) for all \((k,l) \neq (i,j)\).

Else: terminate.

Remark 2.2.2: extensions Extensions to \( G > 1 \) time intervals are straightforward due to the Markov property. It is furthermore possible to deal with the co-evolution of digraphs and other outcome variables (Snijders, Steglich, and Schweinberger, 2006).

2.3 Derivative estimation

The present section begins by briefly discussing the estimation of the parameter vector \( \theta \) and its variance-covariance matrix, which requires the knowledge of the Jacobian matrix of the estimating function. The problem is that the Jacobian matrix of the estimating function cannot be written in closed form and hence must be estimated. Section 2.3.1 outlines a conventional Monte Carlo estimator of the Jacobian matrix based on finite differences, while in Section 2.3.2 three new Monte Carlo estimators are proposed, based on the likelihood ratio / score function method. Attention is restricted to the case of \( G = 1 \) time intervals; the extension to the case \( G > 1 \) is immediate.

Snijders (2001) proposed to estimate the parameter \( \theta \in \Theta \subseteq \mathbb{R}^L \) by the method of moments (Pearson, 1902a,b). Let \( U(t_1) \) be a suitable \( L \times 1 \) vector function (statistic) of the digraph and covariates at time point \( t_1 \) with observed value \( u(t_1) \); the function \( U(t_1) \) may in addition depend on the digraph \( x(t_0) \) and covariates observed at time point \( t_0 \). Let \( \hat{\theta} \) be the solution of the moment equation

\[
E_{\theta}[U(t_1) | X(t_0) = x(t_0)] = u(t_1). \tag{2.9}
\]

In the sequel, the left-hand side of (2.9) is referred to as \( E_{\theta}U \). The estimation problem amounts to finding a root of \( E_{\theta}U - u \) as a function of \( \theta \), where \( u = u(t_1) \). Experience suggests that, for suitable statistics \( U \), Equation (2.9) has a unique root in almost all
cases. The moment estimator \( \hat{\theta} \) is not available in closed form. However, it is possible to simulate the Markov process so that root-finding algorithms based on stochastic approximation (Robbins and Monro, 1951) can be used to find \( \hat{\theta} \) (see Snijders, 2001).

To estimate the variance-covariance matrix of \( \hat{\theta} \), it is inconvenient to use bootstrap (or resampling) methods, because each of the multiple estimation runs required by resampling methods is time-consuming; an alternative is to utilize the delta method (Lehmann, 1999, p. 315) and the implicit function theorem, giving the approximation

\[
\text{Cov}_{\theta} \hat{\theta} \approx \Delta^{-1}(\theta) \Sigma(\theta) [\Delta^{-1}(\theta)]',
\]

where

\[
\Sigma(\theta) = \text{Cov}_{\theta} U
\]

is the \( L \times L \) variance-covariance matrix of \( U \), and

\[
\Delta(\theta) = \frac{\partial}{\partial \theta'} E_{\theta} U
\]

is the Jacobian matrix of \( E_{\theta} U \) at \( \theta \), that is, the \( L \times L \) matrix of first-order partial derivatives of \( E_{\theta} U \) evaluated at \( \theta \). Here again, no closed form expressions are available.

The variance-covariance matrix (2.10) can be estimated by plugging in the moment estimator \( \hat{\theta} \) in (2.11) and (2.12). Monte Carlo estimation of \( \Sigma(\hat{\theta}) \) is straightforward; the issue is how to construct a Monte Carlo estimator of \( \Delta(\hat{\theta}) \). In Section 2.3.1, a conventional Monte Carlo estimator of \( \Delta(\hat{\theta}) \) based on finite differences is described, whereas in Section 2.3.2 three new Monte Carlo estimators are proposed, based on the likelihood ratio / score function method.

### 2.3.1 Finite differences method

By definition,

\[
\Delta_l(\theta) = \lim_{\epsilon \to 0} \frac{E_{\theta + \epsilon \epsilon_l} U - E_{\theta} U}{\epsilon}, \quad l = 1, \ldots, L
\]

where \( \Delta_l(\theta) \) refers to the \( l \)-th column of \( \Delta(\theta) \). The finite differences method to estimate \( \Delta_l(\theta) \) is based on

\[
\Delta_{l,\epsilon}(\theta) = \frac{E_{\theta + \epsilon \epsilon_l} U - E_{\theta} U}{\epsilon}, \quad l = 1, \ldots, L
\]

The expectations \( E_{\theta + \epsilon \epsilon_l} U, \ l = 1, \ldots, L \), and \( E_{\theta} U \) are not available in closed form, but can be estimated by the corresponding Monte Carlo sample averages: given \( \theta + \epsilon_l \epsilon \),
and a pseudo-random number generator (see, e.g., Marsaglia and Zaman, 1991), one can simulate the Markov process as described in Remark 2.2.1 of Section 2.2; having simulated the Markov process multiple times, the Monte Carlo sample average of $U$ can be used as an estimate of $E_{\theta+\epsilon\ell}U$; the expectation $E_{\theta}U$ can be estimated accordingly by simulating the Markov process multiple times given $\theta$ and using the Monte Carlo sample average of $U$ as an estimate of $E_{\theta}U$. Thus, to estimate the $L+1$ expectations, $L+1$ Monte Carlo samples are required, because the parameters are different.

It is well-known (see, e.g., L’Ecuyer, 1991) that the resulting estimator is biased. Under regularity conditions, the bias is of order $\epsilon$, which suggests to make $\epsilon$ as small as possible. On the other hand, it is clear from (2.14) that, when using independent draws of $U$ under the distributions corresponding to $\theta + \epsilon\ell\epsilon$ and to $\theta$, the variance of the resulting estimator is of order $\epsilon^{-2}$, which implies that small values of $\epsilon$ are undesirable. A way out of this dilemma is provided by using common random numbers (Hammersley and Handscomb, 1964, pp. 48—49) for simulating the random variable $U$ under the distributions corresponding to $\theta + \epsilon\ell\epsilon$ and to $\theta$. Denoting by $W$ the random number stream and by $U_{\theta}(W)$ the result of the simulation procedure as a function of $W$ and $\theta$, this means that the same $W$ is used for generating $U$ under $\theta + \epsilon\ell\epsilon$ and under $\theta$, so that the random variable used for generating a value of $\Delta_{l,\epsilon}(\theta)$ is given by

$$\frac{U_{\theta+\epsilon\ell\epsilon}(W) - U_{\theta}(W)}{\epsilon}. \quad (2.15)$$

If $U_{\theta}(w)$ would be a continuously differentiable function of $\theta_l$ for any given $w$, then under regularity conditions the random variable (2.15) would tend to the derivative $\partial U_{\theta}(w)/\partial \theta_l$, its variance would be bounded for $\epsilon \to 0$, and $\epsilon$ could be taken very small to get a finite differences estimator (2.15) which is practically unbiased and $N$-consistent, where $N$ is the size of the Monte Carlo sample. However, the discrete nature of the outcome variable $U$ in the considered model implies that $U_{\theta}(w)$ is a discontinuous function of $\theta$. The following lemma can be used to determine the order of magnitude of the variance of (2.15).

**Lemma 1.** Let $D(\epsilon)$ be a random variable with a finite outcome space that does not depend on $\epsilon$, let $d_{\epsilon} = ED(\epsilon)/\epsilon$, and suppose that $d = \lim_{\epsilon \to 0} d_{\epsilon}$ is finite and non-zero. Then

$$\liminf_{\epsilon \to 0} |\epsilon| \text{Var} \left( \frac{D(\epsilon)}{\epsilon} \right) > 0. \quad (2.16)$$
Proof. Let \( d_0 \) be the smallest non-zero outcome of \(|D(\epsilon)|\). Then \( D^2(\epsilon) \geq d_0 |D(\epsilon)| \) and
\[
E \frac{D^2(\epsilon)}{\epsilon^2} \geq \frac{d_0}{|\epsilon|} E \frac{|D(\epsilon)|}{|\epsilon|}.
\]
(2.17)
From \( \text{Var}(D(\epsilon)/\epsilon) = E[D^2(\epsilon)/\epsilon^2] - d_\epsilon^2 \) it follows that
\[
\lim_{\epsilon \to 0} \inf \left\{ |\epsilon| \text{Var} \left( \frac{D(\epsilon)}{\epsilon} \right) \right\} \geq \lim_{\epsilon \to 0} \inf \left\{ d_0 E \frac{|D(\epsilon)|}{|\epsilon|} - |\epsilon| d_\epsilon^2 \right\}
\]
(2.18)

\( \square \)

Applying Lemma 1 to \( D(\epsilon) = U_{\theta+\epsilon\epsilon}(W) - U_\theta(W) \) shows that if the derivative to be estimated is non-zero, the variance of (2.15) is at least of order \( \epsilon^{-1} \). Note that if the derivative is zero, then \( E_\theta U \) is not sensitive to changes in \( \theta \), and therefore \( U \) is not a sensible choice for estimating \( \theta \) in a method of moments framework (see Snijders, 2001).

In the remainder of the paper, \( D_0(\epsilon) \) refers to the Monte Carlo finite differences estimator with common random numbers which estimates the columns \( \Delta_l(\theta) \) of \( \Delta(\theta) \) by \( \hat{\Delta}_{l,\epsilon}(\theta) \), where \( \hat{\Delta}_{l,\epsilon}(\theta) \) is the Monte Carlo estimator of \( \Delta_{l,\epsilon}(\theta) \) obtained by replacing the expectations \( E_{\theta+\epsilon\epsilon} U \) and \( E_\theta U \) by the corresponding Monte Carlo sample averages. An important practical implication of using \( D_0(\epsilon) \) is that \( L + 1 \) Monte Carlo sample are required. Section 2.3.2 considers an alternative method that produces unbiased and \( N \)-consistent estimators, in contrast to \( D_0(\epsilon) \), and requires only one Monte Carlo sample.

### 2.3.2 Likelihood ratio / score function method

The alternative method is related to the likelihood ratio / score function method of derivative estimation, which can be traced back to Aleksandrov, Sysoyev, and Shemeneva (1968). Some related papers are Rubinstein (1986, 1989) and Glynn and L’Ecuyer (1995).

Denote the complete data—that is, the holding times of the Markov process and the sequence of arc changes in time interval \([t_0, t_1]\)—by \( Z \). Let \( P_\theta \) be the probability law governing \( Z \), admitting a probability density \( p_\theta = dP_\theta / d\mu \) with respect to some dominating measure \( \mu \), and let \( Z_1, Z_2, \ldots \) be Monte Carlo generated random variables with distribution \( P_\theta \).

Three likelihood ratio / score function (LR) estimators of the Jacobian matrix \( \Delta(\theta) \) are derived below, called \( D_I, D_{II}, \) and \( D_{III} \); the dependence of the \( D \)-estimators
on \( \theta \) is left implicit. The LR estimator \( D_I \) (Section 2.3.2) is the basic LR estimator, while LR estimators \( D_{II} \) and \( D_{III} \) (Section 2.3.2) use variance reduction methods and have less Monte Carlo variance than \( D_I \).

**Estimator \( D_I \): the basic LR estimator**

Let

\[
D_I = \frac{1}{N} \sum_{i=1}^{N} U_i \frac{\partial \ln p_{\theta}(Z_i)}{\partial \theta'},
\]

where \( N \) is the size of the Monte Carlo sample.

**Lemma 2.** Let \( U \) be any function of \( x(t_0), x(t_1), \) and covariates that does not depend on \( \theta \). Then \( D_I \) is an unbiased and \( N \)-consistent estimator of the Jacobian matrix \( \Delta(\theta) \) defined in (2.12).

**Proof.** By definition,

\[
\Delta_I(\theta) = \lim_{\epsilon \to 0} \frac{E_{\theta+\epsilon t}U - E_{\theta}U}{\epsilon} = \lim_{\epsilon \to 0} \int_Z U \frac{p_{\theta+\epsilon t}(z) - p_{\theta}(z)}{\epsilon} d\mu(z).
\]

The model definition of Section 2.2 implies that the model for \( Z \) is a family of negative exponential-multinomial distributions, which is an exponential family of distributions. The outcome space of \( U \) is finite, so that \( E_{\theta}U \) exists for all \( \theta \in \Theta \) and is finite. Hence, by Theorem 2.7.1 of Lehmann and Romano (2005, p. 49), it is admissible to interchange the order of differentiation and integration:

\[
\lim_{\epsilon \to 0} \int_Z U \frac{p_{\theta+\epsilon t}(z) - p_{\theta}(z)}{\epsilon} d\mu(z) = \int_Z U \lim_{\epsilon \to 0} \frac{p_{\theta+\epsilon t}(z) - p_{\theta}(z)}{\epsilon} d\mu(z).
\]

Thus the Jacobian matrix can be written as

\[
\Delta(\theta) = \int_Z U \frac{\partial p_{\theta}(z)}{\partial \theta'} d\mu(z) = E_{\theta} \left[ U \frac{\partial \ln p_{\theta}(Z)}{\partial \theta'} \right].
\]

Equation (2.22) proves the unbiasedness of \( D_I \) for estimating \( \Delta(\theta) \). \( N \)-consistency follows from the strong law of large numbers (see, e.g., Ferguson, 1996, p. 21).

Since in practice the aim is to evaluate the Jacobian matrix \( \Delta(\theta) \) at the moment estimate \( \hat{\theta} \) of \( \theta \), \( \hat{\theta} \) is plugged in for \( \theta \). Given \( \hat{\theta} \), the Monte Carlo generation of random variables \( Z_i \) \( (i = 1, 2, \ldots, N) \) with probability law \( P_{\hat{\theta}} \) is straightforward; see Remark 2.2.1 of Section 2.2. The complete-data efficient score \( \partial \ln p_{\hat{\theta}}(Z_i)/\partial \hat{\theta}' \) is derived in Appendix A.

The Monte Carlo variance of the basic LR estimator \( D_I \) may be too large for practical purposes. It is therefore sensible to reduce the variance of \( D_I \) by using variance reduction methods, which is explored in Section 2.3.2.
Estimators $D_{II}$ and $D_{III}$: LR estimators exploiting variance reduction methods based on control variates

In the present section, two LR estimators are proposed which have less variance than $D_I$. Both estimators are based on the idea of reducing the variance of LR estimators by using the complete-data efficient score as a control variate (see Fieller and Hartley, 1954, Rubinstein, 1986, 1989).

It will be convenient to let
\[ S'_\theta = S'_\theta(Z) = \frac{\partial \ln p_\theta(Z)}{\partial \theta}, \]  \hspace{1cm} (2.23)
and to rewrite the Jacobian matrix $\Delta(\theta)$ given by (2.22) using the vec operator as
\[ \text{vec } \Delta(\theta) = E_\theta [\text{vec}(US'_\theta)] = E_\theta [\text{vec}(US'_\theta) - A (S_\theta - E_\theta S_\theta)]. \]  \hspace{1cm} (2.24)

Equation (2.25) suggests that (2.24) can be estimated by
\[ D_C(A) = \frac{1}{N} \sum_{i=1}^{N} [\text{vec} (U_i S'_{\theta,i}) - A S_{\theta,i}] = \text{vec}(D_I) - A \frac{1}{N} \sum_{i=1}^{N} S_{\theta,i}. \]  \hspace{1cm} (2.26)

If $A$ is constant, then $D_C(A)$ as an estimator of (2.24) is both unbiased and $N$-consistent.

The idea of control variates is to exploit the fact that $E_\theta S_\theta$ and hence the Monte Carlo integration error $(1/N) \sum_{i=1}^{N} S_{\theta,i} - E_\theta S_\theta$ is known; let $L = 1$ so that $US_\theta$ and $S_\theta$ are scalars; if $US_\theta$ and $S_\theta$ are correlated, the knowledge of the integration error can be used to (linearly) transform $D_I$ such that $D_I$ gets closer to its expectation, resulting in variance reduction. Two choices of $A$, and hence two control variate estimators, are elaborated below.

**Estimator $D_{II}$: heuristic LR control variate estimator**

A simple, heuristic control variate estimator is obtained as follows. Let
\[ U^* = U - E_\theta U, \]  \hspace{1cm} (2.27)
and observe that at the moment estimate $\hat{\theta}$ of $\theta$, $E_\theta U$ is known and given by $E_{\hat{\theta}} U = u$, where $u$ is the observed value of statistic $U$. The centering (2.27) is equivalent to using
2.3. DERIVATIVE ESTIMATION

$S_\theta$ as a control variate with $A = I_L \otimes u$, where $I_L$ is the $L \times L$ identity matrix, which is evident from the identity

$$\text{vec}(U^* S_\theta') = \text{vec}(US_\theta) - \text{vec}(u S_\theta)$$

(2.28)

and the fact that $\text{vec}(u S_\theta')$ can be written as

$$\text{vec}(u S_\theta') = \text{vec}(u S_\theta' I_L) = (I_L \otimes u) \text{vec}(S_\theta') = A S_\theta.$$  

(2.29)

The resulting estimator, which is (2.26) with $A = I_L \otimes u$, is denoted by $D_{II}$:

$$D_{II} = D_C(I_L \otimes u).$$

(2.30)

Since $A = I_L \otimes u$ is constant, $D_{II}$ as an estimator of (2.24) is both unbiased and $N$-consistent.

To give some insight into the behavior of estimator $D_{II}$, let $L = 1$ so that $U$, $S_\theta$, $D_I$, and $D_{II}$ are scalars. Using $E_\theta[U^* S_\theta] = E_\theta[U S_\theta]$,

$$\text{Var}_\theta(D_I) - \text{Var}_\theta(D_{II}) = \frac{E_\theta[(U S_\theta)^2] - E_\theta[(U^* S_\theta)^2]}{N} = E_\theta\left[\left(\frac{2U}{u} - 1\right) u^2 S_\theta^2\right] / N.$$

(2.31)

Thus, under the (unrealistic) assumption $P_\theta(U/u \geq 1/2) = 1$, the right-hand side of (2.31) is non-negative and $\text{Var}_\theta(D_I) \geq \text{Var}_\theta(D_{II})$. In practice, (2.31) suggests that if the ratio of the standard deviation of $U$ to the absolute value of the expectation $E_\theta U = u$ is less than, say, $1/4$, then the variance of $D_{II}$ is probably much smaller than the variance of $D_I$; by Chebyshev’s inequality, a more conservative guess is that variance reduction is very likely if the ratio is less than $1/10$. Furthermore, when this ratio is close to 0 and thus $U$ is “almost constant”, the right-hand side of (2.31) is roughly $u^2 \text{Var}_\theta(S_\theta)/N$.

However, while the variance of $D_{II}$ may be considerably smaller than the variance of $D_I$, its variance may not be the minimum variance that can be achieved by using $S_\theta$ as a control variate. The estimator proposed below attains the minimum variance by construction.

**Estimator $D_{III}$: minimum variance LR control variate estimator**

Let $\text{Cov}_\theta[D_C(A)]$ be the variance-covariance matrix of the vector valued LR control variate estimator $D_C(A)$ as given by (2.26), and let $|\text{Cov}_\theta[D_C(A)]|$ be the determinant of $\text{Cov}_\theta[D_C(A)]$, called the generalized variance of $D_C(A)$. 

Lemma 3 (Rubinstein and Marcus, 1985). The value of $A$ that minimizes the generalized variance of $D_C(A)$ is given by

$$B = \Gamma_{21}(\theta) \Gamma_{11}^{-1}(\theta),$$

(2.32)

where

$$\Gamma_{11}(\theta) = E_\theta [S_\theta S'_\theta]$$

(2.33)

is the variance-covariance matrix of $S_\theta$, and

$$\Gamma_{21}(\theta) = E_\theta [\{\text{vec} (US'_\theta) - E_\theta [\text{vec} (US'_\theta)]\} S'_\theta]$$

(2.34)

is the covariance matrix of $\text{vec} (US'_\theta)$ and $S_\theta$.

Proof. See Rubinstein and Marcus (1985). \square

Let $L = 1$ so that $S_\theta, US_\theta, DI$, and $D_C(B)$ are scalars, where $D_C(B)$ is given by (2.26) with $A = B$. Then the variance of $D_C(B)$ is equal to $(1 - \rho^2)\text{Var}_\theta(D_I)$, where $\rho$ is the correlation between $US_\theta$ and $S_\theta$: thus, the higher the absolute correlation between $US_\theta$ and $S_\theta$, the greater is the reduction in variance. A similar argument for $L > 1$ follows from Rubinstein and Marcus (1985).

The matrix $B$ can be estimated by

$$\hat{B} = V_{21} V_{11}^{-1},$$

(2.35)

where $V_{11}$ and $V_{21}$ are Monte Carlo estimators of $\Gamma_{11}(\theta)$ and $\Gamma_{21}(\theta)$, respectively, estimated from the Monte Carlo sample $Z_1, Z_2, \ldots, Z_N$.

Let $D_{III}$ be (2.26) with $A = \hat{B}$:

$$D_{III} = D_C(\hat{B}).$$

(2.36)

The estimator $D_{III}$ is $N$-consistent, but because $A = \hat{B}$ depends on $S_\theta$, $D_{III}$ is not unbiased (Fishman, 1996, p. 279); however, the bias is of order $N^{-1}$ (cf. Cochran, 1977, pp. 198—199).

Comparison: estimators $D_0(\epsilon), D_I, D_{II}$, and $D_{III}$

It is evident from Lemma 2 that $D_I$ and $D_{II}$ are unbiased and $N$-consistent, in contrast to $D_0(\epsilon)$, while $D_{III}$ is an $N$-consistent estimator whose bias is of order $N^{-1}$.

To demonstrate how the LR estimators $D_I, D_{II}$, and $D_{III}$ are interrelated, let $L = 1$. It was argued that $D_{II}$ may be superior to $D_I$ in terms of variance, and
2.4. APPLICATIONS

that if the standard deviation of $U$ is small relative to the absolute value of the expectation $E_\theta U = u$ and hence $U$ is “almost constant”, then $\text{Var}_\theta(D_{II}) \approx \text{Var}_\theta(D_I) - u^2 \text{Var}_\theta(S_\theta)/N$. For large $N$, $\hat{B}$ is expected to be close to $B$ as given by (2.32) and hence the variance of $D_{III}$ is roughly $(1 - \varrho^2) \text{Var}_\theta(D_I)$, where $\varrho$ is the correlation between $US_\theta$ and $S_\theta$. Furthermore, if the standard deviation of $U$ is small relative to $|E_\theta U|$, then (2.33) and (2.34) imply that $\Gamma_{21} \approx u \Gamma_{11}$ and thus $B = \Gamma_{21} \Gamma_{11}^{-1} \approx u$, suggesting that $D_{II} \approx D_{III}$ and that the variance of $D_{II}$ and $D_{III}$ is of a similar order of magnitude. In sum, if the standard deviation of $U$ is small relative to $|E_\theta U|$, then it is thought that $D_{II}$ and $D_{III}$ are close in terms of variance and outperform $D_I$.

An important practical motivation for considering $D_I$, $D_{II}$, and in particular $D_{III}$ as alternative estimators of the Jacobian matrix $\Delta(\theta)$ is that using one of the LR estimators instead of $D_0(\epsilon)$ roughly cuts down the computation time by a factor $L + 1$, where $L$ is the dimension of $\theta$.

2.4 Applications

In the present section, the derivative estimators $D_0(\epsilon)$, $D_I$, $D_{II}$, and $D_{III}$ are compared in a situation where the true Jacobian matrix is known (Section 2.4.1) and in addition in the common situation where the true Jacobian matrix is unknown (Section 2.4.2).

In each subsection, one real-world data set is studied, the moment estimate $\hat{\theta}$ of $\theta$ is obtained, and the Jacobian matrix $\Delta(\hat{\theta})$ is estimated from 1,000 Monte Carlo samples of size $N = 1,000$, where $N$ corresponds to the number of terms on which the derivative estimators—which are averages—are based, implying that for $D_I$, $D_{II}$, and $D_{III}$ the Markov process is $N$ times simulated, while for $D_0(\epsilon)$ the Markov process is $(L + 1) \times N$ times simulated, because each term requires $L + 1$ simulations.

2.4.1 Application: Jacobian matrix known

A simple, classical model where Jacobian matrices can be derived analytically is the Independent Arcs (IA) model (see Snijders and Van Duijn, 1997), which is in most empirical applications inadequate because of its simplicity, but provides an opportunity to compare the simulation-based derivative estimators.

The IA model is a continuous-time Markov model, where the rate of change $q_\theta(x^*, i, j)$ depends only on $x^*_i$, so that all arc variables $x_{ij}(t)$ follow independent
Markov processes. To keep the parametrization consistent with the general model of Section 2.2, the rate of change is written as

$$q_\theta(x^*, i, j) = \theta_1 \frac{(1 - x_{ij}^*) \exp(\theta_2) + x_{ij}^* \exp(-\theta_2)}{n - 1}. \quad (2.37)$$

The rate function $\lambda_i$ follows from (2.3) and (2.37), and is given by

$$\lambda_i(\theta, x^*) = \theta_1 \frac{((n - 1) - x_{i+}^*) \exp(\theta_2) + x_{i+}^* \exp(-\theta_2)}{n - 1}, \quad (2.38)$$

where $x_{i+}^* = \sum_{h \neq i} x_{ih}^*$. The conditional probability mass function $r_i$ is, using (2.4), (2.37), and (2.38), given by

$$r_i(\theta, x, j) = \frac{(1 - x_{ij}^*) \exp(\theta_2) + x_{ij}^* \exp(-\theta_2)}{((n - 1) - x_{i+}^*) \exp(\theta_2) + x_{i+}^* \exp(-\theta_2)}. \quad (2.39)$$

Let

$$M_{kl} = \#\{ (i, j) \mid x_{ij}(t_0) = k, \ x_{ij}(t_1) = l \}, \quad (2.40)$$

where $k, l = 0, 1$. The statistic $U = (M_{01} + M_{10}, M_{01} + M_{11})'$ is a sufficient statistic, and thus, to estimate $\theta = (\theta_1, \theta_2)'$ by the method of moments, $E_\theta U - u$ is a natural estimating function; note that, in the case of the IA model, the sufficiency implies that the moment estimator based on estimating equation $E_\theta U - u = 0$ coincides with the maximum likelihood (ML) estimator (see Snijders and Van Duijn, 1997). The expectation $E_\theta U$, the variance-covariance matrix $\Sigma(\theta) = \text{Cov}_\theta U$, and the Jacobian matrix $\Delta(\theta) = \partial E_\theta U/\partial \theta'$ are derived analytically in Appendix B.

Snijders and Van Duijn (1997) applied the IA model to a well-known data set called the EIES data, corresponding to the “communication” among $n = 32$ scholars observed at two time points, where $x_{ij} = 1$ if scholar $i$ met scholar $j$, and $x_{ij} = 0$ otherwise. The moment (and ML) estimate of $\theta = (\theta_1, \theta_2)'$ is $\hat{\theta} = (2.418, 1.557)'$, the expectation of $U$ is $E_\theta U = (154, 653)'$, and, using the results of Appendix B, the exact values of $\Sigma(\hat{\theta})$ and $\Delta(\hat{\theta})$ are given by

$$\Sigma(\hat{\theta}) = \begin{pmatrix} 108.80 & 95.00 \\ 95.00 & 108.80 \end{pmatrix} \quad \text{and} \quad \Delta(\hat{\theta}) = \begin{pmatrix} 52.18 & 114.55 \\ 47.44 & 130.85 \end{pmatrix}. \quad (2.41)$$

The Jacobian matrix $\Delta(\hat{\theta})$ is estimated from each Monte Carlo sample by $D_0(\epsilon)$ with $\epsilon = 0.2$ and $\epsilon = 1.0$; the two values of $\epsilon$ are motivated by the fact that most values of $\epsilon$ used in practice are between .2 and 1.0. In addition, $\Delta(\hat{\theta})$ is estimated from
2.4. APPLICATIONS

Each Monte Carlo sample by $D_I$, $D_{II}$, and $D_{III}$; the complete-data efficient score, on which $D_I$, $D_{II}$, and $D_{III}$ are based, is derived in Appendix A.

The computation time required to evaluate $D_0(2)$ and $D_0(1.0)$ was on average approximately 20 seconds on a PC with Intel Pentium 3.06 GHz processor and 1021 MB RAM. While 20 seconds are negligible, note that in general the computation time is roughly proportional to $N \times L^2 \times C_1 \times C_2$, where $C_1 = \sum_{g=1}^{G} \sum_{i=1}^{n} \sum_{j \neq i}^{n} |x_{ij}(t_g) - x_{ij}(t_{g-1})|$ and $C_2 = (1/G) \sum_{g=0}^{G-1} \sum_{i=1}^{n} \sum_{j \neq i}^{n} x_{ij}(t_g)$. Here, $L = 2$ and $n = 32$, which are very small values: in practice, it is frequently the case that $L > 10$ and $n > 50$ (in fact, $n$ may be in the hundreds), and then computation time is an issue and the computational advantage of $D_I$, $D_{II}$, and $D_{III}$ over $D_0(\epsilon)$ is appreciated.

The average estimates of the Jacobian matrix $\Delta(\hat{\theta})$ are presented in Table 2.1 (p. 25). The average estimate of $\Delta(\hat{\theta})$ based on the biased estimator $D_0(2)$ is close to the true value for the first column, but overestimates the elements of the second column by 7.4% and 5.3%, respectively; the average estimate of $D_0(1.0)$ underestimates the elements of the first column by roughly 7.6%, and overestimates the elements of the second column by 31.8% and 23.1%, respectively. While the bias of $D_0(2)$ may be tolerable, the bias of $D_0(1.0)$ clearly is not. The average estimates of $\Delta(\hat{\theta})$ based on $D_{II}$ and $D_{III}$ are very close to the true value of $\Delta(\hat{\theta})$, but $D_I$, which is known to be unbiased, seems to overestimate the elements of the second row of $\Delta(\hat{\theta})$; this is an inaccuracy stemming from the huge variance of $D_I$ (see below).

The Monte Carlo standard deviations (MC S.D.s) of the estimates of $\Delta(\hat{\theta})$ are shown in Table 2.2 (p. 26). Note that the efficiency of the estimators cannot be evaluated by inspecting the MC S.D.s alone, because $D_0(\epsilon)$ requires $(L + 1) \times N = 3,000$ simulations, while $D_I$, $D_{II}$, and $D_{III}$ require $N = 1,000$ simulations; the efficiency of the estimators is discussed in Section 6. The MC S.D.s of $D_{II}$ and $D_{III}$

<table>
<thead>
<tr>
<th>true value</th>
<th>$D_0(2)$</th>
<th>$D_0(1.0)$</th>
<th>$D_I$</th>
<th>$D_{II}$</th>
<th>$D_{III}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_{11}$</td>
<td>52.18</td>
<td>51.34</td>
<td>48.22</td>
<td>52.60</td>
<td>51.92</td>
</tr>
<tr>
<td>$\delta_{21}$</td>
<td>47.44</td>
<td>46.67</td>
<td>43.83</td>
<td>49.72</td>
<td>47.30</td>
</tr>
<tr>
<td>$\delta_{12}$</td>
<td>114.55</td>
<td>122.98</td>
<td>151.00</td>
<td>115.42</td>
<td>114.29</td>
</tr>
<tr>
<td>$\delta_{22}$</td>
<td>130.85</td>
<td>137.76</td>
<td>161.08</td>
<td>135.43</td>
<td>130.67</td>
</tr>
</tbody>
</table>

Table 2.1: EIES data: average estimates of $\Delta(\hat{\theta}) = (\delta_{ij})$ across 1,000 Monte Carlo samples
CHAPTER 2. ESTIMATING FUNCTIONS: DERIVATIVE ESTIMATION

Table 2.2: EIES data: Monte Carlo standard deviations of estimates of 
\[ \Delta(\hat{\theta}) = (\delta_{ij}) \] 
based on 1,000 Monte Carlo samples

<table>
<thead>
<tr>
<th></th>
<th>(D_0(.2))</th>
<th>(D_0(1.0))</th>
<th>(D_I)</th>
<th>(D_{II})</th>
<th>(D_{III})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\delta_{11})</td>
<td>.53</td>
<td>.21</td>
<td>26.36</td>
<td>2.39</td>
<td>2.40</td>
</tr>
<tr>
<td>(\delta_{21})</td>
<td>.53</td>
<td>.22</td>
<td>111.36</td>
<td>2.28</td>
<td>2.28</td>
</tr>
<tr>
<td>(\delta_{12})</td>
<td>.87</td>
<td>.32</td>
<td>62.27</td>
<td>5.53</td>
<td>5.54</td>
</tr>
<tr>
<td>(\delta_{22})</td>
<td>1.01</td>
<td>.33</td>
<td>262.31</td>
<td>5.84</td>
<td>5.87</td>
</tr>
</tbody>
</table>

for given \(N\) seem to be of a similar order of magnitude, whereas \(D_0(.2)\) and \(D_0(1.0)\) seem to do considerably better and \(D_I\) considerably worse.

The MC S.D.s of the elements of \(D_0(.2)\) are 2.4 to 3.1 times as large as the corresponding MC S.D.s of \(D_0(1.0)\), which is in line with the fact that the variance of \(D_0(\epsilon)\) is at least of order \(\epsilon^{-1}\) (see Lemma 1), and if independent random numbers are used, it is of order \(\epsilon^{-2}\); hence the ratio of MC standard deviations of \(D_0(.2)\) to \(D_0(1.0)\) is expected to be between \(\sqrt{5} = 2.24\) and 5, and because common random numbers are used, closer to 2.24 than to 5, which is indeed the case.

Concerning \(D_I\), note that \(D_I\) estimates the first row of \(\Delta(\hat{\theta})\) much more accurately than the second row, which is not surprising. By (2.41), the standard deviations of coordinates \(U_1\) and \(U_2\) of \(U\) both equal 10.43, while the expectation of \(U\) is given by \(E_{\hat{\theta}}U = (154, 653)'\); therefore, the standard deviations are small relative to the expectations and \(U\) can be considered to be "almost constant" for practical purposes. Then \(\text{Var}_{\hat{\theta}}(D_I) = \text{Var}_{\hat{\theta}}(US_{\hat{\theta}})/N \approx u^2\text{Var}_{\hat{\theta}}(S_{\hat{\theta}})/N\) (in case \(L = 1\), with obvious extension to \(L > 1\)), and the ratio of the MC S.D.s of the second row to the first row is expected to be roughly 653/154 = 4.24, which is indeed the case.

Concerning the LR control variate estimators \(D_{II}\) and \(D_{III}\), it is obvious that the introduction of the complete-data score as a control variate reduces the variance considerably as compared to \(D_I\). The small advantage of \(D_{II}\) over \(D_{III}\) can be explained as follows. Since the standard deviations of the coordinates of \(U\) are small relative to the expectations, \(\hat{B} \approx u\) and \(D_{II} \approx D_{III}\) as argued in Section 2.3.2, and therefore it is not surprising that the MC S.D.s of \(D_{II}\) and \(D_{III}\) are of a similar order of magnitude. The fact that \(D_{II}\) seems to outperform \(D_{III}\) slightly may be due to the additional sampling variance resulting from the estimation of \(B\) given by (2.32).

As was pointed out above, the ultimate aim of the derivative estimators is to
2.4. APPLICATIONS

Figure 2.1: EIES data: kernel density plots of Monte Carlo estimates of s.e.($\hat{\theta}_1$) and s.e.($\hat{\theta}_2$) based on 1,000 Monte Carlo samples

The exact values of the standard errors, s.e.($\hat{\theta}_1$) = .477 and s.e.($\hat{\theta}_2$) = .191, are represented by vertical lines.

produce Monte Carlo estimates of

$$\Delta^{-1}(\hat{\theta}) \Sigma(\hat{\theta}) [\Delta^{-1}(\hat{\theta})]'$$

(2.42)

which is an approximation of the variance-covariance matrix of $\hat{\theta}$ (see (2.10)). For the IA model, (2.42) can be evaluated analytically by using (2.41). The resulting standard errors of $\hat{\theta}_1$ and $\hat{\theta}_2$ are given by s.e.($\hat{\theta}_1$) = .477 and s.e.($\hat{\theta}_2$) = .191, respectively; these values will be referred to as the exact standard errors, exact in the sense that they are based on analytical evaluation of (2.42) and not estimated from Monte Carlo simulations. It is of interest to study the behavior of the Monte Carlo estimators of the standard errors, which are obtained by plugging in estimators $D_0(.2)$, $D_0(1.0)$, $D_I$, $D_{II}$, and $D_{III}$ for $\Delta(\hat{\theta})$ and a Monte Carlo estimator for $\Sigma(\hat{\theta})$. Figure 2.1 (p. 27) shows Gaussian kernel density plots of the Monte Carlo estimates of the standard errors based on $D_0(.2)$, $D_0(1.0)$, $D_{II}$, and $D_{III}$; the plots corresponding to $D_I$ are omitted because the estimated standard errors have huge MC S.D.s. All distributions are fairly symmetric, but the MC S.D.s of the standard errors are larger for $D_0(.2)$ and $D_0(1.0)$ than for the LR estimators $D_{II}$ and $D_{III}$. The estimators of the standard errors based on $D_0(.2)$ and $D_0(1.0)$ seem to be upwards biased, in particular $D_0(1.0)$ leads to a large bias in estimated standard errors. In contrast, the standard errors
based on \(D_{II}\) and \(D_{III}\) seem to be (almost) unbiased.

As a side remark, Snijders and Van Duijn (1997) report for the IA model applied to the EIES data standard errors \(\text{s.e.}(\hat{\theta}_1) = .22\) and \(\text{s.e.}(\hat{\theta}_2) = .24\), which are Monte Carlo estimates based on \(D_0(\epsilon)\); the value of \(\epsilon\) is not reported. Thus the exact standard error \(\text{s.e.}(\hat{\theta}_1) = .477\) is underestimated by 53.9%, which can be explained by (a) bias (if \(\epsilon\) was large) or (b) large MC S.D. (if \(\epsilon\) was small) or both; (a) is not too plausible, because Figure 2.1 suggests an upwards bias; however, no matter what explanation is applicable here, the case illustrates that choosing \(\epsilon\) is hard and can have great practical implications.

### 2.4.2 Application: Jacobian matrix unknown

In the present section, the derivative estimators \(D_0(2), D_0(1.0), D_I, D_{II},\) and \(D_{III}\) are compared in the common situation where the Jacobian matrix is unknown.

Snijders (2001) studied data collected by Van de Bunt (1999), concerning a friendship relation among \(n = 32\) university freshmen enrolled in a common study program. The digraph was observed at 7 time points. Here, the digraph evolution between observation points \(t_2 < t_3 < t_4\) is modeled.

A simple model is specified by constant rate functions and objective function

\[
 f_i(\beta; x, j) = \sum_{k=1}^{4} \beta_k s_{ik}(x, j),
\]

where the statistics \(s_{ik}\) are given by:

\[
 s_{i1}(x, j) = \sum_{h=1}^{n} x_{ih}: \text{the number of outgoing arcs ("outdegree"),}
\]

\[
 s_{i2}(x, j) = \sum_{h=1}^{n} x_{ih}x_{hi}: \text{the number of reciprocated arcs,}
\]

\[
 s_{i3}(x, j) = \sum_{h=1}^{n} (1 - x_{ih}) \max_l x_{il}x_{lh}: \text{the number of indirect connections,}
\]

\[
 s_{i4}(x, j) = c_i s_{i1}(x, j): \text{interaction of outdegree and gender of student } i, \text{ where } s_{i1} \text{ is the outdegree of } i \text{ and } c_i = 1 \text{ if } i \text{ is male and } 0 \text{ otherwise.}
\]

Conditioning, as described in Snijders (2001), on the observed number of changes, which is 60 in \([t_2, t_3]\) and 51 in \([t_3, t_4]\), the parameter \(\theta\) to be estimated by the method of moments reduces to \(\theta = (\beta_1, \beta_2, \beta_3, \beta_4)'\). The coordinate \(U_k\) of statistic vector \(U\), corresponding to coordinate \(\theta_k\) of \(\theta\), is given by \(U_k = \sum_{i=1}^{n} s_{ik} (k = 1, \ldots, 4)\). The moment estimate of \(\theta\) turns out to be \(\hat{\theta} = (-1.058, 2.507, -.535, -.562)'\).
Table 2.3: Van de Bunt data: average estimates of $\Delta(\hat{\theta})$ across 1,000 Monte Carlo samples

<table>
<thead>
<tr>
<th></th>
<th>$D_0(.2)$</th>
<th>$D_0(1.0)$</th>
<th></th>
<th>$D_0(1.0)$</th>
<th>$D_0(1.0)$</th>
<th></th>
<th>$D_0(1.0)$</th>
<th>$D_0(1.0)$</th>
<th></th>
<th>$D_0(1.0)$</th>
<th>$D_0(1.0)$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>78.96 23.93 17.14 2.84</td>
<td>51.86 19.26 22.85 .85</td>
<td></td>
<td>28.88 30.33 37.48 6.23</td>
<td>16.00 26.36 40.10 2.13</td>
<td></td>
<td>150.35 64.81 270.23 15.64</td>
<td>91.13 57.63 342.82 7.79</td>
<td></td>
<td>3.41 4.53 5.33 17.82</td>
<td>4.82 3.65 4.97 17.52</td>
<td></td>
</tr>
<tr>
<td>$D_I$</td>
<td>88.38 25.63 -6.80 2.31</td>
<td>86.74 25.04 7.26 3.24</td>
<td></td>
<td>34.30 31.36 22.27 6.75</td>
<td>33.52 31.08 28.94 7.20</td>
<td></td>
<td>172.60 67.40 203.58 15.29</td>
<td>170.02 66.47 225.66 16.76</td>
<td></td>
<td>2.85 4.60 6.53 17.61</td>
<td>2.93 4.63 5.83 17.57</td>
<td></td>
</tr>
<tr>
<td>$D_{II}$</td>
<td>86.55 24.98 7.27 3.23</td>
<td>33.44 31.02 28.88 7.18</td>
<td></td>
<td>169.65 66.31 225.25 16.70</td>
<td>169.65 66.31 225.25 16.70</td>
<td></td>
<td>2.93 4.62 \ 5.80 17.54</td>
<td>2.93 4.62 \ 5.80 17.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The average estimates of the Jacobian matrix $\Delta(\hat{\theta})$ are presented in Table 2.3 (p. 29). As was pointed out above, estimators $D_0(.2)$ and $D_0(1.0)$ are biased, whereas $D_I$ and $D_{II}$ are unbiased and $D_{III}$ is approximately unbiased. An unbiased estimator of the bias of $D_0(.2)$ is the average of $D_0(.2) - D_{II}$ across Monte Carlo samples; the bias of $D_0(1.0)$ can be estimated accordingly. The bias of $D_0(1.0)$ seems to be large; the bias of $D_0(.2)$ also is non-negligible. The average estimates of $D_{II}$ and $D_{III}$ agree closely, while the average estimate of $D_I$ differs slightly from $D_{II}$ and $D_{III}$.

The Monte Carlo standard deviations (MC S.D.s) of the estimates of $\Delta(\hat{\theta})$ are shown in Table 2.4 (p. 30); to save space, attention is restricted to the diagonal elements of $\Delta(\hat{\theta})$. Note that $D_0(\epsilon)$ requires $(L+1) \times N = 5,000$ simulations, while $D_I$, $D_{II}$, and $D_{III}$ require $N = 1,000$ simulations, so that the efficiency of the estimators cannot be evaluated on the basis of the MC S.D.s alone; see Section 6. The MC S.D.s of $D_{II}$ and $D_{III}$ for given $N$ seem to be of a similar order of magnitude, whereas $D_0(.2)$ and $D_0(1.0)$ seem to do considerably better and $D_I$ considerably worse. Once again, the introduction of the complete-data score as a control variate reduces the
Table 2.4: Van de Bunt data: Monte Carlo standard deviations of the estimated diagonal elements $\delta_{ii}$ of $\Delta(\hat{\theta})$ based on 1,000 Monte Carlo samples

<table>
<thead>
<tr>
<th></th>
<th>$D_0(.2)$</th>
<th>$D_0(1.0)$</th>
<th>$D_I$</th>
<th>$D_{II}$</th>
<th>$D_{III}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_{11}$</td>
<td>1.55</td>
<td>.32</td>
<td>108.17</td>
<td>4.69</td>
<td>4.70</td>
</tr>
<tr>
<td>$\delta_{22}$</td>
<td>1.54</td>
<td>.32</td>
<td>24.67</td>
<td>1.96</td>
<td>1.97</td>
</tr>
<tr>
<td>$\delta_{33}$</td>
<td>5.63</td>
<td>1.43</td>
<td>267.89</td>
<td>19.84</td>
<td>19.95</td>
</tr>
<tr>
<td>$\delta_{44}$</td>
<td>.73</td>
<td>.16</td>
<td>2.46</td>
<td>.94</td>
<td>.95</td>
</tr>
</tbody>
</table>

variance of the LR estimator greatly.

It is of practical interest to assess how the methods perform in terms of the standard errors of the coordinates $\hat{\theta}_k$ of $\hat{\theta}$. Figure 2.2 (p. 31) shows Gaussian kernel density plots of the estimated standard errors s.e.$(\hat{\theta}_k) \ (k = 1, \ldots, 4)$ based on $D_0(.2)$, $D_0(1.0)$, $D_{II}$, and $D_{III}$; the estimates based on $D_I$ have huge MC S.D.s and are omitted. The distributions of estimated standard errors based on $D_{II}$ and $D_{III}$ appear to be very similar; $D_0(.2)$, in turn, produces distributions that are similar to the ones based on $D_{II}$ and $D_{III}$, apart from the distribution of estimates of s.e.$(\hat{\theta}_3)$ which has a smaller mean and a smaller MC S.D. than the corresponding distributions based on $D_{II}$ and $D_{III}$. The estimator $D_0(1.0)$ gives rise to distributions that deviate more (see s.e.$(\hat{\theta}_1)$ and s.e.$(\hat{\theta}_3)$) or less (see s.e.$(\hat{\theta}_2)$ and s.e.$(\hat{\theta}_4)$) from the distributions produced by the other three methods; the estimated standard errors based on $D_0(1.0)$ seem to be biased.

Table 2.5 (p. 32) shows the MC S.D.s of the standard errors s.e.$(\hat{\theta}_k) \ (k = 1, \ldots, 4)$ for Monte Carlo samples of size $N = 100, 200, 500, \text{ and } 1,000$; the Monte Carlo samples of size $N = 100, 200, \text{ and } 500$ are obtained by taking the first 100, 200, and 500 observations of each Monte Carlo sample of size $N = 1,000$, respectively; once again, the standard errors based on $D_I$ are omitted. Table 2.5 indicates that the the performance gap between $D_0(.2)$ on one hand and $D_{II}$ and $D_{III}$ on the other hand is relatively small, leaving aside s.e.$(\hat{\theta}_3)$. The MC standard deviations are roughly proportional to $N^{-1/2}$. However, it should be noted that, when using $D_0(.2)$ and in particular $D_{II}$ and $D_{III}$ and when $N$ is small ($N = 100 \text{ or } N = 200$), increasing $N$ by factor $c$ clearly reduces the MC S.D.s by more than $c^{1/2}$.

A minimum requirement is that the MC S.D. should be less than 5% of the estimand to be useful in practice; the average estimated standard errors s.e.$(\hat{\theta}_k) \ (k = 1, \ldots, 4)$ based on $D_{II}$ are .136, .363, .118, and .263, respectively. According to Table
2.5. DISCUSSION

Figure 2.2: Van de Bunt data: kernel density plots of Monte Carlo estimates of s.e.(\(\hat{\theta}_k\)) \((k = 1, \ldots, 4)\) based on 1,000 Monte Carlo samples

2.5, for \(N = 1,000\) each of the estimators \(D_0(.2), D_{II},\) and \(D_{III}\) meets the 5%-standard for two standard errors, and “almost” meets it for the other two standard errors; therefore, it is sensible to slightly increase the sample size \(N.\)

2.5 Discussion

Three likelihood ratio / score function (LR) estimators of the Jacobian matrix of the estimating function, \(D_I, D_{II},\) and \(D_{III},\) were proposed, and compared with the conventional estimator \(D_0(\epsilon)\) based on finite differences.

Based on theoretical and empirical evidence, it is safe to say that the finite differences estimator \(D_0(\epsilon)\) should be used with much care; in fact, the difficult choice of \(\epsilon\) and the associated bias-variance dilemma, together with the computational disadvantage, are strong arguments against using \(D_0(\epsilon).\)

Concerning the LR estimators, the huge variance of LR estimator \(D_I\) renders \(D_I\) useless for practical applications. The efficiency of the LR control variate estimators \(D_{II}\) and \(D_{III}\) relative to \(D_0(\epsilon)\) can be evaluated by the classical efficiency ratio of Hammersley and Handscomb (1964, p. 51), which can be written as

\[
\frac{\text{Var}_\theta[D_0(\epsilon)]}{\text{Var}_\theta[D_{II} \text{ or } D_{III}]} \times \frac{\text{time}(D_0(\epsilon))}{\text{time}(D_{II} \text{ or } D_{III})} = \frac{\text{Var}_\theta[D_0(\epsilon)]}{\text{Var}_\theta[D_{II} \text{ or } D_{III}]} \times (L + 1), \quad (2.44)
\]
where $\text{Var}_\theta[D]$ refers to some element of derivative estimator $D$—that is, to some partial derivative—and “time($D$)” refers to the amount of computation time required to evaluate $D$. It is evident that the efficiency ratio tends to favor $D_{II}$ and $D_{III}$ if $L$ is moderate or large. In Section 2.4.1, where the simple Independent Arcs model with $L = 2$ parameters was considered, the efficiency ratio of $D_{II}$ relative to $D_0(.2)$ turns out to be .1 for each of the two elements on the main diagonal of the Jacobian matrix. In Section 2.4.2, where $L = 4$, the efficiency ratio of $D_{II}$ relative to $D_0(.2)$ is .5, 3.1, .4, and 3.0 for the four elements on the main diagonal of the Jacobian matrix. However, in practice $L$ is frequently larger than 10, which tends to favor $D_{II}$ and $D_{III}$.

Overall, the conclusion is that using $D_{II}$ or $D_{III}$ is preferable to using $D_0(\epsilon)$.

An alternative approach to derivative estimation is to combine the estimation of $\theta$ and $\Delta(\theta) = \partial E_\theta U/\partial \theta'$ as follows: if the interim estimate $\hat{\theta}_N$ generated by the stochastic approximation algorithm for solving estimating equation (2.9) is in a small neighborhood of the solution $\hat{\theta}$, then the simulations from the distributions corre-

<table>
<thead>
<tr>
<th>$N$</th>
<th>s.e.$(\hat{\theta}_1)$</th>
<th>using $D_0(.2)$</th>
<th>using $D_0(1.0)$</th>
<th>using $D_{II}$</th>
<th>using $D_{III}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.0188</td>
<td>0.0143</td>
<td>0.0257</td>
<td>0.0293</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0813</td>
<td>0.0322</td>
<td>0.0864</td>
<td>0.0925</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0134</td>
<td>0.0061</td>
<td>0.0392</td>
<td>0.0408</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0409</td>
<td>0.0191</td>
<td>0.0454</td>
<td>0.0508</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.0122</td>
<td>0.0100</td>
<td>0.0155</td>
<td>0.0162</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0514</td>
<td>0.0221</td>
<td>0.0505</td>
<td>0.0534</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0084</td>
<td>0.0044</td>
<td>0.0203</td>
<td>0.0206</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0272</td>
<td>0.0136</td>
<td>0.0265</td>
<td>0.0279</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.0074</td>
<td>0.0065</td>
<td>0.0090</td>
<td>0.0091</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0296</td>
<td>0.0138</td>
<td>0.0289</td>
<td>0.0294</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0051</td>
<td>0.0028</td>
<td>0.0119</td>
<td>0.0120</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0167</td>
<td>0.0085</td>
<td>0.0156</td>
<td>0.0160</td>
<td></td>
</tr>
<tr>
<td>1,000</td>
<td>0.0052</td>
<td>0.0046</td>
<td>0.0062</td>
<td>0.0062</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0207</td>
<td>0.0099</td>
<td>0.0208</td>
<td>0.0210</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0036</td>
<td>0.0020</td>
<td>0.0080</td>
<td>0.0081</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0114</td>
<td>0.0060</td>
<td>0.0108</td>
<td>0.0110</td>
<td></td>
</tr>
</tbody>
</table>
sponding to the interim estimates $\hat{\theta}_N$ can be used to estimate the surface of $E_{\theta}U$ as a function of $\theta$, and to produce derivative estimates of $\Delta(\hat{\theta})$ based on fitting a linear model. It will be useful to investigate the computational benefits obtainable by such an approach.

The proposed estimators are implemented in the Windows-based computer program Siena embedded in the program collection StOCNET, which can be downloaded free of charge from http://stat.gamma.rug.nl/stocnet.

Appendix A: Complete-data efficient score

Let $M$ be the total number of changes of $x(t)$ in time interval $[t_0, t_1]$, and let $x_0$ be the digraph $x(t_0)$ observed at time point $t_0$. The Markov process corresponds to holding times $h_1, \ldots, h_M$ and some sequence $x_0, x_1, \ldots, x_M$ of digraphs, where $x_m$ is the digraph to which the process moves by the $m$-th transition, which takes place at time $t_0 + \sum_{i=1}^m h_i$. Conditional on $x(t_0)$, the complete-data probability density $p_\theta = p_\theta(z)$ based on an observed outcome of the Markov process is given by

\[
\prod_{m=1}^M \left[ q_\theta(x_{m-1}) \exp \left( -q_\theta(x_{m-1}) h_m \right) \frac{q_\theta(x_{m-1}, x_m)}{q_\theta(x_{m-1})} \right] \times \exp \left( -q_\theta(x_M) h_{M+1} \right),
\]

where $h_{M+1} = (t_1 - t_0) - \sum_{m=1}^M h_m > 0$. The parameter $q_\theta(x_{m-1})$ of the negative exponential distribution is given by

\[
q_\theta(x_{m-1}) = -q_\theta(x_{m-1}, x_{m-1}) = \sum_{j=1}^n \lambda_j(\theta, x_{m-1}),
\]

while $q_\theta(x_{m-1}, x_m)$ is decomposed according to

\[
q_\theta(x_{m-1}, x_m) = \lambda_{i_m}(\theta, x_{m-1}) r_{i_m}(\theta, x_m, j_m),
\]

where the rate function $\lambda_{i_m}$ and the conditional probability mass function $r_{i_m}$ are given by (2.6) and (2.7), respectively.

Section A.1 gives the complete-data efficient score with respect to $\theta = (\rho, \alpha', \beta')'$. Section A.2 considers some more general models and the associated complete-data efficient score, and Section A.3 considers the IA model and the associated complete-data efficient score.
A.1 Complete-data efficient score with respect to $\theta = (\rho, \alpha', \beta')'$

If $\eta$ represents either $\rho$ or an element of $\alpha = (\alpha_k)$, then the complete-data score corresponding to $\eta$ can be written as

$$
\sum_{m=1}^{M} \left[ \frac{\partial \ln \lambda_i(\theta, x_{m-1})}{\partial \eta} - \frac{\partial q_\theta(x_{m-1})}{\partial \eta} h_m \right] - \frac{\partial q_\theta(x_M)}{\partial \eta} h_{M+1},
$$

(2.48)

where

$$
\frac{\partial \ln \lambda_i(\theta, x_{m-1})}{\partial \rho} = \frac{1}{\rho},
$$

(2.49)

$$
\frac{\partial q_\theta(x_{m-1})}{\partial \rho} = \frac{q_\theta(x_{m-1})}{\rho},
$$

(2.50)

$$
\frac{\partial \ln \lambda_i(\theta, x_{m-1})}{\partial \alpha_k} = a_{i,m,k}(x_{m-1}, c_{i,m}),
$$

(2.51)

and

$$
\frac{\partial q_\theta(x_{m-1})}{\partial \alpha_k} = \sum_{j=1}^{n} a_{jk}(x_{m-1}, c_j) \lambda_j(\theta, x_{m-1}).
$$

(2.52)

The complete-data score with respect to $\beta_k$ is given by

$$
\sum_{m=1}^{M} \left[ s_{im,k}(x_m, j_m) - \sum_{j=1}^{n} s_{im,k}(x_m, j) r_{im}(\theta, x_m, j) \right].
$$

(2.53)

A.2 More general models

The present section discusses briefly how the complete-data efficient score for some selected, more general models deviates from the simple case above.

A.2.1 $G > 1$ time intervals

If the digraph is observed at more than two time points, that is, $G > 1$, then the complete-data score corresponding to parameter coordinates that are constant across time intervals is obtained by simply summing the complete-data score across the time intervals.
A.2.2 “Thinning”

 Modeling social science network data sometimes makes it desirable to give nodes (“actors”), when designated to change something, the freedom not to change anything. In formal terms, the additional freedom is represented by replacing the constraint

\[ \sum_{j \neq i} r_i(\theta, x_m, j) = 1 \] (2.54)

by the constraint

\[ \sum_{j \neq i} r_i(\theta, x_m, j) \leq 1 \] (2.55)

so that

\[ r_i(\theta, x_m, i) = 1 - \sum_{j \neq i} r_i(\theta, x_m, j) \geq 0. \] (2.56)

Therefore, in step (2) of the simulation method described in Remark 2.2.1 of Section 2.2, it is admissible that \( j = i \), and if \( j = i \), then \( x^{(M)} = x^{(M-1)} \). The parameter of the negative exponential distribution after thinning the Markov process (that is, after omitting the “changes” which do not lead to a change of state because \( j = i \)) is given by

\[ q_{\theta}(x_{m-1}) = \sum_{j=1}^{n} \lambda_j(\theta, x_{m-1}) \sum_{h \neq j}^{n} r_j(\theta, x_m, h), \] (2.57)

which leads to inconvenient derivatives due to the dependence of \( q_{\theta}(x_{m-1}) \) on \( \beta \) through the conditional probabilities \( r_j \). It is therefore more appealing to consider the Markov process before thinning. The parameter of the negative exponential distribution before thinning is

\[ q_{\theta}(x_{m-1}) = \sum_{j=1}^{n} \lambda_j(\theta, x_{m-1}). \] (2.58)

Let the complete data correspond to all the holding times, the events which do not result in change, and the events which do result in change. The complete-data likelihood then is proportional to the joint probability of the complete sequence of events, where an event may or may not lead to some change. The same formulae can be used as in Section A.1.
A.2.3 Co-evolution of digraphs and other outcome variables

The model can be extended to include, in addition to the Markov process that shapes the digraph, continuous-time Markov processes that shape other outcome variables; see Snijders, Steglich, and Schweinberger (2006). It is beyond the scope of the present paper to describe such models in detail, but it should be noted that convenient parametrizations lead to simple derivatives.

A.3 Independent Arcs model

Formula (2.45) concerning the complete-data probability density is valid, but, using (2.46) and (2.38),

\[
q_\theta(x_{m-1}) = \sum_{j=1}^{n} \lambda_j(\theta, x_{m-1}) = \theta_1 \left( n(n-1) - x^{(m-1)}_{++} \right) \exp(\theta_2) + x^{(m-1)}_{++} \exp(-\theta_2),
\]

and, by (2.37),

\[
q_\theta(x_{m-1}, x_m) = \theta_1 \left( 1 - x^{(m-1)}_{im,jm} \right) \exp(\theta_2) + x^{(m-1)}_{im,jm} \exp(-\theta_2),
\]

where \(x^{(m-1)}_{im,jm}\) is the arc variable of digraph \(x_{m-1}\) that is changed by the \(m\)-th move of the Markov process, and \(x^{(m-1)}_{++} = \sum_{i=1}^{n} \sum_{h \neq i} x^{(m-1)}_{ih}\).

The complete-data score with respect to \(\theta_k\) \((k = 1, 2)\) can be written as

\[
\sum_{m=1}^{M} \left[ \frac{\partial \ln q_\theta(x_{m-1}, x_m)}{\partial \theta_k} - \frac{\partial q_\theta(x_{m-1})}{\partial \theta_k} h_m \right] - \frac{\partial q_\theta(x_M)}{\partial \theta_k} h_{M+1},
\]

where

\[
\frac{\partial \ln q_\theta(x_{m-1}, x_m)}{\partial \theta_1} = \frac{1}{\theta_1},
\]

\[
\frac{\partial q_\theta(x_{m-1})}{\partial \theta_1} = \frac{q_\theta(x_{m-1})}{\theta_1},
\]

\[
\frac{\partial \ln q_\theta(x_{m-1}, x_m)}{\partial \theta_2} = \frac{(1 - x^{(m-1)}_{im,jm}) \exp(\theta_2) - x^{(m-1)}_{im,jm} \exp(-\theta_2)}{(1 - x^{(m-1)}_{im,jm}) \exp(\theta_2) + x^{(m-1)}_{im,jm} \exp(-\theta_2)},
\]

and

\[
\frac{\partial q_\theta(x_{m-1})}{\partial \theta_2} = \theta_1 \left( n(n-1) - x^{(m-1)}_{++} \right) \exp(\theta_2) - x^{(m-1)}_{++} \exp(-\theta_2).\]
Appendix B: Independent Arcs model: Jacobian matrix

Let $M_{0+} = M_{00} + M_{01}$ and $M_{1+} = M_{10} + M_{11}$, where $M_{kl}$ is defined by (2.40). It can be shown (see Snijders and Van Duijn, 1997) that $M_{01}$ and $M_{11}$ are independent random variables with distribution

$$M_{01} \sim \text{Binomial}(M_{0+}, \xi_0(T))$$  \hspace{1cm} (2.66)

and

$$M_{11} \sim \text{Binomial}(M_{1+}, \xi_1(T)).$$  \hspace{1cm} (2.67)

The parameters $\xi_0(T)$ and $\xi_1(T)$ of the binomial distributions are given by

$$\xi_0(T) = p(\theta_2) \left(1 - \exp[-r(\theta_1, T) q(\theta_2)]\right)$$  \hspace{1cm} (2.68)

and

$$\xi_1(T) = p(\theta_2) + (1 - p(\theta_2)) \exp[-r(\theta_1, T) q(\theta_2)],$$  \hspace{1cm} (2.69)

where $T = t_1 - t_0$,

$$p(\theta_2) = \frac{\exp(\theta_2)}{\exp(\theta_2) + \exp(-\theta_2)},$$  \hspace{1cm} (2.70)

$$q(\theta_2) = \exp(\theta_2) + \exp(-\theta_2),$$  \hspace{1cm} (2.71)

and

$$r(\theta_1, T) = \frac{\theta_1 T}{n - 1}.$$  \hspace{1cm} (2.72)

Due to the conditioning on digraph $x(t_0)$ observed at time point $t_0$, $M_{0+}$ and $M_{1+}$ are known constants. Thus, for given $x(t_0)$ and $\theta = (\theta_1, \theta_2)'$, the expectation $E_\theta U$ is known,

$$E_\theta U = E_\theta \begin{pmatrix} M_{01} + M_{10} \\ M_{01} + M_{11} \end{pmatrix} = \begin{pmatrix} M_{0+} \xi_0(T) + M_{1+} (1 - \xi_1(T)) \\ M_{0+} \xi_0(T) + M_{1+} \xi_1(T) \end{pmatrix},$$  \hspace{1cm} (2.73)

the variance-covariance matrix $\Sigma(\theta) = \text{Cov}_\theta U$ is given by

$$\Sigma(\theta) = \begin{pmatrix} \text{Var}_\theta M_{01} + \text{Var}_\theta M_{11} & \text{Var}_\theta M_{01} - \text{Var}_\theta M_{11} \\ \text{Var}_\theta M_{01} - \text{Var}_\theta M_{11} & \text{Var}_\theta M_{01} + \text{Var}_\theta M_{11} \end{pmatrix},$$  \hspace{1cm} (2.74)
and the Jacobian matrix $\Delta(\theta) = \partial E_\theta U/\partial \theta'$ with respect to $\theta' = (\theta_1, \theta_2)$ is given by

$$
\Delta(\theta) = \begin{pmatrix}
M_0 + \frac{\partial \xi_0(T)}{\partial \theta_1} & M_0 + \frac{\partial \xi_0(T)}{\partial \theta_2} \\
M_0 + \frac{\partial \xi_0(T)}{\partial \theta_1} & M_0 + \frac{\partial \xi_0(T)}{\partial \theta_2}
\end{pmatrix}
\begin{pmatrix}
M_1 + \frac{\partial \xi_1(T)}{\partial \theta_1} & M_1 + \frac{\partial \xi_1(T)}{\partial \theta_2} \\
M_1 + \frac{\partial \xi_1(T)}{\partial \theta_1} & M_1 + \frac{\partial \xi_1(T)}{\partial \theta_2}
\end{pmatrix},
$$
(2.75)

where

$$
\frac{\partial \xi_0(T)}{\partial \theta_1} = p(\theta_2) q(\theta_2) \exp[-r(\theta_1, T) q(\theta_2)] \frac{T}{n - 1},
$$
(2.76)

$$
\frac{\partial \xi_1(T)}{\partial \theta_1} = -(1 - p(\theta_2)) q(\theta_2) \exp[-r(\theta_1, T) q(\theta_2)] \frac{T}{n - 1},
$$
(2.77)

$$
\frac{\partial \xi_0(T)}{\partial \theta_2} = (1 - \exp[-r(\theta_1, T) q(\theta_2)]) \frac{2}{q^2(\theta_2)}
$$
(2.78)

$$
+ p(\theta_2) r(\theta_1, T) \exp[-r(\theta_1, T) q(\theta_2)] (\exp(\theta_2) - \exp(-\theta_2)),
$$

and

$$
\frac{\partial \xi_1(T)}{\partial \theta_2} = (1 - \exp[-r(\theta_1, T) q(\theta_2)]) \frac{2}{q^2(\theta_2)}
$$
(2.79)

$$
- (1 - p(\theta_2)) r(\theta_1, T) \exp[-r(\theta_1, T) q(\theta_2)] (\exp(\theta_2) - \exp(-\theta_2)).
$$
Chapter 3

Tests of goodness-of-fit†

A popular approach to model network panel data is to embed the discrete observations of the network in a latent, continuous-time Markov process. A score-type test statistic for goodness-of-fit tests is proposed, which is useful for studying the goodness-of-fit of a wide range of models. The finite-sample behavior of the test statistic is evaluated by a Monte Carlo simulation study, and its usefulness is demonstrated by an application to empirical data.

Keywords: continuous-time Markov process, regular estimating functions, goodness-of-fit, Lagrange multiplier / Rao score test, Neyman C(α) test.

3.1 Introduction

Social network analysis (Wasserman and Faust, 1994) is concerned with links among entities. The network data considered here correspond to the directed ties among the members of a set of actors. It is common that the ties are binary, but ties may take on arbitrary values.

When modeling networks, the ties between the actors are treated as random variables. The tie variables are, however, not independent. Some well-known examples of dependencies are reciprocity (second-order dependence) and transitivity (Holland and Leinhardt, 1970, 1976), which represents third-order dependence among ties and implies clustering (“group structure”) in social networks.

Since the tie variables are dependent, statistical inference proved to be hard. (Curved) exponential random graph models (ERGMs) (Snijders, Pattison, Robins, and Handcock, 2006, Hunter and Handcock, 2006) have been used to model networks.
CHAPTER 3. TESTS OF GOODNESS-OF-FIT

observed at one time point, but—in spite of recent advances in the specification and estimation of ERGMs—some theoretical and practical issues remain.

Here, the focus is on longitudinal network data. Longitudinal network data come frequently in the form of panel data. There is a large agreement in the literature (see Frank, 1991) that the most promising models for network panel data are continuous-time Markov models which assume that the network was observed at discrete time points, and that between these time points latent, continuous-time Markov processes shape the network. Holland and Leinhardt (1977) and Wasserman (1979, 1980) proposed methods for statistical inference for such Markov models, but these methods are limited to models with second-order dependence, and thus neglect fundamental third- and higher-order dependencies.

Snijders (2001) considered a family of continuous-time Markov models which allows to model third- and higher-order dependencies, and proposed the method of moments to estimate the parameter $\theta$. The probabilistic framework may be described as actor-driven, that is, the nodes are assumed to represent actors who make choices concerning the ties to other actors by either adding or deleting ties, and the choices are assumed to be based on mathematical functions, containing choice “determinants” (tendencies) such as reciprocity, transitivity, and covariate-related effects.

In the tradition of K. Popper’s conception of science, deeply rooted in contemporary social science and statistics (Healy, 1978), a natural question to ask is “whether the observed data support a given specification” (Rao, 2002, p. 9). The study of goodness-of-fit, dating back to Pearson (1900), is considered so natural that model-based inference without goodness-of-fit evaluation is almost inconceivable. However, as is argued below in some more detail, no goodness-of-fit measure has been proposed up to now which is applicable in a wide range of applications.

The present paper proposes a new goodness-of-fit test statistic that (1) has many applications; (2) does not require to estimate the parameters to be tested, which is in practice a decisive advantage as it (a) saves valuable computation time, and (b) allows to test hard-to-estimate parameters (which are not uncommon); (3) admits multi-parameter tests; and (4) has an appealing interpretation in terms of goodness-of-fit, in the sense that it compares the expected value of some function of the data—evaluated under some assumed model—to the observed value of the function; that is, it uses the observed data as an external benchmark to which model predictions are compared.

The paper is structured as follows. Model specification and estimation are sketched in section 3.2. A goodness-of-fit test statistic is proposed in section 3.3. Section 3.4
reviews interesting goodness-of-fit tests. In section 3.5, the finite-sample behavior of
the test statistic is studied by Monte Carlo simulation. The usefulness of the test
statistic for real-world problems is demonstrated in section 3.6 by an application to
the cross-ownership network among more than 400 business firms in Slovenia (EU).

3.2 Model specification and estimation

Due to space restrictions, the discussion is restricted to some basic model specifica-
tions. Extensions to model the co-evolution of networks and other outcome variables
(Snijders, Steglich, and Schweinberger, 2006) are possible.

A binary, directed relation \( i \rightarrow j \) (or digraph) on a finite set of nodes \( N = \{1, 2, \ldots, n\} \)
is considered. The digraph is observed at discrete, ordered time points \( t_1 < t_2 < \cdots < t_M \),
and the observations are represented as binary matrices \( x(t_1), x(t_2), \ldots, x(t_M) \),
where element \( x_{ij}(t_m) \) of \( n \times n \) matrix \( x(t_m) \) is defined by

\[
x_{ij}(t_m) = \begin{cases} 
1 & \text{if } i \rightarrow j \text{ at time point } t_m, \\
0 & \text{otherwise,}
\end{cases} \tag{3.1}
\]

where \( i \rightarrow j \) means that node \( i \) is related to node \( j \); the diagonal elements \( x_{ii}(t_m) \) are regarded as structural zeros.

Although the model described below assumes that \( x_{ij}(t_m) \) is binary, it is possible
to extend the model to the case where \( x_{ij}(t_m) \) takes on discrete, ordered values.

3.2.1 Model specification

The observed digraph \( x(t_1) \) is taken for granted, that is, the model conditions on
\( x(t_1) \). It is postulated that the observed digraphs \( x(t_2), \ldots, x(t_M) \) are generated by
an unobserved Markov process operating in time interval \([t_1, t_M]\). Consider the case
\( M = 2 \); the extension to the case \( M > 2 \) is straightforward due to the Markov
property.

The model is specified by the generator of the Markov process, which corresponds
to a \( W \times W \) matrix \( Q \) indexed by a parameter \( \theta \), where \( W = 2^{n(n-1)} \) is the number
of digraphs on \( N \). The elements \( q_{\theta}(x^*, x) \) of generator \( Q \) are the rates of moving from
digraph \( x^* \) to digraph \( x \). If \( x \) deviates from \( x^* \) in more than one arc variable \( X_{ij}^* \),
then \( q_{\theta}(x^*, x) = 0 \) by assumption (see Snijders, 2001); in other words, it is assumed
that the process moves forward by changing not more than one arc variable \( X_{ij}^* \) at
the time. Let \( x^* \) be an arbitrary digraph on \( N \), and let \( x \) be the digraph that is
obtained from $x^*$ by changing one and only one specified arc variable, say $X_{ij}^*$. Since the transition from $x^*$ to $x$ involves only the ordered pair of nodes $(i, j)$, one can rewrite $q_\theta(x^*, x)$ as $q_\theta(x^*, i, j)$ and decompose $q_\theta(x^*, i, j)$ as follows:

$$q_\theta(x^*, i, j) = \lambda_i(\theta, x^*) r_i(\theta, x, j),$$

where

$$\lambda_i(\theta, x^*) = \sum_{h \neq i} q_\theta(x^*, i, h)$$

(3.3)

is called the rate function of node ("actor") $i$, while

$$r_i(\theta, x, j) = \frac{q_\theta(x^*, i, j)}{\lambda_i(\theta, x^*)}$$

(3.4)

is the conditional probability that $i$ changes $X_{ij}^*$, given that $i$ changes some arc variable $X_{ih}^*$, $h \neq i$.

A simple specification of $\lambda_i$ is

$$\lambda_i(\theta, x^*) = \rho,$$

(3.5)

where $\rho$ is a parameter, while non-constant rate functions are given by

$$\lambda_i(\theta, x^*) = \rho \exp [\alpha' a_i(x^*, c_i)],$$

(3.6)

where $\alpha = (\alpha_k)$ is a vector valued parameter and $a_i = (a_{ik})$ is a vector valued function of node-bound covariates $c_i$ and graph-dependent statistics involving the arcs of node $i$. If there is more than one time interval ($M > 2$), then parameter $\rho$ can be made dependent on time interval $[t_{m-1}, t_m]$.

A convenient, multinomial logit parametrization of $r_i$ is given by

$$r_i(\theta, x, j) = \frac{\exp [f_i(\beta, x, j)]}{\sum_{h \neq i} \exp [f_i(\beta, x, h)]},$$

(3.7)

where the real valued function

$$f_i(\beta, x, j) = \beta' s_i(x, j)$$

(3.8)

is called the objective function, while $\beta = (\beta_k)$ is a vector valued parameter and $s_i = (s_{ik})$ is a vector valued statistic. Examples of statistics $s_{ik}$ are the number of arcs $\sum_{h=1}^n x_{ih}$, the number of transitive triplets $\sum_{h=1}^n x_{ih} x_{hl} x_{il}$, and interactions of covariates with these and other statistics; see section 3.4. Such statistics can be used to define third- and higher-order dependencies.
3.3. TEST STATISTIC

3.2.2 Model estimation

Model estimation is concerned with estimating parameter $\theta$—corresponding to rate parameters $\rho_1, \ldots, \rho_{M-1}, \alpha$ and objective function parameter $\beta$—from the observed data $z$, consisting of the observed digraphs $x(t_1), x(t_2), \ldots, x(t_M)$ and covariates.

Since the Markov process is not observed in continuous time, the likelihood function is intractable. The parameter $\theta$ is therefore estimated by the method of moments (Pearson, 1902a,b). The moment estimate $\hat{\theta}$ is defined here as the solution of the moment equation

$$g_n(z, \theta) = \sum_{m=1}^{M-1} \left[ E_\theta [s(X(t_{m+1})) \mid X(t_m) = x(t_m)] - s(x(t_{m+1})) \right] = 0,$$

where $E_\theta$ denotes the expectation under $\theta$. The coordinates of the vector valued function $s = (s_k)$ in (3.9) correspond to the coordinates of parameter $\theta = (\theta_k)$. In the absence of formal methods to derive statistics with certain optimum properties (neglecting some close-to-trivial models), statistics $s_k$ are chosen heuristically so as to be sensitive to changes of the value of $\theta_k$. To estimate objective function parameter $\beta_k$, the statistic $\sum_{i,j=1}^{n} s_{ik}(x,j)$ is a natural choice; with regard to the rate parameters, see Snijders (2001).

In terms of numerical implementation, finding the moment estimate involves finding the root(s) of $g_n = g_n(z, \theta)$ as a function of $\theta$. A suitable root-finding algorithm, based on iterative, stochastic approximation methods (Robbins and Monro, 1951) and Monte Carlo simulation, is described in Snijders (2001). The variance-covariance matrix of the moment estimator $\hat{\theta}$ can be derived by the delta method (see Snijders, 2001).

3.3 Test statistic

Two papers address significance testing and / or goodness-of-fit in the considered family of models, though both of them have limitations, as is argued in section 3.3.1. A new test statistic is proposed in section 3.3.2, which has important advantages compared to the existing approaches.
3.3.1 Literature review

Snijders (1996) proposed to base significance tests on the pseudo-\(t\)-test statistic

\[
\frac{\hat{\theta}_k}{\text{s.e.}(\hat{\theta}_k)},
\]

where \(\hat{\theta}_k\) is the moment estimator of coordinate \(\theta_k\) of \(\theta\) and s.e.(\(\hat{\theta}_k\)) is its standard error. The distribution of the test statistic is unknown. Snijders (1996) assumed that its sampling distribution under the null hypothesis \(H_0: \theta_k = 0\) is approximately the standard Gaussian distribution.

Snijders (2003) stresses the modeling of (out)degree distributions. A goodness-of-fit plot with confidence intervals is provided as a means to evaluate the fit of the model with respect to the observed (out)degree distribution. However, the goodness-of-fit study is limited to the (out)degree distribution, and while the (out)degree distribution represents a fundamental feature of the data which should be taken into account, the (out)degree distribution in itself is in most applications a nuisance. In the social sciences, the focus is on second- and third-order dependencies among arcs and covariate-related parameters. The goodness-of-fit of corresponding models cannot, however, be addressed within the Snijders (2003) framework.

3.3.2 Goodness-of-fit test statistic

As to the choice of suitable test statistics for goodness-of-fit tests, observe that the choice is constrained, in the first place, by the intractable likelihood function, the absence of saturated models, and the computational burden imposed by estimating non-trivial models. Thus the “holy trinity”, corresponding to the Wald, likelihood ratio (LR), and Lagrange multiplier / Rao score (RS) test (Rao and Poti, 1946, Rao, 1948), is not readily available.

Although the “holy trinity” is not available, it can be instructive to consider some of its features. An important practical concern is computation time. As the LR test requires the estimation of both the restricted and the unrestricted model, it is inferior to the Wald and RS test in terms of computation time. The RS test is most appealing, since only the restricted model must be estimated, while the Wald test requires the more computation-intensive estimation of the unrestricted model. Besides, it is not uncommon to encounter convergence problems in high-dimensional parameter spaces. In fact, it may be argued that, in the considered family of models (given the absence of saturated models), forward model selection, combined with the
RS test, is preferable to backward model selection, because one can start with simple models and is not required to estimate additional parameters in the first place.

In the method of moments framework, the RS test is not available, but a RS-type test can be obtained by generalizing the $C(\alpha)$ test (Neyman, 1959) and the RS test by replacing the Fisher score function by regular estimating functions along the lines of Basawa (1985, 1991). The classical $C(\alpha)$ test is designed to test parametric hypotheses in the presence of nuisance parameters, where the nuisance parameters are replaced by consistent estimates under the null hypothesis. If maximum likelihood estimates (under the null hypothesis) are used to estimate the nuisance parameter, the classical $C(\alpha)$ test reduces to the RS test. The RS test encompasses the classical goodness-of-fit test of Pearson (1900) as a special case (see Rao, 2002). Interrelations among the “holy trinity” test statistics and the $C(\alpha)$ test statistic were studied by, among others, Godfrey (1988, pp. 27—28), Hall and Mathiason (1990), Basawa (1991), and Bera and Bilias (2001a,b).

Partition $\theta' = (\theta'_1, \theta'_2)$, where $\theta_1$ is the (vector valued) nuisance parameter and $\theta_2$ the (vector valued) parameter of primary interest. In the classical Neyman and Pearson tradition, goodness-of-fit can be studied by specifying hypotheses regarding the postulated family of probability distributions $\{P_\theta, \theta \in \Theta\}$, for instance, the null hypothesis

$$H_0 : \theta_2 = \theta_{20}, \quad (3.11)$$

tested against

$$H_1 : \theta_2 \neq \theta_{20}, \quad (3.12)$$

where $\theta_{20}$ is some specified value (commonly, $\theta_{20} = 0$), and $\theta_1$ is unspecified; that is, both $H_0$ and $H_1$ are composite.

A test statistic for testing such composite hypotheses is proposed below, based on the estimating function $g_n$ along the lines of Basawa (1985, 1991). The estimating function and some additional notation is introduced first.

**The estimating function**

For convenience, it is assumed that $M = 2$; the extension to the case $M > 2$ is immediate.

The parameter $\theta$ is estimated by the solution of the estimating equation $g_n = g_n(z, \theta) = 0$, where $g_n$ is the estimating function satisfying some mild smoothness conditions (see Godambe, 1960). Examples of estimating functions include the Fisher
score function, moment functions, or more generally, some function involving parameter $\theta$ and the data. Here, the estimating function $g_n$ is defined in (3.9).

Observe that $g_n$ as defined in (3.9) is an unbiased estimating function in the sense that

$$E_\theta[g_n(Z, \theta) \mid X(t_1) = x(t_1)] = 0$$  \hspace{1cm} (3.13)

holds for all $n$ and all $\theta$.

Let $\Sigma_n$ be the $L \times L$ variance-covariance matrix of $g_n$, and denote the limit of $w_n \Sigma_n$ as $n \rightarrow \infty$ by $\Sigma$, where $w_n$ are appropriate norming constants.

The function $s$ defined in section 3.2.2 does not depend on $\theta$, so that the definition of $g_n$ implies that the derivative of $g_n$ with respect to $\theta$ is deterministic and given by

$$\Delta_n(\theta) = \frac{\partial g_n(z, \theta)}{\partial \theta'} = \frac{\partial}{\partial \theta'} E_\theta [s(X(t_2)) \mid X(t_1) = x(t_1)].$$  \hspace{1cm} (3.14)

Denote the limit of $w_n \Delta_n$ as $n \rightarrow \infty$ by $\Delta$, where $g_n$ is of order $L \times 1$ and $\theta'$ is of order $1 \times L$, so that $\Delta$ is of order $L \times L$.

Last, partition $g_n$, $\Sigma$, and $\Delta$ in accordance with $\theta' = (\theta_1', \theta_2')$:

$$g_n(z, \theta) = \begin{pmatrix} g_{1n}(z, \theta) \\ g_{2n}(z, \theta) \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{pmatrix},$$

$$\Delta(\theta) = \begin{pmatrix} \Delta_{11}(\theta) & \Delta_{12}(\theta) \\ \Delta_{21}(\theta) & \Delta_{22}(\theta) \end{pmatrix}.$$  \hspace{1cm} (3.15)

The test statistic

As was mentioned above, the test statistic is based on estimating function $g_n$. Since the statistics contained in $g_n$ are commonly based on sums of weakly dependent random variables, it is plausible that

$$w_n^{1/2} g_n(Z, \theta) \xrightarrow{d} N_L(0, \Sigma) \text{ as } n \rightarrow \infty$$  \hspace{1cm} (3.16)

for the norming constants $w_n$ introduced above, where $\xrightarrow{d}$ denotes convergence in distribution, $N_L$ refers to the $L$-variate Gaussian distribution, and $\Sigma$ is non-singular. In the case of the Independent Arcs model (Snijders and Van Duijn, 1997), where the arc variables $X_{ij}(t)$ follow independent Markov processes, the asymptotic normality can be proved directly from the Central Limit Theorem. However, in the general case where the arc variables $X_{ij}(t)$ are dependent and the Markov process is non-stationary, a rigorous proof of (3.16) is beyond the scope of the present paper (see
3.3. TEST STATISTIC

The approximate normality is supported by some limited simulation studies (not further reported here); in the remainder of the section, it is assumed that (3.16) holds.

Let $\theta_0 = (\theta'_1, \theta'_{20})$ be the parameter under $H_0 : \theta_2 = \theta_{20}$. To eliminate the impact of the estimated nuisance parameter $\theta_1$ on the test, Neyman’s (1959) orthogonalization method can be exploited, as suggested by Basawa (1985, 1991). Let

$$b_n(z, \theta_0) = g_{2n}(z, \theta_0) - \Gamma(\theta_0) g_{1n}(z, \theta_0),$$

where $\Gamma = \Delta_{21} \Delta_{11}^{-1}$ and $\Delta_{11}$ is non-singular. Since both $g_{2n}$ and $\Gamma g_{1n}$ have zero expectation by (3.13) and are asymptotically Gaussian distributed by (3.16), one obtains

$$w_n^{1/2} b_n(Z, \theta_0) \xrightarrow{d} N_R(0, \Xi) \text{ as } n \to \infty,$$  \hspace{1cm} (3.18)

where the variance-covariance matrix $\Xi$ is given by

$$\Xi = \Sigma_{22} - (\Sigma_{12}' \Gamma(\theta_0) + \Gamma(\theta_0) \Sigma_{12}) + \Gamma(\theta_0) \Sigma_{11} \Gamma(\theta_0)' ,$$  \hspace{1cm} (3.19)

and $R$ is the number of coordinates of $\theta_2$. Thus, under $H_0$,

$$w_n b_n(Z, \theta_0)' \Xi^{-1} b_n(Z, \theta_0) \xrightarrow{d} \chi^2_R \text{ as } n \to \infty$$

is asymptotically central chi-square distributed with $R$ degrees of freedom.

The entities $\Delta$ and $\Sigma$ can be replaced by the constant matrices $w_n \Delta_n$ and $w_n \Sigma_n$, respectively; it is evident that these plug-ins do not change the asymptotic distribution of (3.20). The unknown nuisance parameter $\theta_1$ can be replaced by a consistent estimator under $H_0$ without changing the asymptotic distribution of (3.20). The parameter $\theta_1$ can be estimated by the $\hat{\theta}_1$ that solves $g_n = g_n(z, \theta_0) = 0$ (see section 3.2.2), where $\theta'_0 = (\theta'_1, \theta'_{20})$ and $\theta_{20}$ is the value of $\theta_2$ under $H_0$. It is plausible that $\hat{\theta}_1$ is a consistent estimator of $\theta_1$, although a rigorous proof is beyond the scope of the present paper (see section 3.7). Denote by $C_n$ the test statistic (3.20) obtained by plugging in $\hat{\theta}_1$, $\Delta_n$, and $\Sigma_n$ for $\theta_1$, $\Delta$, and $\Sigma$, respectively.

The entities $g_n$, $\Delta_n$, and $\Sigma_n$ are not available in closed form, but can be estimated by Monte Carlo methods. Crude Monte Carlo estimation (Hammersley and Handscomb, 1964) of the expectations of which $g_n$ is composed and of $\Sigma_n$ is straightforward. Monte Carlo estimators of $\Delta_n$ can be found in Schweinberger and Snijders (2007). These Monte Carlo estimators are simulation-consistent in the sense that the Monte Carlo estimators converge in probability to the desired quantities as the
number of Monte Carlo simulations increases without bound. Therefore, the Monte Carlo estimator of test statistic $C_n$, which is obtained by plugging in the Monte Carlo estimators of $g_n$, $\Delta_n$, and $\Sigma_n$, is simulation-consistent.

**Remarks and extensions**

Observe that, to test restrictions on parameter vector $\theta_2$, $\theta_2$ needs not be estimated, as is generally the case for RS tests.

If $\theta_2$ is a scalar, the test statistic (3.20) can be used both in its quadratic form, as presented above, and in its corresponding linear form

$$-w_n^{1/2} b_n(Z, \theta_0) \Xi^{1/2},$$

where $b_n$ and $\Xi$ are scalars. The linear form is convenient when one-sided one-parameter tests are desired. The minus sign in (3.21) facilitates the interpretation in the sense that, if $s_2$ denotes the statistic corresponding to parameter $\theta_2$ and its conditional expectations are non-decreasing functions of $\theta_2$, then, by the definition of $g_n$ in (3.9), $\theta_2 - \theta_{20} > 0$ is associated with positive values of (3.21). By (3.18), the asymptotic distribution of (3.21) under $H_0: \theta_2 = \theta_{20}$ is standard Gaussian.

Furthermore, tests with $R > 1$ degrees of freedom can be complemented with one degree of freedom tests, testing the restrictions one by one; two-sided one-parameter tests can be based on (3.20), while one-sided one-parameter tests can be based on (3.21).

Note that test statistic (3.20) has an appealing interpretation in terms of goodness-of-fit. Let $M = 2$ and observe that the test statistic is based on

$$g_{2n}(z, \theta) = E_{\theta} [s_2(X(t_2)) \mid X(t_1) = x(t_1)] - s_2(x(t_2)),
\tag{3.22}$$

where $s_2$ is the part of the statistics vector $s$ which corresponds to $\theta_2$. In other words, the test statistic is based on the “distance” between the expected value of the function $s$ of the data—evaluated under the model restricted by $H_0$—and the observed value of $s$. The argument extends to the case $M > 2$.

When the test indicates that there is empirical evidence against $H_0: \theta_2 = \theta_{20}$, it may be desired to estimate $\theta_2$. If $g_n$ is differentiable at $\hat{\theta}_0$, then, by definition (Magnus and Neudecker, 1988, p. 82),

$$g_n(z, \theta) = g_n(z; \hat{\theta}_0) + \Delta_n(\hat{\theta}_0) (\theta - \hat{\theta}_0) + h_{\hat{\theta}_0}(\theta - \hat{\theta}_0),
\tag{3.23}$$

where the Jacobian matrix $\Delta_n$ is given by (3.14), and

$$\lim_{(\theta - \hat{\theta}_0) \to 0} \frac{h_{\hat{\theta}_0}(\theta - \hat{\theta}_0)}{||\theta - \hat{\theta}_0||} = 0.
\tag{3.24}$$
Thus, solving \( g_n = g_n(z, \theta) = 0 \) is asymptotically the same as solving
\[
g_n(z, \hat{\theta}_0) + \Delta_n(\hat{\theta}_0) (\theta - \hat{\theta}_0) = 0, \tag{3.25}
\]
giving rise to the one-step estimator
\[
\theta^* = \hat{\theta}_0 - \Delta_n(\hat{\theta}_0)^{-1} g_n(z, \hat{\theta}_0), \tag{3.26}
\]
where \( \Delta_n \) is non-singular. In general, one-step estimators can be useful as approximations of estimators which are hard to obtain (see Lehmann, 1999). Here, the one-step estimator \( \theta^* \) is an approximation of the unrestricted moment estimator \( \hat{\theta} \), and is useful because all ingredients required to evaluate \( \theta^* \) are available once the ingredients of (3.20) are available, while the estimation of \( \hat{\theta} \) requires an additional, time-consuming estimation run (see section 3.2.2). Note that \( \theta^* = (\theta_1^*, \theta_2^*) \) is an estimator of both \( \theta_1 \) and \( \theta_2 \), and in practice both \( \theta_1^* \) and \( \theta_2^* \) are interesting, because in network models these estimators can be considerably correlated.

### 3.4 Model misspecifications

Two classes of model misspecifications are briefly considered. First, the model is misspecified in the sense that the true parameter \( \theta \) contains other (additional) coordinates than the specified model. Section 3.4.1 deals with such cases. Second, the model is misspecified in the sense that homogeneity assumptions regarding \( \theta \) do not hold. Such homogeneity assumptions are convenient for statistical and computational reasons. Two important cases are the assumption that

1. \( \beta^{(1)} = \beta^{(2)} = \cdots = \beta^{(M-1)} \), where the superscript refers to the period, provided \( M \geq 3 \).

2. \( \beta^{(1)} = \beta^{(2)} = \cdots = \beta^{(n)} \), where the superscript refers to the node.

Testing such homogeneity assumptions is considered in sections 3.4.2 and 3.4.3.

#### 3.4.1 Misspecifications I

The simple case is considered where the specified objective function \( f_i \) by mistake excludes effects inherent to the true \( f_i \).

Besides second- and third-order dependencies, covariates may have an impact on the digraph evolution. As an example, adolescents may favor friendship ties to adolescents of the same gender. Let \( c \) be some
(1) node-bound covariate $c_i$, containing information about node $i$: (a) demographic covariate; (b) behavioral or other covariate (such as smoking).

(2) dyadic covariate $c_{ij}$, containing information about the ordered pair of nodes $(i, j)$: a frequently used form of dyadic covariates is “similarity” of nodes $i$ and $j$ with respect to node-bound covariates.

The objective function can be specified as

$$f_i(\beta, x) = \sum_{k=1}^{K} \beta_k s_{ik}(x, j) + \beta_{K+1} s_{iK+1}(x, c, j).$$ \hspace{1cm} (3.27)

Some possible specifications of statistic $s_{iK+1}$ are

(1) node-bound covariate: $\sum_{h=1}^{n} x_{ih} c_h$ or $c_i \sum_{h=1}^{n} x_{ih}$.

(2) dyadic covariate: $\sum_{h=1}^{n} x_{ih} c_{ih}$.

(3) interaction node-bound covariate $\times$ reciprocated arcs: $\sum_{h=1}^{n} x_{ih} x_{hi} c_h$.

Numerous other interactions between covariates and digraph statistics are conceivable. The covariate $c$ may, or may not, depend on time. If $c$ depends on time, it is—in the presence of panel data—convenient to assume that it is constant in time interval $[t_m, t_{m+1}]$, so that the value of $c(t_m)$ observed at time point $t_m$ can be used to evaluate the corresponding statistic in time interval $[t_m, t_{m+1}]$.

A natural approach is to test the model restricted by

$$H_0 : \beta_{K+1} = 0$$ \hspace{1cm} (3.28)

against the unrestricted model

$$H_1 : \beta_{K+1} \neq 0.$$ \hspace{1cm} (3.29)

### 3.4.2 Misspecifications II: homogeneous periods

An important assumption underlying the Snijders (2001) family of models is that $\beta$ is constant across time intervals $[t_m, t_{m+1}]$, $m = 1, \ldots, M - 1$. To test whether the data-generating mechanism is indeed constant across time intervals, the model can be specified as

$$f_i(\beta, x) = \sum_{k=1}^{K} \beta_k^{(m)} s_{ik}(x, j),$$ \hspace{1cm} (3.30)
where parameter $\beta_k^{(m)}$ depends on time interval $[t_m, t_{m+1}]$. Observe that an equivalent formulation is obtained by including, for the $M-1$ periods, $M-2$ period-dependent dummy variables interacting with statistic $s_{ik}$ (assuming that the main effect of $s_{ik}$ is included), where the dummies are equivalent to node- and time-dependent covariates $c_i^{(m)}$.

To test whether coordinate $\beta_k$ of $\beta$ is constant across time intervals, the model restricted by

$$H_0 : \beta_k^{(1)} = \beta_k^{(2)} = \cdots = \beta_k^{(M-1)}$$

(3.31)

can be tested against the unrestricted model

$$H_1 : \beta_k^{(a)} \neq \beta_k^{(b)} \text{ for some } a \neq b,$$

(3.32)

where $a$ and $b$ indicate periods.

### 3.4.3 Misspecifications II: homogeneous nodes

Another assumption of the Snijders (2001) family of models is that $\beta$ is constant across nodes $i$. Models with node-dependent parameters can be specified, which is equivalent to using $n-1$ node-dependent dummy variables $c_i$ interacting with statistic $s_{ik}$ (assuming that the main effect of $s_{ik}$ is included). Tests can be carried out as in section 3.4.2.

### 3.5 Monte Carlo study

The present section studies the finite-sample behavior of the proposed test statistic (3.20) and the pseudo-$t$-test (3.10). For convenience, the two tests will be referred to as “score test” and “$t$-test”, respectively, although the tests are not equivalent to the tests which are known in the statistical literature under these names.

Data are generated by simulating the Markov process with $n = 30$ (“small data set”) and $n = 60$ nodes (“moderate data set”) in time interval $[0, 2]$, starting with a real-world digraph at time point $t_1 = 0$ and “observing” the random digraph at time points $t_2 = 1$ and $t_3 = 2$. The rate function $\lambda_i$ is constant and equal to $\rho_m$ for period $m = 1, 2$, and the objective function is given by

$$f_i(\beta, x) = \sum_{k=1}^K \beta_k s_{ik}(x, j).$$

(3.33)
The choice of statistics $s_{ik}$ is based on their importance in empirical social science research; the chosen statistics are:

- $s_{i1}(x,j) = \sum_{h=1}^{n} x_{ih}$: the number of arcs,
- $s_{i2}(x,j) = \sum_{h=1}^{n} x_{ih} x_{hi}$: the number of reciprocated arcs,
- $s_{i3}(x,j) = \sum_{h,l=1}^{n} x_{ih} x_{hi} x_{il}$: the number of transitive triplets,
- $s_{i4}(x,j) = \sum_{h=1}^{n} (1 - x_{ih}) \max_{l} x_{il} x_{lh}$: the number of indirect connections,
- $s_{i5}(x,j) = \sum_{h=1}^{n} x_{ih} c_{ih}$: interaction of arcs and dyadic covariate $c_{ih}$,
- $s_{i6}(x,j) = \sum_{h=1}^{n} x_{ih} d_{h}$: interaction of arcs and node-bound covariate $d_{h}$.

The values of the dyadic covariate $c_{ih}$ are generated by independent draws from the Poisson(1) distribution; the values of the node-bound covariate $d_{h}$ are generated by independent draws from the Bernoulli(1/2) distribution.

Two main purposes of goodness-of-fit testing can be distinguished, and hence the simulation study consists of two main parts: (1) testing parameters which capture structural features of the data such as third-order dependence among arcs (“clustering”), and (2) testing the impact of covariates on the digraph evolution. The basic data-generating model corresponds to parameters $\rho_1 = \rho_2 = 4$, $\beta_1 = -1$ (corresponding to $s_{i1}$), and $\beta_2 = 1$ (corresponding to $s_{i2}$), which are common to all models used for generating data. The first part of the simulation study tests hypotheses involving the parameters $\beta_3$ and $\beta_4$, corresponding to $s_{i3}$ and $s_{i4}$, respectively; both parameters capture third-order dependence among arcs. The second part of the simulation study tests hypotheses involving the parameters $\beta_5$ and $\beta_6$, corresponding to $s_{i5}$ and $s_{i6}$, respectively. All tests are two-sided.

**Part I: testing effects capturing third-order dependence** The basic model is $P_{\theta}, \theta = (\rho_1, \rho_2, \beta_1, \beta_2, \beta_3, \beta_4)' = (4, 4, -1, 1, \beta_3, \beta_4)'$; three values of $\beta_3$ (0, .2, and .4) and $\beta_4$ (0, -.3, and -.6) are considered including all combinations, giving 9 models. The values of $\beta_3$ and $\beta_4$ are primarily chosen to study the power of the test against local alternatives. For each model, 500 data sets are generated; the reason for limiting the number of data sets to 500 is the computing time required to estimate models. For each model and each data set, the score test is evaluated in one estimation run (where the 4 unrestricted parameters are estimated) and the $t$-test in another
3.5. MONTE CARLO STUDY

Figure 3.1: Monte Carlo results, model \( P_\theta, \theta = (\rho_1, \rho_2, \beta_1, \beta_2, \beta_3, \beta_4)' = (4, 4, -1, 1, 0, 0)' \), \( n = 30 \) and \( n = 60 \): distribution of test statistics

Note: the curves represent the expected distributions under \( H_0 \); \( c \) and \( t \) refer to score test (3.20) and \( t \)-test (3.10), respectively.

estimation run (where all 6 parameters are estimated); the parameters are estimated by the conditional method of moments of Snijders (2001).

Figure 3.1 (p. 53) shows histograms of the distributions of the score test statistic for testing \( H_0 : \beta_3 = \beta_4 = 0 \), \( H_0 : \beta_3 = 0 \), and \( H_0 : \beta_4 = 0 \), and histograms of the distributions of the \( t \)-test statistic for testing \( H_0 : \beta_3 = 0 \) and \( H_0 : \beta_4 = 0 \), where the true, data-generating model is \( P_\theta, \theta = (\rho_1, \rho_2, \beta_1, \beta_2, \beta_3, \beta_4)' = (4, 4, -1, 1, 0, 0)' \). For the score tests, the distributions agree very well with the chi-square distributions expected under \( H_0 \), but the \( t \)-tests appear to be slightly conservative.

Table 3.1 (p. 54) shows the empirical rejection probabilities for the mentioned hypotheses and the 9 data-generating models using the nominal significance level .05; space restrictions do not allow to show the results for other nominal significance levels (which do not alter the conclusions). Note that if \( H_0 \) is true and the test is of size .05, then the binomial distribution of the number of rejections implies that the empirical rejection rate should be roughly between .03 and .07.

The two-parameter score test behaves reasonable under \( H_0 : \beta_3 = \beta_4 = 0 \) and seems to have (for all practical purposes) sufficient power to detect departures from \( H_0 \).
Table 3.1: Monte Carlo results, model $P_{\theta}$, $\theta = (\rho_1, \rho_2, \beta_1, \beta_2, \beta_3, \beta_4)' = (4, 4, -1, 1, \beta_3, \beta_4)'$, $n = 30$ and $n = 60$: empirical rejection probabilities for tests with nominal significance level .05

<table>
<thead>
<tr>
<th>true model:</th>
<th>$H_0$: $\beta_3 = \beta_4 = 0$</th>
<th>$\beta_3 = 0$</th>
<th>$\beta_4 = 0$</th>
<th>$\beta_3 = 0$</th>
<th>$\beta_4 = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{\theta}$, $\beta_3 = 0$, $\beta_4 = 0$:</td>
<td>$n = 30$</td>
<td>.024</td>
<td>.050</td>
<td>.034</td>
<td>.032</td>
</tr>
<tr>
<td></td>
<td>$n = 60$</td>
<td>.036</td>
<td>.054</td>
<td>.044</td>
<td>.048</td>
</tr>
<tr>
<td>$P_{\theta}$, $\beta_3 = .2$, $\beta_4 = 0$:</td>
<td>$n = 30$</td>
<td>.446</td>
<td>.540</td>
<td>.136</td>
<td>.422</td>
</tr>
<tr>
<td></td>
<td>$n = 60$</td>
<td>.744</td>
<td>.812</td>
<td>.128</td>
<td>.788</td>
</tr>
<tr>
<td>$P_{\theta}$, $\beta_3 = .4$, $\beta_4 = 0$:</td>
<td>$n = 30$</td>
<td>.984</td>
<td>.990</td>
<td>.664</td>
<td>.980</td>
</tr>
<tr>
<td></td>
<td>$n = 60$</td>
<td>1.000</td>
<td>1.000</td>
<td>.650</td>
<td>1.000</td>
</tr>
<tr>
<td>$P_{\theta}$, $\beta_3 = 0$, $\beta_4 = -.3$:</td>
<td>$n = 30$</td>
<td>.492</td>
<td>.078</td>
<td>.596</td>
<td>.018</td>
</tr>
<tr>
<td></td>
<td>$n = 60$</td>
<td>.972</td>
<td>.140</td>
<td>.988</td>
<td>.052</td>
</tr>
<tr>
<td>$P_{\theta}$, $\beta_3 = 0$, $\beta_4 = -.6$:</td>
<td>$n = 30$</td>
<td>.892</td>
<td>.180</td>
<td>.940</td>
<td>.014</td>
</tr>
<tr>
<td></td>
<td>$n = 60$</td>
<td>1.000</td>
<td>.472</td>
<td>1.000</td>
<td>.030</td>
</tr>
<tr>
<td>$P_{\theta}$, $\beta_3 = -.2$, $\beta_4 = -.3$:</td>
<td>$n = 30$</td>
<td>.816</td>
<td>.512</td>
<td>.800</td>
<td>.220</td>
</tr>
<tr>
<td></td>
<td>$n = 60$</td>
<td>1.000</td>
<td>.832</td>
<td>.998</td>
<td>.506</td>
</tr>
<tr>
<td>$P_{\theta}$, $\beta_3 = -.4$, $\beta_4 = -.3$:</td>
<td>$n = 30$</td>
<td>.992</td>
<td>.978</td>
<td>.976</td>
<td>.870</td>
</tr>
<tr>
<td></td>
<td>$n = 60$</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.986</td>
</tr>
<tr>
<td>$P_{\theta}$, $\beta_3 = -.2$, $\beta_4 = -.6$:</td>
<td>$n = 30$</td>
<td>.982</td>
<td>.658</td>
<td>.990</td>
<td>.162</td>
</tr>
<tr>
<td></td>
<td>$n = 60$</td>
<td>1.000</td>
<td>.946</td>
<td>1.000</td>
<td>.316</td>
</tr>
<tr>
<td>$P_{\theta}$, $\beta_3 = -.4$, $\beta_4 = -.6$:</td>
<td>$n = 30$</td>
<td>1.000</td>
<td>.952</td>
<td>1.000</td>
<td>.574</td>
</tr>
<tr>
<td></td>
<td>$n = 60$</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.900</td>
</tr>
</tbody>
</table>
Figure 3.2: Monte Carlo results, model $P_\theta$, $\theta = (\rho_1, \rho_2, \beta_1, \beta_2, \beta_3, \beta_5, \beta_6)' = (4, 4, -1, 1, 2, 0, 0)'$, $n = 30$ and $n = 60$: distribution of test statistics

Note: the curves represent the expected distributions under $H_0$; $c$ and $t$ refer to score test (3.20) and $t$-test (3.10), respectively.

Table 3.1 indicates that the one-parameter score test has considerably more power than the $t$-test, in particular if $n = 30$ and the departure from $H_0$ is small.

Concerning the one-parameter score test, note that the two-parameter test and the two corresponding one-parameter tests are computed from the same estimation run, where $\theta$ was estimated under $H_0: \beta_3 = \beta_4 = 0$. Therefore, the one-parameter score tests in table 3.1 do not control for the other parameter ($\beta_3$ or $\beta_4$). Note that, in principle, it is straightforward to carry out one-parameter score tests of $\beta_3$ and $\beta_4$ where the other parameter is controlled for, but that would require two additional estimation runs.

Concerning the $t$-test, it is notable that the power of the $t$-test for testing $H_0: \beta_3 = 0$ is the smaller the larger the departure of $\beta_4$ from 0.

Part II: testing covariate-related effects The basic model is $P_\theta$, $\theta = (\rho_1, \rho_2, \beta_1, \beta_2, \beta_3, \beta_5, \beta_6)' = (4, 4, -1, 1, 2, \beta_5, \beta_6)'$, which includes transitivity parameter $\beta_3$; three values of $\beta_5$ (0, .1, and .2) and $\beta_6$ (0, .2, and .4) are considered including all combinations, giving 9 models.

Figure 3.2 (p. 55) shows histograms of the distributions of the score test statistic...
Table 3.2: Monte Carlo results, model $P_\theta, \theta = (\rho_1, \rho_2, \beta_1, \beta_2, \beta_3, \beta_5, \beta_6)' = (4, 4, -1, 1, 2, \beta_5, \beta_6)'$, $n = 30$ and $n = 60$: empirical rejection probabilities for tests with nominal significance level .05

<table>
<thead>
<tr>
<th>true model:</th>
<th>$P_{\theta}$, $\beta_5 = 0$, $\beta_6 = 0$:</th>
<th>$P_{\theta}$, $\beta_5 = .1$, $\beta_6 = 0$:</th>
<th>$P_{\theta}$, $\beta_5 = .2$, $\beta_6 = 0$:</th>
<th>$P_{\theta}$, $\beta_5 = 0$, $\beta_6 = .2$:</th>
<th>$P_{\theta}$, $\beta_5 = .1$, $\beta_6 = .4$:</th>
<th>$P_{\theta}$, $\beta_5 = .2$, $\beta_6 = .4$:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$:</td>
<td>$\beta_5 = \beta_6 = 0$</td>
<td>$\beta_5 = 0$</td>
<td>$\beta_6 = 0$</td>
<td>$\beta_5 = 0$</td>
<td>$\beta_6 = 0$</td>
<td>$\beta_5 = 0$</td>
</tr>
<tr>
<td>$n = 30$</td>
<td>.024</td>
<td>.024</td>
<td>.030</td>
<td>.044</td>
<td>.068</td>
<td></td>
</tr>
<tr>
<td>$n = 60$</td>
<td>.012</td>
<td>.014</td>
<td>.032</td>
<td>.058</td>
<td>.062</td>
<td></td>
</tr>
<tr>
<td>$n = 30$</td>
<td>.182</td>
<td>.250</td>
<td>.044</td>
<td>.326</td>
<td>.078</td>
<td></td>
</tr>
<tr>
<td>$n = 60$</td>
<td>.394</td>
<td>.526</td>
<td>.038</td>
<td>.640</td>
<td>.056</td>
<td></td>
</tr>
<tr>
<td>$n = 30$</td>
<td>.690</td>
<td>.780</td>
<td>.034</td>
<td>.846</td>
<td>.058</td>
<td></td>
</tr>
<tr>
<td>$n = 60$</td>
<td>.968</td>
<td>.988</td>
<td>.046</td>
<td>.994</td>
<td>.084</td>
<td></td>
</tr>
<tr>
<td>$n = 30$</td>
<td>.152</td>
<td>.034</td>
<td>.220</td>
<td>.068</td>
<td>.326</td>
<td></td>
</tr>
<tr>
<td>$n = 60$</td>
<td>.336</td>
<td>.030</td>
<td>.512</td>
<td>.066</td>
<td>.626</td>
<td></td>
</tr>
<tr>
<td>$n = 30$</td>
<td>.588</td>
<td>.020</td>
<td>.696</td>
<td>.070</td>
<td>.808</td>
<td></td>
</tr>
<tr>
<td>$n = 60$</td>
<td>.942</td>
<td>.022</td>
<td>.974</td>
<td>.056</td>
<td>.982</td>
<td></td>
</tr>
<tr>
<td>$n = 30$</td>
<td>.344</td>
<td>.276</td>
<td>.190</td>
<td>.374</td>
<td>.292</td>
<td></td>
</tr>
<tr>
<td>$n = 60$</td>
<td>.660</td>
<td>.530</td>
<td>.462</td>
<td>.632</td>
<td>.596</td>
<td></td>
</tr>
<tr>
<td>$n = 30$</td>
<td>.734</td>
<td>.262</td>
<td>.728</td>
<td>.378</td>
<td>.822</td>
<td></td>
</tr>
<tr>
<td>$n = 60$</td>
<td>.980</td>
<td>.500</td>
<td>.982</td>
<td>.616</td>
<td>.992</td>
<td></td>
</tr>
<tr>
<td>$n = 30$</td>
<td>.790</td>
<td>.792</td>
<td>.232</td>
<td>.858</td>
<td>.318</td>
<td></td>
</tr>
<tr>
<td>$n = 60$</td>
<td>.996</td>
<td>.992</td>
<td>.482</td>
<td>.998</td>
<td>.600</td>
<td></td>
</tr>
<tr>
<td>$n = 30$</td>
<td>.932</td>
<td>.812</td>
<td>.712</td>
<td>.886</td>
<td>.818</td>
<td></td>
</tr>
<tr>
<td>$n = 60$</td>
<td>1.000</td>
<td>.984</td>
<td>.984</td>
<td>.994</td>
<td>.990</td>
<td></td>
</tr>
</tbody>
</table>
for testing $H_0: \beta_5 = \beta_6 = 0$, $H_0: \beta_5 = 0$, and $H_0: \beta_6 = 0$, and histograms of the distributions of the $t$-test statistic for testing $H_0: \beta_5 = 0$ and $H_0: \beta_6 = 0$, where the true, data-generating model is $P_\theta$, $\theta = (\rho_1, \rho_2, \beta_1, \beta_2, \beta_3, \beta_5, \beta_6)' = (4, 4, -1, 1, .2, 0, 0)'$. For the score tests, the distributions agree quite well with the chi-square distributions expected under $H_0$, although the two-parameter score test appears to be slightly conservative.

Table 3.2 (p. 56) shows the empirical rejection probabilities for the mentioned hypotheses and the 9 data-generating models using the nominal significance level .05. Table 3.2 indicates that the one-parameter score test has less power than the $t$-test, in particular if $n = 30$ and the departure from $H_0$ is small.

## 3.6 Application

Pahor (2003) studied the directed cross-ownerships among 413 business firms in Slovenia (EU) observed at 5 time points, where directed cross-ownership $A \rightarrow B$ means that firm $A$ holds stock market shares of firm $B$. The present paper re-analyses the data in the light of the new goodness-of-fit test statistic. Once again, it is convenient to refer to the two tests as “score test” and “$t$-test”, respectively, and all tests are two-sided, unless stated otherwise.

The baseline model considered here corresponds to the rate function

$$
\lambda_i(\theta, x^*) = \rho_m \exp \left[ \alpha \left( 1 + \sum_{h=1}^{n} x_{ih}^* \right)^{-1} \right]
$$

for time interval $[t_m, t_{m+1}]$. The objective function is given by

$$
f_i(\beta, x) = \sum_{k=1}^{K} \beta_k s_{ik}(x, j),
$$

where

$$
s_{i1}(x, j) = \sum_{h=1}^{n} x_{ih}: \text{the number of arcs},
$$

$$
s_{i2}(x, j) = \sum_{h=1}^{n} x_{ih}x_{hi}: \text{the number of reciprocated arcs},
$$

$$
s_{i3}(x, j) = \sum_{h=1}^{n} x_{ih}d_{h1}: \text{interaction of arcs and covariate } d_{h1},
$$

$$
s_{i4}(x, j) = \sum_{h=1}^{n} x_{ih}d_{h2}: \text{interaction of arcs and covariate } d_{h2},
$$
CHAPTER 3. TESTS OF GOODNESS-OF-FIT

Table 3.3: Pahor data: testing restricted models

<table>
<thead>
<tr>
<th>nuisance parameters</th>
<th>test</th>
<th>score test</th>
<th>d.f.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \eta )</td>
<td>( H_0 : \beta_5 = \beta_6 = 0 )</td>
<td>61.95</td>
<td>2</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>( \eta )</td>
<td>( H_0 : \beta_5 = 0 )</td>
<td>61.82</td>
<td>1</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>( \eta )</td>
<td>( H_0 : \beta_6 = 0 )</td>
<td>&lt; .01</td>
<td>1</td>
<td>.9905</td>
</tr>
<tr>
<td>( \eta, \beta_5 )</td>
<td>( H_0 : \beta_7 = \beta_8 = \beta_9 = 0 )</td>
<td>127.59</td>
<td>3</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>( \eta, \beta_5 )</td>
<td>( H_0 : \beta_7 = 0 )</td>
<td>43.38</td>
<td>1</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>( \eta, \beta_5 )</td>
<td>( H_0 : \beta_8 = 0 )</td>
<td>70.34</td>
<td>1</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>( \eta, \beta_5 )</td>
<td>( H_0 : \beta_9 = 0 )</td>
<td>14.04</td>
<td>1</td>
<td>.0002</td>
</tr>
<tr>
<td>( \eta, \beta_5, \beta_7, \beta_8, \beta_9 )</td>
<td>( H_0 : \beta_{10} = \beta_{11} = \beta_{12} = 0 )</td>
<td>20.19</td>
<td>3</td>
<td>.0002</td>
</tr>
<tr>
<td>( \eta, \beta_5, \beta_7, \beta_8, \beta_9 )</td>
<td>( H_0 : \beta_{10} = 0 )</td>
<td>10.37</td>
<td>1</td>
<td>.0013</td>
</tr>
<tr>
<td>( \eta, \beta_5, \beta_7, \beta_8, \beta_9 )</td>
<td>( H_0 : \beta_{11} = 0 )</td>
<td>4.60</td>
<td>1</td>
<td>.0319</td>
</tr>
<tr>
<td>( \eta, \beta_5, \beta_7, \beta_8, \beta_9 )</td>
<td>( H_0 : \beta_{12} = 0 )</td>
<td>13.83</td>
<td>1</td>
<td>.0002</td>
</tr>
</tbody>
</table>

The parameter vector \( \eta \) is given by \( \eta = (\alpha', \alpha, \beta_1, \beta_2, \beta_3, \beta_4)' \).

where \( d_{h1} \) indicates whether or not firm \( h \) was quoted on the stock exchange, while \( d_{h2} \) refers to the size of firm \( h \). The parameter \( \theta \) of the baseline model \( P_0 \) corresponds to \( \theta = (\alpha', \alpha, \beta_1, \beta_2, \beta_3, \beta_4)' \), where \( \alpha = (\rho_1, \rho_2, \rho_3, \rho_4)' \).

Pahor (2003) suspected that the data may exhibit third-order dependence, leading to the transitivity parameter \( \beta_5 \) and the indirect connections parameter \( \beta_6 \) as candidates to be tested, which correspond to statistics \( \sum_{h,l=1}^{n} x_{ih} x_{hl} x_{il} \) and \( \sum_{h=1}^{n} (1 - x_{ih}) \max x_{il} x_{ih} \), respectively. According to table 3.3 (p. 58), the two-parameter score test of \( H_0 : \beta_5 = \beta_6 = 0 \) clearly indicates that parameters capturing third-order dependence are required to improve the goodness-of-fit of the model, and the one-parameter score tests suggest to include \( \beta_5 \) but not \( \beta_6 \). A one-sided test of \( H_0 : \beta_5 = 0 \) can be carried out by using the linear form of the score test (see (3.21)), giving 7.86: using the asymptotic standard Gaussian distribution of the linear form under \( H_0 : \beta_5 = 0 \), it seems that \( \beta_5 \) is positive, which is supported by the one-step estimate of \( \beta_5 \) (see (3.26)) given by .996. As a result, \( \beta_5 \) is henceforth included in the model, while \( \beta_6 \) is not.

Pahor (2003) argued that firms tend to hold shares of other firms close to them with respect to region \( (c_{ih1}) \) and industry branch \( (c_{ih2}) \), and to other firms having the same owner \( (c_{ih3}) \). Three parameters are added, \( \beta_7, \beta_8, \) and \( \beta_9 \), corresponding to statistics \( \sum_{h=1}^{n} x_{ih} c_{ihl}, l = 1, 2, 3 \), respectively. The three-parameter score test of \( H_0 : \beta_7 = \beta_8 = \beta_9 = 0 \) and the three corresponding one-parameter tests, shown in table 3.3, suggest to add all three covariates to the model, which is done.
3.6. APPLICATION

Table 3.4: Pahor data, model $P_{\theta}$, $\theta = (\alpha', \alpha, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11}, \beta_{12})'$: estimates, standard errors, and $t$-tests

<table>
<thead>
<tr>
<th>Parameter</th>
<th>One-step Estimate $\theta^*$</th>
<th>Moment Estimate $\hat{\theta}$</th>
<th>s.e.$(\hat{\theta})$</th>
<th>$t$-test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>-2.65</td>
<td>-2.67</td>
<td>.097</td>
<td>-27.54</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>.81</td>
<td>.81</td>
<td>.102</td>
<td>7.91</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>.57</td>
<td>.58</td>
<td>.076</td>
<td>7.60</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>1.01</td>
<td>1.01</td>
<td>.106</td>
<td>9.57</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>.89</td>
<td>.89</td>
<td>.258</td>
<td>3.45</td>
<td>.0006</td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>- .31</td>
<td>- .29</td>
<td>.134</td>
<td>-2.12</td>
<td>.0337</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>- .31</td>
<td>- .28</td>
<td>.158</td>
<td>-1.79</td>
<td>.0735</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>.12</td>
<td>.13</td>
<td>.117</td>
<td>1.14</td>
<td>.2544</td>
</tr>
</tbody>
</table>

To save space, the nuisance parameters $\alpha, \alpha_2, \beta_3$, and $\beta_4$ are omitted. The one-step estimate $\theta^*$ is based on the moment estimate of $\theta$ under the model restricted by $H_0 : \beta_{10} = \beta_{11} = \beta_{12} = 0$; the moment estimate $\hat{\theta}$ is the unrestricted moment estimate of $\theta$.

It is interesting to test whether the values of the parameters are constant across time intervals. A basic parameter for which such homogeneity tests frequently make sense is the “outdegree” parameter $\beta_1$, corresponding to statistic $s_{i1}$. A homogeneity test for $\beta_1$ is conducted by testing $H_0 : \beta_{10} = \beta_{11} = \beta_{12} = 0$, where $\beta_{10}, \beta_{11},$ and $\beta_{12}$ correspond to statistics $e_i^{(m)} s_{i1}, m = 2, 3, 4$, respectively, where $e_i^{(m)}$ is a period-dependent dummy variable with value 1 in time interval $[t_m, t_{m+1}]$ for all $i$ and 0 otherwise. According to table 3.3, the three-parameter score test of $H_0 : \beta_{10} = \beta_{11} = \beta_{12} = 0$ indicates that there is empirical evidence against $H_0$, and the three corresponding one-parameter tests suggest that it is sensible to add $\beta_{10}, \beta_{11},$ and $\beta_{12}$ to the model. The one-sided one-parameter score test statistics (see (3.21)) are $-3.22, -2.15,$ and 3.72, respectively, suggesting that the values of $\beta_{10}$ and $\beta_{11}$ are negative while $\beta_{12}$ is positive. Table 3.4 (p. 59) gives the one-step estimate (see (3.26)) of the parameter $\theta = (\alpha', \alpha, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11}, \beta_{12})'$ based on the moment estimate of $\theta$ under the model restricted by $H_0 : \beta_{10} = \beta_{11} = \beta_{12} = 0$, and shows in addition the unrestricted moment estimate of $\theta$, including standard errors and $t$-tests. The one-step estimates roughly agree with the unrestricted moment estimates. The $t$-tests by and large agree with the score tests (regarding $\beta_5, \beta_7, \beta_8, \beta_9$, and $\beta_{10}$), but slightly disagree about $\beta_{11}$ and clearly disagree about $\beta_{12}$. 
3.7 Discussion

A new goodness-of-fit test statistic was proposed, which has many applications, and, in contrast to the pseudo-\(t\)-test, does not require the estimation of the parameters to be tested (saving computation time and avoiding convergence problems), admits multi-parameter tests, and has an appealing interpretation in terms of goodness-of-fit.

The Monte Carlo study indicated that the finite-sample distribution of the goodness-of-fit test statistic under the considered null hypotheses matches the expected distributions fairly well.

As to the unproven assertion (3.16), observe that the statistical modeling of social networks is in its infancy, and that asymptotic properties of estimators and test statistics are largely unknown; see, for instance, Hunter and Handcock (2006). In addition, asymptotics—despite of being an indispensable tool in mathematical statistics and useful guides in applied statistics—strictly speaking make no statement about samples of given finite size (De Bruijn, 1981, p. 2, Hampel, 1997). In combination with the Monte Carlo study, the approach in the present paper can be recommended for practical use, though care must be taken.

The proposed test statistic is implemented in the Windows-based computer program Siena embedded in the program collection StOCNET, which can be downloaded free of charge from http://stat.gamma.rug.nl/stocnet.
Chapter 4

Random effects models†

Digraph panel data (or longitudinal social network data), which correspond to arcs between nodes observed at discrete time points, are collected in the social sciences and other fields. Conventional models of digraph panel data assume that the data are outcomes of a Markov process which operates in continuous time but is observed at discrete time points. Such models make the implicit assumption that all relevant knowledge with respect to nodes is observed in the form of covariates and correctly incorporated in the model. The present paper proposes Markov models which allow for unobserved heterogeneity across nodes by introducing random variables with latent outcomes, called random effects. Maximum likelihood and Bayesian methods are proposed to estimate such models. Both estimation approaches are implemented with Markov chain Monte Carlo-based data augmentation. The model is illustrated by an application to social network data.

Keywords: social networks, continuous-time Markov process, latent variables, hidden Markov models, data augmentation, Markov chain Monte Carlo.

4.1 Introduction

In various fields of scientific inquiry, such as in the social sciences, graphs have been exploited to represent dyadic data structures that correspond to links between entities (see Wasserman and Faust, 1994). An example is provided by sexual relationships between individuals (Jones and Handcock, 2003), the structure of which is key to understanding the contagion of sexually transmitted diseases such as HIV / AIDS. Some

†Working paper.
other examples are friendships among university freshmen; old-boy networks; cooperation or transmission of information among employees of companies; transactions between companies; and trade relations between countries.

The present paper focuses on directed links (arcs) between entities (nodes), which can be represented by directed graphs (digraphs). It has long been argued in the social sciences (see, e.g., Holland and Leinhardt, 1976) that the arcs—considered as random variables—tend to be dependent. To gain insight into the process that generates such dependent data, it is important to collect longitudinal data. Due to data collection constraints, panel data are the most common form of longitudinal data, that is, the digraph is observed at two or more discrete time points.

Conventional models of digraph panel data, dating back to Holland and Leinhardt (1977), assume that the digraph evolution is governed by a Markov process which operates in continuous time but is observed at discrete time points. Snijders (2001) considered an attractive family of continuous-time Markov models. Its attractiveness stems from the fact that it allows to build models which capture the most important classes of dependencies in social networks and which, at the same time, can readily be communicated to social scientists, because the models have an appealing interpretation in social science terms and can be regarded as directly substantive probability models in the sense of Cox (1990). The basic idea is to model the digraph evolution as a Markov process in continuous time, and let the nodes represent social actors who add and delete arcs with the purpose to obtain the best possible value of some node-specific objective function and random terms. In the model of Snijders (2001), the objective function is a weighted sum of statistics. The statistics depend on the digraph, and may include nodal covariates to account for the fact that nodes are heterogeneous. The weights, regarded as parameters, are assumed to be constant across nodes. The implicit assumption is that all there is to know with regard to nodes is observed in the form of covariates and correctly incorporated in the model. However, it is not unusual that some relevant nodal covariates are unobserved due to data collection constraints and limited prior knowledge of researchers, casting doubt on the constant-weights assumption.

The present paper proposes to account for unobserved heterogeneity across nodes by introducing random variables with latent outcomes, called random effects. Random effects models are widely used in the social sciences (see, e.g., Longford, 1993, Raudenbush and Bryk, 2002, Skrondal and Rabe-Hesketh, 2004); examples of random effects models for non-longitudinal social network data are Hoff (2005), Zijlstra, Van Duijn, and Snijders (2006). The random effects models considered here replace
the constant-weights assumption by the assumption that the weights are unobserved outcomes of nodal random variables, governed by a probability law that is common to all nodes. These models are special cases of hidden Markov models, since the process is, conditional on the random effects (which are assumed to be constant over time), a Markov process.

Maximum likelihood and Bayesian estimation of the parameters is proposed; for maximum likelihood estimation, the non-redundant elements of the random effects variance-covariance matrix are reparametrized so that estimates of the variance-covariance matrix are by construction symmetric and positive definite, and the estimation of very small variances is facilitated. Both maximum likelihood and Bayesian estimation exploit Markov chain Monte Carlo-based data augmentation.

The paper is structured as follows. Section 4.2.1 describes a family of Markov models with fixed effects, while Section 4.2.2 introduces random effects models. Sections 4.3 and 4.4 discuss maximum likelihood and Bayesian estimation, respectively. Section 4.5 applies the model to social network data.

4.2 Model

It is assumed that a binary, directed relation (or digraph) on a finite set of nodes \( N = \{1, \ldots, n\} \) has been observed at discrete, ordered time points \( t_0 < t_1 < \cdots < t_H \). The observations are stored as binary \( n \times n \) matrices \( X(t_0), X(t_1), \ldots, X(t_H) \), where \( X_{ij}(t_h) = 1 \) if there is an arc from node \( i \) to node \( j \) at time point \( t_h \), and \( X_{ij}(t_h) = 0 \) otherwise; the diagonal elements are disregarded and defined as \( X_{ii}(t_h) \equiv 0 \).

The assumption that \( X_{ij}(t_h) \) is binary is made because such data are common and convenient; however, it is possible to extend the model to the case where \( X_{ij}(t_h) \) takes on discrete, ordered values.

4.2.1 Fixed effects models

The digraphs \( X(t_0), X(t_1), \ldots, X(t_H) \) are assumed to be outcomes of a continuous-time Markov process (see, e.g., Resnick, 2002) operating in time interval \( [t_0, t_H] \). Consider the case \( H = 1 \).

The process starts at time \( t \equiv t_0 \) with digraph \( X \equiv X(t_0) \). A holding time \( \Delta t \) is sampled from the negative exponential distribution with parameter \( \phi \), and at time \( t = t + \Delta t \) the digraph \( X \) is allowed to change. Snijders (2001), following Holland and Leinhardt (1977), postulated that one, and only one, element \( X_{ij} \) of \( X \) is allowed
to change, and modeled the change process as driven by nodes $i$—representing social actors—as follows: the parameter $\phi \equiv \phi(X, \theta)$ of the negative exponential distribution is decomposed into node-dependent rates of change $\phi_i(X, \theta)$,

$$\phi \equiv \phi(X, \theta) = \sum_{i=1}^{n} \phi_i(X, \theta),$$

where $\theta$ is a parameter vector. Conditional on the event that $X$ is allowed to change, some node $i \in N$ is chosen with probability

$$\frac{\phi_i(X, \theta)}{\sum_{h=1}^{n} \phi_h(X, \theta)},$$

and the chosen node $i$ is assumed to choose the node $j \in N$ which maximizes

$$f_i(X, j, \theta) + U_{ij}(t),$$

where $f_i(X, j, \theta)$ is called the objective function of $i$ and $U_{ij}(t)$ is a random variable. If the maximum of (4.1) is obtained by choosing $j = i$, then $i$ is assumed to change nothing, otherwise $i$ is assumed to transform the element $X_{ij}$ of $X$ into $1 - X_{ij}$. The process proceeds by updating $t$ and $X$ in the described fashion.

Models can be specified by specifying the rate function $\phi_i(X, \theta)$, the objective function $f_i(X, j, \theta)$, and the distribution of $U_{ij}(t)$.

The rate function $\phi_i(X, \theta)$ may be constant,

$$\phi_i(X, \theta) = \alpha,$$

where $\alpha > 0$ is a parameter, or non-constant,

$$\phi_i(X, \theta) = \alpha \exp [\epsilon' e_i(X)],$$

where $\epsilon$ is a parameter vector and $e_i(X)$ is a statistics vector depending on $X$ and covariates.

In the model of Snijders (2001), the objective function $f_i(X, j, \theta)$ is given by

$$f_i(X, j, \theta) = \eta' s_i(X, j),$$

where $\eta$ is a parameter vector and $s_i(X, j)$ is a statistics vector; statistics can depend on $X$, $j$, and covariates, and can be used to induce dependence among the arc processes $X_{ij}(t)$ (cf. Section 4.5).
4.2. MODEL

It is convenient to assume that the $U_{ij}(t)$ are i.i.d. random variables with Gumbel(0,1) distribution (all $i,j,t$), because then the probability that some given node $i$ chooses node $j \in N$ can be written in closed form (see, e.g., McFadden, 1974) and is given by

$$
\psi_i(j \mid X, \theta) = \frac{\exp[f_i(X, j, \theta)]}{\sum_{h=1}^{n} \exp[f_i(X, h, \theta)]}. \quad (4.3)
$$

Observe that the extension to the case $H \geq 2$ is straightforward because of the Markov property.

4.2.2 Random effects models

The fixed effects models of Section 4.2.1 assume that the weight $\eta$ is constant across nodes. To drop the constant-weights assumption, the objective function (4.2) is replaced by

$$
f_i(X,j, V_i, \theta) = \eta_i s_i(X,j), \quad (4.4)
$$

where the node-specific weight $\eta_i$ is given by

$$
\eta_i = \beta + AV_i.
$$

The vectors $\eta_i$, $\beta$, and $s_i(X,j)$ are of order $L \times 1$, the random effects vectors $V_i$ are of order $K \times 1$ ($K \leq L$), and $A$ is a $L \times K$ design matrix. The parameter vector $\beta$ is constant across nodes, whereas the random effects vectors $V_i$ vary across nodes (but are constant over time), with first- and second-order moments

$$
E[V_i] = 0, \quad i = 1, \ldots, n, \quad (4.5)
$$

$$
E[V_i V_i'] = \Sigma, \quad i = 1, \ldots, n, \quad (4.6)
$$

where assumption (4.5) involves no loss of generality, and $\Sigma$ is positive definite. It is postulated that the random effects are i.i.d. random variables with $K$-variate Gaussian distribution,

$$
V_i \sim N_K(0, \Sigma), \quad i = 1, \ldots, n.
$$

Such models capture both observed nodal heterogeneity—in the form of nodal covariates contained in statistic $s_i(X,j)$—and unobserved nodal heterogeneity in the form of nodal random effects $V_i$. 
CHAPTER 4. RANDOM EFFECTS MODELS

The random effects introduced above depend on the “sender” \( i \) of the arc variable \( X_{ij} \), but random effects may depend in addition on the “receiver” \( j \). Suppose that, if node \( i \) is allowed to change something, \( i \) chooses the node \( j \in N \) which maximizes

\[
f_i(X, j, V_i, \theta) + T_j + U_{ij}(t),
\]

where \( T_j \) is a scalar-valued random effect. The random effect \( T_j \) can be interpreted as the latent popularity of \( j \), and is an alternative to other representations of popularity in terms of \( f_i(X, j, V_i, \theta) \).

4.3 Maximum likelihood estimation

A complete observation of the continuous-time Markov process in time interval \([t_0, t_1]\) corresponds to digraphs \( X_0, X_1, \ldots, X_{M-1}, X_M \), and holding times, where \( M \) denotes the total number of times \( X \) was allowed to change in time interval \([t_0, t_1]\). The digraphs \( X_0 \equiv X(t_0) \) and \( X_M \equiv X(t_1) \) are observed and denoted by \( y \). The digraphs \( X_1, \ldots, X_{M-1} \) are unobserved and can be represented by \( X_0 \) and the sequence \( w = (i_m, j_m)_{m=1}^M \), where \( i_m \) is the node that was allowed to change something, and \( j_m \) is the node chosen by \( i_m \). The holding times are unobserved, but can be disregarded by using the embedded Markov process, corresponding to \( X_0 \) and \( w \). The random effects \( V_i \) are unobserved, and stored as rows of matrix \( V \). The unobserved data \( V \) and \( w \) are referred to as \( z \).

To avoid space-consuming complications which do not alter the basic argumentation, it is assumed that the rate function is given by \( \phi_i(X, \theta) = \alpha \); the case of non-constant rate functions can be obtained along the lines of Snijders, Koskinen, and Schweinberger (2006), who considered maximum likelihood estimation of fixed effects models (see Section 4.2.1). The parameters \( \alpha, \beta, \) and \( \Sigma^{-1} \) are collected in the parameter vector \( \theta \) of dimension \( L \).

Under regularity conditions, the maximum likelihood estimate (MLE) \( \hat{\theta} \) of \( \theta \) solves

\[
\nabla_\theta \ln p_\theta(y) = 0,
\]

where \( \nabla_\theta \ln p_\theta(y) \) is the gradient of \( \ln p_\theta(y) \) with respect to \( \theta \), and \( p_\theta(y) \) is the probability density of \( y \). The problem is that \( \nabla_\theta \ln p_\theta(y) \) is not available in closed form.

In the incomplete-data literature, a result of Fisher (1925) (noted by Efron, 1977) turned out to be useful in dealing with such intractable estimation problems. Observe that

\[
\nabla_\theta p_\theta(y) = \nabla_\theta \int p_\theta(y, z) \, d\mu(z) = \int \nabla_\theta p_\theta(y, z) \, d\mu(z),
\]

(4.8)
where interchanging the order of differentiation and integration is admissible by Theorem 2.7.1 of Lehmann and Romano (2005, p. 49) and the fact that \( p_\theta(y, z) \) is an exponential family density (see Section 4.2). By multiplying (4.8) under the integral sign by \( p_\theta(z \mid y) p_\theta(y) / p_\theta(y, z) \), one obtains the so-called Fisher identity:

\[
\nabla_\theta \ln p_\theta(y) = E_\theta [\nabla_\theta \ln p_\theta(y, Z) \mid y].
\]

Thus, solving (4.7) is equivalent to solving

\[
E_\theta [\nabla_\theta \ln p_\theta(y, Z) \mid y] = 0.
\]

Analytical evaluation of the expectation in (4.10) is infeasible, which rules out the use of standard root-finding methods such as Newton-Raphson to solve (4.10). However, if sampling from \( p_\theta(z \mid y) \) is possible, then the expectation in (4.10) can be approximated and root-finding algorithms based on stochastic approximation (Robbins and Monro, 1951, Chen, 2002) can be used to solve (4.10). Gu and Kong (1998) proposed to use Markov chain Monte Carlo (MCMC) methods (Hastings, 1970, Tierney, 1994) to sample from the conditional distribution of the latent data. Here, MCMC sampling from \( p_\theta(z \mid y) \) is possible, thus stochastic approximation with MCMC-based data augmentation can be used to solve (4.10) by iterating the following steps:

I. Data augmentation step: sample \( z_N \mid y \).

II. Parameter updating step:

\[
\hat{\theta}_N = \hat{\theta}_{N-1} + a_{N-1} B^{-1} \nabla_{\hat{\theta}_{N-1}} \ln p_{\hat{\theta}_{N-1}}(y, z_N),
\]

where \( z_N \) has density \( p_{\theta_{N-1}}(z_N \mid y) \), and \( a_{N-1} \) is a sequence of positive numbers tending to 0. A possible choice of \( L \times L \) matrix \( B \) is given by \( B = -E_\hat{\theta} [\nabla^2_{\hat{\theta}} \ln p_{\hat{\theta}}(y, Z) \mid y] \), where \( \nabla^2_{\hat{\theta}} \ln p_{\hat{\theta}}(y, z) \) is the Hessian matrix of \( \ln p_{\hat{\theta}}(y, z) \) with respect to \( \hat{\theta} \); \( B \) can be estimated by sampling from \( p_{\hat{\theta}_0}(z \mid y) \), where the unknown MLE \( \hat{\theta} \) is replaced by an initial estimate \( \hat{\theta}_0 \), and estimating \( B \) by the corresponding average. A sensible modification of the algorithm is based on the averaging approach (see, e.g., Yin, 1991, and references therein), which is a multi-stage approach that generates in each stage interim estimates \( \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N \) with \( a_1 = a_2 = \cdots = a_N \) according to the iterative scheme above, and replaces \( \hat{\theta}_N \) by \( (1/N) \sum_{i=1}^N \hat{\theta}_i \); stage \( m \geq 2 \) starts with the average of \( \hat{\theta}_i \) based on stage \( m - 1 \), with increased \( N \) and decreased \( a_N \). Under regularity conditions, such stochastic approximation estimators \( \hat{\theta}_N \) converge to the solution of (4.10) (Yin, 1991), which is the MLE \( \hat{\theta} \).
When having obtained an estimate of the MLE $\hat{\theta}$ by stochastic approximation, the observed information matrix can be calculated as follows (Louis, 1982):

$$-\nabla^2_{\theta} \ln p_\theta(y) = -E_{\theta}[\nabla^2_{\theta} \ln p_\theta(y, Z) \mid y]$$

$$- [E_{\theta}[\nabla_{\theta} \ln p_\theta(y, Z)(\nabla_{\theta} \ln p_\theta(y, Z))' \mid y] - \nabla_{\theta} \ln p_\theta(y)(\nabla_{\theta} \ln p_\theta(y))'],$$

(4.11)

where the integrals on the right-hand side of (4.11) can be estimated by sampling from $p_\theta(z \mid y)$ and estimating the integrals by the corresponding averages.

To make the stochastic approximation algorithm with MCMC-based data augmentation operational, MCMC samples from $p_\theta(z \mid y)$ are required, which is discussed in Section 4.3.1; furthermore, three entities are needed in closed form, the probability density $p_\theta(y, z)$ and the gradient and Hessian matrix of $\ln p_\theta(y, z)$ with respect to $\theta$; these are given in Section 4.3.2, along with a reparametrization of $\Sigma^{-1}$.

### 4.3.1 Implementation: data augmentation

To sample $z \mid y$, where $z = (V, w)$, it is convenient to use MCMC cycle algorithms (Hastings, 1970, Tierney, 1994), which iterate, for a given value of $\theta$, the following steps:

(a) Sample $w^{new} \mid y, V$.

(b) Sample $V^{new} \mid y, w^{new}$.

Steps (a) and (b) can be implemented by using the Metropolis-Hastings method (Hastings, 1970) as follows.

Metropolis-Hastings step to sample $w^{new} \mid y, V$. A candidate $w^*$ is generated from a distribution $a(w^* \mid w)$, and $w^*$ is accepted with probability

$$\pi = \min \left[ 1, \frac{p_\theta(w^* \mid y, V)}{p_\theta(w \mid y, V)} \times \frac{a(w \mid w^*)}{a(w^* \mid w)} \right].$$

The candidate-generating distribution $a(w^* \mid w)$ is identical to the candidate-generating distribution of Snijders et al. (2006), which is a probability distribution defined on a discrete set of simple proposals, corresponding to inserting and deleting $(i_m, j_m)$'s in the sequence $w = (i_m, j_m)_{m=1}^M$ or permuting sub-sequences of $w$, subject to the constraint that $X_0 \equiv X(t_0)$ and $X_M \equiv X(t_1)$.

Metropolis-Hastings step to sample $V^{new} \mid y, w^{new}$. For $i = 1, \ldots, n$ independently, a candidate $V_i^*$ is generated from a distribution $b(V_i^* \mid V_i)$, and $V_i^*$ is accepted with
probability
\[ \pi = \min \left[ 1, \frac{p_0(V_i^* \mid y, V_{-i}, w_{\text{new}})}{p_0(V_i \mid y, V_{-i}, w_{\text{new}})} \times \frac{b(V_i \mid V_i^*)}{b(V_i^* \mid V_i)} \right], \]

where \( V_{-i} \) corresponds to the random effects matrix \( V \) without row \( i \). Possible candidates are given by \( V_i^* = a + B(V_i - a) + G_i \), where \( a \) is a constant \( K \times 1 \) vector, \( B \) is a constant \( K \times K \) matrix, and \( G_i \sim N_K(0, \Omega) \). Let \( I_K \) be the \( K \times K \) identity matrix. The choice \( B = I_K \) corresponds to (i) random walk Metropolis-Hastings algorithms; \( B = 0 \) corresponds to (ii) independence samplers, in which case \( a = 0 \) is convenient, the prior expectation of \( V_i \) (see Section 4.2.2); and \( B = -I_K \) corresponds to (iii) first-order autoregressive Metropolis-Hastings algorithms, which reflect \( V_i \) about \( a \) before adding increment \( G_i \) and can reduce MCMC autocorrelations; \( a \) can be calibrated during burn-in iterations. The choice of scale matrix \( \Omega \) can be based on either burn-in iterations or Bayesian point estimates of \( \Sigma \) (see Section 4.4). In most instances, independence samplers (ii) turn out to produce samples with lower MCMC autocorrelations than (i) and (iii).

To make the Metropolis-Hastings algorithms operational, the probability density \( p_0(y, z) \) is required, which is presented in Section 4.3.2.

### 4.3.2 Implementation: closed-form expressions

The probability density \( p_0(y, z) \) and the gradient and Hessian matrix of \( \ln p_0(y, z) \) with respect to \( \theta \) are derived below.

The model of Section 4.2—with rate function \( \phi_i(X, \theta) = \alpha \)—implies that the probability density \( p_0(y, z) \) is given by

\[
\begin{align*}
p_\theta(y, z) &= \alpha^M \exp\left[-n\alpha(t_1 - t_0)\right] \\
&\quad \times \prod_{m=1}^{M} \psi_{i_m}(j_m \mid X_{m-1}, V_{i_m}, \theta) \\
&\quad \times \prod_{i=1}^{n} (2\pi)^{-K/2} \det[\Sigma^{-1}]^{1/2} \exp\left[-\frac{1}{2} V_i' \Sigma^{-1} V_i\right],
\end{align*}
\]

where \( \psi_{i_m}(j_m \mid X_{m-1}, V_{i_m}, \theta) \) is based on \( f_{i_m}(X_{m-1}, j_m, V_{i_m}, \theta) \) as defined by (4.4)—as a natural extension of (4.3)—and \( \det[\Sigma^{-1}] \) is the determinant of \( \Sigma^{-1} \).

The gradient and Hessian matrix of \( \ln p_\theta(y, z) \) with respect to \( \alpha \) and \( \beta \) can readily be obtained.
To impose symmetry and positive definiteness constraints on estimates of $\Sigma^{-1}$, it is sensible to use an unconstrained parametrization of $\Sigma^{-1}$, that is, to reparametrize the non-redundant elements of $\Sigma^{-1}$ so that estimates of $\Sigma^{-1}$ are by construction symmetric and positive definite.

Let $\Sigma^{-1} = \Gamma\Gamma'$, where $\Gamma$ is a $K \times K$ lower triangular matrix restricted by $\gamma_{ij} = 0$ (all $j > i$), called a Cholesky factor of $\Sigma^{-1}$. Estimating $\Gamma$ itself is associated with the drawback that $\Gamma$ is not globally identifiable due to the fact that any column of $\Gamma$ can be multiplied by $-1$ without changing the value of the likelihood.

Therefore, let $\Gamma = \Delta\Lambda^{1/2}$, where $\Delta$ is a $K \times K$ lower triangular matrix constrained by $\delta_{ii} = 1$ (all $i$) and $\delta_{ij} = 0$ (all $j > i$), $\Lambda$ is a $K \times K$ diagonal matrix with elements $\lambda_{ii} = \exp[\xi_{ii}]$ on the main diagonal, and the elements $\xi_{ii}$ are stored on the main diagonal of a $K \times K$ diagonal matrix $\Xi$; for related parametrizations, see Hedeker and Gibbons (1996), Pinheiro and Bates (1996). It is evident that estimating $\xi_{ii} = \ln \lambda_{ii}$ produces positive estimates of $\lambda_{ii}$ by construction, and facilitates the estimation of very small variances. To disregard the constant elements 0, 1 of $\Delta$ and $\Xi$, let $v(\Delta)$ and $v(\Xi)$ be the vectors obtained from vec($\Delta$) and vec($\Xi$) by eliminating all constant elements, respectively, where vec is the vec operator which transforms its matrix argument into a column vector by stacking the columns of the matrix one underneath the other; and let $D(\Delta)$ and $D(\Xi)$ be the unique matrices satisfying $D(\Delta) v(\Delta) + \text{vec}(I_K) = \text{vec}(\Delta)$ and $D(\Xi) v(\Xi) = \text{vec}(\Xi)$, respectively, where $I_K$ is the $K \times K$ identity matrix; note that vec($\Delta$) is an affine rather than a linear function of $v(\Delta)$ because of the constraint $\delta_{ii} = 1$ (all $i$).

The vectors $v(\Delta)$ and $v(\Xi)$ are the parameters to be estimated. The gradient and Hessian matrix of $\ln p_\theta(y, z)$ with respect to $v(\Delta)$ and $v(\Xi)$ are derived in the Appendix and are given by

$$\nabla_{v(\Delta)} \ln p_\theta(y, z) = D'(\Delta) \text{vec}(A(\Sigma) \Gamma\Lambda^{1/2}), \quad (4.13)$$

$$\nabla_{v(\Xi)} \ln p_\theta(y, z) = \frac{1}{2} D'(\Xi) \text{vec}(\Gamma' A(\Sigma) \Gamma), \quad (4.14)$$

$$\nabla^2_{v(\Delta) v(\Xi)} \ln p_\theta(y, z) = \begin{pmatrix} \nabla^2_{\Delta\Delta} & \nabla^2_{\Delta\Xi} \\ \nabla^2_{\Xi\Delta} & \nabla^2_{\Xi\Xi} \end{pmatrix}, \quad (4.15)$$

where

$$\nabla^2_{\Delta\Delta} = -D'(\Delta) [n E_K((\Delta^{-1})' \otimes \Delta^{-1}) + (\Lambda \otimes H)] D(\Delta),$$

$$\nabla^2_{\Delta\Xi} = (\nabla^2_{\Delta\Delta})' = -2D'(\Delta)(\Lambda \otimes H\Delta)D(\Xi),$$

$$\nabla^2_{\Xi\Delta} = \frac{1}{2} D'(\Xi)(\Lambda \otimes \Delta' H\Delta)D(\Xi),$$

$$\nabla^2_{\Xi\Xi} = \frac{1}{2} D'(\Xi)(\Lambda \otimes \Delta H\Delta')D(\Xi).$$
4.4. BAYESIAN ESTIMATION

where \( A(\Sigma) = n\Sigma - H \) and \( H = \sum_{i=1}^{n} V_i V_i' \), while \( E_K \) is the unique permutation matrix satisfying \( E_K \text{vec}(C) = \text{vec}(C') \) for every \( K \times K \) matrix \( C \) (cf. Magnus, 1988, p. 35), and \( \otimes \) is the Kronecker product.

The blocks of the Hessian matrix \( \nabla^2_{\theta} \ln p_\theta(y, z) \) involving \( \alpha, \beta \) on one hand and \( v(\Delta), v(\Xi) \) on the other hand vanish.

When having obtained an estimate of the MLE \( \hat{\theta} \) by stochastic approximation, the observed information matrix (4.11) can be estimated by sampling from \( p_\theta(z \mid y) \). Due to the parametrization-invariance of MLEs, the observed information matrix for interesting functions of \( v(\Delta) \) and \( v(\Xi) \)—such as \( v(\Sigma) \), where \( v(\Sigma) \) is obtained from \( \text{vec}(\Sigma) \) by eliminating the elements \( \sigma_{ij} \) (all \( j > i \))—can be estimated from the same MCMC sample. Estimating the observed information matrix of \( \theta \), where \( v(\Delta) \) and \( v(\Xi) \) are replaced by \( v(\Sigma) \), requires the gradient and Hessian matrix of \( \ln p_\theta(y, z) \) with respect to \( v(\Sigma) \), which are derived in the Appendix and are given by

\[
\begin{align*}
\nabla_{V(\Sigma)} \ln p_\theta(y, z) & = -\frac{1}{2} D'_{(\Sigma)} \text{vec}(\Sigma^{-1} A(\Sigma) \Sigma^{-1}), \\
\nabla^2_{V(\Sigma)} \ln p_\theta(y, z) & = -D'_{(\Sigma)} \left[ (\Sigma^{-1} \otimes \Sigma^{-1} H \Sigma^{-1}) - \frac{n}{2} (\Sigma^{-1} \otimes \Sigma^{-1}) \right] D_{(\Sigma)},
\end{align*}
\]

where \( D_{(\Sigma)} \) is the unique matrix satisfying \( D_{(\Sigma)} v(\Sigma) = \text{vec}(\Sigma) \) for every symmetric \( K \times K \) matrix \( \Sigma \) (cf. Magnus, 1988, p. 55).

4.4 Bayesian estimation

Bayesian inference concerning \( \theta \) is based on the posterior probability density \( p(\theta \mid y) \propto p_\theta(y) p(\theta) \), where \( p(\theta) \) is the prior density of \( \theta \). Concerning \( p(\theta) \), it is, in most applications, reasonable to start with the assumption of prior independence of \( \alpha, \beta, \) and \( \Sigma^{-1} \); convenient families of prior distributions are

\[
\begin{align*}
\alpha & \sim \text{Gamma}(\gamma, \delta), \\
\beta & \sim N_L(0, \Psi), \\
\Sigma^{-1} & \sim \text{Wishart}(\varphi, \Omega).
\end{align*}
\]

The posterior distribution is intractable, but samples from the posterior distribution can be obtained by iterating the following MCMC steps:

I. Data augmentation step: sample \( z_N \mid \alpha_{N-1}, \beta_{N-1}, \Sigma_{N-1}^{-1}, y \).

II. Posterior step:

(a) Sample \( \Sigma_N^{-1} \mid \alpha_{N-1}, \beta_{N-1}, y, z_N \).
(b) Sample \( \alpha_N \mid \beta_{N-1}, \Sigma_{N-1}^{-1}, y, z_N \).

(c) Sample \( \beta_N \mid \alpha_N, \Sigma_N^{-1}, y, z_N \).

The MCMC steps are elaborated below.

Metropolis-Hastings step to sample \( z_N \mid \alpha_{N-1}, \beta_{N-1}, \Sigma_{N-1}^{-1}, y \). One iteration of the cycle algorithm sketched in Section 4.3.1 is sufficient.

Gibbs sampling of \( \Sigma^{-1}_N \mid \alpha_{N-1}, \beta_{N-1}, y, z_N \). If the prior of \( \Sigma^{-1} \) is Wishart(\( \varphi, \Omega \)) parametrized such that \( E_{\varphi, \Omega}[\Sigma^{-1}] = \varphi \Omega \), then the full conditional posterior is Wishart(\( \varphi + n, (H + \Omega^{-1})^{-1} \)) \( (H = \sum_{i=1}^n V_i V_i') \), which can be sampled.

Gibbs sampling of \( \alpha_N \mid \beta_{N-1}, \Sigma_N^{-1}, y, z_N \). If the prior of \( \alpha \) is Gamma(\( \gamma, \delta \)) parametrized such that \( E_{\gamma, \delta}[\alpha] = \gamma/\delta \), then the full conditional posterior of \( \alpha \) is Gamma(\( \gamma + M, \delta + n(t_1 - t_0) \)), which can be sampled.

Metropolis-Hastings step to sample \( \beta_N \mid \alpha_N, \Sigma_N^{-1}, y, z_N \). A candidate \( \beta^* \) is generated from a distribution \( b(\beta^* \mid \beta_{N-1}) \), and \( \beta^* \) is accepted with probability

\[
\pi = \min \left[ 1, \frac{p(\beta^* \mid \alpha_N, \Sigma_N^{-1}, y, z_N)}{p(\beta_{N-1} \mid \alpha_N, \Sigma_N^{-1}, y, z_N)} \times \frac{b(\beta_{N-1} \mid \beta^*)}{b(\beta^* \mid \beta_{N-1})} \right].
\]

Convenient candidates are given by \( \beta^* = a + B(\beta_{N-1} - a) + G \), where \( a \) is a constant \( L \times 1 \) vector, \( B \) is a constant \( L \times L \) matrix, and \( G \sim N_L(0, \Omega) \). Let \( I_L \) be the \( L \times L \) identity matrix. Possible choices of \( B \) are (i) \( B = I_L \) (random walk Metropolis-Hastings algorithms); (ii) \( B = 0 \) (independence samplers); and (iii) \( B = -I_L \) (first-order autoregressive Metropolis-Hastings algorithms), which reflects \( \beta_{N-1} \) about \( a \) before adding increment \( G \) and can reduce MCMC autocorrelations. In case of (ii) and (iii), \( a \) can be based on point estimates of \( \beta \), such as method of moments estimates (see Snijders, 2001, assuming \( \Sigma \equiv 0 \)) or the MLE (see Section 4.3). The choice of scale matrix \( \Omega \) can be based on either burn-in iterations or the inverse observed information matrix of \( \beta \) at the MLE of \( \beta \) (see Section 4.3).

4.5 Application

The model is applied to data collected as part of the Teenage Friends and Lifestyle Study (Pearson and West, 2003), corresponding to friendships among 160 students
of a Scottish school cohort observed at three time points \( t_0 < t_1 < t_2 \) between 1995 and 1997. 129 students were present at all three time points, among which 56 girls. Here, the friendships among the \( n = 56 \) girls are studied, corresponding to \( 56 \times 56 \) matrices \( X(t_0) \), \( X(t_1) \), and \( X(t_2) \), where \( X_{ij}(t_h) = 1 \) if girl \( i \) called girl \( j \) a friend at time point \( t_h \), and \( X_{ij}(t_h) = 0 \) otherwise. At time point \( t_0 \), 156 of \( 56(56-1) = 3,080 \) possible arcs were observed; in time interval \([t_{h-1}, t_h]\) \((h = 1, 2)\), 79 and 65 arcs were added while 75 and 68 arcs were deleted, respectively.

Let \( X \equiv X(t) \) be the digraph at time \( t \in [t_0, t_2] \), let \( i \) be the girl that is allowed to change something, let \( j \) be the girl chosen by \( i \), let \( X_{ij}^* = X_{ij} \) if \( i = j \) and \( X_{ij}^* = 1 - X_{ij} \) otherwise, while \( X_{kl}^* = X_{kl} \) (all \((k, l) \neq (i, j))\), and let \( I_K \) be the \( K \times K \) identity matrix. A simple model is constructed by using rate functions \( \phi_i(X, \theta) = \alpha_h \) in time interval \([t_{h-1}, t_h]\) \((h = 1, 2)\) and assuming that the parameters \( \beta \) and \( \Sigma \) are constant across time intervals. Interesting components \( s_{ik}(X, j) \) of statistics vector \( s_i(X, j) \) in objective function \( f_i(X, j, V_i, \theta) \) are given by

\[
\begin{align*}
    s_{i1}(X, j) &= \sum_{l=1}^{n} X_{il}^* \text{ (number of arcs)}, \\
    s_{i2}(X, j) &= \sum_{l=1}^{n} X_{il}^* X_{lj}^* \text{ (number of reciprocated arcs)}, \\
    s_{i3}(X, j) &= \sum_{h=1}^{n} \sum_{l=1, l \neq h}^{n} X_{ih}^* X_{hl}^* X_{il}^* \text{ (number of transitive triplets)}, \\
    s_{i4}(X, j) &= \sum_{l=1}^{n} X_{il}^* c_i(t_{h-1}) \text{ (number of arcs weighted by covariate } c_i(t_{h-1}))
\end{align*}
\]

which is equivalent to specifying \( f_i(X, j, V_i, \theta) \) as

\[
f_i(X, j, V_i, \theta) = \gamma_{ij} (X_{ij}^* - X_{ij}),
\]

where the weight \( \gamma_{ij} \) is given by

\[
\gamma_{ij} = \eta_{i1} + \eta_{i2} X_{ji} + \eta_{i3} (\sum_{h=1, h \neq j}^{n} X_{ih} X_{hj} + \sum_{l=1, l \neq j}^{n} X_{il} X_{jl}) + \eta_{i4} c_j(t_{h-1}).
\]

The covariate \( c_j(t_{h-1}) \) refers to the amount of financial resources of girl \( j \) as recorded at time point \( t_{h-1} \), which may be regarded as an indicator of social-economic status; in time interval \([t_{h-1}, t_h]\) \((h = 1, 2)\), \( s_{i4}(X, j) \) was computed by using \( c_j(t_{h-1}) \), and \( c_j(t_{h-1}) \) was centered at 0 and rescaled to variance 1. The most important candidate for girl-dependent weights \( \eta_{ik} \) is \( \eta_{i1} \), because \( \eta_{i1} \) captures one of the most fundamental features of the data—the distribution of the number of arcs \( \sum_{l=1}^{n} X_{il}(t_h) \) at time point \( t_h \)—and almost all models used in applications include \( \eta_{i1} \); to keep the model as simple and parsimonious as possible, the remaining weights \( \eta_{ik} \) \((k = 2, 3, 4)\) are assumed to be constant across girls, \( \eta_{ik} = \beta_k \) \((k = 2, 3, 4)\). In the Bayesian framework,
the priors are (a) Gamma(1, 10^{−10}) for \( \alpha_h \) \( (h = 1, 2) \); (b) \( N_4(0, 10^{10} I_4) \) for \( \beta \); and (c) Wishart\( (1, 10 I_1) \) for \( \Sigma^{-1} \), implying that large variances are a priori unlikely, motivated (as in generalized linear mixed models) by the fact that the weights \( \eta_i = \beta + AV_i \) enter the model through exponential functions (see (4.3)) and thus large weights—and, as a result, large variances—are implausible.

The MLE \( \hat{\theta} \) of parameter vector \( \theta \) was estimated by a five-stage stochastic approximation algorithm with 3,300 iterations in total, as sketched in Section 4.3 for random effects models and more detailed in Snijders, Koskinen, and Schweinberger (2006) for fixed effects models. As starting values of \( \Sigma \equiv \sigma^2, .20, .10, \) and \( .05 \) were used, which were thought to be reasonable values. A trace plot of the interim estimates of the MLE \( \hat{\xi} \) of \( \xi = -\ln(\sigma^2) \) is shown in Figure 4.1; note that at the end of each stage (indicated by a vertical line), the present interim estimate is replaced by the average of interim estimates of the preceding stage, and the final stochastic approximation estimate \( \hat{\xi}_N \) is the average of interim estimates of the fifth stage. Figure

**Figure 4.1: Trace plot of interim estimates of MLE \( \hat{\xi} \) using five-stage stochastic approximation algorithms with multiple starting values**

![Trace plot of interim estimates of MLE \( \hat{\xi} \)](image)

A vertical line indicates the end of a stage, at which point the present interim estimate is replaced by the average of interim estimates of the preceding stage.

4.1 shows that, from the second stage on, most interim estimates are in a neighborhood of 2.6. To detect non-convergence of the stochastic approximation estimates \( \hat{\xi}_N \), an alternative to using multiple starting values is based on the Fisher identity
(4.9): the so-called $t$-ratio of the average of the complete-data score $\nabla_{\xi_N} \ln p_{\theta_N}(y, z)$ across augmented sets of data $(y, z)$ to its standard deviation can be used to diagnose non-convergence; for each of the three estimation runs, 2,000 augmented sets of data $(y, z)$ were generated conditional on $y$ and the final stochastic approximation estimate $\hat{\theta}_N$ of $\hat{\theta}$, and the $t$-ratio was computed. The estimation runs with starting values .20, .10, and .05 produced $t$-ratios .002, .012, and .020, respectively, which are small enough for practical purposes and do not suggest non-convergence. It is interesting to see in what region of the marginal posterior density of $\sigma^2$ the three resulting estimates of the MLE $\hat{\sigma}^2$ of $\sigma^2$ are located. The marginal posterior density of $\sigma^2$ was estimated by the Bayesian methods of Section 4.4 with 220,000 iterations, where the first 20,000 iterations were discarded as burn-in iterations and every 40th sampled value of the last 200,000 iterations was recorded; non-convergence was checked by using trace plots and convergence checks of Raftery and Lewis (1996). The marginal

Figure 4.2: Marginal posterior density of $\sigma^2$ under two priors

--- estimates of MLE $\hat{\sigma}^2$; · · · · · · .025, .500, and .975 posterior quantiles of $\sigma^2$. 

posterior density of $\sigma^2$ is shown in Figure 4.2, and it is evident that the three estimates of $\hat{\sigma}^2$ are close to the mode of the marginal posterior density of $\sigma^2$. However, a relevant question is how sensitive the posterior is to the prior. An alternative to the chosen Wishart($1, 10 I_1$) prior of $\sigma^{-2}$ is Wishart($1, 5 I_1$): the resulting marginal posterior density of $\sigma^2$ is depicted in Figure 4.2, and a tentative conclusion is that the posterior seems to be not too sensitive to the prior (within the Wishart family of priors).

To interpret the parameters, 95% posterior intervals, the posterior median, and the MLE of $\theta$ are shown in Table 4.1. The evidence with respect to $\beta$ may be interpreted as follows: if an average girl $i$—average in the sense that $V_{i1} = E[V_{i1}] = 0$ (the prior expectation of $V_{i1}$, see Section 4.2.2) and thus $\eta_{i1} = \beta_1$—does not consider girl $j$ to be a friend ($X_{ij} = 0$), then $i$ will not tend to establish a friendship to $j$ (negative values of $\beta_1$) unless there are good reasons (cf. (4.18)): such as $j$ considering $i$ to be a friend ($X_{ji} = 1$), $j$ being the friend of at least one friend of $i$ ($X_{ih}X_{hj} = 1$), $j$ sharing at least one friend with $i$ ($X_{il}X_{jl} = 1$), or $j$ having high social-economic status—all increasing the likelihood that $i$ establishes a friendship to $j$; if $i$ considers $j$ to be a friend ($X_{ij} = 1$), then the same mechanisms decrease the likelihood that $i$ cancels the friendship to $j$. The magnitude of $\sigma^2$ hints that there is non-negligible variation among girls with respect to the weights $\eta_{i1} = \beta_1 + V_{i1}$, but at the same time the weights $\eta_{i1}$ still appear to be negative, which is confirmed by the posterior medians of $\eta_{i1}$ ($i = 1, \ldots, n$) shown in Figure 4.3.

### Table 4.1: Estimates of $\theta$

<table>
<thead>
<tr>
<th></th>
<th>95% posterior interval</th>
<th>posterior median</th>
<th>MLE (s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_1$</td>
<td>[5.000, 7.861]</td>
<td>6.294</td>
<td>6.421 (.725)</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>[4.193, 6.994]</td>
<td>5.370</td>
<td>5.389 (.653)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>[-2.831, -2.402]</td>
<td>-2.600</td>
<td>-2.560 (.104)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>[1.776, 2.419]</td>
<td>2.089</td>
<td>2.042 (.171)</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>[.351, .503]</td>
<td>.424</td>
<td>.410 (.036)</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>[.003, .233]</td>
<td>.121</td>
<td>.118 (.056)</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>[.024, .264]</td>
<td>.094</td>
<td>.067 (.055)</td>
</tr>
</tbody>
</table>

Bayesian estimates are based on the Wishart($1, 10 I_1$) prior of $\sigma^{-2}$; the MLE of $\theta$ is based on the estimation run with .20 as starting value of $\sigma^2$. 
4.6 Discussion

A family of models for longitudinal social network data was proposed which represents unobserved heterogeneity across nodes by random variables with unobserved outcomes (or random effects), and parameter estimation was considered in a maximum likelihood and a Bayesian framework.

Concerning maximum likelihood estimation, the use of the inverse observed information matrix as an approximation of the variance-covariance matrix of the MLE makes implicit use of large-sample theory. The asymptotic properties of the MLE are unknown, and deriving them is beyond the scope of the present paper. However, for many social network models properties of estimators are still unknown, and much more work needs to be done to clarify theoretical issues arising from the application of statistical models to social network data; see, for instance, the discussion of Hunter and Handcock (2006) in the framework of curved exponential random graph models for social networks.

Another important area of future research is the development of model selection tools: while posterior predictive checks (under fixed effects models) can be used (a) to assess how well (posterior) predicted data match the observed data, using suitable summary statistics, and (b) to suggest what random effects should be included, the choice of summary statistics is arbitrary, and, more importantly, the increase in model complexity by adding random effects is not taken into account.
Appendix: Gradients and Hessian matrices

Using the notation of Section 4.3.2, the gradients and Hessian matrices of $\ln p_\theta(y, z)$ with respect to $v(\Delta)$, $v(\Xi)$, and $v(\Sigma)$ are derived based on the theory of matrix differential calculus of Magnus and Neudecker (1988), abbreviated as “MN”. Twice differentiability is assumed throughout.

Observe that $\ln p_\theta(y, z)$ can be written as

$$\ln p_\theta(y, z) = \frac{n}{2} \ln \det[\Sigma^{-1}] - \frac{1}{2} \text{tr}(H\Sigma^{-1}) + c,$$

where $c$ is a scalar not depending on $\Sigma^{-1}$, and $\text{tr}(H\Sigma^{-1})$ is the trace of $H\Sigma^{-1}$.

**Lemma 1.** The first differential of $\ln p_\theta(y, z)$ with respect to $v(\Delta)$ and $v(\Xi)$ is given by

$$d \ln p_\theta(y, z) = (\text{vec}(A(\Sigma) \Gamma \Lambda^{1/2}))' D_{(\Delta)} d v(\Delta) + \frac{1}{2} (\text{vec}(\Gamma' A(\Sigma) \Gamma))' D_{(\Xi)} d v(\Xi).$$

**Proof of Lemma 1.** Using $d \ln \det[\Sigma^{-1}] = \text{tr}(\Sigma d \Sigma^{-1})$ (MN, Theorem 2, pp. 150–151),

$$d \ln p_\theta(y, z) = \frac{n}{2} d \ln \det[\Sigma^{-1}] - \frac{1}{2} \text{tr}(H d \Sigma^{-1}) = \frac{1}{2} \text{tr}(A(\Sigma) d \Sigma^{-1}).$$

Using $\Sigma^{-1} = \Delta \Lambda \Delta'$, applying Cauchy’s rule of invariance (MN, Theorem 13, p. 96), and using $\text{tr}(BC) = (\text{vec}(B'))' \text{vec}(C)$ for $K \times K$ matrices $B, C$,

$$d \ln p_\theta(y, z) = \frac{1}{2} \text{tr}(A(\Sigma) d(\Delta \Lambda \Delta')) = (\text{vec}(A(\Sigma) \Lambda^{1/2}))' d \text{vec}(\Delta) + \frac{1}{2} (\text{vec}(\Gamma' A(\Sigma) \Gamma))' d \text{vec}(\Xi).$$

Using $\text{vec}(\Delta) = D_{(\Delta)} v(\Delta) + \text{vec}(I_K)$ and $\text{vec}(\Xi) = D_{(\Xi)} v(\Xi)$ and applying Cauchy’s rule of invariance completes the proof. □

**Lemma 2.** The first differential of $\ln p_\theta(y, z)$ with respect to $v(\Sigma)$ is given by

$$d \ln p_\theta(y, z) = -\frac{1}{2} (\text{vec}(\Sigma^{-1} A(\Sigma) \Sigma^{-1}))' D_{(\Sigma)} d v(\Sigma).$$

**Proof of Lemma 2.** Using $d \Sigma^{-1} = -\Sigma^{-1}(d \Sigma)\Sigma^{-1}$ (MN, Theorem 3, p. 151),

$$d \ln p_\theta(y, z) = -\frac{n}{2} d \ln \det[\Sigma] - \frac{1}{2} \text{tr}(H d \Sigma^{-1}) = -\frac{1}{2} (\text{vec}(\Sigma^{-1} A(\Sigma) \Sigma^{-1}))' d \text{vec}(\Sigma).$$
Using vec(Σ) = D_Σ v(Σ) and applying Cauchy’s rule of invariance completes the proof. □

**Lemma 3.** The second differential of \( \ln p_\theta(y, z) \) with respect to \( v(\Delta) \) and \( v(\Xi) \) is given by

\[
\begin{align*}
d^2 \ln p_\theta(y, z) &= -(d v(\Delta))^\prime D_\Delta' \left[ nE_K((\Delta^{-1})' \otimes \Delta^{-1}) + (\Lambda \otimes H) \right] D_\Delta d v(\Delta) \\
&\quad -2(d v(\Delta))^\prime D_\Delta' (\Lambda \otimes H \Delta) D_\Xi d v(\Xi) \\
&\quad -\frac{1}{2} (d v(\Xi))^\prime D_\Xi' (\Lambda \otimes \Delta' H \Delta) D_\Xi d v(\Xi). \\
\end{align*}
\]

**Proof of Lemma 3.** Note that

\[
\begin{align*}
d^2 \ln p_\theta(y, z) &= \frac{n}{2} d(d \ln \det[\Delta \Lambda \Delta']) - \frac{1}{2} d(d \tr(\Delta' H \Delta \Lambda)). \\
\end{align*}
\]

One can show that

\[
\begin{align*}
d(d \ln \det[\Delta \Lambda \Delta']) &= -2 \tr(\Delta^{-1} (d \Delta) \Delta^{-1} d \Delta), \\
d(d \tr(\Delta' H \Delta \Lambda)) &= 2 \tr(\Lambda (d \Delta)' H d \Delta) + 4 \tr(\Lambda (d \Delta)' H \Delta d \Xi) \\
&\quad + \tr(\Lambda (d \Xi) \Delta' H d \Xi). \\
\end{align*}
\]

Collecting terms and using vec(\( \Delta' \)) = \( E_K \) vec(\( \Delta \)) and \( E_K = E_K' \) as well as vec(\( BCD \)) = (\( D' \otimes B \)) vec(\( C \)) for \( K \times K \) matrices \( B, C, D \), one obtains

\[
\begin{align*}
d^2 \ln p_\theta(y, z) &= -(d \text{vec}(\Delta))^\prime \left[ nE_K((\Delta^{-1})' \otimes \Delta^{-1}) + (\Lambda \otimes H) \right] \text{vec}(\Delta) \\
&\quad -2(d \text{vec}(\Delta))^\prime (\Lambda \otimes H \Delta) \text{vec}(\Xi) \\
&\quad -\frac{1}{2} (d \text{vec}(\Xi))^\prime (\Lambda \otimes \Delta' H \Delta) \text{vec}(\Xi). \\
\end{align*}
\]

Since vec(\( \Delta \)) and vec(\( \Xi \)) are affine functions of v(\( \Delta \)) and v(\( \Xi \)), respectively, Theorem 11 of MN (p. 112) can be invoked, which completes the proof. □

**Lemma 4.** The second differential of \( \ln p_\theta(y, z) \) with respect to \( v(\Sigma) \) is given by

\[
\begin{align*}
d^2 \ln p_\theta(y, z) &= -(d \text{vec}(\Sigma))^\prime D_\Sigma' \left[ (\Sigma^{-1} \otimes \Sigma^{-1} H \Sigma^{-1}) - \frac{n}{2} (\Sigma^{-1} \otimes \Sigma^{-1}) \right] D_\Sigma \text{vec}(\Sigma). \\
\end{align*}
\]

**Proof of Lemma 4.**

\[
\begin{align*}
d^2 \ln p_\theta(y, z) &= \frac{n}{2} \tr(\Sigma^{-1} (d \Sigma) \Sigma^{-1} d \Sigma) - \tr(\Sigma^{-1} (d \Sigma) \Sigma^{-1} H \Sigma^{-1} d \Sigma) \\
&= -(d \text{vec}(\Sigma))^\prime \left[ (\Sigma^{-1} \otimes \Sigma^{-1} H \Sigma^{-1}) - \frac{n}{2} (\Sigma^{-1} \otimes \Sigma^{-1}) \right] \text{vec}(\Sigma). \\
\end{align*}
\]
Since \( \text{vec}(\Sigma) \) is a linear function of \( v(\Sigma) \), Theorem 11 of MN (p. 112) applies, completing the proof. □

The gradients of \( \ln p_\theta(y, z) \) with respect to \( v(\Delta) \), \( v(\Xi) \), and \( v(\Sigma) \) follow from the first identification theorem of MN (Theorem 6, p. 87) and Lemmas 1 and 2, and are given by (4.13), (4.14), and (4.16), respectively. The Hessian matrices of \( \ln p_\theta(y, z) \) with respect to \( v(\Delta) \), \( v(\Xi) \), and \( v(\Sigma) \) follow from the second identification theorem of MN (Theorem 6, p. 107) and Lemmas 3 and 4, and are given by (4.15) and (4.17), respectively.
Chapter 5

Bayesian modeling and estimation†

Longitudinal data on social networks and other outcome variables are of special interest to social scientists. Models are considered which assume that the data are outcomes of continuous-time Markov processes, observed at two or more discrete time points. The importance of Bayesian inference is stressed, and Bayesian methods are proposed and demonstrated by applications to empirical data.

Keywords: panel data, continuous-time Markov processes, latent variables, hierarchical Bayesian models, Markov chain Monte Carlo.

5.1 Introduction

Social network analysis studies links between entities (see Wasserman and Faust, 1994); here, directed links (arcs) between entities (nodes) are considered. Examples are friendships among individuals (e.g., adolescents, managers, criminals), transactions between companies, and transmission of information among institutions (e.g., police departments). The arcs between nodes may have an impact on outcome variables of interest: e.g., friendships among adolescents may influence the initiation of smoking, drinking alcohol, or taking drugs. The outcome variables in turn may have an impact on (in the simplest case: the absence or presence; otherwise: the strength of) arcs between nodes: e.g., adolescents may tend to select friends who are similar in terms of alcohol consumption. Thus, the network may affect outcome variables of interest, and the outcome variables in turn may affect the network.

The dependencies between networks and other outcome variables make longitu-

†Working paper.
dinal data preferable to non-longitudinal data, because studying longitudinal observations of the processes that generate such dependencies is thought to increase the scientific understanding of such dependencies. In practice, the most common form of longitudinal data are panel data, i.e., the network and other outcome variables are observed at two or more discrete time points.

Initial work on modeling such panel data focused on networks while treating other variables as exogenous. Snijders (2001) postulated that the network evolution is governed by a continuous-time Markov process. To estimate such continuous-time Markov models from network panel data, Snijders (2001) proposed the method of moments (MM); more recently, Koskinen (2004) proposed Bayesian methods and Snijders, Koskinen, and Schweinberger (2006) the maximum likelihood (ML) method. Motivated by the fact that in the social sciences it is not unusual that some relevant covariates are unobserved due to data collection constraints and limited prior knowledge of researchers, Schweinberger and Snijders (2006) considered continuous-time Markov models which take unobserved heterogeneity across nodes into account by means of latent variables (or random effects), and proposed ML and Bayesian estimation methods. Snijders, Steglich, and Schweinberger (2006) pioneered models where continuous-time Markov processes govern the change of both networks and node-dependent outcome variables, and proposed MM estimation.

The present paper considers continuous-time Markov models which are more general than Schweinberger and Snijders (2006) in that change of both networks and node-dependent outcome variables is modeled, and more general than Snijders, Steglich, and Schweinberger (2006) in that unobserved heterogeneity across nodes is taken into account.

To estimate such models, Bayesian methods are proposed. The importance of Bayesian inference is stressed, because Bayesian inference (1) is well-suited to learning from unique, non-repeatable processes—such as the social processes considered here—and (2) has theoretical and practical advantages compared to MM and ML estimation.

The paper is structured as follows. Section 5.2 introduces the data structure, while Section 5.3 presents models. Section 5.4 motivates Bayesian inference, and proposes Bayesian methods. Section 5.5 applies the Bayesian methods to two small, empirical data sets, and compares MM, ML, and Bayesian estimation.
5.2 Data structure

Let $N = \{1, \ldots, n\}$ be a finite set of nodes. Consider a binary, directed relation (or network) on $N$ and a node-dependent outcome variable, observed at discrete, ordered time points $t_0 < t_1 < \cdots < t_H$. The observations of the network are represented as binary $n \times n$ matrices $X(t_0), X(t_1), \ldots, X(t_H)$, where $x_{ij}(t_h) = 1$ if there is an arc from node $i$ to node $j$ at time point $t_h$, and $x_{ij}(t_h) = 0$ otherwise; the diagonal elements are disregarded and defined as $x_{ii}(t_h) \equiv 0$. The observations of the node-dependent outcome variable are represented as $n \times 1$ vectors $y(t_0), y(t_1), \ldots, y(t_H)$, where $y_i(t_h)$ is binary or integer-valued.

The described data structure is common and convenient, but extensions to more general data structures are possible. First, $x_{ij}(t_h)$ may not be binary, but integer-valued. Second, extensions to more than one node-dependent outcome variable are possible. Note, in addition, that the assumption that $x_{ij}(t_h)$ and $y_i(t_h)$ are binary or integer-valued is, in practice, not too restrictive: in the social sciences, such outcome variables are common.

5.3 Model

The data $(X(t_0), y(t_0)), (X(t_1), y(t_1)), \ldots, (X(t_H), y(t_H))$ are assumed to be generated by a continuous-time Markov process (see, e.g., Resnick, 2002). Let $H = 1$ and note that the extension to the case $H \geq 2$ is straightforward due to the Markov property.

Starting at time $t \equiv t_0$ with $(X, y) \equiv (X(t_0), y(t_0))$, the Markov process corresponds to the following iteration steps: independent holding times $h_X$ and $h_Y$ are sampled from the negative exponential distributions $\text{Exp}(\lambda_X)$ and $\text{Exp}(\lambda_Y)$, respectively, and if $h_X < h_Y$, then $t$ is set to $t = t + h_X$ and $X$ is allowed to change, otherwise $t$ is set to $t = t + h_Y$ and $y$ is allowed to change.

Section 5.3.1 specifies the parameters $\lambda_X$ and $\lambda_Y$ of the negative exponential distributions, and Section 5.3.2 introduces the conditional probability law governing changes of $X$ and $y$. 
5.3.1 Timing of changes: parameters $\lambda_X$ and $\lambda_Y$

The parameters $\lambda_X$ and $\lambda_Y$ of the negative exponential distributions are given by

$$\lambda_X \equiv \lambda_X(X, y, \theta) = \sum_{i=1}^{n} \lambda_{X,i}(X, y, \theta),$$

$$\lambda_Y \equiv \lambda_Y(X, y, \theta) = \sum_{i=1}^{n} \lambda_{Y,i}(X, y, \theta),$$

where $\lambda_{X,i}$ and $\lambda_{Y,i}$ are called the “network rate function” and the “behavior rate function” of node $i$, respectively, interpreted as the rates at which $i$ is allowed to change one of the arc variables $X_{ih} (h \neq i)$ and the behavior $Y_i$, respectively, while $\theta$ is a parameter vector. Convenient rate functions $\lambda_{X,i}$ and $\lambda_{Y,i}$ are given by

$$\lambda_{X,i}(X, y, \theta) = \alpha_X \exp[\epsilon_X e_{X,i}(X, y)],$$

$$\lambda_{Y,i}(X, y, \theta) = \alpha_Y \exp[\epsilon_Y e_{Y,i}(X, y)],$$

where $\alpha_X, \alpha_Y > 0$ are parameters and $\epsilon_X, \epsilon_Y$ are parameter vectors, while the vector functions $e_{X,i}, e_{Y,i}$ depend on $X, y$, and covariates.

5.3.2 Changes: conditional probability law

Motivated by Holland and Leinhardt (1977), it is postulated that changes of $X$ and $y$ proceed in “mini-steps” in the sense that at most one coordinate of $X = (X_{ij})$ or $y = (y_i)$ can change at a time.

If $X$ is allowed to change, a node (called “actor”) $i$ is sampled with probability $\lambda_{X,i}/\lambda_X$. In the actor-driven approach of Schweinberger and Snijders (2006), $i$ is assumed to select some node $j \in N$, and if $i = j$, then $i$ is assumed to change nothing, otherwise $i$ transforms $x_{ij}$ into $x_{ij}^* = 1 - x_{ij}$, while $x_{kl}$ remains unchanged for all $(k, l) \neq (i, j)$. It is assumed that $i$ selects the $j \in N$ which maximizes

$$f_{X,i}(j \mid X, y, \theta) + U_{X,ij}(t), \quad (5.1)$$

where $f_{X,i}$ is called the “network objective function” of $i$ and $U_{X,ij}(t)$ is a random variable.

The network objective function $f_{X,i}$ is given by

$$f_{X,i}(j \mid X, y, \theta) = \eta_{X,i}^j s_{X,i}(j, X, y), \quad (5.2)$$

where $s_{X,i}$ is a vector function of $j$, $X$, $y$, and covariates, and $\eta_{X,i}$ is a weight vector which depends on node $i$,

$$\eta_{X,i} = \beta_X + A_X V_{X,i},$$
where $\beta_X$ is a parameter vector, $A_X$ is a design matrix, and $V_{X,i} \sim N(0, \Sigma_X)$ is a Gaussian distributed random vector. Models with node-dependent weights $\eta_{X,i}$ allow for unobserved heterogeneity across nodes, a frequent phenomenon in social science data, and therefore are viable alternatives to models with constant weights (see Schweinberger and Snijders, 2006).

If $y$ is allowed to change, a node $i$ is sampled with probability $\lambda_{Y,i}/\lambda_Y$. Let $J = \{-1, 0, 1\}$ be the set of possible increments of $y_i$ and $M \subseteq J$ be the set of increments which are admissible given the current value $y_i$ of variable $Y_i$ (if $\{a, a+1, \ldots, b\}$, for two integers $a, b$ such that $a < b$, is the outcome space of variable $Y_i$, then $M = J$ unless $y_i = a$ or $y_i = b$). Node $i$ is assumed to select some increment $j \in M$ and to transform $y_i$ into $y^*_i = y_i + j$, while $y_k$ remains unchanged for all $k \neq i$. It is assumed that $i$ selects the $j \in M$ which maximizes

$$f_{Y,i}(j \mid X, y, \theta) + U_{Y,ij}(t),$$

where $f_{Y,i}$ is called the “behavior objective function” of $i$ and $U_{Y,ij}(t)$ is a random variable.

The behavior objective function $f_{Y,i}$ is given by

$$f_{Y,i}(j \mid X, y, \theta) = \eta_{Y,i}' s_{Y,i}(j, X, y),$$

where $s_{Y,i}$ is a vector function of $j$, $X$, $y$, and covariates, and $\eta_{Y,i}$ is a weight vector which depends on node $i$,

$$\eta_{Y,i} = \beta_Y + A_Y V_{Y,i},$$

where $\beta_Y$ is a parameter vector, $A_Y$ is a design matrix, and $V_{Y,i} \sim N(0, \Sigma_Y)$ is a Gaussian distributed random vector.

If $U_{X,ij}(t)$ and $U_{Y,ij}(t)$ are assumed to be Gumbel(0, 1) distributed, then the conditional probability that $i$ selects $j$ can be written in closed form (see, e.g., McFadden, 1974) and is given by, for changes of $X$ and $y$, respectively,

$$\psi_{X,i}(j \mid X, y, \theta) = \frac{\exp [f_{X,i}(j \mid X, y, \theta)]}{\sum_{h \in N} \exp [f_{X,i}(h \mid X, y, \theta)]},$$

$$\psi_{Y,i}(j \mid X, y, \theta) = \frac{\exp [f_{Y,i}(j \mid X, y, \theta)]}{\sum_{h \in M} \exp [f_{Y,i}(h \mid X, y, \theta)].}$$

Note that the model is more general than Schweinberger and Snijders (2006) because change of both $X$ and $y$ is modeled, and more general than Snijders, Steglich,
and Schweinberger (2006) because \( \eta_{X,i} \) and \( \eta_{Y,i} \) are node-dependent random variables, taking unobserved heterogeneity across nodes into account.

## 5.4 Bayesian estimation

Section 5.4.1 introduces notation. Section 5.4.2 stresses the importance of Bayesian inference, and Section 5.4.3 proposes Bayesian methods.

### 5.4.1 Notation

A complete observation of the continuous-time Markov process in time interval \([t_0, t_1]\) corresponds to data \((X_0, y_0), (X_1, y_1), \ldots, (X_{M-1}, y_{M-1}), (X_M, y_M)\), and holding times, where \( M = M_X + M_Y \), \( M_X \) denotes the total number of times \( X \) was allowed to change, and \( M_Y \) denotes the total number of times \( y \) was allowed to change. The data \((X_0, y_0) \equiv (X(t_0), y(t_0))\) and \((X_M, y_M) \equiv (X(t_1), y(t_1))\), including covariates, are observed and referred to as \( D \). The data \((X_1, y_1), \ldots, (X_{M-1}, y_{M-1})\) are unobserved and can be represented by \((X_0, y_0)\) and the sequence \( W = (a_m, i_m, j_m)_{m=1}^M \), where \( a_m \) indicates whether \( X \) was allowed to change \((a_m = 0)\) or \( y \) was allowed to change \((a_m = 1)\), \( i_m \) is the node that was allowed to change something, and \( j_m \) is the selected alternative (i.e., if \( a_m = 0 \), then \( j_m \) is the selected node, otherwise \( j_m \) is the selected increment). The holding times are unobserved, but can be disregarded by using the embedded Markov process, corresponding to \((X_0, y_0)\) and \( W \). The data \( V_{x,i} \) and \( V_{y,i} \) are unobserved and collected in vector \( V_i \), and the vectors \( V_i \) are stored as rows of matrix \( V \). The unobserved data \( V \) and \( W \) are referred to as \( Z \).

For convenience, it is assumed that the rate functions are constant, i.e., \( \lambda_{X,i} \equiv \alpha_X \) and \( \lambda_{Y,i} \equiv \alpha_Y \); the extension to non-constant rate functions can be obtained along the lines of Snijders, Koskinen, and Schweinberger (2006). The parameter vectors \( \beta_X \) and \( \beta_Y \) are collected in vector \( \beta \). The variance-covariance matrix of \( V_i \) is denoted as \( \Sigma \). The parameters \( \alpha_X, \alpha_Y, \beta, \) and \( \Sigma^{-1} \) are referred to as \( \theta \).

### 5.4.2 Motivation

To estimate continuous-time Markov models from longitudinal data on networks and node-dependent outcome variables, Bayesian inference is a sensible alternative and complement to MM and ML estimation, because Bayesian inference is well-suited to learning from unique, non-repeatable processes, and has theoretical and practical
advantages compared to MM and ML estimation.

**Unique, non-repeatable processes**

In most social science applications, it is, in the first place, desired to assess whether parameters of interest are zero, positive, or negative. Since networks and node-dependent outcome variables change and thus the initial state of the data-generating process changes in the short run, and, indeed, the set of nodes, covariates, and parameters of the data-generating process are suspected to be non-constant in the long run, the data-generating process is unique and non-repeatable. As a result, it is preferable to approach the mentioned set estimation problems without relying on the assumption that the data-generating process can be repeated an infinite number of times, which is possible in a Bayesian framework—using, e.g., \((1 - \alpha)\) posterior intervals—but impossible in a MM or ML framework, since for a given, finite sample, exact \((1 - \alpha)\) confidence intervals do not exist, and constructing conventional \((1 - \alpha)\) confidence intervals makes the implicit assumption that the data-generating process can be repeated an infinite number of times.

**Theoretical and practical advantages**

Bayesian estimation shares with ML estimation some decisive advantages compared to MM estimation, but has additional, practical advantages compared to ML estimation.

**MM estimation**  Likelihood-based inference in the form of Bayesian and ML estimation may have asymptotic advantages compared to MM estimation, though—at the present time—the asymptotic properties of MM, ML, and Bayesian estimators are not fully understood. However, in addition to possible asymptotic advantages, likelihood-based inference has advantages compared to MM estimation for given, finite data.

*Choice of statistic \(U\).* MM estimation (see Snijders, Steglich, and Schweinberger, 2006) amounts to solving estimating equations of the form

\[
E_\theta[U \mid X(t_0), y(t_0)] - u = 0
\]

with respect to \(\theta\), where \(U\) is a function of \((X(t), y(t))\) at time point \(t = t_1\) which is sensitive to changes in \(\theta\), and \(u\) is the value of \(U\) at the observed data \((X(t_1), y(t_1))\) at time point \(t_1\). MM estimation requires choosing a statistic \(U\). However, in the
absence of sufficient statistics, the choice of $U$ is arbitrary: as a result, inefficient choices of $U$ cannot be ruled out.

**Asymmetric distribution of statistic $U$.** Let $\theta$ and $U$ be one-dimensional. If $\theta_0$ is the data-generating value of $\theta$ and the distribution of $U$ under (al)most (all) values of $\theta$, including $\theta_0$, is very left-skewed (right-skewed), then most values $u$ of $U$ generated under $\theta_0$ are larger (smaller) than $E_{\theta_0}[U \mid X(t_0), y(t_0)]$. If $E_{\theta}[U \mid X(t_0), y(t_0)]$ is an increasing function of $\theta$, then the value of $\theta$ which solves $E_{\theta}[U \mid X(t_0), y(t_0)] = u$ is larger (smaller) than $\theta_0$ for most values $u$ of $U$ generated under $\theta_0$, which implies that the MM estimator of $\theta_0$ may be considerably biased; the argument can be extended to the multi-dimensional case.

**Partial conditioning on the observed data $D$.** MM, ML, and Bayesian algorithms sample sequences $W$, corresponding to possible changes of $X$ and $y$. However, while Bayesian and ML estimation condition on the whole of the observed data $D$—on $(X(t_0), y(t_0))$ and $(X(t_1), y(t_1))$—MM estimation conditions on $(X(t_0), y(t_0))$, but not on $(X(t_1), y(t_1))$. The partial conditioning on $D$ has important implications for given, finite data: while Bayesian and ML algorithms sample $W$ subject to the constraint that $(X_0, y_0) \equiv (X(t_0), y(t_0))$ and $(X_M, y_M) \equiv (X(t_1), y(t_1))$ (see Section 5.4.3), MM algorithms sample $W$ subject to the constraint that $(X_0, y_0) \equiv (X(t_0), y(t_0))$, but $(X_M, y_M)$ is not required to be $(X(t_1), y(t_1))$ (see Snijders, Steglich, and Schweinberger, 2006). Since the estimating equation (5.6) is based on the conditional expectation of $U$, which is a function of $W$, the partial conditioning on $D$ implies that a large number of sequences $W$ for which $(X_M, y_M) \neq (X(t_1), y(t_1))$ contributes to the MM estimation of $\theta$, i.e., a large number of hypothetical data sets—which could have been observed but in fact were not—contributes to the MM estimation of $\theta$, which may vitiate conclusions.

**Large, observed numbers of changes of $X$ and $y$.** If the observed number of changes of $X$, $\sum_{i=1}^{n} \sum_{j=1}^{n} |x_{ij}(t_1) - x_{ij}(t_0)|$, and $y$, $\sum_{i=1}^{n} |y_i(t_1) - y_i(t_0)|$, is large, then MM algorithms may not converge, which is discussed in detail by Snijders (2001).

**ML estimation** Bayesian inference has at least two practical advantages compared to ML estimation.
Approximations to the likelihood function. ML estimators are point estimators which maximize the likelihood function, while Bayesian inference gives rise to posterior distributions, which—under vague prior distributions—approximate the entire likelihood function. Therefore, Bayesian estimation can complement ML estimation even when the ML method is the primary method of choice, e.g., by providing insight into what parameter values (in addition to the ML estimates) are plausible in the light of the observed data, and—when the likelihood function is suspected to be asymmetric, multimodal, or having an otherwise non-standard shape—by providing insight into the shape of the likelihood function.

Prior knowledge. In small data sets, it is sometimes suspected that the maxima of the likelihood function are on the boundary of the parameter space; even when the maxima of the likelihood function are not on the boundary of the parameter space, ML estimation may be hard, because the surface of the likelihood function may be flat, multimodal, or otherwise inconvenient. In addition, when—as usual—the objective function $f_{Y,i}$ contains terms of the form $\beta \left( y_i^* - y_i \right)$ ($\beta$ being a parameter), then empirical results suggest that in data sets where, for large numbers of nodes, $y_i(t_0) = a$ or $y_i(t_1) = a$ ($a$ being the lower bound of the outcome space of variable $Y_i$) and $|y_i(t_1) - y_i(t_0)|$ is small, then large, positive values of $\alpha_Y$ and large, negative values of $\beta$ compensate each other; one possible explanation is as follows: an increase in $\alpha_Y$ is associated with an increase in the number of times $y$ is allowed to change, but when $y$ is allowed to change at time $t \in [t_0, t_1]$—and the number of nodes with $y_i(t_0) = a$ or $y_i(t_1) = a$ and small $|y_i(t_1) - y_i(t_0)|$ is large—then the number of nodes with $y_i(t) = a$ is expected to be large, and for such nodes $i$, the probability of selecting increment $j = 0$ (cf. (5.5)) approaches 1 as $\beta \to -\infty$ (holding constant all other parameters). A Bayesian approach may help in such situations by incorporating prior knowledge with respect to critical parameters and pushing the marginal posterior distributions of critical parameters to regions which, according to prior knowledge (e.g., experience from other studies with similar design), seem to be plausible. The most important candidate for informative prior distributions is $\alpha_Y$, because in most applications where convergence problems have been encountered $\alpha_Y$ seems to be involved, and because $\alpha_Y$ is regarded as a nuisance parameter; thus, informative prior distributions of $\alpha_Y$ do not hurt too much, but can help to estimate the parameters of primary interest; in addition, it is expected that in most applications the marginal posterior distributions of the parameters of primary interest are not too sensitive to the prior distribution of $\alpha_Y$, which, in practice, can be examined by sensitivity analyses.
5.4.3 Bayesian methods

Bayesian inference is based on the posterior probability density

\[ p(\theta \mid D) \propto p(D \mid \theta) p(\theta) = \left[ \int p(Z, D \mid \theta) \, dZ \right] p(\theta), \]

where \( p(D \mid \theta) \) is the probability density of \( D \)—which cannot be written in closed form, \( p(Z, D \mid \theta) \) is the joint probability density of \( Z \) and \( D \)—which can be written in closed form, and \( p(\theta) \) is the prior density of \( \theta \).

To derive \( p(Z, D \mid \theta) \) (cf. Section 5.3), note that (1) if \( h_X \) and \( h_Y \) are the holding times sampled from the negative exponential distributions \( \text{Exp}(\lambda_X) \) and \( \text{Exp}(\lambda_Y) \), respectively, then the distribution of \( h_m = \min(h_X, h_Y) \) is \( \text{Exp}(\lambda_X + \lambda_Y) \); (2) given that either \( X \) or \( y \) is allowed to change, the probability that \( X \) is allowed to change is given by \( \lambda_X / (\lambda_X + \lambda_Y) \), while the probability that \( y \) is allowed to change is given by \( \lambda_Y / (\lambda_X + \lambda_Y) \); and (3) given that \( X \) respectively \( y \) is allowed to change, the probability that node \( i \) is allowed to change something is given by \( \lambda_{X,i} / \lambda_X \) respectively \( \lambda_{Y,i} / \lambda_Y \).

The joint probability density of sampling \( h_m \), and allowing \( X \) respectively \( y \) to change and \( i \) to change something, is therefore given by

\[
(\lambda_X + \lambda_Y) \exp[-(\lambda_X + \lambda_Y)h_m] \left[(1 - a_m) \frac{\lambda_X}{\lambda_X + \lambda_Y} \frac{\lambda_{X,i}}{\lambda_X} + a_m \frac{\lambda_Y}{\lambda_X + \lambda_Y} \frac{\lambda_{Y,i}}{\lambda_Y}\right] = \exp[-(\lambda_X + \lambda_Y)h_m] \left[(1 - a_m) \lambda_{X,i} + a_m \lambda_{Y,i}\right].
\]

Under constant rate functions, the joint probability density of sampling \( h_m \) (\( m = 1, \ldots, M + 1 \)) such that \( t_0 + \sum_{m=1}^{M} h_m \leq t_1 < t_0 + \sum_{m=1}^{M+1} h_m \), and allowing \( X_{m-1} \) respectively \( y_{m-1} \) to change and \( i_m \) to change something (\( m = 1, \ldots, M \)), is therefore given by

\[
\left\{ \prod_{m=1}^{M} \exp[-(\lambda_X + \lambda_Y)h_m] \left[(1 - a_m) \lambda_{X,i} + a_m \lambda_{Y,i}\right] \right\} \exp[-(\lambda_X + \lambda_Y)r] = \alpha_X^{M_X} \exp[-n\alpha_X(t_1 - t_0)] \alpha_Y^{M_Y} \exp[-n\alpha_Y(t_1 - t_0)],
\]

where \( r = (t_1 - t_0) - \sum_{m=1}^{M} h_m \). The joint probability density \( p(Z, D \mid \theta) \) is thus given
by

\[
p(Z, D \mid \theta) = \alpha_X^M \exp[-n\alpha_X(t_1 - t_0)] \alpha_Y^M \exp[-n\alpha_Y(t_1 - t_0)] \\
\times \prod_{m=1}^{M} \xi_i(j_m \mid a_m, X_{m-1}, y_{m-1}, \theta) \\
\times \prod_{i=1}^{n} (2\pi)^{-K/2} \det[\Sigma^{-1}]^{1/2} \exp\left[-\frac{1}{2} V_i' \Sigma^{-1} V_i\right],
\]

where

\[
\xi_i(j_m \mid a_m, X_{m-1}, y_{m-1}, \theta) = (1 - a_m) \psi_{X,i}(j_m \mid X_{m-1}, y_{m-1}, \theta) \\
+ a_m \psi_{Y,i}(j_m \mid X_{m-1}, y_{m-1}, \theta).
\]

To specify \( p(\theta) \), it is, in most applications, reasonable to start with the assumption of prior independence of \( \alpha_X, \alpha_Y, \beta, \) and \( \Sigma^{-1} \), while substantive knowledge may suggest prior dependence among components of \( \beta_X \) and \( \beta_Y \); convenient families of prior distributions are

\[
\begin{align*}
\alpha_X &\sim \text{Gamma}(\gamma_X, \delta_X), \\
\alpha_Y &\sim \text{Gamma}(\gamma_Y, \delta_Y), \\
\beta &\sim N(0, \Psi), \\
\Sigma^{-1} &\sim \text{Wishart}(\varphi, \Omega),
\end{align*}
\]

where \( \Psi \) allows to impose prior dependence among components of \( \beta_X \) and \( \beta_Y \).

A Markov chain Monte Carlo (MCMC) algorithm is proposed to sample from the posterior distribution, which can handle (1) networks and node-dependent outcome variables—while Koskinen (2004) and Schweinberger and Snijders (2006) are limited to networks—and (2) node-dependent weights \( \eta_{X,i} \) and \( \eta_{Y,i} \)—while Koskinen (2004) is limited to constant weights; (3) is based on the embedded Markov process, excluding the inconvenient holding times and thus avoiding needless complications encountered by Koskinen (2004); and (4) allows to sample \( \beta \) from the posterior distribution more efficiently (in terms of MCMC autocorrelations) than the (random walk Metropolis-Hastings) methods of Koskinen (2004). The algorithm combines the following Gibbs and Metropolis-Hastings (M-H) steps by means of cycling or mixing (cf. Hastings, 1970, Tierney, 1994):

\[
\text{Gibbs sampling of } \alpha_X \mid \alpha_Y, \beta, \Sigma^{-1}, V, W, D. \text{ If the prior of } \alpha_X \text{ is } \text{Gamma}(\gamma_X, \delta_X) \text{ parametrized such that } E_{\gamma_X, \delta_X}[\alpha_X] = \gamma_X/\delta_X, \text{ then the full conditional posterior of}
\]

\( \alpha_X \) is Gamma\((\gamma_X + M_X, \delta_X + n(t_1 - t_0))\).

Gibbs sampling of \( \alpha_Y | \alpha_X, \beta, \Sigma^{-1}, V, W, D \). If the prior of \( \alpha_Y \) is Gamma\((\gamma_Y, \delta_Y)\) parametrized such that \( E_{\gamma_Y, \delta_Y}[\alpha_Y] = \gamma_Y/\delta_Y \), then the full conditional posterior of \( \alpha_Y \) is Gamma\((\gamma_Y + M_Y, \delta_Y + n(t_1 - t_0))\).

M-H step to sample \( \beta | \alpha_X, \alpha_Y, \Sigma^{-1}, V, W, D \). A candidate \( \beta^* \) is generated from a distribution \( q(\beta^* | \beta) \) and accepted with probability

\[
\pi = \min \left[ 1, \frac{p(\beta^* | \alpha_X, \alpha_Y, \Sigma^{-1}, V, W, D)}{p(\beta | \alpha_X, \alpha_Y, \Sigma^{-1}, V, W, D)} \times \frac{q(\beta^* | \beta)}{q(\beta | \beta^*)} \right].
\]

Convenient candidates are given by \( \beta^* = a + B(\beta - a) + C \), where \( a \) is a constant vector, \( B \) is a constant matrix, and \( C \sim N(0, \Gamma) \). Let \( I \) be the identity matrix of suitable order. Possible choices of \( B \) are (i) \( B = I \), corresponding to random walk M-H algorithms; (ii) \( B = 0 \), corresponding to independence samplers; and (iii) \( B = -I \), corresponding to first-order autoregressive M-H algorithms, which reflect the present parameter vector \( \beta \) about the vector \( a \) before adding increment vector \( C \) and can reduce MCMC autocorrelations. In case of (ii) and (iii), \( a \) can be learned from burn-in iterations or point estimates of \( \beta \); under the simplifying assumption that \( \Sigma \equiv 0 \), MM estimates (see Snijders, Steglich, and Schweinberger, 2006), ML estimates (see Snijders, 2006), or crude Bayesian point estimates based on an explorative run with (i) can be exploited as point estimates of \( \beta \). The scale matrix \( \Gamma \) can be calibrated during burn-in iterations; as an alternative, the inverse information matrix of \( \beta \) at the ML estimate of \( \beta \) (see Snijders, 2006) can be used as scale matrix. In applications, M-H algorithms of class (ii) and (iii) tend to outperform M-H algorithms of class (i) in terms of MCMC autocorrelations.

Gibbs sampling of \( \Sigma^{-1} | \alpha_X, \alpha_Y, \beta, V, W, D \). If the prior of \( \Sigma^{-1} \) is Wishart\((\varphi, \Omega)\) parametrized such that \( E_{\varphi, \Omega}[\Sigma^{-1}] = \varphi \Omega \), then the full conditional posterior is Wishart\((\varphi + n, (H + \Omega^{-1})^{-1})\), where \( H = \sum_{i=1}^{n} V_i V_i' \).

M-H step to sample \( V | \alpha_X, \alpha_Y, \beta, \Sigma^{-1}, W, D \). For \( i = 1, \ldots, n \) independently, a candidate \( V_i^* \) is generated from a distribution \( q(V_i^* | V_i) \) and accepted with probability

\[
\pi = \min \left[ 1, \frac{p(V_i^* | \alpha_X, \alpha_Y, \beta, \Sigma^{-1}, V_{-i}, W, D)}{p(V_i | \alpha_X, \alpha_Y, \beta, \Sigma^{-1}, V_{-i}, W, D)} \times \frac{q(V_i^* | V_i)}{q(V_i^* | V_i)} \right],
\]

where \( V_{-i} \) corresponds to matrix \( V \) without row \( i \). Convenient candidates are given by \( V_i^* = c + B(V_i - c) + D_i \), where \( c \) is a constant vector, \( B \) is a constant matrix, and
5.5. APPLICATIONS

$D_i \sim N(0, \Delta)$. Possible choices of $B$ are (i) $B = I$ (random walk M-H algorithms); (ii) $B = 0$ and $a = 0$ (independence samplers); and (iii) $B = -I$ (first-order autoregressive M-H algorithms) (cf. Schweinberger and Snijders, 2006).

M-H step to sample $W | \alpha_X, \alpha_Y, \beta, \Sigma^{-1}, V, D$. A candidate $W^*$ is generated from a distribution $q(W^* | W)$ and accepted with probability

$$\pi = \min \left[ 1, \frac{p(W^* | \alpha_X, \alpha_Y, \beta, \Sigma^{-1}, V, D)}{p(W | \alpha_X, \alpha_Y, \beta, \Sigma^{-1}, V, D)} \times \frac{q(W | W^*)}{q(W^* | W)} \right].$$

The distribution $q(W^* | W)$ is identical to the proposal distribution of Snijders (2006), which is a probability distribution defined on a discrete set of simple proposals, corresponding to inserting and deleting $(a_m, i_m, j_m)$’s in the sequence $W = (a_m, i_m, j_m)_{m=1}^M$ or permuting sub-sequences of $W$; note that conditioning on the observed data $D$ implies that sampling $W$ is subject to the constraint that $(X_0, y_0) \equiv (X(t_0), y(t_0))$ and $(X_M, y_M) \equiv (X(t_1), y(t_1))$.

5.5 Applications

The model is applied to two small, empirical data sets, because the advantages of Bayesian inference compared to MM and ML estimation may become more evident in small data sets which do not contain much information about parameters of primary interest. The two data sets were selected from the friendship and delinquent behavior study of Knecht, Steglich, Baerveldt, and Snijders (2006), who collected data in 126 Dutch school classes at four time points $t_0 < t_1 < t_2 < t_3$ in 2003 and 2004. Knecht, Steglich, Baerveldt, and Snijders (2006) studied these data by using MM and ML estimation, but sometimes faced considerable estimation problems (e.g., convergence problems). Here, two of the 126 school classes are studied in more detail than in Knecht, Steglich, Baerveldt, and Snijders (2006). In the first of the two selected data sets (Section 5.5.1), MM and ML estimation turn out to be problematic but Bayesian estimation is feasible, while in the second of the two selected data sets (Section 5.5.2), it is not an issue to obtain MM, ML, and Bayesian estimates, but—while likelihood-based inference in the form of Bayesian and ML estimation gives rise to results that agree by and large—MM estimation produces results which deviate from likelihood-based inference.
5.5.1 Application I

School class 1-E of Knecht, Steglich, Baerveldt, and Snijders (2006) with \( n = 21 \) students is studied. The data consist of \( 21 \times 21 \) matrices \( X(t_h) \) \((h = 0, 1, 2, 3)\), where \( x_{ij}(t_h) = 1 \) if student \( i \) considered student \( j \) to be a friend at time point \( t_h \), and \( x_{ij}(t_h) = 0 \) otherwise; and of \( 21 \times 1 \) vectors \( y(t_h) \) \((h = 0, 1, 2, 3)\), where \( y_i(t_h) \) refers to the degree of delinquent behavior of student \( i \) at time point \( t_h \)—measured on a scale with values 1 to 5 and including aspects such as stealing, vandalism, graffiti, and fighting (see Knecht, Steglich, Baerveldt, and Snijders, 2006). Concerning the friendships, at time point \( t_0 \), 98 arcs were present; in time interval \([t_{h-1}, t_h]\) \((h = 1, 2, 3)\), 48, 39, and 46 arcs were deleted while 44, 44, and 39 arcs were added, respectively. Concerning the delinquent behavior, at time point \( t_0 \), 15 students had value 1, and the .025 quantile was 1.0 while the .975 quantile was 4.5; in time interval \([t_{h-1}, t_h]\) \((h = 1, 2, 3)\), 2, 1, and 2 students decreased the delinquent behavior while 7, 5, and 5 students increased the delinquent behavior, respectively.

Motivated by Knecht, Steglich, Baerveldt, and Snijders (2006), the model is specified by rate functions \( \lambda_{X,i} \equiv \alpha_{X,h} \) and \( \lambda_{Y,i} \equiv \alpha_{Y,h} \) in time interval \([t_{h-1}, t_h]\) \((h = 1, 2, 3)\), and objective functions with constant weights \( \eta_{X,i} \equiv \beta_X \) and \( \eta_{Y,i} \equiv \beta_Y \),

\[
f_{X,i}(j \mid X, y, \theta) = \gamma_{X,ij} (x_{ij}^* - x_{ij}),
\]

where

\[
\gamma_{X,ij} = \beta_{X,1} + \beta_{X,2} x_{ji} + \beta_{X,3} \left( \sum_{h=1}^{n} x_{ih} x_{hj} + \sum_{l=1}^{n} x_{il} x_{jl} \right) + \beta_{X,4} y_j + \beta_{X,5} y_i + \beta_{X,6} y_j y_i,
\]

and

\[
f_{Y,i}(j \mid X, y, \theta) = \gamma_{Y,i} (y_i^* - y_i),
\]

where

\[
\gamma_{Y,i} = \beta_{Y,1} + \beta_{Y,2} g_i(X, y),
\]

while

\[
g_i(X, y) = \begin{cases} 
\frac{\sum_{l=1}^{n} x_{il} y_l}{\sum_{l=1}^{n} x_{il}} & \text{if } \sum_{l=1}^{n} x_{il} > 0, \\
0 & \text{otherwise}. 
\end{cases}
\]

Thus, when student \( i \) is allowed to change friendships, then the value attached by the objective function \( f_{X,i} \) to selecting student \( j \)—implying that \( x_{ij}^* - x_{ij} \) is either \(-1\)
(if the friendship to \( j \) is deleted), 0 (if \( j = i \) and thus nothing changes), or 1 (if the friendship to \( j \) is added)—is \( \gamma_{X,ij} \), depending on baseline parameter \( \beta_{X,1} \), reciprocity parameter \( \beta_{X,2} \), transitivity parameter \( \beta_{X,3} \), and parameters \( \beta_{X,4}, \beta_{X,5}, \) and \( \beta_{X,6} \), measuring the impact of the delinquent behavior of \( j \), the delinquent behavior of \( i \), and the interaction of the delinquent behavior of \( j \) and \( i \) on the decision to select student \( j \), respectively; when student \( i \) is allowed to change the delinquent behavior, then the value attached by the objective function \( f_{Y,3} \) to selecting increment \( j \)—implying that \( y_i^* - y_i \) is either \(-1\) (if the delinquent behavior is decreased), 0 (if the delinquent behavior is not changed), or 1 (if the delinquent behavior is increased)—is \( \gamma_{Y,i} \), depending on baseline parameter \( \beta_{Y,1} \) and parameter \( \beta_{Y,2} \), measuring the impact of the average delinquent behavior of the friends of \( i \) on the decision to select increment \( j \).

In the Bayesian framework, prior knowledge can be incorporated in the model, which is sensible in particular with respect to the nuisance parameters \( \alpha_{Y,h} \) \((h = 1, 2, 3)\) (cf. Section 5.4.2). Experience from other studies with similar design—in terms of the length of time \( t_h - t_{h-1} \) between observation points \( t_{h-1}, t_h \) \((h = 1, 2, 3)\), the type of nodes (students), and the type of node-dependent outcome variable (delinquent behavior)—suggest that Gamma(2, 2) priors of \( \alpha_{Y,h} \) \((h = 1, 2, 3)\) make sense, which have expectation 1, variance .5, and place 95% of the probability mass on the interval \([1, 2.8]\). Convenient, vague priors of \( \alpha_{X,h} \) \((h = 1, 2, 3)\) and \( \beta \) are given by Gamma(1, \(10^{-10}\)) and \( N(0, 10^{10}I_8) \), respectively, where \( I_8 \) is the \( 8 \times 8 \) identity matrix.

The parameters of primary interest are \( \beta_{X,k} \) \((k = 4, 5, 6)\) and \( \beta_{Y,2} \). It is sensible to proceed in two steps by first obtaining crude estimates of the nuisance parameters \( \alpha_{X,h} \) and \( \alpha_{Y,h} \) \((h = 1, 2, 3)\), \( \beta_{X,k} \) \((k = 1, 2, 3)\), and \( \beta_{Y,1} \) by estimating the model restricted by \( \beta_{X,k} = 0 \) \((k = 4, 5, 6)\) and \( \beta_{Y,2} = 0 \), and, second—given the estimates of the nuisance parameters under the restricted model—estimate the model without restrictions. MM estimates of the restricted model were computed first, and then ML estimates were computed by using the MM estimates as initial estimates. Concerning Bayesian estimation, the posterior distribution was explored by using the ML estimates as initial estimates and sampling \( \beta \) by random walk M-H algorithms, then exploiting the resulting MCMC output (in particular, estimates of the posterior expectation and variance-covariance matrix of \( \theta \)) for tailoring an independence sampler for sampling \( \beta \), and carrying out 110,000 iterations using the independence sampler; to diagnose possible non-convergence, trace plots and convergence checks of Raftery and Lewis (1996) were used; the first 10,000 iterations were discarded as burn-in iterations, and every 40th sampled value of the last 100,000 iterations was recorded. MM, ML, and Bayesian estimates of \( \theta \) under the restricted model are shown in Table 5.1, and the
Table 5.1: Estimates of $\theta$ under the restricted model

<table>
<thead>
<tr>
<th></th>
<th>MM (s.e.)</th>
<th>ML (s.e.)</th>
<th>posterior median</th>
<th>95% posterior interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{X,1}$</td>
<td>9.582 (1.777)</td>
<td>10.321 (1.801)</td>
<td>10.547</td>
<td>[7.903, 14.009]</td>
</tr>
<tr>
<td>$\alpha_{X,2}$</td>
<td>9.001 (1.649)</td>
<td>8.221 (1.335)</td>
<td>8.415</td>
<td>[6.214, 11.615]</td>
</tr>
<tr>
<td>$\alpha_{X,3}$</td>
<td>8.884 (1.304)</td>
<td>9.463 (1.599)</td>
<td>9.631</td>
<td>[7.074, 13.462]</td>
</tr>
<tr>
<td>$\beta_{X,1}$</td>
<td>-1.521 (.083)</td>
<td>-1.639 (.104)</td>
<td>-1.664</td>
<td>[-1.852, -1.484]</td>
</tr>
<tr>
<td>$\beta_{X,2}$</td>
<td>.761 (.154)</td>
<td>.721 (.134)</td>
<td>.719</td>
<td>[.487, .950]</td>
</tr>
<tr>
<td>$\beta_{X,3}$</td>
<td>.139 (.003)</td>
<td>.183 (.021)</td>
<td>.189</td>
<td>[.155, .223]</td>
</tr>
<tr>
<td>$\alpha_{Y,1}$</td>
<td>3.054 (2.207)</td>
<td>2.854 (.782)</td>
<td>2.877</td>
<td>[1.580, 4.970]</td>
</tr>
<tr>
<td>$\alpha_{Y,2}$</td>
<td>.763 (.511)</td>
<td>.712 (.284)</td>
<td>.828</td>
<td>[.380, 1.640]</td>
</tr>
<tr>
<td>$\alpha_{Y,3}$</td>
<td>1.385 (.819)</td>
<td>1.284 (.456)</td>
<td>1.244</td>
<td>[.620, 2.393]</td>
</tr>
<tr>
<td>$\beta_{Y,1}$</td>
<td>-.446 (.238)</td>
<td>-.360 (.173)</td>
<td>-1.456</td>
<td>[-2.036, -.974]</td>
</tr>
</tbody>
</table>

Marginal posterior densities of $\theta$ under the restricted model are depicted in Figure 5.1, together with the MM and ML estimates. Both MM and ML estimates seem to be close to the modes of the marginal posterior densities, but in almost all cases the ML estimate is closer to the mode than the MM estimate; two notable exceptions are (a) $\beta_{X,3}$—where the MM estimate is in the tail while the ML estimate is in the center of the marginal posterior density, and (b) $\beta_{Y,1}$—where both MM and ML estimate are in the tail of the marginal posterior density; concerning (a), in a number of applications the distribution of statistics of the form $\sum_{h=1}^{H} \sum_{i,j=1}^{n} X_{it}(t_h) X_{ij}(t_h) X_{ij}(t_h)$—which, as usual, was used to estimate $\beta_{X,3}$—has been found to be right-skewed, therefore the argument of Section 5.4.2 concerning asymmetric distributions of statistics can be applied to argue that the MM estimate may considerably underestimate $\beta_{X,3}$, and if the ML estimate and the marginal posterior mode are close to the data-generating value of $\beta_{X,3}$, then (a) does not come as a surprise; concerning (b), no fully satisfying explanation is available at the present time; it is interesting to note that the MM and ML estimates are close to the center of the marginal posterior densities of $\alpha_{Y,h}$ despite of the informative prior distributions of $\alpha_{Y,h}$ ($h = 1, 2, 3$).

The unrestricted model was estimated by MM, ML, and Bayesian algorithms, using as starting values the estimates of the nuisance parameters under the restricted model, and comparable specifications of the algorithms as above. Concerning the MM and ML algorithm, two repeated estimation runs (with starting values obtained from the previous estimation run) reveal that the estimates of $\beta_{Y,1}$ and $\beta_{Y,2}$ (not shown) depend on the starting values: the estimates are the larger the larger the...
Figure 5.1: Marginal posterior densities of $\theta$ under the restricted model

MM and ML estimates are indicated by solid lines; .025, .500, and .975 posterior quantiles are indicated by dotted lines; the marginal prior densities of $\alpha_{Y,h}$ ($h = 1, 2, 3$) are indicated by dashed curves.

starting values are, and have correlations close to $-1.0$; the correlations may be due to the fact that most students show the same level of delinquent behavior—as $y_i(t_h) = 1$ or 2 for 18, 18, 15, and 15 students at time point $t_h$ ($h = 0, 1, 2, 3$), respectively—and therefore $g_i(X, y)$ is “almost constant” across students. Concerning the Bayesian algorithm, trace plots and convergence checks of Raftery and Lewis (1996) indicate that the Markov chain—constructed by the Bayesian algorithm—fails to converge to the posterior distribution within reasonable time; other specifications of the Bayesian algorithm (within the framework sketched in Section 5.4.3) do not solve the problem; one possible explanation is that the posterior distribution is very long-tailed, which is supported by inspecting MCMC output and concerns, in the first place, $\alpha_{Y,h}$ ($h = 1, 2, 3$); note that long-tailed, marginal posterior distributions of $\alpha_{Y,h}$ ($h = 1, 2, 3$) are undesirable, because a large number of iterations is required to estimate the posterior distribution, and because each iteration is time-consuming, since sampling $W$ tends to consume the more time the larger $\alpha_{Y,h}$ ($h = 1, 2, 3$). Therefore, it appears to be sensible to modify the prior distributions of the nuisance parameters $\alpha_{Y,h}$ ($h = 1, 2, 3$) and push the marginal posterior distributions of $\alpha_{Y,h}$
(h = 1, 2, 3) towards the origin (cf. Section 5.4.2): first, Gamma(2, 4) prior distributions were tried, which did not solve the problem; however, under Gamma(2, 6) prior distributions the Bayesian algorithm seems to converge, as can be concluded from the mentioned convergence checks; the following results are based on Gamma(2, 6) prior distributions of α_{Y,h} (h = 1, 2, 3). The marginal posterior densities of θ under the unrestricted model are shown in Figure 5.2. Compared to the restricted model, the marginal posterior densities of α_{X,h} (h = 1, 2, 3) are almost unchanged, while the marginal posterior densities of α_{Y,h} (h = 1, 2, 3) shifted towards the origin, as a result of the modification of the prior distributions of α_{Y,h} (h = 1, 2, 3); the marginal posterior densities of β_{X,1} and β_{Y,1} were spread out by introducing β_{X,k} (k = 4, 5, 6) and β_{Y,2} (and modifying the prior), while the marginal posterior densities of β_{X,2} and β_{X,3} are almost unchanged; and, last, the 95% posterior intervals suggest that there is evidence that β_{X,5} is negative—meaning that delinquent students tend to select less friends than non-delinquent students—and that β_{X,6} is positive—meaning that delinquent students, when selecting friends, tend to prefer delinquent students as friends relative to non-delinquent students. Note that the evidence with respect to the parameters of primary interest, β_{X,k} (k = 4, 5, 6) and β_{Y,2}, agrees by and large...
with the meta-analysis of Knecht, Steglich, Baerveldt, and Snijders (2006)—based on the data sets for which ML estimation was feasible—which is encouraging. An important question is whether the marginal posterior densities of the parameters of primary interest, $\beta_{X,k}$ ($k = 4, 5, 6$) and $\beta_{Y,2}$, are insensitive to the prior distributions of the nuisance parameters $\alpha_{Y,h}$ ($h = 1, 2, 3$), which was one of the grounds on which modifications of the prior distributions of $\alpha_{Y,h}$ ($h = 1, 2, 3$) were defended (cf. Section 5.4.2): Figure 5.3 shows the marginal posterior densities of $\beta_{X,k}$ ($k = 4, 5, 6$) and $\beta_{Y,2}$ under Gamma($2, 6$) and Gamma($2, 6 \times \sqrt{2}$) prior distributions of $\alpha_{Y,h}$ ($h = 1, 2, 3$), where the expectation and variance of the Gamma($2, 6 \times \sqrt{2}$) distribution are $\sqrt{2}$ and 2 times as small as the expectation and variance of the Gamma($2, 6$) distribution, respectively, and thus the Gamma($2, 6 \times \sqrt{2}$) prior distributions are clearly more informative than the Gamma($2, 6$) prior distributions. Figure 5.3 suggests that the marginal posterior densities of $\beta_{X,k}$ ($k = 4, 5, 6$) and $\beta_{Y,2}$ indeed seem to be insensitive to the prior distributions of $\alpha_{Y,h}$ ($h = 1, 2, 3$). Perhaps more interesting than

**Figure 5.3:** Marginal posterior densities of $\beta_{X,k}$ ($k = 4, 5, 6$) and $\beta_{Y,2}$ under Gamma($2, 6$) and Gamma($2, 6 \times \sqrt{2}$) prior distributions of $\alpha_{Y,h}$ ($h = 1, 2, 3$)

$.025, .500, and .975 posterior quantiles are indicated by dotted lines.

Marginal posterior densities are posterior dependencies among parameters: Figure 5.4 shows scatter plots of posterior draws of interesting pairs of parameters; the parameters $\alpha_{X,h}$ and $\alpha_{Y,h}$ ($h = 1, 2, 3$), $\beta_{X,2}$, and $\beta_{X,3}$ are almost uncorrelated with
other parameters and therefore omitted. The parameters $\beta_{X,1}$, $\beta_{X,4}$, $\beta_{X,5}$, and $\beta_{X,6}$

Figure 5.4: Scatter plots of posterior draws of interesting pairs of parameters

![Figure 5.4: Scatter plots of posterior draws of interesting pairs of parameters](image)

are strongly correlated, as are the parameters $\beta_{Y,1}$ and $\beta_{Y,2}$; however, there do not seem to be correlations between parameters of these two subsets of parameters. The results make the estimation problems of the ML algorithm to some degree understandable, and suggest reparametrizing the model. Here, the model—motivated by substantive considerations of Knecht, Steglich, Baerveldt, and Snijders (2006)—was taken for granted, and the aim was to demonstrate that small modifications of the prior distribution of nuisance parameters can help to obtain estimates of parameters of primary interest in situations where other estimates (such as ML estimates) cannot be obtained; possible reparametrizations—which, in the ideal case, combine statistical and substantive insight—are therefore not considered here, but it is intended to consider possible reparametrizations in future work. Note, in addition, that Figure 5.4 demonstrates another potential advantage of the Bayesian approach: while correlations (linear dependencies) between MM or ML estimates can be assessed by (asymptotic approximations of) the variance-covariance matrix of the estimates, the Bayesian approach can give additional insight into non-linear dependencies between parameters—e.g., by scatter plots of posterior draws.
5.5. APPLICATIONS

5.5.2 Application II

School class 2-G of Knecht, Steglich, Baerveldt, and Snijders (2006) with $n = 30$ students is studied. Concerning the friendships, at time point $t_0$, 89 arcs were present; in time interval $[t_{h-1}, t_h]$ ($h = 1, 2, 3$), 33, 51, and 55 arcs were deleted while 65, 35, and 26 arcs were added, respectively. Concerning the delinquent behavior, at time point $t_0$, 19 students had value 1, and the .025 quantile was 1.0 while the .975 quantile was 3.3; in time interval $[t_{h-1}, t_h]$ ($h = 1, 2, 3$), 2, 6, and 4 students decreased the delinquent behavior while 9, 4, and 6 students increased the delinquent behavior, respectively.

The model is specified by rate functions $\lambda_{X,i} \equiv \alpha_{X,h}$ and $\lambda_{Y,i} \equiv \alpha_{Y,h}$ in time interval $[t_{h-1}, t_h]$ ($h = 1, 2, 3$), and objective functions

$$f_{X,i}(j \mid X, y, \theta) = \gamma_{X,ij} (x^*_{ij} - x_{ij}),$$

where

$$\gamma_{X,ij} = \beta_{X,1} + \beta_{X,2} x_{ji} + \beta_{X,3} (\sum_{h=1}^{n} x_{ih} x_{hj} + \sum_{l=1}^{n} x_{il} x_{jl})$$

$$+ \beta_{X,4} g_{ij}(y),$$

and

$$f_{Y,i}(j \mid X, y, \theta) = \beta_{Y,1} (y^*_{i} - y_{i}) + \beta_{Y,2} \sum_{l=1}^{n} x_{il} (g_{il}(y^*) - g_{il}(y)),$$

where

$$g_{il}(y) = \frac{(b - a) - |y_i - y_l|}{(b - a)}$$

can be interpreted as the similarity of students $i$ and $l$ in terms of the delinquent behavior, and $(b - a) = 3$ is the observed range of the delinquent behavior. In the Bayesian framework—guided by the same considerations as in Section 5.5.1—the prior distribution is specified by Gamma$(1, 10^{-10})$ and Gamma$(2, 2)$ for $\alpha_{X,h}$ and $\alpha_{Y,h}$ ($h = 1, 2, 3$), respectively, and $N(0, 10^1 I_6)$ for $\beta$, where $I_6$ is the $6 \times 6$ identity matrix.

The model was estimated by MM, ML, and Bayesian algorithms with similar specifications as in Section 5.5.1. The MM and ML algorithm seem to converge, while the Markov chain constructed by the Bayesian algorithm fails to converge to the posterior distribution within reasonable time. As in Section 5.5.1, there is evidence of long-tailed, marginal posterior distributions of $\alpha_{Y,h}$ ($h = 1, 2, 3$), and therefore the prior distributions of the nuisance parameters $\alpha_{Y,h}$ ($h = 1, 2, 3$) are modified (cf. Sections
first, Gamma(2, 6) and Gamma(2, 12) prior distributions were tried, which did not solve the convergence problem; however, under Gamma(2, 18) prior distributions the convergence problem vanished, and therefore all following results are based on Gamma(2, 18) prior distributions of $\alpha_{Y,h}$ ($h = 1, 2, 3$). MM, ML, and Bayesian estimates of $\theta$ are shown in Table 5.2, and the marginal posterior densities of $\theta$ are depicted in Figure 5.5. The ML estimates seem to be close to the modes of the marginal posterior densities (ignoring $\alpha_{Y,h}$, $h = 1, 2, 3$, having informative prior distributions), with the single exception of $\beta_{Y,1}$—for which, as in Section 5.5.1, there is no fully satisfying explanation; the marginal posterior densities of the parameters of primary interest, $\beta_{X,4}$ and $\beta_{Y,2}$, are fairly flat and indicate considerable uncertainty with respect to $\beta_{X,4}$ and $\beta_{Y,2}$, which is likewise captured by the large standard errors of the ML estimates; the scatter plot of the posterior draws of $\beta_{X,4}$ and $\beta_{Y,2}$—shown in Figure 5.6—suggests that $\beta_{X,4}$ and $\beta_{Y,2}$ are uncorrelated; the MM estimates are sometimes in the center, and sometimes in the tail of the marginal posterior densities; in addition, the confidence intervals, constructed by adding $\pm$ two standard errors to the MM estimates, sometimes overlap considerably with the corresponding 95% posterior intervals, and sometimes do not. The partial disagreement between MM estimation on one hand and likelihood-based inference on the other hand is not surprising (cf. Section 5.4.2), and, indeed, one tentative conclusion based on these and some other data sets is that—while MM estimation of network models (or models

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>MM (s.e.)</th>
<th>ML (s.e.)</th>
<th>posterior median</th>
<th>95% posterior interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{X,1}$</td>
<td>11.75 (3.708)</td>
<td>7.41 (1.124)</td>
<td>7.528</td>
<td>[5.718, 9.908]</td>
</tr>
<tr>
<td>$\alpha_{X,2}$</td>
<td>5.60 (0.739)</td>
<td>9.21 (1.150)</td>
<td>9.297</td>
<td>[7.03, 12.22]</td>
</tr>
<tr>
<td>$\alpha_{X,3}$</td>
<td>5.89 (0.789)</td>
<td>5.81 (0.933)</td>
<td>6.019</td>
<td>[4.36, 8.25]</td>
</tr>
<tr>
<td>$\beta_{X,1}$</td>
<td>-1.87 (0.079)</td>
<td>-1.78 (0.083)</td>
<td>-1.792</td>
<td>[-1.95, -1.63]</td>
</tr>
<tr>
<td>$\beta_{X,2}$</td>
<td>1.35 (0.163)</td>
<td>1.04 (0.133)</td>
<td>1.072</td>
<td>[0.81, 1.32]</td>
</tr>
<tr>
<td>$\beta_{X,3}$</td>
<td>0.17 (0.007)</td>
<td>0.21 (0.025)</td>
<td>0.205</td>
<td>[0.15, 0.26]</td>
</tr>
<tr>
<td>$\beta_{X,4}$</td>
<td>0.30 (0.40)</td>
<td>0.42 (0.42)</td>
<td>0.636</td>
<td>[-1.64, 1.49]</td>
</tr>
<tr>
<td>$\alpha_{Y,1}$</td>
<td>1.59 (0.816)</td>
<td>1.31 (0.397)</td>
<td>0.48</td>
<td>[0.28, 0.76]</td>
</tr>
<tr>
<td>$\alpha_{Y,2}$</td>
<td>1.52 (0.763)</td>
<td>1.57 (0.455)</td>
<td>0.583</td>
<td>[0.35, 0.92]</td>
</tr>
<tr>
<td>$\alpha_{Y,3}$</td>
<td>1.30 (0.630)</td>
<td>1.21 (0.361)</td>
<td>0.457</td>
<td>[0.25, 0.74]</td>
</tr>
<tr>
<td>$\beta_{Y,1}$</td>
<td>-0.49 (0.299)</td>
<td>-0.41 (0.221)</td>
<td>-1.192</td>
<td>[-1.83, -0.60]</td>
</tr>
<tr>
<td>$\beta_{Y,2}$</td>
<td>-0.32 (2.139)</td>
<td>-0.08 (1.205)</td>
<td>0.504</td>
<td>[-2.36, 3.22]</td>
</tr>
</tbody>
</table>
5.6 Discussion

A continuous-time Markov model for longitudinal data on networks and node-dependent outcome variables was presented, where the Markov process is not ob-
served in continuous time but at discrete time points, and where some covariates may be unobserved. The importance of Bayesian inference was stressed, because Bayesian inference has theoretical and practical advantages compared to MM estimation (Section 5.4.2), and because Bayesian inference has practical advantages compared to ML estimation (Sections 5.4.2 and 5.5.1), since in situations where ML algorithms do not converge the incorporation of prior knowledge can enable Bayesian inference, and since Bayesian inference—under vague prior distributions—can complement ML estimation by giving insight into the shape of the likelihood function and non-linear dependencies between parameters. Bayesian methods were proposed and compared to MM and ML estimation in two small, empirical data sets.

A number of issues remains, however: first, an important, practical problem is how to define suitable statistics to be incorporated in the objective functions, suitable in the sense that the statistics capture important mechanisms which are thought to drive the change of networks and node-dependent outcome variables, and which—at the same time—lead to estimable models, which is a non-trivial enterprise (cf. Section 5.5), and it is expected that much progress can be made on these grounds; and second, it is desirable to have model selection tools, so that the range of models that makes sense in the light of the observed data can be narrowed.
Figure 5.7: Marginal posterior densities of $\beta_{X,4}$ and $\beta_{Y,2}$ under $\text{Gamma}(2, 18)$ and $\text{Gamma}(2, 18 \times \sqrt{2})$ prior distributions of $\alpha_{Y,h}$ ($h = 1, 2, 3$).

.025, .500, and .975 posterior quantiles are indicated by dotted lines.
Chapter 6

Discussion

Studying longitudinal network and behavior data is important for understanding social processes, because human beings are interrelated, and the relationships among human beings (human networks) on one hand and human behavior on the other hand are not independent. Examples of dependencies in networks are reciprocity and transitivity (see Wasserman and Faust, 1994). Examples of dependencies between networks and behavior are selection and influence processes. Selection processes operate when actors (e.g., adolescents) change relationships (e.g., friendships), and when selecting relation partners actors take the behavior of potential relation partners (e.g., the consumption of alcohol and drugs) into consideration. Influence processes operate when actors (e.g., adolescents) change behavior (e.g., the consumption of alcohol and drugs), and when changing behavior actors are influenced by the behavior of existing relation partners (e.g., friends). Thus, changes in the network may be affected by behavior, and changes in the behavior may be affected by the network.

As a result of the dependencies, longitudinal network and behavior data are complex. Studying longitudinal network and behavior data therefore demands statistical methods, replacing (complex) data by (simple) summaries of the data (statistics, such as estimates, test statistics) containing as much as possible information about the social processes of interest (cf. Fisher, 1922). Statistical inference for longitudinal network data was pioneered by Holland and Leinhardt (1977) and Snijders (1996, 2001), and extended to longitudinal network and behavior data by Snijders, Steglich, and Schweinberger (2006). These models are based on the assumption that the network and behavior evolution is governed by a Markov process which operates in continuous time but is observed at two or more discrete time points. The Markov process is modeled as driven by actors who, at stochastic times, are allowed to change
either relationships or behavior; and when changing relationships, actors may take into account the behavior of potential relation partners, and when changing behavior, actors may take into account the behavior of existing relation partners. Snijders, Steglich, and Schweinberger (2006) proposed to estimate such models by the method of moments (MM).

The framework of Snijders, Steglich, and Schweinberger (2006) was the point of departure of the present thesis and was advanced by the present thesis both in terms of modeling and statistical inference. The advances made in the present thesis are summed up in Section 6.1. Possible directions of future research are sketched in Section 6.2.

6.1 Summary

The summary of the thesis proceeds chapter by chapter.

Chapter 2: Estimating functions: derivative estimation. In the MM framework of Snijders, Steglich, and Schweinberger (2006), the derivatives of the estimating function with respect to the parameters are required for (1) the standard errors of the parameter estimates; (2) goodness-of-fit test statistics; (3) sensitivity analyses; and (4) linear extrapolation. In the absence of closed-form expressions, the derivatives are estimated by Monte Carlo (MC) methods. Examining the properties of the conventional MC estimator of the derivatives, the finite differences (FD) estimator, revealed that the FD estimator is biased, inconsistent, associated with a bias-variance dilemma (cf. Lemma 1), and expensive in terms of computation time when the number of parameters is moderate or large. Using Lemma 2, three alternative MC estimators were proposed based on the likelihood ratio / score function method of derivative estimation, using variance reduction methods based on control variates. The proposed estimators proved to be consistent and unbiased (see Lemma 2), two of them tend to be more efficient than the FD estimator in terms of computation time, and one of them has minimum variance in a large class of estimators (see Lemma 3). Theoretical insight and two Monte Carlo simulation studies suggested that the FD estimator can bias conclusions considerably, and recommended the use of two of the proposed estimators as alternatives.

Chapter 3: Tests of goodness-of-fit. In the MM framework of Snijders, Steglich, and Schweinberger (2006), goodness-of-fit tests had not been considered before. A test
6.1. SUMMARY

A statistic was proposed that can be regarded as a generalized score test statistic based on regular estimating functions and that admits to test the goodness-of-fit of models which are restricted in one or more parameters. To evaluate the goodness-of-fit test statistic, it is not required to estimate the restricted parameters, which saves computation time and allows to test hard-to-estimate parameters; it was shown that, when the goodness-of-fit of the restricted model is found to be unacceptable, the estimates of all parameters (without restrictions) can be approximated by simplistic one-step estimates, which are crude approximations of the unrestricted estimates, but—other than the unrestricted estimates—require almost no additional computation time. A Monte Carlo simulation study indicated that the finite-sample null distributions of one- and multi-parameter goodness-of-fit tests are close to the expected null distributions; in addition, one-parameter goodness-of-fit tests were compared to conventional one-parameter $t$-tests. A large, empirical data set was studied and the value of the goodness-of-fit test statistic in forward model selection procedures was demonstrated; the one-parameter goodness-of-fit tests and the one-parameter $t$-tests turned out to agree by and large; and the one-step estimates appeared to be reasonable approximations of the unrestricted estimates.

Chapter 4: Random effects models. Snijders, Steglich, and Schweinberger (2006) assume that all relevant knowledge with respect to actors is observed in the form of covariates and correctly incorporated in the model. For longitudinal network data, models were proposed that take unobserved heterogeneity across actors into account by means of latent, actor-dependent variables (or random effects), which are governed by a probability law that is common to all actors. Estimating such models involves estimating the variance-covariance matrix of the random effects subject to symmetry and positive definiteness constraints. Maximum likelihood (ML) and Bayesian estimation methods were elaborated; for ML estimation, the non-redundant elements of the random effects variance-covariance matrix were reparametrized so that estimates of the variance-covariance matrix are by construction symmetric and positive definite, and the estimation of very small variances is facilitated. ML and Bayesian estimation were implemented by Markov chain Monte Carlo (MCMC) methods. The random effects models and methods were illustrated by an application to an empirical data set, where ML and Bayesian estimates of the random effects variance-covariance matrix were compared, and, in the Bayesian framework, the sensitivity of the posterior distribution of the random effects variance-covariance matrix to the prior distribution was examined.
Chapter 5: Bayesian modeling and estimation. Network and behavior models were considered, including extensions of the latent variable model of Chapter 4 to longitudinal network and behavior data. The importance of Bayesian inference was stressed, because it (1) is well-suited to studying unique, non-repeatable social processes, and (2) has practical advantages, since when ML algorithms do not converge, the incorporation of prior knowledge can enable Bayesian estimation, and when ML algorithms do converge, then Bayesian estimation can complement ML estimation by giving—under vague prior distributions—insight into the shape of the likelihood function and non-linear dependencies between parameters. A MCMC algorithm for sampling from the posterior distribution was proposed. MM, ML, and Bayesian estimation were compared on the basis of two small, empirical data sets: one data set where MM and ML estimation turned out to be problematic but Bayesian estimation was possible, and another data set where Bayesian and ML estimation agreed by and large, while MM estimation gave rise to deviating results.

Concluding remarks. The statistical methods of Chapters 2 and 3 have already proved to be useful in a number of applications (see, e.g., Knecht, in preparation, and Mercken, in preparation). The experience with respect to the statistical models and methods of Chapters 4 and 5 is limited, and more applications are required to explore the full potential of these models and methods; in the important special case of network models (with exogeneous behavior), however, the Bayesian methods of Chapter 5 have been applied by the author to a large number of simulated and empirical data sets, and have proved to be useful alternatives to MM and ML estimation.

All proposed models and methods were implemented by the author in the computer program Siena (Snijders, Steglich, Schweinberger, and Huisman, 2006), which is part of the program collection StOCNET (Boer, Huisman, Snijders, Steglich, Wichers, and Zeggelink, 2006) and can be obtained from the website http://stat.gamma.rug.nl/stocnet.

6.2 Future research

The present thesis, and in particular Chapters 4 and 5, have opened the gates to a large number of model extensions and applications. First and foremost, the introduction of latent variables and the corresponding estimation framework allows to build and estimate models which make less restrictive assumptions about the data-
generating process than Snijders, Steglich, and Schweinberger (2006), and thus are expected to improve the goodness-of-fit of models. A great number of other latent variable models are conceivable, which remove one or another restrictive assumption about the data-generating process, such as measurement models, where relationships are allowed to be measured with random errors, or multilevel models, where network and behavior are measured on multiple, non-overlapping sets of actors and parameters are allowed to vary within and between sets of actors; both classes of models could be obtained as extensions of the latent variable framework considered here. Other examples of latent variable models are latent structure models, which assume that there is some latent structure that has an impact on the network and behavior evolution, e.g., latent classes to which actors belong and which influence the probability of changes, or latent metric spaces in which actors are positioned (cf. Schweinberger and Snijders, 2003) and which influence the probability of changes through the distances between actors.

Two remarks are in place, however. First, before considering other latent variable models, some issues need to be addressed for the latent variable models proposed here, in the first place, model selection issues (cf. Chapter 4). Second, since the Markov process, which is assumed to govern the network and behavior evolution, is not observed in continuous time but at discrete time points, one is well-advised to keep the model as parsimonious as possible and avoid the use of latent variables unless there are good reasons.
Appendix A

Guide to \texttt{R} functions

To examine the Markov chain Monte Carlo (MCMC) output of the computer program \texttt{Siena} (Snijders, Steglich, Schweinberger, and Huisman, 2006) for maximum likelihood (ML) and Bayesian estimation, the \texttt{R} functions \texttt{siena.mle} and \texttt{siena.bayes} can be used, respectively. The \texttt{R} functions input files generated by \texttt{Siena} and output, among other things, trace plots and MCMC lag 1, \ldots, 100 autocorrelations of sampled entities (see Chapters 4 and 5), and, in the Bayesian case, in addition 95\% posterior intervals, histograms, and Gaussian kernel density estimates of the marginal posterior densities of the parameters.

The \texttt{R} functions \texttt{siena.mle} and \texttt{siena.bayes} can be downloaded from the website \texttt{http://stat.gamma.rug.nl/stocnet}, and can be used in \texttt{R} as follows:

1. Load the \texttt{R} function:
   - ML estimation: \texttt{source("siena.mle.r")}.
   - Bayesian estimation: \texttt{source("siena.bayes.r").}

2. Call the \texttt{R} function:
   - ML estimation: \texttt{siena.mle(project\_name, full\_output, no\_random\_effects, no\_actors)}.
   - Bayesian estimation: \texttt{siena.bayes(project\_name, full\_output, no\_random\_effects, no\_actors)}.

The arguments are:

- \texttt{project\_name} (string): the name of the \texttt{Siena} project that is to be examined; note that calling \texttt{siena.mle} or \texttt{siena.bayes} presumes that \texttt{Siena} carried out ML or Bayesian estimation of the project \texttt{project\_name}, respectively.
— **full_output** (0 or 1): 1 indicates that the full output is desired, while 0 indicates that selected output is desired.

— **no_random_effects** (non-negative integer): the number of actor-dependent weights (parameters) in the model.

— **no_actors** (positive integer): the number of actors.

Examples are provided by `siena_mle("alcohol", 1, 3, 50)` and `siena_bayes("alcohol", 1, 3, 50)`. 
Sociale relaties tussen actoren (netwerken) en het gedrag van actoren hangen samen. Voorbeelden van samenhang tussen netwerken en gedrag zijn selectie- en invloedsprocessen. Selectieprocessen hebben betrekking op processen waarbij actoren hun relaties veranderen en daarbij rekening houden met het gedrag van andere actoren. Invloedsprocessen hebben betrekking op processen waarbij actoren hun gedrag veranderen en daarbij beïnvloed worden door het gedrag van andere actoren met wie de actoren een relatie hebben. Een voorbeeld zijn vriendschappen tussen scholieren en de consumptie van alcohol: er is sprake van een selectieproces indien scholieren vriendschappen sluiten met scholieren die evenveel alcohol drinken, en er is sprake van een invloedsproces indien scholieren hun alcoholconsumptie aanpassen aan de alcoholconsumptie van hun vrienden.

De statistische modellen van SSS werden door de momentenmethode (MM) geschat. Het uitgangspunt van dit proefschrift waren de statistische modellen en methoden van SSS, en de bedoeling was om zowel de statistische modellen als methoden door te ontwikkelen. De ontwikkelingen werden beschreven in Chapters 2—5 en kunnen als volgt worden samengevat.

Chapter 2: Estimating functions: derivative estimation. MM schattingen vereisen de afgeleiden van de schattingsfunctie naar de parameters voor (1) de standaardfouten van de parameterschattingen; (2) goodness-of-fit toetsen; (3) gevoeligheidsanalyses; en (4) lineaire extrapolatie. Omdat de afgeleiden niet analytisch bepaald kunnen worden, worden de afgeleiden door Monte Carlo (MC) methoden geschat. Er werd wiskundig aangetoond dat de gangbare MC schatter van de afgeleiden, de finite differences (FD) schatter, noch zuiver noch consistent is, met een bias-variantie dilemma verbonden is—omdat een kleine bias met een grote MC variantie samengaat en een kleine MC variantie met een grote bias (vgl. Lemma 1)—en veel rekentijd vergt als het aantal parameters matig of groot is. Drie alternatieve MC schatters werden voorgesteld, gebaseerd op de likelihood ratio / score function methode van afgeleidenschattingen en variantie-reductie methoden met controle-variaten. Er werd wiskundig aangetoond dat de schatters zuiver en consistent zijn (vgl. Lemma 2) en dat een van de drie schatters de kleinste variantie in een grote klas van schatters heeft (vgl. Lemma 3). Daarnaast vergen twee van de drie schatters minder rekentijd dan de FD schatter. Twee Monte Carlo simulatiestudies bevestigden dat de FD schatter conclusies inderdaad serieus kan vertekenen, en bevalen twee van de drie voorgestelde schatters als alternatieven aan.

Chapter 3: Tests of goodness-of-fit. Een toetsinggrootheid werd voorgesteld waar- mee de goodness-of-fit van modellen met restricties op een of meer parameters getoetst kan worden. Om goodness-of-fit toetsen te berekenen, hoeft men de getoetste parameters niet te schatten, waardoor rekentijd wordt bespaard en lastig te schatten parameters getoetst kunnen worden. Er werd wiskundig aangetoond dat, indien de goodness-of-fit van het model met restricties als onaanvaardbaar wordt beschouwd, de parameterschattingen van het volwaardige model benaderd kunnen worden door simplistische een-stap schattingen, die—anders dan de volwaardige schattingen—weinig rekentijd vergen. Uit een Monte Carlo simulatiestudie bleek dat de verdelingen van een- en multi-parameter goodness-of-fit toetsen onder het model met restricties de verwachte verdelingen goed benaderen; daarnaast werden een-parameter goodness-
Summary in Dutch

Chapter 4: Random effects models. Statistische modellen voor longitudinale netwerkd data werden voorgesteld die een kritieke, impliciete veronderstelling van SSS verwijderen, en wel de veronderstelling dat alle relevante kennis met betrekking tot actoren in de vorm van covariaten is waargenomen en in het model correct is verwerkt. De voorgestelde modellen houden rekening met niet-waargenomen, actorspecifieke verschillen door middel van latente, actorspecifieke variabelen (random effects). Schattingen van random-effects modellen moeten voldoen aan de randvoorwaarde dat de random-effects variantie-covariantie matrix symmetrisch en positief-definiet is. Maximum likelihood (ML) en Bayesiaanse schattingsmethoden voor random-effects modellen werden uitgewerkt. In het ML geval werd een herparametrisatie van de random-effects variantie-covariantie matrix gebruikt die automatisch symmetrische en positief-definiete schattingen van de variantie-covariatie matrix oplevert en het schatten van heel kleine varianties vergemakkelijkt. De ML en Bayesiaanse schattingsmethoden werden geïmplementeerd met behulp van Markov chain Monte Carlo (MCMC) methoden. De random-effects modellen werden toegepast op een empirische data set, waarbij de ML en Bayesiaanse schattingen van de random-effects variantie-covariantie matrix met elkaar werden vergeleken en de gevoeligheid van de a posteriori verdeling van de random-effects variantie-covariantie matrix voor de a priori verdeling werd onderzocht.

Chapter 5: Bayesian modeling and estimation. Het belang van Bayesiaanse schatting van longitudinale netwerk- en gedragsmodellen, ingesloten uitbreidingen van de latente-variabelen modellen van Chapter 4, werd benadrukt. Een Bayesiaanse benadering (1) is geschikt om unieke, niet-herhaalbare sociale processen te bestuderen, en (2) heeft praktische voordelen, omdat als ML algoritmes niet convergeren, a priori kennis Bayesiaanse schatting mogelijk kan maken, en als ML algoritmes convergeren, Bayesiaanse schatting ML schatting kan aanvullen door inzicht te geven in de vorm van de likelihood functie en niet-lineaire afhankelijkheden tussen parameters. Een MCMC algoritme voor het trekken uit de a posteriori verdeling werd voorgesteld. MM, ML, en Bayesiaanse schatting werden met elkaar vergeleken op de basis van
twee kleine, empirische data sets. In een data set waren MM en ML schatting problematicch, terwijl Bayesiaanse schatting mogelijk was. In de andere data set leidden Bayesiaanse en ML schatting tot overeenkomende conclusies, terwijl MM schatting tot afwijkende conclusies leidde.

Bibliography


Bera, A. K. and Y. Bilias (2001b). Rao’s score, Neyman’s C(α) and Silvey’s LM tests: an essay on historical developments and some new results. *Journal of Statistical Planning and Inference* 97, 9–44.


Freeman, L. C. (2004). *The Development of Social Network Analysis: A Study in the Sociology of


Pearson, K. (1900). On the criterion that a given system of observations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine* 50, 157–175.


social network dynamics. Working paper.
ICS dissertation series

The ICS series presents dissertations of the Interuniversity Center for Social Science Theory and Methodology. Each of these studies aims at integrating explicit theory formation with state of the art empirical research or at the development of advanced methods for empirical research. The ICS was founded in 1986 as a cooperative effort of the universities of Groningen and Utrecht. Since 1992, the ICS expanded to the University of Nijmegen. Most of the projects are financed by the participating universities or by the Netherlands Organization for Scientific Research (NWO). The international composition of the ICS graduate students is mirrored in the increasing international orientation of the projects and thus of the ICS series itself.


