Modeling Heterogeneity within and between Matrices and Arrays

Bailey Kathryn Fosdick

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Peter D. Hoff, Chair

Adrian E. Raftery

Michael D. Perlman

Program Authorized to Offer Degree:
UW Department of Statistics

University of Washington

Abstract

Modeling Heterogeneity within and between Matrices and Arrays

Bailey Kathryn Fosdick

Chair of the Supervisory Committee:
Professor Peter D. Hoff
Department of Statistics

Datasets in the form of matrices and arrays arise frequently in the social and biological sciences and are characterized by measurements indexed by two or more factors. In this dissertation we address two problems relating to these datasets.

In the case of a single observed array Y, the primary goal in an analysis is often to decompose the array into $Y = M + E$, where $M = f(X, \beta)$ represents a function of covariates X and unknown parameter β and E represents an array of random noise. Typically the errors E are assumed to be independent or dependent along at most one of the array dimensions. Failing to account for other dependencies can lead to inefficient estimates of β, inaccurate standard errors and poor predictions. An alternative to assuming independent errors is to allow for dependence along each dimension of the array using a separable covariance model. However, for many arrays maximum likelihood estimates of the covariance matrices in this model do not exist. We propose a submodel of the separable covariance model that restricts some of the covariance matrices to have factor analytic structure; this model can be viewed as extension of factor analysis to array-valued data.

The second problem we address is specific to matrices that contain network data where the row and column index sets represent a set of actors. Frequently the objective in network analysis is to determine whether dependencies exist between a matrix of network relations and a matrix of actor-specific attributes. Approaches to this problem often condition on either the relations or attributes, require specification of the exact nature of the association

between the network and attributes, and are unable to provide predictions simultaneously for missing attribute and network information. We propose methodology for a unified approach to analysis that allows for testing for dependencies between the relations and attributes, and in the event the test concludes such structure exists, jointly modeling the relations and attributes to conduct inference and make predictions for missing values. We investigate Bayesian estimation procedures for a general class of relational data models, significantly improve the efficiency of a Markov chain Monte Carlo algorithm, and illustrate the inadequacies of a mean-field variational approach for this model class.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I wish to thank Peter Hoff for his enormous patience and commitment to me over the past three years. I have learned so much from him and I know much of my success going forward will be attributable to his great teaching, constructive criticism, and valuable mentoring.

I also want to express my sincere gratitude to Adrian Raftery and Michael Perlman, both of whom I have had the opportunity to learn from as teachers, research collaborators, and mentors.

I cannot say enough wonderful things about the Department of Statistics at the University of Washington. The faculty, specifically Peter Guttorp, June Morita, and Mark Handcock, were amazingly generous with their time, wisdom, and support. In addition, my fellow graduate students have stretched my knowledge in statistics and beyond. I thank Alexander Volfovsky specifically for his insights and helpful discussions on the work presented here.

I wish to acknowledge my amazing friends and family who have supported me throughout graduate school. My husband Tyler is my inspiration and a pillar of strength. My parents, Paul and Ann, and parents-in-law, Mike and Martha, always believed in me. Finally, I wish to thank Connie and Jason, who are truly the best friends I could ever ask for, and the UW Physical Therapy Class of 2013 for keeping me active and being a constant source of laughs.

DEDICATION

To my forever loving husband Tyler.

Chapter 1

INTRODUCTION

Datasets in the form of matrices and arrays arise frequently in the social and biological sciences and are characterized by measurements indexed by two or more factors. Examples of such data include a four-way array of human death rates, the dimensions of which correspond to the four categories age, sex, country and year, and an adolescent friendship network containing measurements of friendship between pairs of students. In this dissertation we present methodology for two problems relating to these datasets: modeling heterogeneity within an array and modeling heterogeneity between a relational data matrix and matrix of actor-specific attributes.

1.1 Modeling heterogeneity within an array

The primary goal in an analysis of a single array Y is often to decompose the array into $Y = M + E$, where $M = f(X, \beta)$ represents a function of covariates X and unknown parameter β and E represents an array of random noise. Typically the errors E are assumed to be independent or dependent along at most one of the array dimensions. Failing to account for other dependencies can lead to inefficient estimates of regression parameters β, inaccurate standard errors and poor predictions. An alternative to assuming independent errors is to allow for dependence along each dimension of the array using a separable covariance model (Hoff (2011)). However, the number of parameters in this model increases rapidly with the dimensions of the array, and for many arrays, maximum likelihood estimates of the covariance parameters do not exist (Manceur and Dutilleul (2013)).

In Chapter 2, we propose a submodel of the separable covariance model that estimates the covariance matrix for each dimension as having factor analytic structure. This model can be viewed as an extension of factor analysis to array-valued data, as it uses a factor model to estimate the covariance along each dimension of the array. We discuss properties

of this model as they relate to ordinary factor analysis, describe maximum likelihood and Bayesian estimation methods, and provide a likelihood ratio testing procedure for selecting the factor model ranks.

Furthermore, we apply this methodology to the analysis of death rates from the Human Mortality Database and show in a cross-validation experiment how it outperforms simpler methods. We use our model to impute mortality rates for countries that have no mortality data for several years. Unlike other approaches, our methodology is able to estimate similarities between the mortality rates of countries, time periods, and sexes and use this information to assist with the imputations.

1.2 Modeling heterogeneity between a relational matrix and attribute matrix

Network analysis is often focused on characterizing the dependencies between a matrix of network relations and a matrix of actor-level attributes. Potential relationships are typically explored by modeling the network as a function of the actor attributes or by modeling the attributes as a function of the network. These methods require specification of the exact nature of the association between the network and attributes, reduce the network data to a small number of summary statistics, and are unable provide predictions simultaneously for missing attribute and network information. Existing methods that model the attributes and network jointly also assume the data are fully observed. In Chapter 3 we introduce a unified approach to analysis that addresses these shortcomings. We use a latent variable model to obtain a low dimensional representation of the network in terms of actor-specific network factors and use a test of dependence between the network factors and attributes as a surrogate for a test of dependence between the network and attributes. We propose a formal testing procedure to determine if dependencies exists between the network factors and attributes. We also introduce a joint model for the network and attributes, for use if the test rejects, that can capture a variety of dependence patterns and be used to make inference and predictions for missing observations.

In Chapter 4 we investigate Bayesian estimation procedures for a general class of relational data models based on the latent variable model discussed in Chapter 3, which can accommodate continuous, ordinal, binary, and censored relations. We discuss a basic

Markov chain Monte Carlo algorithm and propose three modifications to it that significantly improve the efficiency of the sampler and accuracy of the posterior approximation. We also discuss a mean-field variational approach for this model class and present a simulation study which suggests the approximation is relatively accurate for continuous relational observations, yet severely underestimates the parameter uncertainty in the posterior distribution when the observed relations are non-continuous.

Chapter 2

SEPARABLE FACTOR ANALYSIS WITH APPLICATIONS TO MORTALITY DATA

2.1 Introduction

Human mortality data are used extensively by researchers and policy makers to analyze historic and current population trends and assess long-term impacts of public policy initiatives. To enable such inference, numerous regression models have been proposed that estimate mortality rates as a function of age using a small number of parameters (Heligman and Pollard (1980), Mode and Busby (1982), Siler (1983)). Practitioners using these methods typically model the age-specific death rates for each country, year, and sex combination separately and assume independent error distributions. Examples of death rates analyzed by such methods are shown in Figure 2.1 for the United States and Sweden. Each mortality curve is defined by 23 age-specific death rates and the average sex-specific mortality curve from 1960-1980 for thirty-eight countries is also displayed.

From the figure, it is clear that a country's mortality rates in one time period are similar to its rates in adjacent time periods. Acknowledging this fact, several researchers have developed models for "dynamic life tables", i.e. matrices of mortality rates for combinations of ages and time periods, for single country-sex combinations. An example of such a life table is the male death rates in Sweden from 1960 to 1980 shown in Figure 2.1. Some models developed for these data specify ARIMA processes for the time-varying model parameters (McNown and Rogers (1989), Renshaw and Haberman (2003b)), while others smooth the death rates over age and time using a kernel smoother (Felipe et al. (2001)), p-splines (Currie et al. (2004)), nonseparable age-time period covariance functions (Martínez-Ruiz et al. (2010)), or multiplicative effects for age and time (Lee and Carter (1992), Renshaw et al. (1996), Renshaw and Haberman (2003a), Renshaw and Haberman (2003c), Chiou and Müller (2009)).

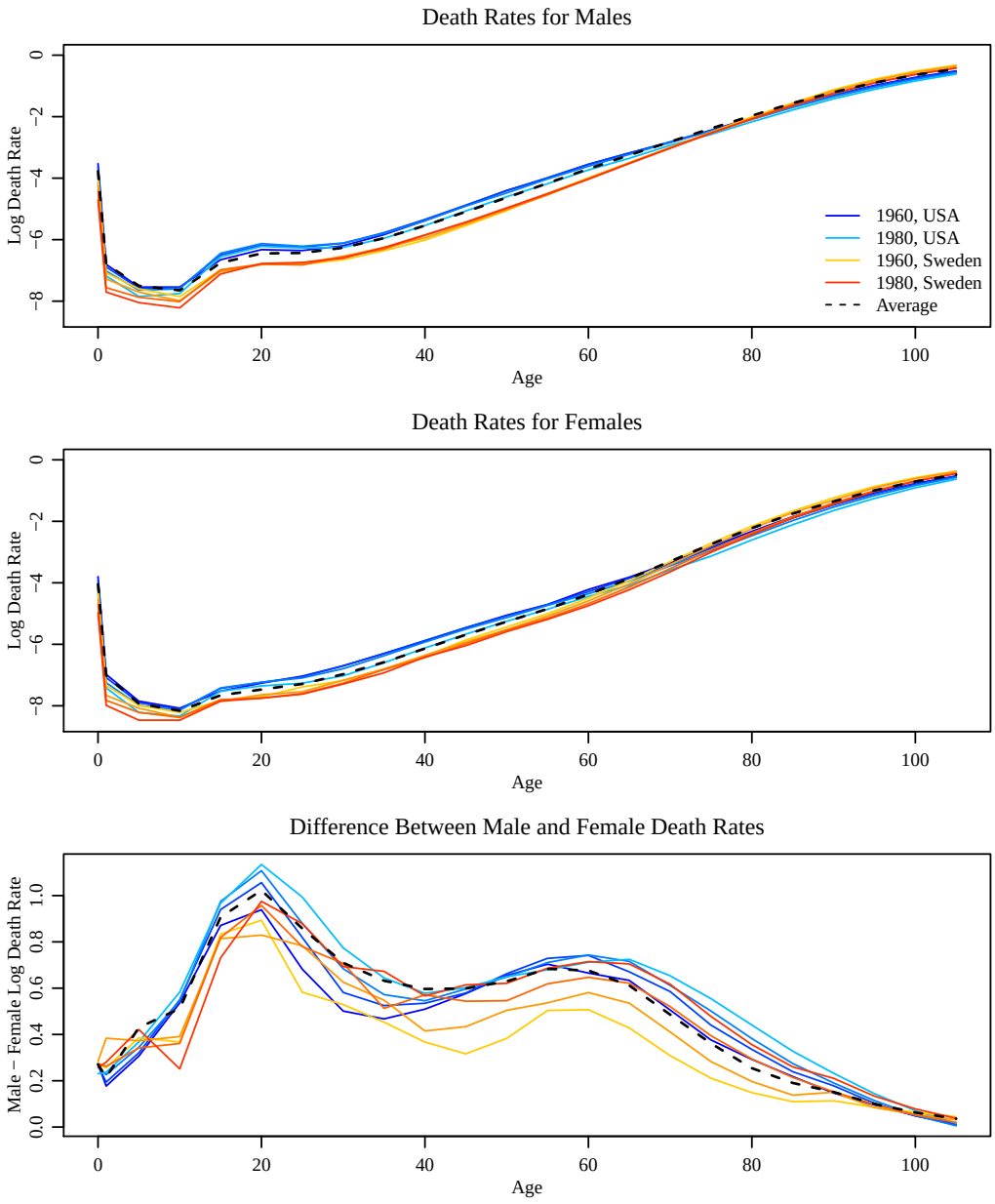

Figure 2.1: Mortality curves for the United States of America and Sweden. The gradient of colors for each country represents the log death rates in the four 5-year time periods from 1960 to 1980. The average sex-specific mortality curve over the four time periods and all countries is shown in black.

Human mortality datasets typically provide mortality rates of populations corresponding to combinations of several factors. For example, the Human Mortality Database (HMD, Human Mortality Database, 2011) provides mortality rates of populations corresponding to combinations of 40 countries, 9 time periods, 23 age groups, and both male and female sexes. As is shown in Figure 2.1, mortality rates of men and women within a country will typically both be higher than or both lower than the sex-specific rates averaged across countries. Furthermore, differences between male and female mortality rates generally show trends that are consistent across countries and time periods. Such patterns suggest joint estimation of mortality rates using a model that can share information across levels of two or more factors. Two models that consider death rates for more than one country or sex are that developed by Li and Lee (2005), which estimates common age and time period effects for a group of countries or both sexes, and Carter and Lee (1992), where male and female death rates within the same country share a time-varying mortality level. Although these methods consider either both sexes or multiple countries, the extreme similarity of the curves in Figure 2.1 for males across countries and for a given country across sexes suggest that separately modeling death rates for different countries or sexes is inefficient, and inference may be improved by using a joint model that shares information across all factors.

With this in mind, we consider a regression model for the HMD data consisting of a mean model that is a piecewise-polynomial in age with additive effects for country, time period, and sex (more details on this model, and its comparison to other models, are provided in Section 2.4). This mean model is extremely flexible: It contains over 370 parameters and an ordinary least squares (OLS) fit accounts for over 99% of the total variation in the data (coefficient of determination, $R^2 > 0.995$). Nonetheless, an analysis of the residuals from the OLS fit indicates that some clear patterns in the data are not captured by the regression model, and in particular, a model of independent errors is a poor representation of these data. To illustrate this, note that the residuals can be represented as a 4-way array, the dimensions of which are given by the number of levels of each of the four factors: country, time period, sex and age. To examine residual correlation across levels of a factor, the 4-way array of residuals can be converted into a matrix whose columns represent the levels

of the factor, and a sample correlation matrix for the factor can be obtained. Figure 2.2 summarizes the patterns in the residual correlations using the first two principal components of each sample correlation matrix. If a model of independent errors were to be adequate, we would expect the sample correlation values to be small and centered about zero, and no discernible patterns to exist in the principal components. However, the sample correlations are substantially more positive than would be expected under independence: 59% of the observed country correlations, 61% of time period correlations and 98% of age correlations are greater than the corresponding 95% theoretical percentiles under the independence assumption. Additionally, there are clear geographic, temporal, and age trends in the principal components in Figure 2.2. For example, the residuals for the Ukrainian mortality rates are positively correlated to those for Russia, and the residuals for the year 2000 are positively correlated with those for 1995. This residual analysis suggests that an assumption of uncorrelated errors is inappropriate on account of the positive correlation exhibited by residuals across levels of the factors.

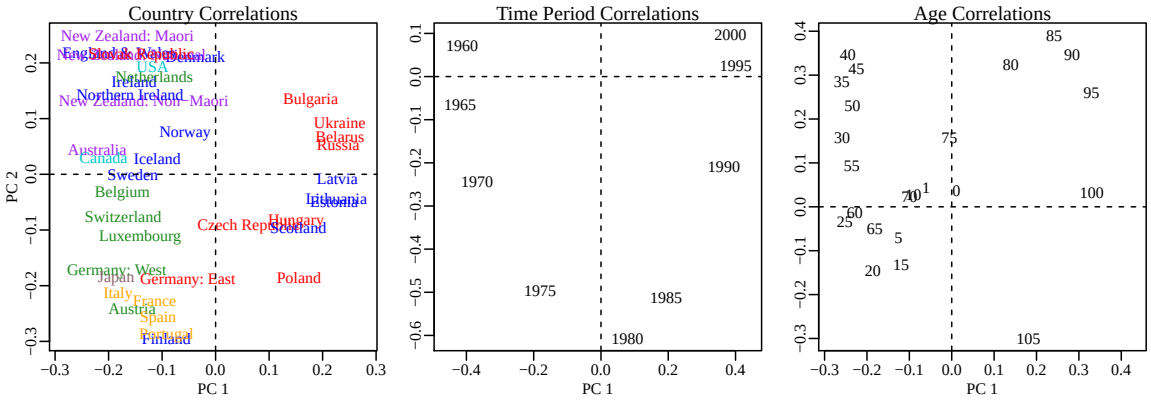

Figure 2.2: The first two principal components of each sample correlation matrix are displayed, and countries in the same United Nations region are shown in the same color. Close proximity in the principal components space away from the origin is indicative of a positive correlation.

Failure to recognize correlated errors can lead to a variety of inferential problems, such as inefficient parameter estimates and inaccurate standard errors. For the analysis of the

mortality data, an additional important consequence is that the accuracy of predictions of missing mortality rates may suffer. Predicting missing death rates is a primary application of modeling mortality data, as developing countries often lack reliable death registration data. It is possible that the residual dependence could be reduced by increasing the flexibility of the mean model, but since this is already fairly complex, we may instead prefer to represent residual dependence with a covariance model, leading to a general linear model for the data in which the mean function and residual covariance are estimated simultaneously.

The mortality data, like the residuals, can be represented as a 4-way array, each dimension of which corresponds to one of the factors of country, time period, sex and age. In the literature on multiway array data (see, for example, Kroonenberg (2008)), each dimension is referred to as a mode of the array, so a 4-way array of mortality data consists of four modes. As described by Hoff (2011), a natural covariance model for a K-way data array is a separable covariance model, parameterized in terms of K covariance matrices, one for each mode of the array. If the array is also assumed to be normally distributed, the model is referred to as the array normal model, and can be seen as an extension of the matrix normal model (Dawid (1981)).

Even though the separable covariance model is not a full, unstructured covariance model, the array normal likelihood is unbounded for many array dimensions, prohibiting the use of maximum likelihood methods (Manceur and Dutilleul (2013)). Estimates of the array normal covariance parameters can still be obtained by taking a Bayesian approach (Hoff (2011)) or by using a penalized likelihood (Allen and Tibshirani (2010)). However, the lack of existence of the maximum likelihood estimates (MLEs) indicates that the data is unable to provide information about all of the parameters. In this chapter we propose an alternative modeling approach that parameterizes the covariance matrix of each mode as a reduced rank matrix plus a diagonal matrix. This covariance structure is commonly associated with factor analysis, and is referred to here as factor analytic covariance structure. We call this new model the Separable Factor Analysis (SFA) model, as it is an extension of factor analysis to array-valued data. The reduction in the number of parameters by using covariance matrices with factor analytic structure leads to existence of MLEs for the SFA parameters in many cases when the MLEs of the array normal parameters do not exist, as well as a parsimonious

representation of mode-specific covariance in an array-valued dataset.

This chapter is outlined as follows: In the next section we introduce and motivate SFA, as well as discuss its properties and similarities to ordinary factor analysis. We describe two estimation procedures in Section 2.3: an iterative maximum likelihood algorithm and a Metropolis-Hastings sampler for inference in a Bayesian framework. A likelihood ratio testing procedure for selecting the rank of the factor model for each mode is also presented. In Section 2.4 the SFA model is used to analyze the HMD mortality data and its performance is compared to simpler covariance models in a simulation study. We illustrate how SFA uses estimated similarities between country mortality rates to provide imputations for countries missing mortality data for several years. This prediction method extends the approach taken in Coale and Demeny (1966), Brass (1971), United Nations (1982), and Murray et al. (2003), where one country's mortality curve is modeled a function of another's. Our approach is novel in that it estimates the covariance between mortality rates across all countries, time periods, and sexes, and uses these relationships to impute missing death rates. We conclude with a discussion in Section 2.5.

2.2 Extending factor analysis to arrays

2.2.1 Motivating separable factor analysis

Suppose Y is a K-way array of dimension $m_1 \times m_2 \times ... \times m_K$. We are interested in relating the data Y to explanatory variables X through the model $Y = M(X, \beta) + E$, where β represents unknown regression coefficients and E represents the deviations from the mean. As was discussed in the preliminary analysis of the mortality data in Section 2.1, it is often unreasonable to assume the elements of E are independent and identically distributed.

In cases where there is no independent replication, estimation of the Cov$[E]$ can be problematic as it must be based on essentially a single sample. One solution is to approximate the covariance matrix with one with simplified structure. A frequently used model in spatio-temporal analysis is a separable covariance model (Stein (2005), Genton (2007)). This model estimates a covariance matrix for each mode of the array and is written Cov$[\text{vec}(E)] = \Sigma_K \otimes \Sigma_{K-1} \otimes ... \otimes \Sigma_1$, where "vec" and "$\otimes$" denote the vectorization

and Kronecker operators, respectively. In the context of the mortality data, this model contains a covariance matrix for country (Σ_c), time period (Σ_t), age (Σ_a), and sex (Σ_s). A separable covariance model with the assumption that the deviations are normally distributed, $\text{vec}(E) \sim \text{normal}(0, \text{Cov}[\text{vec}(E)])$, is an array normal model and was developed by Hoff (2011) as an extension of the matrix normal (Dawid (1981), Browne (1984), Oort (1999)).

The mode covariance matrices in the array normal model are not estimable for certain array dimensions using standard techniques such as maximum likelihood estimation (Manceur and Dutilleul (2013)). However, often the covariance matrices of large modes can be well approximated by matrices with simpler structure. A common approach in the social sciences to modeling the covariance matrix of a high-dimensional random vector is to use factor analysis. The standard k-factor model for a random vector $x \in \mathbb{R}^p$ parameterizes the covariance matrix as $\text{Cov}[x] = \Lambda\Lambda^T + D^2$, where $\Lambda \in \mathbb{R}^{p\times k}$, $k < p$, and D is a diagonal matrix (Spearman (1904), Mardia et al. (1979)). We will refer to this model as single mode factor analysis as it models the covariance among one set of variables. When the number of independent observations n is less than p, the sample covariance matrix is not positive definite and hence cannot be used as an estimate of $\text{Cov}[x]$. Nevertheless, under the assumption that x follows a multivariate normal distribution with known mean, the maximum likelihood estimate of the factor analytic covariance matrix exists if $k < \min(p, n)$ (Robertson and Symons (2007)).

We propose a submodel of the array normal model where each mode covariance matrix potentially has factor analytic structure. We call this model *Separable Factor Analysis (SFA)* and it is written as follows:

$$\begin{aligned} \text{vec}(E) &\sim \text{normal}(0, \text{Cov}[\text{vec}(E)]) \\ \text{Cov}[\text{vec}(E)] &= \Sigma_K \otimes \Sigma_{K-1} \otimes ... \otimes \Sigma_1, \\ \text{where } \Sigma_i &= \Lambda_i\Lambda_i^T + D_i^2 \quad \text{for } 0 \le k_i < m_i \end{aligned} \tag{2.1}$$

and Σ_i is unconstrained (i.e. equals any positive definite matrix) if $k_i = m_i$. SFA models are characterized by the covariance matrix structure chosen for each mode and can be

represented by a K-vector of ranks $(k_1, ..., k_K)$, where k_i equals the rank of Λ_i if mode i's covariance matrix has factor analytic structure and equals m_i if the mode covariance matrix is unstructured. Note that we consider the $k_i = 0$ case where the covariance matrix is diagonal, reflecting independence of entries along the mode. A k_i-factor analytic covariance matrix, $\Lambda_i\Lambda_i^T + D_i^2$, contains $\delta(m_i, k_i) = \left[(m_i - k_i)^2 - (m_i + k_i)\right]/2$ fewer parameters than an unstructured $m_i \times m_i$ covariance matrix. If $\delta(m_i, k_i) \leq 0$, the factor analytic covariance matrix does not provide reduced structure, and to prevent overparameterizing the model, either the factor analytic rank, k_i, should be decreased or an unstructured covariance matrix should be specified.

SFA has advantages over the array normal model that stem from it being an extension of factor analysis to multiple modes. Each mode covariance matrix can be interpreted independently as a decomposition of the mode variability into shared and unique latent components as in classical factor analysis (see Properties below). In addition, empirical evidence has shown that MLEs of the SFA covariance parameters exist for array dimensions where the MLEs of the array normal unstructured covariance matrices do not exist.

2.2.2 Properties of SFA

In this section we relate the SFA parameters to those in ordinary factor analysis, discuss indeterminancies in the model, and interpret the SFA parameters when the true covariance matrix in each mode is unstructured. We use the concept of array matricization to describe how these properties relate to each mode. Here we define matricizing an array in the ith mode as unfolding the array into a matrix $Y_{(i)}$ of dimension $(m_i \times \prod_{j \neq i} m_j)$, where an earlier mode index always varies faster than a later mode index in the columns (Kolda and Bader (2009)). There are multiple ways to matricize an array but here we follow this convention (Kiers (2000), De Lathauwer et al. (2000)).

Latent variable representation

A single mode k-factor model for a sample of n mean zero p-variate random vectors is written $\{x_1, ..., x_n\} \sim$ i.i.d. normal$(0, \Lambda\Lambda^T + D^2)$, where $\Lambda \in \mathbb{R}^{p \times k}$ and D is a diagonal

matrix. Collecting the random vectors in a $p \times n$ matrix $X = [x_1, ..., x_n]$, this model has an equivalent latent variable representation as a decomposition into common latent factors, $Z = [z_1, ..., z_n]$, and variable specific latent factors, $E = [e_1, ..., e_n]$, as follows.

$$\begin{aligned} X_{p\times n} &= \Lambda_{p\times k} Z_{k\times n} + D_{p\times p} E_{p\times n} \\ \{z_1, ..., z_n\} &\sim \text{i.i.d. normal}(0, \mathrm{I}_k) \qquad \text{Cov}[z_i, e_j] = 0_{k\times p} \quad \text{for all } i, j \\ \{e_1, ..., e_n\} &\sim \text{i.i.d. normal}(0, \mathrm{I}_p) \end{aligned} \tag{2.2}$$

In this representation the jth observation of the ith variable X_{ij} is written as a linear combination of common latent factors z_i with coefficients given by the ith row of Λ, plus a single variable specific factor E_{ij}, whose coefficient is the ith diagonal element of D.

A similar representation exists for each mode with a factor analytic covariance structure in the SFA model. Consider a mean zero array Y and an SFA model with a factor analytic covariance matrix in the ith mode. Define $\widetilde{Y}^i$ to be the array obtained by standardizing Y with all but the ith mode's covariance matrix:

$$\text{vec}(\widetilde{Y}^i) := \text{vec}(Y)(\Sigma_K^{-1/2} \otimes ... \otimes \Sigma_{i+1}^{-1/2} \otimes \mathrm{I}_{m_i} \otimes \Sigma_{i-1}^{-1/2} \otimes ... \otimes \Sigma_1^{-1/2}). \tag{2.3}$$

It follows that

$$\widetilde{Y}^i_{(i)} = [y_1, ..., y_{m_{-i}}] \stackrel{d}{=} \Lambda_i Z^i + D_i E^i \quad \text{and} \quad \{y_1, ..., y_{m_{-i}}\} \sim \text{i.i.d. normal}(0, \Lambda_i \Lambda_i^T + D_i^2), \tag{2.4}$$

where $m_{-i} = \prod_{j\neq i} m_j$, and Z^i and E^i are $k_i \times m_{-i}$ and $m_i \times m_{-i}$, respectively, with the same distributional properties as Z and E in (2.2). The superscript i on $\widetilde{Y}^i_{(i)}$ indicates the ith mode has not been standardized and the subscript (i) indicates the array has been matricized along the ith mode. The representation in (2.4) suggests the parameters $\{\Lambda_i, D_i\}$ can be viewed as single mode factor analysis parameters for the ith mode of the array when the covariance in all other modes has been removed.

Model indeterminacies

SFA as parameterized in (2.1) has two indeterminacies, one of which is common to all factor models and one that is common to all array normal models. The first indeterminacy,

which is also present in single mode factor analysis, is the orientation of the Λ matrices. The array covariance matrix in (2.1) is the same with mode i factor analytic parameters $\{\Lambda_i, D_i\}$ as it is with parameters $\{\Lambda_i G_i, D_i\}$, where G_i is any $k_i \times k_i$ orthogonal matrix. The second indeterminacy concerns the scales of the mode covariance matrices and stems from the model's separable covariance structure. The scale of a mode's covariance matrix can be moved to another mode covariance matrix or split among multiple modes' covariance matrices without changing the model. For example, the transformation $\{\Sigma_i, \Sigma_j\} \mapsto \{c\Sigma_i, \Sigma_j/c\}$ does not affect the array covariance matrix for any $c > 0$. This scale non-identifability is eliminated if all mode covariance matrices are restricted to have trace equal to one and a scale parameter is included for the total variance of the array.

Pseudo-true parameters

In single mode factor analysis the goal is to represent the covariance among a large set of variables in terms of a small number of latent factors. However, often it is unlikely the true covariance matrix has factor analytic structure. Consider a $p \times n$ matrix $X = [x_1, ..., x_n]$ and suppose $\{x_1, ..., x_n\} \sim$ i.i.d. normal$(0, \Sigma)$. There is interest in what k-factor analytic parameter values, Λ and D, best approximate the true covariance matrix Σ. These optimal parameter values, denoted $\overline{\Lambda}(\Sigma)$ and $\overline{D}(\Sigma)$, are those that minimize the Kullback-Leibler (KL) divergence between the factor model and the multivariate normal model. Minimizing the KL divergence is equivalent to maximizing the expected value of the k-factor analysis (FA) probability density with respect to the true multivariate normal (MN) distribution. Thus, $\overline{\Lambda}(\Sigma)$ and $\overline{D}(\Sigma)$ can be defined as

$$\begin{aligned}\{\overline{\Lambda}(\Sigma), \overline{D}(\Sigma)\} &:= \underset{\Lambda, D}{\operatorname{argmax}}\, \mathrm{E}_{\mathrm{MN}}[p_{\mathrm{FA}}(X|\Lambda, D)] \\ &= \underset{\Lambda, D}{\operatorname{argmax}}\; c_{FA} - \frac{n}{2}\log(|\Lambda\Lambda^T + D^2|) - \frac{n}{2}\mathrm{tr}[(\Lambda\Lambda^T + D^2)^{-1}\Sigma]\end{aligned}$$

where "tr" represents the trace operator and c_{FA} is a constant not depending on Λ or D. The diagonal matrix D that best approximates the true covariance matrix in the case of $k = 0$ is given by $\overline{D}(\Sigma) = \mathrm{diag}(\Sigma)^{1/2}$, where "diag" is the operator that replaces all off-diagonal entries with zero.

Similar to single mode factor analysis, SFA is an approximation to a separable covariance structure and modes' true covariance matrices are unlikely to have factor analytic structure. Suppose the distribution of Y is array normal with mean zero and covariance matrices $\widetilde{\boldsymbol{\Sigma}} = \{\widetilde{\Sigma}_i : 1 \leq i \leq K\}$. Consider a $(k_1, ..., k_K)$ SFA model for Y with parameters $\boldsymbol{\Lambda} = \{\Lambda_i : 0 < k_i < m_i\}$, $\boldsymbol{D} = \{D_i : 0 \leq k_i < m_i\}$, and $\boldsymbol{\Sigma} = \{\Sigma_j : k_j = m_j\}$. The expected value of the SFA probability density with respect to the true array normal (AN) model is

$$\mathrm{E}_{\mathrm{AN}}[p_{\mathrm{SFA}}(Y|\boldsymbol{\Sigma}, \boldsymbol{D}, \boldsymbol{\Lambda})] = c_{SFA} - \sum_{i=1}^{K} \frac{m}{2m_i} \log(|\Sigma_i|) - \frac{1}{2} \prod_{i=1}^{K} \mathrm{tr}[\Sigma_i^{-1} \widetilde{\Sigma}_i], \tag{2.5}$$

$$\text{where } \Sigma_i = \Lambda_i \Lambda_i^T + D_i^2 \text{ for } 0 \leq k_i < m_i,$$

c_{SFA} is a constant independent of the SFA parameters, and $m = \prod_{i=1}^{K} m_i$. Let $\overline{\overline{\Lambda}}(\widetilde{\boldsymbol{\Sigma}})$, $\overline{\overline{D}}(\widetilde{\boldsymbol{\Sigma}})$, and $\overline{\overline{\Sigma}}(\widetilde{\boldsymbol{\Sigma}})$ denote the SFA parameters that maximize (2.5) and, hence, provide the best approximation to the true separable covariance matrix. It can be shown that for all appropriate i, j, and k

$$\overline{\overline{\Lambda}}_i(\widetilde{\boldsymbol{\Sigma}}) = \overline{\Lambda}(\widetilde{\Sigma}_i), \qquad \overline{\overline{D}}_j(\widetilde{\boldsymbol{\Sigma}}) = \overline{D}(\widetilde{\Sigma}_j), \qquad \text{and} \qquad \overline{\overline{\Sigma}}_k(\widetilde{\boldsymbol{\Sigma}}) = \widetilde{\Sigma}_k. \tag{2.6}$$

This means that the best factor analytic parameters for a given mode in the SFA model are the closest fitting single mode factor analytic parameters for that mode's true covariance matrix. Furthermore, as we might expect, the optimal values of the unstructured covariance matrices in the SFA model are the modes' true covariance matrices. This implies that when the true model is array normal, the optimal SFA parameters for a given mode do not depend on the specified covariance structures in the other modes. Note that the scale indeterminacy of the covariance matrices is still present here. Thus, there is a set of optimal SFA parameter values that provide the same approximation and can be derived from one another by reallocating the covariance matrices' scales. Asymptotically, as the number of replicates of the array increases, these optimal SFA parameter values are the limiting values of the SFA maximum likelihood estimates (White (1982)).

2.3 Estimation and testing

In this section we consider parameter estimation for the SFA model and propose a likelihood ratio testing procedure for selecting the ranks $(k_1, ..., k_K)$. Two estimation methods are

described here: an iterative algorithm for maximum likelihood estimation and a Metropolis-Hastings algorithm which approximates the posterior distribution of the parameters given the data. For notational convenience we present the case where the array has mean zero, however both estimation methods and the testing procedure can be extended to allow for a mean structure. Examples of such extensions are discussed in Section 2.4 for the mortality data.

2.3.1 Maximum likelihood estimation

Simultaneous maximization of the SFA log likelihood with respect to all parameters is difficult. However, the maximization steps are manageable if done separately for each mode's covariance parameters. We propose an iterative algorithm that at each step maximizes the SFA log likelihood over a single mode's covariance parameters using the latest values of all other modes' parameters. This algorithm can be viewed as a form of block coordinate ascent and is guaranteed to increase the log likelihood at each step.

Let $\mathbf{\Lambda} = \{\Lambda_i : 0 < k_i < m_i\}$, $\boldsymbol{D} = \{D_i : 0 \leq k_i < m_i\}$, and $\mathbf{\Sigma} = \{\Sigma_j : k_j = m_j\}$ as in Section 2.2.2. Also let $\mathbf{\Lambda_{-j}} = \mathbf{\Lambda}/\{\Lambda_j\}$ be the set $\mathbf{\Lambda}$ with Λ_j removed, and define $\boldsymbol{D_{-j}}$ and $\mathbf{\Sigma_{-i}}$ analogously. The iterative maximum likelihood algorithm proceeds as follows.

0. Specify initial values for all covariance parameters $\{\mathbf{\Lambda}, \boldsymbol{D}, \mathbf{\Sigma}\}$.
1. For each mode $\{i : k_i = 0\}$, update the estimate of D_i.
2. For each mode $\{i : 0 < k_i < m_i\}$, update the estimates of Λ_i and D_i.
3. For each mode $\{i : k_i = m_i\}$, update the estimate of Σ_i.
4. Repeat steps 1-3 until a desired level of convergence is obtained.

The maximization in steps 1 and 3 over a diagonal matrix and an unstructured covariance matrix, respectively, are straightforward. The SFA log likelihood as a function of D_i, treating all other parameters as fixed, is written

$$\ell(D_i|\mathbf{\Sigma}, \mathbf{\Lambda}, \boldsymbol{D_{-i}}, Y) = b_i - \frac{m}{2m_i}\log(|D_i^2|) - \frac{1}{2}\mathrm{tr}[\widetilde{Y}^i_{(i)}(\widetilde{Y}^i_{(i)})^T D_i^{-2}] \tag{2.7}$$

where b_i is a constant independent of D_i and $m = \prod_i^K m_i$. The maximizing value of D_i and update for step 1 is thus

$$D_i^2 = \text{diag}\left(\frac{m_i}{m}\widetilde{Y}^i_{(i)}(\widetilde{Y}^i_{(i)})^T\right)$$

where the covariance matrices used to standardize Y in $\widetilde{Y}^i$ are the latest covariance matrix estimates. The SFA log likelihood as a function of an unstructured covariance matrix Σ_i is given by (2.7) where D_i^2 is replaced by Σ_i. Hence, the value of Σ_i that maximizes the corresponding log likelihood and the update for step 3 is

$$\Sigma_i = \frac{m_i}{m}\widetilde{Y}^i_{(i)}(\widetilde{Y}^i_{(i)})^T.$$

Estimation of a mode's factor analytic parameters in step 2 is more difficult, but can be accomplished using methods developed for single mode factor analysis. The SFA log likelihood as a function of the ith mode's factor analytic parameters, treating all other modes' parameters as fixed, is

$$\ell(\Lambda_i, D_i|\boldsymbol{\Sigma}, \boldsymbol{\Lambda_{-i}}, \boldsymbol{D_{-i}}, Y) = c_i - \frac{m}{2m_i}\log(|\Lambda_i\Lambda_i^T + D_i^2|) - \frac{1}{2}\text{tr}[(\Lambda_i\Lambda_i^T + D_i^2)^{-1}\widetilde{Y}^i_{(i)}(\widetilde{Y}^i_{(i)})^T] \quad (2.8)$$

where c_i is a constant not depending on Λ_i or D_i. The log likelihood for a single mode k_i-factor model for a $p \times n$ matrix X is written

$$\ell(\Lambda, D|X) = c - \frac{n}{2}\log(|\Lambda\Lambda^T + D^2|) - \frac{1}{2}\text{tr}[(\Lambda\Lambda^T + D^2)^{-1}XX^T]. \quad (2.9)$$

Notice that the SFA log likelihood has the the same form as that for single mode factor analysis where XX^T and n are replaced by $\widetilde{Y}^i_{(i)}(\widetilde{Y}^i_{(i)})^T$ and m/m_i, respectively. Therefore, we can use existing estimation methods for single mode factor analysis to update Λ_i and D_i.

Numerous iterative algorithms have been developed to obtain the single mode factor model maximum likelihood estimates; however many suffer from poor convergence behavior (Lawley (1940), Jöreskog (1967), Jennrich and Bobinson (1969)). An expectation-maximization (EM) algorithm was developed based on the model representation in (2.2) that treats Z as latent variables (Dempster et al. (1977), Rubin and Thayer (1982)). The slow convergence of this algorithm led to expectation/conditional maximization either (ECME) algorithms, some of which rely on numerical optimization procedures (Liu and Rubin (1998),

Zhao et al. (2008)). Zhao et al. (2008) proposed an iterative algorithm that updates Λ, treating D as known, and then sequentially updates each diagonal element of D, treating Λ and all other elements of D as known. This algorithm has closed form expressions for all parameter updates and was shown to outperform the EM algorithm and its extensions in terms of convergence and computation time. For these reasons, we chose to use it for step 2 of the SFA estimation procedure.

Divergence of the SFA maximum likelihood algorithm, where the log likelihood continually grows at a nondecreasing rate, is evidence that the maximum likelihood estimates do not exist. While the update in step 1 for a mode with a diagonal covariance matrix is always well defined (i.e. the SFA log likelihood in (2.7) has a maximum in terms of D_i), step 2 of the algorithm for an unstructured covariance matrix is only well defined if $m_i < \prod_{j \neq i} m_j$. Similarly, step 3 is well defined for a mode i with a factor analytic covariance matrix if $k_i < \text{rank}(\widetilde{Y}^i_{(i)}(\widetilde{Y}^i_{(i)})^T)$. This latter requirement is effectively equivalent to $k_i < \min(m_i, \prod_{j \neq i} m_j)$ since $\widetilde{Y}^i_{(i)}$ is unlikely to be rank deficient for a continuous array Y.

2.3.2 Bayesian estimation

Mortality information is limited for many undeveloped countries that do not have reliable death registration data. Thus, it is not uncommon to be missing a country's death rates for specific ages or at all ages in a given year. A Bayesian approach to parameter estimation can easily accommodate missing data and provide predictive distributions of the missing values. Let Y_o denote the portions of the array Y that are observed and Y_m represent those values that are missing. Inference for the parameters and the missing data can be based on the joint posterior distribution of the parameters and missing data given the observed data, $p(\boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma}, Y_m | Y_o)$. This posterior distribution is written $p(\boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma}, Y_m | Y_o) \propto p(Y | \boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma}) p(\boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma})$, where $p(Y | \boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma})$ is the density of the $(k_1, ..., k_K)$ SFA model for $Y = \{Y_o, Y_m\}$ and $p(\boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma})$ is the joint prior distribution of the parameters. In the case of no missing data, one can consider Y_m to be the empty set and Y_o to equal Y.

Prior specification

In the absence of real prior information, we suggest a convenience prior composed of semi-conjugate distributions for the parameters, which also reflects some of the indeterminacies in the model. For an SFA model in which mode i's covariance matrix is unstructured, the prior distribution for Σ_i^{-1} is Wishart$(\kappa_i, \mathrm{I}_{m_i})$ with hyperparameter κ_i, where $\kappa_i \geq m_i$. For a mode i with a factor analytic covariance matrix, the joint prior distribution of $\{\Lambda_i, D_i\}$ is specified by the marginal distribution of D_i and the conditional distribution of Λ_i given D_i, as follows:

$$\{\mathrm{vec}(\Lambda_i)|D_i\} \sim \mathrm{normal}(0, \mathrm{I}_{k_i} \otimes D_i^2), \tag{2.10}$$

$$\{D_i^{-2}[1,1], ..., D_i^{-2}[m_i, m_i]\} \sim \text{i.i.d. gamma}(\nu_0/2, \text{ rate} = \nu_0 d_0^2/2), \quad \nu_0 > 0, \ d_0^2 > 0. \tag{2.11}$$

A priori each mode's parameters are modeled as independent of all other modes' parameters given the hyperparameters ν_0, d_0^2, and $\{\kappa_i : k_i = m_i\}$. Thus, the joint prior distribution $p(\boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma})$ is equal to the product of the marginal distributions of each mode's parameters.

The prior distribution of the factor analytic parameters given in (2.10) and (2.11) has nice properties related to the rotational indeterminacies in the $\boldsymbol{\Lambda}$ matrices. Recall that the SFA likelihood is invariant to rotation of Λ_i, meaning $\mathrm{L}_{\mathrm{SFA}}(\Lambda_i, D_i, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}_{-i}, \boldsymbol{D}_{-i}|Y) = \mathrm{L}_{\mathrm{SFA}}(\Lambda_i G_i, D_i, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}_{-i}, \boldsymbol{D}_{-i}|Y)$ where $\mathrm{L}_{\mathrm{SFA}}$ is the SFA likelihood and G_i is any $k_i \times k_i$ orthogonal matrix. If the joint prior distribution $p(\Lambda_i, D_i)$ is integrated over the diagonal elements of D_i, the marginal distribution of Λ_i is obtained and can be expressed as

$$p(\Lambda_i) \propto \prod_{j=1}^{m_i} \left[\nu_0 d_0^2 + ||\Lambda_i[j,]||^2\right]^{(k_i+\nu_0)/2},$$

where $||\cdot||^2$ denotes the Frobenius norm. Observe that $p(\Lambda_i) = p(\Lambda_i G_i)$ implying that the prior distribution is also invariant to rotations of Λ_i. This is a desirable property as it indicates the prior does not favor one set of parameters over another if they are equivalent given the data (i.e. have the same SFA likelihood). Namely, all parameter values $\{\Lambda_i G_i : G_i^T G_i = G_i G_i^T = I\}$ are given equal probability in the prior.

Metropolis-Hastings algorithm

The posterior distribution $p(\boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma}, Y_m|Y_o)$ is not a standard distribution and is difficult to sample from directly, so a Metropolis-Hastings algorithm is used to obtain a Monte Carlo approximation of it. The Metropolis-Hastings algorithm produces a Markov chain of $\{\boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma}, Y_m\}$, whose stationary distribution is equal to $p(\boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma}, Y_m|Y_o)$. The algorithm proceeds by iteratively proposing new values of the missing data and each mode's parameters, and accepting or rejecting the proposals based on an acceptance probability. The algorithm can be described as follows:

0. Specify initial values for all covariance parameters $\{\boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma}\}$ and missing data Y_m.
1. For each mode $\{i : k_i = 0\}$, update D_i.
2. For each mode $\{i : 0 < k_i < m_i\}$, update Λ_i and D_i.
3. For each mode $\{i : k_i = m_i\}$, update Σ_i.
4. If elements of Y are missing, update Y_m.
5. Repeat steps 1-4 until a sufficiently accurate approximation of the posterior distribution is obtained.

The update of D_i in step 1 is straightforward since the full conditional distribution of the entires in D_i^{-2} given the data and all other parameters is a product of the following independent gamma distributions:

$$\{D_i^{-2}[j,j]|\boldsymbol{D_{-i}}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, Y\} \sim \text{gamma}\left((\nu_0 + m/m_i)/2, \text{rate} = \left(\nu_0 d_0^2 + S_i[j,j]\right)/2\right) \quad (2.12)$$

for $j \in \{1, ..., m_i\}$ where $S_i = \widetilde{Y}^i_{(i)}(\widetilde{Y}^i_{(i)})^T$. The Metropolis-Hastings acceptance probability is equal to one when sampling from parameters' full conditional distribution. Thus, the update in step 1 is performed by setting the new value of D_i equal to a sample from (2.12), where Y is comprised of Y_o and the most current update of Y_m, and the covariance matrices used to standardize Y in $\widetilde{Y}^i$ are the most current parameter updates.

The updates for the factor analytic parameters $\{\Lambda_i, D_i\}$ in step 2 are based on the latent variable representation of SFA introduced in (2.4), which is that $\widetilde{Y}^i_{(i)} \stackrel{d}{=} \Lambda_i Z^i + D_i E^i$ where the elements of Z^i and E^i are independent standard normal random variables. Conditioning on Z^i, the mode i k_i-factor model can be written

$$\{\text{vec}(\widetilde{Y}^i_{(i)})|Z^i, \Lambda_i, D_i\} \sim \text{normal}(\text{vec}(\Lambda_i Z^i), \text{I}_{m/m_i} \otimes D_i^2).$$

Using this representation of the model, we developed the following Metropolis-Hastings updates:

A. Update Λ_i as follows.

(i) Sample $\{\text{vec}(Z^i)|\Lambda_i, D_i, \widetilde{Y}^i\} \sim \text{normal}(\text{vec}(\phi\Lambda_i^T D_i^{-2}\widetilde{Y}^i_{(i)}), \text{I}_{m/m_i} \otimes \phi)$
where $\phi = (\Lambda_i^T D_i^{-2}\Lambda_i + \text{I})^{-1}$.

(ii) Sample

$$\{\text{vec}(\Lambda_i)|Z^i, Y, \boldsymbol{D}, \boldsymbol{\Lambda_{-i}}, \boldsymbol{\Sigma}\} \sim \text{normal}(\gamma(Z^i_{(i)} \otimes D_i^{-2})\text{vec}(\widetilde{Y}^i_{(i)}), \gamma)$$

where $\gamma = [(Z^i_{(i)}(Z^i_{(i)})^T + \text{I}_{m_i}) \otimes D_i^{-2}]^{-1}$.

B. Update D_i as follows.

(i) Sample $\{\text{vec}(Z^i)|\Lambda_i, D_i, \widetilde{Y}^i\}$ as in A(i).

(ii) Sample the elements of D_i^2 independently according to

$$\{D_i^{-2}[j,j]|Z^i, Y, \boldsymbol{D_{-i}}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}\} \sim \text{gamma}\left(\frac{\nu_m}{2}, \text{rate} = \frac{\left(\nu_0 d_0^2 + J[j,j] + ||\Lambda_i[j,]||^2\right)}{2}\right)$$

where $\nu_m = \nu_0 + m/m_i + k_i$, $J = (\widetilde{Y}^i_{(i)} - \Lambda_i Z^i)(\widetilde{Y}^i_{(i)} - \Lambda_i Z^i)^T$, and $||\cdot||^2$ denotes the Frobenius norm.

These updates for Λ_i and D_i are Metropolis-Hastings proposals with acceptance probabilities equal to one (see Appendix A.1).

Similar to step 1, the update of Σ_i in step 3 can be performed by sampling from the full conditional distribution of Σ_i^{-1} given the data and all other parameters:

$$\{\Sigma_i^{-1}|\boldsymbol{D}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma_{-i}}, Y\} \sim \text{Wishart}(\kappa_i + m/m_i, (\text{I}_{m_i} + \widetilde{Y}^i_{(i)}(\widetilde{Y}^i_{(i)})^T)^{-1}). \tag{2.13}$$

This proposal also has acceptance probability equal to one, and as in step 1 and 2, the latest updates of all other modes' parameters and the missing data are used in the calculation of $\widetilde{Y}^i$.

Step 4 of the algorithm that updates the missing data can be done in one update or as a sequence of updates for any partition of Y_m. All elements of Y_m can be updated together

by sampling from the conditional distribution of vec(Y_m) given the covariance parameters and observed data. Although the conditional distribution is normal, such an update can be expensive due to the large matrices often involved in computing the distribution's covariance matrix. Updating the missing data in partitions of the array known as slices, where one mode index is fixed, avoids the need to work with such large matrices. For the mortality data, examples of slices include the data for country c for all ages, sexes, and years or the data at age a for all sexes, years, and countries. Hoff (2011) shows that for a separable covariance structure the conditional distribution for a slice of an array given the rest of the array can be written as array normal distribution. If the missing data in the slice is then conditioned on the observed data in the slice, a multivariate normal distribution is obtained that can be used to update the missing data. Calculating the conditional distribution of the missing elements in a slice of the array via this two-step conditioning procedure (once for the slice and once for the missing data within the slice) circumvents computation with unnecessarily large matrices. As in the update of Σ_i, the normal distributions mentioned here represent the full conditional distributions of Y_m and a subset of Y_m so the acceptance probabilities equal one. Note that although updating subsets of Y_m may be easier computationally, it has the potential to make the Markov chain less efficient and increase the number of samples needed to obtain an accurate approximation of the posterior distribution.

Unlike in the frequentist setting where divergence of the maximum likelihood estimation procedure indicates a lack of information in the data about the parameters, the posterior distribution of the parameters given the data will always exist. Although Bayesian parameter estimates are available, we should be aware of what information the estimates reflect. The posterior distribution of the parameters given the data is combination of the information in the prior distribution and the information in the data. Extreme similarity between the prior distribution and the posterior distribution suggests that little information is gained from the data and inference based on the posterior distribution is primarily a reflection of the information in the prior.

Hyperparameters

When there is little prior information about the parameters, it is common to choose hyperparameter values that result in diffuse prior distributions. We propose $\nu_0 = 3$ and $\kappa_i = m_i + 2$ for $\{i : k_i = m_i\}$ as default values for the SFA model. These values ensure that the first moments of the prior distributions are finite and represent some of the most diffuse distributions in the Wishart and gamma families, respectively. They also have the following properties.

$$\mathrm{E}[\Sigma_i] = \mathrm{I}_{m_i} \qquad \mathrm{E}[D_i^2[j,j]] = 3d_0^2 \qquad \mathrm{E}[\mathrm{tr}(\Lambda_i \Lambda_i^T)] = 3k_i m_i d_0^2 \tag{2.14}$$

Prior information about specific mode covariance matrices may be limited, however an estimate $\widehat{\psi}$ of the total variance of the array, $\psi = \mathrm{tr}(\mathrm{Cov}[\mathrm{vec}(Y)]) = \prod_{i=1}^{K} \mathrm{tr}(\Sigma_i)$, may be available. This information can improve parameter estimation by centering the prior distribution of the total variance of the array around a reasonable value. Based on the expectations in (2.14) and the independence of the mode covariance matrices in the prior, the prior expected value of the total variance of the array will equal the estimate, $\mathrm{E}[\mathrm{tr}(\mathrm{Cov}[\mathrm{vec}(Y)])] = \widehat{\psi}$, if

$$d_0^2 = \widehat{\psi}^{1/R} \left[\left(\prod_{j:0<k_j<m_j} [k_j + 1] \right) \left(\prod_{i=1}^{K} m_i \right) 3^R \right]^{-1/R} \tag{2.15}$$

where $R = \sum_{i=1}^{K} \mathbb{1}\{0 \leq k_i < m_i\}$ is the number of modes with factor analytic covariance structure. In the event there is no prior knowledge about ψ and it is not of interest in the analysis, we propose taking an empirical Bayes approach and obtaining an estimate of it based on the data. Possible estimates include $\widehat{\psi} = \frac{m}{m_o}||Y_o||^2$ or $\widehat{\psi} = \frac{m}{m_o}||Y_o - \widehat{M}_o(X,\beta)||^2$ if the model has a non-zero mean, where m_o denotes the number of observed entries in Y. In the latter case, $\widehat{M}_o(X,\beta)$ represents an initial estimate of the mean for the observed data, such as the ordinary least squares estimate. Specification of d_0^2, ν_0, and $\{\kappa_i : k_i = m_i\}$ as described here weakly centers the prior distribution for the total variation in Y around the estimate $\widehat{\psi}$.

2.3.3 Testing for the mode ranks

It is often difficult to choose the number of factors for a single mode factor model. This problem is only more pronounced in the array case where the rank k_i must be specified for

each mode. As is done in single mode factor analysis (Mardia et al. (1979)), a likelihood ratio test can be constructed to test between nested SFA models with ranks $(k_1, ..., k_K)$ and $(k_1^*, ..., k_K^*)$, where $k_i \leq k_i^*$ for all i. However, due to the large number of possible combinations of ranks, choosing the ranks using these likelihood ratio tests is challenging. Here we propose an alternative mode-by-mode rank selection procedure that suggests when the rank specified for a given mode is sufficient for capturing the dependence within that mode.

Suppose a K-way array Y is normally distributed, and define $\widetilde{Y}$ to be the array obtained when Y is standardized by its covariance matrix $\Sigma = \text{Cov}[\text{vec}(Y)]$: $\text{vec}(\widetilde{Y}) := \text{vec}(Y)\Sigma^{-1/2}$. The elements of $\widetilde{Y}$ represent independent standard normal random variables. Thus, to determine whether the covariance in mode i is captured by a proposed $(k_1, ..., k_K)$ SFA model, we can compute $\widetilde{Y}$ using the SFA mode covariance matrix estimates as in (2.1) and test whether the covariance matrix of the rows of $\widetilde{Y}_{(i)}$ equals the identity. The likelihood ratio test statistic for this test is

$$t = \frac{m}{m_i}[\text{tr}(\widehat{V}) - \log|\widehat{V}| - m_i], \tag{2.16}$$

where $\widehat{V} = \frac{m_i}{m}\widetilde{Y}_{(i)}\widetilde{Y}_{(i)}^T$, and has an asymptotic $\chi^2_{m_i(m_i+1)/2}$ distribution, as the number of replicates of the array (or equivalently as the dimension of a mode with identity covariance matrix) goes to infinity, under the null hypothesis of an identity row covariance matrix. Note that rejecting this test suggests that a more complex covariance structure is needed for the mode. We propose the following rank selection procedure based on these mode specific tests.

0. Consider an SFA model with all $k_i = 0$. Obtain estimates of the covariance parameters using the maximum likelihood procedure in Section 2.3.1 and compute $\widetilde{Y}$ using the estimates.

1. For each mode i, define $R_i = \text{Cov}[\text{vec}(\widetilde{Y}_{(i)})]$ and test $H_0 : R_i = \text{I}_{m/m_i} \otimes \text{I}_{m_i}$ vs $H_1 : R_i = \text{I}_{m/m_i} \otimes V$, where V is an unstructured $m_i \times m_i$ covariance matrix, using a likelihood ratio test with test statistic given by (2.16).

2. If the test for mode i rejects and $\begin{cases} \delta(m_i, k_i+1) > 0, \text{ increase the rank } k_i \text{ by one.} \\ \delta(m_i, k_i+1) \leq 0, \text{ set the rank equal to } m_i. \end{cases}$

   If the test for mode i does not reject, fix k_i at its current value and perform no further tests on the mode. Obtain maximum likelihood estimates $\{\widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\Lambda}}, \widehat{\boldsymbol{D}}\}$ for an SFA model with the new ranks $(k_1, ..., k_K)$ and compute $\widetilde{Y}$ using these new estimates.

3. Repeat steps 1-2 until each mode has failed to reject a test.

The suggested ranks $(k_1, .., k_K)$ are those that result at the end of this procedure. Recall that $\delta(m, k) = \left[(m-k)^2 - (m+k)\right]/2$ represents the reduction in the number of parameters when using a k-factor analytic covariance matrix instead of an $m \times m$ unstructured covariance matrix. The rank increases in step 2 reflect that an unstructured covariance matrix is specified when a factor analytic covariance structure no longer provides a reduction in the number of covariance parameters.

The maximum number of SFA models that could be considered using this procedure is bounded by largest value of k_l such that $\delta(m_l, k_l) > 0$, where l denotes the array mode with the largest dimension m_l. To control the type I error rate of all mode tests to be α for an iteration of steps 1 and 2, the level of each mode test can be set to α^r where r is the number of modes being tested (i.e. the number that have rejected every test thus far). An example of this procedure is described in Section 2.4.2 for the mortality data.

2.4 Application to Human Mortality Database death rates

In this section we analyze death rates from the Human Mortality Database (HMD) (University of California, Berkeley and Max Planck Institute for Demographic Research, 2009) using an SFA model, compare our model to other covariance models, and obtain predictions for over four hundred missing death rates. We focus on death rates for 5-year time periods for populations corresponding to combinations of sex, age, and country of residence. Specifically, we consider death rates from 1960 to 2005 for 40 countries, both sexes and twenty-three age groups, $\{0, 1-4, 5-9, 10-14, ..., 105+\}$. These data are represented in a 4-way array $Y = \{y_{ctsa}\}$ of dimension $(40 \times 9 \times 2 \times 23)$, where y_{ctsa} is the log death rate

for country c, time period t, sex s and age group a. We will refer to a set of age-specific death rates for a combination of country, time period, and sex as a mortality curve.

We begin this section by introducing a flexible piecewise polynomial mean model and show the residuals from this mean model exhibit dependence within each mode: age, time period, country, and sex. Using the likelihood ratio testing procedure presented in Section 2.3.3, we select ranks for an SFA model. The resulting SFA model is compared to models with simpler covariance structures using out-of-sample cross validation and is used to impute multiple years of missing death rates for Chile and Taiwan.

2.4.1 Mean model selection

As discussed in the Introduction, existing methods for analyzing mortality data model the death rates for different countries, sexes, and/or time periods separately. Such an approach can be inefficient due to the strong similarities between mortality rates within the same country, time period, or sex. For this reason, we propose a new joint mean model for the HMD data that acknowledges the relationships between mortality rates that share levels of one or more of these factors.

Figure 2.1 shows mortality curves defined by the twenty-three age-specific death rates for the United States and Sweden in four time periods. The large spikes at age zero represent infant mortality, and the humps around age twenty, which are especially evident in males, are attributed to teenage and young adult accident mortality. The overall shapes of the mortality curves for each sex are similar across countries and time periods, however Sweden has considerably lower mortality levels during childhood and young adulthood compared to the United States. This suggests that a mean model for the data should allow for different curves across countries and time periods, yet still take advantage of the similarity between death rates within the same country, age group, or sex.

Drawing from the mortality literature and viewing mortality rates as function of age,

we propose the following piecewise polynomial (PP) mean model:

$$\mathrm{E}[y_{cysa}] = \begin{cases} \phi^0 & : a = 0 \\ \phi^1 + a\phi^{11} + a^2\phi^{12} & : 1 \leq a < 20 \\ \phi^2 + a\phi^{21} + a^2\phi^{22} + a^3\phi^{23} & : 20 \leq a \end{cases} \qquad \phi^i = \alpha_c^i + \beta_t^i + \gamma_s^i. \quad (2.17)$$

This model distinguishes between the infant, childhood, and adult stages of mortality by fitting each with a separate polynomial, whose coefficients are composed of additive effects for country, time period, and sex. The constant term at age zero is necessary to model the steep decline from infant mortality to child mortality that is not well represented by a low degree polynomial. Parameter estimates for this model based on the data array can be obtained by minimizing the ordinary the least squares (OLS) criterion $\sum_c \sum_t \sum_s \sum_a [y_{ctga} - \mathrm{E}[y_{ctga}]]^2$, and since the model is linear in its parameters, the OLS estimates can be solved for algebraically.

One of the most commonly used models in demography for age-specific mortality measures is the Heligman-Pollard (HP) model (Heligman and Pollard (1980)). This model also uses eight parameters to parameterize a mortality curve, however it is typically used to model each mortality curve individually and is nonlinear and non-convex in the parameters making estimation extremely difficult (Hartmann (1987), Congdon (1993)). When the HP model is fit separately to the 684 HMD mortality curves for the 38 countries missing no death rates using OLS, it requires over 5,400 parameters and under the assumption of independent, homoscedastic errors has a Bayesian Information Criterion (BIC) value of $-17,288$. However, when the PP model is fit jointly to the same data, it contains 376 parameters and has a BIC of $-52,436$. Due to the relative parsimony of the PP model, its superior fit in terms of BIC, and its straightforward estimation, it was selected as the mean model.

2.4.2 Excess dependence and SFA rank selection

The piecewise polynomial model in (2.17) is extremely flexible. To investigate its fit to the HMD mortality rates, we focused on a subset of the original data that contains no missing observations, specifically the $(38 \times 9 \times 2 \times 23)$ array that does not have death rates for

Chile or Taiwan. The OLS fit this data explains 99.5% of the variation in mortality rates (coefficient of determination, $R^2 = 0.995$). However, there is interest in whether excess correlation exists in the residuals since modeling it can improve both predictions of missing values and the efficiency of parameter estimates. Ordinary least squares estimates of the parameters in (2.17) are equivalent to maximum likelihood estimates assuming independent normal errors. To evaluate this latter assumption, we computed the empirical correlations between the mean model residuals for countries, time periods, and age groups by matricizing the residual array with respect to each mode and computing a sample correlation matrix for the mode.

As mentioned in the Introduction, the distributions of these correlations have substantially more large positive values than would be expected under the assumption of independent errors. For example, speaking specifically to the temporal dependence, the average correlation between adjacent time periods, those one time period apart, and those two periods apart is 0.79, 0.54, and 0.26, respectively. The first two principal components of each correlation matrix are shown in Figure 2.2. The horseshoe pattern in the time period principal components and the clustering of countries within the same region suggests temporal and geographic trends in the data are not captured by the mean (Diaconis et al. (2008)). This indicates that even though the mean model contains several country specific and time period specific parameters, similarities between the mortality curves of certain countries and time periods is not being accounted for. The mean model already contains over 370 parameters and it would likely be nontrivial to modify it to capture all of the dependence seen in the residuals. For this reason, we consider incorporating a covariance structure to model this excess dependence. An array normal separable covariance structure could be specified, however it would add over one thousand parameters to the model. Therefore, we instead consider an SFA model for the data with the PP mean with the belief that some of the residual mode covariance matrices may be well approximated by a low rank factor analytic structure.

As outlined in Section 2.3.3, suggestions for the SFA ranks can be obtained from a repeated likelihood ratio testing procedure. For the mortality data, we consider (k_c, k_t, k_s, k_a) SFA models where the ranks correspond to the country, time period, sex, and age covariance

matrices, respectively. The standardized residual array $\widetilde{Y}$ for a (k_c, k_t, k_s, k_a) SFA model is defined as $\text{vec}(\widetilde{Y}) = (\text{vec}(Y) - \text{vec}(\widehat{M}))(\widehat{\Sigma}_a^{-1/2} \otimes \widehat{\Sigma}_s^{-1/2} \otimes \widehat{\Sigma}_t^{-1/2} \otimes \widehat{\Sigma}_c^{-1/2})$, where $\widehat{M}$ represents the PP mean model maximum likelihood estimate and $\widehat{\Sigma}_i$ is the SFA mode i covariance matrix estimate. The results from the iterative testing procedure are shown in Table 2.1. The first step in this process is to consider a (0,0,0,0) SFA model where all covariance matrices are diagonal. The likelihood ratio test statistics for this model are shown in the first row of Table 2.1 and the corresponding 0.05 level critical values are shown in the last row. Since the test for each mode rejects the null hypothesis of independent, variance one errors, the rank of each mode is increased by one in the subsequent model, except for that for the sex mode. A rank one factor analytic structure for a (2×2) covariance matrix has more parameters than an unstructured covariance matrix so the sex covariance matrix is unstructured in the next model. A box around a test statistic in the table indicates the mode failed to reject the test for the first time. Recall that when a mode's test does not reject, the rank for that mode is fixed and not increased in later models. The table shows where the sex, time period, country, and age ranks become fixed at two, four, nine, and ten, respectively. Observe that after a mode's rank is fixed, the test statistic for that mode stays below the critical value in all subsequent models. Although the mode tests are not independent of the covariance structures fit in the other modes, this consistency supports the suggested ranks.

2.4.3 Out-of-sample cross validation

We evaluate the SFA model by comparing its out-of-sample predictive performance with two simpler covariance models that share the same PP mean model. The three covariance models considered are the following:

M1: Independent and identically distributed (i.i.d.) model

M2: Time covariance model

M3: SFA model (9,4,2,10)

M1 corresponds to the conventional ordinary least squares (OLS) approach where all errors are assumed independent and identically distributed with a common variance parameter. Based on the temporal nature of the data, a natural first step to incorporating a covariance

Table 2.1: Iterative testing procedure for the SFA ranks. Each row represents an SFA model and each entry is the likelihood ratio test statistic based on (2.16). The 0.05 level critical value for each test is given in the last row. A box around a statistic indicates that the mode does not reject the test for the first time and the rank is fixed in subsequent models.

SFA ranks	Likelihood ratio test statistic			
(k_c, k_t, k_s, k_a)	Country	Time period	Sex	Age
(0,0,0,0)	21,852	14,482	702	27,883
(1,1,2,1)	9,526	5,853	$\boxed{0}$	14,451
(2,2,2,2)	4,425	1,722	0	6,374
(3,3,2,3)	2,776	716	0	3,762
(4,4,2,4)	1,946	$\boxed{17}$	0	2,422
(5,4,2,5)	1,556	14	0	1,833
(6,4,2,6)	1,287	10	0	1,340
(7,4,2,7)	1,040	8	0	967
(8,4,2,8)	892	5	0	540
(9,4,2,9)	$\boxed{762}$	8	0	363
(9,4,2,10)	737	8	0	$\boxed{257}$
$\chi^2_{.95}$ critical value	805	62	8	316

model is to consider an unstructured covariance matrix for time as in M2. In general, country mortality rates are relatively stable over time so if the observed mortality for a given country, year, and age deviates from the mean model in one year, it is likely the observations deviate in the same direction in neighboring years.

Fifty cross validations were performed by removing a random 25% of the array, estimating each of the three models on the remaining data, and computing the mean squared error (MSE) between the observed values and the predicted values for the withheld entries. The predicted values for M1 are those from the OLS PP mean estimate. For M2 and M3, the predictions are the posterior mean estimates of the missing values from the Bayesian estimation procedure described in Section 2.3.2.

A prior distribution for the parameters in the PP model is needed to perform simultane-

ous Bayesian estimation for the mean and covariance parameters. The prior on the vector of PP coefficients is a mean zero normal distribution with covariance matrix $m(X^TX)^{-1}$, where X is the design matrix for the PP model for vec(Y) and $m = \prod_{i=1}^{K} m_i$. This is a relatively uninformative prior as it is over 30 times more diffuse than the corresponding unit-information prior (Kass and Wasserman (1995)). The hyperparameters were specified as described in Section 2.3.2 where the mean estimate $\widehat{M_o}$ used in $\widehat{\psi}$ is the OLS estimate of the PP model. Since M2 has no modes with factor analytic structure, the prior on the time covariance matrix is

$$\Sigma_t^{-1} \sim \text{Wishart}\left(n_t = m_t + 2, \frac{m\widehat{\psi}}{m_t} I_{m_t}\right).$$

This specification is necessary to preserve the property that E[tr(Cov[vec(Y)])] = $\widehat{\psi}$ under the prior.

The results from the 50 cross-validations are shown in Table 2.2. The MSE for the SFA model was less than that of the time covariance model for each of the 50 cross-validation runs, and the MSE for the time covariance model was always less than that of the i.i.d. model. In terms of average MSE, both the time covariance model and the SFA model significantly improve upon the i.i.d. model, and the SFA model out performs the time covariance model by nearly a factor of two. This is evidence that even with the extremely flexible PP mean model, the SFA covariance structure still improves model fit as it is able to estimate the similarity between mortality rates across countries, time periods, age groups, and sexes, and use this information in its predictions.

Table 2.2: Average and standard deviation of the mean squared errors from 50 out-of-sample cross-validation experiments.

	M1 (i.i.d.)	M2 (Time)	M3 (SFA)
Average mean squared error	0.02996	0.00729	0.00385
Standard deviation of mean squared errors	0.00084	0.00049	0.00034

2.4.4 Prediction of missing data

The imputation of missing death rates is an important application of modeling mortality data as information is often incomplete for countries lacking accurate death registration data. We now consider the original $(40 \times 9 \times 2 \times 23)$ array of mortality rates with observations for Chile and Taiwan. Seven time periods of mortality information are missing for Chile (1960-1995) and two time periods for Taiwan (1960-1970), combining for a total of 414 missing entries in the array. The maximum likelihood estimation algorithm in Section 2.3.1 cannot accommodate missing data so we are unable to reselect the SFA ranks using the testing procedure. However, this larger data array contains only two additional countries so the SFA ranks $(9, 4, 2, 10)$ selected for the reduced data are used here. Predictions for the missing death rates were based on samples from the Metropolis-Hastings procedure, for which the effective sample sizes for the Monte Carlo estimates of all missing values was greater than 500.

In the left column of Figure 2.3, posterior mean predicted death rates and 95% prediction intervals are shown for Chile in 1990 and Taiwan in 1965. To visualize the impact of the SFA covariance model on the predicted death rates, we investigate the difference between the SFA predicted values and the fitted values based on the PP mean model. The SFA predictions, $\hat{y}_p$, are conditional on the observed mortality rates for all other countries and time periods, while the mean model fitted values, $\hat{y}_m$, are based only on the estimate of the PP model. We call these differences, $\hat{y}_p - \hat{y}_m$, "predictive residuals" since they are based on predicted values instead of observed values. These differences illustrates the changes in the predicted values by using the estimated dependence between residuals within modes of the array and conditioning on the observed mortality rates. The empirical residuals based on the PP mean model, $y - \hat{y}_m$, were computed for the United States and Australia, the two countries most highly correlated with Chile (estimated correlations of around 0.40). These residuals were also computed for Japan and West Germany, the two countries most highly correlated with Taiwan (estimated correlations of around 0.13). The middle column of Figure 2.3 shows the predictive residuals for Chile and Taiwan and the empirical residuals for these select countries. The last column contains the empirical residuals in 1995 and 1970 when mortality

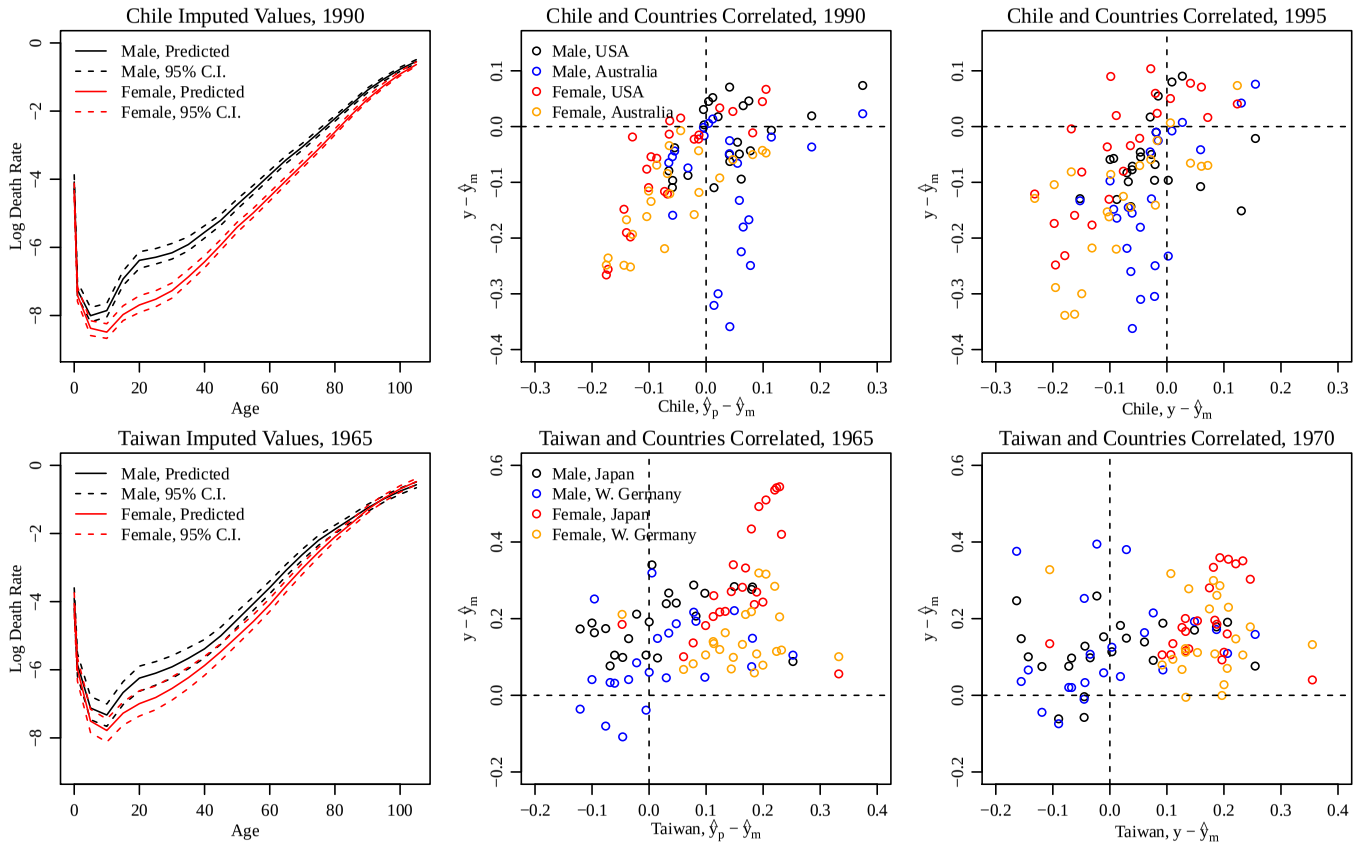

Figure 2.3: The first column of plots shows the predicted values and corresponding 95% prediction intervals for the missing death rates for Chile and Taiwan. The middle column shows the difference between the posterior mean predicted value and the piecewise polynomial mean function fitted value, $\hat{y}_p - \hat{y}_m$, for Chile and Taiwan, along with empirical mean model residuals, $y - \hat{y}_m$, for countries that are highly correlated with them in the posterior mean country covariance matrix. The last column contains empirical residuals for the following time period when Chile and Taiwan mortality is observed.

information is available for all countries. Observe that the plots in the middle column and last column are similar, demonstrating an overall positive association for both sexes and all country pairs. This illustrates how the model uses the relationship between the empirical residuals of Chile and other countries to predict Chile's deviations from the mean model in years when Chile data is missing. The ability to draw information across multiple country, year, and sex residuals to impute missing values is a critical strength of the SFA model that is not shared by other mortality models or simpler covariance structures.

The empirical residuals for Chile shown in the last column may not show as strong of an association with the United States and Australia as one would expect from a posterior mean correlation estimate of 0.4. However, recall that the estimate of the country correlations is based on all time periods, sexes, and ages. Although we show adjacent time periods in this plot, the correlation between the country residuals in the period adjacent to the missing time period and the correlations in time periods furthest away are weighted equally in the estimate of the country correlation, and hence weighted equally in the imputation of the missing data. For example, the correlation between Taiwan and Japan's empirical residuals in 2000 and that in 1970 influence Taiwan's imputations in 1965 equally. This property is a consequence of the separability of the SFA covariance matrix. A more complicated non-separable covariance model would be required for the correlations between countries, ages, and sexes to be differentially weighted in the imputation based on the proximity of the observed data to the missing data.

2.5 Discussion

In this chapter we introduced the separable factor analysis model for array-valued data. Unlike the array normal model where all mode covariance matrices are unstructured, SFA parameterizes mode covariance matrices by those with factor analytic structure. Using covariance matrices with reduced structure decreases the number of parameters in the model considerably and allows mode covariance matrices to be estimated using maximum likelihood methods for any array dimension. Including a covariance structure in a model for multiway data can drastically improve mean model parameter estimation and missing data predictions in situations where dependence exists within modes that is not captured by the mean model.

In an out-of-sample cross validation study with a large set of mortality data, the SFA model was shown to have superior fit compared to models with simpler covariance structures, even in the presence of an extremely flexible mean model. The SFA model was also shown to estimate which countries have similar deviations from the mean model and was able to use this information to predict multiple years of missing death rates.

An alternative extension of factor analysis to arrays was considered that resembles the higher order singular value decomposition (see the appendix for details). This model has a non-separable covariance structure and can be viewed a submodel of the single mode factor analysis model for $\text{vec}(Y)$. We chose the SFA model over this alternative extension due to the interpretability of its parameters as single mode factor model parameters and for its use as an approximation to a separable covariance model with an unstructured covariance matrix in each mode.

Chapter 3

TESTING AND MODELING DEPENDENCIES BETWEEN A NETWORK AND NODAL ATTRIBUTES

3.1 Introduction

A common goal in the analysis of network data is to characterize the dependence between network relations and a set of node-specific attributes. For example in recent years many studies in the social sciences have examined the relationship between individuals' friendship networks and their health measures, such as happiness (Fowler and Christakis (2008)), smoking and drinking behavior (Kiuru et al. (2010)), and obesity (Christakis and Fowler (2007), de la Haye et al. (2010)). Similarly in the biological sciences, scientists are interested in the relationship between how proteins interact and their biological importance (see Butland et al. (2005) for example). In each of these applications, the data consists of two parts: the network relations $\{y_{i,j} : i, j \in \{1, ..., n\}\}$ representing a measure of the directed relationship between each pair of nodes i and j, and p-variate nodal attributes $\{\boldsymbol{x}_i : i \in \{1, ..., n\}\}$. In the case of a social network, the nodes, network relations, and attributes often represent people, their friendships, and their demographic and behavioral characteristics, respectively.

Traditional approaches to describing the dependence between a network and attributes rely on statistical methods that model either the network conditional on the attributes or the attributes conditional on the network. In the social sciences, this first perspective parallels the theory of "social selection", whereby individuals' attributes influence the formation of their social relations, and the second perspective is motivated by "social influence" theory whereby individuals' relations affect their attributes.

Methods that model the network as a function of the attributes commonly specify a regression framework for the dependence: the probability of the relation $y_{i,j}$ is a function of $\boldsymbol{\beta}^T \boldsymbol{x}_{i,j}$ where $\boldsymbol{x}_{i,j} = f(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are the attributes for nodes i and j, and $\boldsymbol{\beta}$ is an unknown parameter vector. The covariate vector $\boldsymbol{x}_{i,j}$ typically includes terms for each

attribute of the sender node i and receiver node j, as well as interaction terms measuring the similarity between the sender and receiver attributes. These interaction terms are frequently defined as the absolute difference between the attributes, an indicator of whether an attribute is the same for both the sender and receiver node (in the case of discrete attributes), or the product of the nodes' attributes. Examples of network models that can accommodate such a regression term are exponentially parameterized random graph models (ERGM) (Frank and Strauss (1986), Wasserman and Pattison (1996), Snijders et al. (2006), Hunter and Handcock (2006)) and latent variables models (Hoff et al. (2002), Hoff (2005)). This latter class of models regresses a function of the network on both attribute terms and node-specific latent variables; Austin et al. (2013) proposed a slight modification to this class where the network is expressed as a function of nodal latent variables and the latent variables are regressed on the attributes.

Methods for assessing the impact of the network on nodal attributes often regress each node's attributes on the attributes of other nodes in the network according to the network relations. For example, Christakis and Fowler (2007) use a logistic regression model to estimate the degree to which an individual's obesity status can be explained by the obesity status of individuals in their social network (children, neighbors, spouse, etc.). Other similar models include the auto-regressive network effects models of Erbring and Young (1979) and Marsden and Friedkin (1993) and the p* social influence models of Robins et al. (2001). All of these models are univariate, focusing a single attribute of interest that is possibly subject to social influence.

While modeling the network and attributes as functions of one another is able to provide some insight into their dependence structure, there are two primary drawbacks to utilizing these methods for analysis. First, neither modeling framework allows for simultaneous inference about the dependencies between and among the network relations and attributes. For example, when analyzing data on an adolescent friendship network and individuals' health behaviors there may be interest in whether smoking habits and obesity status are conditionally independent given the network. Addressing this question of dependence between attributes conditional on the network is impossible using either of the conditional modeling frameworks. A second limitation of these methods is that they are unable to accommodate,

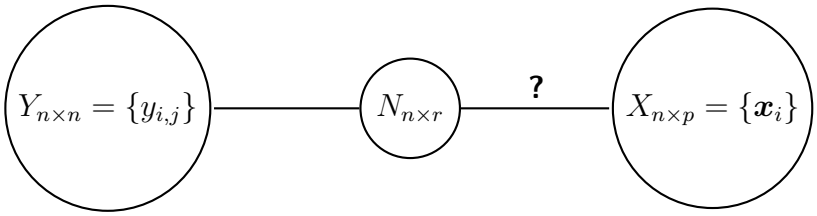

Figure 3.1: The primary patterns in the network Y are represented by r node-specific factors N. To determine if dependencies exist between the network Y and the p nodal attributes X, we propose testing for a the relationship between the network factors and attributes.

and provide predictions for, datasets that have both missing network and attribute information. In the conditional modeling frameworks either the network or attributes are assumed to be fully observed.

Fellows and Handcock (2012) proposed a new class of models called exponential-family random network models which is a combination of an ERGM and a Gibbs random field. This joint network and attribute model addresses the first limitation of the conditional models and could potentially (with modification) address the second limitation of imputing missing data. However these models, like ERGMs, are difficult to estimate and can suffer from model degeneracy problems, where networks simulated from the fitted model are unlike that which was observed (Handcock (2003), Schweinberger (2011)). Kim and Leskovec (2011) and Kim and Leskovec (2012) proposed a simple joint attribute and network model for which mathematical analysis on network connectivity and degree distributions is tractable. However, this model class only accommodates categorical attributes and assumes no missing attribute or network information. Both of these existing joint modeling frameworks lack traditional procedures for testing whether the joint model is appropriate and dependencies even exist between the network and attributes.

In this chapter, we propose a unified approach to the analysis of network and attribute data. This approach allows for testing for dependencies between the network and attributes, and in the event the test concludes such dependencies exists, jointly modeling the network and attributes to make inference and obtain predictions for missing values. Our proposed

methodology can be summarized as follows. Investigating the dependence between network data Y and attribute data X is difficult since network data is often high dimensional, containing relational information on each pair of nodes, and there lacks a one-to-one correspondence between nodes' network relations and their attributes. For these reasons, in Section 3.2 we propose representing the $(n \times n)$ matrix of network relations Y with a low dimensional structure defined by an $(n \times r)$ matrix N of node-specific network factors $(r \ll n)$. These network factors N are not observed directly and hence are estimated from the observed network Y using a network model.

In Section 3.3 we propose evaluating whether dependencies exist between the network Y and attributes X by formally testing for correlation between the estimated network factors N and the attributes X. A conceptual representation of this testing framework is shown in Figure 3.1. If the network is independent of the attributes, then any functions of the network, specifically the network factors, are also independent of the network. Therefore, any test of association between the network factors and attributes will have the correct Type I error rate. A key advantage of this approach is that the overall relationship between the network and an arbitrary number of attributes can be assessed without needing to construct a complex regression model of the network relations on the attributes or perform variable selection. In Section 3.4 we investigate the loss in power for the test between the network factors and attributes as a result of not observing the network factors directly.

A joint model for the network and attributes is presented in Section 3.5 for use when the test of independence between the network factors and attributes rejects. This joint model allows for simultaneous estimation and inference on the dependence between and within the network and attributes, as well as provides methodology for handling and predicting missing network and attribute data. We show that the joint model conditional on the attributes can be viewed as a reduced rank regression of the network relations on attribute interactions. This further motivates the model as a mechanism for parsimoniously characterizing attribute and network dependence. In Section 3.6 our proposed methodology in used to analyze data from the National Longitudinal Study of Adolescent Health. In a cross validation experiment, we demonstrate that predictions of missing attribute data can be improved by basing imputations on both observed network and attribute information instead of attribute

data alone. We conclude with a discussion in Section 3.7.

3.2 Calculation of node-specific network factors

The latent space network models in Hoff et al. (2002) and latent variable models in Hoff (2005) and Hoff (2009) provide parsimonious representations of the patterns in a network using node-specific latent factors. These models have been shown to capture a variety of network dependence patterns such as homophily, transitivity, reciprocity, and heterogeneity in node sociability and popularity. We consider an extension of the model presented in Hoff (2009) that contains additive and multiplicative latent effects, as well as structure for within dyad correlation. Ultimately, we use this model to obtain a low-dimensional representation of the network in terms of interpretable node-specific factors. We describe the model for continuous network data, however at the end of this section we briefly discuss how these methods can be extended to model ordinal or binary relations.

Let $y_{i,j}$ represent a continuous measure of the directed relation between nodes i and j and consider the following model:

$$y_{i,j} = \mu + a_i + b_j + \boldsymbol{u}_i^T \boldsymbol{v}_j + e_{i,j}, \qquad a_i, b_j \in \mathbb{R}, \qquad \boldsymbol{u}_i, \boldsymbol{v}_j \in \mathbb{R}^k. \tag{3.1}$$

The overall mean relation is represented by μ and the random error by $e_{i,j}$. The additive sender effect a_i and receiver effect b_j are often interpreted as a measure of node i's sociability (i.e. outgoingness) and node j's popularity respectively. The multiplicative interaction effect $\boldsymbol{u}_i^T \boldsymbol{v}_j$ can capture higher order dependence, such as network transitivity, balance, and clustering (Hoff (2005)). One interpretation of the these effects comes from the concept of an underlying social space (McFarland and Brown (1973), Faust (1988)), whereby nodes that are close to one another in the underlying space exhibit similar network patterns. In this context, the node-specific sender factors $\boldsymbol{u}_i$ and receiver factors $\boldsymbol{v}_i$ can be interpreted as k-dimensional representations of the underlying outgoing (sending) and incoming (receiving) behaviors of node i. A similar interpretation was used to motivate the latent position models in Hoff et al. (2002).

The random errors are modeled as Gaussian, independent across dyads, and correlated

within a dyad:

$$(e_{i,j}, e_{j,i})^T \overset{\text{iid}}{\sim} \text{normal}_2\Big(\mathbf{0}, \sigma_e^2\Big(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\Big)\Big). \tag{3.2}$$

The additive and multiplicative node-specific factors are also modeled as Gaussian and independent across nodes:

$$(a_i, b_i, \boldsymbol{u}_i^T, \boldsymbol{v}_i^T)^T \overset{\text{iid}}{\sim} \text{normal}_{2+2k}(\mathbf{0}, \Sigma_{abuv}) \qquad \Sigma_{abuv} = \begin{pmatrix} \Sigma_{ab} & \Sigma_{ab,uv} \\ \Sigma_{uv,ab} & \Sigma_{uv} \end{pmatrix}. \tag{3.3}$$

The within dyad correlation ρ is interpreted as a measure of network relation reciprocity and together with the additive effects a_i and b_j induces the covariance structure from the social relations model (Warner et al. (1979), Wong (1982)).

A Bayesian estimation procedure for this network model has been implemented in the 'amen' package in the open source computing software program `R`; however the implementation restricts $\Sigma_{ab,uv} = 0$. Under this restriction the model can capture third-order dependence patterns between relation "cycles", such as $\{y_{i,j}, y_{j,k}, y_{k,i}\}$ or $\{y_{i,j}, y_{j,k}, y_{i,k}\}$ where the edges create a closed loop (ignoring direction), but not between noncyclic relation triples such as $\{y_{i,j}, y_{j,i}, y_{k,i}\}$. By allowing the additive and multiplicative effects to be dependent (i.e. $\Sigma_{ab,uv} \neq 0$) as in (3.3), the model is able to capture a larger class of dependencies. Specifically, it can capture correlation among sets of relation triples where each relation in the set shares at least one node with another relation in the set (i.e. dependence between $\{y_{i,j}, y_{j,k}, y_{k,l}\}$, but not between $\{y_{i,j}, y_{j,k}, y_{m,l}\}$). One justification for allowing such dependence is that latent factors that act additively, affecting node popularity and sociability, plausibly also impact the network in a multiplicative manner. A modified version of the 'amen' `R` package that supports Bayesian parameter estimation for the network model presented here is available at the corresponding author's website.

Motivation via the singular value decomposition

A key strength of the network model in (3.1) is its ability to capture a variety of common network phenomena, however an alternative motivation for the model stems from its relationship to the singular value decomposition (SVD). The singular value decomposition is a

matrix factorization that is commonly used to obtain an approximation of a matrix M by another matrix $\widehat{M}$ which is of reduced rank and contains the main patterns of the original matrix M. The SVD-based approximation $\widehat{M}$ is the optimal matrix approximation of its rank with respect to squared error loss. Here we show that the model in (3.1) is similar to an SVD-based approximation of the network Y, and hence can be viewed as a low dimensional representation of the network that captures the primary patterns in the relations.

The network model in (3.1) can be written in matrix form as $Y = M + E$, where

$$M = \mu \mathbf{1_n} \mathbf{1_n}^T + \boldsymbol{a} \mathbf{1_n}^T + \mathbf{1_n} \boldsymbol{b}^T + UV^T, \tag{3.4}$$

$\boldsymbol{a}$ and $\boldsymbol{b}$ are $(n \times 1)$ vectors of the additive sender and receiver factors, U and V are $(n \times k)$ matrices of multiplicative factors, and E is an $(n \times n)$ matrix of errors.

The singular value decomposition of an arbitrary $(n \times n)$ matrix Y is written $Y = ACB^T$, where A and B are orthogonal $(n \times n)$ matrices and C is an $(n \times n)$ diagonal matrix with non-negative decreasing diagonal elements. The rank-k matrix that best approximates Y based on squared-error loss is given by $\widehat{M} = \widehat{A}\widehat{C}\widehat{B}^T$ where $\widehat{A} = A[, 1\colon k]$, $\widehat{C} = C[1\colon k, 1\colon k]$ and $\widehat{B} = B[, 1\colon k]$. Absorbing $\widehat{C}$ into $\widehat{A}$ and/or $\widehat{B}$, the best rank-k approximation is written $\widehat{M} = \breve{A}\breve{B}^T$. Letting $\boldsymbol{\mu_A}, \boldsymbol{\mu_B} \in \mathbb{R}^k$ contain the columns means of $\breve{A}$ and $\breve{B}$ respectively, $\widehat{M}$ can be expressed:

$$\widehat{M} = \mu_{AB} \mathbf{1_n} \mathbf{1_n}^T + \tilde{\boldsymbol{a}} \mathbf{1_n}^T + \mathbf{1_n} \tilde{\boldsymbol{b}}^T + \widetilde{A}\widetilde{B}^T, \tag{3.5}$$

$$\widetilde{A} = (\breve{A} - \mathbf{1_n} \boldsymbol{\mu_A}^T), \qquad \tilde{\boldsymbol{a}} = \widetilde{A} \boldsymbol{\mu_B}, \qquad \mu_{AB} = \boldsymbol{\mu_A}^T \boldsymbol{\mu_B},$$

$$\widetilde{B} = (\breve{B} - \mathbf{1_n} \boldsymbol{\mu_B}^T), \qquad \tilde{\boldsymbol{b}} = \widetilde{B} \boldsymbol{\mu_A},$$

where $\widetilde{A}$ and $\widetilde{B}$ represent multiplicative factors with mean-zero columns, $\tilde{\boldsymbol{a}}$ and $\tilde{\boldsymbol{b}}$ represent mean-zero row and column factors, and μ_{AB} is an overall mean.

Observe that the representation in (3.5) resembles that in (3.4) for the network model. This illustrates the additive and multiplicative effects structure in the network model is similar to a rank-k matrix approximation of the network. Note that in the decomposition in (3.5), there is functional dependence between the overall mean, additive, and multiplicative effects. Since there are no such restrictions in the network model, the latent network effects represent a slightly larger class of approximations than the set of matrices with rank-k.

From the representation in (3.4), it is evident that the multiplicative network effects individually are nonidentifiable: the probability model for Y is the same with multiplicative latent factors U and V as it is with factors UG^T and VG^{-1} for any nonsingular $(k \times k)$ matrix G. This issue is discussed further in Sections 3.3 and 3.5.

Non-continuous network measures

Observed network information is often not continuous. For example it is common for network data to be binary where $y_{i,j}$ is an indicator of whether the relation between nodes i and j exceeds some threshold, or ordinal where $y_{i,j}$ represents, for instance, the relative rank of node j from the perspective of node i. To model non-continuous relations, the network model in (3.1) can be incorporated into a generalized linear model framework by modeling $y_{i,j} = \ell(z_{i,j})$ where $z_{i,j}$ is a continuous measure of the pairwise relation and ℓ is a link function defining the relationship between $z_{i,j}$ and $y_{i,j}$. The latent continuous network measure $z_{i,j}$ is then modeled using the network model in (3.1) in place of $y_{i,j}$. In the case of binary data, a probit or logit link function may be appropriate, and in the ordinal case the ordered probit can be considered. Hoff et al. (2012) discusses additional link functions which account for censoring of binary and ordinal relations when nodes are restricted on the number of relations they can send (i.e. the number of non-zero relations in a row of Y). Section 3.6 illustrates the use of an appropriate link function for fixed rank nomination data from the National Longitudinal Study of Adolescent Health.

3.3 Testing for dependencies

The goals in an analysis of network and attribute data are often threefold: 1) to determine whether dependencies exist between the network and attributes, 2) to model and estimate these dependencies, and finally 3) to make inference and possibly make predictions for missing data. The first step in any such analysis is to formally test for dependencies between the network and attributes.

A classical approach to determining whether there is an association between the nodal attributes $X_{n\times p}$ and network relations $Y_{n\times n}$ would be to test whether dependencies exist between X and the rows of Y or between X and the columns of Y. This would involve

hypothesizing that each attribute is uncorrelated with each node's outgoing relations (H_0: $\text{Cov}(X[,i], Y[j,]) = 0$ for all i,j) or incoming relations (H_0: $\text{Cov}(X[,i], Y[,j]) = 0$ for all i,j) and investigating the evidence against these claims. However, conventional multivariate analysis tests are not applicable to these problems since these tests address relationships between $p + n$ variables based on n observations.

We propose an alternative testing approach using the the estimated latent network factors $N_{n\times(2k+2)} = [\boldsymbol{a}, \boldsymbol{b}, U, V]$ from the network model in (3.1). The nodal attributes $X = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]^T$ are independent of the network Y if and only if the attributes are independent of any function of the network. As described in Section 3.2, the network factors N provide a simplified representation of the network. Thus, we propose testing for dependence between the latent network factors N and attributes X on the basis that rejecting such a test would imply dependence between the network Y and attributes X (see Figure 3.1). However, the latent network factors N are not observed so in practice they must be estimated from the observed network Y. In this section we propose a test for dependence between the estimated network factors and attributes, discuss invariances in the test, and describe an exact likelihood ratio testing procedure. We also discuss alternative interpretations of the test that do not involve distributional assumptions on both the latent factors and attributes.

Suppose the nodal attributes $\{\boldsymbol{x}_i : i \in \{1, ..., n\}\}$ are continuous and mean-zero, and let $\boldsymbol{n}_i = (a_i, b_i, \boldsymbol{u}_i^T, \boldsymbol{v}_i^T)^T$ denote the (estimated) latent network factors for node i. We propose testing for linear dependence between the network factors and attributes using a classical multivariate test based on the assumption that the network factors and attributes are samples from a multivariate normal distribution:

$$\left(\boldsymbol{x}_i^T, a_i, b_i, \boldsymbol{u}_i^T, \boldsymbol{v}_i^T\right)^T = \left(\boldsymbol{x}_i^T, \boldsymbol{n}_i^T\right)^T \overset{\text{iid}}{\sim} \text{normal}_{p+2+2k}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \Sigma_{XN} = \begin{pmatrix} \Sigma_X & \Sigma_{X,N} \\ \Sigma_{N,X} & \Sigma_N \end{pmatrix}\right). \quad (3.6)$$

The null and alternative hypotheses for this test are

$$H_0 : \Sigma_{X,N} = 0 \qquad \text{vs.} \qquad H_1 : \Sigma_{X,N} \neq 0 \qquad \text{based on (3.6).} \quad (3.7)$$

Network model and test invariances

As mentioned in Section 3.2, the network model in (3.1) is invariant under transformations

of the multiplicative latent factors. Formally, this nonidentifiability can be expressed as a invariance of the probability model under transformations of network factors $\{\boldsymbol{n}_i : i \in \{1, ..., n\}\}$ by elements of group

$$\mathcal{G}_N = \left\{ \boldsymbol{G}_N = \begin{pmatrix} \mathrm{I}_2 & 0 & 0 \\ 0 & A^T & 0 \\ 0 & 0 & A^{-1} \end{pmatrix} : A_{k\times k} \text{ nonsingular} \right\},$$

which act via multiplication on the left:

$$\boldsymbol{n}_i = (a_i, b_i, \boldsymbol{u}_i^T, \boldsymbol{v}_i^T)^T \rightarrow \boldsymbol{G}_N \boldsymbol{n}_i = (a_i, b_i, A^T \boldsymbol{u}_i^T, A^{-1} \boldsymbol{v}_i^T)^T.$$

It would be undesirable for the test in (3.7) to depend on which latent factors in the set $\{\{\boldsymbol{G}_N \boldsymbol{n}_i : i \in \{1, ..., n\}\} : \boldsymbol{G}_N \in \mathcal{G}_N\}$ are selected to represent the network. Define $\mathcal{G}$ to be the extension of group $\mathcal{G}_N$ to transformations of $\left(\boldsymbol{x}_i^T, \boldsymbol{n}_i^T\right)^T$:

$$\mathcal{G} = \left\{ \boldsymbol{G} = \begin{pmatrix} I_p & 0 \\ 0 & \boldsymbol{G}_N \end{pmatrix} : \boldsymbol{G}_N \in \mathcal{G}_N \right\},$$

which acts via left multiplication and leaves $\boldsymbol{x}_i$ unchanged. We define $\mathcal{G}$ in order to relate the invariance in the network model parameterization to the test in (3.7).

The testing problem in (3.7) is itself invariant under left multiplication of $\left(\boldsymbol{x}_i^T, \boldsymbol{n}_i^T\right)^T$ by elements in the group $\mathcal{F}$, where $\mathcal{F}$ is defined

$$\mathcal{F} = \left\{ \boldsymbol{F} = \begin{pmatrix} B^X & 0 \\ 0 & B^N \end{pmatrix} : B^X_{p\times p}, B^N_{(2k+2)\times(2k+2)} \text{ nonsingular} \right\}.$$

An $\mathcal{F}$-invariant test is a test for (3.7) that produces the same results for all attributes and network factors that are equivalent under group $\mathcal{F}$. Observe that $\mathcal{G}$ is a subgroup of $\mathcal{F}$. This implies that an $\mathcal{F}$-invariant test will also respect the $\mathcal{G}$-invariances in the network relations probability model. In other words, all attributes and latent network factors that are equivalent under group $\mathcal{G}$ will generate the same test results for (3.7) under an $\mathcal{F}$-invariant test.

Likelihood ratio test

There is no uniformly most powerful invariant test for (3.7), however the likelihood ratio

test is $\mathcal{F}$-invariant, unbiased (Perlman and Olkin (1980)), and generally performs well. Let $N = [\boldsymbol{a}, \boldsymbol{b}, U, V]$ be the $(n \times (2k+2))$ matrix of network factors. The likelihood ratio test statistic can be written

$$\Lambda = \frac{\max_{\Sigma}\ L(\Sigma|N, X)}{\max_{\Sigma_X, \Sigma_N}\ L_0(\Sigma_X, \Sigma_N|N, X)} = \prod_{i=1}^{p \wedge (2k+2)} (1 - r_i^2)^{-n/2} \tag{3.8}$$

where L_0 and L refer to the likelihood corresponding to the multivariate normal model in (3.6) with and without restricting $\Sigma_{N,X} = 0$. The term r_i^2 is the ith eigenvalue of

$$(X^T X)^{-1/2}(X^T N)(N^T N)^{-1}(N^T X)(X^T X)^{-1/2},$$

and its positive square root is commonly referred to as the ith canonical correlation between N and X. This correlation represents the largest correlation obtainable between a linear combination of attributes and a linear combination of the network factors such that the linear combinations are uncorrelated with the respective combinations used to obtain the first $i-1$ correlations. The test based on (3.8) rejects the null hypothesis for large values of Λ and was shown to have monotonically increasing power as a function of each population canonical correlation (Anderson and Gupta (1964)).

Under the null hypothesis, $W = \Lambda^{-2/n}$ has a Wilks' Lambda $U(p, 2k+2, n-(2k+2))$ distribution, which is equivalent to the product of independent, Beta distributed random variables (Muirhead (1982)):

$$W \sim U(p, 2k+2, n-(2k+2)) = \prod_{i=1}^{p} \text{Beta}\left(\frac{n-(2k+2)-p+i}{2}, \frac{2k+2}{2}\right). \tag{3.9}$$

The α-quantiles for this distribution can be obtained via Monte Carlo estimation and used to perform exact level-α tests for (3.7).

Alternative interpretation of the test

The test in (3.7) was derived as the likelihood ratio test for a model where both the network factors and attributes are samples from a normal distribution. However in some cases these assumptions may not be appropriate. Fortunately, alternative interpretations of the test exist that do not rely on such assumptions. The likelihood ratio test in (3.8) for the test in

(3.7) is the same as the likelihood ratio test to determine whether the coefficients in a linear regression are nonzero, where either the network factors are regressed on the attributes or the attributes are regressed on the network factors. These conditional tests can be expressed

$$\mathrm{H}_0 : \boldsymbol{\beta}_{X|N} = \mathbf{0} \;\text{ vs. }\; \mathrm{H}_1 : \boldsymbol{\beta}_{X|N} \neq \mathbf{0} \;\text{ based on }\; \boldsymbol{x}_i|\boldsymbol{n}_i \overset{\text{iid}}{\sim} \text{normal}\left(\beta_{X|N}\boldsymbol{n}_i, \Sigma_{X|N}\right), \text{ and} \tag{3.10}$$

$$\mathrm{H}_0 : \boldsymbol{\beta}_{N|X} = \mathbf{0} \;\text{ vs. }\; \mathrm{H}_1 : \boldsymbol{\beta}_{N|X} \neq \mathbf{0} \quad \text{based on} \quad \boldsymbol{n}_i|\boldsymbol{x}_i \overset{\text{iid}}{\sim} \text{normal}\left(\beta_{N|X}\boldsymbol{x}_i, \Sigma_{N|X}\right), \tag{3.11}$$

where $\beta_{X|N}$ and $\beta_{N|X}$ are $(p \times (2 + 2k))$ and $((2 + 2k) \times p)$ matrices, respectively. If the nodal attributes were specified as part of the study design or are binary or ordinal, it may be inappropriate to model them as Gaussian as is done in (3.7). Instead it may be preferable to test for dependence between the attributes and network factors via the conditional formulation in (3.11), where no distributional assumptions are placed on X. The likelihood ratio test for the tests in (3.7), (3.10) and (3.11) are identical, so the testing framework presented here is appropriate if the assumption of normality is reasonable for one or both of the network factors and attributes.

Furthermore, it is worth emphasizing that although the network factors N are estimated, the Wilks' Lambda distribution in (3.9) associated with the likelihood ratio test statistic is exact if either the attributes X conditional on the network factor N or network factors conditional on the attributes are samples from a normal distribution. In practice, a central limit theorem argument can be used to claim the distribution in (3.9) is approximately correct when n is large.

3.4 Simulation study

To analyze data with the test outlined in Section 3.3, the network latent factors N must be estimated from the observed network Y. We expect this to result in a decrease in the power of the test in (3.7) compared to if the network factors were able to be observed directly. Furthermore, we expect a greater decrease in power when the observed network relations are less informative (i.e. binary rather than continuous). In this section we present a simulation study that quantifies the degree to which power is lost when the network factors are not observed and must be estimated from observed network relations.

Consider the network model in (3.1) with one multiplicative effect ($k = 1$), zero mean ($\mu = 0$), and independent standard normal errors ($\rho = 0$, $\sigma_e^2 = 1$):

$$y_{i,j} = a_i + b_j + u_i v_j + e_{i,j}, \qquad a_i, b_j, u_i, v_j \in \mathbb{R}, \qquad e_{i,j} \sim \text{normal}(0, 1). \tag{3.12}$$

We consider the case where one nodal attribute is of interest ($p = 1$) and the attribute and latent network factors have one of the following covariance structures:

A) $\text{Cov}[(x_i, a_i, b_i, u_i, v_i)] = \Sigma_{XN} = \text{I} + \gamma E_{x,a}$,

B) $\text{Cov}[(x_i, a_i, b_i, u_i, v_i)] = \Sigma_{XN} = \text{I} + \gamma E_{x,u}$.

$E_{x,a}$ is the (5×5) matrix of zeros with a one in the entires corresponding to $\text{Cov}[x, a]$ and $\text{Cov}[a, x]$, and $E_{x,u}$ is defined analogously. In scenario A the attribute and each network factor are uncorrelated, except the additive sender factor a_i and the attribute x_i which have correlation γ. Similarly, in scenario B correlation γ exists between the sender multiplicative factor u_i and the attribute x_i.

Monte Carlo estimates of the power based on the level-0.05 likelihood ratio test in (3.8) for the test in (3.7) were computed for squared correlation values $\gamma^2 \in \{-0.05, 0, 0.05, 0.1, 0.15, 0.2\}$, network sizes $n \in \{25, 50, 100\}$ and three decreasingly informative observations of the network:

1. $N = [\boldsymbol{a}, \boldsymbol{b}, U, V]$ is observed;
2. N is estimated from a continuous network Y according to (3.1);
3. N is estimated from a binary network B_d, where B_d is defined as $B_d = \{b_{i,j} : b_{i,j} = 1 \text{ if } y_{i,j} > y_d,\ 0 \text{ otherwise}\}$ and y_d is chosen such that the proportion of network relations greater than y_d (i.e. the network density) is d.

Notice that the binary network B_d is a deterministic function of the continuous network Y. We consider the binary networks with density 0.5 and 0.15. The former case represents a relatively dense binary network with many observed relations, whereas the latter case reflects more common network seen in survey data where information about only a small number of nodes' relations are available. For the continuous network Y and binary networks

B_d, the Bayesian estimation procedure was used to obtain estimates of the latent network factors. A probit link function was specified for the binary networks. The additive factors $\boldsymbol{a}$ and $\boldsymbol{b}$ were estimated by their posterior means, and the multiplicative factors U and V were estimated by the first left and right singular vector of the posterior mean of the multiplicative effect UV^T.

Figure 3.2 shows the power estimates for the two covariance structures A and B and the four network observations (N, Y, $B_{0.5}$, $B_{0.15}$). A single power curve is shown for the latent network factors N since the correlation structures A and B are equivalent with respect to the invariances of the test in (3.7). Most notably, Figure 3.2 illustrates there is relatively little power lost when the network factors are estimated from an observed continuous or binary network, even when the network size is small. The power of the test is slightly larger when dependence exists between the attribute and an additive factor compared to when it exists between the attribute and a multiplicative factor for continuous and binary network observations. This is likely a consequence of the relative ease with which additive effects are estimated compared to interaction effects. As expected, the power of the test decreases as the observed network information becomes less informative ($N \rightarrow Y \rightarrow B_{0.5} \rightarrow B_{0.15}$), although for even moderate network sizes the power loss is negligible.

3.5 Joint model for the network and nodal attributes

If the test in Section 3.3 rejects the null hypothesis of independence between the attributes and network factors, there is often interest in estimating and making inference on the dependencies, as well as predicting missing network and attribute information. Addressing such inference objectives requires joint modeling of the network Y and attributes X. We propose jointly modeling the network Y and attributes X via a model composed of the network relations model in (3.1) and (3.2), and the latent factor and attribute model in (3.6). For completeness, we include all components of the joint model below:

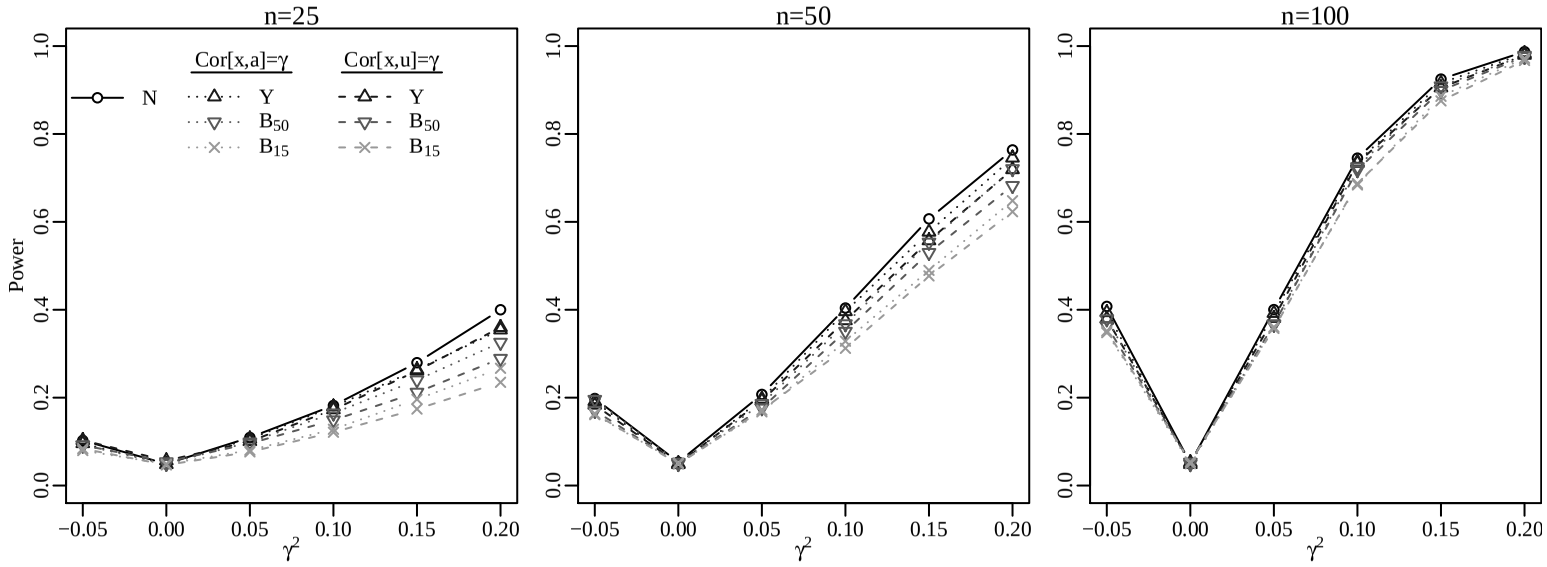

Figure 3.2: Power when testing for independence between a single attribute x_i and network factors $\{a_i, b_i, u_i, v_i\}$ based on four types of network observations (latent network factors N, continuous network Y, binary network $B_{0.50}$, binary network $B_{0.15}$).

$$y_{i,j} = \mu + a_i + b_j + \boldsymbol{u}_i^T \boldsymbol{v}_j + e_{i,j}, \tag{3.13}$$

$$(e_{i,j}, e_{j,i})^T \overset{\text{iid}}{\sim} \text{normal}\Big(\mathbf{0}, \sigma_e^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\Big), \tag{3.14}$$

$$\big(\boldsymbol{x}_i^T, a_i, b_i, \boldsymbol{u}_i^T, \boldsymbol{v}_i^T\big)^T = \big(\boldsymbol{x}_i^T, \boldsymbol{n}_i^T\big)^T \overset{\text{iid}}{\sim} \text{normal}\Big(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \Sigma_{XN} = \begin{pmatrix} \Sigma_X & \Sigma_{X,N} \\ \Sigma_{N,X} & \Sigma_N \end{pmatrix}\Big). \tag{3.15}$$

Inference for the dependence and conditional dependencies between the attributes and network is based on the covariance matrix Σ_{XN}.

Simplified parameterization

The nonidentifiability of the latent factors discussed in Sections 3.2 and 3.3 translates to nonidentifiability of portions of the covariance matrix Σ_{XN}. However, by restricting the covariance matrix to have specific structure, the $\mathcal{G}$-invariance of the network model due to the multiplicative latent factors can be removed.

We propose reparameterizing the model for the latent factors and attributes in (3.15) by

$$\big(\boldsymbol{x}_i^T, a_i, b_i, \boldsymbol{u}_i^T, \boldsymbol{v}_i^T\big) \overset{\text{iid}}{\sim} \text{normal}_{p+2+2k}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \Sigma_{XN} = \begin{pmatrix} \Sigma_{Xab} & \Sigma_{Xab,U} & \Sigma_{Xab,V} \\ \Sigma_{U,Xab} & D & \Sigma_{U,V} \\ \Sigma_{V,Xab} & \Sigma_{V,U} & D \end{pmatrix}\right), \tag{3.16}$$

where D is a diagonal matrix with decreasing elements along the diagonal. This joint model defined by (3.13), (3.14), and (3.16) is not invariant to transformations of the network factors and attributes by elements in the group $\mathcal{G}$, however it continues to posses non-identifiabilty with respect to signs of the entries in U and V. Specifically, the probability of the observed network Y and attributes X is the same with parameters $\{U, V, \Sigma_{XN}\}$ as it is with parameters

$$\left\{US, VS, \begin{pmatrix} I_{p+2} & 0 & 0 \\ 0 & S & 0 \\ 0 & 0 & S \end{pmatrix} \Sigma_{XN} \begin{pmatrix} I_{p+2} & 0 & 0 \\ 0 & S & 0 \\ 0 & 0 & S \end{pmatrix}\right\},$$

where $S_{k\times k}$ is any diagonal matrix with ± 1's along the diagonal.

Relation to reduced rank regression

The expectation of the network relations conditional on the attributes based on (3.13) resembles that of a reduced rank regression on (multiplicative) attribute interaction effects.

This is noteworthy as the motivations underlying reduced rank regression parallel many of the arguments supporting this network modeling framework.

The expectation of the network factors conditional on the attributes can be written

$$\mathrm{E}[(a_i, b_i, \boldsymbol{u}_i^T, \boldsymbol{v}_i^T)^T] = (\beta_{a|X}\boldsymbol{x}_i, \beta_{b|X}\boldsymbol{x}_i, (\beta_{U|X}\boldsymbol{x}_i)^T, (\beta_{V|X}\boldsymbol{x}_i)^T)^T,$$

where $\beta_{a|X}$, $\beta_{b|X}$ are $(p \times 1)$ vectors and $\beta_{U|X}$ and $\beta_{V|X}$ are $((2 + 2k) \times p)$ matrices of coefficients based on Σ_{XN}. Since the latent factors for different nodes are modeled as independent, the expectation of the network relations in (3.13) conditional on the attributes is

$$\mathrm{E}[y_{i,j}|\boldsymbol{x}_i, \boldsymbol{x}_j] = \mu + \beta_{a|X}\boldsymbol{x}_i + \beta_{b|X}\boldsymbol{x}_j + \boldsymbol{x}_i^T\beta_{U|X}^T\beta_{V|X}\boldsymbol{x}_j.$$

The interaction term $\boldsymbol{x}_i^T\beta_{U|X}^T\beta_{V|X}\boldsymbol{x}_j$ represents a linear combination of all possible pairwise products between the p sender and p receiver attributes, resulting in p^2 linear effects. The coefficients on these linear effects are given by the $(k \times k)$ matrix $\beta_{U|X}^T\beta_{V|X}$, whose rank is at most equal to the minimum of p and k. Therefore, if the number of attributes is greater than the number of multiplicative network factors ($p \geq k$), linear constraints will exist among the p^2 effect coefficients. In reduced rank regression the coefficient matrix corresponding to the regression of a multivariate outcome on a multivariate predictor is restricted to be reduced rank (Anderson (1951), see Reinsel and Velu (1998) for a comprehensive review). This approach to parameter dimension reduction is motivated by improvement in parameter estimation and interpretation. A similar goal exists in network modeling and is achieved here using the latent network factors. Modeling dependencies between the latent network factors N and attributes X instead of between the network relations Y and attributes X directly allows us to parsimoniously estimate and characterize complex (multiplicative) dependencies without defining a complicated regression model for the network relations. This approach is especially advantageous when the number of attributes is large and/or it is likely at most a small number of attribute pairs are related to the network.

Estimation

Estimation of the parameters in the joint network and attribute model is straightforward in a Bayesian context, where inference is based on the joint posterior distribution of the

network factors $\{\boldsymbol{a}, \boldsymbol{b}, U, V\}$ and parameters $\{\sigma_e^2, \rho, \Sigma_{XN}\}$ given the data $\{X, Y\}$. Since an analytic expression of the posterior distribution is not available, it is approximated by samples generated from a Markov chain Monte Carlo (MCMC) algorithm. The MCMC procedure implemented in the `R` package 'amen' for the model described in Section 3.2, where the additive and multiplicative factors are uncorrelated, was adapted for the joint model presented here. Details regarding the families of prior distributions considered and the corresponding MCMC algorithm are included in the appendix. Code is provided at the corresponding author's website.

3.6 Analysis of AddHealth data

We consider data from a survey of 389 high-school students from the National Longitudinal Study of Adolescent Health (AddHealth) (Harris et al. (2009)) and investigate whether evidence exists that student friendships are related to student health behaviors and grade point average (GPA). The data we use includes same-sex friendship nomination data, whereby students identified the top five friends of their sex, in addition to demographic and behavioral information. The data considered here can be described as follows:

- **network information -** $R = \{r_{i,j}\}$: $r_{i,j}$ is the rank of student j in student i's listing of friends (5 = highest, 1 = lowest) or 0 if student i did not list student j;

- **nodal attributes -** $X = [\boldsymbol{x}^{\text{exercise}}, \boldsymbol{x}^{\text{drink}}, \boldsymbol{x}^{\text{gpa}}]$: standardized measures of exercise frequency, drinking frequency, and grade point average;

- **nodal covariate -** $W = [\boldsymbol{w}^{\text{grade}}]$: student grade level (9, 10, 11, or 12).

Students in the same grade and adjacent grades are more likely to be friends than students many grades apart. For this reason, we refine our question of interest to be whether students' attributes (exercise, drinking, and GPA) are associated with their network relations' while controlling for their grade.

We use the fixed rank nomination likelihood introduced in Hoff et al. (2012) to model the observed network ranks and restriction that at most five friends could be listed on the

survey. This likelihood assumes each observed network relation $r_{i,j}$ is the function of an underlying (latent) continuous measure $z_{i,j}$ such that the following relation consistencies are satisfied:

$$\begin{aligned} &r_{i,j} > 0 \Rightarrow z_{i,j} > 0, \\ &r_{i,j} > r_{i,k} \Rightarrow z_{i,j} > z_{i,k}, \\ &r_{i,j} = 0 \text{ and student } i \text{ listed} < 5 \text{ friends} \Rightarrow z_{i,j} \leq 0. \end{aligned} \tag{3.17}$$

The first association is the link function used in probit regression which assumes that if a friendship is reported, the latent friendship value must exceed a given threshold (in this case 0). The second relation assures consistency of the ranks with the latent friendship measures. Finally, the last association posits that friendships between a given student and all students he/she did not list as a friend must be below the friendship threshold if the nominating student listed fewer than five friends.

The network model for the latent relations $z_{i,j}$ is that given in (3.1) with additional regression terms for whether students are in the same grade $w^s_{i,j}$ and whether they are in adjacent grades $w^a_{i,j}$:

$$z_{i,j} = \mu + \beta_s w^s_{i,j} + \beta_a w^a_{i,j} + a_i + b_j + \boldsymbol{u}_i^T \boldsymbol{v}_j + e_{i,j}, \qquad a_i, b_i \in \mathbb{R}, \qquad \boldsymbol{u}_i, \boldsymbol{v}_i \in \mathbb{R}^k. \tag{3.18}$$

Selection of factor dimension k

The multiplicative factor dimensions k for the male and female networks were determined using a method analogous to the scree plot method which is commonly used in factor analysis and principal components analysis. The network model in (3.17) and (3.18) was fit to each gender network with $k = 8$. Let M denote the posterior mean estimate of the multiplicative network effect UV^T, and $\widehat{M}$ represent the rank eight matrix approximation of M based on the singular value decomposition. The total variation in $\widehat{M}$ is equal to the sum of the squared singular values: $||\widehat{M}||_F^2 = \sum_{\ell=1}^{8} \lambda_\ell^2$, where $||\cdot||_F$ denotes the Frobenius norm and λ_i is the ith singular value. Figure 3.3 shows the proportion of the total variation in $\widehat{M}$ attributed to each multiplicative effect (i.e. $\lambda_i^2 / \sum_{\ell=1}^{8} \lambda_\ell^2$). For both the male and female network the large majority of the variation in the network relations explained by the eight

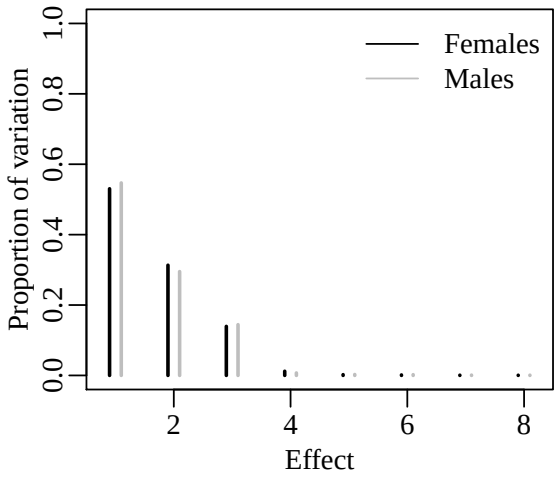

Figure 3.3: Proportion of variation in the posterior mean eight factor multiplicative effect $\widehat{M}$ that is explained by each multiplicative effect.

multiplicative effects is associated with the first three effects. Thus, the multiplicative effect dimension was selected to be three for both networks.

Testing for dependence

As discussed in the Introduction, a traditional approach to modeling dependence between the network relations and nodal attributes would be to include regression terms in the form of sender, receiver, and interaction effects for the three attributes. Including all such effects using this approach would require 15 regression terms. However, by performing the test of independence proposed in Section 3.3 based on the latent network factors, we are able to assess the evidence for any relationship between the attributes and network without creating a potentially unnecessarily complex network model or performing any model selection.

The latent network factors for the model in (3.17) and (3.18) with $k = 3$ were estimated for the male and female networks. The additive factors $\boldsymbol{a}$ and $\boldsymbol{b}$ were estimated by their posterior means, and the multiplicative factors U and V were estimated by the first three left and right singular vectors of the posterior mean of the multiplicative effect UV^T. The

test of independence between the network factors and the three nodal attributes for the female and male network resulted in p-values < 0.001. Therefore, based on a 0.05 level test, we reject the null hypothesis of independence between the student attributes and their network relations after accounting for grade structure.

Jointly modeling the network and attributes

The rejections of the tests of independence between the attributes and network suggests the network factors are informative for nodal attribute data. To investigate this claim we performed a 20-fold cross validation on each sex dataset in which 5% of data for each attribute was treated as missing in each experiment. We compared predictions for the missing attributes based on the observed attributes alone to predictions based on both the network and observed attributes. The predictions based solely on the attributes were the fitted values from a regression of each attribute on all other attributes. The predictions based on the network and attributes were the posterior mean estimates from the Bayesian estimation procedure for the joint network and attribute model introduced in Section 3.5. For each sex dataset, a Markov chain was run for 500 iterations of burn-in followed by an additional 500,000 iterations and samples were thinned to every 25th iteration, resulting in 20,000 simulated values for each missing element. The average effective sample size was 2,607 for the male network and 734 for the female network.

Table 3.1: Mean squared error for predictions from 20-fold cross validation.

	Males			Females		
Method	Exercise	Drinking	GPA	Exercise	Drinking	GPA
Regression (attributes only)	1.89	3.24	2.38	1.67	2.38	2.29
Joint model (attributes & network)	1.75	2.69	2.18	1.61	2.17	1.93
% improvement	7.4	17.0	8.4	3.6	8.8	15.7

Table 3.1 shows the mean squared error over the 20 cross validations for each attribute and each sex dataset. The predictions based on the network and attributes improved upon

the predictions based on the attributes alone for both sexes and all attributes. The improvement was greatest for male drinking frequency and female GPA where prediction mean squared error was reduced by about 15%. This illustrates that when dependence exists between the network and attributes, improvements in the predictions of missing values can be obtained by using both the network and attribute information.

3.7 Discussion

In this chapter we introduced an approach for testing whether dependencies exist between a network and attribute data that relies on a simplified representation of the network in terms of latent node-specific factors. The proposed method tests for dependencies between the network latent factors and attributes as a surrogate for testing for dependencies between the network and attributes. This test was shown to have the correct level under the null hypothesis of independence and have only a slight loss in power due to the fact that the network factors are not directly observed. Methodology for jointly modeling the network and attributes was also introduced, and in a cross validation experiment, we illustrated that predictions for missing attributes can be improved by basing predictions on both observed network and attribute information rather than on attribute information alone.

As discussed in the Introduction, many others have investigated the relationship between network and attribute data. The most common methods involve regressing either the network on functions of the nodal attributes or each node's attributes on functions of the attributes of the node's neighbors in the network. Frequently final models are settled upon after some, often undocumented, model selection procedure which not only alters the interpretation of the results, but also adds an additional element of subjectively to the analysis. The key distinction between the previous methods and that presented here is that our method does not involve any model selection procedures and simultaneously tests and estimates first and second order dependencies between the network and attributes.

A historically difficult problem not addressed here is how to select the number of multiplicative factors for the network model. In Section 3.6 we illustrated a procedure similar to the scree plot method used frequently to choose the number of factors in factor analysis and the number of eigenvectors in principal component analysis. An alternative approach

would be to incorporate the dimension selection into the model by placing a prior on the number of factors similar to that proposed in Hoff (2007) for the singular value decomposition. However this would greatly increase the complexity of the model and computation time of estimation.

The ultimate goal in social network analysis is to understand the causal relationships between the network and attributes, such as whether individuals' relationships cause changes in their behaviors or whether individuals' behaviors dictate their relationship choices (see Christakis and Fowler (2007), Fowler and Christakis (2008)). However the methods developed here solely address whether dependencies exist between the network and attributes. To answer questions about causality, temporal network and attribute information is required, however most relational datasets are cross-sectional and those that are temporal typically contain extremely limited network information. When sufficiently complete temporal network and attribute data is available, the methods developed here could be extended to accurately address questions about the existence of causal relationships.

Chapter 4

BAYESIAN INFERENCE FOR NETWORK AND RELATIONAL MODELS WITH ADDITIVE AND MULTIPLICATIVE EFFECTS

4.1 Introduction

Relational data, also commonly referred to as network data, consists of measurements on pairs of actors and arises in a number of scientific disciplines. In the social sciences a relation $y_{i,j}$ may represent a indicator of whether individual i considers individual j to be a friend, and in the biological sciences, relation $y_{i,j}$ may be the rank of the amount of interaction between protein i and protein j among protein i's interactions. There is often great diversity not only in what the actors and relations represent, but also in the type of the observed relations $y_{i,j}$, which can be continuous, binary, ordinal, and/or censored. In this chapter, we consider directed relations where in general $y_{i,j} \neq y_{j,i}$ and assume the relations $y_{i,i}$ between an actor and itself is undefined.

Different types of relational observations (binary, ranked, etc) can often be interpreted as coarsened measures of some underlying relationship between actors in a pair. For example, consider a relational dataset on a group of high schools students. A continuous dataset may contain entries $y_{i,j}$ which represent the amount of time student i spent with student j during a school year. In a dataset with ranked observations, $y_{i,j}$ may represent the rank of the relative amount of time student i and student j spent together compared to either all other student pairs or to the amount of time student i spent with each other student (i.e. the rank from the perspective of student i). Finally, if the interaction times between students are dichotomized based on whether the times exceeded a given threshold, the observed relations $y_{i,j}$ would be binary. Since each type of observed relation contains information about underlying student interaction, it may be desirable to use methodology that models the observations similarly such that inference based on one observation type is consistent and comparable with that of another type.

Numerous methods have been proposed to model relational data, however most are only able to accommodate specific types of observations. For continuous relations, ANOVA based models have been proposed that decompose observations $y_{i,j}$ into effects of the relation sender i, effects of relation receiver j, and correlation within the dyad $(y_{i,j}, y_{j,i})$ (Warner et al. (1979), Wong (1982)). In the case of binary data where $y_{i,j}$ represents the presence or absence of a relationship between actors, exponential random graph models (ERGM) are popular tools for analysis and model the probability of the presence of a relationship as a function of explanatory variables and network patterns using a small number of sufficient statistics (Frank and Strauss (1986), Wasserman and Pattison (1996), Snijders et al. (2006), Hunter and Handcock (2006)). Stochastic block models represent an additional class of binary relation models (Nowicki and Snijders (2001), Airoldi et al. (2008)), whereby actors are clustered into one or more groups using latent variables. Krivitsky (2012) proposed extensions to ERGMs to accommodate count and ranked data, however these extensions do not directly accommodate censored data and further complicate an already difficult ERGM estimation procedure. Although many of these methods capture similar network patterns, the model formulations and estimation procedures vary drastically. This discontinuity between the models and types of relational observations is unsatisfactory when, as described above for student interactions, different types of relational observations can be only slightly different measures of the same underlying phenomenon.

A flexible class of relational data models are latent variable models, where (non-continuous) observed relations are expressed as a function of underlying continuous relations. These latent relations are then parsimoniously modeled as function of covariates, additive, and multiplicative effects (Hoff et al. (2002), Hoff (2005), Hoff (2009)). In a companion paper, Hoff et al. (2012) extended this class of models to accommodate censored binary and ranked relations to asses the effect of ignoring censoring on inference. A key advantage of these models is that they decouple the modeling of the network phenomena with the modeling of the specific type of relational data: The model for the latent relations posits a decomposition of relational patterns into interpretable effects, while specification of the functional relationship between the latent relations and observed relations reflects the specific type of observed data.

In this chapter, we discuss estimation for a general class of relational data latent variable models, which extends that studied in Hoff et al. (2012), can accommodate a wide variety of observed relation types, and decomposes the patterns in the observed relations into covariate effects and actor-specific additive and multiplicative effects. Our proposed Bayesian estimation procedure requires minor modifications for new forms of relational observations and easily accommodates missing data. In Section 4.2, we introduce the class of relational latent variable models and outline a basic Markov chain Monte Carlo (MCMC) sampler for this class of models based on a Gibbs sampler in Section 4.3. We find obtaining a reasonable approximation to a posterior distribution based on samples from this traditional MCMC procedure can take an undesirable amount of time due to poor mixing of the chain. Thus, in Section 4.4 we propose three small but crucial adjustments to the MCMC procedure which significantly improves the efficiency of the sampler. We illustrate the improvement in estimation efficiency between the original and new MCMC procedures in a simulation study in Section 4.5. We discuss a mean-field variational Bayesian approach to estimation in Section 4.6, which approximates the posterior distribution of the parameters given the observed data with a distribution that assume conditional independencies between parameters. In Section 4.7, we compare the true posterior distribution to the variational Bayesian approximation in a simulation study and conclude with a discussion in Section 4.8.

4.2 Relational models with additive and multiplicative effects

Consider an $(n \times n)$ matrix of observed relations Y with entries $y_{i,j}$ and undefined diagonal entries $y_{i,i}$. In this section we discuss a general class of models for such relations, where Y is modeled as a (potentially) coarsened representation of an $(n \times n)$ matrix of latent (continuous) underlying relations Z. These latent relations are in turn modeled as a function of regression and actor-specific additive and multiplicative effects, which have been shown to capture a variety of relational patterns such as transitivity, clustering, and reciprocity (Hoff (2005), Hoff (2009)).

The class $\mathcal{Y}$ of relational observations Y considered here is an extension of those considered in Hoff et al. (2012). We distinguish between continuous and non-continuous observations and define the class $\mathcal{Y}$ under consideration as consisting of two parts:

1. **Ordinal or binary measures Y.** We assume the observed relations Y provide information about the latent underlying relations Z, such that having observed Y, the set of possible Z values is restricted to a set $S(Y)$. The relations Y we consider here are those whose set $S(Y)$ is closed under multiplication by a positive scalar: $Z \in S(Y) \Rightarrow cZ \in S(Y)$ for any $c > 0$.

2. **Continuous measures, $y_{i,j} \in \mathbb{R}$.** For notational simplicity throughout the paper, we will still discuss relations Z when the observations Y are continuous, however Z will be treated as observed and equal to Y.

A few examples of non-continuous observations Y whose set $S(Y)$ can naturally be defined to satisfy the set constraint of closure under multiplication by a positive scalar are binary, ranked, and ego-centric ranked observations. We briefly discuss each of these observation types below.

- **Binary observations**

  If $y_{i,j} \in \{0, 1\}$ is binary, we model it as an indicator of whether $z_{i,j}$ exceeds a given threshold, in this case zero. Thus Z given Y is restricted to the set

  $$S(Y) = B(Y) \coloneqq \{Z : z_{i,j} > 0 \text{ if } y_{i,j} = 1; z_{i,j} < 0 \text{ if } y_{i,j} = 0\}.$$

  This set restriction corresponds to the probit link function used in binary regression models.

- **Ordinal observations**

  If $y_{i,j}$ is ordinal, possibly representing the rank of the underlying relation $z_{i,j}$ in the dataset, then $S(Y)$ can be defined

  $$S(Y) = R(Y) \coloneqq \{Z : z_{i,j} > z_{k,l} \text{ if } y_{i,j} > y_{k,l} \text{ for all } i, j, k, l\}.$$

  This latent relation formulation has been used in semi-parametric regression modeling in Pettitt (1982) and for copula estimation in Hoff (2007).

- **Ego-centric ranked observations**

  Much relational data in the social sciences is obtained from ego-centric surveys where responses are comparable within a row but not across rows. For example, in fixed rank nomination surveys individuals are often asked to list and rank their top m relationships. A particular example comes from the National Longitudinal Study of Adolescent Health where students identified and ranked their top five friends of each sex (Harris et al. (2009)). In this case, $y_{i,j}$ represents the rank of individual j according to individual i's friendship nominations. The ranks for different nominating individuals are not comparable as the relation attributes which qualify an individual as a third best friend for one individual may be different than for another individual. Therefore, in these cases, Z is restricted to a set

  $$S(Y) = E(Y) \coloneqq \{Z : z_{i,j} > z_{i,k} \text{ if } y_{i,j} > y_{i,k} \text{ for all } i, j, k\},$$

  which enforces a relative ordering of entries in Z in the same row.

See Hoff et al. (2012) for additional examples of sets defined by relational observations which are censored.

Let $p(Z|\theta)$ denote a model for the latent relations Z with parameters θ (a specific model will be discussed in detail below). For non-continuous data Y, the likelihood of the parameters θ given the observed relations Y is defined as an integral over the latent relations Z:

$$L(\theta|Y) = p(Z \in S(Y)|\theta) = \int_{S(Y)} p(Z|\theta) d\mu(Z) \tag{4.1}$$

where μ is a measure that dominates $\{p(Z|\theta) : \theta \in \Theta\}$ (note that in the continuous case, $L(\theta|Y) = p(Z|\theta)$ since $Z = Y$).

The models we consider for Z are a reparameterization and extension of those introduced in Hoff (2009), where each latent relation $z_{i,j}$ is modeled as a function of regression terms and additive and multiplicative effects:

$$z_{i,j} = \boldsymbol{\beta}^T \boldsymbol{x}_{i,j} + a_i + b_j + \boldsymbol{u}_i^T \boldsymbol{v}_j + e_{i,j}, \qquad a_i, b_j \in \mathbb{R} \qquad \boldsymbol{u}_i, \boldsymbol{v}_j \in \mathbb{R}^k \qquad \boldsymbol{\beta}, \boldsymbol{x}_{i,j} \in \mathbb{R}^p. \tag{4.2}$$

In this model $\boldsymbol{\beta}$ represents a vector of (unknown) regression coefficients, a_i is an additive sender effect which reflects correlation among the relations in a row of Z, b_j is an additive receiver effect which reflects correlation among the relations in a column of Z, and $e_{i,j}$ represents random error. The multiplicative term $\boldsymbol{u}_i^T \boldsymbol{v}_j$ is an interaction effect which allows the model to capture higher order dependence patterns in the relations, such as transitivity and clustering (see Hoff (2005) for discussion of a similar effect).

The random errors are modeled as Gaussian, independent across dyads, and correlated within a dyad:

$$(e_{i,j}, e_{j,i})^T \overset{\text{iid}}{\sim} \text{normal}_2\Big(\mathbf{0}, \sigma_e^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\Big), \quad \{1 \le i < j \le n\}, \tag{4.3}$$

where the correlation parameter ρ is often interpreted as a measure of the reciprocity in the relations. When the relations Y are not continuous, the scale of the latent relations Z is non-identifiable so σ_e^2 in (4.3) is fixed to equal 1. The model in (4.2) for continuous relations with only additive row and column effects $\{a_i, b_j\}$ and within-dyad correlation ρ is the known as social relations model (Warner et al. (1979), Wong (1982)).

The additive and multiplicative actor-specific factors $\{a_i, b_i, \boldsymbol{u}_i, \boldsymbol{v}_i\}$ are also modeled as Gaussian and independent across actors:

$$(a_i, b_i)^T \overset{\text{iid}}{\sim} \text{normal}_2(\mathbf{0}, \Sigma_{ab}); \tag{4.4}$$

$$(\boldsymbol{u}_i^T, \boldsymbol{v}_i^T)^T \overset{\text{iid}}{\sim} \text{normal}_{2k}(\mathbf{0}, \Sigma_{UV}), \quad i \in \{1, ..., n\}. \tag{4.5}$$

This model parsimoniously represents the patterns in the relations in terms of covariate effects, actor-level heterogeneity in the sending and receiving of relations via a_i and b_j, and graphical representations of the higher-order dependencies via the multiplicative effects (Hoff (2009)).

Often it will be convenient to express the model in (4.2) in matrix notation. Let X denote the $n \times n \times p$ array of regression covariates and $\langle X, \boldsymbol{\beta} \rangle$ denote the $n \times n$ matrix that results from the inner product of the regression coefficients $\boldsymbol{\beta}$ and the covariates $\boldsymbol{x}_{i,j}$ for each relation. The model in (4.2) can then be expressed

$$Z = \langle X, \boldsymbol{\beta} \rangle + \boldsymbol{a} 1_n^T + 1_n \boldsymbol{b}^T + UV^T + E, \tag{4.6}$$

where $U = [\boldsymbol{u}_1, ..., \boldsymbol{u}_n]^T$ and $V = [\boldsymbol{v}_1, ..., \boldsymbol{v}_n]^T$ are $n \times k$ matrices of actor multiplicative effects, $\boldsymbol{a} = (a_1, ..., a_n)^T$ and $\boldsymbol{b} = (b_1, ..., b_n)^T$ are $n \times 1$ vectors of actor additive effects, and E is a matrix of errors.

When the observed relations are not continuous, some model components in (4.2) are not identifiable. For instance, suppose the set $S(Y)$ is closed under row translation where $Z \in S(Y) \Rightarrow Z + \boldsymbol{w}1^T \in S(Y)$. Model effects which are constant within a row, such as $\boldsymbol{a}$, an intercept β_0 and row covariate effects $\beta_r x_i$, are non-identifiable such that, for example, $P(Z \in S(Y)|\theta, \boldsymbol{a}) = P(Z \in S(Y)|\theta, \boldsymbol{a} + \boldsymbol{w})$ for any $\boldsymbol{w} \in \mathbb{R}^n$. In these cases, there is no information in that data Y about how the relations across rows compare to one another. An example of observations Y whose set $S(Y)$ is closed under such row translation are ego-centric ranked observations. Similarly, model effects constant within a column are non-identifiable if the observations Y correspond to a set $S(Y)$ that is closed under column translation: $Z \in S(Y) \Rightarrow Z + 1\boldsymbol{w}^T \in S(Y)$.

4.3 Markov chain Monte Carlo estimation

As mentioned in the Introduction, an advantage of the latent variable models presented above is that they decompose into two parts: the portion in (4.2) which parsimoniously describes the patterns in latent relations Z and a portion in (4.1) which relates the latent relations Z to the observed relations Y. This decoupling allows the same underlying relational model to be used for a variety of types of relational observations Y. Since the class of relational observations $\mathcal{Y}$ considered includes Y that are continuous or whose set $S(Y)$ is closed under positive scalar multiplication, we require quite general estimation procedures.

Estimation for submodels of the relational data models presented above has been studied for specific types of the relational observations. For example, consider the case of continuous relations Y. Multiple estimation methods have been proposed for the model in (4.2) (where $z_{i,j} = y_{i,j}$) with only row and column additive effects $\boldsymbol{a}$ and $\boldsymbol{b}$, and witin-dyad correlation ρ: Warner et al. (1979) proposed an ANOVA estimation method and Wong (1982) introduced a maximum likelihood estimation procedure based on an expectation-maximization algorithm (EM), which was later extended by Li and Loken (2002). Gill and Swartz (2001) and Li and Loken (2002) suggested Bayesian estimation procedures for these models. Hoff (2005) also

proposed a Bayesian estimation algorithm for ordinal and binary data using Poisson and logit links for the model in (4.2) with symmetric multiplicative effects ($\boldsymbol{u}_i = \boldsymbol{v}_i$). Maximum likelihood estimation procedures for (generalized) linear and bilinear mixed effects models can be used to estimate submodels of that in (4.2) for binary and continuous data, however these procedures often rely on approximations of the likelihood, do not easily generalize to arbitrary observations $Y \in \mathcal{Y}$, and assume the data are fully observed (Schall (1991), Breslow and Clayton (1993), Wolfinger and O'Connell (1993), McGilchrist (1994), Gabriel (1998)).

When the relations Y are not continuous, the likelihood of the parameters θ given the data Y, $L(\theta|Y)$, is a generally intractable integral over a potentially complicated space depending on the observations Y. However, taking a Bayesian approach, inference for the parameters is based on the posterior distribution of the parameters given the data, $p(\boldsymbol{a}, \boldsymbol{b}, U, V, \rho, \boldsymbol{\beta}, Z, \Sigma_{ab}, \Sigma_{UV}, \sigma_e^2|Y, X)$, which although also intractable, can be approximated with samples from a Markov chain Monte Carlo (MCMC) algorithm. Such an algorithm constructs a Markov chain in the parameters whose stationary distribution is equal to the posterior distribution of the parameters given the data. A key strength of this approach is that it is able to take advantage of the decoupling in the relational model between the specific type of observed relations and the model for latent relations as function of covariates, additive and multiplicative effects. MCMC algorithms for different types of observed relations Y can be extremely similar, and estimation for potentially new types of observations in the class $\mathcal{Y}$ requires only minor algorithm modifications.

One of the most commonly used MCMC algorithms is the Gibbs sampler which iteratively samples parameters from their full conditional distribution given all other parameters and the data Y. In this section we present an MCMC algorithm based on a Gibbs sampler for the general class of relational data models presented in Section 4.2. This algorithm can be viewed as a basic sampler, which one might create as a first attempt at a Bayesian estimation procedure.

The prior distributions specified for the covariance matrices and regression coefficients are semi-conjugate such that the full conditional distribution of each covariance matrix and

the coefficients is in the same family as the prior distribution:

$$\begin{aligned} \Sigma_{ab}^{-1} &\sim \text{Wishart}(\text{I}_2, 3), \\ \Sigma_{UV}^{-1} &\sim \text{Wishart}(\text{I}_{2k}, 2k+2), \\ \sigma_e^{-2} &\sim \text{gamma}(1/2, \text{rate} = 1/2), \\ \boldsymbol{\beta} &\sim \text{normal}(\boldsymbol{\mu}_\beta, \Sigma_\beta). \end{aligned} \tag{4.7}$$

The prior distribution on ρ is uniform on the interval $[-1, 1]$. In many cases there may be prior information regarding the reciprocity parameter ρ in which case more informative priors can be considered.

Although the diagonal entires in the relational data Y are undefined and hence there are no corresponding diagonal latent relations, we propose augmenting the Markov chain with diagonal latent relations $\{z_{i,i}, i \in \{1, ..., n\}\}$ to simplify computations of the Markov chain updates. We model the diagonal entires as

$$z_{i,i} = \boldsymbol{\beta}^T x_{i,i} + a_i + b_i + \boldsymbol{u}_i^T \boldsymbol{v}_i + e_{i,i} \qquad e_{i,i} \sim \text{normal}\left(0, \sigma_e^2(1+\rho)\right). \tag{4.8}$$

The likelihood of the parameters given the data $L(\theta|Y)$ is unaltered by the model specification of the diagonal elements since they are integrated over in the model. The variance $\sigma_e^2(1+\rho)$ was selected to ease computations of the full conditional distributions of the parameters (see the appendix for details).

The full conditional distributions for all of the parameters (or subsets of them) are standard distributions which are easily sampled from, except for the full conditional distribution of ρ. Although it is possible to devise a procedure to sample from the full conditional distribution of ρ given the data and other parameters, for simplicity we instead choose to use a Metropolis-Hastings step. We propose the following set of steps for an iteration of basic MCMC algorithm based on a Gibbs sampler:

1. Update the regression coefficients and additive effects
   (a) Sample $\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}$ from its (normal) full conditional distribution.
   (b) Sample Σ_{ab} from its (inverse-Wishart) full conditional distribution.
2. Update the multiplicative effects

(a) For each multiplicative factor $i \in \{1, ..., k\}$,
- Sample $U[, i]$ from its (normal) full conditional distribution.
- Sample $V[, i]$ from its (normal) full conditional distribution.

(b) Sample Σ_{UV} from its (inverse-Wishart) full conditional distribution.

3. Update the dyad correlation

(a) Propose ρ^* from a truncated normal distribution on $[-1, 1]$.

(b) Accept ρ^* based on the appropriate Metropolis-Hastings acceptance probability.

4. Update the latent relations

(a) *If Y is not continuous*, for $i \neq j$, sample $z_{i,j}$ from its (truncated normal) full conditional distribution.

(b) For $i \in \{1, ..., n\}$, sample $z_{i,i}$ from its (normal) full conditional distribution.

5. Update the error variance

(a) *If Y is continuous*, sample σ_e^2 from its (inverse-gamma) full conditional distribution.

The full conditional distribution and Metropolis-Hastings step for ρ are presented in complete detail in the appendix. Combinations of the fives steps in the algorithm can be combined to estimate different models. For example, iterating steps 1, 3, 4, and 5 corresponds to estimating the social relations model of Warner et al. (1979) and Wong (1982) for continuous relations Y. Although the regression coefficients, additive row effects $\boldsymbol{a}$ and additive column effects $\boldsymbol{b}$ are updated together, a model can be estimated that includes any combination of the three effects.

One initially might have been inclined to update the regression coefficients, additive row effects $\boldsymbol{a}$ and additive column effects $\boldsymbol{b}$ each separately from their corresponding full conditional distributions, similar to that done in Gill and Swartz (2001). However, Gelfand et al. (1995) found that an algorithm with such steps will often suffer from poor mixing (i.e. slowly explore the parameter space) when some covariates are constant across rows and/or columns (i.e. an intercept). For this reason, simultaneous sampling of the parameters $(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b})$ from their full conditional distribution is proposed.

4.4 Improving the mixing of the Markov chain

The Markov chain proposed in Section 4.3 mixes rather poorly when the observed relations Y are not continuous due to the strong dependence between the parameters and latent relations Z in the posterior distribution. Thus, obtaining an approximation to the posterior distribution of the parameters based on an effective sample size can take days for a relational dataset with over one thousand actors. In this section we discuss three additional Metropolis-Hastings steps that can be added to the MCMC algorithm to significantly improve the Markov chain mixing and accuracy of the posterior approximation.

Two of the additional steps are group moves which propose transformations to sets of the latent relations Z and mean parameters $\boldsymbol{a}$, $\boldsymbol{b}$, U, V, and $\boldsymbol{\beta}$ along specific directions in the parameter space (Liu and Sabatti (2000)). One of these moves proposes a simultaneous multiplicative shift to the latent relations and all mean parameters and the second step moves either $\{Z, \boldsymbol{a}\}$ or $\{Z, \boldsymbol{b}\}$ together via a translation that leaves the model errors $\{e_{i,j}\}$ in (4.2) unchanged. Both of these steps are applications of Theorem 1 of Liu and Sabatti (2000) which states the following:

> *Suppose* Γ *is a locally compact group of transformations on* $\mathcal{X}$*, and* L *is its left-Haar measure. Let* π *be a probability density for* x*. If* $x \sim \pi(x)$ *and* $\gamma \in \Gamma$ *is drawn from*
>
> $$p_x(\gamma)d\gamma \propto \pi(\gamma(x))\,|J_\gamma(x)|\,L(d\gamma),$$
>
> *where* $J_\gamma(x) = det\{\partial\gamma(x)/\partial x\}$ *is the Jacobian of the transformation, then* $x' = \gamma(x)$ *follows* π*.*

The scale and translation moves proposed are discussed further below.

The third additional MCMC step is a joint update for the latent relations Z and within-dyad correlation ρ when the observed relations Y are binary. Although this last step is not applicable to the entire class of possible relations $Y \in \mathcal{Y}$ introduced in Section 4.2, it is possible modifications of it may make it suitable for additional relation types. We focus on the method for binary observations since they are most prevalent type of relational data, and hence, having an efficient sampler for such data is especially important.

4.4.1 Scale group move for $\{Z, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, U, V\}$

Suppose the observed relations $Y \in \mathcal{Y}$ are not continuous and consider the scale group $\Gamma_\gamma = \{\gamma : \gamma \in \mathbb{R}^+\}$ that acts on the latent relations and parameters $\psi = (Z, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, U, V)$ via $\gamma(\psi) = (\gamma Z, \gamma\boldsymbol{\beta}, \gamma\boldsymbol{a}, \gamma\boldsymbol{b}, \sqrt{\gamma}U, \sqrt{\gamma}V)$. Theorem 1 of Liu and Sabatti (2000) states the stationary distribution of the Markov chain is left invariant when the transformation $\psi \to \gamma(\psi)$ is performed if γ is sampled from

$$p(\gamma|Z, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, U, V, \Sigma_{ab}, \Sigma_{UV}, \rho) \propto p(\gamma Z, \gamma\boldsymbol{\beta}, \gamma\boldsymbol{a}, \gamma\boldsymbol{b}, \sqrt{\gamma}U, \sqrt{\gamma}V, \Sigma_{ab}, \Sigma_{UV}|Y)\ |J_\gamma|\ L(d\gamma), \tag{4.9}$$

where $|J_\gamma| = \gamma^{n^2+p+2n+kn}$ is the Jacobian of the transformation and $L(d\gamma) = \gamma^{-1}d\gamma$ is the unimodular left-Haar measure corresponding to the scale group Γ_γ. Letting $E = Z - \langle X, \boldsymbol{\beta}\rangle - \boldsymbol{a}1_n^T - 1_n\boldsymbol{b}^T - UV^T$ denote the residuals from the model representation in (4.6) and simplifying (4.9), we obtain

$$p(\gamma|Z, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, U, V, \Sigma_{ab}, \Sigma_{UV}, \rho) \propto \gamma^c \exp(-\gamma^2 d - \gamma f)\ d\gamma, \tag{4.10}$$

where

$$\begin{aligned}
c =& n^2 + p + 2n + kn - 1, \\
d =& \frac{1}{2}\left(\frac{\boldsymbol{e}_u^T\boldsymbol{e}_u + \boldsymbol{e}_\ell^T\boldsymbol{e}_\ell - 2\rho\boldsymbol{e}_u^T\boldsymbol{e}_\ell}{1-\rho^2} + \frac{\boldsymbol{e}_d^T\boldsymbol{e}_d}{1+\rho} + \boldsymbol{\beta}^T\Sigma_\beta^{-1}\boldsymbol{\beta}^T + \text{tr}\left[(\boldsymbol{a}, \boldsymbol{b})\Sigma_{ab}^{-1}(\boldsymbol{a}, \boldsymbol{b})^T\right]\right), \\
f =& \frac{1}{2}(\text{tr}\left[(U, V)\Sigma_{UV}^{-1}(U, V)^T\right] - 2\mu_\beta\Sigma_\beta^{-1}\boldsymbol{\beta}^T),
\end{aligned}$$

$\boldsymbol{e}_u = (e_{1,2}, ..., e_{n-1,n})$ denotes the vector of the upper triangular elements of E, $\boldsymbol{e}_\ell = (e_{2,1}, ..., e_{n,n-1})$ represents the lower triangular elements, and $\boldsymbol{e}_d = (e_{1,1}, ..., e_{n,n})$ denotes the vector of diagonal elements of E. Note that σ_e^2 is not present in (4.9)-(4.10) since this update is only applicable in the estimation of models for non-continuous relations Y.

If the specified model has no multiplicative effects ($k = 0$) and the prior mean of the regression coefficients is zero ($\mu_\beta = 0$), then there is no γ term in the exponent in (4.10) ($f = 0$). In this case, we can sample γ^2 from a gamma$((c+1)/2,\ \text{rate} = d)$ distribution and perform the transformation without affecting the stationary distribution of the chain.

However, if the model contains multiplicative effects or $\mu_\beta \neq 0$, the density of γ is not a well known distribution. Therefore, we propose using a Metropolis-Hastings step that

involves the following:

1. Generate a candidate γ from a gamma$((c+1-\lambda)/2,\ \text{rate} = d)$ distribution.
2. Accept the proposal γ and perform the transformation with probability

$$\min\{1, \gamma^{\lambda}\exp\left(-(\gamma-1)f\right)\}.$$

The probability in step 2 corresponds to the traditional Metropolis-Hastings acceptance ratio treating the current γ value as 1. Theorem 2 of Liu and Sabatti (2000) states that such a Metropolis-Hastings step leaves the stationary distribution of the Markov chain invariant if $T_{\psi}(\gamma,\tilde{\gamma}) = T_{\gamma_0^{-1}(\psi)}(\gamma\gamma_0, \tilde{\gamma}\gamma_0)$ where $T_{\psi}(\gamma,\tilde{\gamma})L(d\tilde{\gamma})$ is the Markov transition function corresponding to a current parameter γ. To show this condition is satisfied for the gamma distribution proposal, let γ denote the current value of the scale parameter and let $\tilde{\gamma}$ be a proposal from the gamma distribution in step 1 above. The acceptance probability for the proposal $\tilde{\gamma}$ is

$$\begin{aligned}
\alpha(\gamma,\tilde{\gamma}) &= \min\left\{1, \frac{p(\tilde{\gamma}|Z,\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{b},U,V,\Sigma_{ab},\Sigma_{UV},\rho)q(\gamma|Z,\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{b},U,V,\Sigma_{ab},\Sigma_{UV},\rho)}{p(\gamma|Z,\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{b},U,V,\Sigma_{ab},\Sigma_{UV},\rho)q(\tilde{\gamma}|Z,\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{b},U,V,\Sigma_{ab},\Sigma_{UV},\rho)}\right\}, \\
&= \min\left\{1, \left(\frac{\tilde{\gamma}}{\gamma}\right)^{\lambda}\exp\left(-(\tilde{\gamma}-\gamma)f\right)\right\},
\end{aligned}$$

where q represents the gamma distribution proposal density. The kernel of the Markov transition function for a current parameter value ψ can be written

$$T_{\psi}(\gamma,\tilde{\gamma}) = 2\frac{\tilde{\gamma}^{c+1-\lambda}d^{(c+1-\lambda)/2}}{\Gamma(\frac{c+1-\lambda}{2})}\exp\left(-\tilde{\gamma}^2 d\right)\alpha(\gamma,\tilde{\gamma}). \tag{4.11}$$

Therefore, it is easily seen that

$$\begin{aligned}
T_{\gamma_0^{-1}(\psi)}(\gamma\gamma_0,\tilde{\gamma}\gamma_0) &= \frac{2(\tilde{\gamma}\gamma_0)^{c+1-\lambda}\left(\frac{d}{\gamma_0^2}\right)^{\frac{(c+1-\lambda)}{2}}}{\Gamma(\frac{c+1-\lambda}{2})}\exp\left(-(\tilde{\gamma}\gamma_0)^2\Big(\frac{d}{\gamma_0^2}\Big)\right)\alpha_{\gamma_0^{-1}(x)}(\gamma\gamma_0,\tilde{\gamma}\gamma_0) \quad (4.12) \\
&= T_{\psi}(\gamma,\tilde{\gamma}).
\end{aligned}$$

We suggest $\lambda = kn$ in the proposal for γ as the proposal then corresponds to the conditional density of γ when there are no multiplicative effects. We find this value results in a high acceptance rate since the proposal distribution is extremely similar to the target full

conditional distribution and decreases the autocorrelation of the regression coefficients and additive and multiplicative effects across iterations of the chain.

4.4.2 Translation group move for $\{Z, \boldsymbol{a}\}$ and $\{Z, \boldsymbol{b}\}$

The autocorrelations of the components of Σ_{ab} tend to be high in the estimation of models for non-continuous relations Y due to the high autocorrelation in the additive effects $\boldsymbol{a}$ and $\boldsymbol{b}$. This latter autocorrelation is likely a consequence of the strong dependence between the latent relations Z and the additive effects $\boldsymbol{a}$ and $\boldsymbol{b}$ in the posterior distribution. To address this problem, we propose joint transformations to $(Z, \boldsymbol{a})$ and $(Z, \boldsymbol{b})$ by elements of the translation group $\Gamma_\tau = \{\boldsymbol{\tau} : \boldsymbol{\tau} \in \mathbb{R}^n\}$ which act via $\boldsymbol{\tau}(Z, \boldsymbol{a}) = (Z + \boldsymbol{\tau}1^T, \boldsymbol{a} + \boldsymbol{\tau})$ and via $\boldsymbol{\tau}(Z, \boldsymbol{b}) = (Z + 1\boldsymbol{\tau}^T, \boldsymbol{b} + \boldsymbol{\tau})$. Observe that the probability of transformed latent relations based on (4.2) given the transformed additive effects is equal to the probability of the original latent relations given the original effects: $p(\widetilde{Z}|\tilde{\boldsymbol{a}}, \boldsymbol{b}, U, V, \rho, \boldsymbol{\beta}) = p(Z|\boldsymbol{a}, \boldsymbol{b}, U, V, \rho, \boldsymbol{\beta})$, where $(\widetilde{Z}, \tilde{\boldsymbol{a}}) = \boldsymbol{\tau}(Z, \boldsymbol{a})$ for example.

Consider first the transformation $(Z, \boldsymbol{a}) \to \boldsymbol{\tau}(Z, \boldsymbol{a})$. The Jacobian of the transformation is $|J_\tau| = 1$ and the left-Haar measure of Γ_τ is $L(d\boldsymbol{\tau}) = d\boldsymbol{\tau}$. Therefore, the conditional density of $\boldsymbol{\tau}$ by Theorem 1 is

$$\begin{aligned} p(\boldsymbol{\tau}|Z, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, U, V, \Sigma_{ab}, \Sigma_{UV}, \rho) &\propto p(Z + \boldsymbol{\tau}1^T, \boldsymbol{\beta}, \boldsymbol{a} + \boldsymbol{\tau}, \boldsymbol{b}, U, V, \Sigma_{ab}, \Sigma_{UV}|Y) \;\; |J_\tau| \;\; L(d\boldsymbol{\tau}) \\ &\propto p(Y|Z + \boldsymbol{\tau}1^T) p(Z + \boldsymbol{\tau}1^T|\boldsymbol{\beta}, \boldsymbol{a} + \boldsymbol{\tau}, \boldsymbol{b}, U, V) p(\boldsymbol{a} + \boldsymbol{\tau}, \boldsymbol{b}|\Sigma_{ab}) d\boldsymbol{\tau} \\ &\propto \mathbb{1}\{Z + \boldsymbol{\tau}1^T \in S(Y)\} \exp\left(\frac{-1}{2\sigma^2_{\tau(a)}} (\boldsymbol{\tau} - \boldsymbol{\mu}_{\boldsymbol{\tau}(\boldsymbol{a})})^T (\boldsymbol{\tau} - \boldsymbol{\mu}_{\boldsymbol{\tau}(\boldsymbol{a})}) \right) \end{aligned} \tag{4.13}$$

where $\boldsymbol{\mu}_{\boldsymbol{\tau}(\boldsymbol{a})} = -\boldsymbol{a} + \sigma_{ab}\sigma_b^{-2}\boldsymbol{b}$ and $\sigma^2_{\tau(a)} = \sigma^2_{a|b} = \sigma^2_a - \sigma^2_{ab}/\sigma^2_b$. This is a truncated multivariate normal distribution constrained such that $Z + \boldsymbol{\tau}1^T \in S(Y)$ given Z and Y. Sampling from this distribution is difficult if the truncation bounds for elements of $\boldsymbol{\tau}$ depend on other elements of $\boldsymbol{\tau}$. However, if this is not the case, sampling from (4.13) can be achieved by sampling each element of $\boldsymbol{\tau}$ independently from a univariate truncated normal distribution. Relations Y for which the truncation bounds are strictly functions of Z and Y are those whose set $S(Y)$ is closed under multiplication of the rows by positive scalars: $Z \in S(Y) \Rightarrow$

$DZ \in S(Y)$ for any diagonal matrix D with positive diagonal elements. Examples from Section 4.2 of sets $S(Y)$ that satisfy this property are those based on binary data $B(Y)$ and ego-centric data $E(Y)$. Therefore, for observed relations Y whose set $S(Y)$ is closed under scalings of the rows, the transition $(Z, \boldsymbol{a}) \to \boldsymbol{\tau}(Z, \boldsymbol{a})$ by a $\boldsymbol{\tau}$ sampled from (4.13) is straightforward and leaves the stationary distribution of the Markov chain invariant.

Similarly, it can be shown that to perform an analogous transformation of $(Z, \boldsymbol{b})$, $\boldsymbol{\tau}$ must be sampled from

$$p(\boldsymbol{\tau}|Z,\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{b},U,V,\Sigma_{ab},\Sigma_{UV},\rho) \propto \mathbb{1}\{Z + 1\boldsymbol{\tau}^T \in S(Y)\} \exp\left(\frac{-(\boldsymbol{\tau} - \boldsymbol{\mu}_{\boldsymbol{\tau}(\boldsymbol{b})})^T(\boldsymbol{\tau} - \boldsymbol{\mu}_{\boldsymbol{\tau}(\boldsymbol{b})})}{2\sigma^2_{\tau(b)}}\right) \quad (4.14)$$

where $\boldsymbol{\mu}_{\boldsymbol{\tau}(\boldsymbol{b})} = -\boldsymbol{b} + \sigma_{ab}\sigma_a^{-2}\boldsymbol{a}$ and $\sigma^2_{\tau(b)} = \sigma^2_{b|a} = \sigma^2_b - \sigma^2_{ab}/\sigma^2_a$. Again this distribution is a multivariate truncated normal and is easily sampled from using methods for generating univariate truncation normals if the truncation bounds only depend on Y and Z, which occurs for sets $S(Y)$ closed under scalings of the columns: $Z \in S(Y) \Rightarrow ZD \in S(Y)$ for any diagonal matrix D with positive diagonal elements. Of the relations Y discussed in Section 4.2, only the binary relations set $B(Y)$ satisfies this property. Note that for binary relations Y, it may be possible to create single group translation move of the form $(Z, \boldsymbol{a}, \boldsymbol{b}) \to \boldsymbol{\nu}(Z, \boldsymbol{a}, \boldsymbol{b}) = (Z + \boldsymbol{\nu}_1 1^T + 1\boldsymbol{\nu}_2^T, \boldsymbol{a} + \boldsymbol{\nu}_1, \boldsymbol{b} + \boldsymbol{\nu}_2)$.

4.4.3 Metropolis-Hastings step for $\{Z, \rho\}$

Although the group updates were found to improve the autocorrelation of the regression coefficients $\boldsymbol{\beta}$ and actor-specific effects $\{\boldsymbol{a}, \boldsymbol{b}, U, V\}$, high autocorrelation in the chain continued to be exhibited by the within-dyad correlation parameter ρ. This is a result of both the extreme dependence between the empirical correlation of the latent relations Z and ρ and the fact that $z_{i,j}$, $z_{j,i}$, and ρ for any $i \neq j$ are always updated separately in the MCMC procedure. We now restrict our focus to improving the sampling procedure for binary relations Y for two reasons. First, binary relational data is the most common type of relational data in the social and biological sciences, and second, the latent relations Z are not constrained by one another in $S(Y)$. This limited dependence among the latent

relations in $S(Y)$ allows more efficient sampling procedures to be constructed that update large portions of Z simultaneously.

Let $\boldsymbol{z}_D = (z_{1,1}, ..., z_{n,n})$, $\boldsymbol{z}_U = (z_{1,2}, z_{1,3}, z_{2,3}, ..., z_{n-1,n})$, and $\boldsymbol{z}_L = (z_{2,1}, z_{3,1}, z_{3,2}, ..., z_{n,n-1})$ denote the vector of the diagonal elements, upper triangular elements, and lower triangular elements of Z, respectively. We propose adding a Metropolis-Hastings step to the Markov chain that proposes new values of either $\{\boldsymbol{z}_U, \rho\}$ or $\{\boldsymbol{z}_L, \rho\}$ and accepts the proposals with an appropriate probability. The proposed update for $\{\boldsymbol{z}_U, \rho\}$ proceeds as follows:

1. Generate $\tilde{\rho} \sim \text{truncated normal}_{[-1,1]}(\rho, \sigma_\rho^2)$.
2. Generate $\tilde{\boldsymbol{z}}_U$ from the full conditional truncated normal distribution of $\tilde{\boldsymbol{z}}_U$ given $\tilde{\rho}$ and all other parameters and latent relations.
3. Accept $\{\tilde{\boldsymbol{z}}_U, \tilde{\rho}\}$ with probability

$$\begin{aligned}\alpha(\{\boldsymbol{z}_U, \rho\}, \{\tilde{\boldsymbol{z}}_U, \tilde{\rho}\}) &= \min\Bigg\{1\ ,\ \frac{p(\tilde{\boldsymbol{z}}_U, \tilde{\rho}, \boldsymbol{z}_L, \boldsymbol{z}_D, U, V, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\beta}, \Sigma_{ab}, \Sigma_{UV}|Y)}{p(\boldsymbol{z}_U, \rho, \boldsymbol{z}_L, \boldsymbol{z}_D, U, V, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\beta}, \Sigma_{ab}, \Sigma_{UV}|Y)} \\ &\qquad \times \frac{p(\boldsymbol{z}_U, \rho|\tilde{\boldsymbol{z}}_U, \tilde{\rho}, \boldsymbol{z}_L, U, V, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\beta}, Y)}{p(\tilde{\boldsymbol{z}}_U, \tilde{\rho}|\boldsymbol{z}_U, \rho, \boldsymbol{z}_L, U, V, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\beta}, Y)}\Bigg\} \\ &= \min\Bigg\{1\ ,\ \frac{\Big((1+\tilde{\rho})^{-n/2}\exp\Big(\frac{-\boldsymbol{e}_d^2}{2(1+\tilde{\rho})}\Big)\Big)}{\Big((1+\rho)^{-n/2}\exp\Big(\frac{-\boldsymbol{e}_d^2}{2(1+\rho)}\Big)\Big)} \\ &\qquad \times \frac{\left(\prod\left(\Phi\Big(\frac{\boldsymbol{t}_u^{(U)} - \tilde{\boldsymbol{\mu}}_{z_U}}{\tilde{\sigma}_z}\Big) - \Phi\Big(\frac{\boldsymbol{t}_l^{(U)} - \tilde{\boldsymbol{\mu}}_{z_U}}{\tilde{\sigma}_z}\Big)\right)\right)\left(\Phi\Big(\frac{1-\rho}{\sigma_\rho}\Big) - \Phi\Big(\frac{-1-\rho}{\sigma_\rho}\Big)\right)}{\left(\prod\left(\Phi\Big(\frac{\boldsymbol{t}_u^{(U)} - \boldsymbol{\mu}_{z_U}}{\sigma_z}\Big) - \Phi\Big(\frac{\boldsymbol{t}_l^{(U)} - \boldsymbol{\mu}_{z_U}}{\sigma_z}\Big)\right)\right)\left(\Phi\Big(\frac{1-\tilde{\rho}}{\sigma_\rho}\Big) - \Phi\Big(\frac{-1-\tilde{\rho}}{\sigma_\rho}\Big)\right)}\Bigg\},\end{aligned}$$

   where $\boldsymbol{t}_l^{(U)}$ and $\boldsymbol{t}_u^{(U)}$ are the lower and upper truncation bounds of $\boldsymbol{z}_U$ based on $S(Y)$, $\{\boldsymbol{\mu}_{z_U}, \sigma_z^2\}$ and $\{\tilde{\boldsymbol{\mu}}_{z_U}, \tilde{\sigma}_z^2\}$ are the means and variance of $\boldsymbol{z}_U$ in the full conditional distribution based on ρ and $\tilde{\rho}$ respectively, and $\boldsymbol{e}_d$ are the diagonal elements of E defined by (4.6). See the appendix for more details.

This update proposes moves to the correlation ρ and empirical correlation of the latent relations Z simultaneously in the same direction. Thus, more drastic movements in the

parameters are likely in this joint update than when the latent relations Z and correlation ρ are updated individually conditional on each other. An analogous update is proposed for $\{\boldsymbol{z}_L, \rho\}$. For non-continuous, non-binary relations Y, a similar Metropolis-Hasting step could be constructed for any subset of the relations Z, say $\tilde{\boldsymbol{z}}$, whose constraints in $S(Y)$ depend only on Y and elements of Z not in $\tilde{\boldsymbol{z}}$.

4.4.4 New MCMC algorithm

The tth iteration of the new proposed MCMC algorithm for non-continuous relations Y is comprised of the following steps.

1. Update the regression coefficients and additive effects
   (a) Sample $\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}$ from its full conditional (normal) distribution.
   (b) **If t is even and $S(Y)$ is closed under multiplication of its rows by positive scalars, update $\{Z, a\}$ using the translation group move in Section 4.4.2.**
   (c) **If t is odd and $S(Y)$ is closed under multiplication of its columns by positive scalars, update $\{Z, b\}$ using the translation group move in Section 4.4.2.**
   (d) Sample Σ_{ab} from its full conditional (inverse-Wishart) distribution.
2. Update the multiplicative effects
   (a) For each multiplicative factor $i \in \{1, ..., k\}$,
      - Sample $U[, i]$ from its full conditional (normal) distribution.
      - Sample $V[, i]$ from its full conditional (normal) distribution.
   (b) Sample Σ_{UV} from its full conditional (inverse-Wishart) distribution.
3. Update the dyad correlation and latent relations
   (a) **If t is odd and Y is binary,**
      i. **Update $\{z_U, \rho\}$ using the Metropolis-Hastings procedure in Section 4.4.3.**
      ii. **Update $\{z_L, \rho\}$ using the Metropolis-Hastings procedure in Section 4.4.3.**

iii. For $i \in \{1, ..., n\}$, sample $z_{i,i}$ from its full conditional (normal) distribution.

(b) **If t is even or Y is not binary,**

i. Propose ρ^* from a truncated normal distribution on $[-1, 1]$.

ii. Accept ρ^* based on the appropriate Metropolis-Hastings acceptance probability.

iii. For $i \neq j$, sample $z_{i,j}$ from its full conditional (truncated normal) distribution.

iv. For $i \in \{1, ..., n\}$, sample $z_{i,i}$ from its full conditional (normal) distribution.

4. **Update the scale of $\{\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{U}, \boldsymbol{V}\}$**

(a) **Update $\{\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{U}, \boldsymbol{V}\}$ using the scale group move in Section 4.4.1.**

The bolded steps are the additions to the original algorithm. Although steps 1(a) and 1(b) are alternated every other iteration in the chain, they could both be performed in each iteration. However we find alternating the steps strikes a balance between allowing the parameters to move more quickly without dramatically increasing computation time.

The two different updates of the latent relations Z and within-dyad correlation ρ in 3(a) and 3(b) could also be performed each iteration, however these are some of the more expensive steps in the sampler. We proposing alternating the steps since their advantages complement one another. The updates of $\{\boldsymbol{z}_U, \rho\}$ and $\{\boldsymbol{z}_L, \rho\}$ propose larger changes to the empirical correlation of the latent relations and ρ than that which is observed when the parameters are sampled from their respective full conditional distributions. However, when the Metropolis-Hastings proposal is rejected, Z and ρ remain the same. By including the Gibbs step that samples the parameters from their respective full conditional distributions, the parameters move at least slightly every other iteration of the chain.

4.5 Simulation study: Quantifying the improvement in mixing

To evaluate the improvement in mixing of the Markov chain using the new MCMC algorithm, we simulated 100 binary datasets under the model in (4.2) with two multiplicative effects ($k = 2$) and five regression effects ($p = 5$), which included an intercept, a row covariate $\{x_i^{(r)}\}$, a column covariate $\{x_i^{(c)}\}$, and two dyadic covariates $\{x_{i,j}^{(1)}, x_{i,j}^{(2)}\}$. Datasets

of size 50 and 500 were generated from the model:

$$
\begin{aligned}
y_{i,j} &= 1[z_{i,j} > 0] \\
z_{i,j} &= \beta_0 + \beta_{d1} x_{i,j}^{(1)} + \beta_{d2} x_{i,j}^{(2)} + \beta_r x_i^{(r)} + \beta_c x_j^{(c)} + a_i + b_j + \boldsymbol{u}_i^T \boldsymbol{v}_j + e_{i,j}, \qquad (4.15) \\
(e_{i,j}, e_{j,i})^T &\overset{\text{iid}}{\sim} \text{normal}_2\Big(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\Big), \qquad (a_i, b_i)^T \overset{\text{iid}}{\sim} \text{normal}_2\Big(\mathbf{0}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\Big),
\end{aligned}
$$

where $\rho = 0.7$ and $(\beta_0, \beta_{d1}, \beta_{d2}, \beta_r, \beta_c) = (2, 1, 1, 1, 1)$. The covariates $\{x_i^{(r)}, x_j^{(c)}, x_{i,j}^{(1)}, x_{i,j}^{(2)}\}$ and multiplicative effects $\{\boldsymbol{u}_i, \boldsymbol{v}_i\}$ were generated from a standard normal distribution. For each simulated dataset, a Markov chain was run for 5,000 iterations of burn-in followed by 50,000 additional iterations. The computation time for ten iterations of the new sampler was 22% longer than ten iterations of the original sampler for a relational dataset of size 500.

We compared the mixing of the samplers using the effective sample size of the regression and covariance parameters over the 50,000 saved iterations. The effective sample size, calculated using the 'coda' package in `R`, estimates the number of independent samples from the posterior distribution needed to obtain an approximation to the posterior distribution that is as precise as that based on the given MCMC samples. Figure 4.1 shows the effective sample size for the original MCMC algorithm (ESS) and the effective sample size for the new MCMC (ESS_{NEW}) for the five regression coefficients $\{\beta_0, \beta_r, \beta_c, \beta_{d1}, \beta_{d2}\}$ the additive effects covariance parameters $\{\sigma_a^2, \sigma_b^2, \sigma_{ab}\}$, and the within-dyad correlation ρ. In general the new MCMC procedure improves the efficiency of the original sampler by approximately a factor of two for most regression coefficients and the variances of the additive effects. We see larger improvements for the smaller datasets when $n = 50$ compared to the larger datasets, except for the dyadic regression coefficients and within-dyad correlation. The ESS_{NEW} values displayed above the parameter labels illustrates that these parameters have the largest autocorrelation from one iteration of the chain to the next in the new sampler. This implies the autocorrelations for β_{d1}, β_{d2}, and ρ were extremely high in the original sampler for large datasets and most improved upon (on a multiplicative scale) in the new sampler.

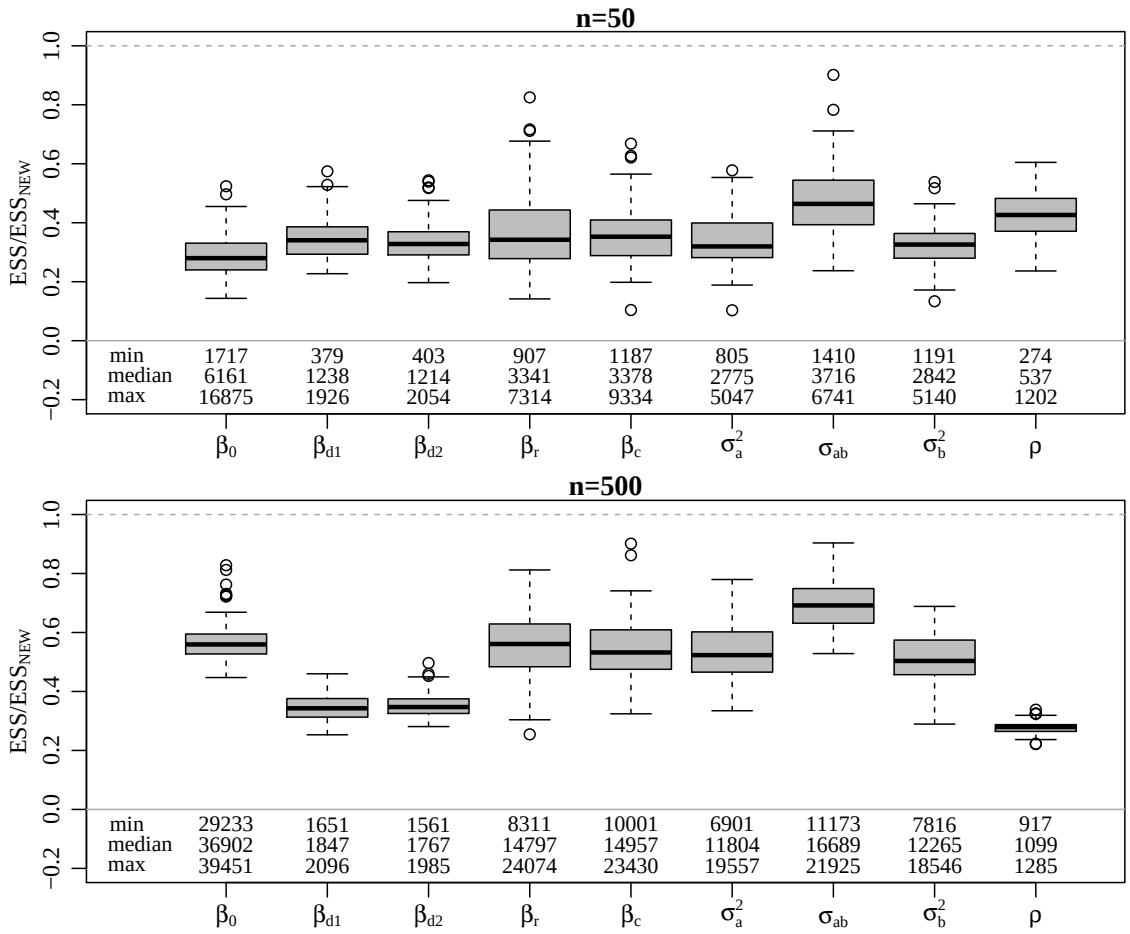

Figure 4.1: Comparison of the effective sample sizes for the original MCMC sampler (ESS) and the new sampler (ESS_{NEW}) for 100 simulations with $n = 50$ and $n = 500$. The minimum, median, and maximum ESS_{NEW} values for the 100 simulations are provided above each parameter.

4.6 Mean-field variational approximation

The MCMC estimation procedure can require over a day to obtain a sufficient approximation of the posterior distribution for relational datasets whose sizes exceed 1000 actors. In cases such as this, an attractive alternative to MCMC based estimation is a variational Bayesian procedure which approximates the posterior distribution of the parameters given the data

with a simpler, often analytically tractable, distribution that satisfies a set of specified constraints (Beal (2003), Jordan et al. (1999); see Jaakkola (2001), Ormerod and Wand (2010) for good reviews). A key advantage of this approach is that frequently inference can be performed analytically on the posterior distribution approximation taking advantage of its simpler, often closed, form and eliminating the need for any sampling procedures.

In this section, we describe a mean-field variational Bayesian approach to parameter estimation for non-continuous relations Y, which approximates the posterior distribution of the parameters and latent relations, $p(\boldsymbol{a}, \boldsymbol{b}, U, V, \rho, \boldsymbol{\beta}, Z, \Sigma_{ab}, \Sigma_{UV}|Y, X)$, with a function $q(\boldsymbol{a}, \boldsymbol{b}, U, V, \rho, \boldsymbol{\beta}, Z, \Sigma_{ab}, \Sigma_{UV})$ that factorizes over sets of the parameters and latent relations (Beal and Ghahramani (2003)). Specifically, we constrain the set of possible approximating distributions q to those that factor as

$$q = q(\boldsymbol{a}, \boldsymbol{b}, U, V, \rho, \boldsymbol{\beta}, Z, \Sigma_{ab}, \Sigma_{UV}) = q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\beta})q(U, V)q(\Sigma_{ab})q(\Sigma_{UV})q(\rho)q(Z). \tag{4.16}$$

This enforces independence in the posterior approximation between parameters in separate components. The variational Bayesian approximation q is defined as the distribution that satisfies (4.16) and minimizes the Kullback-Liebler (KL) divergence between q and the true posterior distribution p. (Equation (4.16) presents the variational posterior approximation for non-continuous relations Y; the variational approximation for continuous relations has the same form as (4.16) where $q(Z)$ is replaced by $q(\sigma_e^2)$. See the appendix for more details.)

Let $\theta = \{\boldsymbol{a}, \boldsymbol{b}, U, V, \rho, \boldsymbol{\beta}, \Sigma_{ab}, \Sigma_{UV}\}$ denote the set of model parameters. The KL divergence between the variational approximation q and true posterior p can be written:

$$\begin{aligned}
\mathrm{KL}(q, p) &= \int_{\theta, Z} \log\left[\frac{q(\theta, Z)}{p(\theta, Z|Y)}\right] q(\theta, Z) d(\theta, Z) \\
&= \log\left[p(Y)\right] - \int_{\theta, Z} \log\left[\frac{p(\theta, Z, Y)}{q(\theta, Z)}\right] q(\theta, Z) d(\theta, Z) \\
&= \log\left[p(Y)\right] - \mathrm{L}(q, p).
\end{aligned}$$

Therefore, minimizing $\mathrm{KL}(q, p)$ is equivalent to maximizing $\mathrm{L}(q, p)$, which is often referred to as the free energy function. Note the $\mathrm{L}(q, p)$ provides a lower bound on the log marginal probability of the data under the model in (4.1) and (4.2).

The posterior approximation q is estimated using a coordinate ascent algorithm, whereby $\mathrm{L}(q, p)$ is iteratively maximized as a function of one of the components of q, say $q(\theta_i)$, treating

all other components as fixed. Taking the functional derivative of $L(q, p)$ with respect to $q(\theta_i)$, it can be shown that the maximizing distribution $q(\theta_i)$ is given by

$$\log[q(\theta_i)] = E_{q/q(\theta_i)}\Big[\log(p(Y, \theta))\Big] + c, \text{or equivalently,}$$
$$q(\theta_i) \propto \exp\left(E_{q/q(\theta_i)}\Big[\log(p(\theta_i|Y, \theta_{-i}))\Big]\right),$$

where $q/q(\theta_i)$ is the posterior distribution approximation q without component $q(\theta_i)$. This optimal distribution is of the form of the full conditional distribution of θ_i given the data Y and all other parameters θ_{-i}. Thus, updates for the components of q which correspond to Gibbs sampling steps in the MCMC algorithm are tractable analytically. This motivated the decomposition of q in (4.16), which mirrors that of the MCMC procedure in Section 4.3. The coordinate ascent algorithm for estimation of the variational posterior approximation q for the class of relational data models presented in Section 4.2 iterates the following sequence of steps until a desired convergence level is achieved:

1. Update $q(\phi) = q(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b})$; calculate expectations $\mathrm{E}_{q(\phi)}[\cdot]$ needed analytically.
2. Update $q(\Sigma_{ab})$; calculate expectations $\mathrm{E}_{q(\Sigma_{ab})}[\cdot]$ needed analytically.
3. Update $q(\rho)$; estimate expectations $\mathrm{E}_{q(\rho)}[\cdot]$ needed using MCMC.
4. Update $q(U, V)$; estimate expectations $\mathrm{E}_{q(U,V)}[\cdot]$ needed using MCMC.
5. Update $q(\Sigma_{UV})$; calculate expectations $\mathrm{E}_{q(\Sigma_{UV})}[\cdot]$ needed analytically.
6. *If Y is continuous (and hence Z is observed):*
   - Update $q(\sigma_e^2)$; calculate expectations $\mathrm{E}_{q(\sigma_e^2)}[\cdot]$ needed analytically.
7. *If Y is not continuous:*
   - Update $q(Z)$; estimate expectations $\mathrm{E}_{q(Z)}[\cdot]$ needed using MCMC.

The stopping criterion is often based on the relative change in the parameters of the component distributions $q(\theta_i)$ or the relative change of $\mathrm{L}(q, p)$. The expectations of functions of ρ, U, V, and Z are not available analytically since the corresponding component distributions in the approximation q, $\{q(\rho), q(U, V), q(Z)\}$, are not standard distributions. Thus, at each iteration of the algorithm we create a Markov chain for each of the parameters and from the

samples obtain estimates of the necessary expectations. See the appendix for the derivation of each update and the required expectations.

There is a trade-off between the accuracy and computational complexity of the variational approximation q as a function of the degree to which q factorizes. If q is assumed to factor over each individual parameter, the updates and expectation computations with respect to the components $q(\theta_i)$ are often straightforward. However, enforcing independencies between the parameters in the posterior approximation q can severely degrade the accuracy of the approximation if there is a large amount of dependence between the parameters in the true posterior.

4.7 Simulation study: Accuracy of the variational approximation

The posterior distribution of the parameters given the data is frequently summarized using a measure of centrality, such as the mean or mode of each parameter distribution, and a measure of uncertainty, such as confidence intervals. Centrality measures of parameters based on the variational approximations of the posterior often closely match that of the true posterior, however, it is known that the variational approximation will frequently underestimate the uncertainty (MacKay (2003), appendix A in Rue et al. (2009)). In some cases, it has been shown that this underestimation is negligible and inference based on the variational approximation is nearly the same as inference based on the true posterior (see Ormerod and Wand (2010) for a pathological example). In this section, we investigate how well the proposed variational estimate q approximates the true posterior distribution in the case of continuous and binary observed relations Y.

We first consider a model for continuous relations Y that contains an intercept, four covariates, additive row and column effects, and two multiplicative effects:

$$y_{i,j} = \beta_0 + \beta_{d1}x_{i,j}^{(1)} + \beta_{d2}x_{i,j}^{(2)} + \beta_r x_i^{(r)} + \beta_c x_j^{(c)} + a_i + b_j + \boldsymbol{u}_i^T\boldsymbol{v}_j + e_{i,j}, \qquad (4.17)$$
$$(e_{i,j}, e_{j,i})^T \overset{\text{iid}}{\sim} \text{normal}_2\Big(\mathbf{0}, \sigma_e^2\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\Big).$$

A relational dataset for 25 actors was simulated from the above model, where the additive effects $\{\boldsymbol{a},\ \boldsymbol{b}\}$, multiplicative effects $\{U,\ V\}$, and regression covariates $\{x_{i,j}^{(1)},\ x_{i,j}^{(2)},\ x_i^{(r)},\ x_j^{(c)}\}$ were sampled from a standard normal distribution. The $\boldsymbol{\beta}$ values were set at

$(\beta_0, \beta_{d1}, \beta_{d2}, \beta_r, \beta_c) = (2, 1, .25, 1, 1)$ and the errors $e_{i,j}$ were simulated from the bivariate normal distribution with $\sigma_e^2 = 1$ and $\rho = 0.7$. The variational Bayesian estimation procedure outlined above was able to obtain an approximation to the posterior distribution within a few seconds. An estimate of the true posterior distribution was obtained by from running the improved MCMC procedure described in Section 4.4 for 5,000 iterations of burn-in followed by an additional 20,000 iterations and thinned to every 5th iteration. This resulted in 4,000 samples from the posterior distribution and took approximately 15 minutes on a standard laptop. The effective sample size of all regression coefficients and covariance parameters was greater than 1,600.

Figure 4.2 compares the true posterior distribution estimate from the MCMC procedure to the variational approximation q. The left plot in the figure shows that the posterior mean estimates of the additive and multiplicative effects are well approximated by the variational posterior, while the right plot shows the same is true for the posterior means of the regression coefficients and covariance parameters. Also in the right plot we see that the uncertainty in the regression and covariance parameters based on the posterior distribution is only slightly underestimated by the variational posterior. This suggests that when the observed relations are continuous, the variational posterior is a reasonable alternative as a basis for inference.

The second model we consider is for binary relations and similar to that (4.17) except with no multiplicative effects:

$$\begin{aligned} y_{i,j} &= \mathbb{1}\{z_{i,j} > 0\}, \\ z_{i,j} &= \beta_0 + \beta_{d1} x_{i,j}^{(1)} + \beta_{d2} x_{i,j}^{(2)} + \beta_r x_i^{(r)} + \beta_c x_j^{(c)} + a_i + b_j + e_{i,j}, \\ (e_{i,j}, e_{j,i})^T &\overset{\text{iid}}{\sim} \text{normal}_2\Big(\mathbf{0}, \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)\Big). \end{aligned} \tag{4.18}$$

We created a binary dataset for 25 actors using the additive effects, covariates, and errors simulated for the normal relations and set $(\beta_0, \beta_{d1}, \beta_{d2}, \beta_r, \beta_c) = (2, 1, .25, 1, 1)$. An estimate of the variational posterior distribution q was obtained in 30 seconds. The MCMC estimation procedure was run for 5,000 iterations of burn-in, followed by an additional 75,000 iterations and thinned every 25th iteration to obtain 3,000 samples from the posterior distribution. The Markov chain took approximately 30 minutes and the effective sample sizes of all regression coefficients and covariance parameters was greater than 1,200.

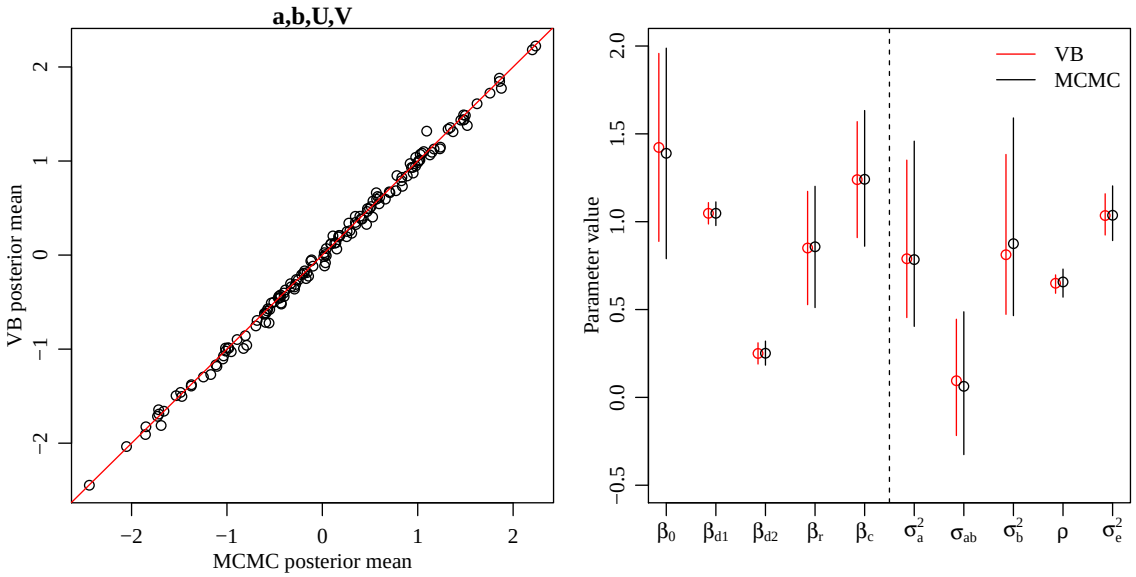

Figure 4.2: Summary of the posterior distribution based on the variational Bayesian approximation (VB) and the MCMC procedure for the model in (4.17) for continuous relations Y. The left plot shows the posterior mean estimates for the additive and multiplicative effects. The right plot shows the posterior mean estimates and the corresponding 95% confidence intervals for the regression coefficients and select covariance parameters.

Figure 4.3 shows summaries of the posterior distribution of the parameters based on samples from the MCMC and the variational approximation q. The variational approximation appears to underestimate the magnitude of the posterior means of the additive effects, regression coefficients, and covariance parameters. More importantly, the left plot shows the variational posterior severely underestimates the uncertainty in the regression coefficients and variance parameters. This result is consistent with that found in Consonni and Marin (2007) (see the follow-up in Armagan and Zaretzki (2011)) for the probit (binary regression) model, which is a submodel of (4.18) where $\rho = 0$ and there are no additive effects $\{\boldsymbol{a}, \boldsymbol{b}\}$. Although the variational approximation is a significantly faster alternative to estimating the posterior distribution using MCMC, its shrunken estimates of the posterior means and drastic underestimation of the parameter uncertainty makes it unsuitable as a basis for

inference for non-continuous observation Y. We could attempt to improve the variational approximation by assuming fewer independencies between parameter (i.e. coarser factorization of q), however this would increase the complexity of the coordinate-ascent algorithm as it would require additional and/or more complicated MCMC procedures to obtain the necessary expectations.

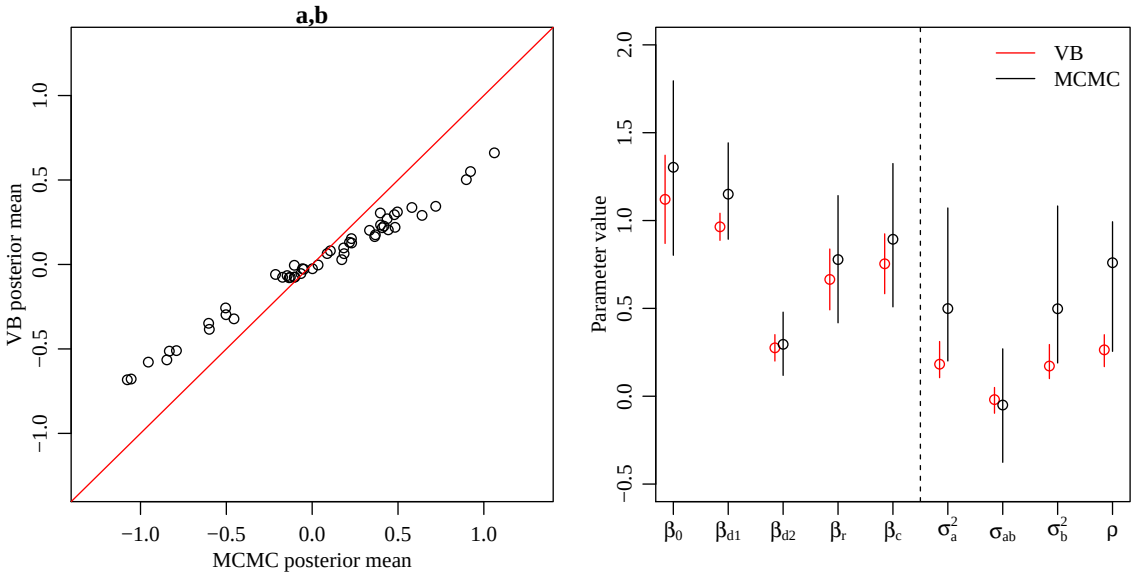

Figure 4.3: Summary of the posterior distributions based on the variational Bayesian approximation (VB) and the MCMC procedure for the model in (4.18) for binary relations Y. The left plot shows the posterior mean estimates for the additive and multiplicative effects. The right plot shows the posterior mean estimates and the corresponding 95% confidence intervals.

4.8 Discussion

In this chapter, we discussed Bayesian estimation for a general class of relational data models. We presented a vanilla MCMC procedure and improved its efficiency using group move updates and a Metropolis-Hasting step for the latent relations Z and within-dyad correlation ρ. The estimation framework presented is extremely general as it can accommodate

variety of relational types, including continuous, binary, ranked, and censored relations. We investigated a mean-field variational Bayesian estimation procedure and found that while its approximation to the posterior distribution was satisfactory when the relations Y are continuous, the approximation severely underestimates parameter uncertainty for non-continuous relations.

Hamiltonian Monte Carlo (HMC) is an additional Monte Carlo estimation method that has recently received much attention for its ability to improve the efficiency of MCMC procedures. The HMC methods developed by Pakman and Paninski (2012) for sampling from a truncated multivariate normal distribution and extended by Kalaitzis and Silva (2013) for Gaussian copulas could potentially be implemented here in place of the updates of the latent variables from their full conditional distributions. A big strength of these methods is that they are exact, such that the Hamiltonian dynamics can be expressed algebraically, removing the need for specification of parameters associated with approximating the dynamics. The computational cost of such an HMC step would increase with the amount of restrictions on the latent relations Z in $S(Y)$ and with the dimension of the network. Thus computations will likely be most burdensome for ranked observations Y, where the constraints on each latent relation is a function of other latent relations. Investigating the usefulness of HMC methods for updating all latent relations jointly in the MCMC is a topic of future work.

Although we discussed methods for non-symmetric datasets Y which contain directed measures of the relations, the models and estimation procedures presented can be extended to datasets of undirected relations, where $y_{i,j} = y_{j,i}$ for all i, j. Code for all estimation methods is available at the author's website.

Chapter 5

CONCLUSIONS AND FUTURE WORK

In this dissertation, we presented methods for modeling heterogeneity within and between matrices and arrays. We proposed a submodel of the array normal model of Hoff (2011) that is able to model dependence within modes of array using factor analytic structured covariance matrices. As a result of the reduction in the number of covariance parameters, maximum likelihood estimates of the SFA submodel exist for arrays where the array normal maximum likelihood estimates do not exist. We also presented a unified approach to the analysis of a relational data matrix and actor-specific attributes that included a test and joint model for dependencies between the dyad and actor-level data. Finally we discussed Bayesian estimation of a general class of relational data models, describing methods for improving Markov chain Monte Carlo procedures and illustrating the accuracy of a mean-field variational approximation for both continuous and non-continuous relational data. We now present some extensions of our methods and ideas for future work.

5.1 Alternative low-dimensional covariance matrix parameterizations

In Chapter 2, we proposed reducing the number of parameters in the array normal model of Hoff (2011) by modeling the covariance matrices with factor analytic structure. Such structure is appealing as it provides a low-dimensional positive definite approximation of a covariance matrix. However, other low-dimensional parameterizations are possible. For example, for an array mode whose indices have a natural ordering (temporal, ordinal, etc.), an autoregressive or banded covariance structure may better approximate the dependence within the mode and contain fewer parameters than the factor analytic approximation. To illustrate this possibility, we consider the four-way array of mortality data discussed in Chapter 2.

Figure 5.1 shows the residual correlations across time period lags and age group lags

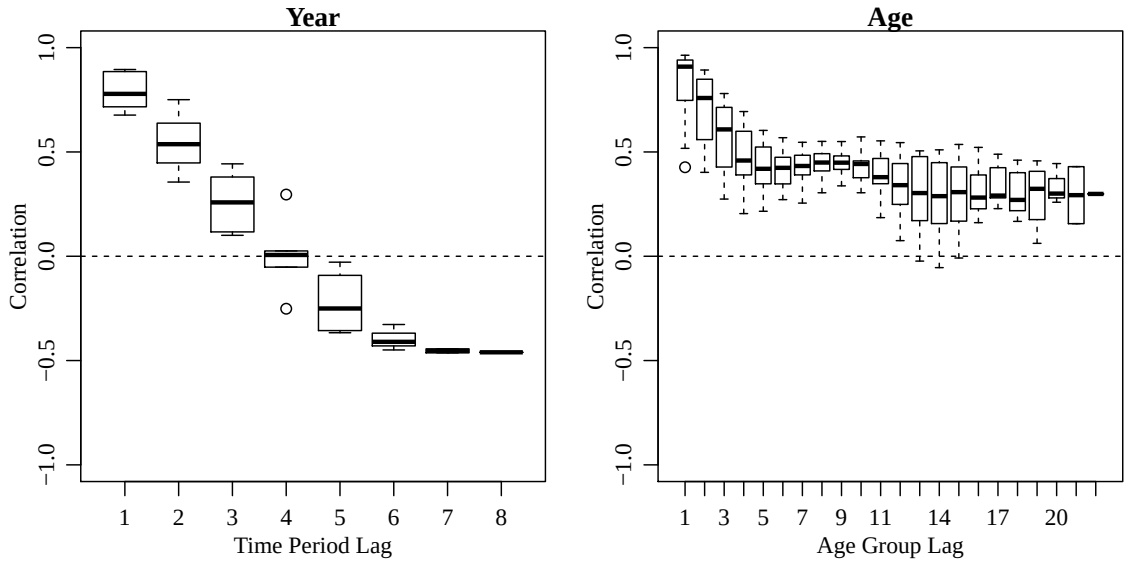

Figure 5.1: Correlations between the time periods and age groups in the mortality data residual array associated with the ordinary least squares fit of the mean model (2.17) grouped by lag.

after the ordinary least squares fit of the mean model in (2.17) is subtracted from the data array. The monotonic decreasing trend with lag in the time period plot suggests the mode correlation matrix could possibly be well approximated with a banded covariance structure. Note that an autoregressive structure with order one would be unable to model the negative correlations seen between time periods at lags greater than four. A less clear pattern is seen in the correlations between the age groups. An exchangeable covariance structure could be assumed, where all age groups pairs have the same correlation, however this would greatly underestimate the correlation seen at small lags. Therefore, while a banded parameterized covariance structure may be appropriate for time periods, an alternative superior parameterization to the factor-analytic structure is less clear for the age groups.

Estimation procedures for Separable Factor Analysis models where some covariance matrices have alternative low-dimensional structures is of immediate future work. An extension of the mode rank testing procedure in Section 2.3.3 to incorporate other covariance param-

eterizations, such as banded, exchangeable, and autoregressive structures, could also be an interesting topic of investigation.

5.2 Relational and attribute data over time and/or across relationship types

The methods presented in Chapter 3 allow for testing and modeling dependencies between a single cross-sectional dataset of actor relations and their attributes. Although these methods are able to address questions about association between attributes and relations, much research in the social and biological sciences is focused on determining causal relationships using network and attribute data measured at multiple time points. As an example, using data on adolescent friendship network and their health behaviors, sociologists are interested in determining whether students' friendships impact their health behaviors or whether health behaviors cause changes in their friendship networks (Bauman and Ennett (1996)). Answering these questions could potentially help guide adolescent drug intervention programs in schools. Therefore, extending the methodology presented in Chapter 3 to test for casual relationships between relational and attribute data, as well as jointly model such data over time would be worth consideration.

Others have considered the problem of modeling networks over time. A popular approach has been to embed models for cross-sectional network data into a continuous (Holland and Leinhardt (1977), Frank (1991)) or discrete time (Robins and Pattison (2001)) Markov process. Snijders (2005) and Snijders (2006) discuss methods for a continuous time stochastic process where network dynamics are either edge-oriented or actor-oriented. Hanneke and Xing (2007), Hanneke et al. (2010) and Krivitsky and Handcock (2013) introduced extensions to exponential family random-graph models (ERGMs) based on discrete time processes. Generalizations of the latent space models of Hoff et al. (2002) that allow actor latent positions to change over time were introduced by Sarkar and Moore (2005), and extended further in Sarkar et al. (2007). A drawback of these methods is that most of them currently only accommodate binary relational data (although theoretically could be modified to accommodate other types) and all make inference for the network conditional on the attributes. Thus, they are unable to asses any effect of the network on attributes and unable to make predictions simultaneously for missing network and attribute information.

Snijders et al. (2007) and Steglich et al. (2010) proposed methods for modeling the co-evolution of networks and behaviors using actor-oriented processes whereby actors dictate changes in their behaviors and outgoing ties. These methods focus on binary data and rely on method of moments estimation procedures as likelihood based inference is computationally prohibitive. A key advantage of the latent variable models is that they provide geometrically interpretable representations of the network using multiplicative effects. Extending the latent variable network and attribute models discussed in Chapter 3 would provide geometric representations of the network over time, allow for testing of casual relationships between the network and attributes, and provide predictions simultaneously for missing network and attribute information.

BIBLIOGRAPHY

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), "Mixed membership stochastic blockmodels," *The Journal of Machine Learning Research*, 9, 1981–2014.

Allen, G. I. and Tibshirani, R. (2010), "Transposable regularized covariance models with an application to missing data imputation," *Annals of Applied Statistics*, 4, 764–790.

Anderson, T. W. (1951), "Estimating linear restrictions on regression coefficients for multivariate normal distributions," *The Annals of Mathematical Statistics*, 327–351.

Anderson, T. W. and Gupta, S. D. (1964), "A Monotonicity Property of the Power Functions of Some Tests of the Equality of Two Covariance Matrices," *Annals of Mathematical Statistics*, 35, 1059–1063.

Armagan, A. and Zaretzki, R. L. (2011), "A note on mean-field variational approximations in Bayesian probit models," *Computational Statistics & Data Analysis*, 55, 641–643.

Austin, A., Linkletter, C., and Wu, Z. (2013), "Covariate-defined latent space random effects model," *Social Networks*.

Bauman, K. E. and Ennett, S. T. (1996), "On the importance of peer influence for adolescent drug use: commonly neglected considerations," *Addiction*, 91, 185–198.

Beal, M. J. (2003), "Variational algorithms for approximate Bayesian inference," Ph.D. thesis, University of London.

Beal, M. J. and Ghahramani, Z. (2003), "The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures," *Bayesian Statistics*, 7.

Brass, W. (1971), "On the scale of mortality," *Biological aspects of demography*, 69–110.

Breslow, N. E. and Clayton, D. G. (1993), "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, 88, 9–25.

Browne, M. W. (1984), "The decomposition of multitrait-multimethod matrices," *British Journal of Mathematical and Statistical Psychology*, 37, 1–21.

Butland, G., Peregrín-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., et al. (2005), "Interaction network containing conserved and essential protein complexes in Escherichia coli," *Nature*, 433, 531–537.

Carter, L. R. and Lee, R. D. (1992), "Modeling and forecasting US sex differentials in mortality," *International Journal of Forecasting*, 8, 393–411.

Chiou, J.-M. and Müller, H.-G. (2009), "Modeling Hazard Rates as Functional Data for the Analysis of Cohort Lifetables and Mortality Forecasting," *Journal of the American Statistical Association*, 104, 572–585.

Christakis, N. A. and Fowler, J. H. (2007), "The Spread of Obesity in a Large Social Network over 32 Years," *New England Journal of Medicine*, 357, 370–379.

Coale, A. and Demeny, P. (1966), *Regional Model Life Tables and Stable Populations*, Princeton University Press.

Congdon, P. (1993), "Statistical Graduation in Local Demographic Analysis and Projection," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 156, 237–270.

Consonni, G. and Marin, J.-M. (2007), "Mean-field variational approximate Bayesian inference for latent variable models," *Computational Statistics & Data Analysis*, 52, 790–798.

Currie, I. D., Durban, M., and Eilers, P. H. (2004), "Smoothing and forecasting mortality rates," *Statistical Modelling*, 4, 279–298.

Dawid, A. P. (1981), "Some Matrix-Variate Distribution Theory: Notational Considerations and a Bayesian Application," *Biometrika*, 68, 265–274.

de la Haye, K., Robins, G., Mohr, P., and Wilson, C. (2010), "Obesity-related behaviors in adolescent friendship networks," *Social Networks*, 32, 161–167.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000), "A multilinear singular value decomposition," *SIAM Journal on Matrix Analysis and Applications*, 21, 1253–1278.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 1–38.

Diaconis, P., Goel, S., and Holmes, S. (2008), "Horseshoes in Multidimensional Scaling and Local Kernel Methods," *The Annals of Applied Statistics*, 2, 777–807.

Erbring, L. and Young, A. A. (1979), "Individuals and Social Structure: Contextual Effects as Endogenous Feedback," *Sociological Methods & Research*, 7, 396–430.

Faust, K. (1988), "Comparison of Methods for Positional Analysis: Structural and General Equivalence," *Social Networks*, 10, 313–341.

Felipe, A., Guillen, M., and Nielsen, J. P. (2001), "Longevity studies based on kernel hazard estimation," *Insurance: Mathematics and Economics*, 28, 191–204.

Fellows, I. and Handcock, M. S. (2012), "Exponential-family Random Network Models," *arXiv preprint arXiv:1208.0121*.

Fowler, J. H. and Christakis, N. A. (2008), "Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study," *British Medical Journal*, 337.

Frank, O. (1991), "Statistical analysis of change in networks," *Statistica Neerlandica*, 45, 283–293.

Frank, O. and Strauss, D. (1986), "Markov graphs," *Journal of the American Statistical Association*, 81, 832–842.

Gabriel, K. R. (1998), "Generalised bilinear regression," *Biometrika*, 85, 689–700.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995), "Efficient parametrisations for normal linear mixed models," *Biometrika*, 82, 479–488.

Genton, M. G. (2007), "Separable approximations of space-time covariance matrices," *Environmetrics*, 18, 681–695.

Gill, P. S. and Swartz, T. B. (2001), "Statistical analyses for round robin interaction data," *Canadian Journal of Statistics*, 29, 321–331.

Handcock, M. S. (2003), *Statistical models for social networks: Inference and degeneracy*, Committee on Human Factors, National Research Council, The National Academies Press, vol. 126, pp. 302–322.

Hanneke, S., Fu, W., and Xing, E. P. (2010), "Discrete temporal models of social networks," *Electronic Journal of Statistics*, 4, 585–605.

Hanneke, S. and Xing, E. P. (2007), "Discrete temporal models of social networks," in *Statistical network analysis: Models, issues, and new directions*, Springer, pp. 115–125.

Harris, K., Halpern, C., Whitsel, E., Hussey, J., Tabor, J., Entzel, P., and Udry, J. (2009), "The National Longitudinal Study of Adolescent Health: Research Design," .

Hartmann, M. (1987), "Past and recent attempts to model mortality at all ages," *Journal of Official Statistics*, 3, 19–36.

Heligman, L. and Pollard, J. (1980), "The age pattern of mortality," *Journal of the Institute of Actuaries*, 107, 49–80.

Hoff, P. D. (2005), "Bilinear Mixed-Effects Models for Dyadic Data," *Journal of the American Statistical Association*, 100, 286–295.

— (2007), "Model averaging and dimension selection for the singular value decomposition," *Journal of the American Statistical Association*, 102, 674–685.

— (2009), "Multiplicative latent factor models for description and prediction of social networks," *Computational & Mathematical Organization Theory*, 15, 261–272.

— (2011), "Separable covariance arrays via the Tucker product, with applications to multivariate relational data," *Bayesian Analysis*, 6, 179–196.

Hoff, P. D., Fosdick, B. K., Volfovsky, A., and Stovel, K. (2012), "Likelihoods for fixed rank nomination networks," *Technical Report 608, Department of Statistics, University of Washington*.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002), "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, 97, 1090–1098.

Holland, P. W. and Leinhardt, S. (1977), "A dynamic model for social networks," *Journal of Mathematical Sociology*, 5, 5–20.

Human Mortality Database (2011), *University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de*.

Hunter, D. R. and Handcock, M. S. (2006), "Inference in Curved Exponential Family Models for Networks," *Journal of Computational and Graphical Statistics*, 15, 565–583.

Jaakkola, T. S. (2001), "10 Tutorial on Variational Approximation Methods," *Advanced mean field methods: theory and practice*, 129.

Jennrich, R. and Bobinson, S. (1969), "A Newton-Raphson algorithm for maximum likelihood factor analysis," *Psychometrika*, 34, 111–123.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999), "An introduction to variational methods for graphical models," *Machine learning*, 37, 183–233.

Jöreskog, K. G. (1967), "Some contributions to maximum likelihood factor analysis," *Psychometrika*, 32, 443–482.

Kalaitzis, A. and Silva, R. (2013), "Flexible Sampling for the Gaussian Copula Extended Rank Likelihood Model," *arXiv preprint arXiv:1306.2685*.

Kass, R. E. and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 90, 928–934.

Kiers, H. A. L. (2000), "Towards a standardized notation and terminology in multiway analysis," *Journal of Chemometrics*, 14, 105–122.

Kim, M. and Leskovec, J. (2011), "Modeling social networks with node attributes using the multiplicative attribute graph model," *arXiv preprint arXiv:1106.5053*.

— (2012), "Multiplicative attribute graph model of real-world networks," *Internet Mathematics*, 8, 113–160.

Kiuru, N., Burk, W. J., Laursen, B., Salmela-Aro, K., and Nurmi, J.-E. (2010), "Pressure to drink but not to smoke: disentangling selection and socialization in adolescent peer networks and peer groups," *Journal of Adolescence*, 33, 801–812.

Kolda, T. G. and Bader, B. W. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500.

Krivitsky, P. N. (2012), "Exponential-family random graph models for valued networks," *Electronic Journal of Statistics*, 6, 1100–1128.

Krivitsky, P. N. and Handcock, M. S. (2013), "A separable model for dynamic networks," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Kroonenberg, P. M. (2008), *Applied multiway data analysis*, Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons].

Lawley, D. (1940), "The estimation of factor loadings by the method of maximum likelihood," *Proceedings of the Royal Society of Edinburgh*, 60, 64–82.

Lee, R. D. and Carter, L. R. (1992), "Modeling and Forecasting U.S. Mortality," *Journal of the American Statistical Association*, 87, 659–671.

Li, H. and Loken, E. (2002), "A unified theory of statistical analysis and inference for variance component models for dyadic data," *Statistica Sinica*, 12, 519–536.

Li, N. and Lee, R. (2005), "Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method," *Demography*, 42, 575–594.

Liu, C. and Rubin, D. (1998), "Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data," *Stat. Sinica*, 8, 729–747.

Liu, J. S. and Sabatti, C. (2000), "Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation," *Biometrika*, 87, 353–369.

MacKay, D. J. (2003), *Information theory, inference and learning algorithms*, Cambridge university press.

Manceur, A. M. and Dutilleul, P. (2013), "Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion," *Journal of Computational and Applied Mathematics*, 239, 37–49.

Mardia, K., Kent, J., and Bibby, J. (1979), *Multivariate Analysis*, Academic Press.

Marsden, P. V. and Friedkin, N. E. (1993), "Network Studies of Social Influence," *Sociological Methods & Research*, 22, 127–151.

Martínez-Ruiz, F., Mateu, J., Montes, F., and Porcu, E. (2010), "Mortality risk assessment through stationary space-time covariance functions," *Stochastic Environmental Research and Risk Assessment*, 24, 519–526.

McFarland, D. and Brown, D. (1973), "Social distance as a metric: a systematic introduction to smallest space analysis," in *Bonds of Pluralism: The Form and Substance of Urban Social Networks*, ed. E. Laumann, New York: Wiley, pp. 213–253.

McGilchrist, C. (1994), "Estimation in generalized mixed models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61–69.

McNown, R. and Rogers, A. (1989), "Forecasting Mortality: A Parameterized Time Series Approach," *Demography*, 26, 645–660.

Mode, C. and Busby, R. (1982), "An eight-parameter model of human mortality - The single decrement case," *Bulletin of Mathematical Biology*, 44, 647–659.

Muirhead, R. (1982), *Aspects of multivariate statistical theory*, Wiley series in probability and mathematical statistics: Probability and mathematical statistics, Wiley.

Murray, C. J. L., Ferguson, B. D., Lopez, A. D., Guillot, M., Salomon, J. A., and Ahmad, O. (2003), "Modified Logit Life Table System: Principles, Empirical Validation, and Application," *Population Studies*, 57, 165–182.

Nowicki, K. and Snijders, T. A. B. (2001), "Estimation and prediction for stochastic block-structures," *Journal of the American Statistical Association*, 96, 1077–1087.

Oort, F. J. (1999), "Stochastic three-mode models for mean and covariance structures," *British Journal of Mathematical and Statistical Psychology*, 52, 243–272.

Ormerod, J. T. and Wand, M. P. (2010), "Explaining Variational Approximations," *The American Statistician*, 64, 140–153.

Pakman, A. and Paninski, L. (2012), "Exact Hamiltonian Monte Carlo for Truncated Multivariate Gaussians," *arXiv preprint arXiv:1208.4118*.

Perlman, M. D. and Olkin, I. (1980), "Unbiasedness of Invariant Tests for Manova and Other Multivariate Problems," *Annals of Statistics*, 8, 1326–1341.

Pettitt, A. (1982), "Inference for the linear model using a likelihood based on ranks," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 234–243.

Reinsel, G. and Velu, R. (1998), *Multivariate Reduced-Rank Regression: Theory and Applications*, Lecture Notes in Statistics, Springer.

Renshaw, A. and Haberman, S. (2003a), "Lee-Carter mortality forecasting: a parallel generalized linear modelling approach for England and Wales mortality projections," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52, 119–137.

— (2003b), "Lee-Carter mortality forecasting with age-specific enhancement," *Insurance: Mathematics and Economics*, 33, 255–272.

— (2003c), "On the forecasting of mortality reduction factors," *Insurance: Mathematics and Economics*, 32, 379–401.

Renshaw, A., Haberman, S., and Hatzopoulos, P. (1996), "The Modelling of Recent Mortality Trends in United Kingdom Male Assured Lives," *British Actuarial Journal*, 2, 449–477.

Robertson, D. and Symons, J. (2007), "Maximum likelihood factor analysis with rank-deficient sample covariance matrices," *Journal of Multivariate Analysis*, 98, 813–828.

Robins, G. and Pattison, P. (2001), "Random graph models for temporal processes in social networks*," *Journal of Mathematical Sociology*, 25, 5–41.

Robins, G., Pattison, P., and Elliott, P. (2001), "Network models for social influence processes," *Psychometrika*, 66, 161–189.

Rubin, D. and Thayer, D. (1982), "EM algorithms for ML factor analysis," *Psychometrika*, 47, 69–76.

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 319–392.

Sarkar, P. and Moore, A. W. (2005), "Dynamic social network analysis using latent space models," *ACM SIGKDD Explorations Newsletter*, 7, 31–40.

Sarkar, P., Siddiqi, S. M., and Gordon, G. J. (2007), "A latent space approach to dynamic embedding of co-occurrence data," in *International Conference on Artificial Intelligence and Statistics*, pp. 420–427.

Schall, R. (1991), "Estimation in generalized linear models with random effects," *Biometrika*, 78, 719–727.

Schweinberger, M. (2011), "Instability, Sensitivity, and Degeneracy of Discrete Exponential Families," *Journal of the American Statistical Association*, 106, 1361–1370.

Siler, W. (1983), "Parameters of mortality in human populations with widely varying life spans," *Statistics in Medicine*, 2, 373–380.

Snijders, T. A. (2005), "Models for longitudinal network data," *Models and methods in social network analysis*, 1, 215–247.

— (2006), "Statistical methods for network dynamics," in *Proceedings of the XLIII Scientific Meeting, Italian Statistical Society*, Padova: CLEUP, Italy, no. 1994, pp. 281–296.

Snijders, T. A., Steglich, C. E., and Schweinberger, M. (2007), "Modeling the co-evolution of networks and behavior," *Longitudinal models in the behavioral and related sciences*, 41–71.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006), "New Specifications for Exponential Random Graph Models," *Sociological Methodology*, 36, 99–153.

Spearman, C. (1904), ""General Intelligence," Objectively Determined and Measured," *The American Journal of Psychology*, 15, 201–292.

Steglich, C., Snijders, T. A., and Pearson, M. (2010), "Dynamic networks and behavior: Separating selection from influence," *Sociological Methodology*, 40, 329–393.

Stein, M. L. (2005), "SpaceTime Covariance Functions," *Journal of the American Statistical Association*, 100, 310–321.

Tucker, L. R. (1966), "Some mathematical notes on three-mode factor analysis," *Psychometrika*, 31, 279–311.

United Nations (1982), "Model Life Tables for Developing Countries," *Population Studies*, 77.

Warner, R., Kenny, D., and Stoto, M. (1979), "A New Round-Robin Analysis of Variance for Social Interaction Data," *Journal of Personality and Social Psychology*, 37, 1742–1757.

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

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

Wolfinger, R. and O'Connell, M. (1993), "Generalized linear mixed models a pseudo-likelihood approach," *Journal of Statistical Computation and Simulation*, 48, 233–243.

Wong, G. Y. (1982), "Round-Robin Analysis of Variance via Maximum Likelihood," *Journal of the American Statistical Association*, 77, 714–724.

Zhao, J. H., Yu, P. L., and Jiang, Q. (2008), "ML estimation for factor analysis: EM or non-EM?" *Statistics and Computing*, 18, 109–123.

Appendix A

SEPARABLE FACTOR ANALYSIS

A.1 Sampling Λ and D from their full conditional distributions

Let Λ_i^* be the proposed value of Λ that results from A(i-ii). The acceptance probability for this proposal is

$$\begin{aligned}\alpha(\Lambda_i^*, \Lambda_i) &= \frac{p(\Lambda_i^*|Y, \Lambda_{-i}, D, \Sigma)p(\Lambda_i|\Lambda_i^*, D, \Sigma, \Lambda_{-i}, Y)}{p(\Lambda_i|Y, \Lambda_{-i}, D, \Sigma)p(\Lambda_i^*|\Lambda_i, D, \Sigma, \Lambda_{-i}, Y)} \\ &= \frac{p(Y|\Lambda_i^*, \Lambda_{-i}, D, \Sigma)p(\Lambda_i^*|D_i)p(\Lambda_i|\Lambda_i^*, D, \Sigma, \Lambda_{-i}, Y)}{p(Y|\Lambda_i, \Lambda_{-i}, D, \Sigma)p(\Lambda_i|D_i)p(\Lambda_i^*|\Lambda_i, D, \Sigma, \Lambda_{-i}, Y)}\end{aligned}$$

The proposal probability can be written

$$\begin{aligned}p(\Lambda_i^*|\Lambda_i, D, \Sigma,\Lambda_{-i}, Y) &= \int p(\Lambda_i^*, Z^i|\Lambda_i, D, \Sigma, \Lambda_{-i}, Y)dZ^i \\ &= \int p(\Lambda_i^*|Z^i, D, \Sigma, \Lambda_{-i}, Y)p(Z^i|\Lambda_i, D, \Sigma, \Lambda_{-i}, Y)dZ^i \\ &= p(\Lambda_i^*|D, \Sigma, \Lambda_{-i}, Y)\int \frac{p(Z^i|\Lambda_i^*, D, \Sigma, \Lambda_{-i}, Y)}{p(Z^i|D, \Sigma, \Lambda_{-i}, Y)}p(Z^i|\Lambda_i, D, \Sigma, \Lambda_{-i}, Y)dZ^i \\ &= \frac{p(Y|\Lambda_i^*, D, \Sigma, \Lambda_{-i})p(\Lambda_i^*, D, \Sigma, \Lambda_{-i})}{p(D, \Sigma, \Lambda_{-i}, Y)} \cdot c(\Lambda_i, \Lambda_i^*|D, \Sigma, \Lambda_{-i}, Y) \\ &= \frac{p(Y|\Lambda_i^*, D, \Sigma, \Lambda_{-i})p(\Lambda_i^*|D_i)p(D)p(\Sigma)p(\Lambda_{-i}|D_{-i})}{p(D, \Sigma, \Lambda_{-i}, Y)} \cdot c(\Lambda_i, \Lambda_i^*|D, \Sigma, \Lambda_{-i}, Y)\end{aligned}$$

where $c(\Lambda_i, \Lambda_i^*|D, \Sigma, \Lambda_{-i}, Y)$ represents the integral, which is symmetric in Λ_i and Λ_i^*. Plugging the last expression into the acceptance probability, we obtain $\alpha(\Lambda_i^*, \Lambda_i) = 1$. Analogous logic can be used to show the acceptance probability for a proposed D_i from B(i-ii) is also one.

A.2 Alternative extension of factor analysis to arrays

An alternative extension of factor analysis to arrays is motivated by the latent variable representation of single model factor analysis in (2.2). First consider extending the single mode factor model to estimate relationships among the rows and the columns of a matrix

X, by writing

$$X_{p\times n} = \Lambda_1 Z \Lambda_2^T + D_1 E D_2^T = Z_{k_1\times k_2} \times \{\Lambda_1, \Lambda_2\} + E_{p\times n} \times \{D_1, D_2\}, \tag{A.1}$$

where $W \times \{A_1, ..., A_K\}$ indicates the first mode of the K-way array W is left multiplied by A_1, the second mode is left multiplied by A_2, etc. As in SFA, Λ_i is $(m_i \times k_i)$, and D_i is diagonal, however we only consider $k_i > 0$. If Z and E have the same distributional properties as in (2.2), the covariance matrix of X is $\text{Cov}[\text{vec}(X)] = (\Lambda_2\Lambda_2^T \otimes \Lambda_1\Lambda_1^T) + (D_2^2 \otimes D_1^2)$.

The analogous model for a K-way array Y has the following equivalent representations:

$$Y = Z \times \{\Lambda_1, ..., \Lambda_K\} + E \times \{D_1, ..., D_K\} \tag{A.2}$$

$$\text{Cov}[\text{vec}(Y)] = (\Lambda_K\Lambda_K^T \otimes ... \otimes \Lambda_1\Lambda_1^T) + (D_K^2 \otimes ... \otimes D_1^2)$$

where Z is $(k_1 \times ... \times k_K)$, E is $(m_1 \times ... \times m_K)$, and these again have the same distributional properties as in (2.2). The second moment of a matricization of the array is written

$$\text{E}[Y_{(i)}Y_{(i)}^T] = \alpha_i \Lambda_i \Lambda_i^T + \gamma_i D_i^2 \qquad \alpha_i = \prod_{j\neq i} \text{tr}(\Lambda_j\Lambda_j^T) \qquad \gamma_i = \prod_{j\neq i} \text{tr}(D_j^2).$$

Observe that the second moment has k_i-factor analytic structure and will be unstructured if $\Lambda_i = D_i = \Sigma_i^{1/2}$, where $\Sigma_i^{1/2}$ is any non-singular $m_i \times m_i$ matrix.

The model in (A.2) has many similarities to the higher-order singular value decomposition (HOSVD) (Tucker (1966), De Lathauwer et al. (2000)). The HOSVD states that any K-way array Y can be written $Y = G \times \{U_1, ..., U_K\}$, where G is an all-orthogonal core matrix of the same dimension as Y and U_i is $(m_i \times m_i)$ satisfying $U_i^T U_i = I$ for $i \in \{1, ..., K\}$. The i^{th} slice of Y in the j^{th} mode is that which results by setting the j^{th} index of Y equal to i. An array is considered to be of reduced rank if slices of the core array G are zero. A K-way array of rank $(r_1, ..., r_K)$ can be expressed by the HOSVD with a core matrix G of dimension $(r_1 \times ... \times r_K)$ where each U_i is of dimension $(m_i \times r_i)$. The alternative array factor model in (A.2) can be written as $Y = M + \widetilde{E}$ where $M = Z \times \{...\}$ is the mean and common factor portion, and $\widetilde{E} = E \times \{...\}$ is the error component. The mean structure of this model resembles the HOSVD if Z is viewed as a core array. Although Z is not

all-orthogonal and Λ_i and Σ_i do not satisfy $\Lambda_i^T \Lambda_i = (\Sigma_i^{1/2})^T \Sigma_i^{1/2} = I$, M can be rewritten using the HOSVD to obtain a new Z, Λ_i and Σ_i that satisfy the constraints. The error component $\widetilde{E}$ is then interpreted as accounting for variation in Y which is not captured by the reduced rank approximation M.

The covariance model in (A.2) is a submodel of the single mode $(\prod k_i)$-factor model for vec(Y) since the covariance matrix is comprised of a reduced rank matrix plus a diagonal matrix. SFA is equivalent to this alternative extension if zero or one mode is specified with factor analytic structure. A drawback of the HOSVD and this alternative extension is that mode parameter values are difficult to interpret and cannot be considered independently of parameters in other modes. For this reason, we chose to focus on SFA as the primary extension of factor analysis to arrays.

Appendix B

JOINT NETWORK AND ATTRIBUTE MODEL

B.1 Bayesian estimation procedure

In this section we outline the Bayesian estimation procedure used to obtain parameter estimates for the joint attribute and network model in (3.15). This procedure is extremely similar to that implemented in the 'amen' package in the statistical computing program `R`. We present the simple case here where the observed network Y is continuous, there are no regression terms in the network model, and there is no missing data. For details on accommodating non-continuous network data see Hoff et al. (2012) and for including regression terms see Hoff (2005).

Model -

$$\begin{aligned} y_{i,j} &= \mu + a_i + b_j + \boldsymbol{u}_i^T \boldsymbol{v}_j + e_{i,j}, \\ (e_{i,j}, e_{j,i})^T &\overset{\text{iid}}{\sim} \text{normal}_2\left(\mathbf{0}, \sigma_e^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \\ \left(\boldsymbol{x}_i^T, a_i, b_i, \boldsymbol{u}_i^T, \boldsymbol{v}_i^T\right)^T &\overset{\text{iid}}{\sim} \text{normal}_{p+2+2k}\left(\mathbf{0}, \Sigma_{XN}\right) \end{aligned}$$

Prior distributions -

$$\begin{aligned} \sigma_e^{-2} &\sim \text{gamma}(1/2, 1/2) \\ \rho &\sim \text{uniform}(-1, 1) \\ \Sigma_{XN}^{-1} &\sim \text{Wishart}\left(p+2+2k+1, \begin{pmatrix} \Sigma_{X0}^{-1} & 0 \\ 0 & \mathrm{I}_{2+2k} \end{pmatrix}\right) \end{aligned}$$

Markov chain Monte Carlo algorithm -

Given initial values of all latent variables $\{\boldsymbol{a}, \boldsymbol{b}, U, V\}$ and parameters $\{\Sigma_{XN}, \rho, \sigma_e^2\}$, the algorithm proceeds as follows:

1. Sample $\boldsymbol{a}, \boldsymbol{b} | Y, X, U, V, \Sigma_{XN}, \rho, \sigma_e^2$ (normal).

2. Sample $\Sigma_{XN} | Y, X, \boldsymbol{a}, \boldsymbol{b}, U, V, \rho, \sigma_e^2$ (inverse-Wishart).

3. Update ρ using a Metropolis-Hastings step with proposal $\rho^*|\rho \sim$ truncated normal$_{[-1,1]}(\rho, \sigma_\rho^2)$;

4. Sample $\sigma_e^2|Y, X, \boldsymbol{a}, \boldsymbol{b}, U, V, \rho, \Sigma_{XN}$ (inverse-gamma).

5. For each latent factor i:

   - Sample $U[,i]|Y, X, \boldsymbol{a}, \boldsymbol{b}, U[,-i], V, \rho, \sigma_e^2, \Sigma_{XN}$ (normal);
   - Sample $V[,i]|Y, X, \boldsymbol{a}, \boldsymbol{b}, U, V[,-i], \rho, \sigma_e^2, \Sigma_{XN}$ (normal).

Although the estimation algorithm is not constructed based on the unique parameterization of the model, each sample of network factors from the posterior distribution can be transformed using the covariance matrix Σ_{XN} sample to represent a sample from (3.16). Inference for the relative likeliness of parameter values is based on the posterior distribution over the parameter equivalence classes associated with representations congruent with (3.16).

Appendix C

ESTIMATION FOR RELATIONAL DATA MODELS

C.1 Markov chain Monte Carlo calculations

C.1.1 Regression and additive effects

Full conditional of $\{\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}\}$

We propose sampling from the full conditional distribution $p(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}|\Sigma_{ab}, Z, Y, U, V, \Sigma_{UV}, \rho, \sigma_e^2)$ by first sampling $\boldsymbol{\beta}$ from

$$p(\boldsymbol{\beta}|\Sigma_{ab}, Z, Y, U, V, \Sigma_{UV}, \rho, \sigma_e^2)$$

and then sampling $(\boldsymbol{a}, \boldsymbol{b})$ from

$$p(\boldsymbol{a}, \boldsymbol{b}|\boldsymbol{\beta}, \Sigma_{ab}, Z, Y, U, V, \Sigma_{UV}, \rho, \sigma_e^2).$$

Analytically expressing the parameters for these distributions and computing them is complicated due to the within-dyad correlation ρ. Thus, we propose standardizing (i.e. decorrelating) the latent relations via a transformation and updating the regression coefficients and additive effects based on the transformed variables.

First, let $\widetilde{Z}$ represent the latent relations in a model with no within-dyad correlation (i.e. $\rho = 0$) and no multiplicative effects ($k = 0$):

$$\tilde{z}_{i,j} = \boldsymbol{\beta}\tilde{x}_{i,j} + \tilde{a}_i + \tilde{b}_j + \tilde{e}_{i,j}, \quad \tilde{e}_{i,j} \sim \text{normal}(0, 1) \quad \text{for all } i, j. \tag{C.1}$$

The model for the vector of $\tilde{z}_{i,j}$'s is written

$$\tilde{\boldsymbol{z}} = \text{vec}(\widetilde{Z}) = \widetilde{X}\boldsymbol{\beta} + G\begin{pmatrix} \tilde{\boldsymbol{a}} \\ \tilde{\boldsymbol{b}} \end{pmatrix} + \tilde{\boldsymbol{e}}, \quad \text{where} \quad G = \begin{pmatrix} (1_n^T \otimes \mathrm{I}_n) \\ (\mathrm{I}_n \otimes 1_n^T) \end{pmatrix}^T,$$

and $\widetilde{X}$ is the appropriate $n^2 \times p$ design matrix. The marginal distribution of $\boldsymbol{\beta}$ given $\widetilde{Z}$ and

$\widetilde{\Sigma}_{ab}$, after integrating out $(\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}})$ has the form

$$p(\boldsymbol{\beta}|\widetilde{Z}, \widetilde{\Sigma}_{ab}) \propto \int p(\widetilde{Z}|\boldsymbol{\beta}, \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}}) p(, \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}}|\widetilde{\Sigma}_{ab}) p(\boldsymbol{\beta}) d\mu(\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}})$$

$$\propto \exp\left(\frac{-1}{2}\left[\boldsymbol{\beta}^T\left(\widetilde{X}^T\left(\mathrm{I}_{n^2} + GVG^T\right)\widetilde{X} + \Sigma_\beta^{-1}\right)\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\left(\widetilde{X}\left(\mathrm{I}_{n^2} + GVG^T\right)\tilde{\boldsymbol{z}} + \Sigma_\beta^{-1}\mu_\beta\right)\right]\right),$$

where $V = (\widetilde{\Sigma}_{ab}^{-1} + n\mathrm{I}_2)^{-1} \otimes \mathrm{I}_n - (\widetilde{\Sigma}_{ab}^{-1} + n\mathrm{I}_2)^{-1} \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right) (\widetilde{\Sigma}_{ab}^{-1} + n1_2 1_2^T)^{-1} \otimes 1_n 1_n^T$

and $\widetilde{\Sigma}_{ab}$ is the covariance matrix of $(\tilde{a}_i, \tilde{b}_i)$. Thus, the distribution of $\boldsymbol{\beta}|\widetilde{Z}, \widetilde{\Sigma}_{ab}$ is multivariate normal$(\tilde{\boldsymbol{m}}_\beta, \widetilde{V}_\beta)$ with

$$\tilde{\boldsymbol{\mu}}_\beta = \tilde{\Sigma}_\beta\left(\widetilde{X}\left(\mathrm{I}_{n^2} + GVG^T\right)\tilde{\boldsymbol{z}} + \Sigma_\beta^{-1}\mu_\beta\right), \quad \widetilde{\Sigma}_\beta = \left(\widetilde{X}^T\left(\mathrm{I}_{n^2} + GVG^T\right)\widetilde{X} + \Sigma_\beta^{-1}\right)^{-1}.$$

The full conditional distribution of $(\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}})$ given $\boldsymbol{\beta}$, $\widetilde{Z}$, and $\widetilde{\Sigma}_{ab}$ is also multivariate normal with mean $\tilde{\boldsymbol{m}}_{ab} = \widetilde{V}_{ab} G^T(\tilde{\boldsymbol{z}} - \widetilde{X}\boldsymbol{\beta})$ and covariance $\widetilde{V}_{ab} = \left(G^T G + \widetilde{\Sigma}_{ab}^{-1} \otimes I_n\right)^{-1}$. Therefore sampling $(\boldsymbol{\beta}, \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}})$ from its full conditional distribution is accomplished by sampling $\boldsymbol{\beta} \sim$ normal$(\tilde{\boldsymbol{m}}_\beta, \widetilde{V}_\beta)$ and then sampling $(\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}}) \sim$ normal$(\tilde{\boldsymbol{m}}_{ab}, \widetilde{V}_{ab})$. Note that the mean and covariance parameters for these distributions can be computed efficiently using matrix algebra to avoid large matrix creation and inversion.

Now consider the original latent relations model in (4.2) with within-dyad correlation ρ and multiplicative effects. To sample from the full conditional distribution of $(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b})$, the latent relations are transformed by subtracting the multiplicative effect and multiplying by the inverse square root of the dyad correlation matrix

$$\Sigma_e^{-1/2} = \sigma_e^{-1} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1/2} = \begin{pmatrix} \gamma & \delta \\ \delta & \gamma \end{pmatrix}, \tag{C.2}$$

where $\gamma = \sigma_e^{-1}((1+\rho)^{-1/2} + (1-\rho)^{-1/2})/2$ and $\delta = \sigma_e^{-1}((1+\rho)^{-1/2} - (1-\rho)^{-1/2})/2$. Using the matrix formulation in (4.6) of the latent relations model, the transformed variables $\widetilde{Z}$ are written

$$\widetilde{Z} = \gamma(Z - UV^T) + \delta(Z - UV^T)^T$$

and follow the uncorrected latent relations model in (C.1) with an appropriately transformed design matrix $\widetilde{X}$, transformed additive effects $\tilde{a}_i = \gamma a_i + \delta b_i$ and $\tilde{b}_i = \gamma b_i + \delta a_i$, and transformed covariance matrix $\widetilde{\Sigma}_{ab} = \Sigma_e^{-1/2} \Sigma_{ab} \Sigma_e^{-1/2}$. Thus to sample from the full conditional of $(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b})$, we propose the following procedure:

1. sample $\boldsymbol{\beta}$ from $p(\boldsymbol{\beta}|\widetilde{Z}, \widetilde{\Sigma}_{ab})$;

2. sample $(\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}})$ from $p(\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}}|\boldsymbol{\beta}, \widetilde{Z}, \widetilde{\Sigma}_{ab})$;

3. transform $(\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}})$ back to $(\boldsymbol{a}, \boldsymbol{b})$.

Modeling the diagonal elements with variance $\sigma_e^2(1+\rho)$ in (4.8) ensures the transformed variables $\widetilde{Z}$ all have variance one.

Full conditional of Σ_{ab}

The full conditional distribution of Σ_{ab} is proportional to $p(\boldsymbol{a}, \boldsymbol{b}|\Sigma_{ab})p(\Sigma_{ab})$, where $p(\Sigma_{ab})$ is the prior distribution specified in (4.7). The full conditional distribution of Σ_{ab} is inverse-Wishart(ν_{ab}, S_{ab}), where $\nu_{ab} = n + 3$ and $S_{ab} = (\boldsymbol{a}, \boldsymbol{b})^T(\boldsymbol{a}, \boldsymbol{b}) + I_2$ (parameterized such that $\mathrm{E}(\Sigma_{ab}) = S_{ab}/(\nu_{ab} - 3)$).

C.1.2 Multiplicative effects

Full conditional of $\{U, V\}$

Each multiplicative effect is sampled separately from its corresponding full conditional distribution. As with the regression coefficients and additive effects, the full conditional distribution is most easily expressed in terms of a transformation of the latent relations Z which removes the within-dyad correlation. With the additive effects, the underlying model was unchanged by the transformation of the latent relations. This is not the case for the multiplicative effects so we start by discussing a slightly different model for the latent relations which has only a single multiplicative effect:

$$\tilde{z}_{i,j} = \gamma u_i v_j + \delta u_j v_i + \tilde{e}_{i,j}, \quad \tilde{e}_{i,j} \sim \text{normal}(0, 1) \quad \text{for all } i, j. \tag{C.3}$$

where γ and δ are that defined in (C.2). In matrix form similar to (4.6), this model is written $\widetilde{Z} = \gamma \boldsymbol{u}\boldsymbol{v}^T + \delta \boldsymbol{v}\boldsymbol{u}^T + E$. If we denote the mean and covariance matrix of the multivariate normal distribution on the multiplicative effects $(\boldsymbol{u}, \boldsymbol{v})$ by $(\boldsymbol{\mu}_U, \boldsymbol{\mu}_V)$ and $\widetilde{\Sigma}_{UV} \otimes I_n$, respectively. It can easily be shown that the conditional distribution of $\boldsymbol{u}$ given $\boldsymbol{v}$ and

$\widetilde{Z}$ is multivariate normal$(\boldsymbol{m}_U, V_U)$, where

$$\boldsymbol{m}_U = V_U \left(\left(\gamma \widetilde{Z} + \delta \widetilde{Z}^T + \tilde{\sigma}_{U|V}^{-2} \tilde{\sigma}_{UV} \tilde{\sigma}_V^{-2} \mathrm{I}_n \right) \boldsymbol{v} + \tilde{\sigma}_{U|V}^{-2} \boldsymbol{\mu}_U \right),$$
$$V_U = \left(2\gamma\delta \boldsymbol{v}\boldsymbol{v}^T + \mathrm{I}_n \left((\gamma^2 + \delta^2) \boldsymbol{v}^T\boldsymbol{v} + \tilde{\sigma}_{U|V}^{-2} \right) \right)^{-1},$$

$\tilde{\sigma}_{UV}$ and $\tilde{\sigma}_V^2$ are components of $\widetilde{\Sigma}_{UV}$, and $\tilde{\sigma}_{U|V}^2 = \tilde{\sigma}_U^2 - \tilde{\sigma}_{UV}^2/\tilde{\sigma}_V^2$. The conditional distribution of $\boldsymbol{v}$ given $\boldsymbol{u}$ and $\widetilde{Z}$ is derived analogously. Thus, sampling from the full conditional of a single multiplicative effect from the model in (C.3) is straightforward.

Now consider the original latent relations model in (4.2). To sample from the full conditional distribution of the kth multiplicative effect, define a transformation $\widetilde{Z}$ of the latent relations which is obtained by subtracting the regression effects, additive effects, and all but the ℓth multiplicative effect, and multiplying by the inverse square root of the within-dyad correlation matrix $\Sigma_e^{-1/2}$. This transformation can be written:

$$\widetilde{Z} = \gamma(Z - \langle X, \boldsymbol{\beta} \rangle - \boldsymbol{a}1_n^T - 1_n\boldsymbol{b}^T - U_{-\ell}V_{-\ell}^T) + \delta(Z - \langle X, \boldsymbol{\beta} \rangle - \boldsymbol{a}1_n^T - 1_n\boldsymbol{b}^T - U_{-\ell}V_{-\ell}^T)^T,$$

where γ and δ are parts of $\Sigma_e^{-1/2}$ as defined in the update for the regression coefficients and additive effects above, and $U_{-\ell}$ and $V_{-\ell}$ denote the $n \times (k-1)$ matrices of multiplicative effects without effect ℓ. The model in (4.2) written in terms of the transformed relations $\widetilde{Z}$ has the form of (C.3), where the mean $(\boldsymbol{\mu}_U, \boldsymbol{\mu}_V)$ and covariance matrix $\widetilde{\Sigma}_{UV}$ of the distribution on the multiplicative effects are the conditional mean and variance of the ℓth components of U and V given $U_{-\ell}$ and $V_{-\ell}$ based on the multivariate normal distribution in (4.5).

Thus, we can use the full conditional multivariate normal distributions for $\boldsymbol{u}$ and $\boldsymbol{v}$ derived based on (C.3) to sample from the full conditional distributions ℓth multiplicative effects corresponding to the model in (4.2). To summarize, samples from the full conditional distributions of the ℓth multiplicative effects can be obtained via

1. sample $\boldsymbol{u}$ from $p(\boldsymbol{u}|\boldsymbol{v}, \widetilde{Z})$;

2. sample $\boldsymbol{v}$ from $p(\boldsymbol{v}|\boldsymbol{u}, \widetilde{Z})$.

Full conditional of Σ_{UV}

The full conditional distribution of Σ_{UV} is proportional to $p(U, V|\Sigma_{UV})p(\Sigma_{UV})$. Using the inverse-Wishart prior specified in (4.7), the full conditional distribution of Σ_{UV} is inverse-Wishart(ν_{UV}, S_{UV}), where $\nu_{UV} = n + 2 + 2k$ and $S_{UV} = (U, V)^T(U, V) + I_{2k}$.

C.1.3 Within-dyad correlation ρ

Let $E = Z - \langle X, \boldsymbol{\beta}\rangle - \boldsymbol{a}1_n^T - 1_n\boldsymbol{b}^T - UV^T$ denote the residuals from the model representation in (4.6). Further, let $\boldsymbol{e}_u = (e_{1,2}, ..., e_{n-1,n})$ denote the vector of the upper triangular portion of E, $\boldsymbol{e}_\ell = (e_{2,1}, ..., e_{n,n-1})$ represent the vector of the lower triangular of E, and $\boldsymbol{e}_d = (e_{1,1}, ..., e_{n,n})$ denote the vector of diagonal elements. The full conditional distribution of ρ is written

$$\begin{aligned} p(\rho|Z, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, U, V, \sigma_e^2) &\propto p(E|\rho, \sigma_e^2)p(\rho) \\ &\propto (1-\rho^2)^{-n(n-1)/4}\exp\left(\frac{-1}{2\sigma_e^2(1-\rho^2)}\left(\boldsymbol{e}_u^T\boldsymbol{e}_u + \boldsymbol{e}_\ell^T\boldsymbol{e}_\ell - 2\rho\boldsymbol{e}_u^T\boldsymbol{e}_\ell\right)\right) \\ &\quad (1+\rho)^{-n/2}\exp\left(\frac{-1}{2\sigma_e^2(1+\rho)}\boldsymbol{e}_d^T\boldsymbol{e}_d\right) \end{aligned}$$

This is not a well-known distribution as a function of ρ so we instead sample ρ using the following Metropolis-Hastings procedure:

1. Sample a candidate value ρ^* from $\rho^*|\rho \sim \text{normal}_{[-1,1]}(\rho, \sigma_\rho^2)$, where $\sigma_\rho^2 = 16 * (1 - \hat{\rho}^2)^2/(n(n-1))$ and $\hat{\rho}$ is the sample correlation estimate from $(\boldsymbol{e}_u, \boldsymbol{e}_\ell)$.

2. Accept ρ^* with probability

$$\alpha(\rho, \rho^*) = \min\left\{1, \frac{p(\rho^*|Z, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, U, V, \sigma_e^2)q(\rho|\rho^*)}{p(\rho|Z, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, U, V, \sigma_e^2)q(\rho^*|\rho)}\right\}$$

   where $q(\rho^*|\rho)$ is the proposal truncated normal density in step 1.

Since updating ρ via the above Metropolis-Hastings step is cheap computationally, we find it is helpful to repeat the step a couple hundred times at each iteration of the Markov chain. Conditioning on the latent relations Z, there is a large amount of information about ρ so typically only proposals ρ^* very similar to the current value ρ are accepted. In our experience, the variance of the proposals σ_ρ^2 results in acceptance rate around 0.35.

C.1.4 Error variance σ_e^2

When the observed relations Y are continuous, the variance of the errors $e_{i,j}$ in the model in (4.2) is σ_e^2. The full conditional distribution of σ_e^2 is proportional to $p(Z|\boldsymbol{a},\boldsymbol{b},U,V,\boldsymbol{\beta},\rho,\sigma_e^2)p(\sigma_e^2)$ where $Z = Y$. Based on the inverse-gamma prior in (4.7), the full conditional distribution of σ_e^2 is inverse-gamma$(\nu_e/2, s/2)$, where

$$\nu_e = n^2 + 1, \qquad s = \frac{1}{(1-\rho^2)}\left(\boldsymbol{e}_u^T\boldsymbol{e}_u + \boldsymbol{e}_\ell^T\boldsymbol{e}_\ell - 2\rho\boldsymbol{e}_u^T\boldsymbol{e}_\ell\right) + \frac{1}{(1+\rho)}\boldsymbol{e}_d^T\boldsymbol{e}_d + 1,$$

and $\boldsymbol{e}_u$, $\boldsymbol{e}_\ell$, and $\boldsymbol{e}_d$ are as defined in description of the ρ Metropolis-Hasting procedure immediately above.

C.1.5 Latent relations Z

The full conditional distribution for a relation $z_{i,j}$ is proportional to

$$p(Z|\boldsymbol{a},\boldsymbol{b},\boldsymbol{\beta},U,V,\rho)p(Y|Z) = p(Z|\boldsymbol{a},\boldsymbol{b},\boldsymbol{\beta},U,V,\rho)\ \mathbb{1}\{Z \in S(Y)\},$$

which is a truncated normal distribution on an interval determined by $S(Y)$ and potentially other relations $z_{k,\ell}$. The distribution has (untruncated) mean and variance given by $m_{i,j} = \mu_{i,j} + \rho(z_{j,i} - \mu_{j,i})$ and $v_{i,j} = \sigma_e^2(1-\rho^2)$, respectively, where $\mu_{i,j} = a_i + b_j + \boldsymbol{\beta}^T\boldsymbol{x}_{i,j} + \boldsymbol{u}_i^T\boldsymbol{v}_j$ is the mean according to the model in (4.2). The truncation bounds are that which constrain Z to the set $S(Y)$ given all but latent relation $z_{i,j}$. These full conditional distribution are similar to those in Bayesian estimation procedure for the semiparametric copula model in Hoff (2007).

As an example, consider the case when the observed relations Y are binary such that $y_{i,j}$ is an indicator of whether $z_{i,j}$ is greater than zero. The full conditional distribution for $z_{i,j}$ is truncated to be above or below zero depending on whether $y_{i,j}$ equals one or zero, respectively. The entire upper triangular portion of Z can be updated simultaneously from its full conditional since the elements are conditionally independent in the model given all other parameters. The entire lower triangular portion of Z can be updated simultaneously as well. In cases where the $z_{i,j}$s must satisfy a relative ordering, the truncation bounds for the full conditional distribution of $z_{i,j}$ will depend on the values of the other $z_{i,j}$s so less fewer relations can be updated simultaneously from their full conditional distribution.

Since there are no relational observations $y_{i,i}$ corresponding to the diagonal elements $z_{i,i}$, the diagonal elements are sampled from the model in (4.8) conditional on the current parameter values. Recall that if the relations Y are continuous, only the diagonal latent relations must be sampled.

C.2 Metropolis-Hastings step for $\{Z, \rho\}$

Consider the update for $\{\boldsymbol{z}_U, \rho\}$. Following the notation in Section 4.4.3, the proposal $\{\tilde{\boldsymbol{z}}_U, \tilde{\rho}\}$ is obtained by sampling

1. $\tilde{\rho} \sim \text{truncated normal}_{[-1,1]}(\rho, \sigma_\rho^2)$
2. $\tilde{\boldsymbol{z}}_U | \tilde{\rho}, \boldsymbol{z}_U, M, Y \sim \text{truncated normal}_{[\boldsymbol{t}_l^{(U)}, \boldsymbol{t}_u^{(U)}]}(\tilde{\boldsymbol{\mu}}_{z_U}, \tilde{\sigma}_z^2 \mathrm{I}_{n(n-1)/2})$

where $\tilde{\boldsymbol{\mu}}_{z_U} = \boldsymbol{m}_U + \tilde{\rho}(\boldsymbol{z}_L - \boldsymbol{m}_L)$, $\tilde{\sigma}_z^2 = (1-\tilde{\rho}^2)$, and $\boldsymbol{m}_U$ and $\boldsymbol{m}_L$ are vectors of the lower and upper triangular elements of $M = \langle X, \boldsymbol{\beta} \rangle + \boldsymbol{a} 1_n^T + 1_n \boldsymbol{b}^T + UV^T$. Since the relations Y are binary, the lower and upper truncation bounds equal $\boldsymbol{t}_l^{(U)} = \log(\boldsymbol{y}_L)$ and $\boldsymbol{t}_u^{(U)} = -\log(1-\boldsymbol{y}_L)$, respectively, where $\boldsymbol{y}_L$ is a vector of the lower triangular elements of Y.

The Metropolis-Hastings acceptance probability is

$$\begin{aligned}
\alpha(\{\boldsymbol{z}_U, \rho\}, \{\tilde{\boldsymbol{z}}_U, \tilde{\rho}\}) &= \min\Bigg\{1 \,,\, \frac{p(\tilde{\boldsymbol{z}}_U, \tilde{\rho}, \boldsymbol{z}_L, \boldsymbol{z}_D, U, V, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\beta}, \Sigma_{ab}, \Sigma_{UV} | Y)}{p(\boldsymbol{z}_U, \rho, \boldsymbol{z}_L, \boldsymbol{z}_D, U, V, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\beta}, \Sigma_{ab}, \Sigma_{UV} | Y)} \\
&\qquad \times \frac{p(\boldsymbol{z}_U, \rho | \tilde{\boldsymbol{z}}_U, \tilde{\rho}, \boldsymbol{z}_L, U, V, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\beta}, Y)}{p(\tilde{\boldsymbol{z}}_U, \tilde{\rho} | \boldsymbol{z}_U, \rho, \boldsymbol{z}_L, U, V, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\beta}, Y)} \Bigg\} \\
&= \min\Bigg\{1 \,,\, \frac{p(Y | \tilde{\boldsymbol{z}}_U, \boldsymbol{z}_L, \boldsymbol{z}_D) p(\tilde{\boldsymbol{z}}_U | \boldsymbol{z}_L, \tilde{\rho}, M) p(\boldsymbol{z}_D | M, \tilde{\rho}) p(\tilde{\rho})}{p(Y | \boldsymbol{z}_U, \boldsymbol{z}_L, \boldsymbol{z}_D) p(\boldsymbol{z}_U | \boldsymbol{z}_L, \rho, M) p(\boldsymbol{z}_D | M, \rho) p(\rho)} \\
&\qquad \times \frac{p(\boldsymbol{z}_U | \boldsymbol{z}_L, \rho, M, Y) p(\rho | \tilde{\rho})}{p(\tilde{\boldsymbol{z}}_U | \boldsymbol{z}_L, \tilde{\rho}, M, Y) p(\tilde{\rho} | \rho)} \Bigg\}
\end{aligned}$$

Note $p(Y | \tilde{\boldsymbol{z}}_U, \boldsymbol{z}_L, \boldsymbol{z}_D) = p(Y | \boldsymbol{z}_U, \boldsymbol{z}_L, \boldsymbol{z}_D) = 1$ since the current and proposed $\boldsymbol{z}_U$ satisfies the constraints in $B(Y)$, and the uniform prior on ρ implies $p(\tilde{\rho}) = p(\rho) = 1$. Thus, the

above expression simplifies to

$$
\begin{aligned}
&= \min\left\{1\ ,\ \frac{p(\tilde{\boldsymbol{z}}_U|\boldsymbol{z}_L,\tilde{\rho},M)p(\boldsymbol{z}_D|M,\tilde{\rho})\cdot\Big(p(\boldsymbol{z}_U|\boldsymbol{z}_L,\rho,M)/p(\boldsymbol{z}_U\in B(Y)|\boldsymbol{z}_L,M)\Big)p(\rho|\tilde{\rho})}{p(\boldsymbol{z}_U|\boldsymbol{z}_L,\rho,M)p(\boldsymbol{z}_D|M,\rho)\cdot\Big(p(\tilde{\boldsymbol{z}}_U|\boldsymbol{z}_L,\tilde{\rho},M)/p(\tilde{\boldsymbol{z}}_U\in B(Y)|\boldsymbol{z}_L,M)\Big)p(\tilde{\rho}|\rho)}\right\}\\
&= \min\left\{1\ ,\ \frac{p(\boldsymbol{z}_D|M,\tilde{\rho})\cdot p(\tilde{\boldsymbol{z}}_U\in B(Y)|\boldsymbol{z}_L,M)p(\rho|\tilde{\rho})}{p(\boldsymbol{z}_D|M,\rho)\cdot p(\boldsymbol{z}_U\in B(Y)|\boldsymbol{z}_L,M)p(\tilde{\rho}|\rho)}\right\}\\
&= \min\Bigg\{1\ ,\ \frac{\Big((1+\tilde{\rho})^{-n/2}\exp\Big(\frac{-\boldsymbol{e}_d^2}{2(1+\tilde{\rho})}\Big)\Big)\left(\prod\left(\Phi\Big(\frac{\boldsymbol{t}_u^{(U)}-\tilde{\boldsymbol{\mu}}_{z_U}}{\tilde{\sigma}_z}\Big)-\Phi\Big(\frac{\boldsymbol{t}_l^{(U)}-\tilde{\boldsymbol{\mu}}_{z_U}}{\tilde{\sigma}_z}\Big)\right)\right)}{\Big((1+\rho)^{-n/2}\exp\Big(\frac{-\boldsymbol{e}_d^2}{2(1+\rho)}\Big)\Big)\left(\prod\left(\Phi\Big(\frac{\boldsymbol{t}_u^{(U)}-\boldsymbol{\mu}_{z_U}}{\sigma_z}\Big)-\Phi\Big(\frac{\boldsymbol{t}_l^{(U)}-\boldsymbol{\mu}_{z_U}}{\sigma_z}\Big)\right)\right)}\\
&\qquad\qquad\times\frac{\Big(\Phi\Big(\frac{1-\rho}{\sigma_\rho}\Big)-\Phi\Big(\frac{-1-\rho}{\sigma_\rho}\Big)\Big)}{\Big(\Phi\Big(\frac{1-\tilde{\rho}}{\sigma_\rho}\Big)-\Phi\Big(\frac{-1-\tilde{\rho}}{\sigma_\rho}\Big)\Big)}\Bigg\},
\end{aligned}
$$

where $\boldsymbol{e}_d$ are the diagonal elements of E defined by (4.6).

C.3 Mean-field variational Bayesian approximation calculations

In order to simplify the notation for analytic expression of the update for each approximating parameter distribution $q(\theta_i)$, we consider a transformation of the latent relations similar to that used in the discussion of the MCMC algorithm. Let $\xi_{i,j} = z_{i,j} + z_{j,i}$ and $\eta_{i,j} = z_{i,j} - z_{j,i}$ for $i < j$. The model in (4.2) and (4.3) can be equivalently expressed as

$$
\begin{pmatrix}\boldsymbol{\xi}\\ \boldsymbol{\eta}\end{pmatrix} = \begin{pmatrix}X_\xi\\ X_\eta\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{a}\\ \boldsymbol{b}\end{pmatrix} + \begin{pmatrix}\Lambda_\xi\\ \Lambda_\eta\end{pmatrix} + \begin{pmatrix}\boldsymbol{\epsilon}_\xi\\ \boldsymbol{\epsilon}_\eta\end{pmatrix}
$$

where

$$
\begin{pmatrix}\boldsymbol{\epsilon}_\xi\\ \boldsymbol{\epsilon}_{\boldsymbol{\eta}}\end{pmatrix} \sim \text{normal}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_\xi^2 = 2\sigma_e^2(1+\rho) & 0\\ 0 & \sigma_\eta^2 = 2\sigma_e^2(1-\rho)\end{pmatrix}\otimes I_{n(n-1)/2}\right),
$$

$\boldsymbol{\xi} = (\xi_{1,2},...,\xi_{n-1,n})$, $\boldsymbol{\eta} = (\eta_{1,2},...,\eta_{n-1,n})$, X_ξ and X_η are the appropriate $(n(n-1)/2 \times (p+2n))$ design matrices, and Λ_ξ and Λ_η are functions of U and V. For notational convenience we will frequently denote the vector of regression coefficients and additive effects by $\boldsymbol{\phi} = (\boldsymbol{\beta}^T, \boldsymbol{a}^T, \boldsymbol{b}^T)^T$. Note that in this variational Bayesian approximation of the posterior distribution, there are no diagonal latent relations $z_{i,i}$. Depending on the type of observed

relations Y, the posterior distribution, and hence, variational approximation can have one of two forms:

1. If Y continuous, there are no latent relations Z and the variational approximating distribution has the form

$$q = q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\beta})q(U, V)q(\Sigma_{ab})q(\Sigma_{UV})q(\rho)q(\sigma_e^2).$$

2. If Y is not continuous, the variance on the latent relations σ_e^2 is fixed at 1 and the approximating distribution has the form

$$q = q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\beta})q(U, V)q(\Sigma_{ab})q(\Sigma_{UV})q(\rho)q(Z).$$

C.3.1 $\boldsymbol{q(a, b, \beta)}$ *(=* $\boldsymbol{q(\phi)}$*)*

$$\begin{aligned}
\log[q(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b})] &= \mathrm{E}_{q/q_{(\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{b})}} \log\Big[p(Y|\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, U, V, \Sigma_e)p(\boldsymbol{a}, \boldsymbol{b}|\Sigma_{ab})p(\boldsymbol{\beta})\Big] + c \\
&= \mathrm{E}_{q/q_{(\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{b})}} \log\Big[\exp\Big(\frac{-1}{2\sigma_\xi^2}\left(\phi^T X_\xi^T X_\xi \phi - 2\phi^T X_\xi^T \boldsymbol{\xi} + 2\phi^T X_\xi^T \Lambda_\xi\right) \\
&- \frac{1}{2\sigma_\eta^2}\left(\phi^T X_\eta^T X_\eta \phi - 2\phi^T X_\eta^T \boldsymbol{\eta} + 2\phi^T X_\eta^T \Lambda_\eta\right)\Big)\exp\Big(\frac{-1}{2}\phi^T \begin{pmatrix} \Sigma_\beta^{-1} & 0 \\ 0 & \Sigma_{ab}^{-1} \otimes I_n \end{pmatrix} \phi\Big)\Big] + c \\
&\Rightarrow q(\phi) = q(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}) = f_{\text{normal}}\left(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b} \middle| \mu_{q(\phi)}, \Sigma_{q(\phi)}\right)
\end{aligned}$$

Here f_{normal} denotes the density of a normal distribution with mean $\mu_{q(\phi)}$ and covariance matrix $\Sigma_{q(\phi)}$ given by

$$\begin{aligned}
\Sigma_{q(\phi)} =& \Big[\frac{1}{2}\mathrm{E}_{q(\sigma_e^2)}\Big[\frac{1}{\sigma_e^2}\Big]\mathrm{E}_{q(\rho)}\Big[\frac{1}{1+\rho}\Big]X_\xi^T X_\xi + \frac{1}{2}\mathrm{E}_{q(\sigma_e^2)}\Big[\frac{1}{\sigma_e^2}\Big]\mathrm{E}_{q(\rho)}\Big[\frac{1}{1-\rho}\Big]X_\eta^T X_\eta \\
&+ \begin{pmatrix} \Sigma_\beta^{-1} & 0 \\ 0 & \mathrm{E}_{q_{(\Sigma_{ab})}}\Big[\Sigma_{ab}^{-1}\Big] \otimes I_n \end{pmatrix}\Big]^{-1}, \\
\mu_{q(\phi)} =& \Sigma_q\Big[\frac{1}{2}\mathrm{E}_{q(\sigma_e^2)}\Big[\frac{1}{\sigma_e^2}\Big]\mathrm{E}_{q(\rho)}\Big[\frac{1}{1+\rho}\Big]\left(X_\xi^T \mathrm{E}_{q_{(Z)}}\Big[\boldsymbol{\xi}\Big] - X_\xi^T \mathrm{E}_{q_{(U,V)}}\Big[\Lambda_\xi\Big]\right) \\
&+ \frac{1}{2}\mathrm{E}_{q(\sigma_e^2)}\Big[\frac{1}{\sigma_e^2}\Big]\mathrm{E}_{q(\rho)}\Big[\frac{1}{1-\rho}\Big]\left(X_\eta^T \mathrm{E}_{q_{(Z)}}\Big[\boldsymbol{\eta}\Big] - X_\eta^T \mathrm{E}_{q_{(U,V)}}\Big[\Lambda_\eta\Big]\right)\Big].
\end{aligned}$$

Expectations needed:

- $\mathrm{E}_{q(\phi)}\Big[\phi\Big]$
- $\mathrm{E}_{q(\phi)}\Big[\phi^T X_\xi^T X_\xi \phi\Big]$, $\mathrm{E}_{q_{(\phi)}}\Big[\phi^T X_\eta^T X_\eta \phi\Big]$
- $\mathrm{E}_{q(\phi)}\Big[(\boldsymbol{a},\boldsymbol{b})^T(\boldsymbol{a},\boldsymbol{b})\Big]$

C.3.2 $\boldsymbol{q(\Sigma_{ab})}$

$$\begin{aligned}\log[q(\Sigma_{ab}^{-1})] &= \mathrm{E}_{q/q_{(\Sigma_{ab})}}\ \log\Big[p(\boldsymbol{a},\boldsymbol{b}|\Sigma_{ab})p(\Sigma_{ab})\Big] + c \\ &= \mathrm{E}_{q/q_{(\Sigma_{ab})}}\ \log\Big[|\Sigma_{ab}|^{-n/2} etr\left(\frac{-1}{2}(\boldsymbol{a},\boldsymbol{b})\Sigma_{ab}^{-1}(\boldsymbol{a},\boldsymbol{b})^T\right) etr(-\Sigma_{ab}^{-1}/2)\Big] + c\end{aligned}$$

$$\Rightarrow q(\Sigma_{ab}^{-1}) = f_{\text{Wishart}}\left(\Sigma_{ab}^{-1}\middle|\nu_{q(\Sigma_{ab})} = n+3, S^{-1}_{q(\Sigma_{ab})} = \Big[\mathrm{E}_{q(a,b)}\Big[(\boldsymbol{a},\boldsymbol{b})^T(\boldsymbol{a},\boldsymbol{b})\Big] + I_2\Big]^{-1}\right)$$

Expectations needed:

- $\mathrm{E}_{q(\Sigma_{ab})}\Big[\Sigma_{ab}^{-1}\Big]$

C.3.3 $\boldsymbol{q(\rho)}$

$$\begin{aligned}\log[q(\rho)] &= \mathrm{E}_{q/q(\rho)}\ \log\Big[p(Y|\boldsymbol{a},\boldsymbol{b},U,V,\boldsymbol{\beta},\Sigma_e)p(\rho)\Big] + c \\ &= \mathrm{E}_{q/q(\rho)}\ \log\Big[(\sigma_\xi^2\sigma_\eta^2)^{-\frac{n(n-1)}{4}}\exp\Big(\frac{-1}{2\sigma_\xi^2}\left(\phi^T X_\xi^T X_\xi\phi - 2\phi^T X_\xi^T\boldsymbol{\xi} + 2\phi^T X_\xi^T\Lambda_\xi - 2\Lambda_\xi^T\boldsymbol{\xi} + \Lambda_\xi^T\Lambda_\xi + \boldsymbol{\xi}^T\boldsymbol{\xi}\right) \\ &\quad - \frac{1}{2\sigma_\eta^2}\left(\phi^T X_\eta^T X_\eta\phi - 2\phi^T X_\eta^T\boldsymbol{\eta} + 2\phi^T X_\eta^T\Lambda_\eta - 2\Lambda_\eta^T\boldsymbol{\eta} + \Lambda_\eta^T\Lambda_\eta + \boldsymbol{\eta}^T\boldsymbol{\eta}\right)\Big)\Big] + c\end{aligned}$$

$$\Rightarrow \log[q(\rho)] = A\cdot\log\Big(1-\rho^2\Big) + \frac{1}{(1+\rho)}B + \frac{1}{(1-\rho)}C + c$$

$$\begin{aligned}A &= -\frac{n(n-1)}{4} \\ B &= \frac{-1}{4}\mathrm{E}_{q_{(\sigma_e^2)}}\Big[\frac{1}{\sigma_e^2}\Big]\Big(\mathrm{E}_{q_{(\phi)}}\Big[\phi^T X_\xi^T X_\xi\phi\Big] - 2\mathrm{E}_{q_{(\phi)}}\Big[\phi^T\Big]X_\xi^T\mathrm{E}_{q_{(Z)}}\Big[\boldsymbol{\xi}\Big] + 2\mathrm{E}_{q_{(\phi)}}\Big[\phi^T\Big]X_\xi^T\mathrm{E}_{q_{(U,V)}}\Big[\Lambda_\xi\Big] \\ &\quad - 2\mathrm{E}_{q_{(U,V)}}\Big[\Lambda_\xi^T\Big]\mathrm{E}_{q_{(Z)}}\Big[\boldsymbol{\xi}\Big] + \mathrm{E}_{q_{(U,V)}}\Big[\Lambda_\xi^T\Lambda_\xi\Big] + \mathrm{E}_{q_{(Z)}}\Big[\boldsymbol{\xi}^T\boldsymbol{\xi}\Big]\Big) \\ C &= \frac{-1}{4}\mathrm{E}_{q_{(\sigma_e^2)}}\Big[\frac{1}{\sigma_e^2}\Big]\Big(\mathrm{E}_{q_{(\phi)}}\Big[\phi^T X_\eta^T X_\eta\phi\Big] - 2\mathrm{E}_{q_{(\phi)}}\Big[\phi^T\Big]X_\eta^T\mathrm{E}_{q_{(Z)}}\Big[\boldsymbol{\eta}\Big] + 2\mathrm{E}_{q_{(\phi)}}\Big[\phi^T\Big]X_\eta^T\mathrm{E}_{q_{(U,V)}}\Big[\Lambda_\eta\Big] \\ &\quad - 2\mathrm{E}_{q_{(U,V)}}\Big[\Lambda_\eta^T\Big]\mathrm{E}_{q_{(Z)}}\Big[\boldsymbol{\eta}\Big] + \mathrm{E}_{q_{(U,V)}}\Big[\Lambda_\eta^T\Lambda_\eta\Big] + \mathrm{E}_{q_{(Z)}}\Big[\boldsymbol{\eta}^T\boldsymbol{\eta}\Big]\Big)\end{aligned}$$

Since the distribution $q(\rho)$ is not a well known distribution, a Markov chain is constructed at each update step and the samples from the chain are used to estimate the expectations of functions of ρ required by the other updates. As in the original MCMC presented in Section 4.3, the Markov chain consists of a Metropolis-Hastings step with a normal proposal distribution that is truncated to the interval $[-1, 1]$ and has an (untruncated) mean equal to the current value of ρ.

Expectations needed:

- $\mathrm{E}_{q(\rho)}\left[\frac{1}{1+\rho}\right]$, $\mathrm{E}_{q(\rho)}\left[\frac{1}{1-\rho}\right]$
- $\mathrm{E}_{q(\rho)}\left[\frac{1}{1-\rho^2}\right]$, $\mathrm{E}_{q(\rho)}\left[\frac{\rho}{1-\rho^2}\right]$ (if Y is not continuous and hence, model has latent relations Z)

C.3.4 $\boldsymbol{q(U, V)}$

$$\begin{aligned}\log[q(U,V)] &= \mathrm{E}_{q/q_(U,V)}\ \log\Big[p(Y|\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, U, V, \Sigma_e)p(U, V|\Sigma_{UV})\Big] + c \\ &= \mathrm{E}_{q/q(U,V)}\ \log\Big[\exp\Big(\frac{-1}{2\sigma_\xi^2}\left(2\phi^T X_\xi^T \Lambda_\xi - 2\Lambda_\xi^T \boldsymbol{\xi} + \Lambda_\xi^T \Lambda_\xi\right) \\ &\quad - \frac{1}{2\sigma_\eta^2}\left(2\phi^T X_\eta^T \Lambda_\eta - 2\Lambda_\eta^T \boldsymbol{\eta} + \Lambda_\eta^T \Lambda_\eta\right)\Big) etr\Big(-\frac{1}{2}(U,V)\Sigma_{UV}^{-1}(U,V)^T\Big)\Big] + c\end{aligned}$$

$$\Rightarrow \log\Big[q(U,V)\Big] = \frac{-1}{2}\Big[-2\Lambda_\xi^T A - 2\Lambda_\eta^T B + C\Lambda_\xi^T \Lambda_\xi + D\Lambda_\eta^T \Lambda_\eta + \mathrm{tr}\Big[(U,V)E(U,V)^T\Big]\Big] + c$$

$$\begin{aligned}A =& \frac{1}{2}\mathrm{E}_{q_{(\sigma_e^2)}}\left[\frac{1}{\sigma_e^2}\right]\mathrm{E}_{q_{(\rho)}}\left[\frac{1}{1+\rho}\right]\left[\mathrm{E}_{q_{(Z)}}\left[\boldsymbol{\xi}\right] - X_\xi \mathrm{E}_{q_{(\phi)}}\left[\phi\right]\right] \\ B =& \frac{1}{2}\mathrm{E}_{q_{(\sigma_e^2)}}\left[\frac{1}{\sigma_e^2}\right]\mathrm{E}_{q_{(\rho)}}\left[\frac{1}{1-\rho}\right]\left[\mathrm{E}_{q_{(Z)}}\left[\boldsymbol{\eta}\right] - X_\eta \mathrm{E}_{q_{(\phi)}}\left[\phi\right]\right] \\ C =& \frac{1}{2}\mathrm{E}_{q_{(\sigma_e^2)}}\left[\frac{1}{\sigma_e^2}\right]\mathrm{E}_{q_{(\rho)}}\left[\frac{1}{1+\rho}\right] \\ D =& \frac{1}{2}\mathrm{E}_{q_{(\sigma_e^2)}}\left[\frac{1}{\sigma_e^2}\right]\mathrm{E}_{q_{(\rho)}}\left[\frac{1}{1-\rho}\right] \\ E =& \mathrm{E}_{q_{(\Sigma_{UV})}}\left[\Sigma_{UV}^{-1}\right]\end{aligned}$$

The form of the $q(U, V)$ is not a known distribution. Thus, in order to approximate the expectations of function of U and V needed for the other updates, we propose creating a Markov chain Gibbs sampler that iteratively samples from the conditional distributions of

each column of U and each column of V given the rest of U and V, similar to that done in the MCMC sampler introduced in Section 4.3. Let the ℓth column of U be denoted $U_{(\ell)}$ and decompose Λ_ξ and Λ_η as $\Lambda_\xi = \Lambda_{\xi,-(\ell)} + \Gamma_{\xi,U(\ell)}U_{(\ell)}$ and $\Lambda_\eta = \Lambda_{\eta,-(\ell)} + \Gamma_{\eta,U(\ell)}U_{(\ell)}$, respectively. Then the conditional distribution of $U_{(\ell)}$ can be written:

$$
\begin{aligned}
q(U_{(\ell)}|U_{-(\ell)},V) &\propto \exp\left(\frac{-1}{2}\Big[-2\Lambda_\xi^T A - 2\Lambda_\eta^T B + C\Lambda_\xi^T\Lambda_\xi + D\Lambda_\eta^T\Lambda_\eta + \mathrm{tr}\Big[(U,V)E(U,V)^T\Big]\Big]\right) \\
&\propto \exp\Big(\frac{-1}{2}\Big[-2\Big(\Lambda_{\xi,-(\ell)} + \Gamma_{\xi,U(\ell)}U_{(\ell)}\Big)^T A - 2\Big(\Lambda_{\eta,-(\ell)} + \Gamma_{\eta,U(\ell)}U_{(\ell)}\Big)^T B \\
&\quad + C\Big(\Lambda_{\xi,-(\ell)} + \Gamma_{\xi,U(\ell)}U_{(\ell)}\Big)^T\Big(\Lambda_{\xi,-(\ell)} + \Gamma_{\xi,U(\ell)}U_{(\ell)}\Big) \\
&\quad + D\Big(\Lambda_{\eta,-(\ell)} + \Gamma_{\eta,U(\ell)}U_{(\ell)}\Big)^T\Big(\Lambda_{\eta,-(\ell)} + \Gamma_{\eta,U(\ell)}U_{(\ell)}\Big) \\
&\quad + \Big[U_{(\ell)}^T E_{\ell,\ell}U_{(\ell)} + 2U_{(\ell)}^T(U_{-(\ell)},V)E_{-\ell,\ell}\Big]\Big]\Big) \\
&\propto \exp\Big(\frac{-1}{2}\Big[U_{(\ell)}^T\Big[C\Gamma_{\xi,U(\ell)}^T\Gamma_{\xi,U(\ell)} + D\Gamma_{\eta,U(\ell)}^T\Gamma_{\eta,U(\ell)} + E_{\ell,\ell}I_n\Big]U_{(\ell)} \\
&\quad - 2U_{(\ell)}^T\Big[\Gamma_{\xi,U(\ell)}^T A + \Gamma_{\eta,U(\ell)}^T B - C\Gamma_{\xi,U(\ell)}^T\Lambda_{\xi,-(\ell)} - D\Gamma_{\eta,U(\ell)}^T\Lambda_{\eta,-(\ell)} - (U_{-(\ell)},V)E_{-\ell,\ell}\Big]\Big]\Big)
\end{aligned}
$$

where $E_{\ell,\ell} = E[\ell,\ell]$, $E_{\ell,-\ell} = E[\ell,-\ell]$.

$$
\Rightarrow q(U_{(\ell)}|U_{-(\ell)},V) = f_{\text{normal}}\left(U_{(\ell)}\middle|\mu_{U_{(\ell)}},\Sigma_{U_{(\ell)}}\right)
$$

$$
\begin{aligned}
\Sigma_{U_{(\ell)}} &= \Big[C\Gamma_{\xi,U(\ell)}^T\Gamma_{\xi,U(\ell)} + D\Gamma_{\eta,U(\ell)}^T\Gamma_{\eta,U(\ell)} + E_{\ell,\ell}I_n\Big]^{-1} \\
\mu_{U_{(\ell)}} &= \Sigma_{U_{(\ell)}}\Big[\Gamma_{\xi,U(\ell)}^T A + \Gamma_{\eta,U(\ell)}^T B - C\Gamma_{\xi,U(\ell)}^T\Lambda_{\xi,-(\ell)} - D\Gamma_{\eta,U(\ell)}^T\Lambda_{\eta,-(\ell)} - (U_{-(\ell)},V)E_{-\ell,\ell}\Big]
\end{aligned}
$$

The conditional distribution of a column of V is derived analogously.

Expectations needed:

- $\mathrm{E}_{q_{(U,V)}}\Big[\Lambda_\xi\Big]$, $\mathrm{E}_{q_{(U,V)}}\Big[\Lambda_\eta\Big]$
- $\mathrm{E}_{q_{(U,V)}}\Big[\Lambda_\xi^T\Lambda_\xi\Big]$, $\mathrm{E}_{q_{(U,V)}}\Big[\Lambda_\eta^T\Lambda_\eta\Big]$
- $\mathrm{E}_{q(U,V)}\Big[(U,V)^T(U,V)\Big]$

C.3.5 $q(\boldsymbol{\Sigma_{UV}})$

$$\begin{aligned}\log[q(\Sigma_{UV}^{-1})] &= \mathrm{E}_{q/q(\Sigma_{UV})}\ \log\Big[p(U,V|\Sigma_{UV})p(\Sigma_{UV})\Big] + c \\ &= \mathrm{E}_{q/q(\Sigma_{UV})}\ \log\Big[|\Sigma_{UV}|^{-n/2} etr\left(\frac{-1}{2}(U,V)\Sigma_{UV}^{-1}(U,V)^T\right)|\Sigma_{UV}|^{-1/2} etr(-\Sigma_{uv}^{-1}/2)\Big] + c \\ \Rightarrow q(\Sigma_{UV}^{-1}) &= f_{\text{Wishart}}\left(\Sigma_{UV}^{-1}\middle| n+2+2k,\ \Big[\mathrm{E}_{q(U,V)}\Big[(U,V)^T(U,V)\Big] + I_{2k}\Big]^{-1}\right)\end{aligned}$$

Expectations needed:

- $\mathrm{E}_{q(\Sigma_{UV})}\Big[\Sigma_{UV}^{-1}\Big]$

C.3.6 $q(\boldsymbol{\sigma_e^2})$

$$\begin{aligned}\log[q(\sigma_e^2)] &= \mathrm{E}_{q/q(\sigma_e^2)}\ \log\Big[p(Y|\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{b},\sigma_e^2,\rho)p(\sigma_e^2)\Big] + c \\ &= \mathrm{E}_{q/q(\sigma_e^2)}\ \log\Big[\exp\Big(\frac{-1}{2\sigma_\xi^2}\left(\phi^T X_\xi^T X_\xi\phi - 2\phi^T X_\xi^T\boldsymbol{\xi} + 2\phi^T X_\xi^T\Lambda_\xi - 2\Lambda_\xi^T\boldsymbol{\xi} + \Lambda_\xi^T\Lambda_\xi + \boldsymbol{\xi}^T\boldsymbol{\xi}\right) \\ &\quad - \frac{1}{2\sigma_\eta^2}\left(\phi^T X_\eta^T X_\eta\phi - 2\phi^T X_\eta^T\boldsymbol{\eta} + 2\phi^T X_\eta^T\Lambda_\eta - 2\Lambda_\eta^T\boldsymbol{\eta} + \Lambda_\eta^T\Lambda_\eta + \boldsymbol{\eta}^T\boldsymbol{\eta}\right)\Big) \\ &\quad (\sigma_\xi^2\sigma_\eta^2)^{-\frac{n(n-1)}{4}}(\sigma_e^{-2})^{\nu_e/2-1}\exp(-\nu_e\sigma_0^2\sigma_e^{-2}/2)\Big] + c\end{aligned}$$

$$\Rightarrow q(\sigma_e^{-2}) = f_{\text{gamma}}\left(\sigma_e^{-2}\middle|\text{shape } = A, \text{rate } = B\right)$$

$$\begin{aligned}A &= n(n-1)/2 + \nu_e/2 \\ B &= \frac{1}{2}\Big[\mathrm{E}_{q_{(\phi)}}\Big[\phi^T\Big[\frac{1}{2}\mathrm{E}_{q(\rho)}\Big[\frac{1}{1+\rho}\Big]X_\xi^T X_\xi + \frac{1}{2}\mathrm{E}_{q(\rho)}\Big[\frac{1}{1-\rho}\Big]X_\eta^T X_\eta\Big]\phi\Big] \\ &\quad - 2\mathrm{E}_{q_{(\phi)}}\Big[\phi^T\Big]\Big[\frac{1}{2}\mathrm{E}_{q(\rho)}\Big[\frac{1}{1+\rho}\Big]X_\xi^T\boldsymbol{\xi} - \frac{1}{2}\mathrm{E}_{q(\rho)}\Big[\frac{1}{1+\rho}\Big]X_\xi^T\mathrm{E}_{q_{(U,V)}}\Big[\Lambda_\xi\Big] \\ &\quad + \frac{1}{2}\mathrm{E}_{q(\rho)}\Big[\frac{1}{1-\rho}\Big]X_\eta^T\boldsymbol{\eta} - \frac{1}{2}\mathrm{E}_{q(\rho)}\Big[\frac{1}{1-\rho}\Big]X_\eta^T\mathrm{E}_{q_{(U,V)}}\Big[\Lambda_\eta\Big]\Big] \\ &\quad + \frac{1}{2}\mathrm{E}_{q(\rho)}\Big[\frac{1}{1+\rho}\Big]\mathrm{E}_{q_{(U,V)}}\Big[\Lambda_\xi^T\Lambda_\xi\Big] + \frac{1}{2}\mathrm{E}_{q(\rho)}\Big[\frac{1}{1-\rho}\Big]\mathrm{E}_{q_{(U,V)}}\Big[\Lambda_\eta^T\Lambda_\eta\Big] \\ &\quad - 2\Big[\frac{1}{2}\mathrm{E}_{q(\rho)}\Big[\frac{1}{1+\rho}\Big]\mathrm{E}_{q_{(U,V)}}\Big[\Lambda_\xi^T\Big]\boldsymbol{\xi} + \frac{1}{2}\mathrm{E}_{q(\rho)}\Big[\frac{1}{1-\rho}\Big]\mathrm{E}_{q_{(U,V)}}\Big[\Lambda_\eta^T\Big]\boldsymbol{\eta}\Big] \\ &\quad + \frac{1}{2}\mathrm{E}_{q(\rho)}\Big[\frac{1}{1+\rho}\Big]\boldsymbol{\xi}^T\boldsymbol{\xi} + \frac{1}{2}\mathrm{E}_{q(\rho)}\Big[\frac{1}{1-\rho}\Big]\boldsymbol{\eta}^T\boldsymbol{\eta} + \nu_e\sigma_0^2\Big]\end{aligned}$$

Note that no expectations are placed on $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ in the above update since σ_e^2 is only in the model when the relations Y are continuous and hence, $Z = Y$ is observed.

Expectations needed:

- $\mathrm{E}_{q(\sigma_e^2)}\left[\frac{1}{\sigma_e^2}\right]$

C.3.7 $\boldsymbol{q(Z)}$

Define $\mu = \boldsymbol{a}1_n^T + 1_n\boldsymbol{b}^T + UV^T + \langle X, \boldsymbol{\beta}\rangle$. Let $\boldsymbol{z}_U = (z_{1,2}, z_{1,3}, z_{2,3}, ..., z_{n-1,n})$ be the vector of the upper triangular portion of Z, $\boldsymbol{z}_L = (z_{2,1}, z_{3,1}, z_{3,2}, ..., z_{n,n-1})$ be the lower triangular portion, and define $\boldsymbol{y}_U$, $\boldsymbol{y}_L$, $\boldsymbol{\mu}_U$ and $\boldsymbol{\mu}_L$ accordingly. The posterior approximation $q(Z)$ is

$$
\begin{aligned}
\log[q(Z)] &= \mathrm{E}_{q/q(Z)}\ \log\Big[p(Y|Z)p(Z|\boldsymbol{a},\boldsymbol{b},U,V,\boldsymbol{\beta})\Big] + c \\
&= \mathrm{E}_{q/q(Z)}\ \log\Big[etr\Big(\frac{-1}{2}((\boldsymbol{z}_L,\boldsymbol{z}_U) - (\boldsymbol{\mu}_L,\boldsymbol{\mu}_U))\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1}((\boldsymbol{z}_L,\boldsymbol{z}_U) - (\boldsymbol{\mu}_L,\boldsymbol{\mu}_U))^T\Big)\ \mathbb{1}\{Z \in S(Y)\}\Big] \\
&= \ \mathrm{tr}\Big[\Big(\frac{-1}{2}\Big((\boldsymbol{z}_L,\boldsymbol{z}_U)\mathrm{E}_{q_{(\rho)}}\left[\begin{pmatrix} \frac{1}{1-\rho^2} & \frac{-\rho}{1-\rho^2} \\ \frac{-\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \end{pmatrix}\right](\boldsymbol{z}_L,\boldsymbol{z}_U)^T \\
&\quad - 2(\boldsymbol{z}_L,\boldsymbol{z}_U)\mathrm{E}_{q_{(\rho)}}\left[\begin{pmatrix} \frac{1}{1-\rho^2} & \frac{-\rho}{1-\rho^2} \\ \frac{-\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \end{pmatrix}\right]\mathrm{E}_{\phi,UV}\Big[(\boldsymbol{\mu}_L,\boldsymbol{\mu}_U)^T\Big]\Big)\Big)\Big] + \log\Big(\mathbb{1}\{Z \in S(Y)\}\Big)
\end{aligned}
$$

where $\mathbb{1}\{A\}$ is an indicator function which is 1 if A is true and 0 otherwise. The distribution $q(Z)$ is potentially quite complex depending on the set $S(Y)$. However, as with $q(U,V)$, expectations of functions of $\boldsymbol{\eta} = \boldsymbol{z}_U - \boldsymbol{z}_L$ and $\boldsymbol{\xi} = \boldsymbol{z}_U + \boldsymbol{z}_L$ needed for the other updates can be approximated by samples from a Gibbs sampling procedure that iteratively samples $z_{i,j}$ from $q(z_{i,j}|Z_{-(i,j)})$, where $Z_{-(i,j)}$ is the matrix of latent relations without $z_{i,j}$. These conditional distribution will have the same form as those in the MCMC full conditional distributions in Section 4.3.

As an example, we consider the case of binary observations Y here further. In this case we can update all of $\boldsymbol{z}_U$ at once from the conditional distribution $q(\boldsymbol{z}_U|\boldsymbol{z}_L)$ and $\boldsymbol{z}_L$ from $q(\boldsymbol{z}_L|\boldsymbol{z}_U)$. The conditional distribution $q(\boldsymbol{z}_L|\boldsymbol{z}_U)$ is a product of univariate truncated normal distributions with vector of means $\mu_{z(L)}$, variances $\sigma^2_{z(L)}$, lower bounds $\ell_{z(L)}$, and

upper bounds $u_{z(L)}$ given by

$$\mu_{z(L)} = \sigma^2_{z(L)}\Big[\mathrm{E}_{q_{(\rho)}}\Big[\frac{\rho}{1-\rho^2}\Big]\Big(\boldsymbol{z}_U - \mathrm{E}_{\phi,UV}\Big[\boldsymbol{\mu}_U\Big]\Big) + \mathrm{E}_{q_{(\rho)}}\Big[\frac{1}{1-\rho^2}\Big]\mathrm{E}_{\phi,UV}\Big[\boldsymbol{\mu}_L\Big]\Big];$$
$$\sigma^2_{z(L)} = 1/\mathrm{E}_{q_{(\rho)}}\Big[\frac{1}{1-\rho^2}\Big]; \quad \ell_{z(L)} = \log(\boldsymbol{y}_L); \quad u_{z(L)} = -\log(1-\boldsymbol{y}_L).$$

The conditional distribution $q(\boldsymbol{z}_U|\boldsymbol{z}_L)$ has a similar form.

Expectations needed:

- $\mathrm{E}_{q(Z)}\Big[\xi\Big]$, $\mathrm{E}_{q(Z)}\Big[\eta\Big]$
- $\mathrm{E}_{q(Z)}\Big[\xi^T\xi\Big]$, $\mathrm{E}_{q(Z)}\Big[\eta^T\eta\Big]$

VITA

Bailey K. Fosdick was born in Steamboat Springs, Colorado. She attended the Colorado School of Mines her freshman year of college and in 2008, she earned a B.S. in Mathematics, with concentrations in General Math and Statistics, along with a minor in Computer Science from Colorado State University in Fort Collins, Colorado. In 2013, she received a Ph.D. in Statistics from the University of Washington in Seattle, Washington and immediately begins a Postdoctoral Fellowship position at the Statistics and Applied Mathematics Institute in North Carolina. In August 2014, she will start as a Assistant Professor at Colorado State University in Fort Collins, Colorado.