Statistical inference using Kronecker structured covariance

Alexander Volfovsky

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

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Peter Hoff, Chair

Peter D. Hoff

Michael D. Perlman

Mathias Drton

Program Authorized to Offer Degree:
UW Statistics Department

University of Washington

Abstract

Statistical inference using Kronecker structured covariance

Alexander Volfovsky

Chair of the Supervisory Committee:
Professor Peter Hoff
UW Statistics and Biostatistics

We present results for testing and estimation in the context of separable covariance models. We concentrate on two types of data: relational data and cross-classified data. Relational data is frequently represented by a square matrix and we are often interested in identifying patterns of similarity between entries in the matrix. Under the assumption of a separable covariance, a natural model for such data is based on the matrix-variate normal distribution. In the context of this model we develop a likelihood ratio test for testing for row and column dependence based on the observation of a single relational data matrix. We provide extensions of the test to accommodate common features of such data, such as undefined diagonal entries, a non-zero mean, multiple observations, and deviations from normality. We then develop an estimation procedure for mean and covariance parameters under this model. In the context of cross-classified data, the separable covariance structure plays a role in relating the different effects in an ANOVA decomposition. Specifically, for many types of categorical factors, it is plausible that levels of a factor that have similar main-effect coefficients may also have similar coefficients in higher-order interaction terms. We introduce a class of hierarchical prior distributions based on the array-variate normal that can adapt to such similarities and hence borrow information from main effects and lower-order interactions in order to improve estimation of higher-order interactions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost I owe a debt of gratitude to my advisor Peter Hoff for introducing me to the general field of special manifolds and in particular to the open problems surrounding the matrix normal distribution.

I wish to thank my committee member and masters thesis advisor Mathias Drton for patiently steering me in the direction of statistics during my career at Chicago. My thanks go out to my committee member Michael Perlman for invaluable advice on invariant tests and estimators as well as for the wealth of anecdotes about statisticians past and present.

Finally, I would like to thank my fellow students at the University of Washington, my mother and my grandparents for helpful discussions on statistical topics and otherwise that have kept me sane these past four years.

DEDICATION

To my grandparents, Ada and Eric

Chapter 1

INTRODUCTION

This thesis is concerned with the study of separable covariance structures, chiefly by considering the matrix-variate and array-variate normal distributions. In the first part of the Introduction we will provide a construction for the matrix-variate normal distribution and discuss some properties that will be employed in the remaining chapters. A summary of the chapter organization follows.

Constructing the matrix-variate normal distribution

One of the most versatile, and by far the most commonly encountered, distribution in statistics is the normal distribution. Its introduction is attributed to Gauss in his "*Theoria motus corporum coelestium in sectionibus conicis solem ambientium*" in 1809 where he proposed the normal law of errors as

$$\varphi\Delta \quad = \quad \frac{h}{\sqrt{\pi}} \exp\left(-hh\Delta\Delta\right), \tag{1.1}$$

where h was the measure of precision of the observation of Δ. Employing slightly more interpretable notation than Gauss, we say that $z \sim N(0,1)$ is a univariate standard normal variable with density function

$$f_z(z) \quad = \quad \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right). \tag{1.2}$$

We can then define a normal variable with mean μ and variance σ^2 as $x \sim N\left(\mu, \sigma^2\right)$ via the linear transformation ($\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$) on z : $\mu + \sigma z$.

We can extend the univariate standard normal to a vector-variate random variable, termed the multivariate standard normal, $\underset{\sim}{z} = (z_1, \ldots, z_m) \sim N_m\left(\underset{\sim}{0}, I\right)$. Each z_i is distributed as a univariate standard normal and due to the independence of the z_is their joiny

density is given by

$$f_{\underset{\sim}{z}}\left(\underset{\sim}{z}\right) = (2\pi)^{-\frac{m}{2}} \exp\left(-\frac{1}{2}\underset{\sim}{z}^t\underset{\sim}{z}\right).$$

We can define a linear transformation on the vector, $\underset{\sim}{z}$, that is similar to the univariate case: $\underset{\sim}{z} : \underset{\sim}{\mu} + A\underset{\sim}{z}$, where $\underset{\sim}{\mu} \in \mathbb{R}^m$ and A is a (symmetric) square root of a symmetric positive definite matrix $\Sigma = AA^t$ of dimension $m \times m$. The random variable that results from this linear map is $\underset{\sim}{x} \sim N_m\left(\underset{\sim}{\mu}, AA^t\right)$, a multivariate normal variable with mean $E\left[\underset{\sim}{x}\right] = \underset{\sim}{\mu}$ and covariance matrix $\text{cov}\left[\underset{\sim}{x}\right] = \Sigma$.

The matrix-variate normal distribution is a special subfamily of the multivariate normal, possessing a separable covariance matrix. While it is possible to gain insight into the Kronecker covariance structure from the above construction of the multivariate normal, a more intuitive construction is available. As the matrix-variate normal distribution is the main focus of this dissertation we state these results formally.

Definition 1. Let $Z \sim N_{m_\text{r} \times m_\text{c}}\left(0, I_{m_\text{r}}, I_{m_\text{c}}\right)$ be a matrix where each entry is an independent standard normal variable. Due to the independence of the entries in the matrix, the density of all the entries in the matrix is a product of the individual densities. Using the standard normal density from Equation (1.2) we have

$$\begin{aligned} f_Z\left(Z\right) &= (2\pi)^{-\frac{m_\text{r} m_\text{c}}{2}} \exp\left(-\frac{1}{2}\sum_{ij} z_{ij}^2\right) \\ &= (2\pi)^{-\frac{m_\text{r} m_\text{c}}{2}} \text{etr}\left(-\frac{1}{2}\text{vec}\left(Z\right)\text{vec}\left(Z\right)^t\right) \\ &= (2\pi)^{-\frac{m_\text{r} m_\text{c}}{2}} \text{etr}\left(-\frac{1}{2}ZZ^t\right), \end{aligned}$$

where $\text{etr} = \exp \text{tr}$ is the exponentiation of the matrix trace. The second and third equalities are direct consequences of basic properties of the matrix trace. We call this the standard matrix-normal distribution.

Definition 2. Define the random variable $Y \sim N_{m_\text{r} \times m_\text{c}}\left(M, \Sigma_\text{r}, \Sigma_\text{c}\right)$ via the bilinear map $Z : M + AZB^t$ where Z has a standard matrix-variate normal distribution and $\Sigma_\text{r} = AA^t$

and $\Sigma_{\mathrm{c}} = BB^t$ are symmetric positive definite matrices of dimensions $m_{\mathrm{r}} \times m_{\mathrm{r}}$ and $m_{\mathrm{c}} \times m_{\mathrm{c}}$, respectively. The matrix M is an arbitrary $m_{\mathrm{r}} \times m_{\mathrm{c}}$ matrix. We call Y a mean M matrix variate normal variable with Kronecker covariance $\Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}}$. The Jacobian of the bilinear map is given by $|\Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}}|^{-\frac{m_{\mathrm{r}} m_{\mathrm{c}}}{2}}$ and so the density of Y is

$$f_Y(Y) = (2\pi)^{-m_{\mathrm{r}} m_{\mathrm{c}}/2} |\Sigma_{\mathrm{r}}|^{-m_{\mathrm{c}}/2} |\Sigma_{\mathrm{c}}|^{-m_{\mathrm{r}}/2} \operatorname{etr}\left(-\frac{1}{2}\Sigma_{\mathrm{r}}^{-1}(Y-M)\Sigma_{\mathrm{c}}^{-1}(Y-M)^t\right). \quad (1.3)$$

As mentioned above, the matrix normal distribution is a subfamily of the multivariate normal. The two can be related as follows:

$$Y \sim N_{m_1 \times m_2}(M, \Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}) \iff \operatorname{vec}(Y) \sim N(\operatorname{vec}(M), \Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}}). \quad (1.4)$$

To interpret this covariance structure, we note that the covariance between two entries in Y is given by $\operatorname{cov}(y_{ij}, y_{kl}) = \Sigma_{\mathrm{r,ik}}\Sigma_{\mathrm{c,jl}}$. The covariance between y_{ij} and y_{kl} is a product of a contribution from Σ_{r} (the ik entry, which corresponds to the rows of Y that y_{ij} and y_{kl} belong to), and a contribution from Σ_{c} (the jl entry, which corresponds to the columns of Y that y_{ij} and y_{kl} belong to). This provides an intuitive interpretation for the two matrices Σ_{r} and Σ_{c} as representing the covariance among the rows and among the columns of a matrix Y, respectively (hence the suggestive subscripts "r" and "c"). Furthermore, it is straightforward to show that the second moments of Y are

$$E\left[YY^t\right] = \Sigma_{\mathrm{r}} \operatorname{tr}(\Sigma_{\mathrm{c}})$$
$$E\left[Y^tY\right] = \Sigma_{\mathrm{c}} \operatorname{tr}(\Sigma_{\mathrm{r}}).$$

These identities provide further justification for the the interpretation of Σ_{r} and Σ_{c} as the covariance of the rows of Y and of the columns of Y.

Matrix variate distributions play an important role in statistics, and the matrix variate normal is no exception. The earliest considerations of the matrix normal concentrated on special cases of the distribution. James [1954] provided a characterization for the matrix normal with $\Sigma_{\mathrm{r}} = I$, that is when Y is left-spherical. Further treatments of left- and right- spherical distributions are available in Dawid [1977, 1981]. The first mention of a

matrix-variate normal distribution with row and column covariance matrices in the form that we have provided above appears in Dawid [1978]. In the 1980s, results on high order expectations of matrix variate distributions appeared in the literature [Neudecker and Wansbeek, 1987, Von rosen, 1988]. These results were closely related to the study of the distribution function of quadratic forms of the matrix variate normal as well as the notion of Wishartiness [Khatri, 1959, 1962]. Many of these results are summarized in Gupta and Nagar [1999].

A brief mention is now given to a natural extension of the matrix-variate normal distribution to the array normal distribution which is discussed in Chapters 2 and 4 of this thesis. The array normal distribution exhibits a separable covariance structure akin to the matrix normal described above. While the matrix normal distribution has only two modes (rows and columns), the array normal can accommodate any number of modes, where the covariance along all the modes is combined via the Kronecker product to form the covariance of the whole array. This is best described via an equivalence to the multivariate normal (as in (1.4)):

$$Y \sim N_{m_1,\ldots,m_p}(M, \Sigma_1, \ldots, \Sigma_p) \iff \text{vec}(Y) \sim N(\text{vec}(M), \Sigma_p \otimes \cdots \otimes \Sigma_1).$$

Many of the properties of the matrix-variate normal distribution extend to the array-variate normal. For example, second moments (where second moments are conveniently written via the matricization operator of Kolda [2006]) lend similar interpretation to the covariance matrices along each mode. The distribution has recently been treated in detail by Hoff [2011].

Summary of chapter organization

The next two chapters concentrate on the study of relational data within the framework of a square matrix variate normal distribution. Relational data are often represented as a square matrix, the entries of which record the relationships between pairs of objects. Many statistical methods for the analysis of such data assume some degree of similarity or dependence between objects in terms of the way they relate to each other. However, formal

tests for such dependence have not been developed. In Chapter 2 we provide a test for such dependence using the framework of the matrix normal model, a type of multivariate normal distribution parameterized in terms of row- and column-specific covariance matrices. We develop a likelihood ratio test (LRT) for row and column dependence based on the observation of a single relational data matrix. We obtain a reference distribution for the LRT statistic, thereby providing an exact test for the presence of row or column correlations in a square relational data matrix. Additionally, we provide extensions of the test to accommodate common features of such data, such as undefined diagonal entries, a non-zero mean, multiple observations, and deviations from normality.

The rejection of the null hypothesis by the test of Chapter 2 leads to an inference problem: how does one account for the row and column correlation that is evident in the data? In Chapter 3 we provide a framework for estimating the separable covariance structure in the context of a single observation from a matrix-variate normal distribution. We first describe covariance estimators in the known mean case. We concentrate on the classes of maximum likelihood estimators and maximum penalized likelihood estimators. We present theoretical results for several subproblems and develop a novel penalty for the similarity between the covariance matrices for the rows and columns of the relational data matrix. Next we extend these results to the case of an unknown mean. In the case of the unpenalized estimators of the covariance, a one-step feasible GLS approach is presented, as the likelihood is unbounded when estimating mean parameters and full row and column covariance matrices. On the other hand, for the penalized methods an iterative estimation procedure is proposed. Theoretical guarantees for the convergence of the optimization procedure and for the unbiasedness of the estimates of the mean parameters are conjectured.

In Chapter 4 we leave the realm of relational data to discuss an application of matrix-variate and array-variate normal distributions to ANOVA decompositions. ANOVA decompositions are a standard method for describing and estimating heterogeneity among the means of a response variable across levels of multiple categorical factors. In such a decomposition, the complete set of main effects and interaction terms can be viewed as a collection of vectors, matrices and arrays that share various index sets defined by the factor levels. For many types of categorical factors, it is plausible that an ANOVA decomposition exhibits

some consistency across orders of effects, in that the levels of a factor that have similar main-effect coefficients may also have similar coefficients in higher-order interaction terms. In such a case, estimation of the higher-order interactions should be improved by borrowing information from the main effects and lower-order interactions. To take advantage of such patterns, this chapter introduces a class of hierarchical prior distributions for collections of interaction arrays that can adapt to the presence of such interactions. These prior distributions are based on a type of array-variate normal distribution, for which a covariance matrix for each factor is estimated. This prior is able to adapt to potential similarities among the levels of a factor, and incorporate any such information into the estimation of the effects in which the factor appears. In the presence of such similarities, this prior is able to borrow information from well-estimated main effects and lower-order interactions to assist in the estimation of higher-order terms for which data information is limited.

A discussion and directions for future work are presented in Chapter 5. Appendix A provides proofs for the theorems of Chapter 2 and Appendix B provides a brief introduction to metrics on the space of positive definite matrices.

Chapter 2

TESTING FOR NODAL DEPENDENCE IN RELATIONAL DATA MATRICES

2.1 Introduction

Networks or relational data among m actors, nodes or objects are frequently presented in the form of an $m \times m$ matrix $Y = \{y_{ij} : 1 \leq i, j \leq m\}$, where the entry y_{ij} corresponds to a measure of the directed relationship from object i to object j. Such data are of interest in a variety of scientific disciplines: Sociologists and epidemiologists gather friendship network data to study social development and health outcomes among children [Fletcher et al., 2011, Pollard et al., 2010, Potter et al., 2012, Van De Bunt et al., 1999], economists study markets by analyzing networks of business interactions among companies or countries [Westveld and Hoff, 2011, Lazzarini et al., 2001], and biologists study gene-gene interaction networks to better understand biological pathways [Bergmann et al., 2003, Stuart et al., 2003].

Often of interest in the study of such data is a description of the variation and similarity among the objects in terms of their relations. Similarities among rows and among columns in empirical networks have long been observed [Sampson, 1968, Leskovec et al., 2008], leading to the development of statistical tools to summarize such patterns. CONCOR (CONvergence of iterated CORrelations) is an early example of a procedure that partitions the rows (or columns) of Y into groups based on a summary of the correlations among the rows (or columns) of Y [White et al., 1976, McQuitty and Clark, 1968]. The procedure yields a "blockmodel" of the objects, a representation of the original data matrix Y by a smaller matrix that identifies relationships among groups of objects. While this algorithm is still commonly used [Lincoln and Gerlach, 2004, Lafosse and Ten Berge, 2006], it suffers from a lack of statistical interpretability [Panning, 1982], as it is not tied to any particular statistical model or inferential goal.

Several model-based approaches presume the existence of a grouping of the objects such

that objects within a group share a common distribution for their outgoing relationships. This is the notion of stochastic equivalence, and is the primary assumption of stochastic blockmodels, a class of models for which the probability of a relationship between two objects depends only on their individual group memberships [Holland et al., 1983, Wang and Wong, 1987, Nowicki and Snijders, 2001, Rohe et al., 2011]. Airoldi et al. [2008] extend the basic blockmodel by allowing each object to belong to several groups. In this model the probability of a relationship between two nodes depends on all the group memberships of each object. This and other variants of stochastic blockmodels belong to the larger class of latent variable models, in which the probability distribution of the relationship between any two objects i and j depends on unobserved object-specific latent characteristics z_i and z_j [Hoff et al., 2002]. Statistical models of this type all presume some form of similarity among the objects in the network. However, while such models are widely used and studied, no formal test for similarities among the objects in terms of their relations has been proposed.

Many statistical methods for valued or continuous relational data are developed in the context of normal statistical models. These include, for example, the widely-used social relations model [Kenny and La Voie, 1984, Li and Loken, 2002] and covariance models for multivariate relational data [Li, 2006, Westveld and Hoff, 2011, Hoff, 2011]. Additionally, statistical models for binary and ordinal relational data can be based on latent normal random variables via probit or other link functions [Hoff, 2005b, 2008]. In this chapter we propose a novel approach to testing for similarities between objects in terms of the row and column correlation parameters of the matrix normal model. The matrix normal model consists of the multivariate normal distributions that have a Kronecker-structured covariance matrix [Dawid, 1981]. Specifically, we say that an $m \times m$ random matrix Y has the mean-zero matrix normal distribution $N_{m\times m}(0, \Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}})$ if $\mathrm{vec}\,(Y) \sim N_{m^2}\,(0, \Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}})$ where "vec" is the vectorization operator and "$\otimes$" denotes the Kronecker product. Under this distribution, the covariance between two relations y_{ij} and y_{kl} is given by $\mathrm{cov}\,(y_{ij}, y_{kl}) = \Sigma_{\mathrm{r},ik}\Sigma_{\mathrm{c},jl}$. Furthermore, it is straightforward to show that

$$E\left[YY^t\right] = \Sigma_{\mathrm{r}}\mathrm{tr}\,(\Sigma_{\mathrm{c}}) \text{ and } E\left[Y^tY\right] = \Sigma_{\mathrm{c}}\mathrm{tr}\,(\Sigma_{\mathrm{r}})\,.$$

These identities suggest the interpretation of $\Sigma_{\rm r}$ and $\Sigma_{\rm c}$ as the covariance of the objects as senders of ties and as receivers of ties, respectively. In this chapter, we evaluate evidence for similarities between objects by testing for non-zero correlations in this matrix normal model. Specifically, we develop a test of

$$H_0 : \ (\Sigma_{\rm r}, \Sigma_{\rm c}) \in \mathcal{D}_+^m \times \mathcal{D}_+^m \ \text{ versus } H_1 : \ (\Sigma_{\rm r}, \Sigma_{\rm c}) \in (\mathcal{S}_+^m \times \mathcal{S}_+^m) \backslash (\mathcal{D}_+^m \times \mathcal{D}_+^m)$$

where $\mathcal{D}_+^m$ is the set of $m \times m$ diagonal matrices with positive entries and $\mathcal{S}_+^m$ is the set of $m \times m$ positive definite symmetric matrices. Model H_0, which we call the Kronecker variance model, represents heteroscedasticity among the rows and the columns while still maintaining their independence. Model H_1, which we call the full Kronecker covariance model, allows for correlations between all of the rows and all the of columns. Rejection of the null of zero correlation would support further inference via a model that allowed for similarities among the objects, such as a stochastic blockmodel, some other latent variable model or the matrix normal model. Acceptance of the null would caution against fitting such a model in order to avoid spurious inferences.

This goal of evaluating the evidence for row or column correlation is in contrast to that of the existing testing literature for matrix normal distributions. This literature has focused on an evaluation of the null hypothesis that $\text{cov}(\text{vec}(Y)) = \Sigma_{\rm c} \otimes \Sigma_{\rm r}$ (our H_1) against an unstructured alternative [Roy and Khattree, 2005, Mitchell et al., 2006, Lu and Zimmerman, 2005, Srivastava et al., 2008]. The tests proposed in this literature are likelihood ratio tests that, in the case of an $m \times m$ square matrix Y, require at least $n > m^2$ replications to estimate the covariance under the fully unstructured model. Such tests are not applicable to most relational datasets, which typically consist of at most a few observed relational matrices.

In the next section we derive a hypothesis test of H_0 versus H_1 in the context of the matrix normal model. We show that given a single observed relational matrix Y, the likelihood is bounded under both H_0 and H_1 and so a likelihood ratio test of H_0 against H_1 can be constructed. We further show how the null distribution of the test statistic can be approximated with an arbitrarily high precision via a Monte Carlo procedure. In Section

2.2.3 we extend these results to the general class of matrix variate elliptically contoured distributions. The power of the test in several different situations is evaluated in Section 2.3.

Although the development of our testing procedure is based on the mean-zero matrix normal distribution, it is straightforward to extend the test to several other scenarios commonly encountered in the study of relational data, including missing diagonal entries, non-zero mean structure, multiple heteroscedastic observations and binary networks. These extensions and two data examples are discussed in Section 2.4. A discussion follows in Section 5.

2.2 Likelihood ratio test

In this section we propose a likelihood ratio test (LRT) for evaluating the presence of correlations among the rows and correlations among the columns of a square matrix. The data matrix Y is modeled as a draw from a mean zero matrix normal distribution $N_{m\times m}(0, \Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}})$. The parameter space under the null hypothesis H_0 is $\Theta_0 = \mathcal{D}_+^m \times \mathcal{D}_+^m$, the space of all pairs of diagonal $m \times m$ matrices with positive entries. Under the alternative H_1, the parameter space is $\Theta_1 = (\mathcal{S}_+^m \times \mathcal{S}_+^m) \backslash (\mathcal{D}_+^m \times \mathcal{D}_+^m)$, the collection of all pairs of positive definite matrices of dimension m for which at least one is not diagonal. To derive the LRT statistic, we first obtain the maximum likelihood estimates (MLEs) under the unrestricted parameter space $\Theta = \Theta_0 \cup \Theta_1$ and under the null parameter space Θ_0. From these MLEs, we construct several equivalent forms of the LRT statistic. While the null distribution of the test statistic is not available in closed form, the statistic is invariant under diagonal rescalings of the data matrix Y, implying that the distribution of the statistic is constant as a function of $(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}) \in \Theta_0$. This fact allows us to obtain null distributions and p-values via Monte Carlo simulation.

2.2.1 Maximum likelihood estimates

The density of a mean zero matrix normal distribution $N_{m\times m}(0, \Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}})$ is given by

$$p\left(Y|\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}\right) \;=\; (2\pi)^{-m^2/2} \left|\Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}}\right|^{-1/2} \exp(-\tfrac{1}{2}\mathrm{tr}\left(\Sigma_{\mathrm{r}}^{-1} Y \Sigma_{\mathrm{c}}^{-1} Y^t\right)),$$

where "tr" is the matrix trace and "$\otimes$" is the Kronecker product. Throughout the chapter we will write $l\left(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}; Y\right)$ as minus two times the log likelihood minus $m^2 \log 2\pi$, hereafter referred to as the scaled log likelihood:

$$l\left(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}; Y\right) \quad = \quad -2 \log p\left(Y | \Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}\right) - m^2 \log 2\pi = \mathrm{tr}\left[\Sigma_{\mathrm{r}}^{-1} Y \Sigma_{\mathrm{c}}^{-1} Y^t\right] - \log\left|\Sigma_{\mathrm{c}}^{-1} \otimes \Sigma_{\mathrm{r}}^{-1}\right|.$$

We will state all the results in this paper in terms of $l\left(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}; Y\right)$. For example, an MLE will be a minimizer of $l\left(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}; Y\right)$ in $\left(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}\right)$. The following result implies that if Y is a draw from an absolutely continuous distribution on $\mathbb{R}^{m \times m}$, then the scaled likelihood is bounded from below, and achieves this bound on a set of nonunique MLEs:

Theorem 1. *If Y is full rank then $l(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}; Y) \geq m^2 + m \log |YY^t/m|$ for all $(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}) \in \Theta$, with equality if $\Sigma_{\mathrm{r}} = Y\Sigma_{\mathrm{c}}^{-1}Y^t/m$, or equivalently, $\Sigma_{\mathrm{c}} = Y^t\Sigma_{\mathrm{r}}^{-1}Y/m$.*

Proof. We first look for MLEs at the critical points of $l\left(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}; Y\right)$. Setting derivatives of l to zero indicates that critical points satisfy

$$\hat{\Sigma}_{\mathrm{r}} \quad = \quad Y\hat{\Sigma}_{\mathrm{c}}^{-1}Y^t/m \tag{2.1}$$

$$\hat{\Sigma}_{\mathrm{c}} \quad = \quad Y^t\hat{\Sigma}_{\mathrm{r}}^{-1}Y/m. \tag{2.2}$$

Note that these equations are redundant: If $(\hat{\Sigma}_{\mathrm{r}}, \hat{\Sigma}_{\mathrm{c}})$ satisfy Equation 2.1, then these values satisfy Equation 2.2 as well. The value of the scaled log likelihood at such a critical point is

$$\begin{aligned} l\left(Y\hat{\Sigma}_{\mathrm{c}}^{-1}Y^t/m, \hat{\Sigma}_{\mathrm{c}}; Y\right) &= \mathrm{tr}\left[\left(Y\hat{\Sigma}_{\mathrm{c}}^{-1}Y^t/m\right)^{-1} Y\hat{\Sigma}_{\mathrm{c}}^{-1}Y^t\right] + \log\left|\hat{\Sigma}_{\mathrm{c}} \otimes Y\hat{\Sigma}_{\mathrm{c}}^{-1}Y^t/m\right| \\ &= m\mathrm{tr}\left[Y^{-t}\hat{\Sigma}_{\mathrm{c}}Y^{-1}Y\hat{\Sigma}_{\mathrm{c}}^{-1}Y^t\right] + m\log\left|\hat{\Sigma}_{\mathrm{c}}\right| \\ &\quad -m\log\left|\hat{\Sigma}_{\mathrm{c}}\right| + m\log\left|YY^t/m\right| \\ &= m\mathrm{tr}\left[I\right] + m\log\left|YY^t/m\right|. \end{aligned} \tag{2.3}$$

In the second and third lines, Y^{-1} exists and $|YY^t/m| > 0$ since Y is square and full rank.

Now we compare the scaled log likelihood at a critical point to its value at any other point.

$$\begin{aligned} l\left(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}; Y\right) - l(Y\hat{\Sigma}_{\mathrm{c}}^{-1}Y^t/m, \hat{\Sigma}_{\mathrm{c}}; Y) &= \operatorname{tr}\left[\Sigma_{\mathrm{r}}^{-1}Y\Sigma_{\mathrm{c}}^{-1}Y^t\right] + m\log\left(|\Sigma_{\mathrm{r}}|\,|\Sigma_{\mathrm{c}}|\right) \\ &\quad - m^2 - m\log\left|YY^t/m\right| \\ &= m^2\frac{1}{m}\operatorname{tr}\left[\Sigma_{\mathrm{r}}^{-1}Y\Sigma_{\mathrm{c}}^{-1}Y^t/m\right] \\ &\quad - m^2\left[\frac{1}{m}\log\left|\Sigma_{\mathrm{r}}^{-1}Y\Sigma_{\mathrm{c}}^{-1}Y^t/m\right| - 1\right] \end{aligned} \tag{2.4}$$

The first equality is a simple combination of equations 2.1 and 2.3. The second equality is a rearrangement of terms that combines all the determinant in the log terms. This difference can be written as $m^2(a - \log g - 1)$, where a is the arithmetic mean and g is the geometric mean of the eigenvalues of $(\Sigma_{\mathrm{r}}^{-1/2}Y\Sigma_{\mathrm{c}}^{-1}Y^t\Sigma_{\mathrm{r}}^{-1/2})/m$. To complete the proof we show that $a - \log g - 1 \geq 0$. Consider $f\left(x\right) = x - 1 - \log x$ and its first and second derivatives with respect to x: $f'\left(x\right) = 1 - \frac{1}{x}$, and $f''\left(x\right) = \frac{1}{x^2}$. The second derivative is positive at the critical point $x = 1$, so $f\left(1\right) = 0$ is a global minimum of the function. Thus $x - \log x - 1 \geq 0$. Now let $\lambda_1, \ldots, \lambda_m$ be the eigenvalues of $(\Sigma_{\mathrm{r}}^{-1/2}Y\Sigma_{\mathrm{c}}^{-1}Y^t\Sigma_{\mathrm{r}}^{-1/2})/m$ and so $a = \frac{1}{m}\sum\lambda_i$ and $g = \left(\prod\lambda_i\right)^{1/m}$. We then have

$$a \geq \log a + 1 = \log\left(\frac{1}{m}\sum x_i\right) + 1 \geq \log\left(\left(\prod x_i\right)^{1/m}\right) + 1 = \log g + 1$$

as $a \geq g$ since $\lambda_i \geq 0\ \forall i$. Since $a - 1 - \log g \geq 0$ we have the desired result. □

Note that the MLE is not unique, nor is the MLE of $\Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}}$. For example. $I \otimes YY^t/m$ is an MLE of $\Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}}$, as is $Y^tY \otimes I/m$. Moreover, there is an MLE for each $\Sigma_{\mathrm{r}} \in \mathcal{S}_+^m$ given by $(\Sigma_{\mathrm{r}}, Y^t\Sigma_{\mathrm{r}}^{-1}Y/m)$, and similarly there is an MLE for each $\Sigma_{\mathrm{c}} \in \mathcal{S}_+^m$ given by $(Y\Sigma_{\mathrm{c}}^{-1}Y^t/m, \Sigma_{\mathrm{c}})$.

Theorem 1 also implies that the likelihood is bounded under the null. Unlike the unrestricted case, the MLE under the null is unique up to scalar multiplication:

Theorem 2. *If Y is full rank then the MLE $\hat{D}_{\mathrm{c}} \otimes \hat{D}_{\mathrm{r}}$ under H_0 is unique, while $\hat{D}_{\mathrm{r}}$ and $\hat{D}_{\mathrm{c}}$ are unique up to a multiplication and division by the same positive scalar.*

A proof is given in the Appendix. To find the MLE under the null model, we obtain

the derivatives of the scaled log likelihood l with respect to $(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}) \in \Theta_0$. For notational convenience, we will refer to diagonal versions of Σ_{r} and Σ_{c} as D_{r} and D_{c} respectively. Setting these derivatives equal to zero, we establish that the critical points of l must satisfy

$$D_{\mathrm{r}} = Y D_{\mathrm{c}}^{-1} Y^t \circ I/m \tag{2.5}$$

$$D_{\mathrm{c}} = Y^t D_{\mathrm{r}}^{-1} Y \circ I/m \tag{2.6}$$

where "$\circ$" is the Hadamard product. The MLE can be found by iteratively solving equations (2.5) and (2.6). This procedure can be seen as a type of block coordinate descent algorithm, decreasing l at each iteration [Tseng, 2001].

2.2.2 Likelihood ratio test statistic and null distribution

Since the scaled log likelihood is bounded below, we are able to obtain a likelihood ratio statistic that is finite with probability 1 when Y is sampled from an absolutely continuous distribution on $\mathbb{R}^{m \times m}$. As usual, a likelihood ratio test statistic can be obtained from the ratio of the unrestricted maximized likelihood to the likelihood maximized under the null. We take our test statistic to be

$$T(Y) = l(\hat{D}_{\mathrm{r}}, \hat{D}_{\mathrm{c}}; Y) - l(\hat{\Sigma}_{\mathrm{r}}, \hat{\Sigma}_{\mathrm{c}}; Y),$$

where $(\hat{\Sigma}_{\mathrm{r}}, \hat{\Sigma}_{\mathrm{c}})$ is any unrestricted MLE and $(\hat{D}_{\mathrm{r}}, \hat{D}_{\mathrm{c}})$ is the MLE under Θ_0. Since the scaled log likelihood l is minus two times the likelihood, our statistic is a monotonically increasing function of the likelihood ratio.

In Theorem 1 we showed that $l(\hat{\Sigma}_{\mathrm{r}}, \hat{\Sigma}_{\mathrm{c}}; Y) = m^2 + m \log |YY^t/m|$ for any unrestricted

MLE $(\hat{\Sigma}_{\mathrm{r}}, \hat{\Sigma}_{\mathrm{c}})$. Similarly, letting $(\hat{D}_{\mathrm{r}}, \hat{D}_{\mathrm{c}})$ be an MLE under H_0, we have

$$\begin{aligned} l(\hat{D}_{\mathrm{r}}, \hat{D}_{\mathrm{c}}; Y) &= \operatorname{tr}\left[Y^t \hat{D}_{\mathrm{r}}^{-1} Y \hat{D}_{\mathrm{c}}^{-1}\right] + \log\left|\hat{D}_{\mathrm{c}} \otimes \hat{D}_{\mathrm{r}}\right| \\ &= \operatorname{tr}\left[(Y^t \hat{D}_{\mathrm{r}}^{-1} Y)(Y^t \hat{D}_{\mathrm{r}}^{-1} Y/m \circ I)^{-1}\right] + \log\left|\hat{D}_{\mathrm{c}} \otimes \hat{D}_{\mathrm{r}}\right| \\ &= \sum_i (Y^t \hat{D}_{\mathrm{r}}^{-1} Y)_{ii} (Y^t \hat{D}_{\mathrm{r}}^{-1} Y/m \circ I)_{ii}^{-1} + \log\left|\hat{D}_{\mathrm{c}} \otimes \hat{D}_{\mathrm{r}}\right| \\ &= m \sum_i (Y^t \hat{D}_{\mathrm{r}}^{-1} Y)_{ii} / (Y^t \hat{D}_{\mathrm{r}}^{-1} Y)_{ii} + \log\left|\hat{D}_{\mathrm{c}} \otimes \hat{D}_{\mathrm{r}}\right| = m^2 + \log\left|\hat{D}_{\mathrm{c}} \otimes \hat{D}_{\mathrm{r}}\right|, \end{aligned}$$

where the second equality stems from MLE satisfying $\hat{D}_{\mathrm{c}} = Y^t \hat{D}_{\mathrm{r}}^{-1} Y/m \circ I$ (Equation (2.6)). The third equality relies on the following identity for traces: for a diagonal matrix A and unstructured matrix B of the same dimension, $\operatorname{tr}[AB] = \sum_i A_{ii} B_{ii}$. The final line is due to the following identity for Hadamard products: if I is the identity matrix and B is an unstructured matrix, then $(B \circ I)_{ii}^{-1} = 1/B_{ii}$.

The maximized likelihoods under the null and alternative give

$$\begin{aligned} T(Y) &= \log\left|\hat{D}_{\mathrm{c}} \otimes \hat{D}_{\mathrm{r}}\right| - m \log\left|YY^t/m\right| \\ &= m\left(\log\left|\hat{D}_{\mathrm{c}}\right| + \log\left|\hat{D}_{\mathrm{r}}\right| - \log\left|YY^t/m\right|\right). \end{aligned} \tag{2.7}$$

Since no closed form solution exists for $\hat{D}_{\mathrm{r}}$ it is not clear how to obtain the null distributions of T in closed form. However, it is possible to simulate from the null distribution of $T(Y)$, as the distribution of the test statistic is the same for all elements of the null hypothesis. To see this, we show that the test statistic itself is invariant under left and right transformations of the data by positive diagonal matrices. Let $\tilde{Y} = D_1 Y D_2$ for positive diagonal matrices D_1 and D_2. Since the MLE $\hat{D}_{\mathrm{c}} \otimes \hat{D}_{\mathrm{r}}$ is unique it is an equivariant function of any matrix Y with respect to left and right multiplication by diagonal matrices (see Eaton [1983] Prop 7.11). In particular, writing $\hat{\theta}(Y)$ for the estimate of $\hat{D}_{\mathrm{c}} \otimes \hat{D}_{\mathrm{r}}$ based on a data matrix Y, we have

$$\hat{\theta}\left(\tilde{Y}\right) = \left(D_2^{1/2} \otimes D_1^{1/2}\right) \hat{\theta}(Y) \left(D_2^{1/2} \otimes D_1^{1/2}\right).$$

Since the determinant is a multiplicative map, we can write the determinant of the above as

$$\left|\hat{\theta}\left(\tilde{Y}\right)\right| \;=\; |(D_2 \otimes D_1)| \left|\hat{\theta}\left(Y\right)\right|.$$

Using the above, the $T\left(\tilde{Y}\right)$ can be written as in Equation (2.7):

$$\begin{aligned} T\left(\tilde{Y}\right) &= m \log\left|\hat{\theta}\left(\tilde{Y}\right)\right| - m \log\left|\tilde{Y}\tilde{Y}^t/m\right| \\ &= m \log\left|\hat{\theta}\left(Y\right)\right| + m \log|D_2 \otimes D_1| - m \log\left|YY^t/m\right| - m \log|D_2 \otimes D_1| = T\left(Y\right). \end{aligned}$$

Since for a matrix normal random variable $Y \sim N_{m\times m}\left(0, \Sigma_1, \Sigma_2\right)$ we have $Y \stackrel{d}{=} \Sigma_1^{1/2} Y_0 \Sigma_2^{1/2}$ for $Y_0 \sim N_{m\times m}\left(0, I, I\right)$, the above argument implies that $T\left(Y\right) \stackrel{d}{=} T\left(Y_0\right)$ under the null. Therefore, the null distribution of T can be approximated via Monte Carlo simulation of Y from any distribution in H_0. For example, a Monte Carlo approximation to the q^{th} quantile, T_q can be obtained from the following algorithm:

1. Simulate $Y_0^1, \ldots, Y_0^S \sim$ i.i.d $N_{m\times m}\left(0, I, I\right)$;

2. Let $\hat{T}_q = \min\{T(Y_0^q) : \sum_{s=1}^{S} 1[T(Y_0^q) \geq T(Y_0^s)]/S \geq q\}$.

2.2.3 Matrix variate elliptically contoured distributions

The results of the previous subsection are immediately extendable to the general class of matrix variate elliptically contoured distributions. In this section we show that under minor regularity conditions on the distributions, the likelihood for a matrix variate elliptically contoured distribution is bounded when the matrix normal distribution is bounded. We provide the form of an MLE for the general class of mean zero square matrix variate elliptically contoured distributions and demonstrate that the likelihood ratio test between H_0 and H_1 has the same form as in Equation (2.7).

We use the notation of Gupta and Varga [1994] for the matrix variate elliptically contoured distribution. We say that Y has a mean zero square matrix variate elliptically

countoured distribution and write $Y \sim E_{m\times m}(0, \Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}, h)$ if its density has the form

$$f_Y(Y) = \frac{1}{|\Sigma_{\mathrm{r}}|^{\frac{m}{2}} |\Sigma_{\mathrm{c}}|^{\frac{m}{2}}} h\left(\mathrm{tr}\left[Y^t \Sigma_{\mathrm{r}}^{-1} Y \Sigma_{\mathrm{c}}^{-1}\right]\right) = \frac{1}{|\Sigma_{\mathrm{r}}|^{\frac{m}{2}} |\Sigma_{\mathrm{c}}|^{\frac{m}{2}}} h\left(\mathrm{vec}(Y)^t \left(\Sigma_{\mathrm{c}}^{-1} \otimes \Sigma_{\mathrm{r}}^{-1}\right) \mathrm{vec}(Y)\right). \tag{2.8}$$

For $h(w) = (2\pi)^{m^2/2} \exp\left(-\frac{1}{2}w\right)$, Y is a mean zero matrix variate normal variable. Gupta and Varga [1994] show that if $Y \sim E(0, \Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}, h)$ then $AYB \sim E\left(0, A\Sigma_{\mathrm{r}}A^t, B\Sigma_{\mathrm{c}}B^t, h\right)$ and if second moments exist, then $\mathrm{cov}(\mathrm{vec}(Y)) = c_h(\Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}})$. In Gupta and Varga [1995], the authors showed that when Σ_{r} and h are known, the MLE of Σ_{c} is proportional to the MLE of Σ_{c} under normality, but they do not provide results for the boundedness of the likelihood or existence of MLEs for the case where only h is known. To find the form of the MLE in this case we state a simplified version of Theorem 1 of Anderson et al. [1986]:

Theorem. *Let Ω be a set in the space of $\mathcal{S}_+^{m^2}$, such that if $V \in \Omega$ then $cV \in \Omega$ $\forall c > 0$ (that is Ω is a cone). Suppose h is such that $h\left(y^t y\right)$ is a density in $\mathbb{R}^{m^2}$ and $x^{m^2/2}h(x)$ has a finite positive maximum x_h. Suppose that on the basis of an observation y from $|V|^{-1/2} h\left(y^t V^{-1} y\right)$ an MLE under normality $\tilde{V} \in \Omega$ exists and $\tilde{V} > 0$ with probability 1. Then an MLE for h is $\hat{V} = \left(m^2/x_h\right)\tilde{V}$ and the maximum of the likelihood is $\left|\hat{V}\right|^{-1/2} h(x_h)$.*

In the previous subsection we proved that for the mean zero square matrix normal distribution, the likelihood is bounded for a single observation. A direct application of the above theorem with $\Omega = \mathcal{S}_+^m \times \mathcal{S}_+^m \subset \mathcal{S}_+^{m^2}$ and $y = \mathrm{vec}(Y)$ proves that for the likelihood of a generic matrix variate elliptically contoured distribution with h defined as in 2.8, the likelihood is bounded and the MLE of cov(vec(Y)) is proportional to the MLE under normality. Clearly the theorem hold for a smaller space $\omega = \mathcal{D}_+^m \times \mathcal{D}_+^m \subset \mathcal{S}_+^{m^2}$ as well, and thus the likelihood ratio statistic can be constructed as follows

$$\begin{aligned} T(Y) &= 2\log\left[\left|\hat{V}_\Omega\right|^{-1/2} h(x_h)\right] - 2\log\left[\left|\hat{V}_\omega\right|^{-1/2} h(x_h)\right] \\ &= \log\left|\hat{V}_\omega\right| - \log\left|\hat{V}_\Omega\right| = \log|\tilde{V}_\omega| - \log|\tilde{V}_\Omega| \end{aligned} \tag{2.9}$$

where $\tilde{V}_\omega$ and $\tilde{V}_\Omega$ are the MLEs under normality that were previously derived. Equation

Dimension m	5	10	15	20	25	30	50	100
95% quantile	43.3	144.3	297.4	502.8	760.0	1064.6	2802.1	10668.4

Table 2.1: 95% quantile of the null distribution of the test statistic for testing H_0 versus H_1. Approximation from 100,000 simulated values.

(2.9) is identical to the original form of the test (*e.g.* Equation (2.7)). As such, to conduct the test for any elliptically contoured distribtuion, we can construct a reference distribution for the null based on a mean zero matrix variate distribution.

2.3 Power calculations

In this section we present power calculations for three different types of covariance models. The three covariance models we consider are: (1) Exchangeable row covariance and exchangeable column covariance; (2) Maximally sparse Kronecker structured covariance; and (3) the covariance induced by a nonseparable stochastic blockmodel with two row groups and two column groups. For each covariance model, we consider the power as a function of parameters that control the total correlation within a covariance matrix as well as in terms of m, the dimension of the matrix. In Table 2.1 we present the 95% quantiles based on the null distributions required for performing level $\alpha = 0.05$ tests.

2.3.1 Exchangeable row and column covariance structure

We first consider a submodel of the matrix normal model in which Σ_{r} and Σ_{c} have exchangeable covariance structure. In this structure, the correlation between any two rows is a constant ρ_{r} and the correlation between any two columns is a constant ρ_{c}. Specifically, $\mathrm{cov}(\mathrm{vec}(Y)) = \Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}}$ where

$$\Sigma_{\mathrm{r}} = (1 - \rho_{\mathrm{r}})\, I + \rho_{\mathrm{r}} \mathbf{1}\mathbf{1}^t \text{ and } \Sigma_{\mathrm{c}} = (1 - \rho_{\mathrm{c}})\, I + \rho_{\mathrm{c}} \mathbf{1}\mathbf{1}^t,$$

and $\mathbf{1}$ is a vector of ones of length m. We first consider a network with $m = 10$ nodes and present the power as a function of ρ_{r} and ρ_{c} ranging from $-1/9$ to 1, where the lower bound

guarantees that the covariance matrices are positive definite. We calculate the power on a 25×25 grid in $[-1/9, 1]^2$ and use a bivariate interpolation to construct the heatmap in the top left panel of Figure 2.1. From the plot it is evident that the power is an increasing function in $|\rho_{\mathrm{r}}|$ and $|\rho_{\mathrm{c}}|$. In particular, keeping ρ_{r} constant, the power is an increasing function of $|\rho_{\mathrm{c}}|$ and vice versa.

In the top left panel of Figure 2.1 we observed that while keeping ρ_{c} constant, the power is an increasing function of $|\rho_{\mathrm{r}}|$. To study the power for higher dimensional matrices, we set $\rho_{\mathrm{c}} = 0$ and vary m and ρ_{r}. The dashed line in the top left panel of Figure 2.1 traces the power function for $m = 10$ and $\rho_{\mathrm{c}} = 0$. This corresponds to the same style dashed line in the top right hand panel of Figure 2.1. The other four lines in the right hand panel represent the calculated power as a function of ρ_{r} for different dimensions m, holding $\rho_{\mathrm{c}} = 0$. As is expected, for each m, the power is an increasing function of $|\rho_{\mathrm{r}}|$. Similarly, for each fixed ρ_{r} value, the power is an increasing function of m. This latter phenomenon is due to the increase in the amount of data information with the increase in the dimension of the sociomatrix.

2.3.2 Maximally sparse Kronecker covariance structured correlation

While the previous example demonstrates the power of the test in the presence of many nonzero off-diagonal entries in the correlation matrices, it is of interest to see if the test has any power against alternatives that do not exhibit a large amount of correlation. For this purpose we consider a maximally sparse Kronecker covariance structure. We set the columns to be independent and only the first two rows to be correlated. This can be written compactly as $\Sigma_{\mathrm{c}} = I$ and $\Sigma_{\mathrm{r}} = I + \rho E_{12} + \rho E_{21}$, where E_{ij} is the 0 matrix with a 1 in the $(i,j)^{\mathrm{th}}$ entry. For each of the matrix sizes $m \in \{5, 10, 25, 50, 100\}$ we computed power functions for values of ρ ranging between -1 and 1. Monte Carlo approximations to the corresponding power functions are presented in the bottom left panel of Figure 2.1. We plot the results for $m = 5$ and $m = 100$ and see that the power increases monotonically as a function of $|\rho|$ for both dimensions. While the power of the test for a fixed ρ appears to decrease as the size of the network m increases, the two curves are nearly identical. We

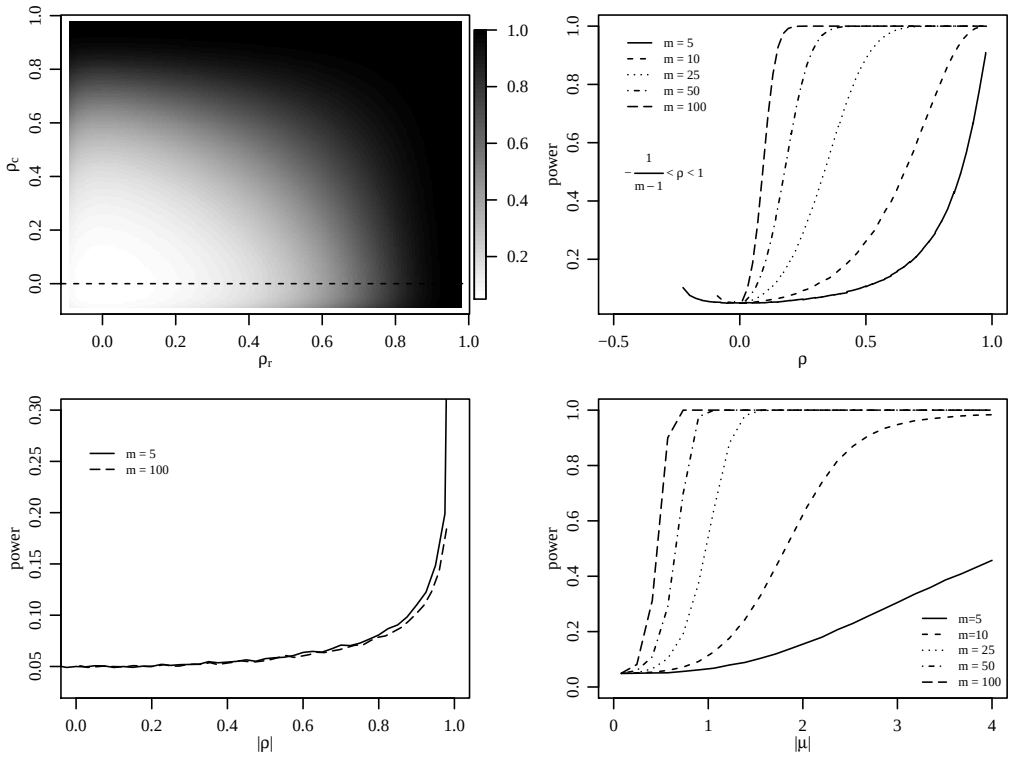

Figure 2.1: The top row of panels displays the power of the test under the exchangeable covariance model of Section 2.3.1. The bottom left panel displays the power of the test for the maximally sparse covariance model of Section 2.3.2 and the bottom right panel displays the power of the test for the nonseparable stochastic blockmodel of Section 2.3.3.

explain this as follows: while one expects that as the dimension m increases there is an increase in data information for identifying the correlation ρ, the power curve is influenced more heavily by the fact that the difference between Σ_{r} and the identity matrix becomes less pronounced. Additional power curves for a range of m values between 5 and 100 were approximated. All the curves were between the $m = 5$ and $m = 100$ curves that are presented in the plot. For all the power calculations, the lowest calculated power for values of ρ close to 0 was always within two Monte Carlo standard errors of 0.05.

2.3.3 Misspecified covariance structure

In the Introduction we discussed a popular model for relational data with an underlying assumption of stochastically equivalent nodes called the stochastic blockmodel. Straightforward calculations show that in general the covariance induced by a stochastic blockmodel is nonseparable, but still induces correlations among the rows and among the columns. As such, we are interested in evaluating the power of our test against such nonseparable alternatives.

A stochastic blockmodel can be represented in terms of multiplicative latent variables. Specifically, we can write the relationship $y_{ij} = u_i^t W v_j + \epsilon_{ij}$ where u_i and v_j are latent vectors representing the row group membership of node i and the column group membership of node j. W is a matrix of means for the different group memberships and ϵ_{ij} is iid random noise. For the purposes of this power calculation we consider a simple setup where each node belongs to one of two row groups and one of two column groups with equal probability. We let $W = \left(\begin{smallmatrix} 0 & -\mu \\ \mu & 0 \end{smallmatrix}\right)$ depend on a single parameter $\mu > 0$. Under this choice of W, we have $E[Y] = 0$ and $E[\mathbf{1}^t Y \mathbf{1}] = 0$. Since there are only two groups, the latent group membership vectors can be written as $u_i = (u_{i1}, 1 - u_{i1})$ and $v_j = (v_{j1}, 1 - v_{j1})$ where u_{i1} and v_{j1} are independent Bernoulli(1/2) random variables.

The bottom right panel of Figure 2.1 presents the power calculations for dimensions $m \in \{5, 10, 25, 50, 100\}$ and $|\mu| \in [0, 4]$. The power of the level $\alpha = 0.05$ test is increasing in $|\mu|$ which is a desirable property for this blockmodel since as $|\mu|$ grows the difference in the means for the groups becomes greater. The power of the test also increases with the dimension m.

2.4 Extensions and Applications

In this section we develop several extensions of the proposed test, and illustrate their use in the context of two data analysis examples. In the first example, we show how the test can be extended to accommodate a missing diagonal, an unknown non-zero mean, and heteroscedastic replications. The second example illustrates the use of the test for binary network data, a common type of relational data.

2.4.1 Extensions and continuous data example

International trade data on the value of exports from country to country is collected by the UN on a yearly basis and disseminated through the UN Comtrade website: `http://comtrade.un.org`. In this section we consider measures of total exports between twenty-six mostly large, well developed countries with high gross domestic product collected from 1996 to 2009 (measured in 2009 dollars). Specifically, we are interested in evaluating evidence for correlations among exporters and among importers. As trade between countries is relatively stable across years, we analyze the yearly change in log trade values, resulting in thirteen measurements (for the fourteen years of data) for every country-country pair. The data takes the form of a three way array $Y = \{Y_{ijk} : i, j \in \{1, \ldots, 26\}, k \in \{1, \ldots, 13\}\}$ where i and j index the exporting and importing countries respectively and k indexes the year. Exports from a country to itself are not defined, and so entries Y_{iik} are "missing".

We consider a model for trade of the form,

$$Y_{ijk} = \beta_1 x_{ik} + \beta_2 x_{jk} + \epsilon_{ijk}, \tag{2.10}$$

where x_{ik} is the difference in log gross domestic product of country i between years k and $k-1$ (theoretical development of this model is available in the economics literature, see Tinbergen et al. [1962], and Bergstrand [1985, 1989]). We use GDP data collected by the World Bank through `http://data.worldbank.org/` to obtain OLS estimates of β_1 and β_2. To investigate the correlations among importers and among exporters we collect the residuals $e_{ijk} = Y_{ijk} - \hat{Y}_{ijk}$ into thirteen matrices, $E_{\cdot\cdot k}$ for $k = 1, \ldots, 13$. Figure 2.2 plots the first two eigenvectors of moment estimates of pairwise row and column correlation matrices based on $E_{\cdot\cdot 1}, \ldots, E_{\cdot\cdot (13)}$. We observe systematic geographic patterns in both panels of the figure suggesting evidence that the ϵ_{ijk} are not independent. To evaluate this evidence formally by testing for dependence of the ϵ_{ijk} we extend the conditions under which the test developed in this chapter is applicable. Specifically, we must accommodate the following features of this data: a missing diagonal (Y_{iik} is not defined for all i and k), multiple observations (13 data points), and a nonzero mean structure (of the form (2.10)).

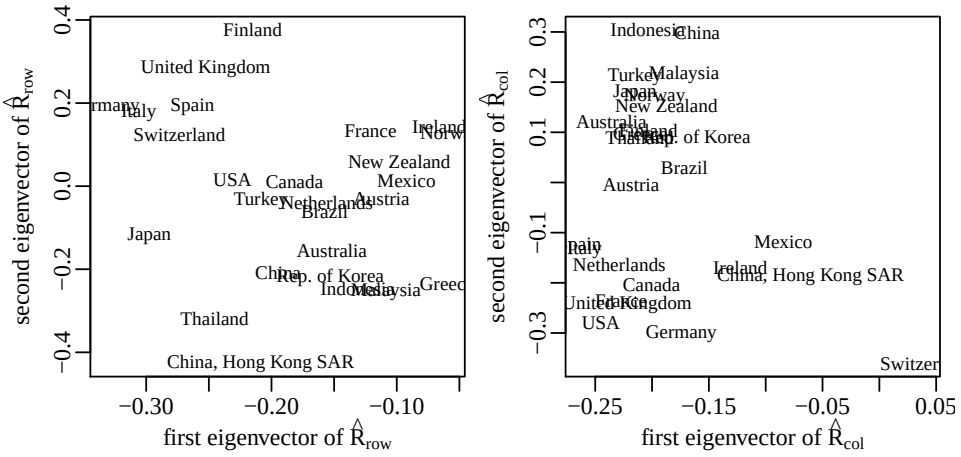

Figure 2.2: Plots of the first two eigenvectors of the estimates $\hat{R}_{\text{row}}$ and $\hat{R}_{\text{col}}$ of the row and column correlation matrices. Proximity of countries in the eigenspace indicates a positive correlation.

Missing diagonal: In relational datasets, the relationship of an actor to himself is typically undefined, meaning that the relational matrix Y has an undefined diagonal. It is common to treat the entries of an undefined diagonal as missing at random and to use a data augmentation procedure to recover a complete data matrix, applying the analysis to the complete data. In the context of this chapter, this approach would allow us to treat the whole data matrix as a draw from a matrix normal distribution and perform our test exactly as outlined in Section 2.2. In this section we describe an augmentation procedure that does not require distributional assumptions for the diagonal elements. The procedure produces a data matrix $\tilde{Y}$ that we use to calculate the test statistic $T(\tilde{Y})$. Specifically, $\tilde{Y}$ replaces the undefined diagonal of Y with zeros, keeping the rest of the data matrix the same. We show that $T(\tilde{Y})$ is invariant under diagonal transformations and thus we can approximate the null distribution of the test statistic based on for data drawn from a matrix normal distribution where the diagonal entries are replaced with zeros.

Consider square matrices Y and $\tilde{Y}$ where $Y \sim N_{m\times m}(0, D_{\text{r}}, D_{\text{c}})$ while $\tilde{Y}$ is distributed identically to Y except the diagonal entries are replaced with zeros. Similarly define square matrices $Y_0 \sim N_{m\times m}(0, I, I)$ and $\tilde{Y}_0$. We showed in Section 2.2 that the distribution of the

likelihood ratio test statistic is invariant under transformations by diagonal matrices on the left and right, that is $T(Y) \stackrel{\mathrm{d}}{=} T(Y_0)$. We now show that $T(\tilde{Y}) \stackrel{\mathrm{d}}{=} T(\tilde{Y}_0)$. It is immediate that since $Y \stackrel{\mathrm{d}}{=} D_{\mathrm{r}}^{1/2} Y_0 D_{\mathrm{c}}^{1/2}$ we have $\tilde{Y} \stackrel{\mathrm{d}}{=} D_{\mathrm{r}}^{1/2} \tilde{Y}_0 D_{\mathrm{c}}^{1/2}$, as zeros on the diagonal are preserved by left and right diagonal transformations and the off diagonal entries (i, j) are normally distributed with variance $D_{\mathrm{r,i}} D_{\mathrm{c,j}}$. As in Section 2.2.2, we appeal to the equivariance of a unique MLE to show that $T(\tilde{Y}) = T(\tilde{Y}_0)$. The argument is identical to the one appearing in the paragraph following Equation (2.7) on page 14 and so we do not reproduce it here. Since $T(\tilde{Y}) \stackrel{\mathrm{d}}{=} T(\tilde{Y}_0)$, we can approximate the null distribution and calculate the relevant quantiles for the test statistic with a simple update to the algorithm at the end of Section 2.2.2:

1. Simulate $\tilde{Y}_0^1, \ldots, \tilde{Y}_0^S \overset{\mathrm{iid}}{\sim} \mathcal{L}\left(\tilde{Y}_0\right)$, where $\mathcal{L}(\tilde{Y}_0)$ denotes the distribution of $\tilde{Y}_0$;

2. Let $\hat{T}_q = \min\{T(\tilde{Y}_0^q) : \sum_{s=1}^{S} 1[T(\tilde{Y}_0^q) \geq T(\tilde{Y}_0^s)]/S \geq q\}$.

Repeated observations: The test we discussed in this chapter is designed for a single observation. However, the test conveniently generalizes to the situation in which multiple observations are available. We will consider two types of additional observations: independent homoscedastic observations and independent heteroscedastic observations. First, if there are p independent identically distributed observations, we note that likelihood equations of Section 2 can be rewritten as

$$mp\hat{D}_{\mathrm{r}} = \sum Y_i \hat{D}_{\mathrm{c}}^{-1} Y_i^t \circ I \qquad mp\hat{D}_{\mathrm{c}} = \sum Y_i^t \hat{D}_{\mathrm{r}}^{-1} Y_i \circ I$$
$$mp\hat{\Sigma}_{\mathrm{r}} = \sum Y_i \hat{\Sigma}_{\mathrm{c}}^{-1} Y_i^t \qquad mp\hat{\Sigma}_{\mathrm{c}} = \sum Y_i^t \hat{\Sigma}_{\mathrm{c}}^{-1} Y_i.$$

The likelihood remains bounded and the form of the test statistic is identical to Equation (2.7). When the observations are heteroscedastic the likelihood equations (included in the proof of Theorem 3 in the Appendix) are more complicated because of the need to estimate the variability along the replications (we refer the reader to Hoff [2011] for an exposition on the general class of array normal distributions and estimation procedures).

Theorem 3. *Let $Y_1, \ldots, Y_p$ be independent random matrices distributed as $Y_i \sim N_{m\times m}(0, d_i\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}})$. Then $\forall p \geq 1$ the likelihood is bounded as a function of the covariance matrices Σ_{r} and Σ_{c} and the variance parameters $d_1, \ldots, d_p$.*

A proof is in the Appendix. Theorem 3 extends the literature on maximum likelihood estimation for proportional covariance models from natural exponential families to the matrix normal family which is a curved exponential family [Eriksen, 1987, Flury, 1986, Jensen and Johansen, 1987, Jensen and Madsen, 2004]. Due to Theorem 3, we can modify the test statistic to test $H_0' : D_{\mathrm{obs}} \in \mathcal{D}_+^p,\ D_{\mathrm{c}}, D_{\mathrm{r}} \in \mathcal{D}_+^m$ vs $H_1' : D_{\mathrm{obs}} \in \mathcal{D}_+^p,\ \Sigma_{\mathrm{c}}, \Sigma_{\mathrm{r}} \in \mathcal{S}_+^m$:

$$\begin{aligned} T(Y_1, \ldots, Y_p) &= l(\hat{D}_{\mathrm{obs}}^{\mathrm{null}}, \hat{D}_{\mathrm{r}}, \hat{D}_{\mathrm{c}}; Y) - l(\hat{D}_{\mathrm{obs}}^{\mathrm{alt}}, \hat{\Sigma}_{\mathrm{r}}, \hat{\Sigma}_{\mathrm{c}}; Y) \\ &= \log\left|\hat{D}_{\mathrm{obs}}^{\mathrm{null}} \otimes \hat{D}_{\mathrm{c}} \otimes \hat{D}_{\mathrm{r}}\right| - \log\left|\hat{D}_{\mathrm{obs}}^{\mathrm{alt}} \otimes \hat{\Sigma}_{\mathrm{c}} \otimes \hat{\Sigma}_{\mathrm{r}}\right|. \end{aligned}$$

We can again approximate the null distribution of the test statistic due to the invariance of the test statistic $T(Y_1, \ldots, Y_p)$ under diagonal transformations of the data along all three modes. The results on missing diagonal elements (above) and a non-zero mean (below) are also immediately applicable to $T(Y_1, \ldots, Y_p)$ above.

Relaxing the mean zero assumption: The reference distributions for the test statistics developed in Section 2.2 are based on the assumption that $\mathrm{E}[Y] = 0$, a strong assumption that is unlikely to be true for any observed dataset. While treating the mean as a nuisance parameter is tempting, the likelihood function under the alternative model is unbounded when estimating a mean matrix and two covariance matrices simultaneously. We propose to first fit a regression based mean to the data assuming the entries in the data matrix are independently distributed with the same variance parameter and to then perform the test based on the demeaned data. We consider the following regression framework for the mean: $y_{ij} = \beta^t x_{ij} + \epsilon_{ij}$. The regressor x_{ij} is a p-dimensional vector that can include features of node i, features of node j and dyadic features for nodes i and j. The ϵ_{ij} are assumed to be independent and identically distributed errors. Writing this in vector notation as $\mathrm{vec}(Y) = X\beta + \mathrm{vec}(\epsilon)$ where $X = \left(x_{12}^t \cdots x_{(m-1)m}^t\right)^t$ and ϵ is an $m \times m$ matrix, the OLS estimate of β is $\hat{\beta} = \left(X^t X\right)^{-1} X^t \mathrm{vec}(Y)$. Under mild regularity conditions on the distribu-

tion of the explanatory variables X and the row and column variances for new nodes, the OLS estimate $\hat{\beta}$ is a consistent estimate of β. This motivates us to base the test statistic on $\text{vec}\,(\hat{\epsilon}) = \text{vec}\,(Y) - X\hat{\beta}$, the residuals of the regression, as we expect the distribution of $\text{vec}\,(\hat{\epsilon})$ to be close to the distribution of $\text{vec}\,(\epsilon)$ for large m. The null distribution of the test statistic T based on the residuals from the regression is not identical to the one derived in Section 2.2 and no explicit computation of the new null distribution is readily available. However, we have observed via simulation that the level of the test based on the estimated residuals $\hat{\epsilon}$ appears to be asymptotically correct and that the test statistic based on ϵ and $\hat{\epsilon}$ appear to have the same limiting distributions.

Application to international trade data: Figure 2.2 suggested that there is evidence of residual dependence among exporters and among importers based on additional information about the data (the relative geographic positions of the countries). Above we developed the tools to test for independence of ϵ_{ijk} in (2.10) using the likelihood ratio test proposed in Section 2.2. Formally, we are testing the null hypothesis $H_0' : D_{\text{time}} \in \mathcal{D}_+^{13},\ D_{\text{imp}}, D_{\text{exp}} \in \mathcal{D}_+^{26}$ versus the alternative hypothesis $H_1' : D_{\text{time}} \in \mathcal{D}_+^{13},\ \Sigma_{\text{imp}}, \Sigma_{\text{exp}} \in \mathcal{S}_+^{26}$. The approximate 95% quantile of the distribution of the test statistic when the data are missing diagonal entries under the null is 729.8. Setting $E_{iik} = 0$ for all i and k, the test statistic for the data is $T(E_{\cdot\cdot 1}, \ldots, E_{\cdot\cdot 13}) = 3354$. This value is much greater than the 95% quantile confirming that we should reject the independence of the ϵ_{ijk}. It is thus inappropriate to assume that the exporters and importers are independent.

2.4.2 Application to binary protein-protein interaction network

So far we have developed a testing procedure for the presence of row and column correlations in relational matrices within the framework of matrix normal and general matrix variate elliptically contoured distributions. In this section we propose a methodology that allows us to evaluate the presence of row and column correlations for binary relational data, where the observed network is represented by a sociomatrix matrix A where a_{ij} describes the relationship from node i to node j. When the entries of a_{ij} are binary indicators of a relationship from i to j, the matrix A can be viewed as the adjacency matrix of a directed

graph. In this example we use the protein-protein interaction data of Butland et al. [2005], which consists of a record of interactions between $m = 270$ essential proteins of *E. coli*. The data are organized into a 270×270 binary matrix A, where $a_{ij} = 1$ if protein j binds to protein i and $a_{ij} = 0$ otherwise. The network has one large connected component with 234 nodes as shown in the left hand side of Figure 2.3. In this case diagonal elements of the matrix A are meaningful since proteins may bind to themselves.

A popular class of models for the analysis of such data is based on representing the relations a_{ij} as functions of latent normal random variables [Hoff, 2005b, 2008]. For the protein-protein interaction data we propose to use an asymmetric version of the eigenmodel of Hoff [2008], a type of reduced rank latent variable model where the relationship between nodes i and j is characterized by multiplicative latent sender and receiver effects. The model can be written as:

$$a_{ij} = 1[y_{ij} > \gamma], \qquad y_{ij} = u_i^t v_j + \epsilon_{ij}, \qquad Y = UV^t + E,$$

where $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0,1)$ and $u_i, v_j \in \mathbb{R}^R$ for $R < 270$. Considering Y as a matrix variate normal variable it is immediate that $E[YY^t] = U(V^tV)U^t + I$ and $E[Y^tY] = V(U^tU)V^t + I$ and so the heterogeneity in U describes the row covariance Σ_{r} while the heterogeneity in V describes the column covariance Σ_{c}.

We propose using the test developed in this chapter to evaluate how well models of rank R capture the dependence in the data. Specifically, we fit the above model for multiple values of R, and for each value we approximate the posterior distribution of the test statistic. If the rank-R model is sufficient for capturing the row and column correlations found in the data, we do not expect to have evidence to reject the null of independence. Hoff [2008] outlines a Markov chain Monte Carlo algorithm for fitting the above model. Following the procedure described in Thompson and Geyer [2007] we apply the testing procedure to draws from the posterior distribution of $Y - UV^t$ constructed via MCMC and for each test statistic we calculate a p-value. These p-values are termed "fuzzy p-values" as their distribution provides a description of the uncertainty about the p-value that results from not observing Y, U and V.

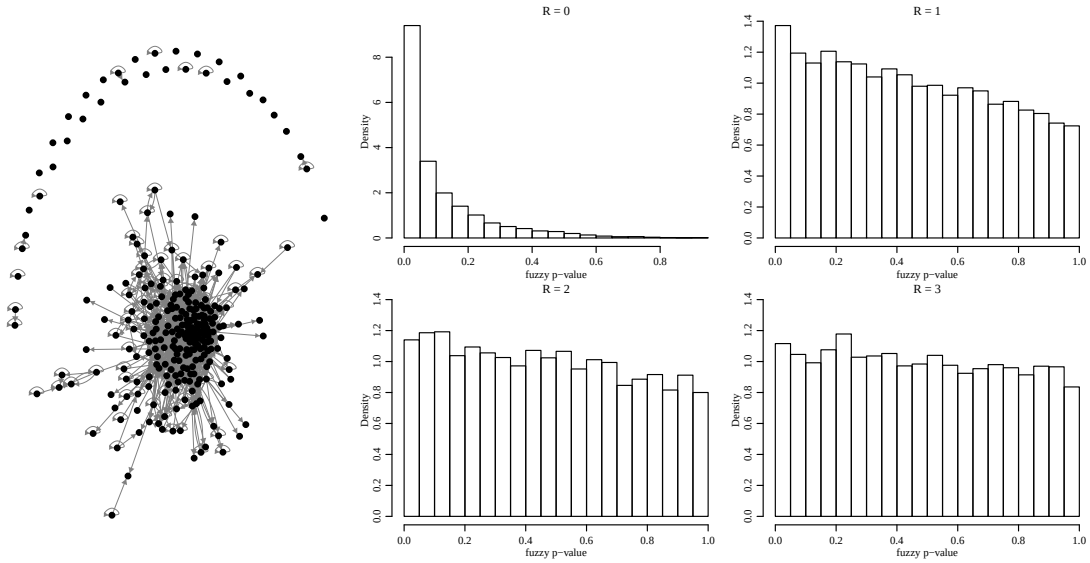

Figure 2.3: Protein-protein interaction data and histograms of fuzzy p-values for models of ranks $R \in \{0, 1, 2, 3\}$. The bins have a width of 0.05.

As this is a very sparse network (the interaction rate is $\bar{A} = 0.03$), we expect a low rank approximation to be appropriate. In fact, analysis using cross validation of a symmetrized version of this data identified $R = 3$ to be an appropriate rank in Hoff [2008]. In the right hand side of Figure 2.3 we present the distributions of the fuzzy p-values for $R \in \{0, 1, 2, 3\}$. A visual inspection of the the fuzzy p-values in the four panels of the figure provides evidence about the rank of the latent factors. For example, under the $R = 0$ model, the y_{ij}s are independent and identically distributed, and so the graph represented by the adjacency matrix A is a simple random graph. The fuzzy p-values are concentrated at a value lower than 0.05 suggesting a high probability of rejecting the null if Y were observed. For $R \in \{1, 2\}$ the fuzzy p-values are no longer concentrated lower than 0.05, but the distribution is skewed to the right, which we take as evidence that there is correlation in Y that is not captured by the rank 1 and rank 2 models. The fuzzy p-values provide little evidence of residual dependence in Y for models of rank $R \geq 3$.

2.5 Discussion

In this chapter we presented a likelihood ratio test for relational datasets. Unlike the previous testing literature for matrix normal models that required multiple observations, and concentrated on testing a null of separable covariances versus an unstructured alternative, we proposed testing a null of no row or column correlations versus an alternative of full row and column correlations using a single observation of a network. While the form of the null distribution of the test statistic is intractable, we are able to simulate a reference distribution for the test statistic under the null due to its invariance to left and right diagonal transformations of the data. In the power simulations of Section 2.3 we demonstrated the power of the test against maximally sparse and nonseparable alternatives.

This test can be applied to a wide variety of relational data. While the test was developed using the matrix-normal model, we have shown that this distributional assumptions can be greatly relaxed. Specifically, if we consider a data matrix Y with an arbitrary matrix variate elliptically contoured distribution that is centered at the zero matrix, the test statistic for testing for correlation among the rows and among the columns of Y is identical to that of the matrix normal case. We have also demonstrated that the test can accommodate frequently observed features of relational data such as non-zero mean, missing diagonal and multiple observations. In Section 2.4.2 we demonstrated an application of the theory developed in this paper to binary network data where the matrix Y is an adjacency matrix. The method we describe for binary data can be extended to ordinal and discrete data that can be modeled via a latent matrix variate elliptically contoured distribution.

Once we reject the null hypothesis of independence among the rows and among the columns of a relational matrix, we are faced with the challenge of modeling the dependence in the data. As shown in Section 2.2.1, for a mean zero matrix normal distribution, the MLE is not unique. Specifically, there is an MLE for each $\Sigma_{\mathrm{r}} \in \mathcal{S}_{+}^{m}$ given by $(\Sigma_{\mathrm{r}}, Y^{t}\Sigma_{\mathrm{r}}^{-1}Y/m)$. We have observed in separate work that it is possible to distinguish between MLEs by considering their risk. However, other than in very specialized cases (such as equal eigenvalues of Σ_{r} and Σ_{c}), obtaining analytic results for identifying risk optimal MLEs is difficult. The presence of a non-zero mean leads to an unbounded likelihood and further complicates the

problem of estimation. Several authors have recently considered Bayesian and penalized likelihood approaches this estimation problem. Bonilla et al. [2008] and Yu et al. [2007] studied hierarchical Gaussian Process priors in the context of a classification problem. In our context, this approach results in a matrix normal prior for the mean parameters and inverse Wishart priors for the row and column covariance matrices. A second approach based on a mixture of independent L_1 and L_2 penalties on the row and column precision matrices was proposed by Allen and Tibshirani [2010].

Computer code and data for the results in Sections 3 and 4 are available at the authors' websites.

Chapter 3

COVARIANCE AND MEAN PARAMETER ESTIMATION FOR THE SQUARE MATRIX NORMAL

3.1 Introduction

In Chapter 2 we developed a test for dependence among the rows and among the columns of a relational dataset within the framework of matrix-variate elliptically contoured distributions. The rejection of the null hypothesis by this test leads to an inference problem: how does one account for the row and column correlation that is evident in the data? This problem can be further broken down into cases: the simple case where the mean is known (and without loss of generality will be assumed to be zero) and the more difficult one, where the mean is unknown and needs to be estimated.

This chapter discusses several methods for estimating covariances (and means and covariances) in the context of a square matrix-variate normal distribution. Section 2 describes covariance estimators in the known mean case. We concentrate on the classes of maximum likelihood estimators and maximum penalized likelihood estimators. We present theoretical results for several subproblems and develop a novel penalty for the similarity between the covariance matrices for the rows and columns of the relational data matrix. In Section 3 we extend these results to the case of an unknown mean. In the case of the unpenalized estimators of Section 2, a one-step feasible GLS approach is presented, as the likelihood is unbounded when estimating mean parameters and full row and column covariance matrices. On the other hand, for the penalized methods an iterative estimation procedure is proposed. Theoretical guarantees for the convergence of the optimization procedure and for the unbiasedness of the estimates of the mean parameters are conjectured. We conclude with a simulation that studies the risks of three estimators: OLS, feasible GLS based on the risk optimal MLE of Section 2, and our penalized estimator. A discussion follows in Section 4.

3.2 Covariance estimation when the mean is known

In this section we discuss estimating covariances for the known mean case for a single obesrvation of a square matrix normal variable. We provide the form of the maximum likelihood estimators and describe a subclass of MLEs that is equivariant. The remainder of the section is dedicated to penalized likelihood approaches. We propose two types of penalties: one that penalizes differences in the eigenvalues of the covariance matrices and one that penalizes the differences between the full covariance matrices. In the former case we demonstrate that there exists an equivariant MLE for which any symmetric penalty on the eigenvalues is zero. This estimator is then shown to be Bayes risk optimal in a special class of equivariant MLEs as well as being the one to minimize the Kullback Leibler divergence from the moment estimators of the row and column covariances. For the penalty on the similarity between the full covariance matrices we describe a reparametrization of the likelihood in terms of fixed-trace row and column covariance matrices and an overall scale term. We propose the use of a penalty function based on the symmetrized Kullback Leibler divergence between the two covariance matrices and develop an optimization procedure for finding the optimizer of the penalized likelihood function. We conclude by describing a procedure for selecting the penalty term.

3.2.1 Maximum likelihood estimators

In this subsection we reproduce the results on the maximum likelihood estimator (MLE) from Chapter 2. Recall, that the density of a mean zero matrix normal distribution $N_{m\times m}(0,\Sigma_{\mathrm{r}},\Sigma_{\mathrm{c}})$ is given by

$$p\left(Y|\Sigma_{\mathrm{r}},\Sigma_{\mathrm{c}}\right) \;=\; (2\pi)^{-m^2/2}\left|\Sigma_{\mathrm{c}}\otimes\Sigma_{\mathrm{r}}\right|^{-1/2}\exp(-\tfrac{1}{2}\mathrm{tr}\left(\Sigma_{\mathrm{r}}^{-1}Y\Sigma_{\mathrm{c}}^{-1}Y^{t}\right)),$$

where "tr" is the matrix trace and "$\otimes$" is the Kronecker product. Throughout this chapter we will write $l\left(\Sigma_{\mathrm{r}},\Sigma_{\mathrm{c}};Y\right)$ as minus two times the log likelihood minus $m^2\log 2\pi$, hereafter referred to as the scaled log likelihood:

$$l\left(\Sigma_{\mathrm{r}},\Sigma_{\mathrm{c}};Y\right) \;=\; -2\log p\left(Y|\Sigma_{\mathrm{r}},\Sigma_{\mathrm{c}}\right)-m^2\log 2\pi=\mathrm{tr}\left[\Sigma_{\mathrm{r}}^{-1}Y\Sigma_{\mathrm{c}}^{-1}Y^{t}\right]-\log\left|\Sigma_{\mathrm{c}}^{-1}\otimes\Sigma_{\mathrm{r}}^{-1}\right|.$$

We will state all the results in this chapter in terms of $l(\Sigma_r, \Sigma_c; Y)$. For example, an MLE will be a minimizer of $l(\Sigma_r, \Sigma_c; Y)$ in (Σ_r, Σ_c). The following result implies that if Y is a draw from an absolutely continuous distribution on $\mathbb{R}^{m\times m}$, then the scaled likelihood is bounded from below, and achieves this bound on a set of nonunique MLEs:

Theorem 4. *If Y is full rank then $l(\Sigma_r, \Sigma_c; Y) \geq m^2 + m \log |YY^t/m|$ for all $(\Sigma_r, \Sigma_c) \in \Theta$, with equality if $\Sigma_r = Y\Sigma_c^{-1}Y^t/m$, or equivalently, $\Sigma_c = Y^t\Sigma_r^{-1}Y/m$.*

Proof. We first look for MLEs at the critical points of $l(\Sigma_r, \Sigma_c; Y)$. Setting derivatives of l to zero indicates that critical points satisfy

$$\hat{\Sigma}_r = Y\hat{\Sigma}_c^{-1}Y^t/m \tag{3.1}$$

$$\hat{\Sigma}_c = Y^t\hat{\Sigma}_r^{-1}Y/m. \tag{3.2}$$

Note that these equations are redundant: If $(\hat{\Sigma}_r, \hat{\Sigma}_c)$ satisfy Equation 3.1, then these values satisfy Equation 3.2 as well. The value of the scaled log likelihood at such a critical point is

$$\begin{aligned} l\left(Y\hat{\Sigma}_c^{-1}Y^t/m, \hat{\Sigma}_c; Y\right) &= \operatorname{tr}\left[\left(Y\hat{\Sigma}_c^{-1}Y^t/m\right)^{-1} Y\hat{\Sigma}_c^{-1}Y^t\right] + \log\left|\hat{\Sigma}_c \otimes Y\hat{\Sigma}_c^{-1}Y^t/m\right| \\ &= m\operatorname{tr}\left[Y^{-t}\hat{\Sigma}_c Y^{-1}Y\hat{\Sigma}_c^{-1}Y^t\right] + m\log\left|\hat{\Sigma}_c\right| - m\log\left|\hat{\Sigma}_c\right| \\ &\quad + m\log\left|YY^t/m\right| \\ &= m\operatorname{tr}[I] + m\log\left|YY^t/m\right|. \end{aligned} \tag{3.3}$$

In the second and third lines, Y^{-1} exists and $|YY^t/m| > 0$ since Y is square and full rank. Now we compare the scaled log likelihood at a critical point to its value at any other point.

$$\begin{aligned} l(\Sigma_r, \Sigma_c; Y) - l(Y\hat{\Sigma}_c^{-1}Y^t/m, \hat{\Sigma}_c; Y) &= \operatorname{tr}\left[\Sigma_r^{-1}Y\Sigma_c^{-1}Y^t\right] + m\log\left(|\Sigma_r|\,|\Sigma_c|\right) \\ &\quad - m^2 - m\log\left|YY^t/m\right| \\ &= m^2(\operatorname{tr}\left[\Sigma_r^{-1}Y\Sigma_c^{-1}Y^t/m\right]/m) \\ &\quad - m^2(\log\left|\Sigma_r^{-1}Y\Sigma_c^{-1}Y^t/m\right|/m) - m^2 \end{aligned}$$

The first equality is a simple combination of equations 3.1 and 3.3. The second equality is a

rearrangement of terms that combines all the determinant in the log terms. This difference can be written as $m^2(a - \log g - 1)$, where a is the arithmetic mean and g is the geometric mean of the eigenvalues of $(\Sigma_{\mathrm{r}}^{-1/2} Y \Sigma_{\mathrm{c}}^{-1} Y^t \Sigma_{\mathrm{r}}^{-1/2})/m$. To complete the proof we show that $a - \log g - 1 \geq 0$. Consider $f(x) = x - 1 - \log x$ and its first and second derivatives with respect to x: $f'(x) = 1 - \frac{1}{x}$, and $f''(x) = \frac{1}{x^2}$. The second derivative is positive at the critical point $x = 1$, so $f(1) = 0$ is a global minimum of the function. Thus $x - \log x - 1 \geq 0$. Now let $\lambda_1, \ldots, \lambda_m$ be the eigenvalues of $(\Sigma_{\mathrm{r}}^{-1/2} Y \Sigma_{\mathrm{c}}^{-1} Y^t \Sigma_{\mathrm{r}}^{-1/2})/m$ and so $a = \frac{1}{m} \sum \lambda_i$ and $g = (\prod \lambda_i)^{1/m}$. We then have

$$a \geq \log a + 1 = \log \left(\frac{1}{m} \sum x_i \right) + 1 \geq \log \left(\left(\prod x_i \right)^{1/m} \right) + 1 = \log g + 1$$

as $a \geq g$ since $\lambda_i \geq 0 \; \forall i$. Since $a - 1 - \log g \geq 0$ we have the desired result. □

Note that the MLE is not unique, nor is the MLE of $\Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}}$. For example. $I \otimes YY^t/m$ is an MLE of $\Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}}$, as is $Y^tY \otimes I/m$. Moreover, there is an MLE for each $\Sigma_{\mathrm{r}} \in \mathcal{S}_+^m$ given by $(\Sigma_{\mathrm{r}}, Y^t \Sigma_{\mathrm{r}}^{-1} Y/m)$, and similarly there is an MLE for each $\Sigma_{\mathrm{c}} \in \mathcal{S}_+^m$ given by $(Y \Sigma_{\mathrm{c}}^{-1} Y^t/m, \Sigma_{\mathrm{c}})$.

Maximum likelihood estimators as equivariant estimators

In full exponential families maximum likelihood estimators are equivariant [Eaton, 1983]. When dealing with curved exponential families where MLEs might not be unique, this property is not guaranteed. For the matrix normal distribution, orthogonal equivariance of the covariance estimators implies that

$$\hat{\Sigma}_{\mathrm{r}}(UYV^t) = U \hat{\Sigma}_{\mathrm{r}}(Y) U^t \tag{3.4}$$

$$\hat{\Sigma}_{\mathrm{c}}(UYV^t) = V \hat{\Sigma}_{\mathrm{c}}(Y) V^t. \tag{3.5}$$

It is easy to see that the MLE formed from $\hat{\Sigma}_{\mathrm{r}}(Y) = Y^tY$ does not satisfy equation (3.4): $\hat{\Sigma}_{\mathrm{r}}(UYV^t) = (UYV^t)^t(UYV^t) = VY^tYV^t$. As such not all MLEs derived above are equivariant. However equivariance is a desirable property that we might wish to impose on an estimator.

Theorem 5. *If Y is full rank with a singular value decomposition given by $Y = ADB^t$, then the set of orthogonally equivariant estimators that minimize the scaled log likelihood is*

$$\{(AL_{\mathrm{r}}A^t, BL_{\mathrm{c}}B^t) : L_{\mathrm{r}}L_{\mathrm{c}} = D^2/m\}$$

Proof. It is easy to verify that any element of this set is an MLE that satisfies the likelihood equations (3.1) and (3.2). We now show that all orthogonally equivariant MLEs are contained in this set. Recall the definition of equivariance from Equations (3.4) and (3.5). Since the estimators must be symmetric positive definite matrices, there exist $\tilde{U}$ and $\tilde{V}$ orthogonal such that $\tilde{U}\hat{\Sigma}_{\mathrm{r}}(Y)\tilde{U}^t = \Delta_{\mathrm{r}}$ and $\tilde{V}\hat{\Sigma}_{\mathrm{c}}(Y)\tilde{V}^t = \Delta_{\mathrm{c}}$. We now substitute this identity into the likelihood equation (3.1)

$$\begin{aligned}\Delta_{\mathrm{r}} = \hat{\Sigma}_{\mathrm{r}}(\tilde{U}Y\tilde{V}^t) =& (\tilde{U}Y\tilde{V}^t)\hat{\Sigma}_{\mathrm{c}}(\tilde{U}Y\tilde{V}^t)^{-1}(\tilde{U}Y\tilde{V}^t)^t/m \\ =& (\tilde{U}Y\tilde{V}^t)\Delta_{\mathrm{c}}^{-1}(\tilde{U}Y\tilde{V}^t)^t/m.\end{aligned}$$

Since the above must hold for any data matrix Y, let Y be a diagonal matrix D then we must have $\Delta_{\mathrm{r}} = (\tilde{U}D\tilde{V}^t)\Delta_{\mathrm{c}}^{-1}(\tilde{U}D\tilde{V}^t)^t$. Since $\Delta_{\mathrm{r}}, \Delta_{\mathrm{c}}$ and D are diagonal matrices by definition, $\tilde{U}D\tilde{V}^t$ must be a weighted permutation matrix and in turn $\tilde{U}$ and $\tilde{V}$ are permutation matrices. This means that $\hat{\Sigma}_{\mathrm{r}}(D)$ and $\hat{\Sigma}_{\mathrm{c}}(D)$ are diagonal. Now, for a (non-diagonal) matrix $Y = ADB^t$ we have that $\hat{\Sigma}_{\mathrm{r}}(ADB^t) = A\hat{\Sigma}_{\mathrm{r}}(D)A^t$ and $\hat{\Sigma}_{\mathrm{c}}(ADB^t) = B\hat{\Sigma}_{\mathrm{c}}(D)B^t$. By using the orthogonal equivariance of the estimator and the fact that $\hat{\Sigma}_{\mathrm{r}}(D)$ is diagonal we have (by plugging into Equation (3.1)) that $\hat{\Sigma}_{\mathrm{r}}(D) = D\hat{\Sigma}_{\mathrm{c}}(D)^{-1}D/m$. This shows that all orthogonally equivariant MLEs must belong to the set $\{(A\hat{\Sigma}_{\mathrm{r}}(D)A^t, B\hat{\Sigma}_{\mathrm{c}}(D)B^t : \hat{\Sigma}_{\mathrm{r}}(D)\hat{\Sigma}_{\mathrm{c}}(D) = D^2/m\}$. □

3.2.2 Penalized methods

Penalized methods are prevalent in statistics. They have been developed to assist in estimating probability density functions [Silverman, 1982], regression coefficients [Green, 1987, 1990, Tibshirani, 1996] and for inducing sparsity in covariance and precision matrices [Friedman et al., 2008, Allen and Tibshirani, 2010]. Furthermore, many Bayesian procedure can

be viewed as a penalty approach to estimation. The benefit of penalized methods is the regularization that they induce in the estimator via the penalty.

We propose two novel penalties that take take into account the particular structure of relational data (that of square matrices) and the special relationship between the index sets of the rows and the columns. Specifically, we note that in a relational dataset the rows and columns have the same index sets. Recall that the row covariance matrix describes the relationships among the rows of the relational data matrix, that is the similarity in the distribution of the relations for different senders. Similarly, the column covariance matrix describes the similarity in distribution of the relations for different receivers. Since the senders and the receivers are the same, it is plausible that if two individuals are similar as senders they will also be similar as receivers. This would imply a similarity in the row and column covariance matrices. We propose two types of penalties: the first penalty states that the similarity between the row and column covariance matrices is only in their scale - that is they have similar eigenvalues. The second type of penalty is much more flexible and penalizes dissimilarities between the full row and column covariance matrices. We discuss the exact penalty functions in detail below.

Whenever discussing penalized methods we will consider minimizing the augmented objective function

$$f(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}; Y) = l(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}; Y) + \theta d(g(\Sigma_{\mathrm{r}}), g(\Sigma_{\mathrm{c}})), \tag{3.6}$$

where $l(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}; Y)$ is the scaled log likelihood, $d(g(\Sigma_{\mathrm{r}}), g(\Sigma_{\mathrm{c}}))$ is a (symmetric) penalty function that is zero only when $g(\Sigma_{\mathrm{r}}) = g(\Sigma_{\mathrm{c}})$ and θ is a positive penalty parameter. We choose to only consider penalty functions that are invariant to inversion, that is $d(A, B) = d(A^{-1}, B^{-1})$ as we have no apriori preference for penalizing similarity among the precision or among the covariance matrices. Under mild conditions, the penalty function $\theta d(g(\Sigma_{\mathrm{r}}), g(\Sigma_{\mathrm{c}}))$ can be interpreted as a proper prior on the covariances. While the penalty functions that we consider define proper priors, we cannot easily simulate from them as they do not correspond to standard distributions. If the choice of θ can be made from the data then the penalty is called adaptive, as is the case in our development.

Similar eigenvalues

Consider the augmented objective function (3.6) and let $g(\Sigma)$ be the function that extracts the eigenvalues of the symmetric positive definite matrix Σ. We note that the estimator is identical for all symmetric penalties on the eigenvalues $d(g(\Sigma_{\mathrm{r}}), g(\Sigma_{\mathrm{c}}))$. To see this write $\Sigma_{\mathrm{r}} = U\Lambda U^t$ for the eigenvalue decomposition of Σ_{r} and $Y = ADB^t$ for the singular value decomposition of the data matrix Y and consider the first derivative of the augmented objective function with respect to Σ_{r}^{-1}:

$$\begin{aligned}
\frac{\partial f(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}; Y)}{\partial \Sigma_{\mathrm{r}}^{-1}} =& \frac{\partial l(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}; Y)}{\partial \Sigma_{\mathrm{r}}^{-1}} + \theta \frac{\partial d(g(\Sigma_{\mathrm{r}}), g(\Sigma_{\mathrm{c}}))}{\partial \Sigma_{\mathrm{r}}^{-1}} \\
=& Y\Sigma_{\mathrm{c}}^{-1}Y^t - m\Sigma_{\mathrm{r}} + \theta \frac{\partial d(g(\Sigma_{\mathrm{r}}^{-1}), g(\Sigma_{\mathrm{c}}^{-1}))}{\partial \Sigma_{\mathrm{r}}^{-1}} \\
=& Y\Sigma_{\mathrm{c}}^{-1}Y^t - m\Sigma_{\mathrm{r}} + \theta \sum \frac{\partial d(g(\Sigma_{\mathrm{r}}^{-1}), g(\Sigma_{\mathrm{c}}^{-1}))}{\partial g(\Sigma_{\mathrm{r}}^{-1})_i} \frac{\partial g(\Sigma_{\mathrm{r}}^{-1})_i}{\partial \Sigma_{\mathrm{r}}^{-1}} \\
=& Y\Sigma_{\mathrm{c}}^{-1}Y^t - m\Sigma_{\mathrm{r}} + \theta \sum \frac{\partial d(g(\Sigma_{\mathrm{r}}^{-1}), g(\Sigma_{\mathrm{c}}^{-1}))}{\partial g(\Sigma_{\mathrm{r}}^{-1})_i} u_i u_i^t \\
=& Y\Sigma_{\mathrm{c}}^{-1}Y^t - m\Sigma_{\mathrm{r}} + \theta U(\mathrm{diag}(\frac{\partial d(g(\Sigma_{\mathrm{r}}^{-1}), g(\Sigma_{\mathrm{c}}^{-1}))}{\partial g(\Sigma_{\mathrm{r}}^{-1})_i}))U^t.
\end{aligned}$$

The second equality follows from the invariance of the penalty function to inversions. The third equality is an application of the chain rule, while the fourth equality is due to Magnus [1985]. The final equality aggregates the derivative into matrix form. Setting this equal to 0, it is immediate that letting $U = A$ satisfies this equation (similarly, letting the eigenvectors of Σ_{c} be equal to B in the partial derivative with respect to Σ_{c}^{-1} yields the same result). Thus we have determined the eigenvectors of the penalized estimator of the row and column covariance matrices. We also note that the equivariant maximum likelihood estimators discussed in Section 2.1 have this eigenstructure. Among these MLEs, there exists a unique MLE for which the eigenvalues of the row and column covariances are equal which leads to the penalty term $d(g(\Sigma_{\mathrm{r}}), g(\Sigma_{\mathrm{c}}))$ vanishing completely. This MLE is of the form $(A(D/\sqrt{m})A^t, B(D/\sqrt{m})B^t)$ and is both Bayes risk optimal among a subclass of equivariant MLEs and it is the MLE that minimizes the KL divergence from the moment estimators of the row and column covariances. We show these properties below.

Bayes risk optimality. We are able to demonstrate that the estimator derived above minimizes the Bayes risk among a class of equivariant MLEs. We show this for a general class of loss functions and for exchangeable priors on the covariance matrices. We first define a useful operator:

Definition 3. The switch operator on Kronecker products acts as $r(A \otimes B) = B \otimes A$.

Theorem 6. *Let* $Y \sim N_{m\times m}(0, \Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}})$ *and define a class of estimators* $\Delta(Y) = \{\delta_s(Y) = (Y^tY/m)^s \otimes (YY^t/m)^{1-s} : \ s \in [0,1]\}$. *For the general class of switch invariant losses that are convex in* s *and exchangeable priors on the covariance matrices* Σ_{r} *and* Σ_{c}*, the estimator* $\delta_{1/2}(Y)$ *minimizes the Bayes risk among the class* $\Delta(Y)$.

Proof. We first note that $\delta_{1-s}(Y) = r(\delta_s(Y^t))$. This allows us to write the loss functions as follows:

$$\begin{aligned} L(\delta_{1-s}(Y), \Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}}) &= L(r(\delta_s(Y^t)), \Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}}) \\ &= L(\delta_s(Y^t), \Sigma_{\mathrm{r}} \otimes \Sigma_{\mathrm{c}}), \end{aligned} \tag{3.7}$$

where the first equality is due to the definition of the estimator and the switch operator (that leads to $\delta_{1-s}(Y) = r(\delta_s(Y^t))$). The second equality is due to the invariance of the loss function under the switch operator (as the switch operator is applied to both arguments). To prove the theorem we need to show that

$$\mathrm{E}_{Y,\Sigma_{\mathrm{c}}\otimes\Sigma_{\mathrm{r}}}[L(\delta_{1-s}(Y), \Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}})] = \mathrm{E}_{Y,\Sigma_{\mathrm{c}}\otimes\Sigma_{\mathrm{r}}}[L(\delta_s(Y), \Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}})], \tag{3.8}$$

since if the loss function is convex in s, then so is the Bayes risk and so the above equality would imply a symmetry about $s = 1/2$, which would have to be the minimizer. The expectation on the left hand side of (3.8) is equal to the expectation of (3.7). This means that it is enough to show that

$$\mathrm{E}_{Y,\Sigma_{\mathrm{c}}\otimes\Sigma_{\mathrm{r}}}[L(\delta_s(Y^t), \Sigma_{\mathrm{r}} \otimes \Sigma_{\mathrm{c}})] = \mathrm{E}_{Y,\Sigma_{\mathrm{c}}\otimes\Sigma_{\mathrm{r}}}[L(\delta_s(Y), \Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}})].$$

This equality is guaranteed if the joint distribution of $\{Y, \Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}}\}$ is equal to the joint

distribution of $\{Y^t, \Sigma_{\mathrm{r}} \otimes \Sigma_{\mathrm{c}}\}$. To show this equality in distribution we note that for a matrix-variate normal distribution, the following equality in conditional distributions holds: $Y|\Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}} \stackrel{\mathrm{d}}{=} Y^t|\Sigma_{\mathrm{r}} \otimes \Sigma_{\mathrm{c}}$ [Gupta and Nagar, 1999]. Also, by the assumption of the exchangeability we have $\Sigma_{\mathrm{c}} \otimes \Sigma_{\mathrm{r}} \stackrel{\mathrm{d}}{=} \Sigma_{\mathrm{r}} \otimes \Sigma_{\mathrm{c}}$. By the definition of conditional distributions, the above two equalities in distribution are sufficient for having the joint equality in distribution that we need. We have shown that the Bayes risk is symmetric about $s = 1/2$ which means that $\delta_{1/2}(Y)$ minimizes the Bayes risk in the class of estimators $\Delta(Y)$ as desired. □

This result suggests that lacking any outside information about the eigenvalues of the row and column covariance matrices, one should choose the estimator $\delta_{1/2}(Y)$ to estimate the covariance structure for an observation from a matrix-variate normal distribution.

MLE minimizing the KL divergence to the moment estimators. We now show that the above estimator minimizes the Kullback Leibler (KL) divergence from the moment estimators of the row and column covariance (YY^t and Y^tY respectively) in the complete class of equivariant MLEs. The KL divergence between two normal distributions is given by

$$\mathrm{tr}\left(\hat{\Sigma}_{\mathrm{r,mom}}^{-1}\hat{\Sigma}_{\mathrm{r,mle}}\right)\mathrm{tr}\left(\hat{\Sigma}_{\mathrm{c,mom}}^{-1}\hat{\Sigma}_{\mathrm{c,mle}}\right) - m\log\left|\hat{\Sigma}_{\mathrm{c,mom}}^{-1}\hat{\Sigma}_{\mathrm{c,mle}}\right| - m\log\left|\hat{\Sigma}_{\mathrm{r,mom}}^{-1}\hat{\Sigma}_{\mathrm{r,mle}}\right|.$$

We can write the estimators in terms of the SVD of the the data matrix $\sqrt{m}Y = ADB^t$ as $\hat{\Sigma}_{\mathrm{r,mom}} = AD^2A^t$ and $\hat{\Sigma}_{\mathrm{r,mle}} = A\hat{\Sigma}_{\mathrm{r,mle}}(D)A^t$.

$$\begin{aligned}\mathrm{tr}\left(\hat{\Sigma}_{\mathrm{r,mom}}^{-1}\hat{\Sigma}_{\mathrm{r,mle}}\right) &= \mathrm{tr}\left(AD^{-2}A^tA\hat{\Sigma}_{\mathrm{r,mle}}(D)A^t\right)\\ &= \mathrm{tr}\left(D^{-2}\hat{\Sigma}_{\mathrm{r,mle}}(D)\right)\end{aligned}$$

and

$$\begin{aligned}\log\left|\hat{\Sigma}_{\mathrm{r,mom}}^{-1}\hat{\Sigma}_{\mathrm{r,mle}}\right| &= \log\left|AD^{-2}A^tA\hat{\Sigma}_{\mathrm{r,mle}}(D)A^t\right|\\ &= \log\left|D^{-2}\hat{\Sigma}_{\mathrm{r,mle}}(D)\right|\end{aligned}$$

and so

$$\begin{aligned} KL\left(\hat{\Sigma}_{\text{mom}}, \hat{\Sigma}_{\text{mle}}\right) &= \operatorname{tr}\left(D^{-2}\hat{\Sigma}_{\text{r,mle}}(D)\right)\operatorname{tr}\left(D^{-2}\hat{\Sigma}_{\text{c,mle}}(D)\right) && (3.9)\\ &\quad -m\log\left|D^{-2}\hat{\Sigma}_{\text{r,mle}}(D)\right| - m\log\left|D^{-2}\hat{\Sigma}_{\text{c,mle}}(D)\right| && (3.10)\\ &= \sum_{ij}\frac{\hat{d}_{\text{r,i}}\hat{d}_{\text{c,j}}}{(d_i d_j)^2} - m\sum\log\frac{\hat{d}_{\text{r,i}}}{d_i^2} - m\sum\log\frac{\hat{d}_{\text{c,j}}}{d_j^2} \end{aligned}$$

where $\hat{\Sigma}_{\text{r,mle}}(D)\,\hat{\Sigma}_{\text{c,mle}}(D) = D^2/m$. If $\hat{\Sigma}$ is a power transformation, that is, each element of D^2/m is raised to the power p_i for $\hat{\Sigma}_{\text{r,mle}}$ and $q_i = 1-p_i$ for $\hat{\Sigma}_{\text{c,mle}}$ it is clear that the choice of $p_i = \frac{1}{2}$ yields the minimal KL distance since Eq 3.9 simplifies to $\frac{1}{m}\operatorname{tr}\left(D^{2(\text{p}-1)}\right)\operatorname{tr}\left(D^{-2\text{p}}\right) - m\sum\log d_i^{2(p_i-1)} - m\sum\log d_i^{-2p_i}$ which is symmetric in p about $\text{p} = (1/2,\ldots,1/2)$. If $\hat{\Sigma}$ is allowed to include multiplicative transformations, that is $\hat{\Sigma}_{\text{r,mle,i}} = \left(d^2/m\right)/s_i$ and $\hat{\Sigma}_{\text{c,mle}} = \left(d^2/m\right)\times s_i$ it is clear that the minimizer of the KL divergence is going to be $s_i = 1$ since Eq 3.9 simplifies to $\frac{1}{m}\operatorname{tr}(S)\operatorname{tr}\left(S^{-1}\right)$ where S is the diagonal matrix of s_i (arithmetic mean is equal to the harmonic mean only if all the entries over which the mean is taken are equal). This demonstrates that the orthogonally equivariant MLE closest to the moment estimators in terms of KL divergence must be of the form $((YY^t/m)^{1/2}, (Y^tY/m)^{1/2})$.

Similar covariance matrices

In cases where we want to penalize the dissimilarity between the row and column covariances beyond the magnitude of the eigenvalues, we must consider a penalty function $d(g(\Sigma_{\text{r}}), g(\Sigma_{\text{c}}))$ that jointly penalizes the eigenvalues and the eigenvectors of the two matrices (we now drop the dependence of the penalty on the function g as we will only consider penalties where g is the identity map). The first problem that we must consider is the scale nonidentifiability of the Kronecker structured covariance. Specifically, any penalty we consider must either respect the nonidentifiability (that is $d(\Sigma_{\text{r}}, \Sigma_{\text{c}}) = d(a\Sigma_{\text{r}}, \Sigma_{\text{c}}/a)$) or we must decouple the scale of the individual covariance matrices. We will reparametrize the problem in the following way: let $\operatorname{cov}(\operatorname{vec}(Y)) = \Sigma$ and write $\Sigma = \Sigma_{\text{c}} \otimes \Sigma_{\text{r}}\sigma^2$ where $\operatorname{tr}(\Sigma_{\text{r}}) = \operatorname{tr}(\Sigma_{\text{c}}) = 1$. In this parametrization there is no longer a scale nonidentifiability for the covariance matrices. However, this introduces a new scale parameter, σ^2, that needs to

be estimated, and possibly penalized.

We rewrite the augmented objective function using this new parametrization (previously Equation (3.6):

$$f(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}, \sigma^2) = l(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}, \sigma^2) + \theta_1 d(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}) + \theta_2 m(\sigma^2), \tag{3.11}$$

where $d(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}})$ is a symmetric penalty function on trace 1 symmetric positive definite matrices that is zero only when the matrices are identical and $m(\sigma^2)$ is some penalty function on the magnitude of the total covariance matrix. The choice of penalty function for the scale parameter is reasonably straightforward. Specifically, there exist a plethora of functions $\theta_2 m(\sigma^2)$ that correspond to proper priors on the positive real line and so a choice of any of those yields a penalty on the scale that has been studied in detail in the Bayesian literature. For the purposes of our exposition, we will write $\theta_2 m(\sigma^2) = \theta_2/\sigma^2$ corresponding to an exponential prior on $1/\sigma^2$ with scale parameter θ_2.

The choice of penalty function d is less obvious. We propose using the symmetrized Kullback Leibler divergence between two mean zero multivariate normal distributions. The form for this is

$$d(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}) = \mathrm{tr}(\Sigma_{\mathrm{r}}\Sigma_{\mathrm{c}}^{-1} + \Sigma_{\mathrm{c}}\Sigma_{\mathrm{r}}^{-1} - 2). \tag{3.12}$$

This penalty function possesses several qualities that are desirable when comparing two covariance matrices. First, First, the function is invariant to inversion, that is $d(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}) = d(\Sigma_{\mathrm{r}}^{-1}, \Sigma_{\mathrm{c}}^{-1})$. Secondly, the function is invariant to transformations by the general linear group: for A an invertible matrix, $d(A\Sigma_{\mathrm{r}}A^t, A\Sigma_{\mathrm{c}}A^t) = d(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}})$. The first property implies that this metric penalizes the distance between the row and column covariance matrices as well as similarly penalizing the distance between the row and column precision matrices. The second property indicates the dependence of the distance function on the eigenvalues of $\Sigma_{\mathrm{r}}^{-1/2}\Sigma_{\mathrm{c}}\Sigma_{\mathrm{r}}^{-1/2}$ (simply let $A = \Sigma_{\mathrm{r}}^{-1/2}$) and provides an indication that this penalty is somewhat reasonable for the space of positive definite matrices. Specifically, it turns out that there are several natural metrics on the space of symmetric positive definite matrices

which depend exclusively on the eigenvalues of $\Sigma_{\mathrm{r}}^{-1/2}\Sigma_{\mathrm{c}}\Sigma_{\mathrm{r}}^{-1/2}$, chief among them being the natural Riemannian metric. The discussion of this and several other distance functions is reserved for Appendix B, but the similarity between the proposed penalty function and these geometrically and algebraically motivated distances suggests that the measurements of similarity that we acquire via our penalty correspond to the geometry of the space. Having defined the augmented objective function including all the parameters, we develop an optimization procedure.

Direct Optimization

We write the augmented objective function explicitly (we abuse notation a bit by writing its arguments as the precision matrices):

$$\begin{aligned} f(\Sigma_{\mathrm{r}}^{-1}, \Sigma_{\mathrm{c}}^{-1}, \sigma^2) =& l(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}, \sigma^2) + \theta_1 d(\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}) + \theta_2/\sigma^2 \\ =& \operatorname{tr}(\Sigma_{\mathrm{r}}^{-1} Y \Sigma_{\mathrm{c}}^{-1} Y^t \sigma^{-2}) - m \log |\Sigma_{\mathrm{r}}^{-1}| - m \log |\Sigma_{\mathrm{c}}^{-1}| - m^2 \log \sigma^{-2} \\ & + \theta_1 \operatorname{tr}(\Sigma_{\mathrm{r}}\Sigma_{\mathrm{c}}^{-1} + \Sigma_{\mathrm{r}}^{-1}\Sigma_{\mathrm{c}} - 2) + \theta_2/\sigma^2, \end{aligned}$$

where $\operatorname{tr}(\Sigma_{\mathrm{r}}) = \operatorname{tr}(\Sigma_{\mathrm{c}}) = 1$. The optimization problem is convex in each of the arguments individually and we propose optimizing sequentially, first in Σ_{r}^{-1}, then in Σ_{c}^{-1} and then in σ^2. For fixed values of $\Sigma_{\mathrm{c}_\star}^{-1}$ and $\sigma_\star^2$, write the augmented objective as

$$\begin{aligned} f(\Sigma_{\mathrm{r}}^{-1}, \Sigma_{\mathrm{c}\star}^{-1}, \sigma_\star^2) =& \operatorname{tr}(\Sigma_{\mathrm{r}}^{-1} Y \Sigma_{\mathrm{c}\star}^{-1} Y^t \sigma_\star^{-2}) - m \log |\Sigma_{\mathrm{r}}^{-1}| + \theta_1 \operatorname{tr}(\Sigma_{\mathrm{r}}\Sigma_{\mathrm{c}\star}^{-1} + \Sigma_{\mathrm{r}}^{-1}\Sigma_{\mathrm{c}\star} - 2) \\ & + \operatorname{const}(\Sigma_{\mathrm{c}\star}, \sigma_\star^2). \end{aligned} \tag{3.13}$$

In order to minimize this function we must construct the Lagrangian with a Lagrange term for the constraint $\operatorname{tr}(\Sigma_{\mathrm{r}}) = 1$: $h(\Sigma_{\mathrm{r}}^{-1}) = f(\Sigma_{\mathrm{r}}^{-1}, \Sigma_{\mathrm{c}\star}^{-1}, \sigma_\star^2) + \eta_1(\operatorname{tr}(\Sigma_{\mathrm{r}}) - 1)$. Taking a derivative with respect to Σ_{r}^{-1} and setting it equal to zero we are able to construct a quadratic equation

in matrix argument that is solveable for Σ_{r} in terms of $\eta_1, \Sigma_{\mathrm{c}\star}$ and $\sigma^2_\star$.

$$\frac{\partial h}{\partial \Sigma_{\mathrm{r}}^{-1}} = Y\Sigma_{\mathrm{c}}^{-1}Y^t/\sigma^2 - m\Sigma_{\mathrm{r}} + \theta_1\Sigma_{\mathrm{c}} - \theta_1\Sigma_{\mathrm{r}}\Sigma_{\mathrm{c}}^{-1}\Sigma_{\mathrm{r}} - \eta_1\Sigma_{\mathrm{r}}^2 = 0 \tag{3.14}$$

$$0 = (Y\Sigma_{\mathrm{c}}^{-1}Y^t/\sigma^2 + \theta_1\Sigma_{\mathrm{c}}) - m\Sigma_{\mathrm{r}} - \Sigma_{\mathrm{r}}(\theta_1\Sigma_{\mathrm{c}}^{-1} + \eta_1 I)\Sigma_{\mathrm{r}} \tag{3.15}$$

$$0 = \Sigma_{\mathrm{r}}^{-1}(Y\Sigma_{\mathrm{c}}^{-1}Y^t/\sigma^2 + \theta_1\Sigma_{\mathrm{c}})\Sigma_{\mathrm{r}}^{-1} - m\Sigma_{\mathrm{r}}^{-1} - (\theta_1\Sigma_{\mathrm{c}}^{-1} + \eta_1 I), \tag{3.16}$$

where in Equation (3.15) we combined terms and in Equation (3.16) we multiply on the left and right by Σ_{r}^{-1}. We observe that $W = (Y\Sigma_{\mathrm{c}}^{-1}Y^t/\sigma^2 + \theta_1\Sigma_{\mathrm{c}})$ necessarily admits a square root and so we can rewrite (3.16) by multiplying on the left and right by $W^{1/2}$ (unique symmetric square root of W):

$$0 = (W^{1/2}\Sigma_{\mathrm{r}}^{-1}W^{1/2})(W^{1/2}\Sigma_{\mathrm{r}}^{-1}W^{1/2}) - mW^{1/2}\Sigma_{\mathrm{r}}^{-1}W^{1/2} - W^{1/2}(\theta_1\Sigma_{\mathrm{c}}^{-1} + \eta_1 I)W^{1/2}.$$

We can now define a new variable $X := (W^{1/2}\Sigma_{\mathrm{r}}^{-1}W^{1/2})$ and note that the above is a quadratic equation in X: $0 = X^2 - mX - W^{1/2}(\theta_1\Sigma_{\mathrm{c}}^{-1} + \eta_1 I)W^{1/2}$. It is well known that a matrix quadratic equation with a leading identity matrix has a unique root if the matrix coefficient on the "linear" term commutes with the constant term and the "discriminant" of the quadratic equation admits a matrix square root [Higham and Kim, 2001]. These conditions are satisfied for the above equation and the unique solution is given by $2X = mI + (m^2 I + 4W^{1/2}(\theta_1\Sigma_{\mathrm{c}}^{-1} + \eta_1 I)W^{1/2})^{1/2}$. We can map this solution back to Σ_{r} via the map $\Sigma_{\mathrm{r}} = W^{1/2}X^{-1}W^{1/2}$. Plugging this into the constraint, $\mathrm{tr}(\Sigma_{\mathrm{r}}) = 1$, allows us to solve for the Lagrange multiplier η_1 via numerical optimization. Following this optimization we can recover Σ_{r}. An equivalent step can be defined for Σ_{c}.

We can update σ^2 via a similar approach: take derivatives of the objective function with respect to σ^{-2} and set equal to zero,

$$\begin{aligned}
\frac{\partial f}{\partial \sigma^{-2}} &= \mathrm{tr}(\Sigma_{\mathrm{r}}^{-1}Y\Sigma_{\mathrm{c}}^{-1}Y^t) - m^2\sigma^{-2} + \theta_2 \\
0 &= \mathrm{tr}(\Sigma_{\mathrm{r}}^{-1}Y\Sigma_{\mathrm{c}}^{-1}Y^t) - m^2\sigma^2 + \theta_2 \\
\sigma^2 &= (\mathrm{tr}(\Sigma_{\mathrm{r}}^{-1}Y\Sigma_{\mathrm{c}}^{-1}Y^t) + \theta_2)/m^2.
\end{aligned}$$

By iterating the steps for Σ_r, Σ_c and σ^2 we define a procedure that finds the minimum of the augmented objective function. Since the scaled log likelihood is bounded below, this iterative procedure is guaranteed to converge.

Selecting the penalty terms θ_1 and θ_2 To choose θ_1 and θ_2 we propose to perform cross validation by excluding 20% of the data, and estimating the covariance parameters using a data augmentation step to account for the missing data. We perform this for multiple (θ_1, θ_2) pairs and select the pair with the smallest cross validation error. For this approach to find the optimizer of the objective function by integrating over the missing values we would need to implement an EM algorithm. The EM algorithm for this particular problem requires calculating $E[\text{tr}(Y\Sigma_c^{-1}Y^t\Sigma_r^{-1})|Y_{\text{observed}}, \Sigma_r, \Sigma_c]$. Simplifying the above expression by considering $\text{tr}(E[Y\Sigma_c^{-1}Y^t|Y_{\text{observed}}, \Sigma_r, \Sigma_c]\Sigma_r^{-1})$ we note that the computation still requires conditional second moments. Allen and Tibshirani [2010] have previously derived this E step and provide a compact form for this expectation. However, even in its' compact form it requires the construction of m^2 $m \times m$ sparse matrices at each iteration. This is effectively intractable for most matrix sizes. Instead of performing a full E step, we propose to perform an approximation: at each iteration we update the unobserved values by setting them equal to their conditional expectations. A one step version of this is used by Allen and Tibshirani [2010] and they demonstrate that it performs similar to the complete EM.

We demonstrate the performance of this method for choosing the penalty parameter in the context of an unknown mean below.

3.3 Mean model

Known mean problems are rarely available outside the realm of classroom exercises. As such, it is desirable to incorporate mean parameter estimation in any estimation procedure we develop for the covariance parameters. We illustrate the importance of accounting for the covariance parameters when estimating the mean in the context of a matrix normal

distribution with a single mean parameter μ. Specifically, let

$$\begin{aligned} Y &= \mu \mathbf{1}\mathbf{1}^t + E \\ E &\sim N_{m\times m}(0, \Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}), \end{aligned} \tag{3.17}$$

where μ is a scalar, $\mathbf{1}$ is a vector of ones of dimension m, and Σ_{r} and Σ_{c} are $m \times m$ positive definite matrices. A naive unbiased estimate of the mean parameter μ is given by the moment estimator $\hat{\mu}^{(0)} = \bar{Y}_{\cdot\cdot\cdot}$. Alternatively, when Σ_{r} and Σ_{c} are known, the oracle estimator of μ is

$$\hat{\mu}^{(\mathrm{ora})} = \frac{\mathbf{1}^t\Sigma_{\mathrm{r}}^{-1}Y\Sigma_{\mathrm{c}}^{-1}\mathbf{1}}{\mathbf{1}^t\Sigma_{\mathrm{r}}^{-1}\mathbf{1}\mathbf{1}^t\Sigma_{\mathrm{c}}^{-1}\mathbf{1}}.$$

This estimator is also unbiased, but has lower variance than $\hat{\mu}^{(0)}$:

$$\mathrm{var}(\hat{\mu}^{(0)}) = \mathrm{E}[\hat{\mu}^{(0)}\hat{\mu}^{(0)}] - \mathrm{E}[\hat{\mu}^{(0)}]^2 = \mathrm{E}[\mathbf{1}^tY\mathbf{1}\mathbf{1}^tY\mathbf{1}]/m^4 - (\mathbf{1}^t\mathbf{1}\mathbf{1}^t\mathbf{1}\mu)^2/m^4 = \mathbf{1}^t\Sigma_{\mathrm{c}}\mathbf{1}\mathbf{1}^t\Sigma_{\mathrm{r}}\mathbf{1}/m^4,$$

versus

$$\begin{aligned} \mathrm{var}(\hat{\mu}^{(\mathrm{ora})}) &= \mathrm{E}[\hat{\mu}^{(\mathrm{ora})}\hat{\mu}^{(\mathrm{ora})}] - \mathrm{E}[\hat{\mu}^{(\mathrm{ora})}]^2 \\ &= \mathrm{E}[\frac{\mathbf{1}^t\Sigma_{\mathrm{r}}^{-1}Y\Sigma_{\mathrm{c}}^{-1}\mathbf{1}}{\mathbf{1}^t\Sigma_{\mathrm{r}}^{-1}\mathbf{1}\mathbf{1}^t\Sigma_{\mathrm{c}}^{-1}\mathbf{1}}\frac{\mathbf{1}^t\Sigma_{\mathrm{r}}^{-1}Y\Sigma_{\mathrm{c}}^{-1}\mathbf{1}}{\mathbf{1}^t\Sigma_{\mathrm{r}}^{-1}\mathbf{1}\mathbf{1}^t\Sigma_{\mathrm{c}}^{-1}\mathbf{1}}] - (\frac{\mathbf{1}^t\Sigma_{\mathrm{r}}^{-1}\mathbf{1}\mathbf{1}^t\Sigma_{\mathrm{c}}^{-1}\mathbf{1}}{\mathbf{1}^t\Sigma_{\mathrm{r}}^{-1}\mathbf{1}\mathbf{1}^t\Sigma_{\mathrm{c}}^{-1}\mathbf{1}}\mu)^2 \\ &= \frac{\mathbf{1}^t\Sigma_{\mathrm{r}}^{-1}\mathrm{E}[Y\Sigma_{\mathrm{c}}^{-1}\mathbf{1}\mathbf{1}^t\Sigma_{\mathrm{r}}^{-1}Y]\Sigma_{\mathrm{c}}^{-1}\mathbf{1}}{(\mathbf{1}^t\Sigma_{\mathrm{r}}^{-1}\mathbf{1}\mathbf{1}^t\Sigma_{\mathrm{c}}^{-1}\mathbf{1})^2} - \mu^2 \\ &= \frac{\mathbf{1}^t\Sigma_{\mathrm{r}}^{-1}[\Sigma_{\mathrm{r}}\Sigma_{\mathrm{r}}^{-1}\mathbf{1}\mathbf{1}^t\Sigma_{\mathrm{c}}^{-1}\Sigma_{\mathrm{c}} + \mu^2\mathbf{1}\mathbf{1}^t\Sigma_{\mathrm{c}}^{-1}\mathbf{1}\mathbf{1}^t\Sigma_{\mathrm{r}}^{-1}\mathbf{1}\mathbf{1}^t]\Sigma_{\mathrm{c}}^{-1}\mathbf{1}}{(\mathbf{1}^t\Sigma_{\mathrm{r}}^{-1}\mathbf{1}\mathbf{1}^t\Sigma_{\mathrm{c}}^{-1}\mathbf{1})^2} - \mu^2 \\ &= 1/(\mathbf{1}^t\Sigma_{\mathrm{r}}^{-1}\mathbf{1}\mathbf{1}^t\Sigma_{\mathrm{c}}^{-1}\mathbf{1}). \end{aligned}$$

In the more general regression framework we consider models of the form

$$\begin{aligned} Y &= <\mathbf{X}, \beta> + E \\ E &\sim N_{m\times m}(0, \Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}), \end{aligned} \tag{3.18}$$

where $< \mathbf{X}, \beta >$ is the inner product of the array of fixed covariates $\mathbf{X}$ and the vector of coefficients β (*i*.e. $y_{ij} = x_{ij}^t \beta + \epsilon_{ij}$). In this setting, the oracle estimator of β when Σ_{r} and Σ_{c} are known is given by

$$\hat{\beta}^{(\mathrm{ora})} = (\mathrm{mat}(\mathbf{X})^t (\Sigma_{\mathrm{c}}^{-1} \otimes \Sigma_{\mathrm{r}}^{-1}) \mathrm{mat}(\mathbf{X}))^{-1} \mathrm{mat}(\mathbf{X}) \mathrm{vec}(Y),$$

where mat is the matricization operator of Kolda [2006] and vec is the vectorization operator. This can also be viewed as the generalized least squares (GLS) estimator as $\hat{\beta}^{(\mathrm{ora})}$ is the value that optimizes the function $\mathrm{vec}(Y - < \mathbf{X}, \beta >)^t (\Sigma_{\mathrm{c}}^{-1} \otimes \Sigma_{\mathrm{r}}^{-1}) \mathrm{vec}(Y - < \mathbf{X}, \beta >)$. We will write $\hat{\beta}^{(\mathrm{gls})}$ to indicate this explicitly. The variance of this estimator is given by $(\mathrm{mat}(\mathbf{X})^t (\Sigma_{\mathrm{c}}^{-1} \otimes \Sigma_{\mathrm{r}}^{-1}) \mathrm{mat}(\mathbf{X}))^{-1}$ and it is an immediate consequence of the Gauss-Markov theorem that this estimator is the best linear unbiased estimator of β.

In the particular case of relational data, we are able to explore the properties of $\hat{\beta}^{(\mathrm{gls})}$ for specific types of covariates. It is common in relational data analysis to have covariates that are specific to the rows of the data matrix (called sender covariates), the columns of the data matrix (called receiver covariates) or to each element of the matrix (called dyadic covariates). The covariates array $\mathbf{X}$ is thus an $m \times m \times p$ array where each $m \times m$ slice corresponds to a specific covariate. The different types of covariates correspond to the following slices of $\mathbf{X}$:

- Overall mean: $\mathbf{X}_1 = \mathbf{1}\mathbf{1}^t$
- Row: $\mathbf{X}_{\mathrm{r}} = x_{\mathrm{r}} \mathbf{1}^t$, $x_{\mathrm{r}} \in \mathbb{R}^m$
- Column: $\mathbf{X}_{\mathrm{c}} = \mathbf{1} x_{\mathrm{c}}^t$, $x_{\mathrm{c}} \in \mathbb{R}^m$
- Dyadic (multiplicative): $\mathbf{X}_{\mathrm{md}} = x_{\mathrm{rd}} x_{\mathrm{cd}}^t$
- Dyadic (general): $\mathbf{X}_{\mathrm{dyad}}$ is an unstructured $m \times m$ matrix.

It is well known that the oracle GLS estimator for the regression parameters (when the covariance matrices are known) is more efficient than the OLS estimator. Specifically, it is easy to see that $\mathrm{var}(\hat{\beta}^{(\mathrm{ols})}) - \mathrm{var}(\hat{\beta}^{(\mathrm{gls})}) \geq 0$. However, this requires conditioning on both the covariates $\mathbf{X}$ and the covariance matrices Σ_{r} and Σ_{c}. Figure (3.1) explores this relative efficiency of OLS to GLS for independent normally distributed covariates and

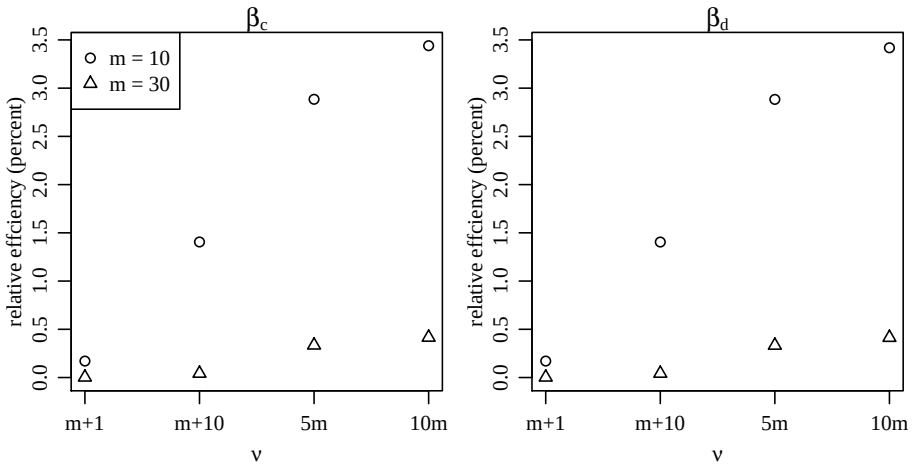

Figure 3.1: Relative efficiency of OLS to oracle GLS estimator. For both panels, the covariates are independent and identically distributed normal and the two covariance matrices are distributed as inverse Wishart variables concentrated around the same matrix A with concentration parameter ν (the matrix A is in turn also generated from a Wishart with concentration parameter $m+1$ centered around the identity).

varying degrees of similarity among the trace 1 covariance matrices Σ_{r} and Σ_{c}. Each point in the figure corresponds to the average across 100,000 simulated datasets of the ratio of $\mathrm{var}(\hat{\beta}^{(\mathrm{ols})})/\mathrm{var}(\hat{\beta}^{(\mathrm{gls})})$ where for each simulated dataset we have:

1. $A \sim \mathrm{Wish}(I, m+2)$,

2. $\tilde{\Sigma}_{\mathrm{r}}, \tilde{\Sigma}_{\mathrm{c}} \sim \mathrm{inverseWish}(A, \nu)$, $\Sigma_{\mathrm{r}} = \tilde{\Sigma}_{\mathrm{r}}/\mathrm{tr}(\tilde{\Sigma}_{\mathrm{r}})$ and $\Sigma_{\mathrm{c}} = \tilde{\Sigma}_{\mathrm{c}}/\mathrm{tr}(\tilde{\Sigma}_{\mathrm{c}})$

3. $x_c, x_r \sim N_m(0, I)$

4. Calculate

   - $\mathrm{var}(\hat{\beta}_{\mathrm{c}}^{(\mathrm{ols})}) = \frac{(\mathbf{1}^t\Sigma_{\mathrm{r}}\mathbf{1})(x_c^t\Sigma_{\mathrm{c}}x_c)}{((x_c^t x_c)^2 m^2)}$
   - $\mathrm{var}(\hat{\beta}_{\mathrm{c}}^{(\mathrm{gls})}) = \frac{1}{(\mathbf{1}^t\Sigma_{\mathrm{r}}^{-1}\mathbf{1})(x_c^t\Sigma_{\mathrm{c}}^{-1}x_c)}$
   - $\mathrm{var}(\hat{\beta}_{\mathrm{dyad}}^{(\mathrm{ols})}) = \frac{(x_r^t\Sigma_{\mathrm{r}}x_r)(x_c^t\Sigma_{\mathrm{c}}x_c)}{((x_c^t x_c)(x_r^t x_r))^2}$
   - $\mathrm{var}(\hat{\beta}_{\mathrm{dyad}}^{(\mathrm{gls})}) = \frac{1}{(x_r^t\Sigma_{\mathrm{r}}^{-1}x_r)(x_c^t\Sigma_{\mathrm{c}}^{-1}x_c)}$

The panels of Figure (3.1) provide the relative efficiency of OLS to GLS in percents. We note that the two panels, representing the relative efficiencies of a column covariate and a

multiplicative dyadic covariate (left and right panels respectively), are very similar suggesting that the estimators are expected to perform similarly for different types of covariates. We note that there is a reduction in the relative efficiency of GLS as the parameter ν, which controls the similarity between the two covariance matrices, increases . However this is greatly tampered by an increase in the matrix size from $m = 10$ to $m = 30$. It is clear that when the covariance matrices themselves need to be estimated, the efficiency of the estimator of the regression parameters will not be as high as that of GLS. We explore the behavior of several such estimators below.

Feasible GLS

A feasible GLS estimator involves substituting an estimate of the covariance matrix into the formula for the estimate of the regression parameters. For example, for the overall mean from before, the FGLS estimator based on some estimates of the row and column covariances $\hat{\Sigma}_{\mathrm{r}}$ and $\hat{\Sigma}_{\mathrm{c}}$ is given by

$$\hat{\mu}^{(\mathrm{gls})}_{(\hat{\Sigma}_{\mathrm{r}},\hat{\Sigma}_{\mathrm{c}})} = \frac{\mathbf{1}^t\hat{\Sigma}_{\mathrm{r}}^{-1}Y\hat{\Sigma}_{\mathrm{c}}^{-1}\mathbf{1}}{\mathbf{1}^t\hat{\Sigma}_{\mathrm{r}}^{-1}\mathbf{1}\mathbf{1}^t\hat{\Sigma}_{\mathrm{c}}^{-1}\mathbf{1}}.$$

We will explicitly denote dependence on an estimate of the covariances via a subscript. As such, to assess the quality of the FGLS estimator, we must specify the estimators of the covariance. It is important to note that if the form of the estimator is incorrectly specified, the FGLS estimator might be less efficient than the OLS estimator (see for example Amemiya and Fuller [1967]). However, if there exists a consistent estimator of the covariance matrix, the FGLS estimator and the oracle GLS estimator have the same distribution asymptotically [Maddala, 1971, Zellner, 1962]. Similarly, approximate normality [Rothenberg, 1984], convergence of iterative procedures and unbiasedness of estimates of regression parameters [Andrews, 1986, Henk Don and Magnus, 1980] have all been demonstrated under different regularity conditions (the most important condition being that the likelihood is bounded).

In the case of relational data, when there is only one observed matrix, the regression problem in Equation (3.18) has an unbounded likelihood if $\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}$ are full matrices and β must be jointly estimated. A possible estimator for β in this case is a one step feasible GLS

estimator. The procedure for it is as follows:

1. Construct the FGLS estimator, $\hat{\beta}^{(0)}$, based on an initial estimate of the covariance matrices (for example, independence).

2. Estimate the covariance matrices based on $Y - < \mathbf{X}, \hat{\beta}^{(0)} >$.

3. Construct the new FGLS estimator $\hat{\beta}^{(\text{gls})}_{\hat{\Sigma}_\text{r},\hat{\Sigma}_\text{c}}$ based on the estimates from 2.

When step 2 above is based on the likelihood equations (3.1) and (3.2) for the covariances, it is not possible to repeat the procedure indefinitely for the relational data problem (hence calling an estimator based on a single run through steps 1-3 a "one step" estimator). It does not appear that there are any theoretical guarantees for these estimators. If instead of full covariance matrices, the covariance matrices can be written in terms of only a few parameters, then iterating steps 2 and 3 (where step 2 is again based on the likelihood equations) yields an iterative procedure that finds the maximum likelihood estimators. A similar result holds when step 2 is based on a single step of the the penalized likelihood approach of Section 2. We note that while we cannot verify all of the assumptions of Henk Don and Magnus [1980] to guarantee the convergence of the iterative estimation procedure for the penalized approach, in practice our algorithm converges. As such we have the following conjecture about the iterative penalized procedure:

Conjecture 7. *For the iterative procedure defined above with step 2 based on the penalized likelihood of Section 2 we have:*

1. *The distribution of the estimator at the j^{th} iteration of step 3, denoted by $\hat{\beta}^{(\text{gls})}_{\hat{\Sigma}_\text{r},\hat{\Sigma}_\text{c},j}$, is symmetric about the true parameter β.*

2. *If the expectation of $\hat{\beta}^{(\text{gls})}_{\hat{\Sigma}_\text{r},\hat{\Sigma}_\text{c},j}$ exists then it is an unbiased estimator of β.*

3. *If the ratio of the maximal to minimal eigenvalues of the Kronecker structured covariance matrix is bounded with probability 1, then the iterative procedure converges to a minimum of the augmented objective function and the estimator of β at convergence is unbiased.*

Part 1 of the Conjecture is an immediate consequence of Proposition 1 of Henk Don and Magnus [1980] and can be easily verified by noting that the estimator of the covariance matrices at step $j-1$ is an even function of the residual $Y - < \mathbf{X}, \hat{\beta}^{(\text{gls})}_{\hat{\Sigma}_\text{r}, \hat{\Sigma}_\text{c}, j-1} >$. Parts 2 and 3 are complicated: To verify part 2, we must show that the expectation of the trace of the estimate of the precision matrices is finite at each iteration of the iterative procedure. Part 3 of the Conjecture rests on demonstrating that the penalty that we proposed keeps the ratio of eigenvalues bounded. To verify these two parts we would need to find the explicit form of the eigenvalues of the precisions and covariance matrices or verify Assumption 5 of Henk Don and Magnus [1980] which the authors describe as being effectively impossible to verify in practice (and we have not been able to verify in this case).

Simulation We now demonstrate that the penalized procedure proposed above performs well in terms of the risk of estimating mean parameters (under quadratic loss) when compared to OLS and a one step FGLS estimator. We simulate datasets from the following model:

- Let $(\beta_0, \beta_\text{r}, \beta_\text{c}, \beta_\text{dyad}) = (5, 1, 1, 1)$.
- Generate $x_\text{r} = x_\text{c} \sim N_m(0, I)$ and $X_\text{dyad} \sim N_{m\times m}(0, I, I)$.
- Generate the covariance matrices:
  - Generate $A \sim \text{Wish}(I, m+2)$,
  - Generate $\tilde{\Sigma}_\text{r}, \tilde{\Sigma}_\text{c} \sim \text{inverseWish}(A, \nu)$, $\Sigma_\text{r} = \tilde{\Sigma}_\text{r}/\text{tr}(\tilde{\Sigma}_\text{r})$ and $\Sigma_\text{c} = \tilde{\Sigma}_\text{c}/\text{tr}(\tilde{\Sigma}_\text{c})$
  - Generate $\sigma^{-2} \sim \text{Gamma}(10, 1)$
- Generate $E \sim N_{m\times m}(0, I, I)$
- Set $Y = \beta_0 \mathbf{1}\mathbf{1}^t + \beta_\text{r} x_\text{r} \mathbf{1}^t + \beta_\text{c} \mathbf{1} x_\text{c}^t + \beta_\text{dyad} X_\text{dyad} + \sigma \Sigma_\text{r}^{1/2} E \Sigma_\text{c}^{1/2}$

We conduct the simulation for $m \in \{10, 20\}$ and different parameters ν that control the similarity between the two covariance matrices. For example, for $m = 10$, when $\nu = 11$, the

average distance (according to the symmetric KL penalty we use for estimation) between two covariance matrices is 1300 while when $\nu = 50$ the average distance is 5.6.

Recall from Section 2 that we must choose penalty parameters θ_1 and θ_2 in order to perform the penalized procedure. We note that there is abundant data information for estimating $\sigma^2 = \text{tr}(\text{cov}(Y))$ and propose to use a plug-in estimate of σ^2, thus eliminating the need for choosing a parameter θ_2. In this case, rather than solving the equation for σ^2 on page 44 we let $\hat{\sigma}^2$ be the moment estimate of the scale of the variability based on a one step estimator of the mean parameters: $\hat{\sigma}^2 = \text{tr}((Y - < \mathbf{X}, \hat{\beta} >)^t (Y - < \mathbf{X}, \hat{\beta} >))$. We have observed that this simplification performs well in practice.

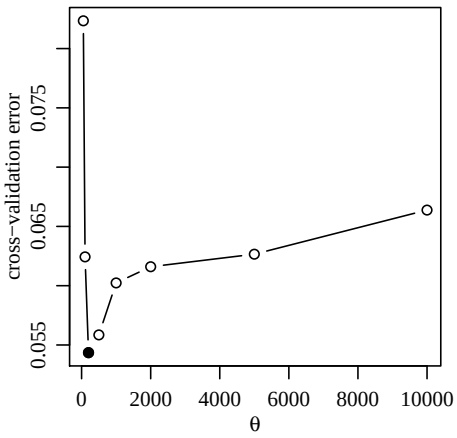

Figure 3.2: Choice of the penalty parameter θ based on five fold cross validation for a 10×10 matrix.

As such we are required to pick θ (we drop the subscript as there is now only one penalty parameter). To do so we perform the cross validation procedure described at the end of Section 2. Figure 3.2 illustrates the results of the cross validation for a single dataset with $m = 10$ and $\nu = 11$. We chose the range of θ to be big (10 to 10000) as we have not established any criteria that would provide an *a*priori magnitude for θ. For this particular dataset the cross validation error is smallest for $\theta = 200$.

Following the proper choice of θ for each dataset based on cross validation, we reran the analysis on the full data. For each dataset we then constructed estimates $\hat{\beta}^{(\text{ols})}$,$\hat{\beta}^{(\text{FGLS})}$ and

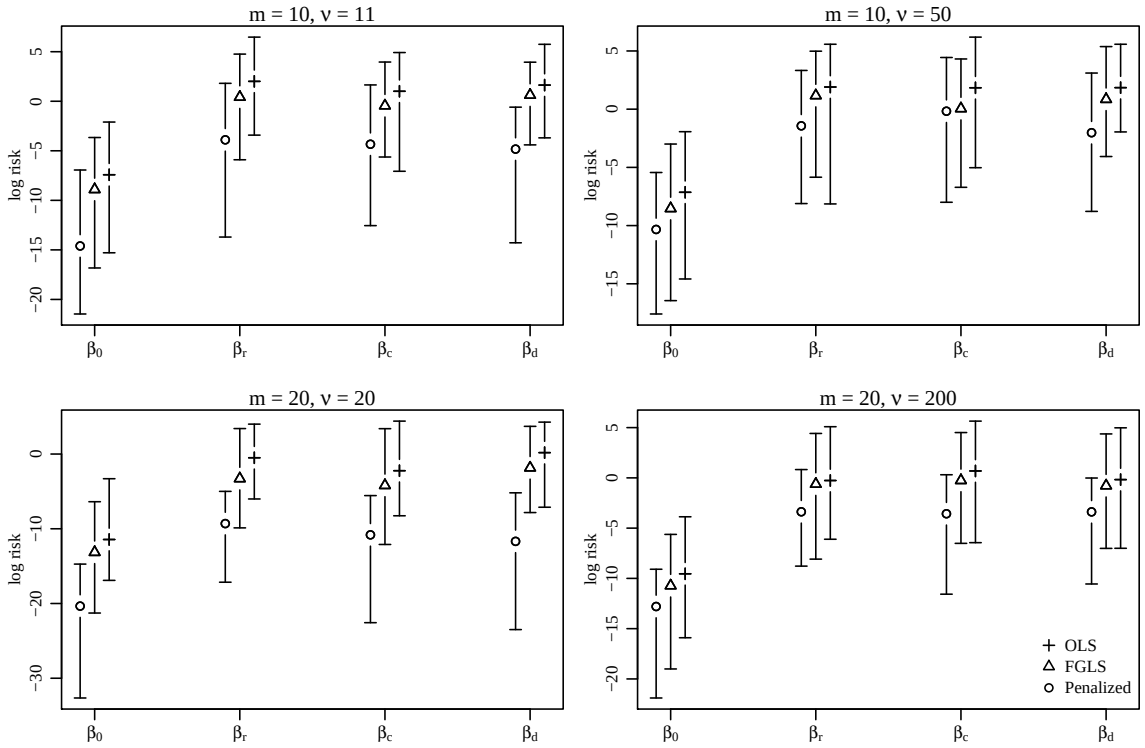

Figure 3.3: 95% confidence intervals of the estimated log risk under quadratic loss for each regression parameter. Each confidence interval in the top row of panels is based on 50 simulated 10×10 datasets while each confidence interval in the bottom row of panels is based on 35 20×20 datasets.

$\hat{\beta}^{(pen)}$. The feasible GLS estimator was based on a one step FGLS where the covariances were estimated from the OLS errors, $\hat{E}$, as $((\hat{E}\hat{E}^t/m)^{1/2}, (\hat{E}^t\hat{E}/m)^{1/2})$ - which is the estimator that was discussed in detail in Section 2. We then calculate the quadratic loss for each individual coefficient. Figure 3.3 plots the 95% confidence intervals of the log transformed quadratic losses for each dataset and each term in β. We use the log transformed risk as the distribution of the untransformed risk is highly skewed. The leftmost bar in each series of three represents the risk for the penalized estimates and we see that those are consistently smaller than those for the FGLS and OLS estimator. Since the penalized estimates outperform FGLS and OLS for both large and small ν, it appears that the procedure for choosing the penalty parameter θ is reasonable, as it appropriately adapts to scenarios where the

true matrices are very similar (the right hand column of the Figure) or very dissimilar (the left hand column of the Figure). We note that the improvement presented by the penalized procedure appears to be greater when the matrices are dissimilar, which confirms what was suggested by Figure 3.1 that the relative efficiency of GLS methods decreases as the covariance matrices become more similar.

In all four panels we see that the grand mean is estimated very well by all the estimators (with the upper bounds of the 95% confidence intervals all being less than 0). These plots do not provide strong evidence that the one step estimator provides an improvement over the naive OLS estimator.

3.4 Discussion

In this chapter we developed a procedure for estimating mean parameters and covariance matrices for relational data. We leverage the fact that the rows and columns of a relational data matrix share the same index set. As such, similarity among rows or a relational data matrix might reasonably imply a similarity among the corresponding columns of the relational data matrix. This structure suggests modeling the row and column covariance matrices of a relational dataset as similar. We propose incorporating this similarity into the estimation procedure via two different penalties. The first penalty acts on the eigenvalues of the two matrices and leads to a unique maximum likelihood estimator of the covariances (for any choice of penalty) in the case when the mean of the distribution is known. This MLE minimizes the Bayes risk among a class of equivariant MLEs for a general class of priors and is the MLE that minimizes the Kullback Leibler divergence from the moment estimators in this class. The second type of penalty deals directly with similarity between the two covariance matrices via the symmetric Kullback Leibler divergence. The estimation procedure based on this penalty does not yield a maximum likelihood estimator. However, this procedure is adaptive, in the sense that a penalty parameter can be used to vary the influence of the penalty and so affect the similarity of the two covariance matrices. Thus, by letting the penalty parameter approach zero, we can recover the set of MLEs when the mean is known. We demonstrated that when the mean parameters of the relational data is unknown, estimators of the mean parameters that do not take account of the covariance

structure are inefficient. We then outline an iterative procedure for constructing the feasible generalized least squares estimator for the mean parameters using the dissimilarity penalty we developed in Section 2. The estimates based on this procedure are shown to have smaller risk than the naive OLS estimates or estimates based on a one step procedure.

Frequently relational data is not measured on a continuous scale. In these cases it is common to model this data via a continuous latent variable [Nowicki and Snijders, 2001, Hoff et al., 2002, Hoff, 2005a]. The approach developed in this chapter can be naturally extended to these latent variable models via a Bayesian procedure. Recall that the proposed penalty corresponds to a proper prior on the product space of two symmetric positive definite matrices. As such, after establishing the proper likelihood to use for the non-continuous relational data, we can incorporate this prior into the Markov chain Monte Carlo procedure outlined in Hoff et al. [2012]. We note that the prior distribution for the covariance matrices does not have a familiar form and so sampling from it currently requires rejection sampling or a Metropolis-Hastings step. This is an ongoing area of research.

Chapter 4

HIERARCHICAL ARRAY PRIORS FOR ANOVA DECOMPOSITIONS

4.1 Introduction

Cross-classified data are prevalent in many disciplines, including the social and health sciences. For example, a survey or observational study may record health behaviors of its participants, along with a variety of demographic variables such as age, ethnicity and education level, by which the participants can be classified. A common data analysis goal in such settings is the estimation of the health behavior means for each combination of levels of the demographic factors. In a three-way layout, for example, the goal is to estimate the three-way table of population cell means, where each cell corresponds to a particular combination of factor levels. A standard estimator of the table is provided by the table of sample means, which can alternatively be represented by its ANOVA decomposition into additive effects and two- and three-way interaction terms.

The cell sample means provide an unbiased estimator of the population means, as long as there are observations available for each cell. However, if the cell-specific sample sizes are small then it may be desirable to share information across the cells to reduce the variance of the estimator. Perhaps the simplest and most common method of information sharing is to assume that certain mean contrasts among levels of one set of factors are equivalent across levels of another set of factors, or equivalently, that certain interaction terms in the ANOVA decomposition of population cell means are exactly zero. This is a fairly large modeling assumption, and can often be rejected via plots or standard F-tests. If such assumptions are rejected, it still may be desirable to share information across cell means, although perhaps in a way that does not posit exact relationships among them.

As a concrete example, consider estimating mean macronutrient intake across levels of age (binned in 10 year increments), ethnicity and education from the National Health and Nutrition Examination Survey (NHANES). Table 4.1 summarizes the cell-specific sample

sizes for intake of overall carbohydrates as well as two subcategories (sugar and fiber) by age, ethnicity, and education levels for male respondents (more details on these data are provided in Section 4). Studies of carbohydrate intake have been motivated by a frequently cited relationship between carbohydrate intake and health outcomes [Chandalia et al., 2000, Moerman et al., 1993]. Studies of obesity in the US, have shown an overall increase in caloric intake primarily due to a drastic increase in carbohydrate intake from 44 to 48.7 percent of total calories from 1971 to 2006 Austin et al. [2011]. Recently, the types of carbohydrates that are being consumed has become of primary interest. For example, in the study of cardiovascular disease, simple sugars are associated with raising triglycerides and overall cholesterol while dietary fiber has been associated with lowering triglycerides [Albrink and Ullrich, 1986, Yang et al., 2003]. Total carbohydrates and the types of carbohydrates have also been targeted in recent studies of effective weight loss (*e.g.* sugar consumption in the form of HFCS in drinks, Nielsen et al. [2004]).

However, these studies generally report on marginal means of carbohydrate intake across demographic variables, and do not take into account potential non-additivity, or interaction terms, between them [Park et al., 2011, Montonen et al., 2003, Basiotis et al., 1989, Verly Junior et al., 2010, Johansson et al., 2001]. In a study where non-additivity was considered, the authors only tested for the presence of a small subset of possible interactions and did not consider any interactions of more than two effects [Austin et al., 2011]. A more detailed understanding of the relationship between mean carbohydrate intake and the demographic variables can be obtained from a MANOVA decomposition of the means array into main-effects, two- and three-way interactions. Evidence for interactions for multivariate data can be assessed with approximate F-tests based on the Pillai trace statistics [Olson, 1976].

For our data, the F-tests presented in Table 4.2 indicate strong evidence that the two- and three-way interactions are not zero. Based on these results, standard practice would be to retain the full model and describe the interaction patterns via various contrasts of cell sample means. Often this is done by visual examination of interaction plots, that is, plots of cell means by various combinations of factors. For example, Figure 4.1 gives the age by education interaction plots for each of the four ethnicity groups. The three-way interaction between ethnicity, age and education can be described as the inconsistency of

	Mexican					Hispanic					White					Black				
Age	P	S	HD	AD	BD	P	S	HD	AD	BD	P	S	HD	AD	BD	P	S	HD	AD	BD
31-40	21	24	23	17	13	12	8	10	11	1	3	37	56	55	56	1	13	31	35	16
41-50	26	10	19	14	6	11	9	10	9	3	10	25	56	57	50	2	25	21	25	17
51-60	29	11	10	14	10	17	6	12	13	11	10	24	46	57	57	3	23	23	24	14
61-70	31	7	5	11	5	19	4	11	6	7	15	23	56	46	54	16	34	20	33	14
71-80	27	2	3	1	3	10	8	5	2	7	61	37	93	72	68	16	10	11	7	12

Table 4.1: Cross-tabulation of the sample sizes for the demographic variables. "Hispanic" is coded as "Hispanic, not Mexican".

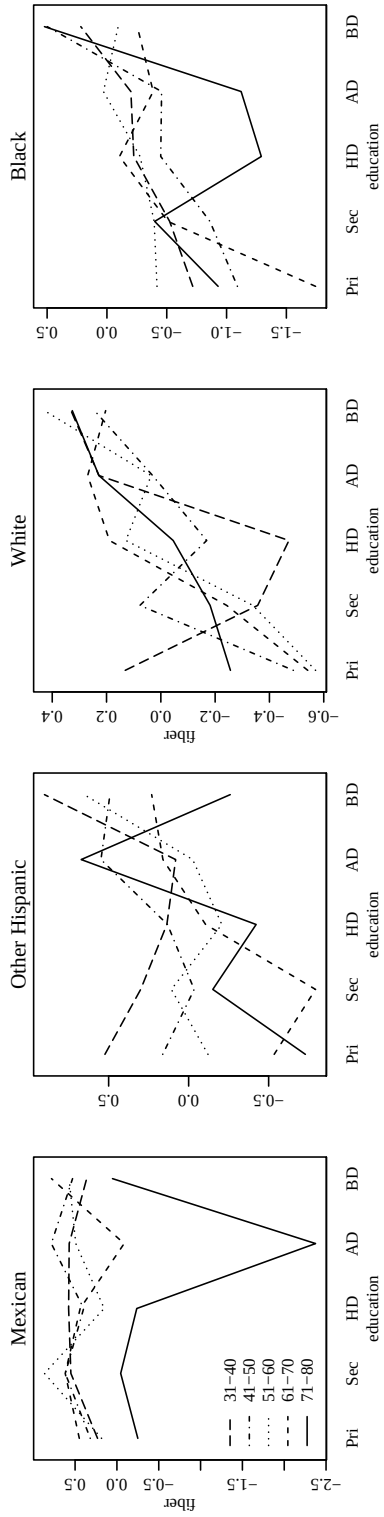

Figure 4.1: Three-way interaction plot of fiber cell means by ethnicity, age and education level.

	approx F	num df	den df	p-value
Education	11.15	15	6102	< 0.01
Ethnicity	18.07	9	6102	< 0.01
Age	21.38	12	6102	< 0.01
Education:Ethnicity	1.67	36	6102	0.01
Education:Age	1.60	48	6102	0.01
Ethnicity:Age	2.05	36	6102	< 0.01
Education:Ethnicity:Age	1.44	144	6102	< 0.01

Table 4.2: MANOVA testing of interaction terms via Pillai's trace statistic.

the two-way interactions across levels of ethnicity. Visually, there is some indication that Mexican respondents have a different age by education interaction than the other ethnicities, but it is difficult to say anything more specific. Indeed, it is difficult to even describe the two-way interactions, due to the high variability of the cell sample means.

Much of the heterogeneity in these plots can be attributed to the low sample sizes in many cells and the resulting sampling variability of the cell sample means. A cleaner picture of the three way interactions could possibly be obtained via cell mean estimates with lower variability. A variety of penalized least squares procedures have been proposed in order to reduce estimate variability and mean squared error (MSE), such as ridge regression and the lasso. Recent variants of these approaches allow for different penalties on ANOVA terms of different orders, including the ASP method of Beran [2005], and grouped versions of the lasso [Yuan and Lin, 2007, Friedman et al., 2010]. Corresponding Bayesian approaches include Bayesian lasso procedures [Yuan and Lin, 2005, Genkin et al., 2007, Park and Casella, 2008] and multilevel hierarchical priors [Pittau et al., 2010, Park et al., 2006, Hodges et al., 2007, Cui et al., 2010].

While these procedures attain a reduced MSE by shrinking linear model coefficient estimates towards zero, they do not generally take full advantage of the structure that is often present in cross-classified datasets. In the data analysis example above, two of the three factors (age and education) are ordinal, with age being a binned version of a continuous predictor. Considering factors such as these more generally, suppose a categorical factor x is a binned version of some underlying continuous or ordinal explanatory variable $\tilde{x}$ (such as

income, age, number of children or education level). If the mean of the response variable y is smoothly varying in the underlying variable $\tilde{x}$, we would expect that adjacent levels of the factor x would have similar main effects and interaction terms. Similarly, for non-ordinal factors (such as ethnic group or religion) it is possible that two levels represent similar populations, and thus may have similar main-effects and interaction terms as well. We refer to such similarities across the orders of the effects as *order consistent interactions*.

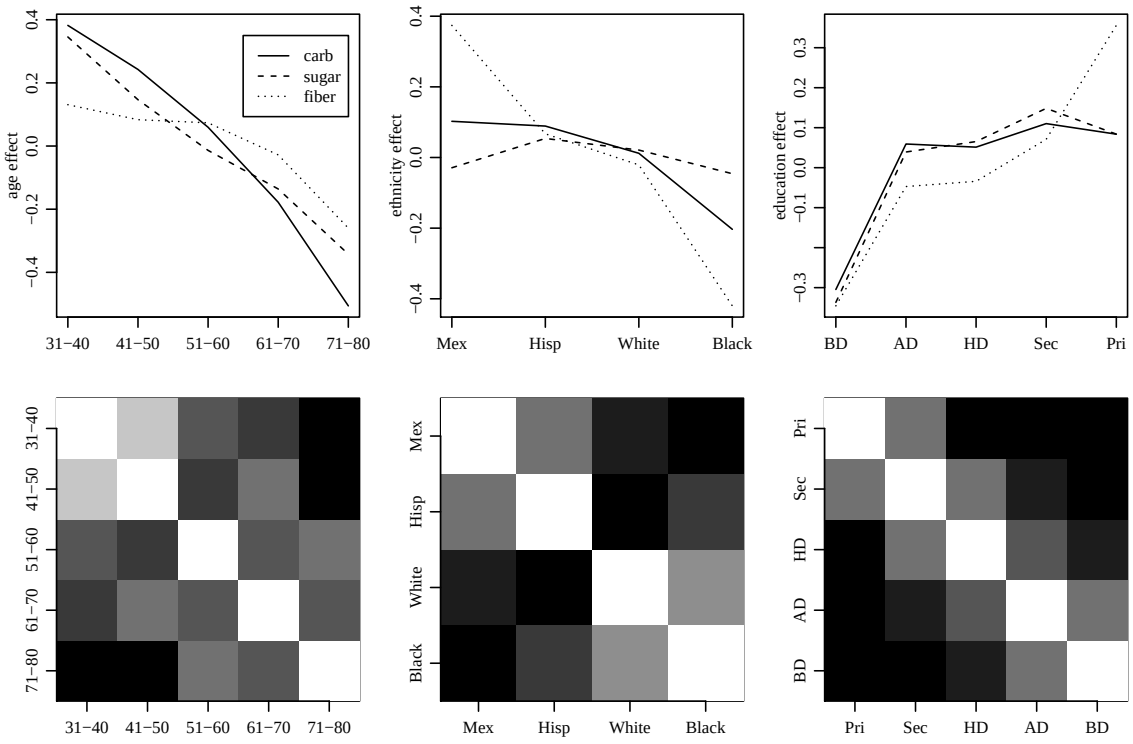

Figure 4.2: Plots of main effects and interaction correlations for the three outcome variables (carbohydrates, sugar and fiber). The first row of plots gives OLS estimates of the main effects for each factor. The second row of plots gives correlations of effects between levels of each factor, with white representing 1 and black representing -1. The interactions are calculated based on OLS estimates of the main effects and two-way interactions of each factor.

Returning to the NHANES data, Figure 4.2 summarizes the OLS estimates of the main effects and two-way interactions for the three outcome variables (carbohydrates, sugar and fiber). Not surprisingly, the main effects for the ordinal factors (age and education) are

"smooth," in that the estimated main effect for a given level is generally similar to the effect for an adjacent level. Additionally, some similarities among the ethnic groups appear consistent across the three outcome variables. To assess consistency of such similarities between main effects and two-way interactions, we computed correlations of parameter estimates for these effects between levels of each factor. For example, there are $3 \times 10 = 30$ main-effect and two-way interaction estimates involving each level of age: For each of the three outcome variables, there is 1 main-effect estimate for each age level, 4 estimates from the age by ethnicity interaction and 5 estimates from the age by education interaction. We compute a correlation matrix for the five levels of age based on the resulting 30×5 matrix of parameter estimates, and similarly compute correlations among levels of ethnicity and among levels of education. The second row of Figure 4.2 gives grayscale plots of these correlation matrices. The results suggest some degree of order consistent interactions: For the ordinal factors, the highest correlations are among adjacent pairs. For the ethnicity factor, the results suggest that on average, the effects for the Mexican category are more similar to the Hispanic (not Mexican) category than to the other ethnic categories, as we might expect.

The OLS estimates of the main effects and three-way interactions presented above, along with the fact that two of the three factors are ordinal, suggest the possibility of order consistent interactions among the array of population cell means. More generally, order consistent interactions may be present in a variety of datasets encountered in the social and health sciences, especially those that include ordinal factors, or factors for which some of the levels may represent very similar populations. In this paper, we propose a novel class of hierarchical prior distributions over main effects and interaction arrays that can adapt to the presence of order consistent interactions. The hierarchical prior distribution provides joint estimates of a covariance matrix for each factor, along with the factor main effects and interactions. Roughly speaking, the covariance matrix for a given factor is estimated from the main effects and interactions in which the factor appears. Conversely, an estimate of a factor's covariance matrix can assist in the estimation of higher-order interactions, for which data information is limited. We make this idea more formal in the next section, where we construct our prior distribution from a set of related array normal distributions

with separable covariance structures [Hoff, 2011], and provide a Markov chain Monte Carlo algorithm for inference under this prior. In Section 3 we provide a simulation study comparing estimation under our proposed prior to some standard estimators. As expected, our approach outperforms others when the data exhibit order consistent interactions. Additionally, for data lacking any interactions, our approach performs comparably to the OLS estimates obtained from the additive model (i.e. the oracle estimator). In Section 4 we extend this methodology to MANOVA models in order to analyze the multivariate NHANES data presented above. In addition to estimates of main effects and interactions, our analysis provides measures of similarity between levels of each of the factors. We conclude in Section 5 with a summary of our approach and a discussion of possible extensions.

4.2 A hierarchical prior for interaction arrays

In this section we introduce the hierarchical array (HA) prior, and present a Markov chain Monte Carlo (MCMC) algorithm for posterior approximation and parameter estimation. The HA prior is constructed from several semi-conjugate priors, and so the MCMC algorithm can be based on a straightforward Gibbs sampling scheme.

4.2.1 The hierarchical array prior

For notational convenience we consider the case of three categorical factors, and note that the HA prior generalizes trivially to accommodate a greater number of factors. Suppose the three categorical factors have levels $\{1,\ldots,m_1\}$, $\{1,\ldots,m_2\}$ and $\{1,\ldots,m_3\}$ respectively. The standard ANOVA model for a three-way factorial dataset is

$$\begin{aligned} y_{ijkl} &= \mu + a_i + b_j + c_k + (ab)_{ij} + (ac)_{ik} + (bc)_{jk} + (abc)_{ijk} + \epsilon_{ijkl} \qquad (4.1) \\ \{\epsilon_{ijkl}\} &\sim \text{i.i.d. normal}(0,\sigma^2). \end{aligned}$$

Let a denote the $m_1 \times 1$ vector of main effects for the first factor, (ab) denote the $m_1 \times m_2$ matrix describing the two-way interaction between the first two factors, (abc) denote the $m_1 \times m_2 \times m_3$ three-way interaction array, and let b, c, (ac), and (bc) be defined similarly. Bayesian inference for this model proceeds by specifying a prior distribution for the ANOVA

decomposition $\theta = \{\mu, a, b, c, (ab), (ac), (bc), (abc)\}$ and the error variance σ^2.

As described in the Introduction, if two levels of a factor represent similar populations, we would expect that coefficients of the decomposition involving these two levels would have similar values. For example, suppose levels i_1 and i_2 of the first factor correspond to similar populations. We might then expect a_{i_1} to be close to a_{i_2}, the vector $\{(ab)_{i_1,j}, j = 1, \ldots, m_2\}$ to be close to the vector $\{(ab)_{i_2,j}, j = 1, \ldots, m_2\}$, and so on. We represent this potential similarity between levels of the first factor with a covariance matrix Σ_a, and consider a mean zero prior distribution on the ANOVA decomposition such that

$$\begin{aligned} \mathrm{Cov}[a] = \mathrm{E}[aa^T] &= \Sigma_a, \\ \mathrm{E}[(ab)(ab)^T] &= k_{ab}\Sigma_a, \\ \mathrm{E}[(ac)(ac)^T] &= k_{ac}\Sigma_a, \\ \mathrm{E}[(abc)_{(1)}(abc)_{(1)}^T] &= k_{abc}\Sigma_a, \end{aligned}$$

where k_{ab}, k_{ac} and k_{abc} are scalars. Here, $(abc)_{(1)}$ is the *matricization* of the array (abc), which converts the $m_1 \times m_2 \times m_3$ array into an $m_1 \times (m_2 m_3)$ matrix [Kolda and Bader, 2009]. To accommodate similar structure for the second and third factors, we propose the following prior covariance model for the main effects and interaction terms:

$$\begin{gathered} \mathrm{Cov}[a] = \Sigma_a \qquad \mathrm{Cov}[b] = \Sigma_b \qquad \mathrm{Cov}[c] = \Sigma_c \\ \mathrm{Cov}[\mathrm{vec}(ab)] = \Sigma_b \otimes \Sigma_a/\gamma_{ab} \qquad \mathrm{Cov}[\mathrm{vec}(bc)] = \Sigma_c \otimes \Sigma_b/\gamma_{bc} \qquad \mathrm{Cov}[\mathrm{vec}(ac)] = \Sigma_c \otimes \Sigma_a/\gamma_{ac} \\ \mathrm{Cov}[\mathrm{vec}(abc)] = \Sigma_c \otimes \Sigma_b \otimes \Sigma_a/\gamma_{abc}, \end{gathered}$$

where "$\otimes$" is the Kronecker product. The covariance matrices Σ_a, Σ_b and Σ_c represent the similarities between the levels of each of the three factors, while the scalars $\gamma_{ab}, \gamma_{ac}, \gamma_{bc}, \gamma_{abc}$ represent the relative (inverse) magnitudes of the interaction terms as compared to the main effects. Further specifying the priors on the ANOVA decomposition parameters as being mean-zero and Gaussian, the prior on a is then the multivariate normal distribution $N_{m_1}(0, \Sigma_a)$, and the prior on $\mathrm{vec}(ab)$ is $N_{m_1 m_2}(0, \Sigma_b \otimes \Sigma_a/\gamma_{ab})$. This latter distribution is sometimes referred to as a matrix normal distribution [Dawid, 1981]. Similarly, the prior on

vec(abc) is $N_{m_1 m_2 m_3}(0, \Sigma_c \otimes \Sigma_b \otimes \Sigma_a / \gamma_{abc})$, which has been referred to as an array normal distribution [Hoff, 2011].

In classical ANOVA decompositions, it is common to impose an identifiability constraint on the different effects. In a Bayesian analysis it is possible to place priors over identifiable sets of parameters, but this is cumbersome and not frequently done in practice [Gelman and Hill, 2007, Kruschke, 2010]. The priors we propose for the effects in the ANOVA decomposition in this chaper induce a prior over the cell means, which are identifiable. These priors have an intuitive interpretation and do not negatively affect the convergence of MCMC chains generated by the proposed procedure as can be seen in the Simulation and Application sections.

In most data analysis situations the similarities between the levels of a given factor and magnitudes of the interactions relative to the main effects will not be known in advance. We therefore consider a hierarchical prior so that Σ_a, Σ_b, Σ_c and the γ-parameters are estimated from the data. Specifically, we use independent inverse-Wishart prior distributions for each covariance matrix, e.g. $\Sigma_a \sim$ inverse-Wishart(η_{a0}, S_{a0}^{-1}) and gamma priors for the γ-parameters, e.g. $\gamma_{ab} \sim$ gamma$(\nu_{ab0}/2, \tau_{ab0}^2/2)$, where η_a, S_a, ν_{ab0} and τ_{ab0}^2 are hyperparameters to be specified (some default choices for these parameters are discussed at the end of this section). This hierarchical prior distribution can be viewed as an adaptive penalty, which allows for sharing of information across main effects and interaction terms. For example, estimates of the three-way interaction will be stabilized by the covariance matrix $\Sigma_c \otimes \Sigma_b \otimes \Sigma_a$, which in turn is influenced by similarities between levels of the factors that are consistent across the main effects, two-way and three-way interactions.

4.2.2 Posterior approximation

Due to the semi-conjugacy of the HA prior, posterior approximation can be obtained from a straightforward Gibbs sampling scheme. Under this scheme, iterative simulation of parameter values from the corresponding full conditional distributions generates a Markov chain having a stationary distribution equal to the target posterior distribution. For computational simplicity, we consider the case of a balanced dataset in which the sample size in

each cell is equal to some common value n, in which case the data can be expressed as an $m_1 \times m_2 \times m_3 \times n$ four-way array Y. A modification of the algorithm to accommodate unbalanced data is discussed in the next subsection.

Derivation of the full conditional distributions of the grand mean μ and the error variance σ^2 are completely standard: Under a $N(\mu_0, \tau_0^2)$ prior for μ, the corresponding full conditional distribution is $N(\mu_1, \tau_1^2)$, where $\tau_1^2 = (1/\tau_0^2 + nm_1m_2m_3/\sigma^2)^{-1}$ and $\mu_1 = \tau_1^2(\mu_0/\tau_0^2 + nm_1m_2m_3\bar{r}/\sigma^2)$, where $\bar{r} = \sum_{ijkl}(y_{ijkl} - [a_i + b_j + c_k + (ab)_{ij} + (ac)_{ik} + (bc)_{jk} + (abc)_{ikj}])/n$. Under an inverse-gamma$\big(\nu_0/2, \nu_0\sigma_0^2/2\big)$ prior distribution, the full conditional distribution of σ^2 is an inverse-gamma$\big(\nu_1/2, \nu_1\sigma_1^2/2\big)$ distribution, where $\nu_1 = \nu_0 + nm_1m_2m_3$, $\nu_1\sigma_1^2 = \nu_0\sigma_0^2 + \sum_{ijkl}\left(y_{ijkl} - \mu_{ijk}\right)^2$ and $\mu_{ijk} = \mu + a_i + b_j + c_k + (ab)_{ij} + (ac)_{ik} + (bc)_{jk} + (abc)_{ikj}$. Derivation of the full conditional distributions of parameters other than μ and σ^2 is straightforward, but slightly non-standard due to the use of matrix and array normal prior distributions for the interaction terms. In what follows, we compute the full conditional distributions for a few of these parameters. Full conditional distributions for the remaining parameters can be derived in an analogous fashion.

Full conditionals of a and (abc): To identify the full conditional distribution of the vector a of main effects for the first factor, let

$$\begin{aligned} r_{ijkl} &= y_{ijkl} - \Big(\mu + b_j + c_k + (ab)_{ij} + (ac)_{ik} + (bc)_{jk} + (abc)_{ijk}\Big) \\ &= a_i + \epsilon_{ijkl}, \end{aligned}$$

i.e., r_{ijkl} is the "residual" obtained by subtracting all effects other than a from from the data. Since $\{\epsilon_{ijkl}\} \sim$ i.i.d. normal$(0, \sigma^2)$, we have

$$p(Y|\theta, \sigma^2) \quad \propto_a \quad \exp\Big\{-\frac{m_2m_3n}{2\sigma^2}(a^Ta - 2a^T\bar{r})\Big\},$$

where $\bar{r} = (\bar{r}_1, \ldots, \bar{r}_{m_1})$ with $\bar{r}_i = \sum_{jkl} r_{ijkl}/(m_2m_3n)$, $\theta = \{a, b, c, (ab), (ac), (bc), (abc)\}$ and "$\propto_a$" means "proportional to as a function of a." Combining this with the $N_{m_1}(0, \Sigma_a)$

prior density for a, we have

$$p(a|Y,\theta_{-a},\sigma^2) \propto_a \exp\left(-\frac{m_2 m_3 n}{2\sigma^2}\left[a^T a - 2a^T\bar{r}\right] - \frac{1}{2}a^T\Sigma_a^{-1}a\right)$$

and so the full conditional distribution of a is multivariate normal with

$$\begin{aligned}\mathrm{Var}[a|Y,\theta_{-a},\sigma^2] &= \left(\Sigma_a^{-1} + Im_2m_3n/\sigma^2\right)^{-1}\\ \mathrm{E}[a|Y,\theta_{-a},\sigma^2] &= \left(\Sigma_a^{-1} + Im_2m_3n/\sigma^2\right)^{-1}\bar{r}\times(m_2m_3n/\sigma^2),\end{aligned}$$

where I is the $m_1 \times m_1$ identity matrix.

Derivation of the full conditional distributions for the interaction terms is similar. For example, to obtain the full conditional distribution of (abc) let r_{ijkl} be the residual obtained after subtracting all other components of θ from the data, so that $r_{ijkl} = (abc)_{ijk} + \epsilon_{ijkl}$. Let $\bar{r}$ be the three-way array of cell means of $\{r_{ijkl}\}$, so that $\bar{r}_{ijk} = \sum_l r_{ijkl}/n$. Combining the likelihood in terms of $\bar{r}$ with the $N_{m_1m_2m_3}(0, \Sigma_c \otimes \Sigma_b \otimes \Sigma_a/\gamma_{abc})$ prior density for $\mathrm{vec}(abc)$ gives

$$p\left((abc)\,|Y,\sigma^2,\Sigma_a,\Sigma_b,\Sigma_c,\gamma_{abc},\theta_{-(abc)}\right) \propto_{(abc)} \exp\left(-\frac{n}{2}\left[\mathrm{vec}\,(abc)^T\,\mathrm{vec}\,(abc) - 2\mathrm{vec}\,(abc)^T\,\mathrm{vec}(\bar{r})\right]\right)\times \exp\left(-\frac{1}{2}\mathrm{vec}\,(abc)^T\left(\Sigma_c\otimes\Sigma_b\otimes\Sigma_c/\gamma_{abc}\right)^{-1}\mathrm{vec}\,(abc)\right)$$

and so $\mathrm{vec}(abc)$ has a multivariate normal distribution with variance and mean given by

$$\begin{aligned}\mathrm{Var}[\mathrm{vec}(abc)|Y,\theta_{-(abc)},\sigma^2] &= \left(\left(\Sigma_c\otimes\Sigma_b\otimes\Sigma_a/\gamma_{abc}\right)^{-1} + In/\sigma^2\right)^{-1}\\ \mathrm{E}[\mathrm{vec}(abc)|Y,\theta_{-(abc)},\sigma^2] &= \left(\left(\Sigma_c\otimes\Sigma_b\otimes\Sigma_a/\gamma_{abc}\right)^{-1} + In/\sigma^2\right)^{-1}\mathrm{vec}(\bar{r})\times n/\sigma^2.\end{aligned}$$

Full conditional distributions for the remaining effects can be derived analogously.

Full conditional of Σ_a: The parameters in the ANOVA decomposition whose priors depend on Σ_a are a, (ab), (ac) and (abc). For example, the prior density of (ab) given Σ_a,

Σ_b and γ_{ab} can be written as

$$\begin{aligned} p((ab)|\Sigma_a,\Sigma_b,\gamma_{ab}) &= |2\pi\Sigma_b\otimes\Sigma_a/\gamma_{ab}|^{-1/2}\exp\left(-\mathrm{vec}(ab)^T[\Sigma_b\otimes\Sigma_a/\gamma_{ab}]^{-1}\mathrm{vec}(ab)/2\right) \\ &\propto_{\Sigma_a} |\Sigma_a|^{-m_2/2}\mathrm{etr}\left(-\Sigma_a^{-1}\gamma_{ab}(ab)^T\Sigma_b^{-1}(ab)/2\right) \\ &= |\Sigma_a|^{-m_2/2}\mathrm{etr}(-\Sigma_a^{-1}S_{ab}/2), \end{aligned}$$

where $S_{ab}=\gamma_{ab}(ab)^T\Sigma_b^{-1}(ab)$ and $\mathrm{etr}(A)=\exp\{\mathrm{trace}(A)\}$ for a square matrix A. Similarly, the priors for a, (ac) and (abc) are proportional to $|\Sigma_a|^{-d_i/2}\mathrm{etr}\left(-\Sigma_a^{-1}S_i/2\right)$ (as a function of Σ_a) for $i\in\{a,ac,abc\}$ where

$$\begin{aligned} S_a &= aa^T \\ S_{ac} &= \gamma_{ac}(ac)^T\Sigma_c^{-1}(ac) \\ S_{abc} &= \gamma_{abc}(abc)_{(1)}(\Sigma_c\otimes\Sigma_b)^{-1}(abc)_{(1)}, \end{aligned}$$

and $d_a=1$, $d_{ac}=m_3$ and $d_{abc}=m_2m_3$. The inverse-Wishart(η_{a0},S_{a0}^{-1}) prior density for Σ_a can be written in a similar fashion: it is proportional to $|\Sigma_a|^{-(\eta_{a0}+m_1+1)/2}\mathrm{etr}\left(-\Sigma_a^{-1}S_{a0}/2\right)$. Multiplying together the prior densities for a, (ab), (ac), (abc) and Σ_a and simplifying by the additivity of exponents and the linearity of the trace gives

$$p(\Sigma_a|\theta,\Sigma_b,\Sigma_c,\gamma)\propto|\Sigma_a|^{-(1+m_1+\eta_{a0}+1+m_2+m_3+m_2m_3)/2}\,\mathrm{etr}\left(-\Sigma_a^{-1}(S_{a0}+S_a+S_{ab}+S_{ac}+S_{abc})/2\right).$$

It follows that the full conditional distribution of Σ_a is inverse-Wishart$\left(\eta_{a1},S_{a1}^{-1}\right)$ where $\eta_{a1}=\eta_{a0}+(1+m_2+m_3+m_2m_3)$ and $S_{a1}=S_{a0}+S_a+S_{ab}+S_{ac}+S_{abc}$. The full conditional expectation of Σ_a is therefore $S_{a1}/\left(\eta_{a1}-m_1-1\right)$, which combines several estimates of the similarities among the levels of the first factor, based the main effects and the interactions.

Full conditional of γ_{abc}: The full conditional distribution of γ_{abc} depends only on the (abc) interaction term. The normal prior for (abc) can be written as

$$p((abc)|\Sigma_a,\Sigma_b,\Sigma_c,\gamma_{abc})\propto_{\gamma_{abc}}\gamma_{abc}^{m_1m_2m_3/2}\exp\{-\gamma_{abc}\mathrm{vec}(abc)^T[\Sigma_c\otimes\Sigma_b\otimes\Sigma_a]^{-1}\mathrm{vec}(abc)^T/2\}$$

Combining this density with a gamma$(\nu_{abc0}/2, \tau^2_{abc0}/2)$ prior density yields a full conditional for γ_{abc} that is gamma$\big(\nu_{abc1}/2, \tau^2_{abc1}/2\big)$, where

$$\begin{aligned}\nu_{abc1} &= \nu_{abc0} + m_1 m_2 m_3 \\ \tau^2_{abc1} &= \tau^2_{abc0} + \text{vec}(abc)^T [\Sigma_c \otimes \Sigma_b \otimes \Sigma_a]^{-1} \text{vec}(abc).\end{aligned}$$

4.2.3 Balancing unbalanced designs

For most survey data we expect the sample sizes $\{n_{ijk}\}$ to vary across combinations of factors. As a result, the full conditional distributions of the ANOVA decomposition parameters are more difficult to compute. For example, the conditional variance of the three-way interaction vec(abc) changes from $\left(\gamma_{abc} \left(\Sigma_c \otimes \Sigma_b \otimes \Sigma_a\right)^{-1} + In/\sigma^2\right)^{-1}$ in the balanced case to $\left(\gamma_{abc} \left(\Sigma_c \otimes \Sigma_b \otimes \Sigma_a\right)^{-1} + D/\sigma^2\right)^{-1}$ in the general case, where D is a diagonal matrix with diagonal elements vec$(\{n_{ijk}\})$. Even for moderate numbers of levels of the factors, the matrix inversions required to calculate the full conditional distributions in the unbalanced case can slow down the Markov chain considerably. As an alternative, we propose the following data augmentation procedure to "balance" an unbalanced design. Let $\bar{Y}^o$ be the three-way array of cell means based on the observed data, i.e. $\bar{y}^o_{ijk} = \sum y_{ijkl}/n_{ijk}$. Letting $n = \max\left(\{n_{ijk}\}\right)$, for each cell ijk with sample size $n_{ijk} < n$ and at each step of the Gibbs sampler, we impute a cell mean based on the "missing" $n - n_{ijk}$ observations as $\bar{y}^m_{ijk} \sim \text{normal}(\mu_{ijk}, \sigma^2/[n_{\max} - n_{ijk}])$, where μ_{ijk} is the population mean for cell ijk based on the current values of the ANOVA decomposition parameters. We then combine $\bar{y}^o_{ijk}$ and $\bar{y}^m_{ijk}$ to form the "full sample" cell mean $\bar{y}^f_{ijk} = (n_{ijk}\bar{y}^o_{ijk} + (n - n_{ijk})\bar{y}^m_{ijk})/n$. This array of cell means provides the sufficient statistics for a balanced dataset, for which the full conditional distributions derived above can be used.

4.2.4 Setting hyperparameters

In the absence of detailed prior information about the parameters, we suggest using a modified empirical Bayes approach to hyperparameter selection based on the maximum likelihood estimates (MLEs) of the error variance and mean parameters. Priors for μ and σ^2

can be set as unit information priors [Kass and Wasserman, 1995], whereby hyperparameters are chosen so that the prior means are near the MLEs but the prior variances are set to correspond roughly to only one observation's worth of information. For the covariance matrices Σ_a, Σ_b and Σ_c, recall that the prior for the main effect a of the first factor is $N_{m_1}(0, \Sigma_a)$. Based on this, we choose the prior for Σ_a to be inverse-Wishart(ν_{a0}, S_{a0}^{-1}) with $\nu_{a0} = m_1 + 2$ and $S_{a0} = \|\hat{a}\|^2 I_{m_1}/m_1$, where $\hat{a}$ is the MLE of a and $\|\hat{a}\|$ is the L_2 norm of $\hat{a}$. Under this prior, $\mathrm{E}[\mathrm{tr}(\Sigma_a)] = \|\hat{a}\|^2$, and so the scale of the prior matches the empirical estimates. Finally, the γ-parameters can be set analogously, using diffuse gamma priors but centered around values to match the magnitude of the OLS estimates of the interaction terms they correspond to, relative to the magnitude of the main effects. For example, in the next section we use a gamma$(\nu_{ab0}/2, \tau_{ab0}^2/2)$ prior for γ_{ab} in which $\nu_{ab0} = 1$ and $\tau_{ab0}^2 = \|\hat{a}\|^2\|\hat{b}\|^2/\|\hat{(ab)}\|^2$, where $\hat{a}$, $\hat{b}$ and $\hat{(ab)}$ are the OLS estimates.

The above procedure can be modified to accommodate an incomplete design, where not all the OLS estimates are available for a complete model. For example, in a two-way example, if exactly one cell is empty then the OLS estimates are available for all effect levels except for the two-way interaction for the missing cell. Abusing notation a bit, let $\|\hat{(ab)}\|$ be the L_2 norm of available OLS estimates for the two-way interaction. There are $m_1 m_2 - 1$ of these. Note, that this will likely underestimate $\|(ab)\|$ as it is missing the component contributed by the missing cell. To correct for this underestimate we propose the following modification for setting the hyperparameters: $\|\tilde{(ab)}\|^2 = \|\hat{(ab)}\|(m_1 m_2)/(m_1 m_2 - 1)$. The choice of τ_{ab0}^2 above becomes $\|\hat{a}\|^2\|\hat{b}\|^2/\|\tilde{(ab)}\|^2$.

4.3 Simulation study

This section presents the results of four simulation studies comparing the HA prior to several competing approaches. The first simulation study uses data generated from a means array that exhibits order consistent interactions. Estimates based on the HA prior outperform standard OLS estimates as well as estimates from a standard Bayesian approach that is similar to the one in Gelman [2005], and is also related to a grouped version of the lasso procedure [Yuan and Lin, 2006]. The second simulation study uses data from a means array that exhibits "order inconsistent" interactions, i.e. interactions without consistent

similarities in parameter values between levels of a factor. In this case the HA prior still outperforms the OLS and standard Bayes approaches, although not by as much as in the presence of order consistent interactions. In the third simulation we study the Bayes risk of the HA procedure when data is generated directly from the SB prior. Unlike the second simulation study, where interactions were "order inconsistent" but had potential similarities, in this case all effects were completely independent and so the oracle SB approach that imposes independence on the interaction effects outperforms HA, though not by much. The fourth simulation study uses data from a means array that has an exact additive decomposition, i.e. there are no interactions. The HA prior procedure again outperforms the standard Bayes and OLS approaches, although it does not do as well as OLS and Bayes oracle estimators that assume the correct additive model.

We ran our analysis on a 16 node cluster with 128 cores, with a total of 128GB RAM. The additive Bayes approach is significantly faster than the other two Bayesian procedures since it contains the fewest parameters. The other two procedures are comparable, but with SB being somewhat faster than HA on average. However, the relative speed of SB decreased with sample size, with HA running an estimated 21%, 17%, 10% and 7% slower than SB for sample sizes 400, 1000, 5000 and 10000, respectively.

4.3.1 Data with order consistent interactions

The data in this simulation study is generated from a model where the means array exhibits order consistent interactions. The dimensions of the means array M were chosen to be $m_1 \times m_2 \times m_3 = 15 \times 7 \times 3$, which could represent, for example, the number of categories we might have for age, education level and political affiliation in a cross-classified survey dataset. The means array was generated from a cubic function of three variables that was then binned. Figure 4.3 plots the mean array across the third factor, demonstrating the nonadditivity present in M. By decomposing M into the main, two-way and three-way effects in the same manner as described in Section 4.2, we can summarize the nonadditivity of M through the magnitudes of the different sums of squares. The magnitudes of the main effects, given by the squared L_2 norm of the effects, $\|a\|^2/m_1, \|b\|^2/m_2,$ and $\|c\|^2/m_3$ are

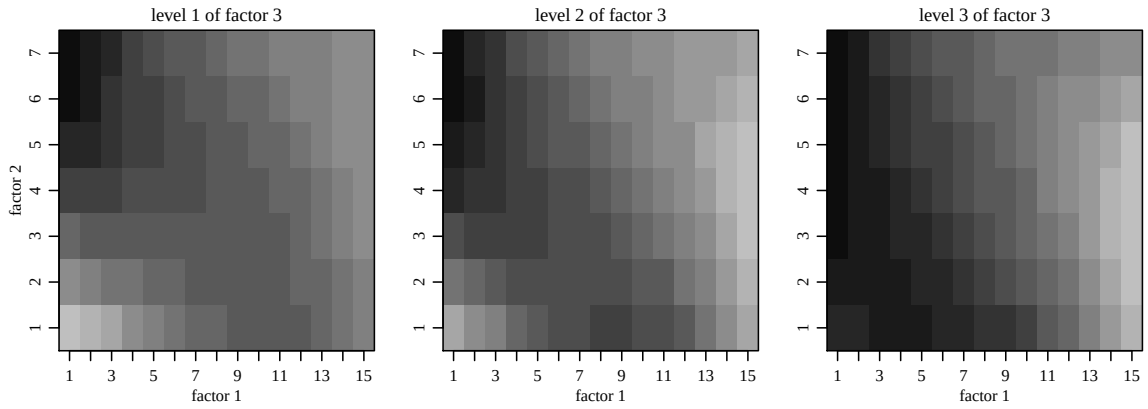

Figure 4.3: The means array M across levels of the third factor.

$5.267, 0.012, 0.004$ respectively. The magnitudes of the two-way interactions $\|ab\|^2/(m_1m_2)$, $\|ac\|^2/(m_1m_3)$, and $\|bc\|^2/(m_2m_3)$ are $1.365, 1.312,$ and 0.384, and the magnitude of the three-way interaction $\|abc\|^2/(m_1m_2m_3)$ is 0.474. For each sample size {400, 1000, 5000, 10000}, we simulated 50 datasets using the mean array M and independent standard normal errors. In order to make a comparison to OLS possible, we first allocated one observation to each cell of the means array. We then distributed the remaining observations uniformly at random (with replacement) among the cells of the means array. This leads to a complete but potentially unbalanced design. The average number of observations per cell under the sample sizes {400,1000,5000,10000} was $\{1.3, 3.2, 15.9, 31.7\}$.

For each simulated dataset we obtained estimates under the HA prior (using the hyperparameter specifications described in Section 2.4), as well as ordinary least squares estimates (OLS) and posterior estimates under a standard Bayesian prior (SB). The SB approach is essentially a simplified version of the HA prior in which the parameter values are conditionally independent given the hyperparameters: $\{a_i\} \sim$ i.i.d. $N\left(0, \sigma_a^2\right)$, $\{(ab)_{ij}\} \sim$ i.i.d. $N\left(0, \sigma_{ab}^2\right)$ and $\{(abc)_{ijk}\} \sim$ i.i.d. $\left(0, \sigma_{abc}^2\right)$ and similarly for all other main effects and interactions. To facilitate comparison to the HA prior, the hyperpriors for these σ^2-parameters are the same as the hyperpriors for the inverses of the γ-parameters in the HA approach. As a result, this standard Bayes prior can be seen as the limit of a sequence of HA priors where the inverse-Wishart prior distributions for the Σ-matrices converge to point-masses

on the identity matrices of the appropriate dimension.

For each simulated dataset, the Gibbs sampler described in Section 4.2 was run for 11,000 iterations, the first 1,000 of which were dropped to allow for convergence to the stationary distribution. Parameter values were saved every 10th scan, resulting in 1,000 Monte Carlo samples per simulation. Starting values for all the mean effects were set to zero and all variances set to identity matrices of the proper dimensions. We examined the convergence and autocorrelation of the marginal samples of the error variance σ^2 using Geweke's z-test and the effective sample size. The minimum effective sample size across all simulations was 233 out of the 1000 recorded scans, and the average effective sample size was 895. Geweke's z-statistic was less than 2 in absolute value in 93, 93, 97, and 95 percent of the Markov chains for the four sample sizes (with the percentages being identical for both Bayesian methods). While the cases in which $|z| > 2$ were not extensively examined, it is assumed that running the chain longer would have yielded improved estimation.

For each simulated data set, the posterior mean estimates $\hat{M}_{\text{HA}}$ and $\hat{M}_{\text{SB}}$ were obtained by averaging their values across the 1,000 saved iterations of the Gibbs sampler. The average squared error (ASE) of estimation was calculated as $\text{ASE}(\hat{M}) = \|\hat{M} - M\|^2/(m_1 m_2 m_3)$ where M is the means array that generated the data. These values were then compared across the three approaches. The left panel of Figure 4.4 demonstrates that the SB estimator

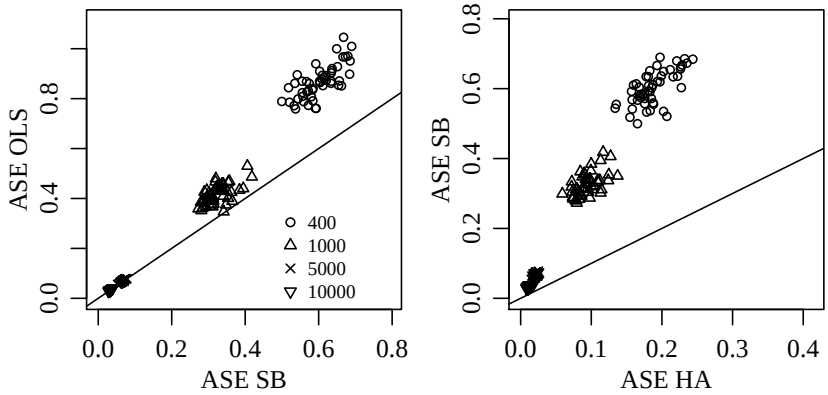

Figure 4.4: Comparison of ASE for different estimation methods when the true means array exhibits order consistent interactions.

provided a reduction in ASE when compared to the OLS estimator for all data sets with sample sizes 400 and 1000, 96% of the data sets with sample size 5000 and 90% of data sets with sample size 10000. The second panel demonstrates that the HA estimator provides a substantial further reduction in ASE for all data sets. As we would expect, the reduction in ASE is dependent on the sample size and decreases as the sample size increases.

These results are not surprising: By estimating the variances $\sigma_a^2, \sigma_{ab}^2$, etc. from the data, the SB approach provides adaptive shrinkage and so we expect these SB estimates to outperform the OLS estimates in terms of ASE. However, the SB approach does not use information on the similarity among the levels of an effect, and so its estimation of higher order interactions relies on the limited information available directly in the corresponding sufficient statistics. As such, we expect the SB estimates to perform less well than the HA estimates, which are able to borrow information from well-estimated main effects and low-order interactions to assist in the estimation of higher-order terms for which data information is limited.

This behavior is further illustrated in Figure 4.5 that provides an ASE comparison for the effects in the decomposition of the means array. To produce these plots we decomposed each estimated means array and considered the ASE for each effect when compared to the decomposition of the true means array. It is immediate that the gains in ASE are primarily from improved estimation of the higher order interaction terms. The top row of Figure 4.5 demonstrates that the SB estimator performs at least as well as the OLS estimator in terms of ASE for the main effect a, and provides a significant reduction in ASE for two- and three-way interactions. The reduction in ASE for the higher order terms is due to the shrinkage provided by SB. The second row of Figure 4.5 demonstrates that the HA estimator provides a moderate reduction in ASE for the main effect a and a substantial further reduction in ASE for the higher order terms. This is exactly the behavior we expect as the HA procedure is able to borrow information from lower order terms in order to further shrink hight order interactions. We have also evaluated the width and coverage of nominal 95% confidence intervals for the cell means. The results for HA and SB are presented in Table **??**. The confidence intervals for the entries in the means array were smaller for the HA procedure than for SB, while the coverage was approximately 95% for both.

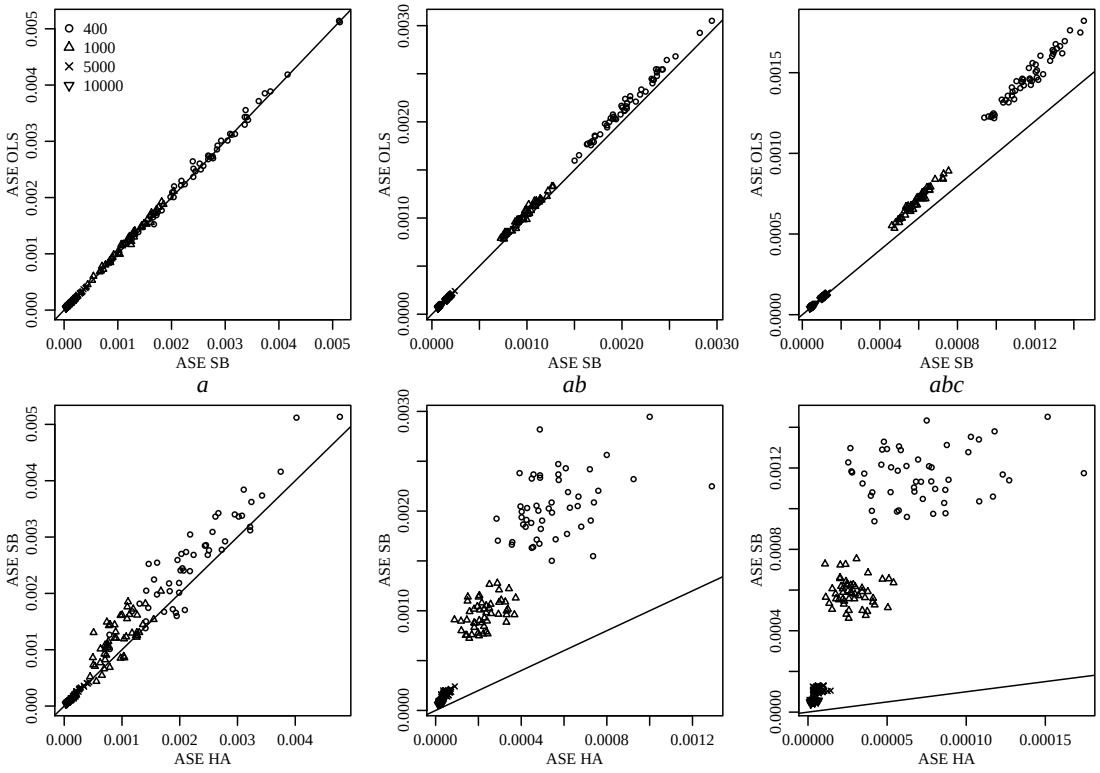

Figure 4.5: ASE comparisons for the main effect, a two-way interaction and a three-way interaction that involve a are in the three columns respectively. The first row compares ASE between SB and OLS and the second row compares ASE between HA and SB.

	Coverage		Width	
OBS	HA	SB	HA	SB
400	0.94	0.93	1.55	3.18
1000	0.93	0.95	1.04	2.32
5000	0.94	0.94	0.49	1.00
10000	0.95	0.95	0.36	0.70

Table 4.3: Actual coverage and interval widths of 95% nominal confidence intervals for the cell means as estimated by HA and SB when order consistent interactions are present.

Recall that the parameters in the mean array M were generated by binning a third-degree polynomial, and were not generated from array normal distributions, i.e. the HA prior is "incorrect" as a model for M. Even so, the HA prior is able to capture the similarities

between adjacent factor levels, resulting in improved estimation. However, we note that not all of the improvement in ASE achieved by the HA prior should be attributed to the identification of order-consistent interactions. The simulation study that follows suggests some of the performance of the HA prior is due to additional parameter shrinkage provided by the inverse-Wishart distributions on the Σ-matrices.

4.3.2 Data with order inconsistent interactions

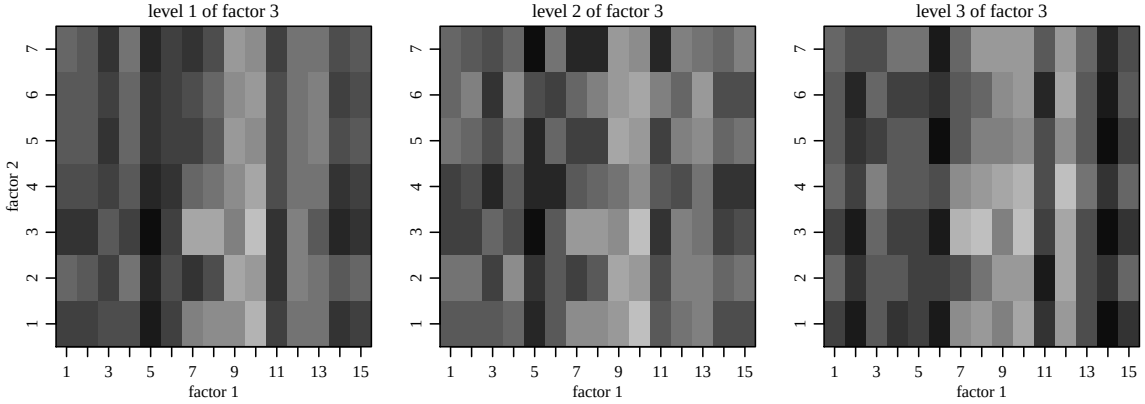

Figure 4.6: The means array M for the second simulation study, across levels of the third factor.

In this subsection we evaluate the HA approach for populations which exhibit interactions that are order inconsistent. The means array M is constructed by taking the means array from Section 4.3.1, decomposing it into main effects, two- and three-way interactions, permuting the levels of each factor within each effect, and reconstructing a means array. That is, if $\{a_i : i = 1, \ldots, m_1\}$ is the collection of parameters for the first main effect and $\left\{(ab)_{ij} : i = 1 \ldots, m_1, j = 1, \ldots, m_2\right\}$ is the collection of parameters for the two way interaction between the first and second factors in Section 4.3.1 then $\left\{a_{\pi_1(i)}\right\}$ and $\left\{(ab)_{\pi_2(i)\pi_3(j)}\right\}$ are the main effect and two-way interaction parameters for the means array in this section, where π_1, π_2 and π_3 are independent permutations. The remaining effects were permuted analogously. Due to this construction, the magnitudes of the main effects, two- and three-way interactions remain the same, but the process becomes less "smooth," as can be seen

in Figure 4.6.

Again, 50 data sets were generated for each sample size, and estimates $\hat{M}_{\text{HA}}$, $\hat{M}_{\text{SB}}$ and $\hat{M}_{\text{OLS}}$ were obtained for each data set, where the Bayesian estimates were obtained using the same MCMC approximation procedure as in the previous subsection. Figure 4.7 compares ASE across the different approaches. The left panel of Figure 4.7, as with the left panel of Figure 4.4, demonstrates that the SB estimator provides a reduction in ASE when compared to the OLS estimator. As expected, since neither of these approaches take advantage of the structure of the order consistent interactions, this plot is nearly identical to the corresponding plot in Figure 4.4.

The second panel demonstrates that the HA estimator provides a further reduction in ASE for all data sets, although this reduction is less substantial than in the presence of order consistent interactions. The lower ASE of the HA estimates may be initially surprising, as there are no order consistent interactions for the HA prior take advantage of. We conjecture that the lower ASE is due to the additional shrinkage on the parameter estimates that the inverse-Wishart priors on the Σ-parameters provide. For example, under both the SB and HA priors we have $\text{Cov}[\text{vec}(ab)] = \Sigma_b \otimes \Sigma_a / \gamma_{ab}$, but under the former the covariance matrices are set to the identity whereas under the latter they have inverse-Wishart distributions.

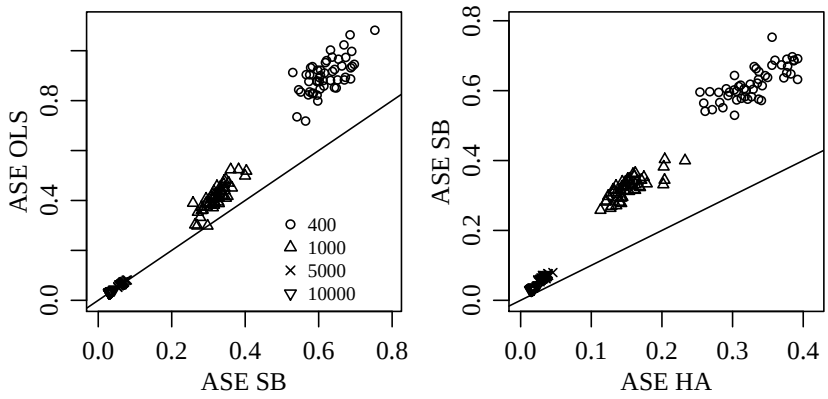

Figure 4.7: Comparison of ASE for different estimation methods when the true means array exhibits order inconsistent interactions that have the same magnitude as the order consistent interactions of Section 4.3.1.

As with the previous simulation, we evaluated the width and coverage of nominal 95%

confidence intervals for the cell means. The results for HA and SB are presented in Table **??**. As in the previous simulation, the coverage for both procedures is approximately 95%. The confidence intervals are wider for SB than for HA, but the differences between the two procedures are much smaller in this simulation as compared to the previous one.

	Coverage		Width	
OBS	HA	SB	HA	SB
400	0.95	0.94	2.26	2.98
1000	0.95	0.95	1.56	2.15
5000	0.96	0.95	0.73	0.98
10000	0.96	0.94	0.53	0.69

Table 4.4: Actual coverage and interval widths of 95% nominal confidence intervals for the cell means as estimated by HA and SB when order inconsistent interactions are present.

4.3.3 Data with order inconsistent interactions: Bayes risk

The surprising outcome of the previous section requires further study of the behavior of the HA approach when order inconsistent interactions are present. To get a more complete picture of this behavior, we evaluate the Bayes risk of the procedure when data is generated directly from the SB prior. We construct 200 means arrays $M_1, \ldots, M_{200}$ of the same dimensions as in the previous subsections using the following procedure:

1. Generate $\gamma_a, \gamma_b, \gamma_c, \gamma_{ab}, \gamma_{ac}, \gamma_{bc}, \gamma_{abc} \overset{\text{iid}}{\sim} \text{gamma}\left(\nu/2, \tau^2/2\right)$ with shape paramter $\nu = 4$ and rate parameter $\tau^2 = 2$. These are the precision components for the 3 main effects, 3 two-way interactions, and 1 three-way interaction respectively

2. Generate effect levels as follows: $\{a_1, \ldots, a_{15}\} \sim N\left(0, I/\gamma_a\right)$, $\{ab_{1,1}, \ldots, ab_{15,7}\} \sim N\left(0, I/\gamma_{ab}\right)$, and similarly for the remaining 5 effects.

3. Combine the effects from (2) into a means array M_i according to Equation (4.1).

For each sample size {400,1000,5000,10000} we generated 50 datasets, each using a unique means array M_i, in the same manner as in the previous two simulation studies. We obtained

estimates $\widehat{M}_{i\mathrm{HA}}$, $\widehat{M}_{i\mathrm{SB}}$ and $\widehat{M}_{i\mathrm{OLS}}$ for each data set, where the Bayesian estimates were obtained using the same MCMC procedure as in the previous two subsections.

ASE represents the posterior quadratic loss of an estimation procedure for a particular dataset, and so by varying the true means array M_i between simulated datasets, we can estimate the Bayes risk of an estimation procedure by taking the average of ASE across simulated datasets. The Bayes risk for the SB procedure is guaranteed to be smaller than that for OLS and HA for all sample sizes and so we report the results of the simulation study as ratios of estimated Bayes risk for SB to the estimated Bayes risk of the other procedures in Table 4.5. For example, the first entry in the top row of Table 4.5 states that the Bayes risk for SB is 41% lower than the Bayes risk for the OLS procedure for a sample size of 400. As is expected, the difference in Bayes risk shrinks with increasing sample size for both OLS and HA. The results of this simulation study suggest that even for moderately sized datasets, the relative risk of using the HA procedure when compared to SB is rather small even when all effects are completely independent. Additionally, the posterior estimates of all of the effects in the decomposition of the means array had similar variances under both SB and HA priors. This suggests that using the HA procedure is not detrimental even when the "order consistency" of the interactions cannot be verified.

Sample size	400	1000	5000	10000
OLS	0.59	0.69	0.93	0.97
HA	0.78	0.91	0.97	0.98

Table 4.5: Ratio of estimated Bayes risk for SB to OLS and HA by sample size.

4.3.4 Data without interactions

In this subsection we evaluate the HA approach for populations in which interactions are not present. The data in this simulation is generated from a model where the means array M is exactly additive, and was constructed by binning a linear function of three variables. As in the previous simulations, M is of dimension $m_1 \times m_2 \times m_3 = 15 \times 7 \times 3$. The magnitudes of the three main effects are $\|a\|^2/m_1 = 3.0$, $\|b\|^2/m_2 = 1.3$ and $\|c\|^2/m_3 = 0.3$ while

all interactions are exactly zero. In addition to the SB and OLS estimators, we compare the HA approach to two "oracle" estimators: the additive model least squares estimator (AOLS) and Bayes estimator under the additive model (ASB). The prior used by the ASB approach is the same as the SB prior, but does not include terms other than main effects in the model.

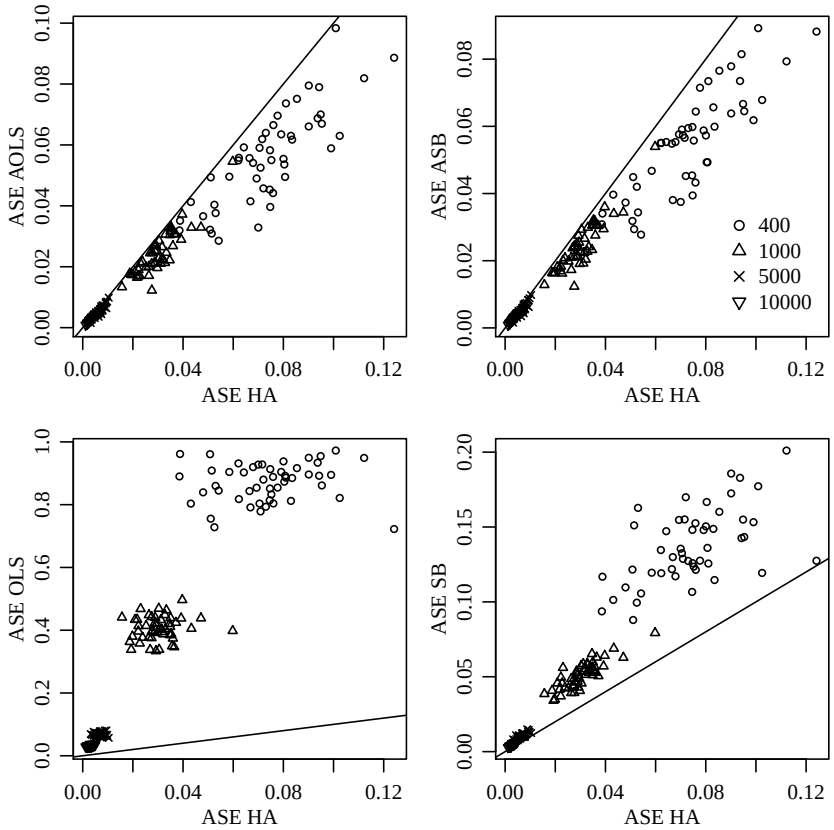

Figure 4.8: Comparison of ASE for different estimation methods when the true means array is additive.

As before, 50 data sets were generated for each sample size, and estimates $\hat{M}_{\text{HA}}$, $\hat{M}_{\text{SB}}$, $\hat{M}_{\text{OLS}}$, $\hat{M}_{\text{ASB}}$ and $\hat{M}_{\text{AOLS}}$ were obtained for each data set, where the Bayesian estimates were obtained using the same MCMC approximation procedure as in the previous two subsection. Some results are shown in Figure 4.8, which compares ASE across the different approaches. In the top row of Figure 4.8 we see that the performance of the HA estimates

is comparable to but not as good as the the oracle least squares and Bayesian estimates in terms of ASE. Specifically, the ASE for the HA estimates is 24.2, 18.6, 20.1 and 17.4 percent higher than for the AOLS estimates for data sets with sample sizes 400, 1000, 5000 and 10000 respectively. Similarly, the ASE for the HA estimates is 25, 19.7, 20.3 and 17.8 percent higher than for the ASB estimates for data sets with sample sizes 400, 1000, 5000 and 10000 respectively. However, the bottom row of Figure 4.8 shows that the HA prior is superior to the other non-oracle OLS and SB approaches that attempt to estimate the interaction terms.

These results, together with those of the last two subsections, suggest that the HA approach provides a competitive method for fitting means arrays in the presence or absence of interactions. When order consistent interactions are present, the HA approach is able to make use of the similarities across levels of the factors, thereby outperforming approaches that cannot adapt to such patterns. Additionally, the HA approach does not appear to suffer when interactions are not order consistent. Finally, in the absence of interactions altogether, the HA approach adapts well, providing estimates similar to those that assume the correct additive model.

4.4 Analysis of carbohydrate intake

In this section we estimate average carbohydrate, sugar and fiber intake by education, ethnicity and age using the HA procedure described in Section 4.2. Our estimates are based on data from 2134 males from the US population, obtained from the 2007-2008 NHANES survey. Nutrient intake is self reported on two non-consecutive days. Each day's data concerns food and beverage intake from the preceding 24 hour period only, and is calculated using the USDA's Food and Nutrient Database for Dietary Studies 4.1 [USDA, 2010]. All intake was measured in grams, and we average the intake over the two days to yield a single measurement per individual. When intake information is only available for one day, we treat that as the observation (we do not perform any reweighing to account for this partial information). We are interested in relating the intake data to the following demographic variables:

- Age: $(31-40), (41-50), (51-60), (61-70), (71-80)$.
- Education: Primary (P), Secondary (S), High School diploma (HD), Associates degree (AD), Bachelors degree (BD).
- Ethnicity: Mexican (Hispanic), other Hispanic, white (not Hispanic) and black (not Hispanic).

Sample sizes for age-education-ethnicity combination were presented in Table 4.1 in Section 1. Of the 2234 male respondents within the above demographic groups, 100 were missing their nutrient intake information for both days, with similar rates of missingness across the demographic variables and were excluded from the analysis. For the 2134 individuals included in the analysis, 291 were missing nutrient intake information one of the two days. For those individuals, the available day's information was used as their nutrient intake, while for the remaining 1843 individuals an average over the two days was used.

The data on the original scale are somewhat skewed and show heteroscedasticity across the demographic variables. Since different variances across groups can lead to bias in the sums of squares, making F-tests for interactions anti-conservative [Miller and Brown, 1997], stabilizing the variance is desirable. Figure 4.9 provides two-way scatterplots of the response variables after applying a quarter power transformation to each variable, which we found stabilized the variances across the groups better than either a log or square-root transformation. Additionally, following the quarter power transformation, we centered and scaled each response variable to have mean zero and variance one.

4.4.1 MANOVA model and parameter estimation

As presented in Table 4.2 of Section 1, F-tests indicate evidence for the presence of interactions in the array of population cell means. However, 12% of all age-education-ethnicity categories have sample sizes less than 5, and so we are concerned with overfitting of the OLS estimates. As an alternative, we extend the HA methodology described in Section 4.2 to accommodate a MANOVA model. Our MANOVA model has the same form as the ANOVA model given by Equation (4.1), except that each effect listed there is a three-dimensional

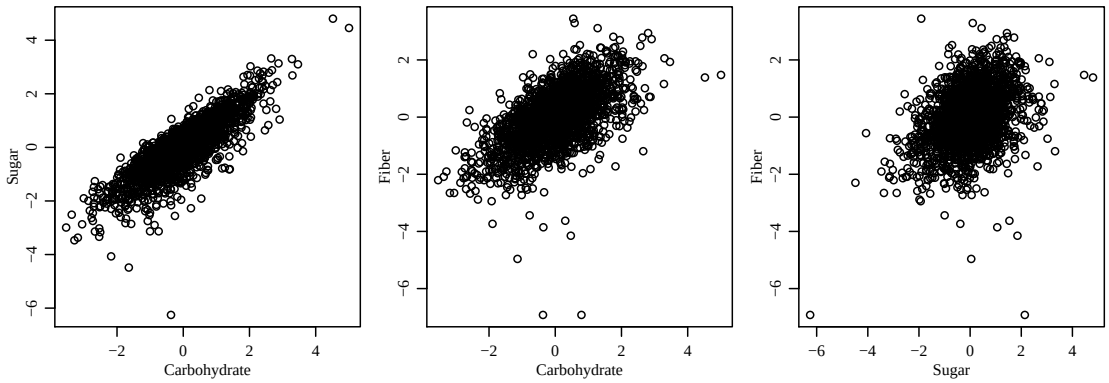

Figure 4.9: Two-way plots of the transformed data.

vector corresponding to the separate effects for each of the three response variables. Additionally, the error terms now have a multivariate normal distribution with zero-mean and unknown covariance matrix Σ_y.

We extend the hierarchical array prior discussed above to accommodate the p-variate MANOVA model as follows: Our prior for the $m_1 \times p$ matrix a of main effects for the first factor is $\text{vec}(a) \sim N_{m_1 p}(0, I \otimes \Sigma_a)$ where Σ_a is as before. Our prior for the $m_1 \times m_2 \times p$ array (ab) of two-way interaction terms is given by $\text{vec}(ab) \sim N_{m_1 m_2 p}(0, \Gamma_{ab}^{-1} \otimes \Sigma_b \otimes \Sigma_a)$. Here, Γ_{ab} is a $p \times p$ diagonal matrix whose terms determine the scale of the two-way interactions for each of the p response variables. If we consider only the first response, then $(\Gamma_{ab})_{11}$ is exactly the γ_{ab} scalar described in the ANOVA setup. Similarly, our prior for the four-way array (abc) of three-way interaction terms is $\text{vec}(abc) \sim N_{m_1 m_2 m_3 p}(0, \Gamma_{abc}^{-1} \otimes \Sigma_c \otimes \Sigma_b \otimes \Sigma_a)$. Priors for other main effects and interaction terms are defined similarly. The hyperpriors for each diagonal entry of Γ are independent gamma distributions, chosen as in Section 2.4 so that the prior magnitude of the effects for each response is centered around the sum of squares of the effect from the OLS decomposition.

An alternative prior would be to include a covariance matrix representing similarities of effects across the three variables. This would be achieved by replacing $I \otimes \Sigma_a$ in the prior for a with $\Sigma_p \otimes \Sigma_a$, Γ_{ab}^{-1} with $\Sigma_p \Gamma_{ab}^{-1}$ in the prior for ab, and so on. Such a covariance term might be approriate for data in which marginal correlations between the p response

variables were driven by similarities in the cell means, rather than by within-cell correlations. In such a case we would expect, for example, that if a_1, the main effects for variable 1, were positively correlated with a_2, the main effects for variable 2, then b_1 and b_2 would be positively correlated, as would c_1 and c_2, as well as any other pair of effects in the decompositions of variables 1 and 2. However, such consistency does not appear in our NHANES data: For example, considering correlations between the ANOVA decomposition parameters for sugar and carbohydrates, we observe positive correlations for the main effects of age and education and negative correlations for the interaction terms age×ethnicity and age×ethnicity×education. These observations support the choice of $\Sigma_p = I$ in the prior for the analysis of these data, although estimating Σ_p might be warranted for other datasets.

A Gibbs sampling scheme similar to the one outlined in Section 4.2 was iterated 200,000 times with parameter values saved every 10 scans, resulting in 20,000 simulated values of the means array M and the covariance matrices $\{\Sigma_{\text{eth}}, \Sigma_{\text{age}}, \Sigma_{\text{edu}}\}$ for posterior analysis. Mixing of the Markov chain for M was good: Figure 4.10 shows MCMC samples of 4 out of 300 entries of M (chosen so that their trace plots were visually distinct). The autocorrelation across the saved scans was low, with the lag-10 autocorrelation for the thinned chain less than 0.14 in absolute value for each element of M (97.3% of entries have lag-10 autocorrelation less than 0.07 in absolute value) and effective sample sizes between 1929 and 13520. The mixing for the elements of the covariance matrices $\Sigma_{\text{eth}}, \Sigma_{\text{age}}, \Sigma_{\text{edu}}$ is not as good as that of the means array M: The maximum absolute value of lag-10 autocorrelation of the saved scans for the three rescaled covariance matrices is 0.18, 0.12, and 0.19 respectively. The effective sample sizes for the elements of the covariance matrices are at least 1684.

4.4.2 Posterior inference on M and Σs

We obtain a Monte Carlo approximation to $\hat{M} = E[M|Y]$ by averaging over the saved scans of the Gibbs sampler. Figure 4.11 provides information on the shrinkage and regularization of the estimates due to the HA procedure, as compared to OLS. The first panel plots the difference between the OLS and Bayes estimates of the cell means versus cell-specific

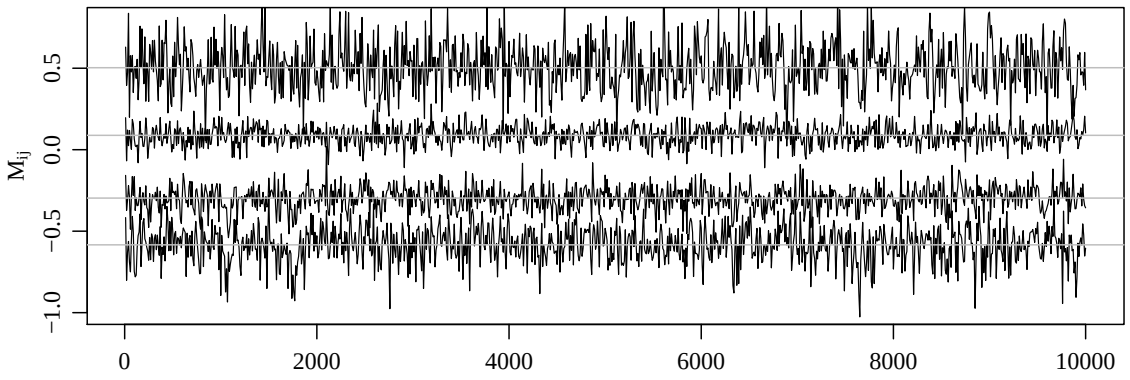

Figure 4.10: MCMC samples of 4 out of 300 entries of the means array M.

sample sizes. For small sample sizes, the Bayes estimate for a given cell is affected by the data from related cells, and can generally be quite different from the OLS estimate (the cell sample mean). For cells with large sample sizes the difference between the two estimates is generally small. The second panel of the figure plots the OLS estimates of the cell means for carbohydrate intake of black survey participants across age and education levels. Note that there appears to be a general trend of decreasing intake with increasing age and education level, although the OLS estimates themselves are not consistently ordered in this way. In contrast, these trends are much more apparent in the Bayes estimates plotted in the third panel. The HA prior allows the parameter estimates to be close to additive, while not enforcing strict additivity in this situation where we have evidence of non-additivity via the F-tests. The smoothing provided by the HA prior is attributed to its ability to share information across levels of an effect and across interactions. When more levels are present for a particular effect, the smoothing of the HA prior closely resembles the behavior one would expect from an unbinned continuous effect. On the other hand, OLS will continue to model each cell-specific mean separately, ignoring the similarities among levels and failing to recognize the continuous nature of the effect. The third panel of the figure was also more consistently ordered than a similar analysis performed with the SB prior, suggesting that the added shrinkage due to the inverse-Wishart priors and the ability to share information across effect levels leads to more realistic behavior of the estimates.

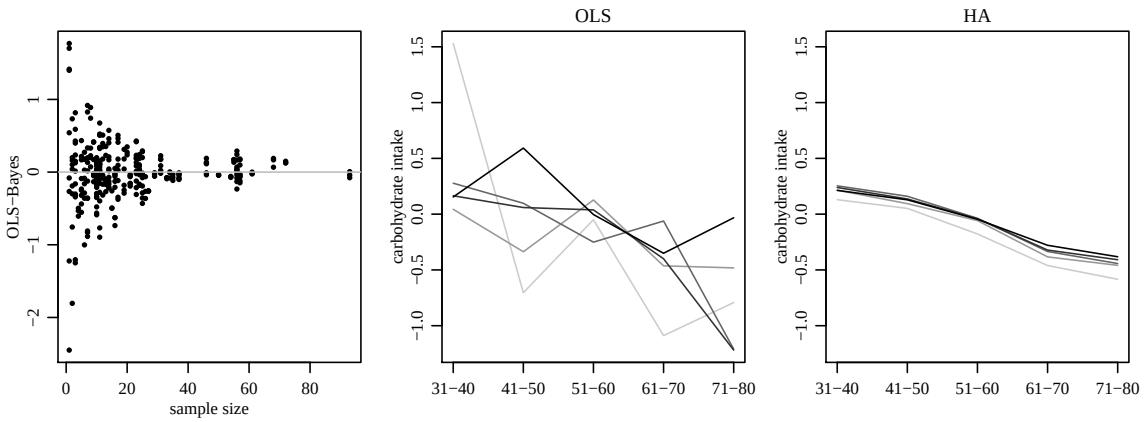

Figure 4.11: Shrinkage and regularization plots. The first panel plots the difference between the OLS and HA estimates of a cell-mean against the cell-specific sample sizes. The second a third panels plot estimated cell-means for black survey participants across age and education levels, where lighter shades represent higher levels of education.

The range of cell means for the standardized effects is -0.58 to 0.4 for carbohydrates, -0.38 to 0.38 for sugar and -1 to 0.51 for fiber. The average standard errors for the cell means for the three responses are 0.08, 0.09 and 0.13 respectively. When fitting the data with the SB prior (analysis not included here), the average standard errors for the cell means were substantially larger: 0.12, 0.13 and 0.15 for the three responses, respectively. The first row of Figure 4.12 provides the estimates of the main effects from the HA procedure. The second row of Figure 4.12 summarizes covariance matrices $\{\Sigma_{\text{eth}}, \Sigma_{\text{age}}, \Sigma_{\text{edu}}\}$ via the posterior mean estimates of the correlation matrices $\{C_{d,ij}\} = \left\{\Sigma_{d,ij}/\sqrt{\Sigma_{d,ii}\Sigma_{d,jj}}\right\}$ for $d \in \{\text{eth}, \text{age}, \text{edu}\}$. In this figure, the diagonal elements are all 1, and darker colors represent a greater departure from one. The range of the estimated correlations was -0.34 to 0.42 for age categories, -0.30 to 0.35 for ethnic groups, and -0.17 to 0.38 for educational categories. For the two ordered categorical variables, age and education, we see that closer categories are generally more positively correlated than ones that are further apart. While the ethnicity variable is not ordered, its correlation matrix informs us of which categories are more similar in terms of these response variables. The middle panel of the second row of Figure 4.12 confirms the order-consistent interactions we observed in Figure 4.2: Mexican survey participants are more similar to Hispanic participants in terms of carbohydrate intake patterns than to

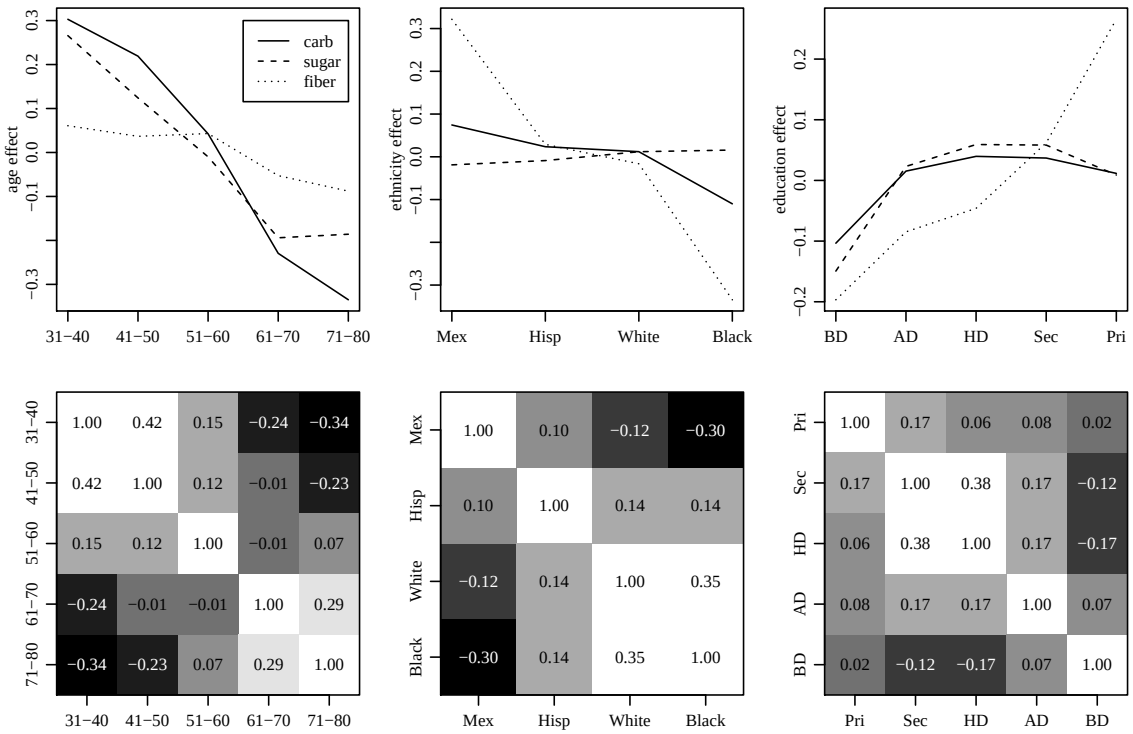

Figure 4.12: Plots of main effects and interaction correlations for the three outcome variables (carbohydrates, sugar and fiber). The first row of plots gives HA estimates of the main effects for each factor. The second row of plots gives correlations of effects between levels of each factor, with white representing 1 and darker colors representing a greater departure from one.

white or black participants.

For fiber intake, the top row of Figure 4.13 gives age by education interaction plots for each level of ethnicity, using cell mean estimates obtained from the HA procedure. Comparing these plots to the analogous plots for the OLS estimates presented in Figure 4.1, we see that the smoother HA estimates allow for a more interpretable description of the three-way interaction. Recall that a three-way interaction can be described as the variability of a two-way interaction across levels of a third factor. Based on the plots, the two-way age by education interactions for the Mexican and Black groups seem quite small. In contrast, the White and Other Hispanic groups appear to have interactions that can be described as heterogeneity in the education effect across levels of age. For both of these groups, this

heterogeneity is ordered by age: For the Other Hispanic group, the education effects seem similar for the three youngest age groups. For the White group, the education effects seem similar for the two youngest age groups.

This similarity in education effects for similar levels of age is more apparent in these HA estimates than in the corresponding parameter estimates from the SB procedure, presented in the second row of Figure 4.13, particularly for the White ethnicity. In contrast to the SB approach, the HA procedure was able to recognize the similarity of parameters corresponding to adjacent age levels, and use this information to assist with estimation. Our expectations that age effects are likely to be smooth, as well as the good performance of the HA procedure in the simulation study of the previous section, give us confidence that the HA procedure is providing more realistic and interpretable cell mean estimates than either the OLS or SB approaches.

4.5 Discussion

This chapter has presented a novel hierarchical Bayes method for parameter estimation of cross-classified data under ANOVA and MANOVA models. Unlike least-squares estimation, a Bayesian approach provides for regularized estimates of the potentially large number of parameters in a MANOVA model. Unlike the non-hierarchical Bayesian approach, the hierarchical approach provides a data-driven method of regularization, and unlike the standard hierarchical Bayes, the hierarchical array prior can identify similarities among categories and share this information across interaction effects to assist in the estimation of higher-order terms for which data information is limited. In a simulation study the HA approach was able to detect interactions when they were present, and to estimate the means array better than a full least squares or standard Bayesian approaches (in terms of mean squared error). When the true means array was completely additive, the HA prior was able to adapt to this smaller model better than the other full model estimation approaches under consideration.

An immediate extension to our approach modifies the priors on the covariance matrices to incorporate known structure. For example, in the case of observations for different time periods, an autoregressive covariance model might be desirable. In the simplest case of an AR(1) model, Berger and Yang [1994] suggest the use of a reference prior $\pi_R(\rho)$

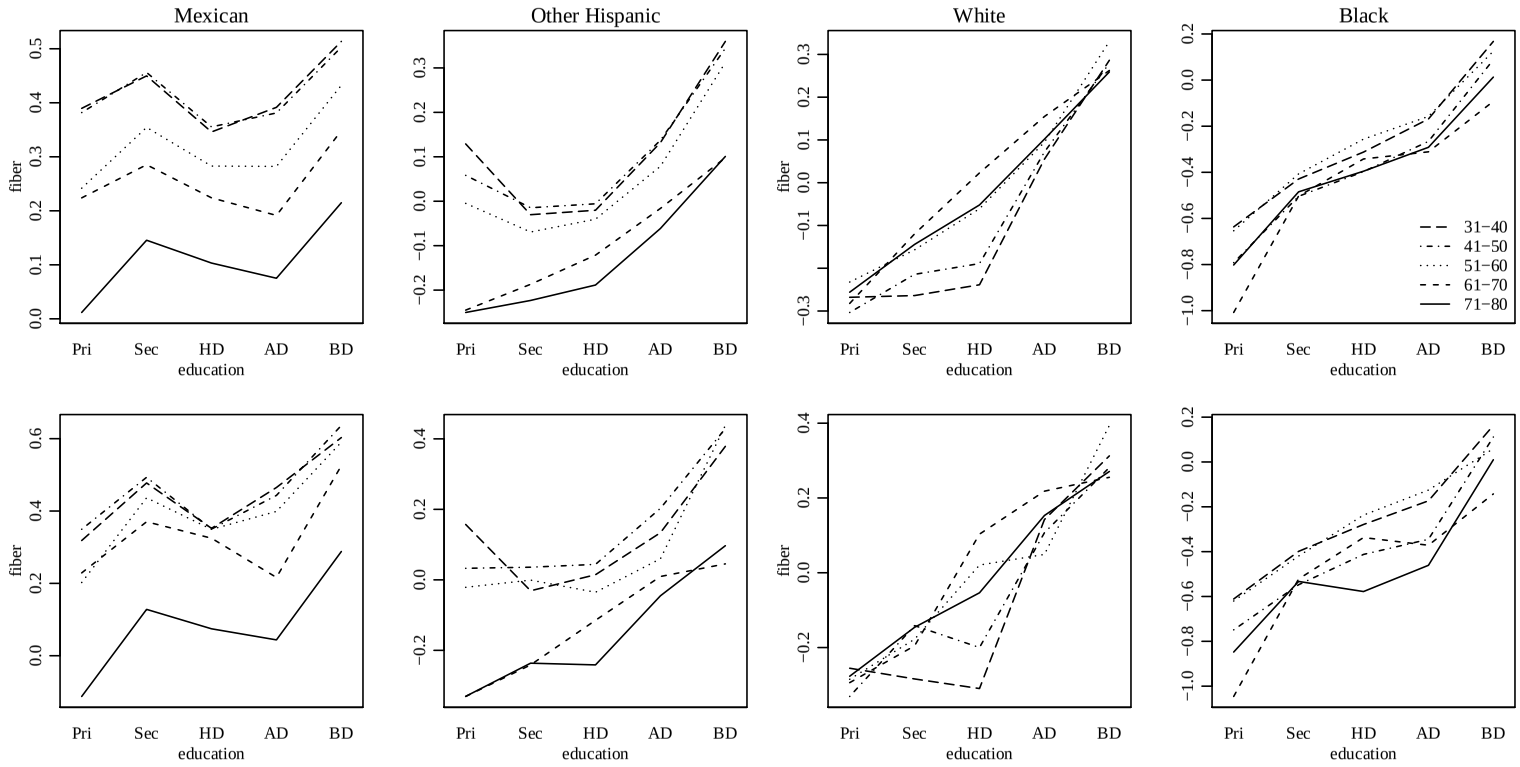

Figure 4.13: HA and SB interaction plots of estimated mean fiber intake by ethnicity, age and education level. HA and SB estimates are in the top and bottom rows, respectively.

for the single parameter ρ. We also note that due to the scale nonidentifiability of the Kronecker product we can assume that the variance parameter is equal to 1. The posterior approximation follows the outline of Section 4.2.2: the full conditionals for the effects and the full conditionals for the covariance matrices that do not exhibit a specific structure remain the same. The only difference is in the posterior approximation procedure for the structured covariance matrix, where a posterior sample of ρ can be obtained by importance sampling. The HA procedure can easily accommodate other structured covariances as well, with the only changes to the posterior approximation steps reflecting this additional prior information for the covariance matrix.

Generalizations of the HA prior are applicable to any model whose parameters consist of vectors, matrices and arrays for which some of the index sets are shared. This includes generalized linear models with categorical factors, as well as ANCOVA models that involve interactions between continuous and categorical explanatory variables. As an example of the latter case, suppose we are interested in estimating the linear relationship between an outcome and a set of explanatory variables for every combination of three categorical factors. The regression parameters then consist of an $m_1 \times m_2 \times m_3 \times p$ array, where m_1, m_2, m_3 are the numbers of factor levels and p is the number of continuous regressors. The usual ANCOVA decomposition can be used to parametrize this array in terms of main effects and interactions arrays, for which a hierarchical array prior may be used.

Computer code and data for the results in Sections 4.3 and 4.4 are available at the authors' websites.

Chapter 5

DISCUSSION AND FUTURE WORK

This thesis has provided novel insight into the study of relational datasets and ANOVA decomposition by leveraging properties of the matrix-variate and array-variate normal distributions. In Chapter 2 we presented a likelihood ratio test for relational datasets. In contrast to the previous testing literature for matrix normal models that required multiple observations and concentrated on testing a null hypothesis of separability versus an unstructured alternative, we proposed testing a null of no row or column correlations versus an alternative of full row and column correlations using a single observation of a relational dataset. Two natural extensions exist to this treatment: First, there is a growing literature for an asymptotic treatment of network data Bickel and Chen [2009], Rohe and Yu [2012], and so studying the asymptotic properties of the proposed test is of interest. Secondly, while the test is developed for square data matrices, a great number of applications have rectangular data matrices and testing for independence among the rows and columns of those is of interest. Chapter 3 developed a framework for answering the inference problem of how to account for the row and column correlation in relational data in case the test of Chapter 2 rejects the null. We discussed in detail the class of maximum likelihood estimators for the square matrix normal and extended these results to estimating mean parameters and covariance parameters jointly. This was all done in the context of the square matrix normal, and just as with the testing framework, it is desirable to extend the results to rectangular data matrices. In the remainder of this chapter we propose several avenues for developing the aforementioned open questions:

Asymptotic treatment of the likelihood ratio test. When studying likelihood ratio tests it is frequently of interest to study their asymptotic behavior. Knowing the asymptotic distribution of the likelihood ratio test statistic in the case of Chapter 2 could reduce the

computational burden for conducting the test as we would be able to rely on the asymptotic approximation to the quantiles rather than constructing the null distribution for every sample size. We are interested in the asymptotic behavior of the test for a single relational data matrix as the number of actors (the dimension of the matrix) increases:

Conjecture 8. *Let $Y \sim N_{m\times m}(0, I, I)$, and write the likelihood ratio statistic as $T_m(Y) = m\left(\log\left|Y^t \hat{D}_{\mathrm{r}}^{-1}(Y)\, Y \circ I\right| - \log\left|Y^t \hat{D}_{\mathrm{r}}^{-1}(Y)\, Y\right|\right)$. Then*

$$\frac{T_m(Y) - m\left((m+1)\log m - \log m!\right)}{m\sqrt{\log m}} \xrightarrow{\mathrm{d}} Q$$

for some random variable Q.

We have observed that for the case of increasing m, the centering and scaling values proposed in Conjecture 8 stabilize the asymptotic distribution of the LRT statistic under the null from Chapter 2. The centering and scaling values are motivated by the the mean and variance of $m(\log|Y^tY \circ I| - \log\left|Y^tY\right|)$ where $Y \sim N_{m\times m}(0, I, I)$.

Testing for independence along the rows and columns of rectangular data matrices. Other than in the case of relational data where the data is frequently bound to square matrices due to the restriction that the index sets of the rows and columns must be the same (that is, senders in the network must also be receivers), data matrices are frequently rectangular. Classical statistics considered the problem of inference for n independent observations of p attributes where the assumption was that $n > p$. In modern data streams both the assumption of independent observations and of more replicates than attributes is frequently violated (for example in the study of microarrays [Leek and Storey, 2008, Efron, 2009]). It is thus desirable to test for independence among the rows and among the columns of these non-square datasets to determine proper inference approaches. Since the likelihood of the rectangular full matrix normal distribution is unbounded for a single observation, the likelihood ratio test developed in Chapter 2 cannot be applied immediately. This problem can be approached either by restricting the alternative model such that the likelihood of the restricted model is bounded or by considering a penalized likelihood ratio test. For the first approach, a possible restriction of the alternative model is to a class of separable factor

analytic models Fosdick and Hoff [2012]. For the second approach, recent literature has employed penalties that induce sparsity in the row and column precision matrices [Allen and Tibshirani, 2010]. Since the estimator under the null hypothesis of independence can be viewed as the limiting case for sparse estimators, it would appear natural to consider a test based on the ratio of the likelihood under the null and the penalized likelihood under the alternative. It remains to be shown that a test based on the penalized likelihood ratio in this case is properly behaved, but results of Fan and Peng [2004] on the asymptotic behavior of a penalized likelihood ratio test with a growing number of parameters are encouraging.

Maximum likelihood estimation for rectangular matrix-variate normal distributions. As mentioned, many types of data do not come in square matrix form and so the theory we developed for maximum likelihood estimation must be extended to these rectangular matrices. The first step in this direction is the development of necessary and sufficient conditions for the existence of maximum likelihood estimates for rectangular matrix normal distributions. The current results in the literature do not appear to be correct. Potential avenues for finding the necessary and sufficient conditions appear to rely on arguments from linear algebra. In personal communication, Peter Hoff proposed that finding the minimal rank of a specific matrix product is equivalent to determining if the matrix normal likelihood for a particular triplet $(n, m_{\mathrm{r}}, m_{\mathrm{c}})$, the number of observations, the number of rows and the number of columns respectively, is bounded. This condition appears to be difficult to verify in practice. When these results are available they should be extendable to the array-variate normal distribution.

Beyond maximum likelihood estimators Consider the regression framework of Chapter 3, but with a single observation from a rectangular matrix normal distribution. In this case, the likelihood is unbounded and so other approaches must be employed. One such approach would be to leverage the results of Chételat and Wells [2012] on estimating the mean of a multivariate normal distribution when the covariance is unknown and there are more features than independent observations ($p > n$) to improve the estimates of our regression parameters. The estimators proposed in Chételat and Wells [2012] employ a low

rank estimate of the covariance matrix to improve the estimation of the mean. As such, it is plausible that a similar approach using low rank estimates of the row and column covariance matrices can yield improved estimates of regression parameters.

BIBLIOGRAPHY

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

Margaret J. Albrink and Irma H. Ullrich. Interaction of dietary sucrose and fiber on serum lipids in healthy young men fed high carbohydrate diets. *The American journal of clinical nutrition*, 43(3):419–428, 1986.

Genevera I. Allen and Robert Tibshirani. Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics*, 4(2):764–790, 2010.

Takeshi Amemiya and Wayne A. Fuller. A comparative study of alternative estimators in a distributed lag model. *Econometrica, Journal of the Econometric Society*, pages 509–529, 1967.

T.W. Anderson, Huang Hsu, and Kai-Tai Fang. Maximum-likelihood estimates and likelihood-ratio criteria for multivariate elliptically contoured distributions. *Canadian Journal of Statistics*, 14(1):55–59, 1986.

Donald W.K. Andrews. A note on the unbiasedness of feasible gls, quasi-maximum likelihood, robust, adaptive, and spectral estimators of the linear model. *Econometrica: Journal of the Econometric Society*, pages 687–698, 1986.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM journal on matrix analysis and applications*, 29(1):328–347, 2007.

Gregory L. Austin, Lorraine G. Ogden, and James O. Hill. Trends in carbohydrate, fat, and protein intakes and association with energy intake in normal-weight, overweight, and

obese individuals: 1971–2006. *The American journal of clinical nutrition*, 93(4):836–843, 2011.

P. Peter Basiotis, Robin G. Thomas, June L. Kelsay, and Walter Mertz. Sources of variation in energy intake by men and women as determined from one year's daily dietary records. *The American journal of clinical nutrition*, 50(3):448–453, 1989.

R. Beran. ASP fits to multi-way layouts. *Annals of the Institute of Statistical Mathematics*, 57(2):201–220, 2005.

James O. Berger and Ruo-Yong Yang. Noninformative priors and bayesian testing for the ar (1) model. *Econometric Theory*, 10(3-4):461–482, 1994.

S. Bergmann, J. Ihmels, and N. Barkai. Similarities and differences in genome-wide expression data of six organisms. *PLoS biology*, 2(1):e9, 2003.

J.H. Bergstrand. The gravity equation in international trade: some microeconomic foundations and empirical evidence. *The review of economics and statistics*, pages 474–481, 1985.

J.H. Bergstrand. The generalized gravity equation, monopolistic competition, and the factor-proportions theory in international trade. *The review of economics and statistics*, pages 143–153, 1989.

P.J. Bickel and A. Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

Edwin Bonilla, Kian Ming Chai, and Christopher Williams. Multi-task gaussian process prediction. 2008.

Gareth Butland, José Manuel Peregrín-Alvarez, Joyce Li, Wehong Yang, Xiaochun Yang, Veronica Canadien, Andrei Starostine, Dawn Richards, Bryan Beattie, Nevan Krogan, et al. Interaction network containing conserved and essential protein complexes in escherichia coli. *Nature*, 433(7025):531–537, 2005.

M. Chandalia, A. Garg, D. Lutjohann, K. von Bergmann, S.M. Grundy, and L.J. Brinkley. Beneficial effects of high dietary fiber intake in patients with type 2 diabetes mellitus. *New England Journal of Medicine*, 342(19):1392–1398, 2000.

Anoop Cherian. *Similarity Search in Visual Data*. PhD thesis, UNIVERSITY OF MINNESOTA, 2013.

Anoop Cherian, Suvrit Sra, Arindam Banerjee, and Nikolaos Papanikolopoulos. Efficient similarity search for covariance matrices via the jensen-bregman logdet divergence. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2399–2406. IEEE, 2011.

Didier Chételat and Martin T Wells. Improved multivariate normal mean estimation with unknown covariance when p is greater than n. *The Annals of Statistics*, 40(6):3137–3160, 2012.

R. Courant. *Dirichlet's principle, conformal mapping, and minimal surfaces*. Interscience Publishers, Inc, New York, 1950.

Yue Cui, James S Hodges, Xiaoxiao Kong, and Bradley P Carlin. Partitioning degrees of freedom in hierarchical and other richly parameterized models. *Technometrics*, 52(1): 124–136, 2010.

A.P. Dawid. Spherical matrix distributions and a multivariate model. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 254–261, 1977.

A.P. Dawid. Extendibility of spherical matrix distributions. *Journal of Multivariate Analysis*, 8(4):559–566, 1978.

A.P. Dawid. Some matrix-variate distribution theory: notational considerations and a bayesian application. *Biometrika*, 68(1):265–274, 1981.

M.L. Eaton. *Multivariate statistics: a vector space approach*. Wiley New York, 1983.

Bradley Efron. Are a set of microarrays independent of each other? *The annals of applied statistics*, 3(3):922, 2009.

P. Svante Eriksen. Proportionality of covariance matrices. *The Annals of Statistics*, pages 732–748, 1987.

Jianqing Fan and Heng Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

A. Fletcher, C. Bonell, and A. Sorhaindo. You are what your friends eat: systematic review of social network analyses of young people's eating behaviours and bodyweight. *Journal of epidemiology and community health*, 65(6):548–555, 2011.

Bernhard K. Flury. Proportionality of¡ i¿ k¡/i¿ covariance matrices. *Statistics & probability letters*, 4(1):29–33, 1986.

Bailey K. Fosdick and Peter D. Hoff. Separable factor analysis with applications to mortality data. *arXiv preprint arXiv:1211.3813*, 2012.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. *Arxiv preprint arXiv:1001.0736*, 2010.

Andrew Gelman. Analysis of variance—why it is more important than ever. *The Annals of Statistics*, 33(1):1–53, 2005.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel Hierarchical Models*. 2007.

A. Genkin, D.D. Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

Peter J. Green. Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique*, pages 245–259, 1987.

Peter J. Green. On use of the em for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 443–452, 1990.

A.K. Gupta and D.K. Nagar. *Matrix variate distributions*, volume 104. Chapman & Hall/CRC, 1999.

A.K. Gupta and T. Varga. A new class of matrix variate elliptically contoured distributions. *Statistical Methods & Applications*, 3(2):255–270, 1994.

A.K. Gupta and T. Varga. Some inference problems for matrix variate elliptically contoured distributions. *Statistics: A Journal of Theoretical and Applied Statistics,*, 26(3):219–229, 1995.

F.J. Henk Don and Jan R. Magnus. On the unbiasedness of iterated gls estimators. *Communications in Statistics-Theory and Methods*, 9(5):519–527, 1980.

Nicholas J. Higham and Hyun-Min Kim. Solving a quadratic matrix equation by newton's method with exact line searches. *SIAM Journal on Matrix Analysis and Applications*, 23 (2):303–316, 2001.

James S. Hodges, Yue Cui, Daniel J. Sargent, and Bradley P. Carlin. Smoothing balanced single-error-term analysis of variance. *Technometrics*, 49(1):12–25, 2007.

Peter D. Hoff. Bilinear mixed-effects models for dyadic data. *Journal of the american Statistical association*, 100(469):286–295, 2005a.

Peter D. Hoff. Bilinear mixed-effects models for dyadic data. *J. Amer. Statist. Assoc.*, 100(469):286–295, 2005b. ISSN 0162-1459. URL `http://dx.doi.org/10.1198/016214504000001015`.

Peter D. Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 657–664. MIT Press, Cambridge, MA, 2008. URL `http://cran.r-project.org/web/packages/eigenmodel/`.

Peter D. Hoff. Separable covariance arrays via the tucker product, with applications to multivariate relational data. *Bayesian Analysis*, 6(2):179–196, 2011.

Peter D. Hoff, A.E. Raftery, and M.S. Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

Peter D. Hoff, Bailey K. Fosdick, Alexander Volfovsky, and Katherine Stovel. Likelihoods for fixed rank nomination networks. *arXiv preprint arXiv:1212.6234*, 2012.

P.W. Holland, K.B. Laskey, and S. Leinhardt. Stochastic blockmodels: first steps. *Social networks*, 5(2):109–137, 1983.

Alan T. James. Normal multivariate analysis and the orthogonal group. *The Annals of Mathematical Statistics*, pages 40–75, 1954.

Soøren Tolver Jensen and Jesper Madsen. Estimation of proportional covariances in the presene of certain linear restrictions. *The Annals of Statistics*, 32(1):219–232, 2004.

Søren Tolver Jensen and Søren Johansen. Estimation of proportional covariances. *Statistics & probability letters*, 6(2):83–85, 1987.

Gunnar Johansson, Asa Wikman, Ann-Mari Ahren, Goran Hallmans, Ingegerd Johansson, et al. Underreporting of energy intake in repeated 24-hour recalls related to gender, age, weight status, day of interview, educational level, reported food intake, smoking habits and area of living. *Public health nutrition*, 4(4):919–928, 2001.

Robert E. Kass and Larry Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934, 1995. ISSN 0162-1459.

D.A. Kenny and L. La Voie. The social relations model. *Advances in experimental social psychology*, 18:142–182, 1984.

C.G. Khatri. On conditions for the forms of the type xax'to be distributed independently or to obey wishart distribution. *Calcutta Statist. Assoc. Bull*, 8:162–168, 1959.

C.G. Khatri. Conditions for wishartness and independence of second degree polynomials in a normal vector. *The Annals of Mathematical Statistics*, 33(3):1002–1007, 1962.

T.G. Kolda. *Multilinear operators for higher-order decompositions*. United States. Department of Energy, 2006.

T.G. Kolda and B.W. Bader. Tensor decompositions and applications. *SIAM review*, 51 (3):455, 2009.

John Kruschke. *Doing Bayesian data analysis: A tutorial introduction with R and BUGS*. Academic Press, 2010.

R. Lafosse and J.M.F. Ten Berge. A simultaneous concor algorithm for the analysis of two partitioned matrices. *Computational statistics & data analysis*, 50(10):2529–2535, 2006.

S.G. Lazzarini, F.R. Chaddad, and M.L. Cook. Integrating supply chain and network analyses: the study of netchains. *Journal on chain and network science*, 1(1):7–22, 2001.

Jeffrey T. Leek and John D. Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723, 2008.

J. Leskovec, K.J. Lang, A. Dasgupta, and M.W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proceeding of the 17th international conference on World Wide Web*, pages 695–704. ACM, 2008.

Heng Li. The covariance structure and likelihood function for multivariate dyadic data. *J. Multivariate Anal.*, 97(6):1263–1271, 2006. ISSN 0047-259X. URL `http://dx.doi.org/10.1016/j.jmva.2005.06.004`.

Heng Li and Eric Loken. A unified theory of statistical analysis and inference for variance component models for dyadic data. *Statist. Sinica*, 12(2):519–535, 2002. ISSN 1017-0405.

J.R. Lincoln and M.L. Gerlach. *Japan's network economy: structure, persistence, and change*. Cambridge University Press, 2004.

N. Lu and D.L. Zimmerman. The likelihood ratio test for a separable covariance matrix. *Statistics & probability letters*, 73(4):449–457, 2005.

G.S. Maddala. Generalized least squares with an estimated variance covariance matrix. *Econometrica: Journal of the Econometric Society*, pages 23–33, 1971.

Jan R. Magnus. On differentiating eigenvalues and eigenvectors. *Econometric Theory*, 1: 179–191, 1985.

L.L. McQuitty and J.A. Clark. Clusters from iterative, intercolumnar correlational analysis. *Educational and psychological measurement*, 1968.

R.G. Miller and B.W. Brown. *Beyond ANOVA: basics of applied statistics*. Chapman & Hall/CRC, 1997.

M.W. Mitchell, M.G. Genton, and M.L. Gumpertz. A likelihood ratio test for separability of covariances. *Journal of Multivariate Analysis*, 97(5):1025–1043, 2006.

C.J. Moerman, H.B.A.S. De Mesquita, and S. Runia. Dietary sugar intake in the aetiology of biliary tract cancer. *International journal of epidemiology*, 22(2):207–214, 1993.

J. Montonen, P. Knekt, R. Järvinen, A. Aromaa, and A. Reunanen. Whole-grain and fiber intake and the incidence of type 2 diabetes. *The American journal of clinical nutrition*, 77(3):622–629, 2003.

Heinz Neudecker and Tom Wansbeek. Fourth-order properties of normally distributed random matrices. *Linear Algebra and Its Applications*, 97:13–21, 1987.

Samara Joy Nielsen, Barry M Popkin, et al. Changes in beverage intake between 1977 and 2001. *American journal of preventive medicine*, 27(3):205–210, 2004.

K. Nowicki and T.A.B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

C.L. Olson. On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 83(4):579, 1976.

W.H. Panning. Fitting blockmodels to data. *Social Networks*, 4(1):81–101, 1982.

D.K. Park, A. Gelman, and J. Bafumi. State level opinions from national surveys: Poststratification using multilevel logistic regression. *Public Opinion in State Politics*, pages 209–28, 2006.

T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Y. Park, A.F. Subar, A. Hollenbeck, and A. Schatzkin. Dietary fiber intake and mortality in the nih-aarp diet and health study. *Archives of internal medicine*, 171(12):1061, 2011.

M.G. Pittau, R. Zelli, and A. Gelman. Economic disparities and life satisfaction in european regions. *Social indicators research*, 96(2):339–361, 2010.

M.S. Pollard, J.S. Tucker, H.D. Green, D. Kennedy, and M.H. Go. Friendship networks and trajectories of adolescent tobacco use. *Addictive behaviors*, 35(7):678–685, 2010.

G.E. Potter, M.S. Handcock, I.M. Longini, and M.E. Halloran. Estimating within-school contact networks to understand influenza transmission. *The Annals of Applied Statistics*, 6(1):1–26, 2012.

K. Rohe and B. Yu. Co-clustering for directed graphs; the stochastic co-blockmodel and a spectral algorithm. *arXiv preprint arXiv:1204.2296*, 2012.

K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Thomas J Rothenberg. Approximate normality of generalized least squares estimates. *Econometrica: Journal of the Econometric Society*, pages 811–825, 1984.

A. Roy and R. Khattree. On implementation of a test for kronecker product covariance structure for multivariate repeated measures data. *Statistical Methodology*, 2(4):297–306, 2005.

S.F. Sampson. *A novitiate in a period of change: An experimental and case study of social relationships*. PhD thesis, Cornell University, September, 1968.

Bernard W Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, pages 795–810, 1982.

Suvrit Sra. Positive definite matrices and the symmetric stein divergence. *arXiv preprint arXiv:1110.1773*, 2011.

M.S. Srivastava, T. von Rosen, and D. Von Rosen. Models with a kronecker product covariance structure: estimation and testing. *Mathematical Methods of Statistics*, 17 (4):357–370, 2008.

J.M. Stuart, E. Segal, D. Koller, and S.K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, 2003.

E.A. Thompson and C.J. Geyer. Fuzzy p-values in latent variable problems. *Biometrika*, 94(1):49–60, 2007.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

J. Tinbergen et al. *Shaping the world economy: Suggestions for an international economic policy*. Twentieth Century Fund New York, 1962.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

USDA. *Food and Nutrient Database for Dietary Studies 4.1*. U.S. Department of Agriculture, Agricultural Research Service, Food Surveys Research Group, Beltsville, MD, 2010.

G.G. Van De Bunt, M.A.J. Van Duijn, and T.A.B. Snijders. Friendship networks through time: An actor-oriented dynamic statistical network model. *Computational & Mathematical Organization Theory*, 5(2):167–192, 1999.

Eliseu Verly Junior, Regina Mara Fisberg, Chester Luis Galvão Cesar, and Dirce Maria Lobo Marchioni. Sources of variation of energy and nutrient intake among adolescents in são paulo, brazil. *Cadernos de Saúde Pública*, 26(11):2129–2137, 2010.

Dietrich Von rosen. Moments for matrix normal variables. *Statistics: A Journal of Theoretical and Applied Statistics*, 19(4):575–583, 1988.

Y.J. Wang and G.Y. Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19, 1987.

A.H. Westveld and P.D. Hoff. A mixed effects model for longitudinal relational and network data, with applications to international trade and conflict. *The Annals of Applied Statistics*, 5(2A):843–872, 2011.

H.C. White, S.A. Boorman, and R.L. Breiger. Social structure from multiple networks. i. blockmodels of roles and positions. *American journal of sociology*, pages 730–780, 1976.

Eun Ju Yang, Hae Kyung Chung, Wha Young Kim, Jean M Kerver, and Won O Song. Carbohydrate intake is associated with diet quality and risk factors for cardiovascular disease in us adults: Nhanes iii. *Journal of the American College of Nutrition*, 22(1): 71–79, 2003.

Kai Yu, Wei Chu, Shipeng Yu, Volker Tresp, and Zhao Xu. Stochastic relational models for discriminative link prediction. *Advances in neural information processing systems*, 19: 1553, 2007.

M. Yuan and Y. Lin. Efficient empirical bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, 100(472):1215–1225, 2005.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Arnold Zellner. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, 57(298):348–368, 1962.

Appendix A

PROOFS OF RESULTS IN CHAPTER 2

Proof of Theorem 2. To show that the solutions to the likelihood equations provide a unique minimizer to the scaled log likelihood function (Equation 2.1) we will show that the Hessian of l evaluated at the solutions is strictly positive definite and then demonstrate that only a single solution is possible. We rewrite Equation 2.1 here, explicitly stating that we will be considering diagonal matrices

$$l\left(D_{\mathrm{r}}, D_{\mathrm{c}}; Y\right) = -2\log L\left(D_{\mathrm{r}}, D_{col}; Y\right) \quad = \quad \mathrm{tr}\left[D_{\mathrm{r}}^{-1} Y D_{\mathrm{c}}^{-1} Y^{t}\right] - \log\left|D_{\mathrm{c}}^{-1} \otimes D_{\mathrm{r}}^{-1}\right| + c,$$

writing for simplicity $\Psi = D_{\mathrm{r}}^{-1}$ and $\Gamma = D_{\mathrm{c}}^{-1}$ we take first derivatives with respect to the diagonal matrices of Γ and Ψ:

$$\frac{\partial l\left(D_{\mathrm{r}}, D_{\mathrm{c}}; Y\right)}{\partial_{\Psi}} \quad = \quad Y\Gamma Y^{t} \circ I - m\Psi^{-1} \tag{A.1}$$

$$\frac{\partial l\left(D_{\mathrm{r}}, D_{\mathrm{c}}; Y\right)}{\partial_{\Gamma}} \quad = \quad Y^{t}\Psi Y \circ I - m\Gamma^{-1}, \tag{A.2}$$

yielding the familiar equations used to find the maximizers of the likelihood. Considering the singular value decomposition of $Y = ALB^{t}$, the above can also be written as partial derivatives with respect to the entries of Ψ and Γ (since these are diagonal matrices, the index k refers to the k^{th} row, k^{th} column entry in the matrix):

$$\frac{\partial l\left(D_{\mathrm{r}}, D_{\mathrm{c}}; Y\right)}{\partial_{\Psi_j}} \quad = \quad \sum_{ikm} L_i L_m \Gamma_k \left(A_{jm} A_{ji} B_{km} B_{ki}\right) - \frac{m}{\Psi_j} \tag{A.3}$$

$$\frac{\partial l\left(D_{\mathrm{r}}, D_{\mathrm{c}}; Y\right)}{\partial_{\Gamma_k}} \quad = \quad \sum_{ijm} L_i L_m \Psi_j \left(A_{jm} A_{ji} B_{km} B_{ki}\right) - \frac{m}{\Gamma_k} \tag{A.4}$$

To compute the Hessian, we take derivatives of equations A.3 and A.4, yielding the second partial derivatives of l_{D}:

$$\begin{aligned}
\frac{\partial l\left(D_{\mathrm{r}}, D_{\mathrm{c}}; Y\right)}{\partial_{\Psi_j \Gamma_k}} &= f\left(j,k\right) = \sum_{im} L_i L_m \left(A_{jm} A_{ji} B_{km} B_{ki}\right) = Y_{jk}^2 \\
\frac{\partial l\left(D_{\mathrm{r}}, D_{\mathrm{c}}; Y\right)}{\partial_{\Psi_j \Psi_l}} = \frac{\partial l\left(D_{\mathrm{r}}, D_{\mathrm{c}}; Y\right)}{\partial_{\Gamma_k \Gamma_m}} &= 0 \\
\frac{\partial l\left(D_{\mathrm{r}}, D_{\mathrm{c}}; Y\right)}{\partial_{\Psi_j \Psi_j}} &= \frac{m}{\Psi_j^2} \\
\frac{\partial l\left(D_{\mathrm{r}}, D_{\mathrm{c}}; Y\right)}{\partial_{\Gamma_k \Gamma_k}} &= \frac{m}{\Gamma_k^2}.
\end{aligned}$$

As such, we can write the Hessian matrix H as

$$H = \begin{bmatrix} m\Psi^{-2} & F \\ F^t & m\Gamma^{-2} \end{bmatrix}$$

where $F = \left[f\left(j,k\right)\right]_{j,k}$. Our first observation is that F is an everywhere positive matrix since $f\left(j,k\right) = Y_{jk}^2 > 0 \forall j,k$ (since $P\left(Y \neq 0\right) = 1$).

To show that l_{F} is minimized at the solutions to the likelihood equations A.1 and A.2 we will show that the Hessian H is strictly positive definite at the solutions. For that, we will verify Sylvester's criterion: a matrix H is positive definite if and only if all of its leading minors are positive (or equivalently its trailing minors). First we note that the Kronecker product of the covariances leads to a nonidentifiability in the scale of the individual matrices, and so WLOG we let $\Psi_1 = 1$. We consider the reparametrized problem and its' Hessian $\tilde{H} = H_{-1,-1}$, the Hessian of the original problem with the first row and column removed. Now, the boundedness of the likelihood function implies that the $m-1$ leading minors are positive (since $\hat{\Psi}$ is a positive diagonal matrix) and so we are left with verifying that the remaining m minors are positive. Abusing notation a bit and writing Ψ to correspond to the reparametrized version of row precisions, we get the first minor that includes entries

other than those in Ψ is

$$\begin{aligned}\begin{vmatrix} m\hat{\Psi}^{-2} & F_{\cdot,1} \\ F_{1,\cdot} & m\hat{\Gamma}_1^{-2} \end{vmatrix} &= \left| \frac{m}{\hat{\Gamma}_1^2} - F_{\cdot,1}\frac{\hat{\Psi}^2}{m}F_{1,\cdot} \right| \left| m\hat{\Psi}^{-2} \right| \\ &:= |a|\,|b|\end{aligned}$$

Clearly, $|b| > 0$ as it is simply the previous minor. Now, a does not satisfy the first derivative of $l\left(D_{\mathrm{r}}, D_{\mathrm{c}}; Y\right)$ with respect to Γ_1 (Equation A.4) and more so we note that

$$\frac{m}{\hat{\Gamma}_1^2} = \frac{1}{m}\left(\sum_j \hat{\Psi}_j F_{j,1}\right)^2 = \frac{1}{m}\left(F_{\cdot,1}\hat{\Psi}^2 F_{1,\cdot} + c\right)$$

where c is always positive since it is a sum of positive values. Thus, $|a| = c/m > 0$. We can apply this approach to the remaining $m-1$ minors, where the kth minor is given by $M_k = |a|M_{k-1}$ where $|a| > 0$. Thus we have verified Sylvester's criterion and have demonstrated that the Hessian of $l\left(D_{\mathrm{r}}, D_{\mathrm{c}}; Y\right)$ is strictly positive definite at the solution to the likelihood equations which means we have a local minimum of the scaled log likelihood function (or a local maximum of the likelihood function).

To show uniqueness we apply the Mountain Pass Theorem [page 223 of Courant, 1950]: Since $l\left(D_{\mathrm{r}}, D_{\mathrm{c}}; Y\right)$ is smooth, differentiable and coercive (that is $l\left(D_{\mathrm{r}}, D_{\mathrm{c}}; Y\right) \to \infty$ as $|D_{\mathrm{c}} \otimes D_{\mathrm{r}}| \to \infty$), we have that if there are two critical points x_1 and x_2 that are strict minima (strictly positive definite Hessian), then there must be another critical point, distinct from x_1 and x_2, that is *not* a relative minimizer of l. This contradicts the above notion that the Hessian is strictly positive definite for every critical point. □

Proof of Theorem 3. For p random variables $Y_1, \ldots, Y_p$ where $Y_i \sim N_{m\times m}\left(0, d_i\Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}\right)$, we write $D = (d_1, \ldots, d_p)$ and the scaled log likelihood function as

$$l\left(D, \Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}} | Y_1, \ldots, Y_p\right) \propto \sum_{i=1}^{p} \frac{1}{d_i}\mathrm{tr}\left(Y_i\Sigma_{\mathrm{c}}^{-1}Y_i^t\Sigma_{\mathrm{r}}^{-1}\right) - mp\log\left|\Sigma_{\mathrm{c}}^{-1}\right| - mp\log\left|\Sigma_{\mathrm{c}}^{-1}\right| - m^2\log\left|D^{-1}\right|.$$

Taking first derivatives with respect to Σ_{r}^{-1}, Σ_{c}^{-1} and d_i^{-1} and setting them equal to zero

yields the likelihood equations:

$$\begin{aligned} mp\Sigma_{\mathrm{r}} &= \sum \frac{1}{d_i} Y_i \Sigma_{\mathrm{c}}^{-1} Y_i^t \\ mp\Sigma_{\mathrm{c}} &= \sum \frac{1}{d_i} Y_i^t \Sigma_{\mathrm{r}}^{-1} Y_i \\ m^2 d_i &= \mathrm{tr}\left(Y_i \Sigma_{\mathrm{c}}^{-1} Y_i^t \Sigma_{\mathrm{r}}^{-1}\right). \end{aligned}$$

We note that when holding Σ_{r} and Σ_{c} constant, $l\left(D, \Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}\right)$ is a strictly convex function of D^{-1} and so with probability 1 it attains a global minimum at points that satisfy the first derivative condition $m^2 d_i = \mathrm{tr}\left(Y_i \Sigma_{\mathrm{c}}^{-1} Y_i^t \Sigma_{\mathrm{r}}^{-1}\right)$. We define the profile likelihood

$$\begin{aligned} g\left(\Sigma_{\mathrm{r}}^{-1}, \Sigma_{\mathrm{c}}^{-1}\right) &= \inf_{D^{-1} \in \mathbb{R}_+^p} l\left(D, \Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}\right) \\ &= m^2 p - mp \log\left|\Sigma_{\mathrm{r}}^{-1}\right| - mp \log\left|\Sigma_{\mathrm{c}}^{-1}\right| + m^2 \sum \log\left|\mathrm{tr}\left(Y_i \Sigma_{\mathrm{c}}^{-1} Y_i^t \Sigma_{\mathrm{r}}^{-1}\right)\right| + c. \end{aligned}$$

There are two nonidentifiabilities in the scaled log likelihood given by $l\left(D, \Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}\right) = l\left(aD, b\Sigma_{\mathrm{r}}, \frac{1}{ab}\Sigma_{\mathrm{c}}\right)$ for $a, b > 0$ and so we restrict our domain to consider minimization the of $l\left(D, \Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}\right)$ over $\mathbb{R}_+^p \times \mathcal{S}_2 \times \mathcal{S}_2$ where $\mathcal{S}_2$ is the bounded subset of $\mathcal{S}_+^m$ of positive definite matrices whose largest eigenvalue is 1. This makes the model identifiable. Since minimization of $l\left(D, \Sigma_{\mathrm{r}}, \Sigma_{\mathrm{c}}\right)$ is equivalent to minimization of $g\left(\Sigma_{\mathrm{r}}^{-1}, \Sigma_{\mathrm{c}}^{-1}\right)$, we restrict minimizing $g\left(\Sigma_{\mathrm{r}}^{-1}, \Sigma_{\mathrm{c}}^{-1}\right)$ to $\mathcal{S}_2 \times \mathcal{S}_2$. The continuity of g on $\mathcal{S}_+^m \times \mathcal{S}_+^m \supset \mathcal{S}_2 \times \mathcal{S}_2$ guarantees that it will attain a minimum on $\mathcal{S}_2 \times \mathcal{S}_2$ as long as for $\overline{\Sigma_{\mathrm{r}}^{-1}}, \overline{\Sigma_{\mathrm{c}}^{-1}} \in \overline{\mathcal{S}_2}$

$$\begin{aligned} \lim_{\Sigma_{\mathrm{r}}^{-1} \to \overline{\Sigma_{\mathrm{r}}^{-1}}} g\left(\Sigma_{\mathrm{r}}^{-1}, \Sigma_{\mathrm{c}}^{-1}\right) &= g\left(\overline{\Sigma_{\mathrm{r}}^{-1}}, \Sigma_{\mathrm{c}}^{-1}\right) \\ \lim_{\Sigma_{\mathrm{c}}^{-1} \to \overline{\Sigma_{\mathrm{c}}^{-1}}} g\left(\Sigma_{\mathrm{r}}^{-1}, \Sigma_{\mathrm{c}}^{-1}\right) &= g\left(\Sigma_{\mathrm{r}}^{-1}, \overline{\Sigma_{\mathrm{c}}^{-1}}\right) \\ \lim_{\Sigma_{\mathrm{c}}^{-1} \to \overline{\Sigma_{\mathrm{c}}^{-1}}} \lim_{\Sigma_{\mathrm{r}}^{-1} \to \overline{\Sigma_{\mathrm{r}}^{-1}}} g\left(\Sigma_{\mathrm{r}}^{-1}, \Sigma_{\mathrm{c}}^{-1}\right) &= g\left(\overline{\Sigma_{\mathrm{r}}^{-1}}, \overline{\Sigma_{\mathrm{c}}^{-1}}\right). \end{aligned}$$

All three conditions are met for positive definite boundary points, so all that remains to show is that $g\left(\Sigma_{\mathrm{r}}^{-1}, \Sigma_{\mathrm{c}}^{-1}\right) \to \infty$ when a subset of the eigenvalues of Σ_{r}^{-1} or Σ_{c}^{-1} approach 0 (behavior near positive semidefinite boundary points). It is immediate that when the eigenvectors of Σ_{r}^{-1} do not match the left eigenvectors of Y_i and the eigenvectors of Σ_{c}^{-1}

do not match the right eigenvectors of Y_i, $\mathrm{tr}\left(Y_i\Sigma_{\mathrm{c}}^{-1}Y_i^t\Sigma_{\mathrm{r}}^{-1}\right) \to c_i > 0$, thus the $\log\mathrm{tr}\left(\cdot\right)$ term converges to a finite constant. As such, the behavior of $g\left(\Sigma_{\mathrm{r}}^{-1}, \Sigma_{\mathrm{c}}^{-1}\right)$ when subsets of eigenvalues approach zero is completely governed by the log determinant terms, both of which will converge to $+\infty$. $\square$

Appendix B

METRICS ON SYMMETRIC POSITIVE DEFINITE MATRICES

An important task in studying covariance matrices is the ability to determine if two covariance matrices are similar. Symmetric positive definite matrices can be embedded in Euclidean space, and traditional Euclidean distances can be employed. However, Euclidean distance (L_2 norm for example) does not respect the non-Euclidean geometry of symmetric positive definite matrices.

Positive definite matrices form a "self-dual convex cone whose strict interior is a Riemannian manifold" [Sra, 2011]. This manifold comes with a natural distance function

$$\delta_R(X, Y) := \|\log(Y^{-1/2} X Y^{-1/2})\|_F, \tag{B.1}$$

where $X, Y > 0$ and $\log(\cdot)$ is the matrix logarithm. Note that the above can be viewed as the sum of squared logarithms of the generalized eigenvalues of X and Y. This distance is the Riemannian metric on the manifold of positive definite matrices. Alternative distances have been proposed for evaluating "nearness" that respect the geometric properties of the manifold. A distance that has been getting a lot of attention in the machine learning literature is the Symmetric Stein Divergence (or Jensen-Bregman LogDet Divergence) of Cherian et al. [2011],

$$\delta_S^2(X, Y) := \log|(X + Y)/2| - \log|X||Y|/2. \tag{B.2}$$

Sra [2011] proves that $\delta_S(X, Y)$ is in fact a metric.

The reason for introducing Stein Divergence reflects the needs in machine learning to speed up computation. Computing the natural distance of Equation (B.1) requires the computation of generalized eigenvalues whereas to compute the Stein divergence of Equation (B.2) requires computing three Cholesky decompositions. In Cherian [2013], the author

suggests that the computational burden of the former is 12 times greater (though both remain of order $O(m^3)$ for m the dimension of the matrix).

Some similarities among the two functions were enumerated in Sra [2011], which proves these results for δ_S. Some properties that hold for both are stated below (δ represents either metric):

1. GL(m) invariance: For A invertible, $\delta(AXA^t, AYA^t) = \delta(X, Y)$ (the map $X \to AXA^t$ is an isometry for the metric).

2. Inversion: $\delta(X^{-1}, Y^{-1}) = \delta(X, Y)$ (inversion is an involutive isometry on SPD for this metric).

3. Kronecker expansion: For $A > 0$ we have $\delta(A \otimes X, A \otimes Y) = m\delta(X, Y)$.

4. Matrix geometric mean $X\#Y := X^{1/2}(X^{-1/2}YX^{-1/2})^{1/2}X^{1/2}$ has the variational characterization $X\#Y = \arg\min_{A>0} \delta^2(X, A) + \delta^2(Y, A)$.

5. Power relationship: For $t \in [0, 1]$, $\delta(X^t, Y^t) \leq t\delta(X, Y)$

A possibly undesirable property of δ_S^2 is its limited convexity/concavity. That is (Corollary 3 of Sra [2011]), for fixed Y, $\delta_S^2(X, Y)$ is convex in X when $X \leq (1+\sqrt{2})Y$ and concave if $X \geq (1 + \sqrt{2})Y$. More so, it is not geodesically convex (unlike δ_R).

The "log-Euclidean" distance proposed by Arsigny et al. [2007] is given by

$$\delta_L(X, Y) := \|\log X - \log Y\|_F. \tag{B.3}$$

The authors argue that this is necessary in order to fully benefit from the structure of symmetric positive definite matrices. However, it can be seen that this definition is not extremely useful for truly understanding the similarity between two matrices. Specifically, Equation (B.3) maps the symmetric positive definite matrices to a flat Riemannian manifold (that is a manifold that looks like Euclidean space in terms of distance). By doing so it only provides a lower bound for the Riemannian metric δ_R with equality of the two metrics only for commuting matrices.

VITA

Alexander Volfovsky grew up on three continents. In 2009 he received a Bachelors degree in Mathematics and a Masters degree in Statistics from the University of Chicago. In the same year he joined the Statistics Department at the University of Washington. In August 2013 he became a Doctor of Philosophy.