Joint Mean and Covariance Modeling of Matrix-Variate Data

by

Michael David Hornstein

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2018

Doctoral Committee:

Professor Kerby Shedden, Co-chair
Assistant Professor Shuheng Zhou, Co-chair
Professor Alfred Hero
Professor Douglas Richstone

Michael Hornstein

mdhorn@umich.edu

ORCID id: 0000-0003-2721-807X

ACKNOWLEDGEMENTS

First I want to thank my advisors Professors Shuheng Zhou and Kerby Shedden for their research and personal mentorship; it has been a great privilege to be advised by them. They served as role models of intellectual curiosity and the drive to gain insight into research questions, and they have been deeply invested in my success in research and in life. I appreciate the extensive time they spent with me discussing everything from research ideas to the detailed aspects of writing and clear communication. I would like to extend additional thanks to Shuheng for supporting me through difficult times and being a role model in terms of character and persistence. I also enjoyed getting to know Shuheng's wonderful family.

I would like to thank Professors Doug Richstone and Al Hero on my thesis committee for insightful discussion of future research directions.

I would like to thank my friend and colleague Roger Fan. I thoroughly enjoyed working with him on joint mean and covariance estimation. Chapters 2 and 3 of the thesis are joint work with Roger Fan, Kerby Shedden, and Shuheng Zhou. I would like to thank Professor David Ruppert for serving as editor during the review process in the Journal of the American Statistical Association. Chapter 4 is joint work with Shuheng Zhou and Kerby Shedden. I would like to acknowledge research funding from the NSF under Grant DMS-1316731.

I would like to thank Seyoung Park, Byoung Jang, Joey Dickens, and Yuekai Sun for interesting research discussions and for contributing to a lively intellectual environment. I would like to thank Adam Hall, Karen Nielsen, Josh Errickson, Nhat Ho,

Teal Guidici, and Wyliena Guan for interesting discussions of research and statistics.

I would like to thank Brad Efron and Susan Holmes for undergraduate advising. I would like to thank Deanna Needell for supervising my undergraduate research and introducing me to mathematical and statistical research; her mentorship was instrumental. I would like to thank Helen Tombropoulos for inviting me to departmental undergraduate pizza parties and helping me throughout my undergraduate years.

I would like to thank my friends Adam Hall and Ruffa Arguelles. Spending time with them was one of my great joys in Michigan, and attending their wedding was one of the highlights of my time here.

I would like to thank Shirley, Thomas, Carlos, and Monchie for being my family in Michigan. I would like to thank Jeremy Brightbill for inspiring me to pursue math, and encouraging me throughout my time in the PhD program. I would like to thank Jeffrey Spiro, Daniel Shifren, Chris Marten, Nico Clayton, and Daniel Teplitz for lifelong friendship and good times. I would like to thank Sari Spiro, Randy Spiro, and Barbara Shifren for being my second family in Los Angeles. I would like to thank my teachers Mrs.. Haenschke, Mr. Vriesman, Mr. Piligian, Mr. Rutschman, Mr. Laderman, Mr. Lieberman, Mr. Davisson, Mrs. Rogers, and Mr. Monarch for instilling a love of learning and providing an academic foundation. I would like to thank Jaclyn for her friendship.

I would like to thank my parents Rona and Bruce for our many joyous memories and phone conversations, as well as my grandparents Max, Jennie, Bobbie, and Marv, Uncle Steve, Aunt Bryna, cousins Ethan and Ella, Aunt Paula, Uncle Richard, and cousins Ivan and Michael.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

This dissertation addresses theory, methodology, and applications for joint mean and covariance estimation with matrix-variate data. Chapters 2 and 3 consider joint mean and covariance estimation in the Kronecker product model, which has natural methodological connections to large-scale screening and differential mean analysis in various application areas including genomics. It has been proposed that complex populations, such as those that arise in genomics studies, may exhibit dependencies among observations as well as among variables. This gives rise to the challenging problem of analyzing unreplicated high-dimensional data with unknown mean and dependence structures. Matrix-variate approaches that impose various forms of (inverse) covariance sparsity allow flexible dependence structures to be estimated, but cannot directly be applied when the mean and covariance matrices are estimated jointly. We present a practical method utilizing generalized least squares and penalized (inverse) covariance estimation to address this challenge. We establish consistency and obtain rates of convergence for estimating the mean parameters and covariance matrices. The advantages of our approaches are: (i) dependence graphs and covariance structures can be estimated in the presence of unknown mean structure, (ii) the mean structure becomes more efficiently estimated when accounting for the dependence structure among observations; and (iii) inferences about the mean parameters become correctly calibrated. We use simulation studies and analysis of genomic data from a twin study of ulcerative colitis to illustrate the statistical convergence and the performance of our methods in practical settings. Several lines of evidence show that the test statistics for differential gene expression produced by our

methods are correctly calibrated and improve power over conventional methods.

Chapter 4 uses matrix-variate techniques to gain insight into pitch curve data that plays an important role in linguistics research. These curves can be viewed as large multi-indexed data arrays with distinct covariance behaviors along each index. We estimate covariance and inverse covariance matrices and graphs, and we connect edge structures to word properties.

CHAPTER I

Introduction

1.1 Matrix-variate data

In the setting of matrix-variate data, correlations exist between both rows (observations) and columns (variables) of a data matrix (*Efron*, 2009; *Allen and Tibshirani*, 2012). Data with correlations along multiple axes exists in a broad range of research fields, including environmental statistics (spatial and temporal correlations), neuroscience (correlations among experimental trials, neurons, and time), and genomics (correlations between people and genes). Such correlations affect both the accuracy and calibration of inferences, resulting in under- or over-estimates of standard errors (*Allen and Tibshirani*, 2012). We focus on the problem of jointly estimating mean and covariance structures while accounting for such correlations.

In Chapters 2 and 3, we consider data in which the covariance matrix of each column is proportional to a common matrix B. This allows information to be pooled across the columns in order to estimate B. We present algorithms for estimation and inference in this setting, including associated theory on rates of convergence of mean and covariance parameters. A special case of this model is the Kronecker product model, in which correlations between entries of the data matrix are decomposed into factors that depend on rows and factors that depend on columns. Our method builds on the Gemini estimator introduced by *Zhou* (2014a), which estimates covariance

matrices when both rows and columns of the data matrix are dependent. In the setting where correlations exist along only one axis of the array, researchers have proposed various covariance estimators and studied their theoretical and numerical properties (*Banerjee et al.*, 2008; *Fan et al.*, 2009; *Friedman et al.*, 2008; *Lam and Fan*, 2009; *Meinshausen and Bühlmann*, 2006; *Peng et al.*, 2009; *Ravikumar et al.*, 2011; *Rothman et al.*, 2008; *Yuan and Lin*, 2007; *Zhou et al.*, 2010; *Ren et al.*, 2015). We build on this work to jointly estimate mean and covariance parameters for matrix-variate data. For matrix-variate data with two way dependencies, e.g., in the space-time data, prior work depended on a large number of replicates to obtain certain convergence guarantees, see for example *Dutilleul* (1999), *Werner et al.* (2008) and *Tsiligkaridis et al.* (2013).

In Chapter 4, we investigate a tensor modeling framework which accounts for mean and trial specific variations in a large scale linguistic data, where non-i.i.d. replicates are available. In particular, we analyze linguistics pitch curve data using a Kronecker product covariance model while allowing individual mean matrices. The goals are to examine word-word and time-time correlation matrices, inverse correlation matrices, and associated graphical models. By contrast with the previous chapters, the pitch curve data contains a limited number of replicates, which allows us to use a novel trial differencing idea to remove the complex mean matrices. We investigate whether edges are associated with characteristics of the words, including initial consonant, vowel type, and voicing using rigorous statistical methods to be introduced in Section 1.2.1 and 1.2.2. In particular, we hierarchically decompose the words by consonants and/or by vowels while analyzing edges between individual words as well as word groups categorized by initial consonant or vowel properties.

In Chapter 5, we discuss future work. One direction for future work is to consider hypothesis testing for edges in linguistics pitch curve data, as well as cross-validation for model selection. Another direction is to apply additive covariance models to

pitch curve data, including the Kronecker sum model for the precision matrix of the vectorized data matrix. The precision matrix is sparser than in the case of the Kronecker product model, and the graph (for normally distributed data) has a Cartesian product structure, which has a simple interpretation. Prior work on optimization algorithms for the Kronecker sum model of the precision matrix includes the Biraphical Lasso (*Kalaitzis et al.*, 2013) and Tensor Graphical Lasso (*Greenwald et al.*, 2017). Another avenue for future work is to apply the decorrelation procedure proposed by *Zhou* (2014a), in which we use the estimated time-time inverse correlation matrix to decorrelate the data along the time axis, with the aim of improving the estimate of the word-word covariance and inverse covariance matrices.

1.2 Organization of the thesis

- In Chapter 2, we present two algorithms for joint mean and covariance estimation in the setting of matrix-variate data. We assess the performance of the algorithms using simulations, and we apply the algorithms to data arising from a genomic study of ulcerative colitis in twin pairs.

- In Chapter 3, we present theoretical results for the algorithms defined in Chapter 2. We prove rates of convergence of the estimated mean and covariance parameters.

- In Chapter 4, we analyze linguistics pitch curve data with trial replicates.

- In Chapter 5, we discuss future work, including cross-validation and applying additional matrix-variate methods to linguistics pitch curve data.

Chapters 2 and 3 were accepted for publication in the Journal of the American Statistical Association (*Hornstein et al.*, 2018). We aim to send chapter 4 to NIPS this May. With all future work which entails further analysis using cross validation,

permutation and another linguistics dataset in Chapter 5, we aim to eventually send the paper to a journal.

In the remaining two subsections of the introduction, we introduce matrix-variate graphical modeling and nodewise regression.

1.2.1 Matrix-variate graphical modeling

Graphical modeling plays a key role in the thesis, in particular in Chapter 4. Consequently, we now provide a definition of matrix-variate graphical models. The following paragraphs in this subsection defining matrix-variate graphical modeling are quoted verbatim from *Zhou* (2014a).

First recall the following definition concerning the classical Gaussian graphical model for a random vector.

Definition 1.2.1. Let $V = (V_1, \ldots, V_n)^T$ be a random Gaussian vector, which we represent by an undirected graph $G = (\mathcal{V}, F)$. The vertex set $\mathcal{V} := \{1, \ldots, n\}$ has one vertex for each component of the vector V. The edge set F consists of pairs (j, k) that are joined by an edge. If V_j is independent of V_k given the other variables, then $(j, k) \notin F$.

Now let $\mathcal{V} = \{1, \ldots, n\}$ be an index set which enumerates rows of X according to a fixed order. For all $i = 1, \ldots, m$, we assign to each variable of a column vector x^i exactly one element of the set $\mathcal{V}$ by a rule of correspondence $g : x^i \to \mathcal{V}$ such that $g(x^i_j) = j, j = 1, \ldots, n$. The graphs $G_i(\mathcal{V}, F)$ constructed for each random column vector $x^i, i = 1, \ldots, m$ according to Definition 1.2.1 will share an identical edge set F, because the normalized column vectors $x^1/\sqrt{a_{11}}, \ldots, x^m/\sqrt{a_{mm}}$ follow the same multivariate normal distribution $\mathcal{N}_n(0, B)$. Hence, graphs $G_1, \ldots, G_m$ are isomorphic and we write $G_i \simeq G_j, \forall i, j$. Due to the isomorphism, we use $G(\mathcal{V}, F)$ to represent the family of graphs $G_1, \ldots, G_m$. Hence, a pair (ℓ, k) which is absent in F encodes conditional independence between the ℓth row and the kth row given all

other rows. Similarly, let $\Gamma = \{1, \ldots, m\}$ be the index set which enumerates columns of X according to a fixed order. We use $H(\Gamma, E)$ to represent the family of graphs $H_1, \ldots, H_n$, where H_i is constructed for row vector y^i, and $H_i \simeq H_j, \forall i, j$. Now $H(\Gamma, E)$ is a graph with adjacency matrix $\Upsilon(H) = \Upsilon(A^{-1})$ as edges in E encode nonzeros in A^{-1}. And $G(\mathcal{V}, F)$ is a graph with adjacency matrix $\Upsilon(G) = \Upsilon(B^{-1})$. The Kronecker product, $H \otimes G$, is defined as the graph with adjacency matrix $\Upsilon(H) \otimes \Upsilon(G)$ (*Weichsel*, 1962), where clearly missing edges correspond to zeros in the inverse covariance $A^{-1} \otimes B^{-1}$, and $H \otimes G$ represents the graph of the p-variate Gaussian random vector vec$\{X\}$, where $p = mn$.

1.2.2 Nodewise regression

In addition to using Glasso, we also estimate edges using nodewise regression. *Meinshausen and Bühlmann* (2006) proposed variable selection via nodewise regression, in which each variable is regressed on each other variable via ℓ_1 penalized regression. The edges correspond to the nonzero entries of the regression coefficients (i.e. an edge exists between vertices i and j if either the regression coefficient of variable i on j is nonzero, or the regression coefficient of variable j on i is nonzero). *Meinshausen and Bühlmann* (2006) proved variable selection consistency of nodewise regression.

We now explain nodewise regression in more detail. Let $\widetilde{X} \in \mathbb{R}^{n \times m}$ denote a centered and scaled data matrix,so that the sample correlation matrix $\widehat{\Gamma} \in \mathbb{R}^{m \times m}$ can be expressed as

$$\widehat{\Gamma} = \frac{1}{n} \widetilde{X}^T \widetilde{X}. \tag{1.1}$$

Let $\widehat{\Gamma}^{(i)} \in \mathbb{R}^{(m-1) \times (m-1)}$ denote the submatrix of $\widehat{\Gamma}$ obtained by excluding the ith column and ith row. Let $\widehat{\gamma}^{(i)}$ denote the ith column of $\widehat{\Gamma}$ excluding the diagonal entry. The regression coefficient for the ith variable is obtained by solving the ℓ_1

penalized least squares problem,

$$\widehat{\beta}^i = \arg\min_{\beta:\beta\in\mathbb{R}^{m-1}} \left\{ \frac{1}{2}\beta^T\widehat{\Gamma}^{(i)}\beta - \langle\widehat{\gamma}^{(i)},\beta\rangle + \lambda\|\beta\|_1 \right\}. \tag{1.2}$$

Afterwards, the inverse correlation matrix is reconstructed by first obtaining a matrix $\widetilde{\Theta}$,

$$\widetilde{\Theta}_{-j,-j} = -(\widehat{\Gamma}_{jj} - \widehat{\Gamma}_{j,-j}\widehat{\beta}^j)^{-1}\widehat{\beta}^j, \quad \text{and} \quad \widetilde{\Theta}_{jj} = (\widehat{\Gamma}_{jj} - \widehat{\Gamma}_{j,-j}\widehat{\beta}^j)^{-1}, \tag{1.3}$$

then projecting $\widetilde{\Theta}$ onto the space of symmetric matrices.

Using nodewise regression with a refit to obtain an estimate of the inverse covariance matrix was proposed by *Yuan* (2010); *Loh and Wainwright* (2012). In *Zhou et al.* (2011), they combine a multiple regression approach with ideas of thresholding and refitting: first they infer a sparse undirected graphical model structure via thresholding of each among many ℓ_1-norm penalized regression functions of (1.2); they then estimate the covariance matrix and its inverse using the maximum likelihood estimator. They show that under suitable conditions, this approach yields consistent estimation in terms of graphical structure and fast convergence rates with respect to the operator and Frobenius norm for the covariance matrix and its inverse. Finally, they also derive an explicit bound for the Kullback Leibler divergence.

In the present work, our nodewise regression with thresholding procedure follows from ideas of *Zhou et al.* (2011) and *Zhou* (2010); in future work, we plan to further exploit the MLE refit procedure using the model (edge set) obtained through nodewise regression in combination with thresholding. See also *Dempster* (1972); *Zhou* (2010).

Since our input matrix is positive semidefinite, the methods of *Loh and Wainwright* (2012), *Yuan* (2010), and *Zhou et al.* (2011) would all work to obtain Θ.

CHAPTER II

Joint mean and covariance estimation of matrix-variate data

This chapter is joint work with Roger Fan, Kerby Shedden, and Shuheng Zhou.

2.1 Introduction

Understanding how changes in gene expression are related to changes in biological state is one of the fundamental tasks in genomics research, and is a prototypical example of "large scale inference" (*Efron*, 2010). While some genomics datasets have within-subject replicates or other known clustering factors that could lead to dependence among observations, most are viewed as population cross-sections or convenience samples, and are usually analyzed by taking observations (biological samples) to be statistically independent of each other. Countering this conventional view, *Efron* (2009) proposed that there may be unanticipated correlations between samples even when the study design would not suggest it. To identify and adjust for unanticipated sample-wise correlations, *Efron* (2009) proposed an empirical Bayes approach utilizing the sample moments of the data. In particular, sample-wise correlations may lead to inflated evidence for mean differences, and could be one explanation for

the claimed lack of reproducibility in genomics research (*Leek et al.*, 2010; *Allen and Tibshirani*, 2012; *Sugden et al.*, 2013).

A persistent problem in genomics research is that test statistics for mean parameters (e.g. t-statistics for two-group comparisons) often appear to be incorrectly calibrated (*Efron*, 2005; *Allen and Tibshirani*, 2012). When this happens, for example when test statistics are uniformly overdispersed relative to their intended reference distribution, this is usually taken to be an indication of miscalibration, rather than reflecting a nearly global pattern of differential effects (*Efron*, 2007). Adjustments such as genomic control (*Devlin and Roeder*, 1999) can be used to account for this; a related approach is that of *Allen and Tibshirani* (2012). In this work we address unanticipated sample-wise dependence, which can exhibit a strong effect on statistical inference. We propose a new method to jointly estimate the mean and covariance with a single instance of the data matrix, as is common in genetics. The basic idea of our approach is to alternate for a fixed number of steps between mean and covariance estimation. We exploit recent developments in two-way covariance estimation for matrix-variate data (*Zhou*, 2014a). We crucially combine the classical idea of generalized least squares (GLS) (*Aitken*, 1936) with thresholding for model selection and estimation of the mean parameter vector. Finally, we use Wald-type statistics to conduct inference. We motivate this approach using differential expression analysis in a genomics context, but the method is broadly applicable to matrix-variate data having unknown mean and covariance structures, with or without replications. We illustrate, using theory and data examples, including a genomic study of ulcerative colitis, that estimating and accounting for the sample-wise dependence can systematically improve the calibration of test statistics, therefore reducing or eliminating the need for certain post-hoc adjustments.

With regard to variable selection, another major challenge we face is that variables (e.g. genes or mRNA transcripts) have a complex dependency structure that

exists together with any dependencies among observations. As pointed out by *Efron* (2009) and others, the presence of correlations among the samples makes it more difficult to estimate correlations among variables, and vice versa. A second major challenge is that due to dependence among both observations and variables, there is no independent replication in the data, that is, we have a single matrix to conduct covariance estimation along both axes. This challenge is addressed in *Zhou* (2014a) when the mean structure is taken to be zero. A third major challenge that is unique to our framework is that covariance structures can only be estimated after removing the mean structure, a fact that is generally not considered in most work on high dimensional covariance and graph estimation, where the population mean is taken to be zero. We elaborate on this challenge next.

2.1.1 Our approach and contributions

Two obvious approaches for removing the mean structure in our setting are to globally center each column of the data matrix (containing the data for one variable), or to center each column separately within each group of sample points to be compared (subsequently referred to as "group centering"). Globally centering each column, by ignoring the mean structure, may result in an estimated covariance matrix that is not consistent. Group centering all genes, by contrast, leads to consistent covariance estimation, as shown in Theorem II.3 with regard to Algorithm 1. However, group centering all genes introduces extraneous noise when the true vector of mean differences is sparse. We find that there is a complex interplay between the mean and covariance estimation tasks, such that overly flexible modeling of the mean structure can introduce large systematic errors in the mean structure estimation. To mitigate this effect, we aim to center the data using a model selection strategy. More specifically, we adopt a model selection centering approach in which only mean parameters having a sufficiently large effect size (relative to the dimension of the data) are tar-

geted for removal. This refined approach has theoretical guarantees and performs well in simulations. The estimated covariance matrix can be used in uncertainty assessment and formal testing of mean parameters, thereby improving calibration of the inference.

In Section 2.2, we define the two group mean model, which is commonly used in the genomics literature, and introduce the GLS algorithm in this context. We bound the statistical error for estimating each column of the mean matrix using the GLS procedure so long as each column of X shares the same covariance matrix B, for which we have a close approximation. It is commonly known that genes are correlated, so correlations exist across columns as well as rows of the data matrix. In particular, in Theorem II.1 in Section 2.3.1, we establish consistency for the GLS estimator given a deterministic $\widehat{B}$ which is close to B in the operator norm, and present the rate of convergence for mean estimation for data generated according to a subgaussian model to be defined in Definition 2.2.1. Moreover, we do not impose a separable covariance model in the sense of (2.1).

What distinguishes our model from those commonly used in the genomics literature is that we do not require that individuals are independent. Our approach to covariance modeling builds on the Gemini method (*Zhou*, 2014a), which is designed to estimate a separable covariance matrix for data with two-way dependencies. For matrices $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{n\times n}$, the Kronecker product $A \otimes B \in \mathbb{R}^{mn\times nm}$ is the block matrix for which the (i,j)th block is $a_{ij}B$, for $i,j \in \{1,\ldots,m\}$. We say that an $n \times m$ random matrix X follows a matrix variate distribution with mean $M \in \mathbb{R}^{n\times m}$ and a separable covariance matrix

$$X_{n\times m} \sim \mathcal{L}_{n,m}(M, A_{m\times m} \otimes B_{n\times n}), \tag{2.1}$$

if $\mathrm{vec}\{\,\mathrm{X}\,\}$ has mean $\mathrm{vec}\{\,\mathrm{M}\,\}$ and covariance $\Sigma = A \otimes B$. Here $\mathrm{vec}\{\,\mathrm{X}\,\}$ is formed by

stacking the columns of X into a vector in $\mathbb{R}^{mn}$. For the mean matrix M, we focus on the two-group setting to be defined in (2.4). Intuitively, A describes the covariance between columns while B describes the covariance between rows of X. Even with perfect knowledge of M, we can only estimate A and B up to a scaling factor, as $A\eta \otimes \frac{1}{\eta}B = A \otimes B$ for any $\eta > 0$, and hence this will be our goal and precisely what we mean when we say we are interested in estimating covariances A and B. However, this lack of identifiability does not affect the GLS estimate, because the GLS estimate is invariant to rescaling the estimate of B^{-1}.

2.1.2 Related work

Efron (2009) introduced an approach for inference on mean differences in data with two-way dependence. His approach uses empirical Bayes ideas and tools from large scale inference, and also explores how challenging the problem of conducting inference on mean parameters is when there are uncharacterized dependences among samples. We combine GLS and variable selection with matrix-variate techniques. *Allen and Tibshirani* (2012) also consider this question and develop a different approach that uses ordinary least squares (OLS) through the iterations, first decorrelating the residuals and then using OLS techniques again on this adjusted dataset. The confounder adjustment literature in genomics, including *Sun et al.* (2012) and *Wang et al.* (2015), can also be used to perform large-scale mean comparisons in similar settings that include similarity structures among observations. These methods use the same general matrix decomposition framework, where the mean and noise are separated. They exploit low-rank structure in the mean matrix, as well as using sparse approximation of OLS estimates, for example where thresholding. Our model introduces row-wise dependence through matrix-variate noise, while the confounder adjustment literature instead assumes that a small number of latent factors also affect the mean expression, resulting in additional low-rank structure in the mean matrix. Section 2.9 contains

detailed comparisons between our approach and these alternative methods.

Our inference procedures are based on Z-scores and associated FDR values for mean comparisons of individual variables. While we account for sample-wise correlations, gene-gene correlations remain, which we regard as a nuisance parameter. Our estimated correlation matrix among the genes can be used in future work in combination with the line of work that addresses FDR in the presence of gene correlations. This relies on earlier work for false discovery rate estimation using correlated data, including *Owen* (2005); *Benjamini and Yekutieli* (2001); *Cai et al.* (2011); *Li and Zhong* (2014); *Benjamini and Hochberg* (1995); *Storey* (2003). Taking a different approach, *Hall et al.* (2010) develop the innovated higher criticism test statistics to detect differences in means in the presence of correlations between genes. Our estimated gene-gene correlation matrix can be used in combination with this approach; we leave this as future work. Another line of relevant research has focused on hypothesis testing of high-dimensional means, exploiting assumed sparsity of effects, and developing theoretical results using techniques from high dimensional estimation theory. Work of this type includes *Cai and Xia* (2014); *Chen et al.* (2014); *Bai and Saranadasa* (1996); *Chen et al.* (2010). *Hoff* (2011) adopts a Bayesian approach, using a model that is a generalization of the matrix-variate normal distribution.

Our method builds on the Gemini estimator introduced by *Zhou* (2014a), which estimates covariance matrices when both rows and columns of the data matrix are dependent. In the setting where correlations exist along only one axis of the array, researchers have proposed various covariance estimators and studied their theoretical and numerical properties (*Banerjee et al.*, 2008; *Fan et al.*, 2009; *Friedman et al.*, 2008; *Lam and Fan*, 2009; *Meinshausen and Bühlmann*, 2006; *Peng et al.*, 2009; *Ravikumar et al.*, 2011; *Rothman et al.*, 2008; *Yuan and Lin*, 2007; *Zhou et al.*, 2010; *Ren et al.*, 2015). Although we focus on the setting of Kronecker products, or separable covariance structures, *Cai et al.* (2016) proposed a covariance estimator for a model

with several populations, each of which may have a different variable-wise covariance matrix. Our methods can be generalized to this setting. *Tan and Witten* (2014) use a similar matrix-variate data setting as in (2.1), but perform biclustering instead of considering a regression problem with a known design matrix.

2.1.3 Notation and organization

Before we leave this section, we introduce the notation needed for the technical sections. Let $e_1, \ldots, e_p$ be the canonical basis of $\mathbb{R}^p$. For a matrix $A = (a_{ij})_{1 \leqslant i,j \leqslant m}$, let $|A|$ denote the determinant and $\mathrm{tr}(A)$ be the trace of A. Let $\|A\|_{\max} = \max_{i,j} |a_{ij}|$ denote the entry-wise max norm. Let $\|A\|_1 = \max_j \sum_{i=1}^m |a_{ij}|$ denote the matrix ℓ_1 norm. The Frobenius norm is given by $\|A\|_F^2 = \sum_i \sum_j a_{ij}^2$. Let $\varphi_i(A)$ denote the ith largest eigenvalue of A, with $\varphi_{\max}(A)$ and $\varphi_{\min}(A)$ denoting the largest and smallest eigenvalues, respectively. Let $\kappa(A)$ be the condition number for matrix A. Let $|A|_{1,\mathrm{off}} = \sum_{i \neq j} |a_{ij}|$ denote the sum of the absolute values of the off-diagonal entries and let $|A|_{0,\mathrm{off}}$ denote the number of non-zero off-diagonal entries. Let $a_{\max} = \max_i a_{ii}$. Denote by $r(A)$ the stable rank $\|A\|_F^2 / \|A\|_2^2$. We write $\mathrm{diag}(A)$ for a diagonal matrix with the same diagonal as A. Let I be the identity matrix. We let $C, C_1, c, c_1, \ldots$ be positive constants which may change from line to line. For two numbers a, b, $a \wedge b := \min(a, b)$ and $a \vee b := \max(a, b)$. Let $(a)_+ := a \vee 0$. For sequences $\{a_n\}, \{b_n\}$, we write $a_n = O(b_n)$ if $|a_n| \leqslant C|b_n|$ for some positive absolute constant C which is independent of n and m or sparsity parameters, and write $a_n \asymp b_n$ if $c|a_n| \leqslant |b_n| \leqslant C|a_n|$. We write $a_n = \Omega(b_n)$ if $|a_n| \geqslant C|b_n|$ for some positive absolute constant C which is independent of n and m or sparsity parameters. We write $a_n = o(b_n)$ if $\lim_{n \to \infty} a_n / b_n = 0$. For random variables X and Y, let $X \sim Y$ denote that X and Y follow the same distribution.

The remainder of the paper is organized as follows. In Section 2.2, we present our matrix-variate modeling framework and methods on joint mean and covariance esti-

mation. In particular, we propose two algorithms for testing mean differences based on two centering strategies. In Section 2.3, we present convergence rates for these methods. In Theorems II.3 and II.4, we provide joint rates of convergence for mean and covariance estimation using Algorithms 1 and 2, respectively. We also emphasize the importance of the design effect (c.f. equation (2.15)) in testing and present theoretical results for estimating this quantity in Corollary II.2 and Corollary II.5. In Section 2.4, we demonstrate through simulations that our algorithms can outperform OLS estimators in terms of accuracy and variable selection consistency. In Section 2.5, we analyze a gene expression dataset, and show that our method corrects test statistic overdispersion that is clearly present when using sample mean based methods (c.f. Section 2.4.2). Sections 2.6 and 2.7 contain additional simulation and data analysis results. We conclude in Section 2.8. Proofs are presented in Chapter 3. In Section 2.9 we provide additional comparisons between our method and some related methods on both simulated and real data.

2.2 Models and methods

In this section we present our model and method for joint mean and covariance estimation. Our results apply to subgaussian data. Before we present the model, we define subgaussian random vectors and the ψ_2 norm. The ψ_2 condition on a scalar random variable V is equivalent to the subgaussian tail decay of V, which means $P(|V| > t) \leqslant 2\exp(-t^2/c^2)$, for all $t > 0$. For a vector $y = (y_1, \ldots, y_p) \in \mathbb{R}^p$, denote by $\|y\|_2 = \sqrt{\sum_{i=1}^p y_i^2}$.

Definition 2.2.1. Let Y be a random vector in $\mathbb{R}^p$. (a) Y is called isotropic if for every $y \in \mathbb{R}^p$, $E[|\langle Y, y\rangle|^2] = \|y\|_2^2$. (b) Y is ψ_2 with a constant α if for every $y \in \mathbb{R}^p$,

$$\|\langle Y, y\rangle\|_{\psi_2} := \inf\{t : E[\exp(\langle Y, y\rangle^2/t^2)] \leqslant 2\} \leqslant \alpha\|y\|_2.$$

Our goal is to estimate the group mean vectors $\beta^{(1)}, \beta^{(2)}$, the vector of mean differences between two groups $\gamma = \beta^{(1)} - \beta^{(2)} \in \mathbb{R}^m$, the row-wise covariance matrix $B \in \mathbb{R}^{n \times n}$, and the column-wise covariance matrix $A \in \mathbb{R}^{m \times m}$. In our motivating genomics applications, the people by people covariance matrix B is often incorrectly anticipated to have a simple known structure, for example, B is taken to be diagonal if observations are assumed to be uncorrelated. However, we show by example in Section 2.5 that departures from the anticipated diagonal structure may occur, corroborating earlier claims of this type by *Efron* (2009) and others. Motivated by this example, we define the two-group mean model and the GLS algorithm, which takes advantage of the covariance matrix B.

The model. Our model for the matrix-variate data X can be expressed as a mean matrix plus a noise term,

$$X = M + \varepsilon, \tag{2.2}$$

where columns (and rows) of ε are subgaussian. Let $u, v, \in \mathbb{R}^n$ be defined as

$$u = (\underbrace{1, \ldots, 1}_{n_1}, \underbrace{0, \ldots, 0}_{n_2}) \in \mathbb{R}^n \quad \text{and} \quad v = (\underbrace{0, \ldots, 0}_{n_1}, \underbrace{1, \ldots, 1}_{n_2}) \in \mathbb{R}^n. \tag{2.3}$$

Let $\mathbf{1}_n \in \mathbb{R}^n$ denote a vector of ones. For the two-group model, we take the mean matrix to have the form

$$M = D\beta = \begin{bmatrix} \mathbf{1}_{n_1} \beta^{(1)T} \\ \mathbf{1}_{n_2} \beta^{(2)T} \end{bmatrix} \in \mathbb{R}^{n \times m}, \quad \text{where} \quad D = \begin{bmatrix} u & v \end{bmatrix} \in \mathbb{R}^{n \times 2} \tag{2.4}$$

is the design matrix and $\beta = (\beta^{(1)}, \beta^{(2)})^T \in \mathbb{R}^{2 \times m}$ is a matrix of group means. Let $\gamma = \beta^{(1)} - \beta^{(2)} \in \mathbb{R}^m$ denote the vector of mean differences. Let $d_0 = |\operatorname{supp}(\gamma)| = |\{j : \gamma_j \neq 0\}|$ denote the size of the support of γ. To estimate the group means, we

use a GLS estimator,

$$\widehat{\beta}(\widehat{B}^{-1}) := (D^T \widehat{B}^{-1} D)^{-1} D^T \widehat{B}^{-1} X \in \mathbb{R}^{2\times m}, \tag{2.5}$$

where $\widehat{B}^{-1}$ is an estimate of the observation-wise inverse covariance matrix. Throughout the paper, we denote by $\widehat{\beta}(B^{-1})$ the oracle GLS estimator, since it depends on the unknown true covariance B. Also, we denote the estimated vector of mean differences as $\widehat{\gamma}(\widehat{B}^{-1}) = \delta^T \widehat{\beta}(\widehat{B}^{-1}) \in \mathbb{R}^m$, where $\delta = (1, -1) \in \mathbb{R}^2$.

2.2.1 Matrix-variate covariance modeling

In the previous section, we have not yet explicitly constructed an estimator of B^{-1}. To address this need, we model the data matrix X with a matrix-variate distribution having a separable covariance matrix, namely, the covariance of $\mathrm{vec}\{\,\mathrm{X}\,\}$ follows a Kronecker product covariance model. When ε (2.2) follows a matrix-variate normal distribution $\mathcal{N}_{n,m}(0, A \otimes B)$, as considered in *Zhou* (2014a), the support of B^{-1} encodes conditional independence relationships between samples, and likewise, the support of A^{-1} encodes conditional independence relationships among genes. The inverse covariance matrices A^{-1} and B^{-1} have the same supports as their respective correlation matrices, so edges of the dependence graphs are identifiable under the model $\mathrm{Cov}(\mathrm{vec}(\varepsilon)) = A \otimes B$. When the data is subgaussian, the method is still valid for obtaining consistent estimators of A, B, and their inverses, but the interpretation in terms of conditional independence does not hold in general.

Our results do not assume normally distributed data; we analyze the subgaussian correspondent of the matrix variate normal model instead. In the Kronecker product covariance model we consider in the present work, the noise term has the form $\varepsilon = B^{1/2} Z A^{1/2}$ for a mean-zero random matrix Z with independent subgaussian entries satisfying $1 = \mathbb{E} Z_{ij}^2 \leqslant \|Z_{ij}\|_{\psi_2} \leqslant K$. Clearly, $\mathrm{vec}\{\,\varepsilon\,\} = \mathrm{A} \otimes \mathrm{B}$. Here, the matrix A

represents the shared covariance among variables for each sample, while B represents the covariance among observations which in turn is shared by all genes.

For identifiability, and convenience, we define

$$A^* = \frac{m}{\mathrm{tr}(A)}A \quad \text{and} \quad B^* = \frac{\mathrm{tr}(A)}{m}B, \tag{2.6}$$

where the scaling factor is chosen so that A^* has trace m. For the rest of the paper A and B refer to A^* and B^*, as defined in (2.6). Let S_A and S_B denote sample covariance matrices to be specified. Let the corresponding sample correlation matrices be defined as

$$\widehat{\Gamma}_{ij}(A) = \frac{(S_A)_{ij}}{\sqrt{(S_A)_{ii}(S_A)_{jj}}} \quad \text{and} \quad \widehat{\Gamma}_{ij}(B) = \frac{(S_B)_{ij}}{\sqrt{(S_B)_{ii}(S_B)_{jj}}}. \tag{2.7}$$

The baseline Gemini estimators (*Zhou*, 2014a) are defined as follows, using a pair of penalized estimators for the correlation matrices $\rho(A) = (a_{ij}/\sqrt{a_{ii}a_{jj}})$ and $\rho(B) = (b_{ij}/\sqrt{b_{ii}b_{jj}})$,

$$\widehat{A}_\rho = \underset{A_\rho \succ 0}{\arg\min} \left\{ \mathrm{tr}\left(\widehat{\Gamma}(A)A_\rho^{-1}\right) + \log|A_\rho| + \lambda_B |A_\rho^{-1}|_{1,\mathrm{off}} \right\}, \quad \text{and} \tag{2.8a}$$

$$\widehat{B}_\rho = \underset{B_\rho \succ 0}{\arg\min} \left\{ \mathrm{tr}\left(\widehat{\Gamma}(B)B_\rho^{-1}\right) + \log|B_\rho| + \lambda_A |B_\rho^{-1}|_{1,\mathrm{off}} \right\}, \tag{2.8b}$$

where the input are a pair of sample correlation matrices as defined in (2.7).

Let $\widehat{M}$ denote the estimator of the mean matrix M in (2.1). Denote the centered data matrix and the sample covariance matrices as

$$\begin{aligned} X_{\mathrm{cen}} &= X - \widehat{M}, \quad \text{for } \widehat{M} \text{ to be specified in Algorithms 1 and 2,} \\ S_B &= X_{\mathrm{cen}}X_{\mathrm{cen}}^T/m, \quad \text{and} \quad S_A = X_{\mathrm{cen}}^T X_{\mathrm{cen}}/n. \end{aligned} \tag{2.9}$$

Define the diagonal matrices of sample standard deviations as

$$\widehat{W}_1 = \sqrt{n}\mathrm{diag}(S_A)^{1/2} \in \mathbb{R}^{m\times m}, \quad \widehat{W}_2 = \sqrt{m}\mathrm{diag}(S_B)^{1/2} \in \mathbb{R}^{n\times n}, \tag{2.10}$$

$$\text{and } \widehat{A\otimes B} = \left(\widehat{W}_1\widehat{A}_\rho\widehat{W}_1\right)\otimes\left(\widehat{W}_2\widehat{B}_\rho\widehat{W}_2\right)/\|X_{\mathrm{cen}}\|_F^2. \tag{2.11}$$

2.2.2 Group based centering method

We now discuss our first method for estimation and inference with respect to the vector of mean differences $\gamma = \beta^{(1)} - \beta^{(2)}$, for $\beta^{(1)}$ and $\beta^{(2)}$ as in (2.4). Our approach in Algorithm 1 is to remove all possible mean effects by centering each variable within every group.

Algorithm 1: GLS-Global group centering

Input: X; and $\mathcal{G}(1), \mathcal{G}(2)$: indices of group one and two, respectively.
Output: $\widehat{A}^{-1}$, $\widehat{B}^{-1}$, $\widehat{A\otimes B}$, $\widehat{\beta}(\widehat{B}^{-1})$, $\widehat{\gamma}$, T_j for all j

---

1. Group center the data. Let Y_i denote the ith row of the data matrix. To estimate the group mean vectors $\beta^{(1)}, \beta^{(2)} \in \mathbb{R}^m$: Compute sample mean vectors

$$\widetilde{\beta}^{(1)} = \frac{1}{n_1}\sum_{i\in\mathcal{G}(1)} Y_i \quad \text{and} \quad \widetilde{\beta}^{(2)} = \frac{1}{n_2}\sum_{i\in\mathcal{G}(2)} Y_i; \quad \text{set} \quad \widehat{\gamma}^{\mathrm{OLS}} = \widetilde{\beta}^{(1)} - \widetilde{\beta}^{(2)} \tag{2.12}$$

$$\text{Center the data by} \quad X_{\mathrm{cen}} = X - \widehat{M}, \text{ with } \widehat{M} = \begin{bmatrix} 1_{n_1}\widetilde{\beta}^{(1)T} \\ 1_{n_2}\widetilde{\beta}^{(2)T} \end{bmatrix}.$$

2. Obtain regularized correlation estimates. (2a) The centered data matrix used to calculate S_A and S_B for Algorithm 1 is $X_{\mathrm{cen}} = (I - P_2)X$, where P_2 is

the projection matrix that performs within-group centering,

$$P_2 = \begin{bmatrix} n_1^{-1} 1_{n_1} 1_{n_1}^T & 0 \\ 0 & n_2^{-1} 1_{n_2} 1_{n_2}^T \end{bmatrix} = uu^T/n_1 + vv^T/n_2, \tag{2.13}$$

with u and v as defined in (2.3). Compute sample covariance matrices based on group-centered data: $S_A = \frac{1}{n} X_{\text{cen}}^T X_{\text{cen}} = \frac{1}{n} X^T (I - P_2) X$ and $S_B = \frac{1}{m} X_{\text{cen}} X_{\text{cen}}^T = \frac{1}{m} (I - P_2) X X^T (I - P_2)$.

(2b) Compute (2.7) to obtain penalized correlation matrices $\widehat{A}_\rho$ and $\widehat{B}_\rho$ using the Gemini estimators as defined in (2.8a) and (2.8b) with tuning parameters to be defined in (2.23).

3. Rescale the estimated correlation matrices to obtain penalized covariance

$$\widehat{B}^{-1} = m\widehat{W}_2^{-1} \widehat{B}_\rho \widehat{W}_2^{-1} \quad \text{and} \quad \widehat{A}^{-1} = (\|X_{\text{cen}}\|_F^2/m) \widehat{W}_1^{-1} \widehat{A}_\rho \widehat{W}_1^{-1}. \tag{2.14}$$

4. Estimate the group mean matrix using the GLS estimator as defined in (2.5).

5. Obtain test statistics. The jth test statistic is defined as

$$T_j = \frac{\widehat{\gamma}_j(\widehat{B}^{-1})}{\sqrt{\delta^T (D^T \widehat{B}^{-1} D)^{-1} \delta}}, \qquad \text{with } \delta = (1, -1) \in \mathbb{R}^2, \tag{2.15}$$

and $\widehat{\gamma}_j(\widehat{B}^{-1}) = \delta^T \widehat{\beta}_j(\widehat{B}^{-1})$, for $j = 1, \ldots, m$. Note that T_j as defined in (2.15) is essentially a Wald test and the denominator is a plug-in standard error of $\widehat{\gamma}_j(B^{-1})$.

2.2.3 Model selection centering method

In this section we present Algorithm 2, which aims to remove mean effects that are strong enough to have an impact on covariance estimation. The strategy here is to use a model selection step to identify variables with strong mean effects.

Algorithm 2: GLS-Model selection centering

Input: X, and $\mathcal{G}(1), \mathcal{G}(2)$: indices of group one and two, respectively.

Output: $\widehat{A}^{-1}$, $\widehat{B}^{-1}$, $\widehat{A \otimes B}$, $\widehat{\beta}(\widehat{B}^{-1})$, $\widehat{\gamma}$, T_j for all j

---

1. **Run Algorithm 1.** Use the group centering method to obtain initial estimates $\widehat{\gamma}_j^{\text{init}} = \widehat{\beta}_j^{(1)} - \widehat{\beta}_j^{(2)}$ for all $j = 1, \ldots, m$. Let $\widehat{B}_{\text{init}}^{-1}$ and $\widehat{B}_{\text{init}}$ be as obtained in (2.14).

2. **Select genes with large estimated differences in means.** Let $\widetilde{J}_0 = \{j : |\widehat{\gamma}_j^{\text{init}}| > 2\widehat{\tau}_{\text{init}}\}$ denote the set of genes which we consider as having strong mean effects, where

$$\widehat{\tau}_{\text{init}} \asymp \left(\frac{\log^{1/2} m}{\sqrt{m}} + \frac{\|\widehat{B}_{\text{init}}\|_1}{n_{\min}} \right) \sqrt{\frac{n_{\text{ratio}} |\widehat{B}_{\text{init}}^{-1}|_{0,\text{off}}}{n_{\min}}} + \sqrt{\log m} \|(D^T \widehat{B}_{\text{init}}^{-1} D)^{-1}\|_2^{1/2}, \tag{2.16}$$

   with $n_{\min} = n_1 \wedge n_2$, $n_{\max} = n_1 \vee n_2$, and $n_{\text{ratio}} = n_{\max}/n_{\min}$.

3. **Calculate Gram matrices based on model selection centering.** Global centering can be expressed in terms of the projection matrix $P_1 = n^{-1}\mathbf{1}_n\mathbf{1}_n^T$. Compute the centered data matrix

$$X_{\text{cen},j} = \begin{cases} X_j - P_2 X_j & \text{if } j \in \widetilde{J}_0 \\ X_j - P_1 X_j & \text{if } j \in \widetilde{J}_0^c, \end{cases}$$

where $X_{\text{cen},j}$ denotes the jth column of the centered data matrix X_{cen}. Compute the sample covariance and correlation matrices with X_{cen} following (2.9) and (2.7).

4. **Estimate covariances and means.** (4a) Obtain the penalized correlation matrices $\widehat{B}_\rho$ and $\widehat{A}_\rho$ using Gemini estimators as defined in (2.8a) and (2.8b) with tuning parameters of the same order as those in (2.23).

   (4b) Obtain inverse covariance estimates $\widehat{B}^{-1}$, $\widehat{A}^{-1}$ using (2.14).

   (4c) Calculate the GLS estimator $\widehat{\beta}(\widehat{B}^{-1})$ as in (2.5), as well as the vector of mean differences $\widehat{\gamma}(\widehat{B}^{-1}) = \delta^T \widehat{\beta}(\widehat{B}^{-1})$, for $\delta = (1, -1) \in \mathbb{R}^2$.

5. **Obtain test statistics.** Calculate test statistics as in (2.15), now using $\widehat{B}^{-1}$ as estimated in Step 4.

Remarks. In the case that γ is sparse, we show that this approach can perform better than the approach in Section 2.2.2, in particular when the sample size is small. We now consider the expression $\widehat{\tau}_{\text{init}}$ in (2.16) as an upper bound on the threshold in the sense that it is chosen to tightly control false positives. In Section 2.4.2 we show in simulations that with this plug-in estimate $\widehat{\tau}_{\text{init}}$, Algorithm 2 can nearly reach the performance of GLS with the true B. Since this choice of $\widehat{\tau}_{\text{init}}$ acts as an order on the threshold we need, the plug-in method can also be applied with a multiplier between 0 and 1. When we set $\widehat{\tau}_{\text{init}}$ at its lower bound, namely,

$$\sqrt{\log m}\|(D^T \widehat{B}_{\text{init}}^{-1} D)^{-1}\|_2^{1/2}, \quad \text{where } \widehat{B}_{\text{init}}^{-1} \text{ is obtained as in Step 3 from Algorithm 1,}$$

we anticipate many false positives. In Figure 2.3, we show that the performance of Algorithm 2 is stable in the setting of small n and sparse γ for different values of $\widehat{\tau}_{\text{init}}$, demonstrating robustness of our methods to the multiplier; there we observe that the performance can degrade if the threshold is set to be too small, eventually reaching

the performance of Algorithm 1.

Second, if an upper bound on the number of differentially expressed genes is known a priori, one can select a set of genes $\check{J}_0$ to group center such that the cardinality $|\check{J}_0|$ is understood to be chosen as an upper bound on $d_0 = |\operatorname{supp}(\gamma)|$ based on prior knowledge. We select the set $\check{J}_0$ by ranking the components of the estimated vector of mean differences $\widehat{\gamma}$. In the data analysis in Section 2.5 we adopt this strategy in an iterative manner by successively halving the number of selected genes, choosing at each step the genes with largest estimated mean differences from the previous step. We show in this data example and through simulation that the proposed method is robust to the choice of $|\check{J}_0|$.

Finally, it is worth noting that these algorithms readily generalize to settings with more than two groups, in which case we simply group center within each group. This is equivalent to applying the method with a different design matrix D. In fact, we can move beyond group-wise mean comparisons to a regression analysis with a fixed design matrix D, which includes the k-group mean analysis as a special case.

2.3 Theoretical results

We first state Theorem II.1, which provides the rate of convergence of the GLS estimator (2.5) when we use a fixed approximation of the covariance matrix B. We then provide in Theorems II.3 and II.4 the convergence rates for estimating the group mean matrix $\beta \in \mathbb{R}^{2\times m}$ for Algorithms 1 and 2 respectively. In Theorem II.3 we state rates of convergence for the Gemini estimators of B^{-1} and A^{-1} when the input sample covariance matrices use the group centering approach as defined in Algorithm 1, while in Theorem II.4, we state only the rate of convergence for estimating B^{-1}, anticipating that the rate for A^{-1} can be similarly obtained, using the model selection centering approach as defined in Algorithm 2.

2.3.1 GLS under fixed covariance approximation

We now state a theorem on the rate of convergence of the GLS estimator (2.5), where we use a fixed approximation $B_{n,m}^{-1}$ to B^{-1}, where the operator norm of $\Delta_{n,m} = B_{n,m}^{-1} - B^{-1}$ is small in the sense of (2.17). We will specialize Theorem II.1 to the case where B^{-1} is estimated using the baseline method in *Zhou* (2014a) when X follows subgaussian matrix-variate distribution as in (2.1). We prove Theorem II.1 in Section 3.2.

Theorem II.1. *Let Z be an $n \times m$ random matrix with independent entries Z_{ij} satisfying $\mathbb{E}Z_{ij} = 0$, $1 = \mathbb{E}Z_{ij}^2 \leqslant \|Z_{ij}\|_{\psi_2} \leqslant K$. Let $Z_1, \ldots, Z_m \in \mathbb{R}^n$ be the columns of Z. Suppose the jth column of the data matrix satisfies $X_j \sim B^{1/2}Z_j$. Suppose $B_{n,m} \in \mathbb{R}^{n\times n}$ is a positive definite symmetric matrix. Let $\Delta_{n,m} := B_{n,m}^{-1} - B^{-1}$. Suppose*

$$\|\Delta_{n,m}\|_2 < \frac{1}{(n_{\max}/n_{\min})\,\|B\|_2}, \text{ where } n_{\min} = n_1 \wedge n_2 \quad \text{and} \quad n_{\max} = n_1 \vee n_2. \tag{2.17}$$

Then with probability at least $1 - 8/(m \vee n)^2$, for some absolute constants C, C',

$$\forall j, \quad \|\widehat{\beta}_j(B_{n,m}^{-1}) - \beta_j^*\|_2 \;\leqslant\; r_{n,m} := s_{n,m} + t_{n,m}, \qquad \text{where} \tag{2.18}$$

$$s_{n,m} = C\sqrt{\log m \|B\|_2/n_{\min}} \quad \text{and} \quad t_{n,m} = C'\|\Delta_{n,m}\|_2/n_{\min}^{1/2}; \tag{2.19}$$

$$\text{and} \qquad \|\widehat{\gamma}(B_{n,m}) - \gamma\|_\infty \leqslant \sqrt{2}\left(C\sqrt{\frac{\log m\|B\|_2}{n_{\min}}} + C'n_{\min}^{-1/2}\|\Delta_{n,m}\|_2\right). \tag{2.20}$$

Remarks. If the operator norm of B is bounded, that is $\|B\|_2 < W$, then condition (2.17) is equivalent to $\|\Delta_{n,m}\|_2 < 1/(Wn_{\text{ratio}})$. The term $t_{n,m}$ in (2.19) reflects the error due to approximating B^{-1} with $B_{n,m}^{-1}$, whereas $s_{n,m}$ reflects the error in estimating the mean matrix (2.5) using GLS with the true B^{-1} for the random design X. The term $s_{n,m}$ is $O(\sqrt{\log m/n})$, whereas $t_{n,m}$ is $O(1/\sqrt{n})$. The dominating term $s_{n,m}$ in (2.19) can be replaced by the tighter bound, namely, $s'_{n,m} =$

$C' \log^{1/2}(m)\sqrt{\delta^T(D^T B^{-1} D)^{-1}\delta}$, with $\delta = (1, -1) \in \mathbb{R}^2$. This bound correctly drops the factor of $\|B\|_2$ present in (2.19) and (2.20), while revealing that variation aligned with the column space of D is especially important in mean estimation.

Note that the condition (2.17) is not stringent, and that the $\widehat{B}$ estimates used in Algorithms 1 and 2 have much lower errors than this. When $M = 0$ is known, S_A and S_B can be the usual Gram matrices, and the theory in *Zhou* (2014a) guarantees that $t_{n,m}$ as defined in (2.19) has rate $C_A\sqrt{\log m/m}$, with $C_A = \sqrt{m}\|A\|_F/\operatorname{tr}(A)$. However in our setting, M in general is nonzero. In Sections 2.2.2 and 2.2.3 we provide two constructions for S_A and S_B, which differ in how the data are centered. These constructions have a different bound $t_{n,m}$, as we will discuss in Theorems II.3 and II.4.

In Section 2.4, we present simulation results that demonstrate the advantage of the oracle GLS and GLS with estimated $\widehat{B}$ (2.5) over the sample mean based (OLS) method (c.f. (2.12) and (2.32)) for mean estimation as well as the related variable selection problem with respect to γ. There, we scrutinize this quantity and its estimation procedure in detail.

Design effect. The "design effect" is the variance of the "oracle" GLS estimator (2.5) of γ_j using the true B, that is,

$$\delta^T(D^T B^{-1} D)^{-1}\delta = \operatorname{Var}(\widehat{\gamma}_j(B^{-1})),\ \forall j = 1, \ldots, m. \tag{2.21}$$

The design effect reflects the potential improvement of GLS over OLS. It appears as a factor above in $s'_{n,m}$, so it contributes to the rate of mean parameter estimation as characterized in Theorem II.1. Lower variance in the GLS estimator of the mean difference contributes to greater power of the test statistics relative to OLS. The design effect also appears as a scale factor in the test statistics for $\widehat{\gamma}$ (2.15), and therefore it is particularly important that the design effect is accurately estimated in

order for the test statistics to be properly calibrated. In a study focusing on mean differences, it may be desirable to assess the sample size needed to detect a given effect size using our methodology. Given the design effect, our tests for differential expression are essentially Z-tests based on the GLS fits, followed by some form of multiple comparisons adjustment.

Corollary II.2. *Let* $\Omega = (D^T B^{-1} D)^{-1}$, $\widehat{\Omega} = (D^T \widehat{B}^{-1} D)^{-1}$, *and* $\Delta = \widehat{\Omega} - \Omega$. *Under the conditions of Theorem II.1, the relative error in estimating the design effect is bounded as*

$$\frac{\left|\delta^T \widehat{\Omega} \delta - \delta^T \Omega \delta\right|}{\delta^T \Omega \delta} \leqslant 2C' \frac{\kappa(B) \|B\|_2 \|\Delta\|_2}{n_{\text{ratio}}}, \tag{2.22}$$

with probability $1 - C/(m \vee n)^d$, *for some absolute constants* C, C'.

We prove Corollary II.2 in Section 3.2.2. Corollary II.2 implies that given an accurate estimator of B^{-1}, the design effect is accurately estimated and therefore suggests that traditional techniques can be used to gain an approximate understanding of the power of our methods. We show that B^{-1} can be accurately estimated under conditions in Theorems 3 and 4. If pilot data are available that are believed to have similar between-sample correlations to the data planned for collection in a future study, Corollary II.2 also justifies using this pilot data to estimate the design effect. If no pilot data are available, it is possible to conduct power analyses based on various plausible specifications for the B matrix.

2.3.2 Rates of convergence for Algorithms 1 and 2

We state the following assumptions.

(A1) The number of nonzero off-diagonal entries of A^{-1} and B^{-1} satisfy

$$\begin{aligned} \left|A^{-1}\right|_{0,\text{off}} &= o(n/\log(m \vee n)) \qquad (n, m \to \infty) \quad \text{and} \\ \left|B^{-1}\right|_{0,\text{off}} &= o\left([m/\log(m \vee n)] \vee \left[n^2/\|B\|_1^2\right]\right) \qquad (n, m \to \infty). \end{aligned}$$

(A2) The eigenvalues of A and B are bounded away from 0 and $+\infty$. We assume that the stable ranks satisfy $r(A), r(B) \geqslant 4\log(m \vee n)$, where $r(A) = \|A\|_F^2 / \|A\|_2^2$.

Theorem II.3. *Suppose that (A1) and (A2) hold. Consider the data as generated from model* (2.2) *with* $\varepsilon = B^{1/2} Z A^{1/2}$, *where* $A \in \mathbb{R}^{m\times m}$ *and* $B \in \mathbb{R}^{n\times n}$ *are positive definite matrices, and* Z *is an* $n \times m$ *random matrix as defined in Theorem II.1. Let* $C, C', C_1 C_2, C'', C'''$ *be some absolute constants. Let* $C_A = \sqrt{m}\|A\|_F / \operatorname{tr}(A)$ *and* $C_B = \sqrt{n}\|B\|_F / \operatorname{tr}(B)$. ***(I)*** *Let* λ_A *and* λ_B *denote the penalty parameters for* (2.8b) *and* (2.8a) *respectively. Suppose*

$$\lambda_A \geqslant C\left(C_A K \frac{\log^{1/2}(m\vee n)}{\sqrt{m}} + \frac{\|B\|_1}{n_{\min}}\right) \quad \text{and} \quad \lambda_B \geqslant C'\left(C_B K \frac{\log^{1/2}(m\vee n)}{\sqrt{n}} + \frac{\|B\|_1}{n_{\min}}\right) \tag{2.23}$$

Then with probability at least $1 - C''/(m\vee n)^2$, *for* $\widehat{A\otimes B}$ *as define in* (2.11),

$$\begin{aligned} &\|\widehat{A\otimes B} - A\otimes B\|_2 \leqslant \|A\|_2\|B\|_2\delta, \\ &\|\widehat{A\otimes B}^{-1} - A^{-1}\otimes B^{-1}\|_2 \leqslant \|A^{-1}\|_2\|B^{-1}\|_2\delta', \\ \textit{where} \qquad &\delta, \delta' = O\left(\lambda_A\sqrt{|B^{-1}|_{0,\mathit{off}}\vee 1} + \lambda_B\sqrt{|A^{-1}|_{0,\mathit{off}}\vee 1}\right). \end{aligned}$$

Furthermore, with probability at least $1 - C'''/(m\vee n)^2$,

$$\|\widehat{A\otimes B} - A\otimes B\|_F \leqslant \|A\|_F\|B\|_F\eta, \tag{2.24}$$

$$\textit{where} \qquad \eta = O\left(\lambda_A\sqrt{|B^{-1}|_{0,\mathit{off}}\vee n}/\sqrt{n} + \lambda_B\sqrt{|A^{-1}|_{0,\mathit{off}}\vee m}/\sqrt{m}\right). \tag{2.25}$$

The same conclusions hold for the inverse estimate, with η *being bounded in the same order as in* (2.25). ***(II)*** *Let* $\widehat{\beta}$ *be defined as in (2.5) with* $\widehat{B}^{-1}$ *being defined as in* (2.14) *and* D *as in (2.4). Then, with probability at least* $1 - C/m^d$ *the following holds*

for all j,

$$\|\widehat{\beta}_j(\widehat{B}^{-1}) - \beta_j^*\|_2 \leqslant C_1 \lambda_A \sqrt{\frac{n_{\text{ratio}}\left(|B^{-1}|_{0,off} \vee 1\right)}{n_{\min}}} + C_2 \sqrt{\log m} \|(D^T B^{-1} D)^{-1}\|_2^{1/2}. \tag{2.26}$$

We prove Theorem II.3 part I in Section 3.3; this relies on rates of convergence of $\widehat{B}^{-1}$ and $\widehat{A}^{-1}$ in the operator and the Frobenius norm, which are established in Lemma III.7. We prove part II in Section 3.3.2.

Remarks. We find that the additional complexity of estimating the mean matrix leads to an additional additive term of order $1/n$ appearing in the convergence rates for covariance estimation for B and A. In part I of Theorem II.3, λ_A is decomposed into two terms, one term reflecting the variance of S_B, and one term reflecting the bias due to group centering. The variance term goes to zero as m increases, and the bias term goes to zero as n increases. To analyze the error in the GLS estimator based on $\widehat{B}^{-1}$, we decompose $\|\widehat{\beta}_j(\widehat{B}^{-1}) - \beta_j^*\|_2$ as

$$\|\widehat{\beta}_j(\widehat{B}^{-1}) - \beta_j^*\|_2 \leqslant \|\widehat{\beta}_j(\widehat{B}^{-1}) - \widehat{\beta}_j(B^{-1})\|_2 + \|\widehat{\beta}_j(B^{-1}) - \beta_j^*\|_2,$$

where the first term is the error due to not knowing B^{-1}, and the second term is the error due to not knowing β_j^*. The rate of convergence given in (2.26) reflects this decomposition. For Algorithm 2, we have analogous rates of convergence for both mean and covariance estimation. Simulations suggest that the constants in the rates for Algorithm 2 are smaller than those in (2.26).

We state the following assumptions for Theorem II.4 to hold on Algorithm 2.

(A1') Suppose (A1) holds. Let the number of nonzero off-diagonal entries of B^{-1} satisfy

$|B^{-1}|_{0,\text{off}} \leqslant \max\left(m, \frac{n^2}{\|B\|_1^2}, n \log m\right)$.

(A2') Suppose (A2) holds, and $n \geqslant \log m \left(\|A\|_2 \|B\|_2 b_{\max}/C_A^2\right)$.

(A3) Let $\text{supp}(\gamma) = \{j : \gamma_j \neq 0\}$. Let $s = |\text{supp}(\gamma)|$ denote the sparsity of γ. Assume that $s = O\left(\frac{C_A}{\|B\|_2} n \sqrt{\frac{m}{\log m}}\right)$.

Remarks. When B is dense in the sense that $\|B\|_1 \asymp \sqrt{n}\,\|B\|_2$, the new condition $|B^{-1}|_{0,\text{off}} \leqslant n \log m$ is vacuous. Condition (A2') is mild, because the condition on the stable rank of B already implies that $n \geqslant \log m$.

Theorem II.4. *Suppose that (A1'), (A2'), and (A3) hold. Consider the data as generated from model* (2.4) *with* $\varepsilon = B^{1/2} Z A^{1/2}$, *where* $A \in \mathbb{R}^{m \times m}$ *and* $B \in \mathbb{R}^{n \times n}$ *are positive definite matrices, and* Z *is an* $n \times m$ *random matrix as defined in Theorem II.3. Let* λ_A *denote the penalty parameter for estimating* B. *Suppose* λ_A *is as defined in* (2.23). *Let*

$$\tau_{\text{init}} \asymp \sqrt{\log m} \|(D^T B^{-1} D)^{-1}\|_2^{1/2}. \tag{2.27}$$

Then with probability at least $1 - C''/(m \vee n)^2$, *for output of Algorithm 2,*

$$\left\| \text{tr}\,(A) \left(\widehat{W}_2 \widehat{B}_\rho \widehat{W}_2 \right)^{-1} - B^{-1} \right\|_2 \leqslant \frac{C' \lambda_A \sqrt{|B^{-1}|_{0,\text{off}} \vee 1}}{b_{\min} \varphi^2_{\min}(\rho(B))}, \quad \textit{and} \tag{2.28}$$

$$\|\widehat{\beta}_j(\widehat{B}^{-1}) - \beta_j^*\|_2 \leqslant C_2 \sqrt{\log m} \|(D^T B^{-1} D)^{-1}\|_2^{1/2}, \tag{2.29}$$

for all j, *for absolute constants* C, C_2, C', *and* C''.

We prove Theorem II.4 in Section 3.6.5. In Section 3.6.4 we also show a standalone result, namely Theorem III.21, for the case of fixed sets of group and globally centered genes. This result shows how the algorithm used in the preliminary step to choose which genes to group center can be decoupled from the rest of the estimation procedure in Algorithm 2, so long as certain conditions hold. The proof of Theorem II.4 indeed validates that such conditions hold for the output of Algorithm 1. It is worth noting that a similar rate of convergence for estimating A could also be derived, but we focus on B in our methodology and applications, and therefore leave this as an exercise for interested readers.

We specialize Corollary II.2 to the case where B^{-1} is estimated using Algorithm 2.

Corollary II.5. *Under the conditions of Theorem II.4, we have with probability* $1 - C/m^2$

$$\frac{\left|\delta^T \widehat{\Omega}\delta - \delta^T \Omega \delta\right|}{\delta^T \Omega \delta} \leqslant 2C' \frac{n_{\text{ratio}}}{\lambda_{\min}(B)} \kappa(B) \lambda_A \sqrt{|B^{-1}|_{0,\text{off}} \vee 1}, \tag{2.30}$$

for some absolute constants C *and* C'.

Remarks. The right-hand-side of (2.30) goes to zero because of the assumptions (A1'), (A2'), and (A3), which ensure that the factor $\lambda_A \sqrt{|B^{-1}|_{0,\text{off}} \vee 1}$ goes to zero. We conduct simulations to assess the accuracy of estimating the design effect in Section 2.4.2.

2.4 Simulations

We present simulations to compare Algorithms 1 and 2 to both sample mean based analysis and oracle algorithms that use knowledge of the true correlation structures A and B. We show these results for a variety of population structures and sample sizes. We construct covariance matrices for A and B from one of:

- AR1(ρ) model. The covariance matrix is of the form $B = \{\rho^{|i-j|}\}_{i,j}$, and the graph corresponding to B^{-1} is a chain.

- Star-Block model. The covariance matrix is block-diagonal with equal-sized blocks whose inverses correspond to star structured graphs, where $B_{ii} = 1$, for all i. In each subgraph, a central hub node connects to all other nodes in the subgraph, with no additional edges. The covariance matrix for each block S in B is generated as in *Ravikumar et al.* (2011): $S_{ij} = \rho = 0.5$ if $(i, j) \in E$ and $S_{ij} = \rho^2$ otherwise.

- Erdős-Rényi model. We use the random concentration matrix model in *Zhou et al.* (2010). The graph is generated according to a type of Erdős-Rényi random graph. Initially we set $B^{-1} = 0.25 I_{n\times n}$. Then, we randomly select d edges and update B^{-1} as follows: for each new edge (i, j), a weight $w > 0$ is chosen uniformly at random from $[w_{\min}, w_{\max}]$ where $w_{\min} = 0.6$ and $w_{\max} = 0.8$; we subtract w from B^{-1}_{ij} and B^{-1}_{ji}, and increase B^{-1}_{ii} and B^{-1}_{jj} by w. This keeps B^{-1} positive definite. We then rescale so that B^{-1} is an inverse correlation matrix.

2.4.1 Accuracy of $\widehat{\gamma}$ and its implication for variable ranking

Table 2.1 displays metrics that reflect how the choice of different population structures B can affect the difficulty of the mean and covariance estimation problems. Column 2 is a measure discussed by *Efron* (2007). Column 3 appears directly in the theoretical analysis, reflecting the entry-wise error in the sample correlation $\widehat{\Gamma}(B)$. Columns 4 analogously reflects the entry-wise error for the Flip-Flop procedure in *Zhou* (2014a), and is included here for completeness. Column 5 displays the value of $\sqrt{\delta^T (D^T B^{-1} D)^{-1} \delta}$, where $\delta = (1, -1) \in \mathbb{R}^2$, which represents the standard deviation of the difference in means estimated using GLS with the true B^{-1}. Column 6 displays what we call the standard deviation ratio, namely

$$\sqrt{\frac{u^T B u}{\delta^T (D^T B^{-1} D)^{-1} \delta}}, \tag{2.31}$$

where $u = (\underbrace{1/n_1, \ldots, 1/n_1}_{n_1}, \underbrace{-1/n_2, \ldots, -1/n_2}_{n_2}) \in \mathbb{R}^n$ and $\delta = (1, -1) \in \mathbb{R}^2$, which reflects the potential efficiency gain for GLS over sample mean based method (2.12) for estimating γ. Note that the standard deviation ratio depends on the relationship between the covariance matrix B and the design matrix D. In Table 2.1, the first $n/2$ individuals are in group one, and the following $n/2$ are in group two. The values in Column 6 show that substantial improvement is possible in mean estimation. For

	B	ρ_B^2	$\|B\|_F/\mathrm{tr}(B)$	$\|\rho(B)^{-1}\|_{1,\mathrm{off}}$	sd GLS	sd ratio
$n = 80$						
1	AR1(0.2)	0.00	0.12	32.92	0.27	1.00
2	AR1(0.4)	0.00	0.13	75.24	0.33	1.02
3	AR1(0.6)	0.01	0.16	148.12	0.40	1.07
4	AR1(0.8)	0.04	0.24	351.11	0.46	1.32
5	StarBlock(4, 20)	0.02	0.18	101.33	0.35	1.51
6	ER(0.6, 0.8)	0.01	0.14	92.75	0.17	1.21
$n = 40$						
1	AR1(0.2)	0.00	0.16	16.25	0.38	1.01
2	AR1(0.4)	0.01	0.19	37.14	0.45	1.03
3	AR1(0.6)	0.03	0.23	73.12	0.53	1.12
4	AR1(0.8)	0.08	0.33	173.33	0.53	1.47
5	StarBlock(2, 20)	0.04	0.25	50.67	0.50	1.51
6	ER(0.6, 0.8)	0.02	0.21	47.24	0.25	1.23

Table 2.1: Assessment of the difficulty of estimating B^{-1} and the potential gain from GLS. The total correlation ρ_B is the average squared off-diagonal value of the correlation matrix $\rho(B)$. The fourth column is the design effect as defined in (2.21). The last column (sd ratio) presents the ratio of the standard deviation of the difference in sample means in (2.12) to the standard deviation of the GLS estimator of the difference in means. The first three columns of the table reflect the difficulty of estimating B, whereas the last two columns reflect the potential improvement of GLS over the sample mean based method (2.12). In the notation StarBlock(a, b), a refers to the number of blocks, and b refers to the block size.

an AR1 covariance matrix, the standard deviation ratio increases as the AR1 parameter increases; as the correlations get stronger, the potential improvement in mean estimation due to GLS grows. For the Star Block model with fixed block size, the standard deviation ratio is stable as n increases.

In Figure 2.1, we use ROC curves to illustrate the sensitivity and specificity for variable selection in the sense of how well we can identify the support for $\{i : \gamma_i \neq 0\}$ when we threshold $\widehat{\gamma}_i$ at various values. To evaluate and compare different methods, we let $\widehat{\gamma}$ be the output of Algorithm 1, Algorithm 2, the oracle GLS, and the sample mean based method (2.12). These correspond to the four curves on each plot of the top two rows of plots. We find that Algorithm 1 and Algorithm 2 perform better than the sample mean based method (2.12), and in some cases perform comparably to the

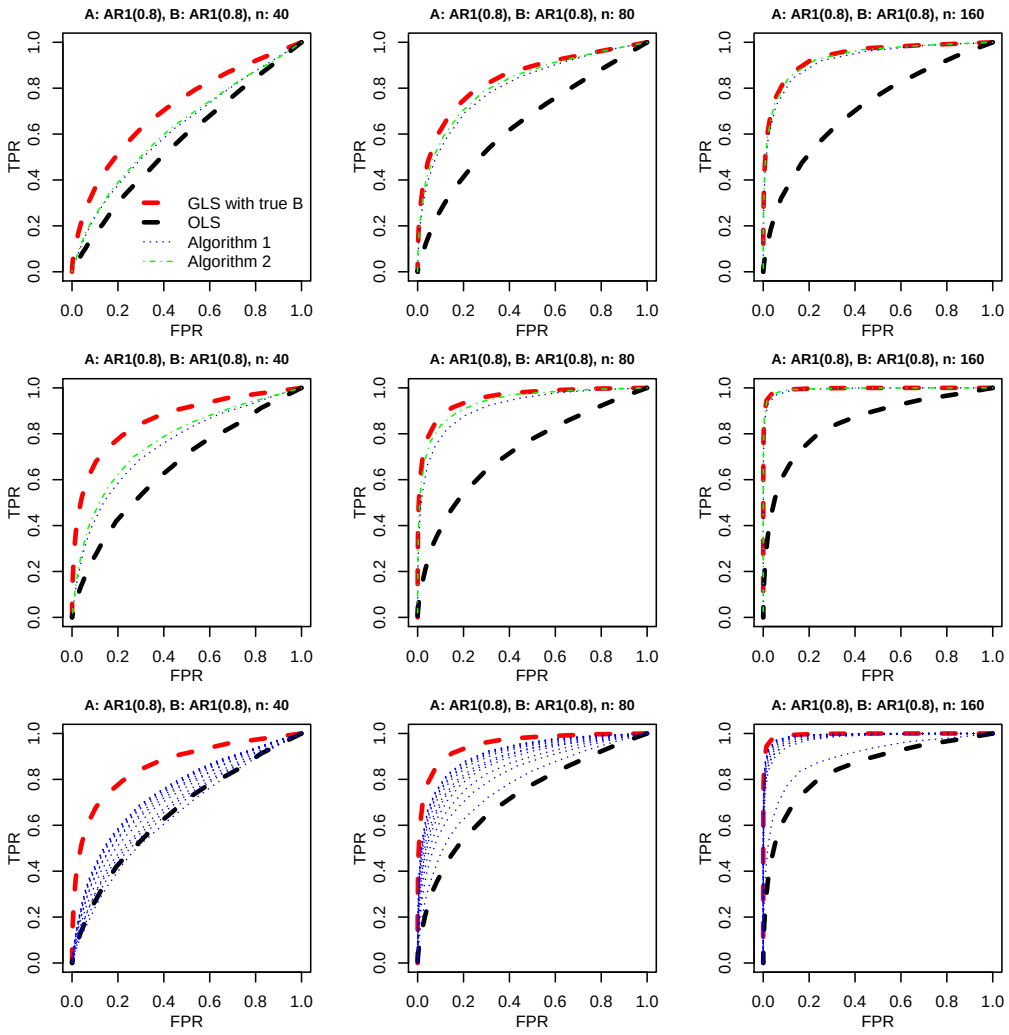

Figure 2.1: ROC curves. For each plot, the horizontal axis is false positive rate (FPR) and the vertical axis is true positive rate (TPR), as we vary a threshold for classifying variables as null or non-null. The covariance matrices A and B are both AR1 with parameter 0.8, with $m = 2000$ and $n = 40$, 80, and 160 in column one, two, and three, respectively. Ten variables in γ have nonzero entries. On each trial, the group labels are randomly assigned, with equal sample sizes. The marginal variance of each entry of the data matrix is equal to one. For the first row of plots, the magnitude of each nonzero entry of γ is 0.2, and for the second and third rows of plots, the magnitude of each nonzero entry of γ is 0.3. In the first two rows we display ROC curves for Algorithms 1 and 2 with penalty parameters chosen to maximize area under the curve. The third row displays an ROC curves for Algorithm 1, sweeping out penalty parameters.

oracle GLS. Plots in the third row of Figure 2.1 illustrate the sensitivity of Algorithm 1 to the choice of the graphical lasso (GLasso) penalty parameter (2.23); the simulations

are run using the `glasso` R package (*Friedman et al.*, 2008) to estimate B via (2.8b). The performance can degenerate to that of the sample mean based method (2.12), if the penalty is too high.

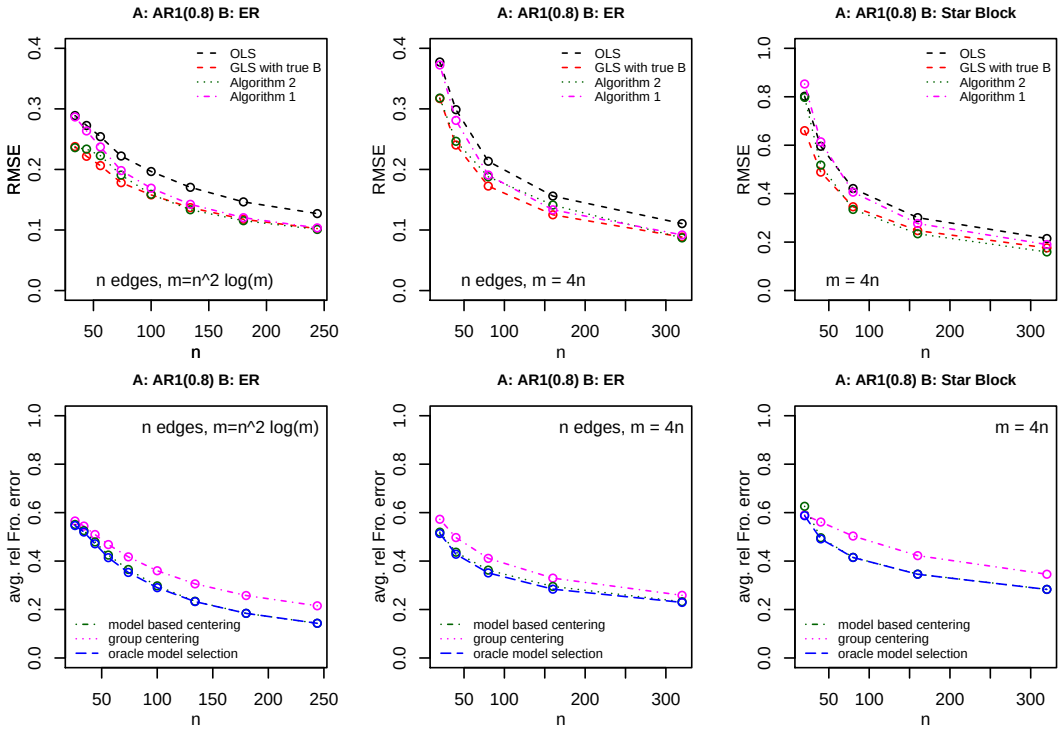

Figure 2.2: Performance of centering methods as n and m are varied, with n shown on the horizontal axis. In the first column of plots, the number of edges is proportional to $\sqrt{m/\log(m)}$. In the second and third columns of plots, the number of edges is proportional to m. In the first two columns of plots, B^{-1} is an Erdős-Rényi inverse covariance matrix. In the third column, B^{-1} is star block with blocks of size 10. The first row of plots shows RMSE for estimating γ, whereas the second row shows average relative Frobenius error in estimating B^{-1}. All panels are based on 250 simulation replications.

In the top row of Figure 2.2 we plot the root mean squared error (RMSE) when estimating the mean differences γ for Algorithm 1, Algorithm 2, OLS (i.e. sample means) and the oracle GLS estimate. The population structures for B are Erdős-Rényi and Star Block. Both Algorithms 1 and 2 consistently outperform the sample mean based method (2.12) for mean estimation, and Algorithm 2 even achieves comparable performance to the oracle GLS in some settings. The bottom row displays the

relative Frobenius error for estimating B^{-1}. Algorithm 2 outperforms Algorithm 1 in terms of covariance estimation and is comparable to oracle model selection, which only centers the columns with a true mean difference.

In Figure 2.3, we illustrate that Algorithm 2 can perform well using a plug-in estimator $\widehat{\tau}_{\text{init}}$ as in (2.16). We compare the methods when the true mean structure is a decaying exponential; we display the correlation of the ranks of the entries of γ to the ranks of the estimates of γ. Algorithm 2 with a plugin estimator $\widehat{\tau}_{\text{init}}$ can nearly reach the performance of GLS with the true B. Furthermore, the plug-in version of Algorithm 2 also consistently outperforms Algorithm 1. We also assess sensitivity to the choice of threshold: the curve labeled "Algorithm 2" uses the plug-in estimate $\widehat{\tau}_{\text{init}}$, whereas "Algorithm 2 with threshold multiplier" uses a plug-in estimate of the lower bound given in (2.27) in Theorem II.4. These two-plug in estimators exhibit similar performance, showing robustness of Algorithm 2 to the choice of the threshold parameter. In real data analysis, we validate this further. For the top row (AR1), the ratio of thresholds (2.27) to (2.16) is 0.75, and for the bottom row (UC), the ratio is 0.17.

In Section 2.9, we perform additional simulations to compare Algorithm 2 to two similar methods using ROC curves, namely, the sphering method of *Allen and Tibshirani* (2012), which uses a matrix-variate model similar to ours, and the confounder adjustment method of *Wang et al.* (2015), which uses a latent factor model. Our simulations show that Algorithm 2 consistently outperforms these competing methods in a variety of simulation settings using matrix-variate data.

2.4.2 Inference for the mean difference $\widehat{\gamma}$

Two basic approaches to conducting inference for mean differences are paired and unpaired t statistics. The unpaired t statistic is defined as follows. Let $X = (X_{ij})$.

Then the jth unpaired t statistic is

$$\begin{aligned} T_j &= \left(\widetilde{\beta}_j^{(1)} - \widetilde{\beta}_j^{(2)}\right) \widehat{\sigma}_j^{-1} (n_1^{-1} + n_2^{-1})^{-1/2}, \text{ where} \\ \widehat{\sigma}_j^2 &= (n_1 + n_2 - 2)^{-1} \sum_{k=1}^{2} \sum_{i \in \mathcal{G}_k} \left(X_{ij} - \widetilde{\beta}_j^{(k)}\right)^2, \end{aligned} \tag{2.32}$$

where $\widetilde{\beta}_j^{(k)}$, $k = 1, 2$, and $j = 1, \ldots, m$, denotes the sample mean of group k and variable j as defined in (2.12), and $\mathcal{G}_k$ is the set of indices corresponding to group k. When there is a natural basis for pairing the observations, and paired units are anticipated to be positively correlated, we can calculate paired t statistics. For the paired t statistic, suppose observations i and $i' = i + n/2$ are paired, for $i \in \{1, \ldots, n/2\}$. Note that samples can always be permuted so as to be paired in this way. Define the paired differences $d_{ij} = X_{ij} - X_{i'j}$, for $i \in \{1, \ldots, n/2\}$. Then the paired t statistic is $\overline{d}_j (n/2 - 1)^{1/2} / \left(\sum_{i=1}^{n/2} (d_{ij} - \overline{d}_j)^2\right)^{1/2}$, where $\overline{d}_j = (n/2)^{-1} \sum_{i=1}^{n/2} d_{ij}$.

Figure 2.4 considers estimation of the "design effect" $\delta^T (D^T B^{-1} D)^{-1} \delta$, as previously defined in (2.21), with $\delta = (1, -1)^T$. The importance of this object is discussed in Sections 2.3.1 and 2.3.2. The design effect is estimated via $\delta^T (D^T \widehat{B}^{-1} D)^{-1} \delta$, with $\widehat{B}^{-1}$ from Algorithm 1 or 2. The GLasso penalty parameters are chosen as

$$\lambda_A = f_A \left(C_A K \frac{\log^{1/2}(m \vee n)}{\sqrt{m}} + \frac{\|B\|_1}{n_{\min}} \right) \tag{2.33}$$

where we sweep over the factor f_A, referred to as the penalty multiplier. Figure 2.4 displays boxplots of the ratio $\delta^T (D^T \widehat{B}^{-1} D)^{-1} \delta / \delta^T (D^T B^{-1} D)^{-1} \delta$ over 250 replications for each setting of the penalty multiplier f_A. In Figure 2.4, B^{-1} follows the Erdős-Rényi model, and A is AR1(0.8), with $m = 2000$, and $n = 40$ and 80. Figure 2.4 shows that Algorithm 2 (plots B and D) estimates the design effect to high accuracy and is quite insensitive to the penalty multiplier as long as it is less than 1, as predicted by the theoretical analysis. Algorithm 1 also estimates the design effect

with high accuracy, but with somewhat greater sensitivity to the tuning parameter. The best penalty parameter for Algorithm 1 is around 0.1, whereas reasonable penalty parameters for Algorithm 2 are in the range 0.01 to 0.1. This is consistent with smaller entrywise error in the sample covariance for model selection centering than for group centering.

We next compare the results from Algorithm 2 to results obtained using paired and unpaired t statistics. Figure 2.5 illustrates the calibration and power of plug-in Z-scores, $\widehat{\gamma}_j/\widehat{\mathrm{SE}}(\widehat{\gamma}_j)$ derived from Algorithm 2 for three population settings. The standard error is calculated as $\sqrt{\delta^T(D^T\widehat{B}^{-1}D)^{-1}\delta}$, with $\delta = (1,-1)$. In the first and second plots, the data was simulated from AR1(0.8) and Erdős-Rényi, respectively. In the third plot, the data was simulated from $\widehat{B}$ for ulcerative colitis data described in Section 2.5. To obtain $\widehat{B}$, we apply Algorithm 2 to the ulcerative colitis data, using a Glasso penalty of $\lambda \approx 0.5[(\log(m)/m) + 3/n]$ in step 1, followed by group centering the top ten genes in step 2, and using a Glasso penalty of $\lambda \approx 0.1[(\log(m)/m) + 3/n]$ in step 4. In all cases A is AR1(0.8). In each case, we introduce 10 variables with different population means in the two groups, by setting $\gamma = 0.8$ for those variables, with the remaining γ values equal to zero. The ideal Q-Q plot would follow the diagonal except at the upper end of the range, as do our plug-in GLS test statistics. The t statistics (ignoring dependence) are seen to be overly dispersed throughout the range, and are less sensitive to the real effects.

2.4.3 Covariance estimation for A

Figure 2.6 shows the relative Frobenius error in estimating A^{-1} as n grows, for fixed m. The horizontal axis is $n/(d\log(m))$, scaled so that the curves align, where d is the maximum node degree. Because $\|A^{-1}\|_F$ is of order $\sqrt{m}$, the vertical axis essentially displays $\|\widehat{A}^{-1} - A^{-1}\|_F/\sqrt{m}$. For estimating A^{-1}, the rate of convergence is of order $\sqrt{\log(m)/n}$. For each of the three population structures, accuracy increases

with respect to n.

2.5 Genomic study of ulcerative colitis

Ulcerative colitis (UC) is a chronic form of inflammatory bowel disease (IBD), resulting from inappropriate immune cell infiltration of the colon. As part of an effort to better understand the molecular pathology of UC, *Lepage et al.* (2011) reported on a study of mRNA expression in biopsy samples of the colon mucosal epithelium, with the aim of being able to identify gene transcripts that are differentially expressed between people with UC and healthy controls. The study subjects were discordant identical twins, that is, monozygotic twins such that one twin has UC and the other does not. This allows us to simultaneously explore dependences among samples (both within and between twins), dependences among genes, and mean differences between the UC and non-UC subjects. The data set is available on the Gene Expression Omnibus, GEO accession GDS4519 (*Edgar et al.*, 2002).

The data consist of 10 discordant twin pairs, for a total of 20 subjects. Each subject's biopsy sample was assayed for mRNA expression, using the Affymetrix UG 133 Plus 2.0 array, which has 54,675 distinct transcripts. Previous analyses of this data did not consider twin correlations or unanticipated non-twin correlations, and used very different methodology (e.g. Wilcoxon testing). Roughly 70 genes were found to be differentially expressed (*Lepage et al.*, 2011).

We applied our Algorithm 2 to the UC genomics data as follows. First we selected the 2000 most variable genes based on marginal variance and then rescaled each gene to have unit marginal variance. We then applied step 1 of Algorithm 2, setting $\lambda = 0.1 \approx 0.5\left(\sqrt{\frac{\log(m)}{m}} + \frac{3}{n}\right)$, with $m = 2000$ and $n = 20$. For step 2 of the algorithm, we ranked the estimated mean differences, group centered the top ten, and globally centered the remaining genes. We then re-calculated the Gram matrix S_B using the centered data. In step 3, following the Gemini approach, we applied

the GLasso to S_B using a regularization parameter $\lambda \approx 0.25(\sqrt{\log(m)/m} + 3/n)$. We obtain estimated differences in means and test statistics via steps 4 through 6. A natural analysis of these data using more standard methods would be a paired t-test for each mRNA transcript (paired by twin pair). Such an approach is optimized for the situation where there is a constant level of correlation within all of the twin pairs, with no non-twin correlations. However as in Efron (2008), we wish to accommodate unexpected correlations, which in this case would be correlations between non-twin subjects or a lack of correlation between twin subjects. Our approach, developed in Section 2.2, does not require pre-specification or parameterization of the dependence structure, thus we were able to consider twin and non-twin correlations simultaneously. Lepage et al. note that UC has lower heritability than other forms of IBD. If UC has a relatively stronger environmental component, this could explain the pattern of correlations that we uncovered, as shown in Figure 2.7. The samples are ordered so that twins are adjacent, corresponding to 2 by 2 diagonal blocks. The penalized inverse sample correlation matrix contains nonzero entries both within twin pairs and between twin pairs.

To also handle these unexpected non-twin correlations, we performed testing using Algorithm 2. We found only a small amount of evidence for differential gene expression between the UC and non-UC subjects. Four of the adjusted p-values fell below a threshold of 0.1, using the Benjamini-Hochberg adjustment; that is, four genes satisfied $2000\widehat{p}_{(i)}/i < 0.1$, where $\widehat{p}_{(i)}$ is the i^{th} order statistic of the p-values calculated using Algorithm 2, for $i = 1, \ldots, 2000$. Based on our theoretical and simulation work showing that our procedure can successfully recover and accommodate dependence among samples, we argue that this is a more meaningful representation of the evidence in the data for differential expression compared to methods that do not adapt to dependence among samples. Specifically, in Section 2.5.1 we demonstrate that our test statistics are properly calibrated and as a result have weaker (but

more accurate) evidence for differential expression results. Below we argue that the sample-wise correlations detected by our approach would be expected to artificially inflate the evidence for differential expression.

2.5.1 Calibration of test statistics

As noted above, based on the test statistics produced by Algorithm 2, we find evidence for only a small number of genes being differentially expressed. This conclusion, however, depends on the test statistics conforming to the claimed null distribution whenever the group-wise means are equal. In this section, we consider this issue in more detail.

The first plot of Figure 2.8 compares the empirical quantiles of $\Phi^{-1}(T_j)$ to the corresponding quantiles of a standard normal distribution, where Φ is the standard normal cdf and the T_js are as defined in (2.32). Plots 2 and 3 show the same information for successive non-overlapping blocks of two thousand genes sorted by marginal variance. Since this is a discordant twins study, we also show results for the standard paired t statistics, pairing by twin. In all cases, the paired and unpaired statistics are more dispersed relative to the reference distribution. By contrast, the central portion of the GLS test statistics coincide with the reference line. Overdispersion of test statistics throughout their range is often taken to be evidence of miscalibration (*Devlin and Roeder*, 1999). In this setting the GLS statistics are calibrated correctly under the null hypothesis, but the paired and unpaired t statistics are not.

2.5.2 Stability of gene sets

The motivation of our Algorithm 2 is that in many practical settings a relatively small fraction of variables may have differential means, and therefore it is advantageous to avoid centering variables presenting no evidence of a strong mean difference. Here we assess the stability of the estimated mean differences as we vary the number

of group centered genes in Algorithm 2. To do so, we successively group center fewer genes, globally centering the remaining genes.

The iterative process is as follows. Let $\widehat{B}^{-1}_{(i)} \in \mathbb{R}^{n\times n}$ denote the estimate of B^{-1} at iteration i, let $\widehat{\beta}_{(i)} \in \mathbb{R}^{2\times m}$ denote the estimates of the group means β on the ith iteration, let $\widehat{\gamma}_{(i)} \in \mathbb{R}^m$ denote the vector of differences in group means between the two groups, and let $\widehat{\mu}_{(i)} \in \mathbb{R}^m$ denote vector of global mean estimates. Let $\widehat{\mu}(B^{-1}) \in \mathbb{R}^m$ denote the result of applying GLS with design matrix $D = 1_n$ to estimate the global means.

Initialize $\widehat{\beta}_{(1)}$, $\widehat{\mu}_{(1)}$ and $\widehat{\gamma}_{(1)}$ using the sample means. On the ith iteration,

1. Rank the genes according to $|\widehat{\gamma}_{(i-1)}|$. Center the highest ranked n_i' genes around $\widehat{\beta}_{(i-1)}$. Center the remaining genes around $\widehat{\mu}_{(i-1)}$.

2. Obtain $\widehat{B}^{-1}_{(i)}$ by applying GLasso to the centered data matrix from step 1.

3. Set $\widehat{\beta}_{(i)} = \widehat{\beta}(\widehat{B}^{-1}_{(i)})$, $\widehat{\mu}_{(i)} = \widehat{\mu}(\widehat{B}^{-1}_{(i)})$, and $\widehat{\gamma}_{(i)} = (1, -1)\widehat{\beta}_{(i)}$.

We assess the stability of the mean estimates by comparing the rankings of the genes across iterations of the algorithm. Table 2.2 displays the number of genes in common out of the top ten genes on each pair of iterations of the algorithm. For example, three genes ranked in the top ten on the first iteration of the algorithm are also ranked in the top ten on the last iteration. Iterations six through nine produce the same ranking of the top ten genes. Three genes are ranked among the top ten on every iteration of the algorithm: DPP10-AS1, OLFM4, and PTN. Table 2.4 shows simulations confirming these results.

2.5.3 Stability analysis

Table 2.3 shows the number of genes that fall below an FDR threshold of 0.1 on each iteration, for several values of the GLasso penalty λ. The number of genes below the threshold is more sensitive to the number of group-centered genes than to

Table 2.2: Each iteration k of the algorithm produces a ranking of all 2000 genes. For the top ten genes on each iteration, entry (i, j) of the table shows the number of genes in common in iterations i and j of the algorithm. Note that the maximum possible value for any entry of the table is 10; if entry (i, j) is 10, then iterations i and j selected the same top ten genes.

	1	2	3	4	5	6	7	8	9
1	10	10	7	5	5	3	3	3	3
2	10	10	7	5	5	3	3	3	3
3	7	7	10	6	5	3	3	3	3
4	5	5	6	10	8	5	5	5	5
5	5	5	5	8	10	7	7	7	7
6	3	3	3	5	7	10	10	10	10
7	3	3	3	5	7	10	10	10	10
8	3	3	3	5	7	10	10	10	10
9	3	3	3	5	7	10	10	10	10

Table 2.3: For the algorithm, this table shows the number of genes that are significant at an FDR level of 0.1 on each iteration of the algorithm, for different values of the GLasso penalty λ. The top row shows the number of genes group centered on each iteration.

n.group	2000	1024	512	256	128	64	32	16	8
$\lambda = 0.1$	1006	913	327	14	3	1	1	1	1
$\lambda = 0.2$	865	806	262	2	1	1	1	1	0
$\lambda = 0.3$	778	789	303	3	1	1	0	0	0
$\lambda = 0.4$	706	774	452	3	1	0	0	0	0
$\lambda = 0.6$	657	751	587	19	1	1	0	0	0
$\lambda = 0.8$	628	699	493	30	1	1	1	1	1

the GLasso penalty parameter. This is consistent with the first plot of Figure 2.10 where the design effect (in the denominator of the test statistics) is likewise more sensitive to the number of group centered genes than to the GLasso penalty. When fewer than 128 genes are group centered, the number of genes below an FDR threshold of 0.1 is stable across the penalty parameters from $\lambda = 0.1$ to $\lambda = 0.8$.

2.6 Additional simulation results

Figure 2.9 demonstrates the effect of mean structure on covariance estimation. As expected, when there is no mean structure Gemini performs competitively. As

more mean structure is added, however, its performance quickly decays to be worse than Algorithm 2. This also provides evidence that the plug-in estimator $\widehat{\tau}_{\text{init}}$ used in Algorithm 2 is appropriately selecting genes to group center, as when there are no or very few differentially expressed genes Algorithm 2 is still never worse than Gemini. Algorithm 1 does not perform as well as Algorithm 2 but still tends to eventually outperform Gemini as more mean structure is added. As the sample size increases, the difference between Algorithm 2 and Algorithm 1 decreases as the added noise from group centering becomes less of a factor. We still recommend using Algorithm 2 in most realistic scenarios, but this reinforces our theoretical finding that the two algorithms have the same error rates.

2.7 Additional data analysis

As discussed in Section 2.3.1, it is particularly important that the design effect is accurately estimated in order for the test statistics to be properly calibrated. The first plot of Figure 2.10 displays the sensitivity of the estimated design effect (2.21) for Algorithm 2 to the GLasso penalty parameter and the number of group centered columns. In the case that all columns are group centered, Algorithm 2 reduces to Algorithm 1. If we group center all genes, the estimated design effect is sensitive to the penalty parameter, but if we group center a small proportion of genes, it is less sensitive to the penalty parameter. This is further evidence that it may be advantageous to avoid over-centering the data when the true mean difference vector γ may be sparse. The second plot of Figure 2.10 shows a quantile plot comparing the distribution of test statistics from the UC data to test statistics from a simulation whose population correlation structure is matched to the UC data. The quantile plot reveals that we can reproduce the pattern of overdispersion in the test statistics using simulated data having person-person as well as gene-gene correlations. Such correlations therefore provide a possible explanation for the overdispersion of the test

statistics.

Figure 2.11 displays a quantile plot and inverse covariance graph for $\lambda = 0.4$ and 128 group centered genes. Under these settings the test statistics appear correctly calibrated, coinciding with the central portion of the reference line. Furthermore, the inverse covariance graph is sparse (38 edges). In the inverse covariance graph, there are more edges between subjects with UC than between the healthy subjects, which could be explained by the existence of subtypes of UC inducing correlations between subsets of subjects. The third plot of Figure 2.11 displays a sparser inverse covariance graph, corresponding to a larger penalty $\lambda = 0.5$. There are three edges between twin pairs, and there are more edges between subjects with UC than between those without UC.

2.7.1 Stability simulation

Table 2.4 shows the results from a simulation analogous to Table 2.2, demonstrating stability across iterations of the procedure. Iteration 1 begins by group centering 1280 genes and this number is halved in each successive iteration. We can see from the table that the gene rankings generated by Algorithm 2 are robust to misspecifying the number of differentially expressed genes. When the number of group centered genes is 160 or below (iterations 4 through 8), the commonly selected genes among the top 20 genes are stable. Furthermore, the true positives remain stable as we decrease the amount of genes centered, while the false positives decrease.

2.8 Conclusion

It has long been known that heteroscedasticity and dependence between observations impacts the precision and degree of uncertainty for estimates of mean values and regression coefficients. Further, data that are modeled for convenience as being independent observations may in fact show unanticipated dependence (*Kruskal*, 1988).

Table 2.4: Number of genes in common among genes ranked in the top 20 when different numbers of genes are group centered. This simulation is analogous to Table 2.2. Note that the maximum possible value for any entry of the table is 20; if entry (i, j) is 20, then iterations i and j selected the same top twenty genes. The first 10 genes have a difference of 1.5 and the second 10 have a difference of 1. All remaining genes have a true mean difference of zero. We use B as estimated from the UC data, and A is from an AR1(0.8) model. These simulations have $n = 20$ individuals and 2000 genes and are averaged over 200 replications. The last two rows display the average number of true and false positives among the genes ranked in the top 20 of each iteration.

	1	2	3	4	5	6	7	8
1	20.0	17.6	15.8	14.8	14.3	14.0	14.0	13.9
2	17.6	20.0	17.9	16.8	16.2	15.9	15.8	15.8
3	15.8	17.9	20.0	18.7	18.1	17.8	17.7	17.6
4	14.8	16.8	18.7	20.0	19.3	19.0	18.9	18.8
5	14.3	16.2	18.1	19.3	20.0	19.6	19.5	19.4
6	14.0	15.9	17.8	19.0	19.6	20.0	19.8	19.7
7	14.0	15.8	17.7	18.9	19.5	19.8	20.0	19.8
8	13.9	15.8	17.6	18.8	19.4	19.7	19.8	20.0
TP	12.7	14.3	15.6	16.4	16.7	16.8	16.8	16.8
FP	7.3	5.7	4.4	3.6	3.3	3.2	3.2	3.2

This has motivated the development of numerous statistical methods, including generalized/weighted least squares (GLS/WLS), mixed effect models, and generalized estimating equations (GEE). Our approach utilizes recent advances in high dimensional statistics to permit estimation of an inter-observation dependence structure (reflected in the matrix B in our model). Like GLS/GEE, we use an approach that alternates between mean and covariance estimation, but limit it in Algorithm 1 to a mean estimation step, followed by a covariance update, followed by a mean update, with an additional covariance and mean update if Algorithm 2 is used. We provide convergence guarantees and rates for both algorithms.

Estimation of dependence or covariance structures usually requires some form of replication, and/or strong models. We require a relatively weak form of replication and a relatively weak model. In our framework, the dependence among observations must be common (up to proportionality) across a set of "quasi-replicates" (the

columns of X, or the genes in our UC example). These quasi-replicates may be statistically dependent, and may have different means. We also require the precision matrices for the dependence structures to be sparse, which is a commonly used condition in recent high-dimensional analyses.

In addition to providing theoretical guarantees, we also show through simulations and a genomic data analysis that the approach improves estimation accuracy for the mean structure, and appears to mitigate test statistic overdispersion, leading to test statistics that do not require post-hoc correction. The latter observation suggests that undetected dependence among observations may be one reason that genomic analyses are sometimes less reproducible than traditional statistical methods would suggest, an observation made previously by *Efron* (2009) and others.

Although our theoretical analysis guarantees the convergence of our procedure even with a single observation of the random matrix X, there are reasons to expect this estimation problem to be fundamentally challenging. One reason for this as pointed out by *Efron* (2009) and subsequently explored by *Zhou* (2014a), is that the row-wise and column-wise dependence structures are somewhat non-orthogonal, in that row-dependence can "leak" into the estimates of column-wise dependence, and vice-versa. Our results suggest that while row-wise correlations make it more difficult to estimate column-wise correlations (and vice-versa), when the emphasis is on mean structure estimation, even a somewhat rough estimate of the dependence structure (B) can substantially improve estimation and inference.

We provide additional simulation and data analysis results in Section 2.6 and 2.7. We state some preliminary results and notation in Section 3.1. We prove Theorem II.1 in Section 3.2 and Corollary II.2 in Section 3.2.2. We prove Theorem II.3 in Section 3.3, with additional lemmas proved in Section 3.4. We prove entrywise convergence of the sample correlation matrices for Algorithm 1 in Section 3.5. We prove Theorem II.4 in Section 3.6, and we prove additional lemmas used in the proof of Theorem II.4

in Section 3.7. In Section 2.9 we provide additional comparisons between our method and some related methods on both simulated and real data.

2.9 Comparisons to related methods

The most similar existing method to ours is the sphering approach from *Allen and Tibshirani* (2012). Both methods use a preliminary demeaned version of the data to generate covariance estimates, then use these estimates to adjust the gene-wise t-tests. The largest difference between the procedures lies in this last step. The sphering approach produces an adjusted data set based on decorrelating residuals from a preliminary mean estimate and performs testing and mean estimation on this adjusted data using traditional OLS techniques. Though their approach is well-motivated at the population level, they do not provide theoretical support for their plug-in procedure, and in particular do not explore how noise in the initial mean estimate may complicate their decorrelation procedure. In contrast, our approach uses a generalized least squares approach motivated by classical statistical results including the Gauss Markov theorem.

The sphering approach also involves decorrelating a data matrix along both axes. Our work, including the theoretical analysis in *Zhou* (2014a), suggests that when the data matrix is non-square, attempting to decorrelate along the longer axis generally degrades performance. For genetics applications, where there are usually many more genes than samples, this suggests that decorrelating along the genes may hurt the performance of the sphering method. Fortunately, for gene-level analyses it is not necessary to decorrelate along the gene axis, since inference methods like false discovery rate are robust to a certain level of dependence among the variables (genes) (*Benjamini and Yekutieli*, 2001). Therefore, we also consider a modification of the sphering algorithm that only decorrelates along the samples.

Confounder adjustment is another related line of work that deals with similar

issues when attempting to discover mean differences. In particular, a part of that literature posits models where row-wise connections arise from the additive effects of potential latent variables. *Sun et al.* (2012) and *Wang et al.* (2015) use models of the form

$$X_{n\times m} = D_{n\times 1}\beta^T_{m\times 1} + Z_{n\times r}\Gamma^T_{m\times r} + E_{n\times m}$$
$$Z_{n\times r} = D_{n\times 1}\alpha^T_{r\times 1} + W_{n\times r}$$

where Z is an unobserved matrix of r latent factors. Rewriting these equations into the following form lets us better contrast the confounder model to our matrix-variate setup:

$$X = D(\beta + \Gamma\alpha)^T + W\Gamma^T + E. \tag{2.34}$$

These models are generally estimated by using some form of factor analysis to estimate Γ and then using regression methods with additive outlier detection to identify β, methodology that is quite different from our GLS-based methods.

For the two-group model, in the case of a globally centered data matrix X, the design matrix D in (2.34) takes the form

$$D^T_{n\times 1} = \begin{bmatrix} -1 & \cdots & -1 & 1 \cdots & 1 \end{bmatrix} = \begin{bmatrix} -1^T_{n_1} & 1^T_{n_2} \end{bmatrix}, \tag{2.35}$$

and 2β represents the vector of true mean differences between the groups. The vector β is estimated via OLS, yielding $\widehat{\beta}_{\text{OLS}}$, and CATE considers whether the residual $X - D_{n\times 1}\widehat{\beta}_{\text{OLS}}$ has a low-rank covariance structure plus noise. If so, $\widehat{\Gamma}\widehat{\alpha}$ aims to take out the residual low-rank structure through $D(\widehat{\Gamma\alpha})^T$. As illustrated in simulation and data analysis, this improves upon inference based only on $\widehat{\beta}_{\text{OLS}}$. When applying the CATE and related methods to data originated from the generative model as described in the present paper, CATE (and in particular, the related LEAPP) method essentially seeks

a sparse approximation of $\widehat{\beta}_{\text{OLS}}$; Moreover in LEAPP, this is essentially achieved via hard thresholding of coefficients of $\widehat{\beta}_{\text{OLS}}$, leading to improvements in performance in variable selection and its subsequence inference when the vector of true mean differences is presumed to be sparse. In our setting, we improve upon OLS using GLS.

2.9.1 Simulation results

Figure 2.12 compares the performance of Algorithm 2 to the sphering method of *Allen and Tibshirani* (2012) and the robust regression confounder adjustment method of *Wang et al.* (2015) on simulated matrix variate data motivated by the ulcerative colitis dataset described in Section 2.5. Note that this robust regression confounder adjustment is a minor modification of the LEAPP algorithm introduced in *Sun et al.* (2012). As discussed above, we also consider a modification of *Allen and Tibshirani* (2012) that only decorrelates along the rows.

We can see that across a range of dataset sizes our method consistently outperforms sphering in terms of sensitivity and specificity for identifying mean differences. In some settings, CATE improves on Tsphere and t-statistics despite being applied on misspecified models, because CATE takes out the additional rank two structure from the mean after OLS regression and does some approximate thresholding on the coefficients. Our method using GLS performs significantly better than CATE in the setting of non-identity B, with edges present both within and between groups.

Figure 2.14 fixes the sample size and repeats these comparisons on different sample correlation structures (which are described in Section 2.4). Figure 2.15 is analogous to Figure 2.14, but with A as the identity matrix. Algorithm 2 is competitive or superior to the competing methods across a range of topologies.

2.9.2 Comparison on UC data

We apply both Algorithm 2 and CATE on the ulcerative colitis data to compare their respective findings on real data. Figure 2.16 presents the test statistics from these algorithms. The test statistics have a correlation of 0.75. As expected, both methods find that the bulk of genes have small test statistics. Note that the regression line of the CATE test statistics on Algorithm 2's test statistics has a slope less than 1. This implies that Algorithm 2 generates more dispersed test statistics than CATE, and, given that we have shown in Figures 2.5 and 2.8 that Algorithm 2 provides well-calibrated test statistics, that it also has more power in this situation.

Using a threshold of FDR adjusted p-values smaller than 0.1, both methods find four genes with significant mean differences. However, there is only one gene (DPP10-AS1) that both methods identify. So, although there is significant correlation between the test statistics, the methods do not necessarily identify the same genes.

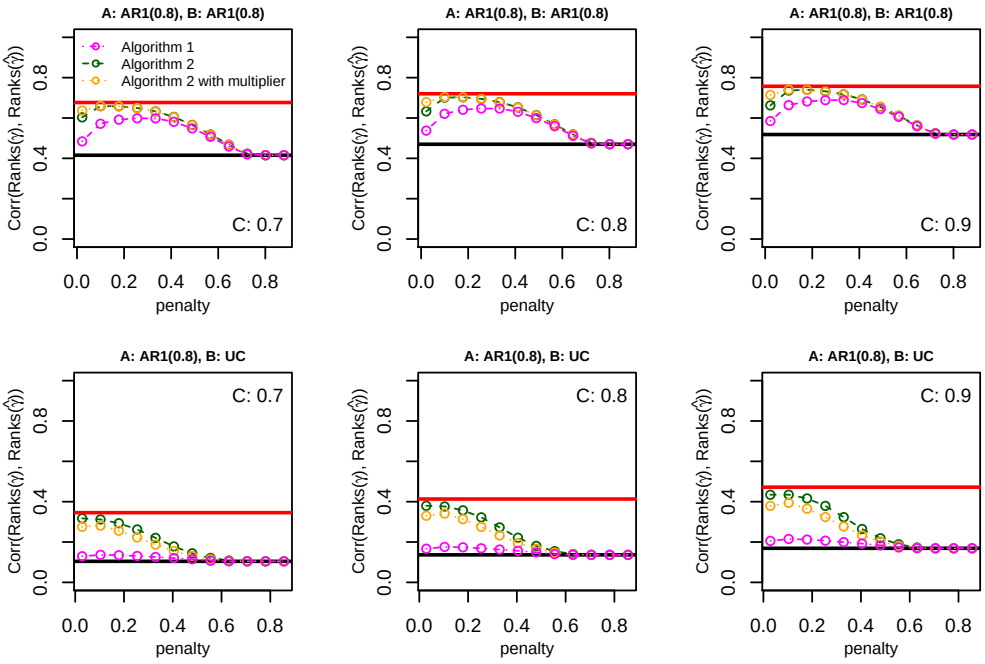

Figure 2.3: This figure displays the correlation between the rankings of the components of γ and $\widehat{\gamma}$, sorted by magnitude, denoted Corr(Ranks(γ), Ranks($\widehat{\gamma}$)) in the axis label. The vector of mean differences is chosen as $\gamma_j = C\exp(-(3/2000)j)$, for $j = 1, \ldots, 2000$. We also present the Algorithm 2 results with a multiplier on the threshold as described in Section 2.2.3. In the top row, the true B is AR1(0.8), with $n = 40$ and $m = 2000$. In the bottom row, the true B is chosen as an estimate from the UC data, with $n = 20$ and $m = 2000$. For the top row, the group labels are randomly assigned; for the bottom row, the first ten rows of the data are in group one, and the other ten are in group two. The figure is averaged over 200 replications. The top and bottom horizontal lines represent GLS with true B and OLS, respectively. The vertical axis displays the correlation of ranks between $\widehat{\gamma}$ and γ, and the horizontal axis displays the GLasso penalty parameter.

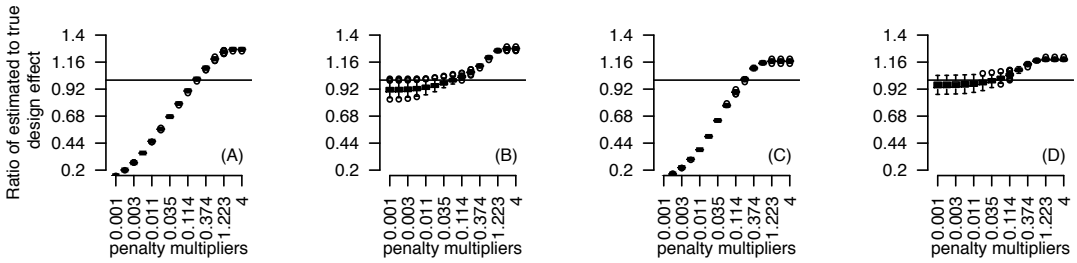

Figure 2.4: Ratio of estimated design effect to true design effect when B^{-1} is Erdős-Rényi, and A is AR1(0.8). Figures (A) and (B) correspond to sample size $n = 80$; (C) and (D) correspond to $n = 40$. Figures (A) and (C) correspond to Algorithm 1; Figures (B) and (D) correspond to Algorithm 2, with ten columns group centered. These results are based on dimension parameter $m = 2000$ and 250 simulation replications.

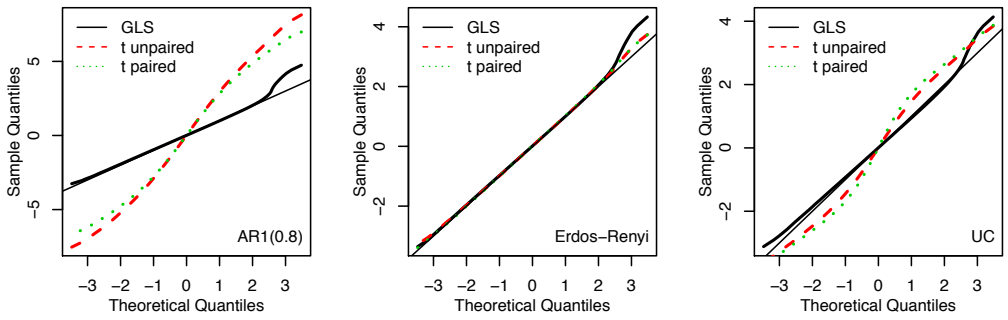

Figure 2.5: Quantile plots of test statistics. Ten genes have nonzero mean differences equal to 2, 0.8, and 1 in the three plots, respectively. In each plot A is AR1(0.8). Covariance structures for B are as indicated. In the third plot, the true B is set to $\widehat{B}$ for the ulcerative colitis data, described in Section 2.5. For the first two plots there are $n = 40$ samples and $m = 2000$ variables. For the third plot there are $n = 20$ samples and $m = 2000$ variables. Each plot has 250 simulation replications.

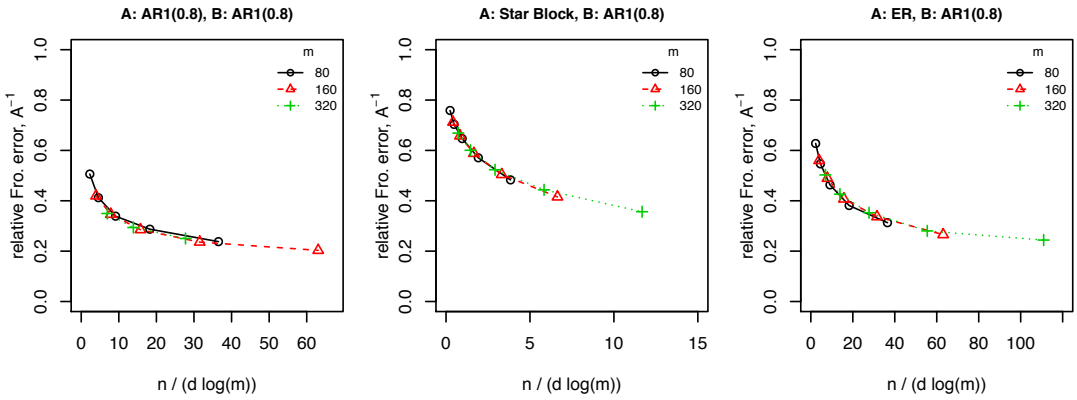

Figure 2.6: Relative Frobenius error in estimating A^{-1}, as n varies. In each plot the matrix B is AR1(0.8) and A is as indicated. The vertical axis is relative Frobenius error, and the horizontal axis $n/(d\log(m))$, where d is the maximum node degree. The GLasso penalty is chosen to minimize the relative Frobenius error. Each point is based on 250 Monte Carlo replications.

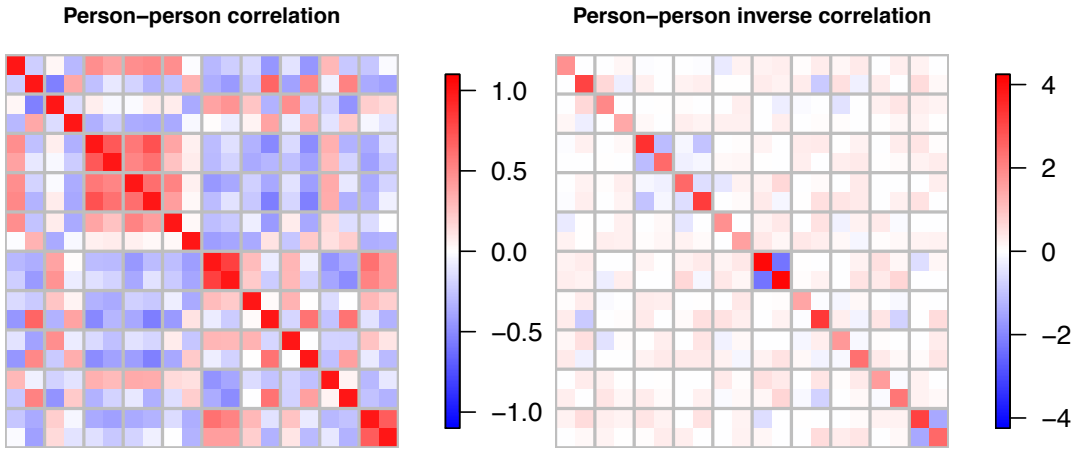

Figure 2.7: Estimated person-person correlation matrix and its inverse, estimated using the 2000 genes with largest marginal variance.

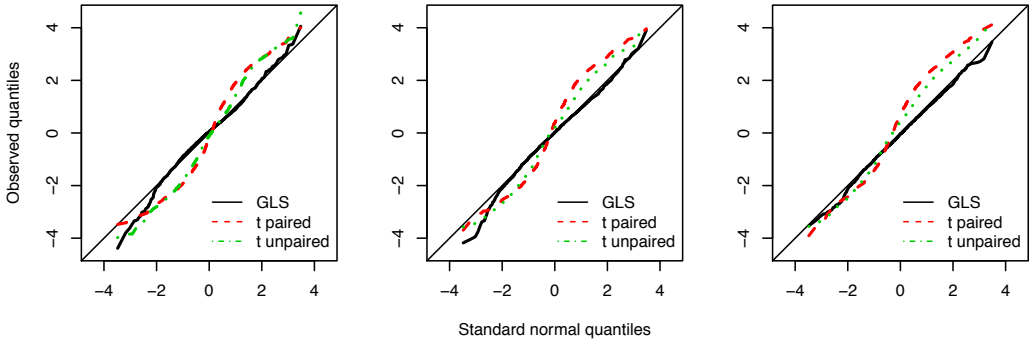

Figure 2.8: Quantile plots of test statistics for three disjoint gene sets, each consisting of 2000 genes. The genes are partitioned based on marginal variance. GLS statistics are taken from step 5 of Algorithm 2; in step 2, the ten genes with greatest mean differences are selected for group centering.

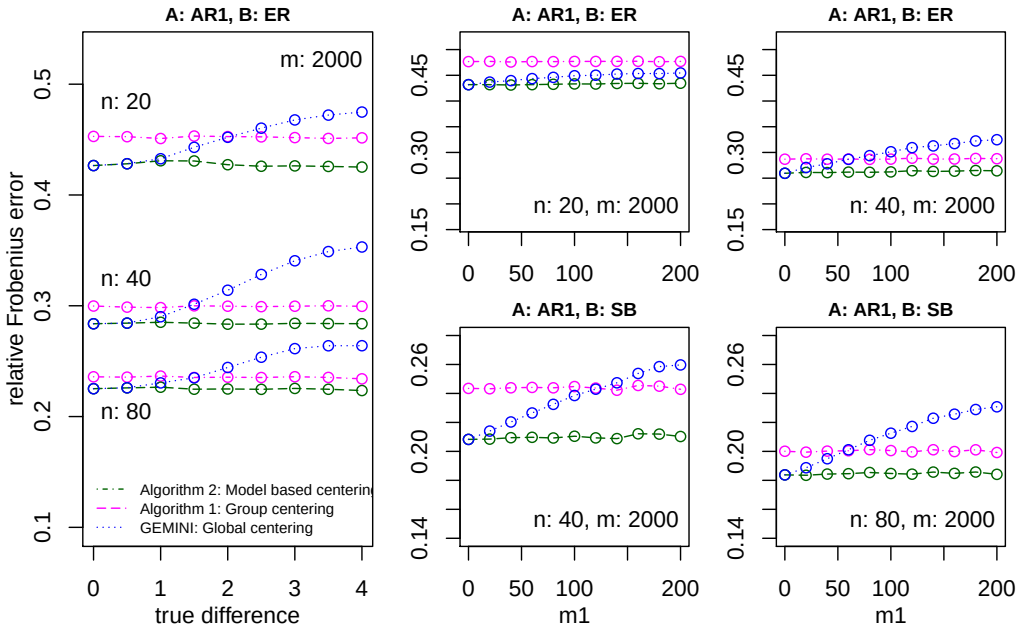

Figure 2.9: Performance of Gemini, Algorithm 1, and Algorithm 2 for estimating B under different mean and covariance structures. As the sample size increases, we can see that Algorithm 1 improves relative to Gemini and begins to catch up to Algorithm 2. Gemini's performance always degrades as the true differences grow or more differentially expressed genes are added, while Algorithm 1 and 2 are stable. We set B^{-1} as Erdős-Rényi (ER) or star-block with blocks of size 4 (SB). All plots use A from an AR1(0.8) model with $m = 2000$ and are averaged over 200 replications. In the left plot the first 50 genes are differentially expressed at the level specified on the x-axis. As indicated, the three groups of lines correspond to $n = 20$, 40, and 80. In the right two columns there are `m1` number of genes with exponentially decaying true differences between groups, scaled so that the largest difference is 5 (resulting in an average difference of approximately 1).

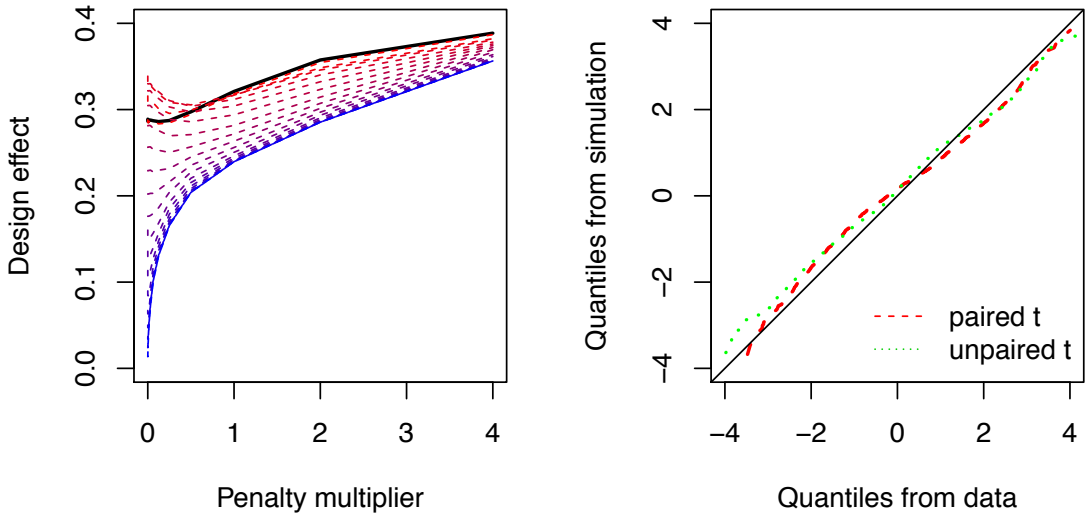

Figure 2.10: The first plot displays the estimated design effect vs. the penalty multiplier for Algorithm 2. Each curve corresponds to a different number of columns being group centered. As the curves go from top to bottom, the number of group centered columns increases from 10 to 2000. The second plot shows a quantile plot of test statistics from the data vs. simulated test statistics; in the simulation, the population person-person covariance matrix is $\widehat{B}$, as estimated from the UC data.

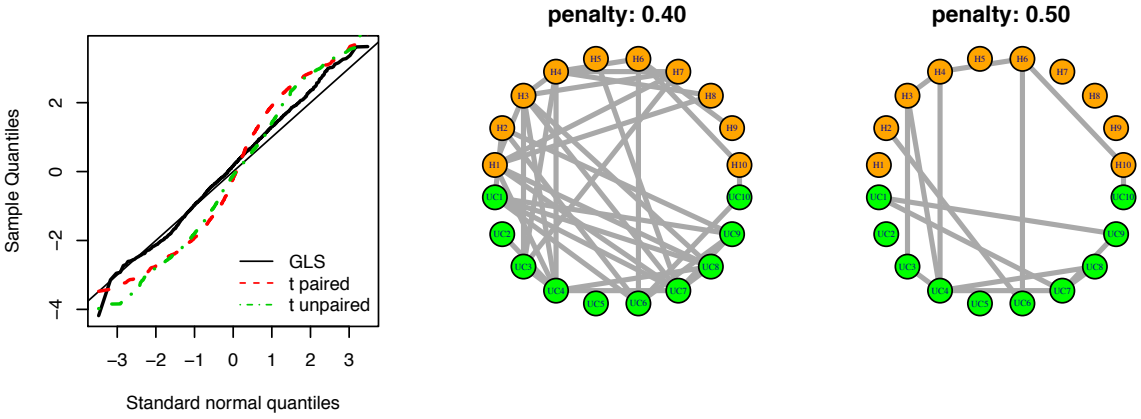

Figure 2.11: Quantile plot and inverse covariance graphs. The first two plots correspond to $\lambda = 0.4$ and 128 group centered genes. The third plot corresponds to $\lambda = 0.5$ and 128 group centered genes. Green circles correspond to twins with UC, orange circles to twins without UC. Twins are aligned vertically.

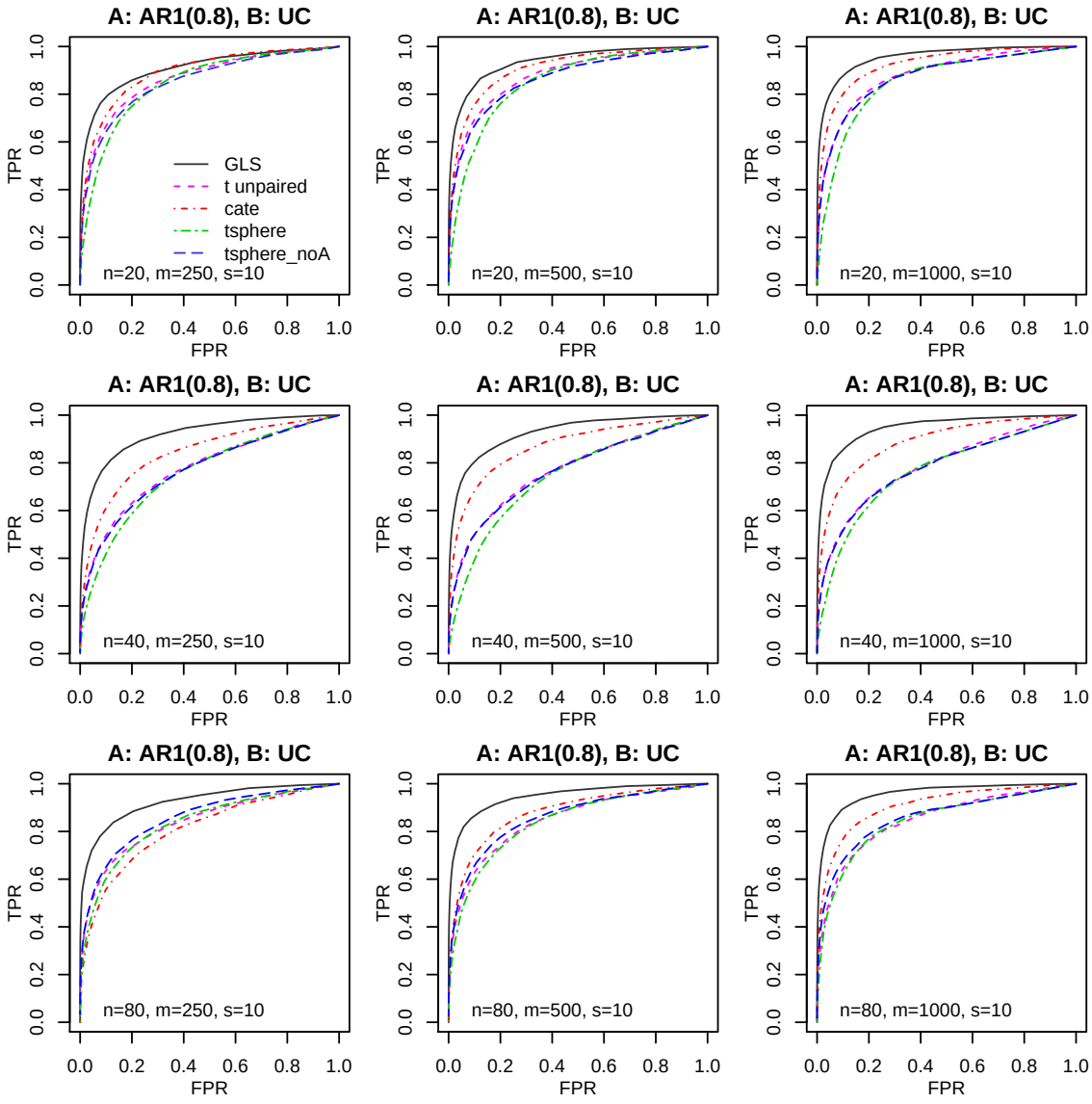

Figure 2.12: Performance of Algorithm 2 (GLS) relative to sphering and confounder adjustment methods, labeled as `tsphere` and `cate`, respectively. These are ROC curves for identifying true mean differences. An implementation of the sphering algorithm that does not adjust for A is also included, labeled as `tsphere_noA`. Each panel shows the average ROC curves over 200 simulations. We simulate matrix variate data with gene correlations from an AR1(0.8) model and let $s = 10$ genes have true mean differences of 0.8, 0.6, and 0.4 for the first, second and third rows, respectively. For all of these the true B is set to $\widehat{B}$ from the ulcerative colitis data (using a repeated block structure for larger n values), described in Section 2.5 and evenly-sized groups are assigned randomly.

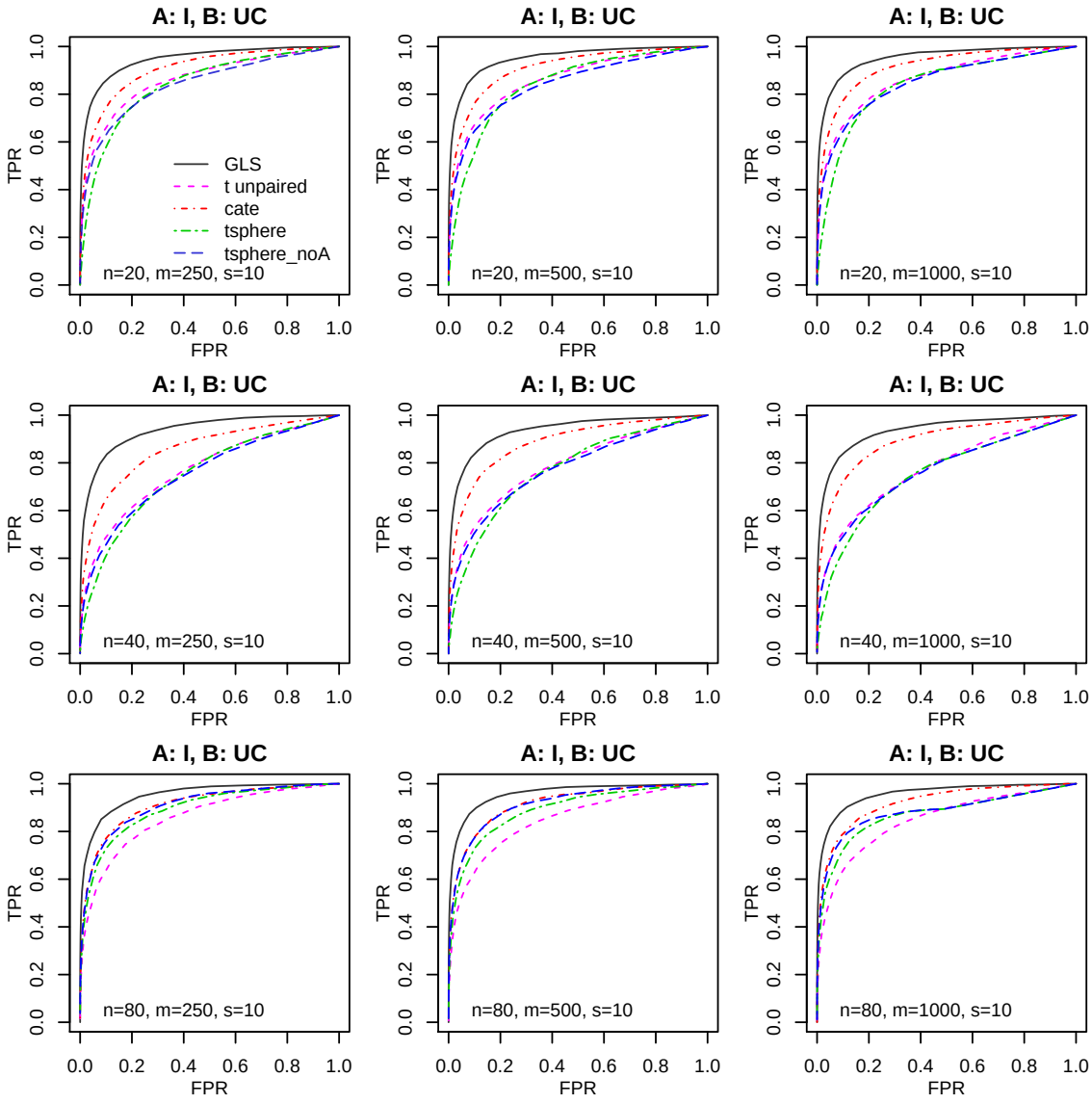

Figure 2.13: Performance of Algorithm 2 (GLS) relative to sphering and confounder adjustment, labeled as `tsphere` and `cate`, respectively. These are ROC curves for identifying true mean differences. An implementation of the sphering algorithm that does not adjust for A is also included, labeled as `tsphere_noA`. Each panel shows the average ROC curves over 200 simulations. We simulate matrix variate data with no gene-wise correlations ($A = I$) and let $s = 10$ genes have true mean differences of 0.8, 0.6, and 0.4 for the first, second and third rows, respectively. For all of these the true B is set to $\widehat{B}$ from the ulcerative colitis data (using a repeated block structure for larger n values), described in Section 2.5 and evenly-sized groups are assigned randomly.

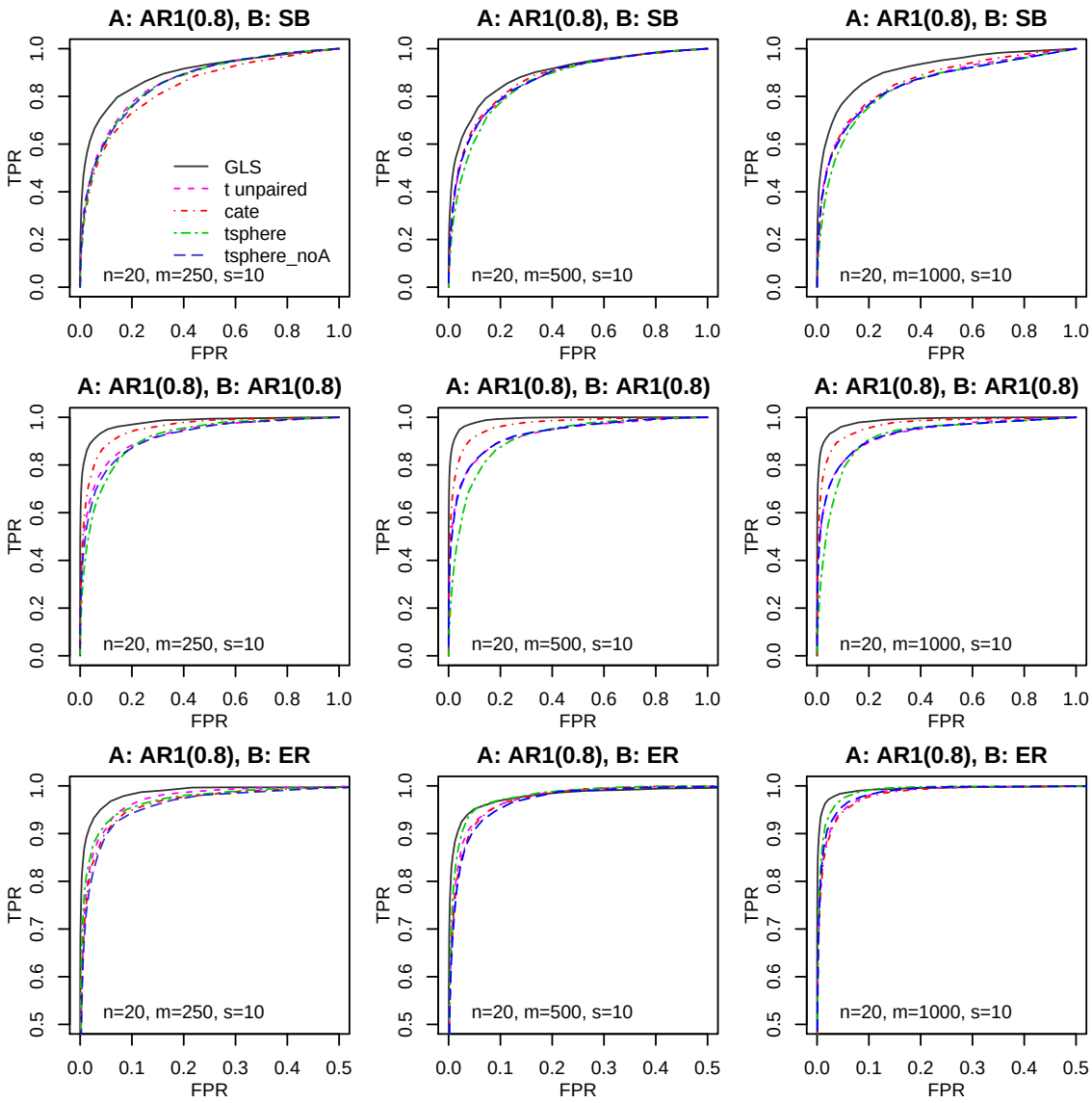

Figure 2.14: Performance of Algorithm 2 (GLS) relative to sphering and confounder adjustment, labeled as `tsphere` and `cate`, respectively. These are ROC curves for identifying true mean differences. An implementation of the sphering algorithm that does not adjust for A is also included, labeled as `tsphere_noA`. Each panel shows the average ROC curves over 200 simulations. We simulate matrix variate data with an AR1(0.8) model for A and let $s = 10$ genes have true mean differences of 0.8. B is constructed according to a Star-Block model with blocks of size 4, an AR1(0.8), and an Erdős-Rényi random graph with $d = n \log n$ edges. All of these use $n = 20$ and randomly assign 10 observations to each group.

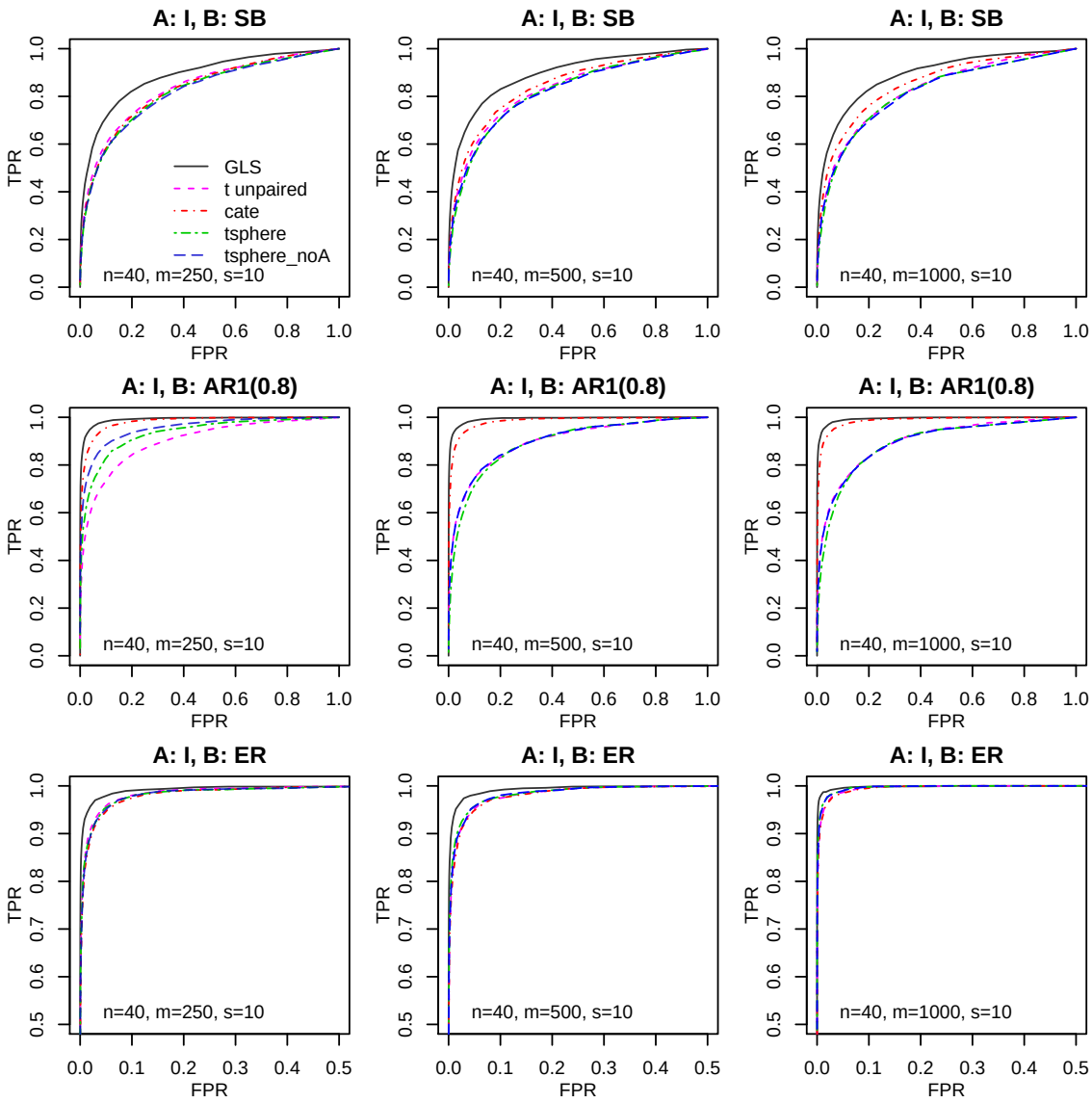

Figure 2.15: Performance of Algorithm 2 (GLS) relative to sphering and confounder adjustment, labeled as `tsphere` and `cate`, respectively. These are ROC curves for identifying true mean differences. An implementation of the sphering algorithm that does not adjust for A is also included, labeled as `tsphere_noA`. Each panel shows the average ROC curves over 200 simulations. We simulate matrix variate data with no gene-wise correlations ($A = I$) and let $s = 10$ genes have true mean differences of 0.6. B is constructed according to a Star-Block model with blocks of size 4, an AR1(0.8), and an Erdős-Rényi random graph with $d = n \log n$ edges. All of these use $n = 40$ and randomly assign 20 observations to each group.

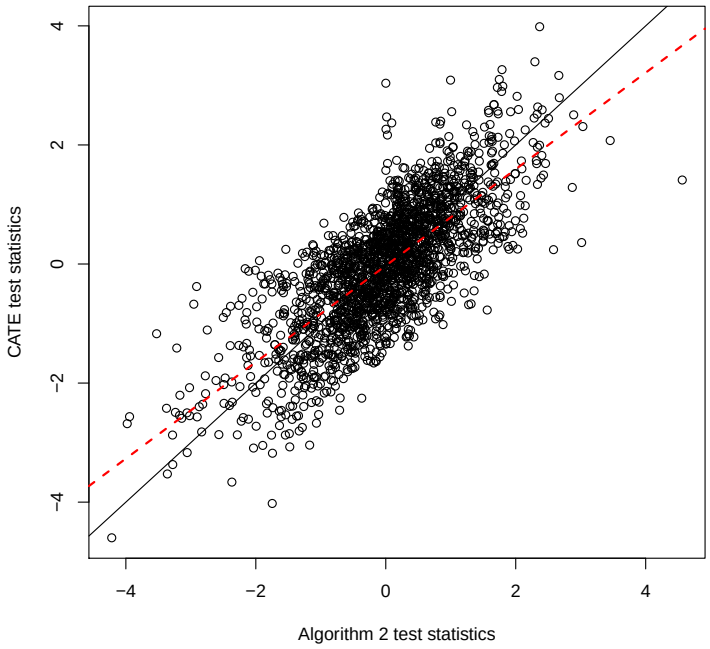

Figure 2.16: Scatterplot of t-statistics for CATE and Algorithm 2 applied on the ulcerative colitis data. The 45-degree line is included in black while red dashed line is the linear fit.

CHAPTER III

Theoretical results for joint mean and covariance estimation

This chapter is joint work with Roger Fan, Kerby Shedden, and Shuheng Zhou.

In this chapter, we provide proofs of the theorems presented in Chapter 2. Section 3.1 contains some preliminary results and notation. In Section 3.2, we prove Theorem II.1. In Sections 3.3 and 3.4 we prove Theorem II.3. In Section 3.5, we derive entry-wise rates of convergence for the sample covariance matrices. In Sections 3.6 and 3.7 we prove Theorem II.4 and its auxiliary results.

3.1 Preliminary results

In this section, we refresh notation and introduce propositions that are shared in the proofs of the theorems. For convenience, we first restate some notation.

$$D = \begin{bmatrix} 1_{n_1} & 0 \\ 0 & 1_{n_2} \end{bmatrix} \in \mathbb{R}^{n\times 2} \tag{3.1}$$

$$\Omega = (D^T B^{-1} D)^{-1} \text{ and } \Omega_{n,m} = (D^T B_{n,m}^{-1} D)^{-1} \tag{3.2}$$

$$\Delta = B_{n,m}^{-1} - B^{-1} \tag{3.3}$$

$$\widehat{\beta}(\widehat{B}^{-1}) = (D^T \widehat{B}^{-1} D)^{-1} D^T \widehat{B}^{-1} X \in \mathbb{R}^{2\times m} \tag{3.4}$$

When D has the form (3.1), the singular values are $\sigma_{\max}(D) = \sqrt{n_{\max}}$ and $\sigma_{\min}(D) = \sqrt{n_{\min}}$. The condition number is $\kappa(D) = \sigma_{\max}(D)/\sigma_{\min}(D) = \sqrt{n_{\text{ratio}}}$ where $n_{\text{ratio}} = \max(n_1, n_2)/\min(n_1, n_2)$.

We first state some convenient notation and bounds.

$$r_a := a_{\max}/a_{\min} \text{ and } r_b := b_{\max}/b_{\min};$$

$$1/\varphi_{\min}(A) = \|A^{-1}\|_2 \leqslant \|\rho(A)^{-1}\|_2/a_{\min} = \frac{1}{a_{\min}\varphi_{\min}(\rho(A))}, \tag{3.5}$$

$$1/\varphi_{\min}(B) = \|B^{-1}\|_2 \leqslant \|\rho(B)^{-1}\|_2/b_{\min} = \frac{1}{b_{\min}\varphi_{\min}(\rho(B))}, \tag{3.6}$$

$$1/\varphi_{\min}(\rho(A)) = \|\rho(A)^{-1}\|_2 \leqslant a_{\max}\|A^{-1}\|_2, \tag{3.7}$$

$$1/\varphi_{\min}(\rho(B)) = \|\rho(B)^{-1}\|_2 \leqslant b_{\max}\|B^{-1}\|_2 \tag{3.8}$$

$$\|A\|_2 \leqslant a_{\max}\|\rho(A)\|_2, \quad \|B\|_2 \leqslant b_{\max}\|\rho(B)\|_2, \tag{3.9}$$

$$\|\rho(A)\|_2 \leqslant \|A\|_2/a_{\min}, \quad \text{and} \quad \|\rho(B)\|_2 \leqslant \|B\|_2/b_{\min}. \tag{3.10}$$

The eigenvalues of the correlation matrices satisfy

$$0 < \varphi_{\min}(\rho(A)) \leqslant 1 \leqslant \varphi_{\max}(\rho(A)) \text{ and } 0 < \varphi_{\min}(\rho(B)) \leqslant 1 \leqslant \varphi_{\max}(\rho(B)). \tag{3.11}$$

In the remainder of this section, we state preliminary results and highlight important intermediate steps that are used in the proofs of Theorems II.1 and II.3. First we state propositions used in mean estimation for Theorems II.1 and II.3.

3.1.1 Propositions

We now state propositions used in the proofs of Lemmas III.5 and III.6. We defer the proof of Proposition III.1 to Section 3.2.5.

Proposition III.1. *For Ω as defined in (3.2) and some design matrix D,*

$$\|\Omega\|_2 \leqslant \|B\|_2/\sigma^2_{\min}(D)$$

In the case that D is defined as in (3.1), we have $\|\Omega\|_2 \leqslant \|B\|_2/n_{\min}$.

Furthermore,

$$\lambda_{\min}(\Omega) \geqslant \frac{\lambda_{\min}(B)}{n_{\max}}. \tag{3.12}$$

We state the following perturbation bound.

Theorem III.2 (Golub & Van Loan, Theorem 2.3.4)**.** *If A is invertible and $\|A^{-1}E\|_p < 1$, then $A + E$ is invertible and*

$$\|(A+E)^{-1} - A^{-1}\|_p \leqslant \frac{\|E\|_p\|A^{-1}\|_p^2}{1 - \|A^{-1}E\|_p} \leqslant \frac{\|E\|_p\|A^{-1}\|_p^2}{1 - \|A^{-1}\|_p\|E\|_p}.$$

In Proposition III.3, we provide auxiliary upper bounds that depend on $\|\Delta\|_2$, $\|B\|_2$, $\kappa(D)$, and $\sigma_{\min}(D)$. We defer the proof of Proposition III.3 to the end of this section, for clarity of presentation.

Proposition III.3. *Let* $\Delta = B_{n,m}^{-1} - B^{-1}$.

$$\delta_0(\Delta) := \|\Omega_{n,m} - \Omega\|_2 \leqslant \frac{1}{\sigma_{\min}^2(D)} \frac{\|B\|_2^2 \|\Delta\|_2}{1/\kappa^2(D) - \|B\|_2 \|\Delta\|_2} \tag{3.13}$$

$$\delta_1(\Delta) := \left\|\Omega D^T \Delta\right\|_2 \leqslant \sigma_{\max}(D) \|B\|_2 \|\Delta\|_2 / \sigma_{\min}^2(D) = \frac{\sqrt{n_{\max}}}{n_{\min}} \|B\|_2 \|\Delta\|_2. \tag{3.14}$$

If $\|(D^T B^{-1} D)^{-1} D^T \Delta D\|_2 < 1$, *then*

$$\delta_2(\Delta) := \left\|(\Omega_{n,m} - \Omega)\, D^T \Delta\right\|_2 \leqslant \frac{\kappa(D)}{\sigma_{\min}(D)} \frac{\|B\|_2^2 \|\Delta\|_2^2}{1/\kappa^2(D) - \|B\|_2 \|\Delta\|_2} \tag{3.15}$$

$$\delta_3(\Delta) := \left\|(\Omega_{n,m} - \Omega)\, D^T B^{-1}\right\|_2 \leqslant \frac{\kappa(D)}{\sigma_{\min}(D)} \frac{\|B\|_2^2 \|B^{-1}\|_2 \|\Delta\|_2}{1/\kappa^2(D) - \|B\|_2 \|_2 \|\Delta\|_2} \tag{3.16}$$

The following proposition is a corollary of Proposition III.3.

Proposition III.4. *When* D *has the form (3.1), and* Ω *is as defined in* (3.2),

$$\begin{aligned}
\delta_0(\Delta) &= \|\Omega_{n,m} - \Omega\|_2 \leqslant \frac{1}{n_{\min}} \frac{\|B\|_2^2 \|\Delta\|_2}{1/n_{ratio} - \|B\|_2 \|\Delta\|_2} \\
\delta_1(\Delta) &= \left\|\Omega D^T \Delta\right\|_2 \leqslant \frac{\sqrt{n_{ratio}}}{\sqrt{n_{\min}}} \|B\|_2 \|\Delta\|_2 \\
\delta_2(\Delta) &= \left\|(\Omega_{n,m} - \Omega)\, D^T \Delta\right\|_2 \leqslant \frac{\sqrt{n_{ratio}}}{\sqrt{n_{\min}}} \frac{\|B\|_2^2 \|\Delta\|_2^2}{1/n_{ratio} - \|B\|_2 \|\Delta\|_2}
\end{aligned}$$

Let K be defined as in Theorem II.1. We express the entrywise rates of convergence of the sample correlation matrices $\widehat{\Gamma}(B)$ and $\widehat{\Gamma}(A)$, respectively, in terms of the following quantities:

$$\widetilde{\alpha} = C_A K \frac{\log^{1/2}(m)}{\sqrt{m}} \left(1 + \frac{\|B\|_1}{n}\right) + \frac{\|B\|_1}{n_{\min}} \quad \text{and} \quad \widetilde{\eta} = C_B K \frac{\log^{1/2}(m \vee n)}{\sqrt{n}} + \frac{\|B\|_1}{n} \tag{3.17}$$

3.2 Proof of Theorem II.1 and Corollary II.2

3.2.1 Proof of Theorem II.1

Let $B_{n,m} \in \mathbb{R}^{n\times n}$ denote a fixed positive definite matrix. Let D be as defined as in (2.4). Define $\Delta_{n,m} = B_{n,m}^{-1} - B^{-1}$ and

$$\Omega = (D^T B^{-1} D)^{-1} \text{ and } \Omega_{n,m} = (D^T B_{n,m}^{-1} D)^{-1}. \tag{3.18}$$

Note that we can decompose the error for all j as

$$\|\widehat{\beta}_j(B_{n,m}^{-1}) - \beta_j^*\|_2 \leqslant \|\widehat{\beta}_j(B^{-1}) - \beta_j^*\|_2 + \|\widehat{\beta}_j(B_{n,m}^{-1}) - \widehat{\beta}_j(B^{-1})\|_2 =: \mathrm{I} + \mathrm{II}. \tag{3.19}$$

We will use the following lemmas, which are proved in subsections 3.2.4 and 3.2.3, to bound these two terms on the right-hand side, respectively.

Lemma III.5. *Let $\mathcal{E}_2$ denote the event*

$$\mathcal{E}_2 = \left\{ \|\widehat{\beta}_j(B^{-1}) - \beta_j^*\|_2 \leqslant s_{n,m} \right\}, \quad \textit{with} \quad s_{n,m} = C_3 d^{1/2} \sqrt{\frac{\log(m)\|B\|_2}{n_{\min}}}. \tag{3.20}$$

Then $P(\mathcal{E}_2) \geqslant 1 - 2/m^d$.

Lemma III.6. *Let $B_{n,m} \in \mathbb{R}^{n\times n}$ denote a fixed matrix such that $B_{n,m} \succ 0$. Let $X_j \in \mathbb{R}^n$ denote the jth column of X, where X is a realization of model (2.2). Let $\mathcal{E}_3$ denote the event*

$$\mathcal{E}_3 = \left\{ \|\widehat{\beta}_j(B_{n,m}^{-1}) - \widehat{\beta}_j(B^{-1})\|_2 \leqslant t_{n,m} \right\}, \quad \textit{with} \quad t_{n,m} = \widetilde{C} n_{\min}^{-1/2} \|\Delta_{n,m}\|_2. \tag{3.21}$$

for some absolute constant $\widetilde{C}$. Then $P(\mathcal{E}_3) \geqslant 1 - 2/m^d$.

The proof of (2.18) follows from the union bound $P(\mathcal{E}_2 \cap \mathcal{E}_3) \geqslant 1 - P(\mathcal{E}_2) - P(\mathcal{E}_3) \geqslant 1 - 4/m^d$. Next we prove (2.20). Let $r_{n,m} = s_{n,m} + t_{n,m}$, as defined in (2.18). Let

$\delta = (1, -1) \in \mathbb{R}^2$. Then

$$|\widehat{\gamma}_j(B_{n,m}^{-1}) - \gamma_j| = \left|\delta^T \left(\widehat{\beta}_j(B_{n,m}^{-1}) - \beta_j^*\right)\right| \leqslant \|\delta\|_2 \|\widehat{\beta}_j(B_{n,m}^{-1}) - \beta_j^*\|_2 = \sqrt{2}\|\widehat{\beta}_j(B_{n,m}^{-1}) - \beta_j^*\|_2,$$

where we used the Cauchy-Schwarz inequality. Hence if $\|\widehat{\beta}_j(B_{n,m}^{-1}) - \beta_j\|_2 \leqslant r_{n,m}$, it follows that $|\widehat{\gamma}_j(B_{n,m}^{-1}) - \gamma_j| \leqslant \sqrt{2} r_{n,m}$. The result holds by applying a union bound over the variables $j = 1, \ldots, m$. $\square$

This completes the proof of Theorem II.1.

3.2.2 Proof of Corollary II.2 and Corollary II.5

First note that by Proposition III.4,

$$\begin{aligned}
\left|\delta^T (D^T \widehat{B}^{-1} D)^{-1} \delta - \delta^T (D^T B^{-1} D)^{-1} \delta\right| &= \left|\delta^T \left((D^T \widehat{B}^{-1} D)^{-1} - (D^T B^{-1} D)^{-1}\right) \delta\right| \\
&\leqslant \|\delta\|_2^2 \left\|(D^T \widehat{B}^{-1} D)^{-1} - (D^T B^{-1} D)^{-1}\right\|_2 \\
&= 2 \left\|(D^T \widehat{B}^{-1} D)^{-1} - (D^T B^{-1} D)^{-1}\right\|_2 \\
&\leqslant 2 \frac{\|B\|_2^2 \, \|\Delta\|_2}{n_{\min}}. \qquad (3.22)
\end{aligned}$$

Note that by Proposition III.1,

$$|\delta^T \Omega \delta| \geqslant \frac{\lambda_{\min}(B)}{n_{\max}}. \qquad (3.23)$$

Corollary II.2 follows from (3.22) and (3.23), which provide an upper bound on the numerator and lower bound on the denominator, respectively.

Corollary II.5 holds because by (2.28) of Theorem II.4,

$$\left|\delta^T \left(\widehat{\Omega} - \Omega\right) \delta\right| \leqslant 2 \frac{\|B\|_2^2}{n_{\min}} \left(\frac{C' \lambda_A \sqrt{|B^{-1}|_{0,\text{off}} \vee 1}}{b_{\min} \varphi_{\min}^2(\rho(B))}\right) \leqslant 2C' \frac{\kappa(B)}{n_{\min}} \lambda_A \sqrt{|B^{-1}|_{0,\text{off}} \vee 1} \qquad (3.24)$$

3.2.3 Proof of Lemma III.5

First, we show that

$$\|\Omega^{1/2}\|_F + d^{1/2}K^2\sqrt{\log(m)}\|\Omega\|_2^{1/2}/\sqrt{c} \leqslant s_{n,m}, \tag{3.25}$$

with $s_{n,m}$ as defined in (2.19). Because $\|\Omega^{1/2}\|_F \leqslant \sqrt{2}\|\Omega^{1/2}\|_2$, it follows that

$$\begin{aligned}\|\Omega^{1/2}\|_F + d^{1/2}K^2\sqrt{\log(m)}\|\Omega\|_2^{1/2}/\sqrt{c} &\leqslant \left(\sqrt{2} + d^{1/2}K^2\sqrt{\log(m)}/\sqrt{c}\right)\|\Omega\|_2^{1/2} \\ &\leqslant C_3 d^{1/2}\sqrt{\log(m)}\|\Omega\|_2^{1/2} \leqslant C_3 d^{1/2}\sqrt{\frac{\log(m)\|B\|_2}{n_{\min}}},\end{aligned}$$

where the last step follows from Proposition III.1. Next, we express $\widehat{\beta}_j(B^{-1}) - \beta_j^*$ as

$$\widehat{\beta}_j(B^{-1}) - \beta_j^* = \Omega^{1/2}\eta_j, \quad \text{where} \quad \eta_j = \Omega^{-1/2}\left(\widehat{\beta}_j(B^{-1}) - \beta_j^*\right).$$

By the bound (3.25), event $\mathcal{E}_2^c$ implies $\{\|\Omega^{1/2}\eta_j\|_2 > \|\Omega^{1/2}\|_F + d^{1/2}K^2\sqrt{\log(m)}\|\Omega\|_2^{1/2}/\sqrt{c}\}$. Therefore,

$$\begin{aligned}P\left(\|\Omega\eta_j\|_2 \geqslant s_{n,m}\right) &\leqslant P\left(\|\Omega\eta_j\|_2 > \|\Omega^{1/2}\|_F + d^{1/2}K^2\sqrt{\log(m)}\|\Omega\|_2^{1/2}/\sqrt{c}\right) \\ &\leqslant P\left(\left|\|\Omega^{1/2}\eta_j\|_2 - \|\Omega^{1/2}\|_F\right| > d^{1/2}K^2\sqrt{\log(m)}\|\Omega\|_2^{1/2}/\sqrt{c}\right) \\ &\leqslant 2\exp\left(\frac{-c\left(d^{1/2}K^2\sqrt{\log(m)}\|\Omega\|_2^{1/2}/\sqrt{c}\right)^2}{K^4\|\Omega^{1/2}\|_2^2}\right) \\ &= 2\exp\left(\frac{-d\log(m)\|\Omega\|_2}{\|\Omega^{1/2}\|_2^2}\right) = 2\exp\left(-d\log(m)\right) = 2/m^d.\end{aligned}$$

□

3.2.4 Proof of Lemma III.6

The proof will proceed in the following steps. First, we show that $\widehat{\beta}_j(B_{n,m}^{-1}) - \widehat{\beta}_j(B^{-1})$ can be expressed as VZ_j, where

$$V = \left(\Omega_{n,m} D^T B_{n,m}^{-1} - \Omega D^T B^{-1}\right) B^{1/2} \in \mathbb{R}^{2\times m}$$

is a fixed matrix, and $Z_j = B^{-1/2} X_j$. Second, we show that

$$\|V\|_F + d^{1/2} K^2 \log^{1/2}(m) \|V\|_2 / \sqrt{c} \leqslant \widetilde{C} n_{\min}^{-1/2} \|\Delta\|_2.$$

Third, we use the first and second steps combined with the Hanson-Wright inequality to show that with high probability, $\|VZ_j\|_2$ is at most $\widetilde{C} n_{\min}^{-1/2} \|\Delta\|_2$.

For the first step of the proof, let $Z_j = B^{-1/2} X_j$, and note that $\widehat{\beta}_j(B_{n,m}^{-1}) - \widehat{\beta}_j(B^{-1}) = VZ_j$, where $V \in \mathbb{R}^{2\times m}$ is a fixed matrix, because

$$\begin{aligned}
\widehat{\beta}_j(B_{n,m}^{-1}) - \widehat{\beta}_j(B^{-1}) &= \left[(D^T B_{n,m}^{-1} D)^{-1} D^T B_{n,m}^{-1} - \Omega D^T B^{-1}\right] B^{1/2} (B^{-1/2} X_j) \\
&= \left[(D^T B_{n,m}^{-1} D)^{-1} D^T B_{n,m}^{-1} - \Omega D^T B^{-1}\right] B^{1/2} Z_j.
\end{aligned}$$

For the second step of the proof, we show that $\|V\|_F + d^{1/2} K^2 \log^{1/2}(m) \|V\|_2 / \sqrt{c} \leqslant \widetilde{C} n_{\min}^{-1/2} \|\Delta\|_2$. First we obtain an upper bound on V. By the triangle inequality,

$$\begin{aligned}
\|\Omega_{n,m} D^T B_{n,m}^{-1} - \Omega D^T B^{-1}\|_2 &= \left\|\Omega_{n,m} D^T B_{n,m}^{-1} - \Omega D^T B^{-1}\right\|_2 \\
&\leqslant \left\|(\Omega_{n,m} - \Omega) D^T (B_{n,m}^{-1} - B^{-1})\right\|_2 + \left\|(\Omega_{n,m} - \Omega) D^T B^{-1}\right\|_2 + \left\|\Omega D^T \Delta\right\|_2 \\
&= \delta_2(\Delta) + \delta_3(\Delta) + \delta_1(\Delta).
\end{aligned}$$

We bound each of the three terms using Proposition III.3,

$$\begin{aligned}
\delta_2(\Delta) &= \left\|\left(\Omega_{n,m} - \Omega\right) D^T \Delta\right\|_2 \leqslant \frac{\sqrt{n_{\text{ratio}}}}{\sqrt{n_{\min}}} \frac{\|B\|_2^2 \|\Delta\|_2^2}{1/n_{\text{ratio}} - \|B\|_2 \|\Delta\|_2} \\
\delta_3(\Delta) &= \left\|\left(\Omega_{n,m} - \Omega\right) D^T B^{-1}\right\|_2 \leqslant \frac{\sqrt{n_{\text{ratio}}}}{\sqrt{n_{\min}}} \frac{\|B\|_2^2 \|B^{-1}\|_2 \|\Delta\|_2}{1/n_{\text{ratio}} - \|B\|_2 \|_2 \|\Delta\|_2} \\
\delta_1(\Delta) &= \left\|\Omega D^T \Delta\right\|_2 \leqslant \frac{\sqrt{n_{\text{ratio}}}}{\sqrt{n_{\min}}} \|B\|_2 \|\Delta\|_2.
\end{aligned}$$

Applying the above bounds yields

$$\begin{aligned}
\|V\|_2 &\leqslant \frac{\sqrt{n_{\text{ratio}}}}{\sqrt{n_{\min}}} \|\Delta\|_2 \|B\|_2^{1/2} \left(\frac{\|B\|_2^2 \|\Delta\|_2}{1/\kappa^2(D) - \|B\|_2 \|\Delta\|_2} + \frac{\|B\|_2^2 \|B^{-1}\|_2}{1/\kappa^2(D) - \|B\|_2 \|_2 \|\Delta\|_2} + \|B\|_2 \right) \\
&\leqslant \widetilde{C} n_{\min}^{-1/2} \|\Delta\|_2.
\end{aligned}$$

For the third step of the proof, we use the Hanson-Wright inequality to bound $\|VZ_j\|_2$:

$$\begin{aligned}
P\left(\|VZ_j\|_2 > \widetilde{C} n_{\min}^{-1/2} \|\Delta\|_2\right) &\leqslant P\left(\|VZ_j\|_2 > \|V\|_F + d^{1/2} K^2 \log^{1/2}(m) \|V\|_2 / \sqrt{c}\right) \\
&= P\left(\|VZ_j\|_2 - \|V\|_F > d^{1/2} K^2 \log^{1/2}(m) \|V\|_2 / \sqrt{c}\right) \\
&\leqslant P\left(\left|\|VZ_j\|_2 - \|V\|_F\right| > d^{1/2} K^2 \log^{1/2}(m) \|V\|_2 / \sqrt{c}\right) \\
&\leqslant 2 \exp\left(- \frac{c\left(d^{1/2} K^2 \log^{1/2}(m) \|V\|_2 / \sqrt{c}\right)^2}{K^4 \|V\|_2^2} \right) \qquad \text{(Hanson-Wright inequality)} \\
&= 2 \exp\left(-d \log(m)\right) = 2/m^d.
\end{aligned}$$

□

3.2.5 Proof of Proposition III.1

Let $D = U \Psi V^T$ be the singular value decomposition of D, with $U \in \mathbb{R}^{n \times 2}$, $\Psi \in \mathbb{R}^{2 \times 2}$, and $V \in \mathbb{R}^{2 \times 2}$. Then $(D^T B^{-1} D)^{-1} = (V \Psi U^T B^{-1} U \Psi V^T)^{-1} = V \Psi^{-1} (U^T B^{-1} U)^{-1} \Psi^{-1} V^T$.

Thus

$$\begin{aligned}
\|(D^TB^{-1}D)^{-1}\|_2 &= \|\Psi^{-1}(U^TB^{-1}U)^{-1}\Psi^{-1}\|_2 \qquad \text{(because } V \text{ is square, orthonormal)} \\
&\leqslant \|\Psi^{-1}\|_2^2\|(U^TB^{-1}U)^{-1}\|_2 \qquad \text{(sub-multiplicative property)} \\
&= \sigma_{\max}^2(\Psi^{-1})\|(U^TB^{-1}U)^{-1}\|_2 \\
&= \|(U^TB^{-1}U)^{-1}\|_2/\sigma_{\min}^2(\Psi) = \|(U^TB^{-1}U)^{-1}\|_2/\sigma_{\min}^2(D),
\end{aligned}$$

where $\sigma_{\min}(D) = \sigma_{\min}(\Psi)$, because Ψ is the diagonal matrix of singular values of D. Next, note that $\|(U^TB^{-1}U)^{-1}\|_2 = 1/\varphi_{\min}(U^TB^{-1}U)$ and

$$\varphi_{\min}(U^TB^{-1}U) \;=\; \min_{\eta\in\mathbb{R}^2}\eta^TU^TB^{-1}U\eta/\eta^T\eta.$$

We perform the change of variables $\gamma = U\eta$, under which $\eta^T\eta = \gamma^TU^TU\gamma = \gamma^T\gamma$ (that is, U preserves the length of η because the columns of U are orthonormal). Hence

$$\begin{aligned}
\varphi_{\min}(U^TB^{-1}U) &= \min_{\gamma\in\mathrm{col}(U),\gamma\neq 0}\gamma^TB^{-1}\gamma/\gamma^T\gamma \\
&\geqslant \min_{\gamma\neq 0}\gamma^TB^{-1}\gamma/\gamma^T\gamma \\
&= \varphi_{\min}(B^{-1}) = 1/\|B\|_2.
\end{aligned}$$

We have shown that $1/\varphi_{\min}(U^TB^{-1}U) \leqslant \|B\|_2$, which implies that

$$\|(U^TB^{-1}U)^{-1}\|_2 \leqslant \|B\|_2.$$

Therefore

$$\|(D^TB^{-1}D)^{-1}\|_2 \leqslant \|B\|_2/\sigma_{\min}^2(D).$$

In the special case of the two-group design matrix, $\sigma^2_{\min}(D) = n_{\min}$, so $\|(D^T B^{-1} D)^{-1}\|_2 \leqslant \|B\|_2 / n_{\min}$.

The proof of (3.12) is as follows:

$$\lambda_{\min}(\Omega) = \frac{1}{\lambda_{\max}(\Omega^{-1})} = \frac{1}{\lambda_{\max}(D^T B^{-1} D)} \geqslant \frac{1}{\|D\|_2^2 \, \lambda_{\max}(B^{-1})} = \frac{\lambda_{\min}(B)}{\|D\|_2^2} = \frac{\lambda_{\min}(B)}{n_{\max}}.$$

□

3.2.6 Proof of Proposition III.3

By the definitions of $\Omega_{n,m}$ in (3.2) and $\Delta = B_{n,m}^{-1} - B^{-1}$, we have by Theorem III.2

$$\begin{aligned}
\|\Omega_{n,m} - \Omega\|_2 &= \|(D^T B_{n,m} D)^{-1} - \Omega\|_2 \\
&= \left\| \left(D^T B_{n,m}^{-1} D - D^T B^{-1} D + D^T B^{-1} D \right)^{-1} - \Omega \right\|_2 \\
&= \left\| \left(D^T B^{-1} D + D^T \Delta D \right)^{-1} - \Omega \right\|_2 \\
&\leqslant \frac{\|D^T \Delta D\|_2 \|\Omega\|_2^2}{1 - \|\Omega\|_2 \|D^T \Delta D\|_2} \qquad \text{(by Theorem III.2)} \\
&\leqslant \frac{(\sigma^2_{\max}(D)/\sigma^4_{\min}(D)) \, \|B\|_2^2 \|\Delta\|_2)}{1 - \kappa^2(D) \|B\|_2 \|\Delta\|_2}.
\end{aligned}$$

In the last step we apply Proposition III.1. Thus

$$\begin{aligned}
\|\Omega_{n,m} - \Omega\|_2 &\leqslant \frac{1}{\sigma^2_{\min}(D)} \frac{\kappa^2(D) \|B\|_2^2 \|\Delta\|_2}{1 - \kappa^2(D) \|B\|_2 \|\Delta\|_2} \\
&= \frac{1}{\sigma^2_{\min}(D)} \frac{\|B\|_2^2 \|\Delta\|_2}{(1/\kappa^2(D)) - \|B\|_2 \|\Delta\|_2}.
\end{aligned}$$

We prove (3.14) using the submultiplicative property of the operator norm and Proposition III.1:

$$\left\| \Omega D^T \Delta \right\|_2 \leqslant \frac{\|B\|_2}{\sigma^2_{\min}(D)} \sigma_{\max}(D) \|\Delta\|_2 = \frac{\kappa(D)}{\sigma_{\min}(D)} \|B\|_2 \|\Delta\|_2.$$

We prove (3.15), as follows:

$$\begin{aligned}
\left\|\left(\Omega_{n,m}-\Omega\right)D^T\Delta\right\|_2 &\leqslant \left\|\Omega_{n,m}-\Omega\right\|_2 \left\|D^T\right\|_2 \|\Delta\|_2 \\
&\leqslant \left[\frac{1}{\sigma^2_{\min}(D)}\frac{\|B\|_2^2\|\Delta\|_2}{(1/\kappa^2(D))-\|B\|_2\|\Delta\|_2}\right]\sigma_{\max}(D)\|\Delta\|_2 \qquad \text{(by Proposition III.3)} \\
&= \frac{\kappa(D)}{\sigma_{\min}(D)}\frac{\|B\|_2^2\|\Delta\|_2^2}{(1/\kappa^2(D))-\|B\|_2\|\Delta\|_2}.
\end{aligned}$$

The proof of (3.16) is analogous. □

3.3 Proof of Theorem II.3

Note that the proof in the current Section follows exactly the same steps as the proof of Theorems 3.1 and 3.2 in *Zhou* (2014a). Theorem II.3 **Part II** is proved in Section 3.3.2. To prove Theorem II.3 **Part I**, we first state Lemma III.7, which establishes rates of convergence for estimating A^{-1} and B^{-1} in the operator and the Frobenius norm. We then state the auxiliary Lemma III.8, which is identical to that for Theorems 11.1 and 11.2 of *Zhou* (2014a), except that we plug in $\widetilde{\alpha}$ and $\widetilde{\eta}$ as defined in (3.17). Putting these results together proves Theorem II.3, **Part I.** We prove these auxiliary results in Section 3.4.

Let $\mathcal{X}_0$ denote the event

$$\forall i,j \qquad \left|\frac{(e_i-p_i)^T XX^T(e_j-p_j)}{\operatorname{tr}(A^*)\sqrt{b^*_{ii}b^*_{jj}}}-\rho_{ij}(B)\right| \leqslant \widetilde{\alpha} \tag{3.26}$$

$$\forall i,j \qquad \left|\frac{X_i^T(I-P_2)X_j}{\operatorname{tr}(B^*)\sqrt{a^*_{ii}a^*_{jj}}}-\rho_{ij}(A)\right| \leqslant \widetilde{\eta}, \tag{3.27}$$

with $\mathcal{X}_0(B)$ and $\mathcal{X}_0(A)$ denoting the events defined by equations (3.26) and (3.27), respectively.

Let $\widetilde{\alpha}$ and $\widetilde{\eta}$ be as defined in (3.17). On event $\mathcal{X}_0(A)$, for all j, $\widehat{\Gamma}_{jj}(A)=\rho_{jj}(A)=1$

and

$$\max_{j,k,j\neq k} |\widehat{\Gamma}_{jk}(A) - \rho_{jk}(A)| \leqslant \frac{2\widetilde{\eta}}{1-\widetilde{\eta}} \tag{3.28}$$

On event $\mathcal{X}_0(B)$, for all j, $\widehat{\Gamma}_{jj}(B) = \rho_{jj}(B) = 1$ and

$$\max_{j,k,j\neq k} |\widehat{\Gamma}_{jk}(B) - \rho_{jk}(B)| \leqslant \frac{2\widetilde{\alpha}}{1-\widetilde{\alpha}}. \tag{3.29}$$

Lemma III.7. *Suppose (A1) and (A2) hold. Let $\widehat{W}_1$ and $\widehat{W}_2$ be as defined in* (2.10)*. Let $\widehat{A}_\rho$ and $\widehat{B}_\rho$ be as defined in (2.8a) and (2.8b). For some absolute constants $18 < C, C' < 36$, the following events hold with probability at least $1 - 2/(n \vee m)^2$,*

$$\delta_{A,2} := \|\widehat{W}_1\widehat{A}_\rho\widehat{W}_1/\operatorname{tr}(B) - A\|_2 \leqslant Ca_{\max}\kappa(\rho(A))^2\lambda_B\sqrt{|A^{-1}|_{0,off} \vee 1} \tag{3.30}$$

$$\delta_{B,2} := \|\widehat{W}_2\widehat{B}_\rho\widehat{W}_2/\operatorname{tr}(A) - B\|_2 \leqslant C'b_{\max}\kappa(\rho(B))^2\lambda_A\sqrt{|B^{-1}|_{0,off} \vee 1} \tag{3.31}$$

$$\delta_{A,F} := \|\widehat{W}_1\widehat{A}_\rho\widehat{W}_1/\operatorname{tr}(B) - A\|_F \leqslant Ca_{\max}\kappa(\rho(A))^2\lambda_B\sqrt{|A^{-1}|_{0,off} \vee m} \tag{3.32}$$

$$\delta_{B,F} := \|\widehat{W}_2\widehat{B}_\rho\widehat{W}_2/\operatorname{tr}(A) - B\|_F \leqslant C'b_{\max}\kappa(\rho(B))^2\lambda_A\sqrt{|B^{-1}|_{0,off} \vee n}; \tag{3.33}$$

and for some $10 < C, C' < 19$,

$$\begin{aligned}
\delta^-_{A,2} &:= \left\|\operatorname{tr}(B)\left(\widehat{W}_1\widehat{A}_\rho\widehat{W}_1\right)^{-1} - A^{-1}\right\|_2 \leqslant \frac{C\lambda_B\sqrt{|A^{-1}|_{0,off} \vee 1}}{a_{\min}\varphi^2_{\min}(\rho(A))} \\
\delta^-_{B,2} &:= \left\|\operatorname{tr}(A)\left(\widehat{W}_2\widehat{B}_\rho\widehat{W}_2\right)^{-1} - B^{-1}\right\|_2 \leqslant \frac{C'\lambda_A\sqrt{|B^{-1}|_{0,off} \vee 1}}{b_{\min}\varphi^2_{\min}(\rho(B))} \\
\delta^-_{A,F} &:= \left\|\operatorname{tr}(B)\left(\widehat{W}_1\widehat{A}_\rho\widehat{W}_1\right)^{-1} - A^{-1}\right\|_F \leqslant \frac{C\lambda_B\sqrt{|A^{-1}|_{0,off} \vee m}}{a_{\min}\varphi^2_{\min}(\rho(A))} \\
\delta^-_{B,F} &:= \left\|\operatorname{tr}(A)\left(\widehat{W}_2\widehat{B}_\rho\widehat{W}_2\right)^{-1} - B^{-1}\right\|_F \leqslant \frac{C'\lambda_A\sqrt{|B^{-1}|_{0,off} \vee n}}{b_{\min}\varphi^2_{\min}(\rho(B))}.
\end{aligned}$$

Lemma III.8 follows from Theorems 11.1 and 11.2 of *Zhou* (2014a,b), where we now plug in $\widetilde{\alpha}$ and $\widetilde{\eta}$ as defined in (3.17). For completeness, we provide a sketch in

Section 3.4.2.

Lemma III.8. *Suppose (A1) and (A2) hold. For $\varepsilon_1, \varepsilon_2 \in (0,1)$, let*

$$\lambda_A = \widetilde{\eta}/\varepsilon_1, \quad \lambda_B = \widetilde{\alpha}/\varepsilon_2,$$

for $\widetilde{\alpha}$, $\widetilde{\eta}$ as defined in (3.17), and suppose $\lambda_A, \lambda_B < 1$. Then on event $\mathcal{X}_0$, for $18 < C, C' < 36$,

$$\begin{aligned}
&\|\widehat{A \otimes B} - A \otimes B\|_2 \leqslant \frac{\lambda_A \wedge \lambda_B}{2}\|A\|_2\|B\|_2 + C\lambda_B a_{\max}\|B\|_2\kappa(\rho(A))^2\sqrt{|A^{-1}|_{0,off} \vee 1} \\
&+C'\lambda_A b_{\max}\|A\|_2\kappa(\rho(B))^2\sqrt{|B^{-1}|_{0,off} \vee 1} \\
&+2\left[C'\lambda_A b_{\max}\kappa(\rho(B))^2\sqrt{|B^{-1}|_{0,off} \vee 1}\right]\left[C\lambda_B a_{\max}\kappa(\rho(A))^2\sqrt{|A^{-1}|_{0,off} \vee 1}\right],
\end{aligned}$$

and for $10 < C, C' < 19$,

$$\begin{aligned}
\|\widehat{A \otimes B}^{-1} - A^{-1} \otimes B^{-1}\|_2 &\leqslant \frac{\lambda_A \wedge \lambda_B}{3}\|A^{-1}\|_2\|B^{-1}\|_2 + C\lambda_B\|B^{-1}\|_2\frac{\sqrt{|A^{-1}|_{0,off} \vee 1}}{a_{\min}\varphi^2_{\min}(\rho(A))} \\
&+ C'\lambda_A\|A^{-1}\|_2\frac{\sqrt{|B^{-1}|_{0,off} \vee 1}}{b_{\min}\varphi^2_{\min}(\rho(B))} + \frac{3}{2}\left[C\lambda_B\frac{\sqrt{|A^{-1}|_{0,off} \vee 1}}{a_{\min}\varphi^2_{\min}(\rho(A))}\right]\left[C'\lambda_A\frac{\sqrt{|B^{-1}|_{0,off} \vee 1}}{b_{\min}\varphi^2_{\min}(\rho(B))}\right];
\end{aligned}$$

For $18 < C, C' < 36$,

$$\begin{aligned}
&\|\widehat{A \otimes B} - A \otimes B\|_F \leqslant \frac{\lambda_A \wedge \lambda_B}{2}\|A\|_F\|B\|_F + C\lambda_B a_{\max}\|B\|_F\kappa(\rho(A))^2\sqrt{|A^{-1}|_{0,off} \vee m} \\
&+C'\lambda_A b_{\max}\|A\|_F\kappa(\rho(B))^2\sqrt{|B^{-1}|_{0,off} \vee n} \\
&+2\left[C'\lambda_A b_{\max}\kappa(\rho(B))^2\sqrt{|B^{-1}|_{0,off} \vee n}\right]\left[C\lambda_B a_{\max}\kappa(\rho(A))^2\sqrt{|A^{-1}|_{0,off} \vee m}\right],
\end{aligned}$$

and for $10 < C, C' < 19$,

$$\|\widehat{A \otimes B}^{-1} - A^{-1} \otimes B^{-1}\|_F \leqslant \frac{\lambda_A \wedge \lambda_B}{3}\|A^{-1}\|_2\|B^{-1}\|_F + C\lambda_B\|B^{-1}\|_F\frac{\sqrt{|A^{-1}|_{0,off} \vee m}}{a_{\min}\varphi^2_{\min}(\rho(A))}$$
$$+ C'\lambda_A\|A^{-1}\|_F\frac{\sqrt{|B^{-1}|_{0,off} \vee n}}{b_{\min}\varphi^2_{\min}(\rho(B))} + \frac{7}{5}\left[C\lambda_B\frac{\sqrt{|A^{-1}|_{0,off} \vee m}}{a_{\min}\varphi^2_{\min}(\rho(A))}\right]\left[C'\lambda_A\frac{\sqrt{|B^{-1}|_{0,off} \vee n}}{b_{\min}\varphi^2_{\min}(\rho(B))}\right].$$

3.3.1 Proof of Theorem II.3, Part I

We state additional helpful bounds:

$$(a_{\min} \vee \varphi_{\min}(A))\sqrt{m} \leqslant \|A\|_F = \left(\sum_{i=1}^m \varphi_i^2(A)\right)^{1/2} \leqslant \sqrt{m}\|A\|_2, \tag{3.34}$$

$$(b_{\min} \vee \varphi_{\min}(B))\sqrt{n} \leqslant \|B\|_F = \left(\sum_{i=1}^m \varphi_i^2(B)\right)^{1/2} \leqslant \sqrt{n}\|B\|_2, \tag{3.35}$$

$$\sqrt{m}/a_{\max} = \left(\frac{1}{a_{\max}} \vee \frac{1}{\varphi_{\max}(A)}\right)\sqrt{m} \leqslant \|A^{-1}\|_F \leqslant \sqrt{m}\|A^{-1}\|_2, \tag{3.36}$$

and

$$\sqrt{n}/b_{\max} = \left(\frac{1}{b_{\max}} \vee \frac{1}{\varphi_{\max}(B)}\right)\sqrt{n} \leqslant \|B^{-1}\|_F \leqslant \sqrt{n}\|B^{-1}\|_2. \tag{3.37}$$

Proof of Theorem II.3, **Part I**. We plug in bounds as in (3.9) and (3.10) into Lemma III.8 to obtain under (A1) and (A2), $\left\|\widehat{A \otimes B} - A \otimes B\right\|_2 \leqslant \|A\|_2\|B\|_2\delta$, where

$$\begin{aligned}\delta &= \frac{\lambda_A \wedge \lambda_B}{2} + \frac{Cr_a\kappa(\rho(A))}{\varphi_{\min}(\rho(A))}\lambda_B\sqrt{|A^{-1}|_{0,\text{off}} \vee 1} + \frac{C'r_b\kappa(\rho(B))}{\varphi_{\min}(\rho(B))}\lambda_A\sqrt{|B^{-1}|_{0,\text{off}} \vee 1} \\ &\quad + 2\left[\frac{Cr_a\kappa(\rho(A))}{\varphi_{\min}(\rho(A))}\lambda_B\sqrt{|A^{-1}|_{0,\text{off}} \vee 1}\right]\left[\frac{C'r_b\kappa(\rho(B))}{\varphi_{\min}(\rho(B))}\lambda_A\sqrt{|B^{-1}|_{0,\text{off}} \vee 1}\right] \\ &= \frac{\lambda_A \wedge \lambda_B}{2} + \log^{1/2}(m \vee n)\left(\sqrt{\frac{|A^{-1}|_{0,\text{off}} \vee 1}{m}} + \sqrt{\frac{|B^{-1}|_{0,\text{off}} \vee 1}{n}}\right) + o(1).\end{aligned}$$

For the inverse, we plug in bounds as in (3.7) and (3.8) into Lemma III.8 to obtain

under (A1) and (A2), $\left\| \widehat{A \otimes B}^{-1} - A^{-1} \otimes B^{-1} \right\|_2 \leqslant \|A^{-1}\|_2 \|B^{-1}\|_2 \delta'$, where

$$\begin{aligned}
\delta' &= \frac{\lambda_A \wedge \lambda_B}{3} + \frac{C r_a \lambda_B \sqrt{|A^{-1}|_{0,\text{off}} \vee 1}}{\varphi_{\min}(\rho(A))} + \frac{C' r_b \lambda_A \sqrt{|B^{-1}|_{0,\text{off}} \vee 1}}{\varphi_{\min}(\rho(B))} \\
&\quad + \frac{3}{2} \left[\frac{C r_a \lambda_B \sqrt{|A^{-1}|_{0,\text{off}} \vee 1}}{\varphi_{\min}(\rho(A))} \right] \left[\frac{C' r_b \lambda_A \sqrt{|B^{-1}|_{0,\text{off}} \vee 1}}{\varphi_{\min}(\rho(B))} \right] \\
&\asymp \frac{\lambda_A \wedge \lambda_B}{3} + \log^{1/2}(m \vee n) \left(\sqrt{\frac{|A^{-1}|_{0,\text{off}} \vee 1}{m}} + \sqrt{\frac{|B^{-1}|_{0,\text{off}} \vee 1}{n}} \right) + o(1).
\end{aligned}$$

The bounds in the Frobenius norm are proved in a similar manner; see *Zhou* (2014a) to finish. □

3.3.2 Proof of Theorem II.3, Part II

Let $\widehat{B}^{-1} = \widehat{W}_2 \widehat{B}_\rho \widehat{W}_2$. Let $\widehat{\Delta} = \widehat{B}^{-1} - B^{-1}$. Let $\mathcal{E}_0(B)$ denote the event given by equations (3.34) and (3.34), which we know has probability at least $1 - 2/(n \vee m)^2$ from Lemma III.7, and define the event

$$\mathcal{E}_4 = \left\{ \|\widehat{\beta}_j(\widehat{B}^{-1}) - \beta_j^*\|_2 \leqslant s_{n,m} + t'_{n,m} \right\}, \tag{3.38}$$

where $s_{n,m}$ is as defined in (2.19) and

$$t'_{n,m} := C \lambda_A \sqrt{\frac{n_{\text{ratio}} \left(|B_0^{-1}|_{0,\text{off}} \vee 1 \right)}{n_{\min}}}. \tag{3.39}$$

Under $\mathcal{E}_0(B)$, we see that

$$\|\widehat{\Delta}\|_2 \leqslant \frac{C' \lambda_A \sqrt{|B^{-1}|_{0,\text{off}} \vee 1}}{b_{\min} \varphi_{\min}^2(\rho(B))} = o(1). \tag{3.40}$$

Using Proposition III.1 and the fact that $\|D\|_2 = \sqrt{n_{\max}}$, we get that

$$\|\Omega D^T \widehat{\Delta} D\|_2 \leqslant n_{\text{ratio}} \|B\|_2 \|\widehat{\Delta}\|_2, \tag{3.41}$$

From (3.40) we know that $\|\widehat{\Delta}\|_2 \leqslant 1/(n_{\text{ratio}}\|B\|_2)$, which we can plug into (3.41) to show that $\|\Omega D^T \widehat{\Delta} D\|_2 < 1$. This implies that $\widetilde{C} n_{\min}^{-1/2} \|\widehat{\Delta}\|_2 \leqslant t'_{n,m}$. Therefore, we can apply Theorem II.1 to get that the conditional probability of $\mathcal{E}_4$ given $\mathcal{E}_0(B)$ is at least $1 - 4/(n \vee m)^2$.

We can then bound the unconditional probability,

$$\begin{aligned} P(\mathcal{E}_4^c) &\leqslant P\left(\mathcal{E}_4^c \mid \mathcal{E}_0(B)\right) P\left(\mathcal{E}_0(B)\right) + P\left(\mathcal{E}_0(B)^c\right) \\ &\leqslant P\left(\mathcal{E}_4^c \mid \mathcal{E}_0(B)\right) + P\left(\mathcal{E}_0(B)^c\right) \\ &\leqslant \frac{4}{(n \vee m)^2} + \frac{2}{(n \vee m)^2}. \end{aligned}$$

□

3.4 More proofs for Theorem II.3

The proof of Lemma III.7 appears in Section 3.4.1. The proofs of auxiliary lemmas appear in Section 3.4.2.

3.4.1 Proof of Lemma III.7

In order to prove Lemma III.7, we need Theorem III.9, which shows explicit non-asymptotic convergence rates in the Frobenius norm for estimating $\rho(A)$, $\rho(B)$, and their inverses. Theorem III.9 follows from the standard proof; see *Rothman et al.* (2008); *Zhou et al.* (2011) We also need Proposition III.11 and Lemma III.10, which are stated below and proved in Section 3.4.2.

Theorem III.9. *Suppose that (A2) holds. Let $\widehat{A}_\rho$ and $\widehat{B}_\rho$ be the unique minimizers defined by* (2.8a) *and* (2.8b) *with sample correlation matrices $\widehat{\Gamma}(A)$ and $\widehat{\Gamma}(B)$ as their input.*

Suppose that event $\mathcal{X}_0$ holds, with

$$\widetilde{\eta}\sqrt{|A^{-1}|_{0,off} \vee 1} = o(1) \quad \textit{and} \quad \widetilde{\alpha}\sqrt{|B^{-1}|_{0,off} \vee 1} = o(1).$$
$$\textit{Set for some } \ 0 < \epsilon, \varepsilon < 1, \ \ \lambda_B = \widetilde{\alpha}/\varepsilon \ \ \textit{and} \ \ \lambda_A = \widetilde{\eta}/\epsilon. \tag{3.42}$$

Then on event $\mathcal{X}_0$, we have for $9 < C < 18$

$$\left\|\widehat{A}_\rho - \rho(A)\right\|_2 \leqslant \left\|\widehat{A}_\rho - \rho(A)\right\|_F \leqslant C\kappa(\rho(A))^2\lambda_B\sqrt{|A^{-1}|_{0,off} \vee 1},$$
$$\left\|\widehat{B}_\rho - \rho(B)\right\|_2 \leqslant \left\|\widehat{B}_\rho - \rho(B)\right\|_F \leqslant C\kappa(\rho(B))^2\lambda_A\sqrt{|B^{-1}|_{0,off} \vee 1},$$

and

$$\left\|\widehat{A}_\rho^{-1} - \rho(A)^{-1}\right\|_2 \leqslant \left\|\widehat{A}_\rho^{-1} - \rho(A)^{-1}\right\|_F < \frac{C\lambda_B\sqrt{|A^{-1}|_{0,off} \vee 1}}{2\varphi^2_{\min}(\rho(A))}, \tag{3.43}$$

$$\left\|\widehat{B}_\rho^{-1} - \rho(B)^{-1}\right\|_2 \leqslant \left\|\widehat{B}_\rho^{-1} - \rho(B)^{-1}\right\|_F \leqslant \frac{C\lambda_A\sqrt{|B^{-1}|_{0,off} \vee 1}}{2\varphi^2_{\min}(\rho(B))}. \tag{3.44}$$

We now state an auxiliary result, Lemma III.10, where we prove a bound on the error in the diagonal entries of the covariance matrices, and on their reciprocals. The following Lemma provides bounds analogous to those in Claim 15.1 *Zhou* (2014a,b).

Lemma III.10. *Let $\widehat{W}_1$ and $\widehat{W}_2$ be as defined in* (2.10). *Let $W_1 = \sqrt{\mathrm{tr}(B)}\,\mathrm{diag}(A)^{1/2}$ and $W_2 = \sqrt{\mathrm{tr}(A)}\,\mathrm{diag}(B)^{1/2}$. Suppose event $\mathcal{X}_0$ holds, as defined in* (3.26), (3.27).

For $\eta' := \frac{\widetilde{\eta}}{\sqrt{1-\widetilde{\eta}}} \leqslant \frac{\lambda_B}{6}$ *and* $\alpha' := \frac{\widetilde{\alpha}}{\sqrt{1-\widetilde{\alpha}}} \leqslant \frac{\lambda_A}{6}$,

$$\left\|\widehat{W}_1 - W_1\right\|_2 \leqslant \widetilde{\eta}\sqrt{\operatorname{tr}(B)}\sqrt{a_{\max}}, \qquad \left\|\widehat{W}_1^{-1} - W_1^{-1}\right\|_2 \leqslant \frac{\widetilde{\eta}}{1-\widetilde{\eta}}/\sqrt{\operatorname{tr}(B)}\sqrt{a_{\min}},$$
$$\left\|\widehat{W}_2 - W_2\right\|_2 \leqslant \widetilde{\alpha}\sqrt{\operatorname{tr}(A)}\sqrt{b_{\max}}, \text{ and } \left\|\widehat{W}_2^{-1} - W_2^{-1}\right\|_2 \leqslant \frac{\widetilde{\alpha}}{1-\widetilde{\alpha}}/\sqrt{\operatorname{tr}(A)}\sqrt{b_{\min}}.$$

Proposition III.11. *(Zhou, 2014a). Let* $\widehat{W}$ *and* W *be diagonal positive definite matrices. Let* $\widehat{\Psi}$ *and* Ψ *be symmetric positive definite matrices. Then*

$$\begin{aligned}\left\|\widehat{W}\widehat{\Psi}\widehat{W} - W\Psi W\right\|_2 &\leqslant \left(\left\|\widehat{W} - W\right\|_2 + \|W\|_2\right)^2 \left\|\widehat{\Psi} - \Psi\right\|_2 \\ &\quad + \left\|\widehat{W} - W\right\|_2 \left(\left\|\widehat{W} - W\right\|_2 + 2\right) \|\Psi\|_2 \\ \left\|\widehat{W}\widehat{\Psi}\widehat{W} - W\Psi W\right\|_F &\leqslant \left(\left\|\widehat{W} - W\right\|_2 + \|W\|_2\right)^2 \left\|\widehat{\Psi} - \Psi\right\|_F \\ &\quad + \left\|\widehat{W} - W\right\|_2 \left(\left\|\widehat{W} - W\right\|_2 + 2\right) \|\Psi\|_F.\end{aligned}$$

Proof of Lemma III.7. Assume that event $\mathcal{X}_0$ holds. The proof follows exactly that of Lemma 15.3 in *Zhou* (2014a,b), in view of Theorem III.9, Lemma III.10 and Proposition 15.2 from *Zhou* (2014a,b), which is restated immediately above in Proposition III.11. □

It remains to prove Lemma III.10.

Proof of Lemma III.10. Suppose that event $\mathcal{X}_0$ holds. Then

$$\max_{i=1,\ldots,m} \left| \frac{\sqrt{X_i^T(I-P_2)X_i}}{\sqrt{a_{ii}\operatorname{tr}(B)}} - 1 \right| \leqslant \left(1 - \sqrt{1-\widetilde{\eta}}\right) \bigvee \left(\sqrt{1+\widetilde{\eta}} - 1\right) \leqslant \widetilde{\eta}.$$

Thus for all i,

$$\frac{1}{\sqrt{1+\widetilde{\eta}}} \leqslant \frac{\sqrt{a_{ii}\operatorname{tr}(B)}}{\sqrt{X_i^T(I-P_2)X_i}} \leqslant \frac{1}{\sqrt{1-\widetilde{\eta}}},$$

so

$$\left| \frac{\sqrt{a_{ii}\operatorname{tr}(B)}}{\sqrt{X_i^T(I-P_2)X_i}} - 1 \right| \leqslant \left(\frac{1-\sqrt{1-\widetilde{\eta}}}{\sqrt{1-\widetilde{\eta}}} \right) \bigvee \left(\frac{\sqrt{1+\widetilde{\eta}}-1}{\sqrt{1+\widetilde{\eta}}} \right) \leqslant \frac{\widetilde{\eta}}{\sqrt{1-\widetilde{\eta}}}.$$

□

3.4.2 Proof of Lemma III.8

In order to prove Lemma III.8, we state Lemma III.12, Lemma III.13, and Proposition III.14. Let $\|\cdot\|$ denote a matrix norm such that $\|A \otimes B\| = \|A\|\|B\|$. Let

$$\Delta := \widehat{W}_1 \widehat{A}_\rho \widehat{W}_1 \otimes \widehat{W}_2 \widehat{B}_\rho \widehat{W}_2 / \operatorname{tr}(A)\operatorname{tr}(B) - A \otimes B, \tag{3.45}$$

$$\Delta' := \operatorname{tr}(A)\operatorname{tr}(B) \left(\widehat{W}_1 \widehat{A}_\rho \widehat{W}_1 \right)^{-1} \otimes \left(\widehat{W}_2 \widehat{B}_\rho \widehat{W}_2 \right)^{-1} - A^{-1} \otimes B^{-1}. \tag{3.46}$$

Lemma III.12 is identical to Lemma 15.5 of *Zhou* (2014a), except that we now plug in quantities $\widetilde{\alpha}$ and $\widetilde{\eta}$ as defined in (3.17). Likewise, Proposition III.14 is analogous to (20) in Theorem 4.1 of *Zhou* (2014a), except that we now use the centered data matrix $(I - P_2)X$, together with the rates $\widetilde{\alpha}$, $\widetilde{\eta}$.

Lemma III.12. *Let $\widehat{A \otimes B}$ be as in (2.11). Then for $\Sigma = A \otimes B$,*

$$\left\| \widehat{A \otimes B}^{-1} - \Sigma^{-1} \right\| \leqslant (\widetilde{\alpha} \wedge \widetilde{\eta}) \|A^{-1}\| \|B^{-1}\| + (1 + \widetilde{\alpha} \wedge \widetilde{\eta}) \|\Delta'\| \tag{3.47}$$

$$\left\| \widehat{A \otimes B} - \Sigma \right\| \leqslant \frac{\lambda_A \wedge \lambda_B}{2} \|A\| \|B\| + (1 + \frac{\lambda_A \wedge \lambda_B}{2}) \|\Delta\|. \tag{3.48}$$

Lemma III.13 is a helpful bound on the difference of Kronecker products.

Lemma III.13. *(Zhou, 2014a). For matrices A_1 and B_1, let $\Delta_A := A_1 - A$ and $\Delta_B := B_1 - B$. Then*

$$\|A_1 \otimes B_1 - A \otimes B\| \leqslant \|\Delta_A\| \|B\| + \|\Delta_B\| \|A\| + \|\Delta_A\| \|\Delta_B\|.$$

Proposition III.14. *Under the event $\mathcal{X}_0$, as defined in as defined in* (3.26), (3.27),

$$\left|\|(I-P_2)X\|_F^2 - \mathrm{tr}(A)\mathrm{tr}(B)\right| \leqslant (\widetilde{\alpha}\wedge\widetilde{\eta})\mathrm{tr}(A)\mathrm{tr}(B).$$

Proof of Lemma III.8. Assume that event $\mathcal{X}_0$ as defined in (3.26), (3.27) holds. The proof follows exactly the steps in Theorems 11.1 and 11.2 in Supplementary Material of *Zhou* (2014a,b). □

Proof of Lemma III.12. By the triangle inequality and the sub-multiplicativity of the norm $\|\cdot\|$, with Δ and Δ' as defined in (3.45) and (3.46),

$$\mathrm{tr}(A)\,\mathrm{tr}(B)\left\|\left(\widehat{W}_1^{-1}\widehat{A}_\rho^{-1}\widehat{W}_1^{-1}\right)\otimes\left(\widehat{W}_2^{-1}\widehat{B}_\rho^{-1}\widehat{W}_2^{-1}\right)\right\| \leqslant \|A^{-1}\|\|B^{-1}\| + \|\Delta'\| \quad (3.49)$$

$$\left\|\left(\widehat{W}_1\widehat{A}_\rho\widehat{W}_1\right)\otimes\left(\widehat{W}_2\widehat{B}_\rho\widehat{W}_2\right)/\,\mathrm{tr}(A)\,\mathrm{tr}(B)\right\| \leqslant \|A\|\|B\| + \|\Delta\|. \quad (3.50)$$

Following proof of Lemma 15.5 *Zhou* (2014a,b), we have by definition of Δ', and Proposition III.14, and (3.49),

$$\left\|\widehat{A\otimes B}^{-1} - A^{-1}\otimes B^{-1}\right\| \;\leqslant\; (\widetilde{\alpha}\wedge\widetilde{\eta})\left(\|A^{-1}\|\|B^{-1}\| + \|\Delta'\|\right) + \|\Delta'\|.$$

By Proposition III.14, we have for $\lambda_A \geqslant 3\widetilde{\alpha}$, $\lambda_B \geqslant 3\widetilde{\eta}$, where $\widetilde{\alpha}\wedge\widetilde{\eta} \leqslant \frac{\lambda_A\wedge\lambda_B}{3}$,

$$\begin{aligned}
&\left|\frac{1}{\|(I-P_2)X\|_F^2} - \frac{1}{\mathrm{tr}(A)\,\mathrm{tr}(B)}\right| = \left|\frac{\|(I-P_2)X\|_F^2 - \mathrm{tr}(A)\,\mathrm{tr}(B)}{\|(I-P_2)X\|_F^2\,\mathrm{tr}(A)\,\mathrm{tr}(B)}\right| \\
&\leqslant \left|\frac{\widetilde{\alpha}\wedge\widetilde{\eta}}{\|(I-P_2)X\|_F^2}\right| \leqslant \frac{\widetilde{\alpha}\wedge\widetilde{\eta}}{\mathrm{tr}(A)\,\mathrm{tr}(B)(1-\widetilde{\alpha}\wedge\widetilde{\eta})} \\
&\text{thus } \left|\frac{\mathrm{tr}(A)\,\mathrm{tr}(B)}{\|(I-P_2)X\|_F^2} - 1\right| \leqslant \frac{\widetilde{\alpha}\wedge\widetilde{\eta}}{1-\widetilde{\alpha}\wedge\widetilde{\eta}} \leqslant \frac{\lambda_A\wedge\lambda_B}{2}. \quad (3.51)
\end{aligned}$$

By the triangle inequality, the definition of Δ in (3.45), and (3.50) and (3.51),

$$\left\|\widehat{A\otimes B} - A\otimes B\right\| \;\leqslant\; \frac{\lambda_A+\lambda_B}{2}\|A\|\|B\| + (1+\frac{\lambda_A+\lambda_B}{2})\|\Delta\|;$$

See the proof of Lemma 15.5 *Zhou* (2014a,b). □

Proof of Proposition III.14. Suppose event $\mathcal{X}_0$ holds. Note that

$$E[\|(I-P_2)X\|_F^2] = \mathrm{tr}\left((I-P_2)E[XX^T](I-P_2)\right) = \mathrm{tr}(A)\mathrm{tr}(\widetilde{B})$$

Decomposing by columns, we obtain the inequality,

$$\begin{aligned}\left|\|(I-P_2)X\|_F^2 - \mathrm{tr}(A)\mathrm{tr}(B)\right| &= \left|\sum_{j=1}^m \|(I-P_2)X_j\|_2^2 - a_{jj}\mathrm{tr}(B)\right| \\ \leqslant \sum_{j=1}^m \left|X_j^T(I-P_2)X_j - a_{jj}\mathrm{tr}(B)\right| &\leqslant \sum_{j=1}^m \widetilde{\eta}_{jj}a_{jj}\mathrm{tr}(B) \leqslant \widetilde{\eta}\mathrm{tr}(A)\mathrm{tr}(B).\end{aligned}$$

Decomposing by rows, we obtain the inequality,

$$\begin{aligned}\left|\|(I-P_2)X\|_F^2 - \mathrm{tr}(A)\mathrm{tr}(B)\right| &= \left|\sum_{i=1}^n \|(e_i-p_i)^T X\|_2^2 - b_{ii}\mathrm{tr}(A)\right| \\ \leqslant \sum_{i=1}^n \left|(e_i-p_i)^T XX^T(e_i-p_i) - b_{ii}\mathrm{tr}(A)\right| &\leqslant \sum_{i=1}^n \widetilde{\alpha}_{ii}b_{ii}\mathrm{tr}(A) \leqslant \widetilde{\alpha}\mathrm{tr}(A)\mathrm{tr}(B).\end{aligned}$$

Therefore $|\|(I-P_2)X\|_F^2 - \mathrm{tr}(A)\mathrm{tr}(B)| \leqslant (\widetilde{\alpha}\wedge\widetilde{\eta})\mathrm{tr}(A)\mathrm{tr}(B)$. □

3.5 Entrywise convergence of sample correlations

In this section we prove entrywise rates of convergence for the sample correlation matrices in Theorem III.15. The theorem applies to the Kronecker product model, $\mathrm{Cov}(\mathrm{vec}(X)) = A^* \otimes B^*$, where for identifiability we define the sample covariance matrices as

$$A^* = \frac{m}{\mathrm{tr}(A)}A \quad \text{and} \quad B^* = \frac{\mathrm{tr}(A)}{m}B,$$

with the scaling chosen so that A^* has trace m. Let $\rho(A) \in \mathbb{R}^{m\times m}$ and $\rho(B) \in \mathbb{R}^{n\times n}$ denote the correlation matrices corresponding to covariance matrices A^* and B^*, respectively. Assume that that the mean of X satisfies the two-group model (2.4).

Let P_2 be as defined in (2.13). The matrix $I - P_2$ is a projection matrix of rank $n-2$ that performs within-group centering. The sample covariance matrices are defined as

$$S(B^*) = \frac{1}{m}\sum_{j=1}^{m}(I - P_2)X_j X_j^T(I - P_2), \tag{3.52}$$

$$S(A^*) = X^T(I - P_2)X/n, \tag{3.53}$$

where $S(B^*)$ has null space of dimension two.

Theorem III.15. *Consider a data generating random matrix as in (2.2). Let C be some absolute constant. Let $\widetilde{\alpha}$ and $\widetilde{\eta}$ be as defined in* (3.17)*. Let $m \vee n \geqslant 2$. Then with probability at least $1 - \frac{3}{(m\vee n)^2}$, for $\widetilde{\alpha}, \widetilde{\eta} < 1/3$, and $\widehat{\Gamma}(A)$ and $\widehat{\Gamma}(B)$ as in (2.7),*

$$\forall i \neq j,\ \left|\widehat{\Gamma}_{ij}(B) - \rho_{ij}(B)\right| \leqslant \frac{\widetilde{\alpha}}{1-\widetilde{\alpha}} + |\rho_{ij}(B)|\,\frac{\widetilde{\alpha}}{1-\widetilde{\alpha}} \leqslant 3\widetilde{\alpha},$$
$$\forall i \neq j,\ \left|\widehat{\Gamma}_{ij}(A) - \rho_{ij}(A)\right| \leqslant \frac{\widetilde{\eta}}{1-\widetilde{\eta}} + |\rho_{ij}(A)|\,\frac{\widetilde{\eta}}{1-\widetilde{\eta}} \leqslant 3\widetilde{\eta}.$$

We state three results used in the proof of Theorem III.15: Proposition III.16 provides an entrywise rate of convergence of $S(B^*)$, Proposition III.17 provides an entrywise rate of convergence of $S(A^*)$, and Lemma III.18 states that these entrywise rates imply $\mathcal{X}_0$. Let

$$\widetilde{B} := (I - P_2)B^*(I - P_2) = \mathrm{Cov}((I - P_2)X_j), \tag{3.54}$$

where X_j is the jth column of X. Let $\widetilde{b}_{ij}$ denote the (i,j)th entry of $\widetilde{B}$.

Proposition III.16. *Let $d > 2$. Then with probability at least $1 - 2/m^{d-2}$,*

$$\forall i, j\ \left|S_{ij}(B^*) - b^*_{ij}\right| \leqslant \phi_{B,ij}, \tag{3.55}$$

with

$$\phi_{B,ij} = C\frac{\log^{1/2}(m)}{\sqrt{m}}\frac{\|A^*\|_F}{\sqrt{m}}\sqrt{\widetilde{b}_{ii}\widetilde{b}_{jj}} + \frac{3\|B^*\|_1}{n_{\min}}. \tag{3.56}$$

Proposition III.17. *Let $d > 2$. Then with probability at least $1 - 2/n^{d-2}$,*

$$\forall i, j \ \left|S_{ij}(A^*) - a^*_{ij}\,\mathrm{tr}\,(B^*)/n\right| > \phi_{A,ij}, \tag{3.57}$$

with

$$\phi_{A,ij} = (a^*_{ij}/n)\left|\mathrm{tr}\left(\widetilde{B}\right) - \mathrm{tr}\,(B^*)\right| + d^{1/2}K\log^{1/2}(n\vee m)(1/n)\sqrt{a^{*2}_{ij} + a^*_{ii}a^*_{jj}}\|\widetilde{B}\|_F. \tag{3.58}$$

Lemma III.18. *Suppose that (A2) holds and that $m \vee n \geqslant 2$. The event* (3.57) *defined in Proposition III.17 implies that $\mathcal{X}_0(A)$ holds. Similarly, the event* (3.55) *defined in Proposition III.16 implies $\mathcal{X}_0(B)$. Hence $\mathbb{P}(\mathcal{X}_0) \geqslant 1 - \frac{3}{(m\vee n)^2}$.*

Proposition III.16 is proved in section 3.5.1. Proposition III.17 is proved in section 3.5.2. Lemma III.18 is proved in section 3.5.3. Note that Lemma III.18 follows from Propositions III.16 and III.17. We now prove Theorem III.15, which follows from Lemma III.18.

Proof of Theorem III.15. Let q_i denote the ith column of $I - P_2$, so that $q_i^T XX^T q_j$ is the (i,j)th entry of $(I - P_2)XX^T(I - P_2)$. Under $\mathcal{X}_0(B)$, the sample

correlation $\widehat{\Gamma}(B)$ satisfies the following bound:

$$\begin{aligned}
\left|\widehat{\Gamma}_{ij}(B) - \rho_{ij}(B)\right| &= \left|\frac{q_i^T X X^T q_j}{\sqrt{q_i^T X X^T q_i}\sqrt{q_j^T X X^T q_j}} - \rho_{ij}(B)\right| \\
&= \left|\frac{q_i^T X X^T q_j / \left(\mathrm{tr}(A^*)\sqrt{b_{ii}^* b_{jj}^*}\right)}{\sqrt{q_i^T X X^T q_i / \left(b_{ii}^* \mathrm{tr}(A^*)\right)}\sqrt{q_j^T X X^T q_j / \left(b_{jj}^* \mathrm{tr}(A^*)\right)}} - \rho_{ij}(B)\right| \\
&\leqslant \left|\frac{q_i^T X X^T q_j / \left(\mathrm{tr}(A^*)\sqrt{b_{ii}^* b_{jj}^*}\right) - \rho_{ij}(B)}{\sqrt{q_i^T X X^T q_i / \left(b_{ii}^* \mathrm{tr}(A^*)\right)}\sqrt{q_j^T X X^T q_j / \left(b_{jj}^* \mathrm{tr}(A^*)\right)}}\right| \\
&+ \left|\frac{\rho_{ij}(B)}{\sqrt{q_i^T X X^T q_i / \left(b_{ii}^* \mathrm{tr}(A^*)\right)}\sqrt{q_j^T X X^T q_j / \left(b_{jj}^* \mathrm{tr}(A^*)\right)}} - \rho_{ij}(B)\right| \\
&\leqslant \frac{\widetilde{\alpha}}{1-\widetilde{\alpha}} + |\rho_{ij}(B)|\left|\frac{1}{1-\widetilde{\alpha}} - 1\right| \\
&\leqslant 3\widetilde{\alpha},
\end{aligned}$$

where the first inequality holds by $\mathcal{X}_0(B)$ and the second inequality holds for $\widetilde{\alpha} \leqslant 1/3$. Similarly, under $\mathcal{X}_0(A)$ we obtain an entrywise bound on the sample correlation $\widehat{\Gamma}(A)$:

$$\begin{aligned}
\left|\widehat{\Gamma}_{ij}(A) - \rho_{ij}(A)\right| &= \left|\frac{X_i^T (I-P_2) X_j}{\sqrt{X_i^T (I-P_2) X_i}\sqrt{X_j^T (I-P_2) X_j}} - \rho_{ij}(A)\right| \\
&= \left|\frac{X_i^T (I-P_2) X_j / \left(\mathrm{tr}(B^*)\sqrt{a_{ii}^* a_{jj}^*}\right)}{\sqrt{X_i^T (I-P_2) X_i / \left(a_{ii}^* \mathrm{tr}(B^*)\right)}\sqrt{X_j^T (I-P_2) X_j / \left(a_{jj}^* \mathrm{tr}(B^*)\right)}} - \rho_{ij}(A)\right| \\
&\leqslant \left|\frac{X_i^T (I-P_2) X_j / \left(\mathrm{tr}(B^*)\sqrt{a_{ii}^* a_{jj}^*}\right) - \rho_{ij}(A)}{\sqrt{X_i^T (I-P_2) X_i / \left(a_{ii}^* \mathrm{tr}(B^*)\right)}\sqrt{X_j^T (I-P_2) X_j / \left(a_{jj}^* \mathrm{tr}(B^*)\right)}}\right| \\
&+ \left|\frac{\rho_{ij}(A)}{\sqrt{X_i^T (I-P_2) X_i / \left(a_{ii}^* \mathrm{tr}(B^*)\right)}\sqrt{X_j^T (I-P_2) X_j / \left(a_{jj}^* \mathrm{tr}(B^*)\right)}} - \rho_{ij}(A)\right| \\
&\leqslant \frac{\widetilde{\eta}}{1-\widetilde{\eta}} + |\rho_{ij}(A)|\left|\frac{1}{1-\widetilde{\eta}} - 1\right| \leqslant 3\widetilde{\eta},
\end{aligned}$$

where the first inequality holds by $\mathcal{X}_0(A)$, and the second inequality holds for $\widetilde{\eta} < 1/3$.

By Lemma III.18, the event $\mathcal{X}_0 = \mathcal{X}_0(B) \cap \mathcal{X}_0(A)$ holds with probability at least $1 - 3/(n \vee m)^2$, which completes the proof. $\square$

3.5.1 Proof of Proposition III.16

We first present Lemma III.19 and Lemma III.20, which decompose the rate of convergence into a bias term and a variance term, respectively. We then combine the rates for the bias and variance terms to prove the entrywise rate of convergence for the sample covariance. Define

$$\mathcal{B}(B^*) := E[S(B^*)] - B^* \quad \text{and} \tag{3.59}$$

$$\sigma(B^*) := S(B^*) - E[S(B^*)]. \tag{3.60}$$

We state maximum entrywise bounds on $\mathcal{B}(B^*)$ and $\sigma(B^*)$ in Lemma III.19 and Lemma III.20, respectively. Proofs for these lemmas are provided in Section 3.5.4 and 3.5.5 respectively.

Lemma III.19. *For $\mathcal{B}(B^*)$ as defined in* (3.59)*,*

$$\|\mathcal{B}(B^*)\|_{\max} \leqslant \frac{3\|B^*\|_1}{n_{\min}}. \tag{3.61}$$

Lemma III.20. *Let $\sigma(B^*)$ be as defined in* (3.60)*. With probability at least $1-2/m^d$,*

$$|\sigma_{ij}(B^*)| = \left|S_{ij}(B^*) - b^*_{ij}\right| < C \log^{1/2}(m) \frac{\|A^*\|_F}{\operatorname{tr}(A^*)} \sqrt{\widetilde{b}_{ii}\widetilde{b}_{jj}}.$$

We now prove the entrywise rate of convergence for the sample covariance $S(B^*)$.

Proof of Proposition III.16. By the triangle inequality,

$$\begin{aligned}\left|S_{ij}(B^*) - b^*_{ij}\right| &\leqslant |S_{ij}(B^*) - E[S_{ij}(B^*)]| + \left|E[S_{ij}(B^*)] - b^*_{ij}\right| \\ &= |\mathcal{B}_{ij}(B^*)| + |\sigma_{ij}(B^*)| \\ &\leqslant \phi_{B,ij},\end{aligned}$$

where the last step follows from Lemmas III.19 and III.20. □

Remark. Note that the first term of (3.56) is of order $\log^{1/2}(m)/\sqrt{m}$, and the second term is of order $\|B^*\|_1/n_{\min}$.

3.5.2 Proof of Proposition III.17

We express the (i,j)th entry of $S(A^*)$ as a quadratic form in order to apply the Hanson-Wright inequality to obtain an entrywise large deviation bound. Without loss of generality, let $i = 1$, $j = 2$. The $(1,2)$ entry of $S(A^*)$ can be expressed as a quadratic form, as follows,

$$\begin{aligned}S_{12}(A^*) &= X_1^T(I - P_2)X_2/n \\ &= (1/2)\begin{bmatrix} X_1^T & X_2^T \end{bmatrix}\begin{bmatrix} 0 & (I-P_2) \\ (I-P_2) & 0 \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}/n \\ &= (1/2)\begin{bmatrix} X_1^T & X_2^T \end{bmatrix}\left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \otimes (I-P_2)\right)\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}/n.\end{aligned}$$

We decorrelate the random vector $(X_1, X_2) \in \mathbb{R}^{2n}$ so that we can apply the Hanson-Wright inequality. The covariance matrix used for decorrelation is

$$\mathrm{Cov}\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\right) = \begin{bmatrix} a^*_{11} & a^*_{12} \\ a^*_{21} & a^*_{22} \end{bmatrix} \otimes B^* =: A^*_{\{1,2\}} \otimes B^*,$$

with

$$A^*_{\{1,2\}} = \begin{bmatrix} a^*_{11} & a^*_{12} \\ a^*_{21} & a^*_{22} \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

Decorrelating the quadratic form yields

$$S_{12}(A^*) = Z^T \Phi Z,$$

where $Z \in \mathbb{R}^{2n}$, with $E[Z] = 0$ and $\mathrm{Cov}(Z) = I_{2n\times 2n}$, and

$$\Phi = (1/2n) \left((A^*_{\{1,2\}})^{1/2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} (A^*_{\{1,2\}})^{1/2} \right) \otimes B^{1/2}(I - P_2)B^{1/2}. \tag{3.62}$$

To apply the Hanson-Wright inequality, we first find the trace and Frobenius norm of Φ. For the trace, note that

$$\mathrm{tr}\left((A^*_{\{1,2\}})^{1/2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} (A^*_{\{1,2\}})^{1/2} \right) = \mathrm{tr}\left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} A^*_{\{1,2\}} \right) = 2a^*_{12}. \tag{3.63}$$

For the Frobenius norm, note that

$$\begin{aligned} \left\| (A^*_{\{1,2\}})^{1/2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} (A^*_{\{1,2\}})^{1/2} \right\|_F^2 &= \mathrm{tr}\left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} A^*_{\{1,2\}} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} A^*_{\{1,2\}} \right) \\ &= \mathrm{tr}\left(\begin{bmatrix} a^{*2}_{12} + a^*_{11}a^*_{22} & 2a^*_{12}a^*_{22} \\ 2a^*_{12}a^*_{22} & a^{*2}_{12} + a^*_{11}a^*_{22} \end{bmatrix} \right) \\ &= 2a^{*2}_{12} + 2a^*_{11}a^*_{22}, \end{aligned}$$

Therefore the trace of Φ is

$$\mathrm{tr}\,(\Phi) = a^*_{12}\,\mathrm{tr}\left(\widetilde{B}\right)/n, \tag{3.64}$$

and the Frobenius norm of Φ is

$$\|\Phi\|_F = (1/n)\sqrt{a_{12}^{*2} + a_{11}^* a_{22}^*}\|\widetilde{B}\|_F. \tag{3.65}$$

Applying the Hanson-Wright inequality yields

$$\begin{aligned}
&P\left(|S_{12}(A^*) - a_{12}^* \operatorname{tr}(B^*)/n| > \phi_{A,12}\right) \\
&\leqslant P\left(\left|S_{12}(A^*) - a_{12}^* \operatorname{tr}\left(\widetilde{B}\right)/n\right| + (a_{12}^*/n)\left|\operatorname{tr}\left(\widetilde{B}\right) - \operatorname{tr}(B^*)\right| > \phi_{A,12}\right) \\
&= P\left(\left|S_{12}(A) - a_{12}^* \operatorname{tr}\left(\widetilde{B}\right)/n\right| > d^{1/2}K\log^{1/2}(n \vee m)\|\Phi\|_F\right) \\
&\leqslant 2/(n \vee m)^d.
\end{aligned}$$

By the union bound,

$$\begin{aligned}
&P\left(\forall i, j\, |S_{ij}(A^*) - a_{ij}\operatorname{tr}(B^*)/n| < \phi_{A,ij}\right) \\
&\geqslant \quad 1 - \sum_{i=1}^{m}\sum_{j=1}^{m} P\left(|S_{ij}(A^*) - a_{ij}\operatorname{tr}(B^*)/n| > \phi_{A,ij}\right) \\
&\geqslant \quad 1 - 2m^2/(n \vee m)^d \geqslant 2/(n \vee m)^{d-2}.
\end{aligned}$$

□

3.5.3 Proof of Lemma III.18

For the event (3.55) from Proposition III.16,

$$\left|S_{ij}(B^*) - b_{ij}^*\right| < \phi_{B,ij} = K^2 d\frac{\log^{1/2}(m)}{\sqrt{m}}C_A\sqrt{\widetilde{b}_{ii}\widetilde{b}_{jj}} + \left|b_{ij}^* - \widetilde{b}_{ij}\right|,$$

dividing by $\sqrt{b_{ii}^* b_{jj}^*}$ yields

$$\left|\frac{q_i X X^T q_j}{\operatorname{tr}(A^*)\sqrt{b_{ii}^* b_{jj}^*}} - \rho_{ij}(B)\right| < K^2 d C_A\frac{\log^{1/2}(m)}{\sqrt{m}}\sqrt{\frac{\widetilde{b}_{ii}\widetilde{b}_{jj}}{b_{ii}^* b_{jj}^*}} + \frac{\left|b_{ij} - \widetilde{b}_{ij}\right|}{\sqrt{b_{ii}^* b_{jj}^*}}. \tag{3.66}$$

By Lemma III.19,

$$\widetilde{b}_{ij} = b_{ij}\left[1 + O\left(\frac{\|B\|_1}{n}\right)\right],$$

so the right-hand side of (3.66) is less than or equal to $\widetilde{\alpha}$. Hence event (3.55) implies $\mathcal{X}_0(B)$. Therefore, we know that $P(\mathcal{X}_0(B)) \geqslant 1 - 2/m^{d-2}$.

Similarly, event (3.57) in Proposition III.17:

$$\begin{aligned}
&\left|S_{ij}(A^*) - a^*_{ij}\,\mathrm{tr}\,(B^*)/n\right| < \phi_{A,ij} \\
&= (a^*_{ij}/n)\left|\mathrm{tr}\left(\widetilde{B}\right) - \mathrm{tr}\,(B)\right| + d^{1/2}K\log^{1/2}(n\vee m)(1/n)\sqrt{a^{*2}_{ij} + a^*_{ii}a^*_{jj}}\|\widetilde{B}\|_F,
\end{aligned}$$

implies that

$$\begin{aligned}
&\left|\frac{X_j^T(I-P_2)X_t}{\mathrm{tr}(B^*)\sqrt{a^*_{jj}a^*_{tt}}} - \rho_{jt}(A)\right| \\
&< |\rho_{jt}(A)|\frac{\left|\mathrm{tr}\left(\widetilde{B}\right) - \mathrm{tr}\,(B^*)\right|}{\mathrm{tr}(B^*)} + d^{1/2}K\log^{1/2}(n\vee m)\sqrt{\rho_{jt}(A)^2+1}\frac{\|\widetilde{B}\|_F}{\mathrm{tr}(B^*)} \\
&= |\rho_{jt}(A)|\frac{\left|\mathrm{tr}\left(\widetilde{B}\right) - \mathrm{tr}\,(B^*)\right|}{\mathrm{tr}(B^*)} + d^{1/2}KC_B\frac{\|\widetilde{B}\|_F}{\|B^*\|_F}\sqrt{\rho_{jt}(A)^2+1}\frac{\log^{1/2}(n\vee m)}{\sqrt{n}} \\
&\leqslant \widetilde{\eta},
\end{aligned}$$

which is the event $\mathcal{X}_0(A)$. Therefore, we get that $P(\mathcal{X}_0(A)) \geqslant 1 - 2/(n\vee m)^d$.

We can obtain the $P(\mathcal{X}_0)$ by using a union bound put together $P(\mathcal{X}_0(B))$ and $P(\mathcal{X}_0(A))$, completing the proof. □

3.5.4 Proof of Lemma III.19

Recall that $\widetilde{B} = (I-P_2)B^*(I-P_2)$. The matrix $\widetilde{B} - B^*$ can be expressed as

$$\widetilde{B} - B^* = (I-P_2)B^*(I-P_2) - B^* = -P_2B^* - B^*P_2 + P_2B^*P_2.$$

By the triangle inequality, $\|\widetilde{B} - B^*\|_{\max} \leqslant \|P_2 B^*\|_{\max} + \|B^* P_2\|_{\max} + \|P_2 B^* P_2\|_{\max}$. We bound each term on the right-hand side.

First we bound $\|P_2 B^*\|_{\max}$ and $\|B^* P_2\|_{\max}$. Let p_i denote the ith column of P_2. The (i,j)th entry satisfies

$$|p_i^T b_j^*| \leqslant \|B^* p_i\|_\infty \leqslant \|B^*\|_\infty \|p_i\|_\infty = \|B^*\|_1 \|p_i\|_\infty = \|B^*\|_1 / n_{\min},$$

so $\|P_2 B^*\|_{\max} \leqslant \|B^*\|_1 / n_{\min}$. Because P_2 and B^* are symmetric, $\|P_2 B^*\|_{\max} = \|B^* P_2\|_{\max}$.

We now bound $\|P_2 B^* P_2\|_{\max}$. Let $B^{1/2}$ denote the symmetric square root of B^*. We can express $p_i^T B^* p_j$ as an inner product $(B^{1/2} p_i)^T (B^{1/2} p_j)$, so

$$|(P_2 B^* P_2)_{ij}| = |(B^{1/2} p_i)^T (B^{1/2} p_j)| \leqslant \left(p_i^T B^* p_i\right)^{1/2} \left(p_j^T B^* p_j\right)^{1/2} \tag{3.67}$$

$$\leqslant \|p_i\|_2 \|p_j\|_2 \|B\|_2 \leqslant \|B^*\|_2 / n_{\min}, \tag{3.68}$$

where (3.67) follows from the Cauchy Schwarz inequality, and (3.68) holds because

$$\|p_i\|_2 = \begin{cases} 1/\sqrt{n_1} & \text{if } i \in \{1, \ldots, n_1\} \\ 1/\sqrt{n_2} & \text{if } i \in \{n_1 + 1, \ldots, n\}. \end{cases}$$

□

3.5.5 Proof of Lemma III.20

Let $B^{1/2}$ denote the symmetric square root of B^*. Let $Z_j = (a_{jj}^* B^*)^{-1/2} X_j$. We express $S_{ij}(B^*)$ as a quadratic form in order to use the Hanson-Wright inequality to prove a large deviation bound. That is, we show that $S_{ij}(B^*) = \operatorname{vec}(Z)^T \Phi^{ij} \operatorname{vec}(Z)$,

with

$$\Phi^{ij} \quad = \quad (1/m)A^* \otimes B^{1/2}(e_j - p_j)(e_i - p_i)^T B^{1/2}. \tag{3.69}$$

We express $S_{ij}(B^*)$ as a quadratic form, as follows:

$$\begin{aligned} S_{ij}(B^*) &= \frac{1}{m}\sum_{k=1}^{m}(e_i - p_i)^T X_k X_k^T (e_j - p_j) = \frac{1}{m}\sum_{k=1}^{m}\mathrm{tr}\left[(e_i - p_i)^T X_k X_k^T (e_j - p_j)\right] \\ &= \frac{1}{m}\sum_{k=1}^{m} X_k^T (e_j - p_j)(e_i - p_i)^T X_k \\ &= \frac{1}{m}\,\mathrm{vec}(X)^T \left(I_{m\times m} \otimes (e_j - p_j)(e_i - p_i)^T\right)\mathrm{vec}(X) \\ &= \mathrm{vec}(Z)^T \Phi^{ij}\,\mathrm{vec}(Z) \end{aligned}$$

where

$$\mathrm{tr}(\Phi^{ij}) \quad = \quad \mathrm{tr}(B^{1/2}(e_j - p_j)(e_i - p_i)^T B^{1/2}) = (e_i - p_i)^T B^*(e_j - p_j) = \widetilde{b}_{ij}, \tag{3.70}$$

$$\begin{aligned} \|\Phi^{ij}\|_F &= \frac{1}{m}\|A^*\|_F \|B^{1/2}(e_j - p_j)(e_i - p_i)^T B^{1/2}\|_F \qquad (3.71) \\ &= \frac{1}{m}\|A^*\|_F \left((e_i - p_i)^T B^*(e_i - p_i)\right)^{1/2}\left((e_j - p_j)^T B^*(e_j - p_j)\right)^{1/2} = \frac{1}{m}\|A^*\|_F \sqrt{\widetilde{b}_{ii}\widetilde{b}_{jj}}. \end{aligned}$$

Therefore, we get that

$$\begin{aligned} & P\left(\forall i,j \ \left|S_{ij}(B^*) - \widetilde{b}_{ij}\right| \leqslant K^2 d\log^{1/2}(m)\|\Phi^{ij}\|_F/c'\right) \\ &= P\left(\forall i,j \ \left|\mathrm{vec}(Z)^T\Phi^{ij}\,\mathrm{vec}(Z) - \mathrm{tr}\left(\Phi^{ij}\right)\right| \leqslant K^2 d\log^{1/2}(m)\|\Phi^{ij}\|_F/c'\right) \\ &\geqslant 1 - 2m^2\exp\left(-c\min\left(d^2\log(m)/c'^2, \frac{d\log^{1/2}(m)\|\Phi^{ij}\|_F/c'}{\|\Phi^{ij}\|_2}\right)\right) \\ &\geqslant 1 - 2/m^{d-2}. \end{aligned}$$

If the event $\left\{\forall i,j \left|S_{ij}(B^*) - \widetilde{b}_{ij}\right| \leqslant K^2 d \log^{1/2}(m) \|\Phi^{ij}\|_F / c'\right\}$ holds, it follows that

$$\left|S_{ij}(B^*) - b^*_{ij}\right| \leqslant \left|S_{ij}(B^*) - \widetilde{b}_{ij}\right| + |b^*_{ij} - \widetilde{b}_{ij}| \leqslant K^2 d \log^{1/2}(m) \|\Phi^{ij}\|_F / c' + |b_{ij} - \widetilde{b}_{ij}|.$$

The Lemma is thus proved. $\square$

3.6 Proof of Theorem II.4

3.6.1 Notation

Notation	Meaning
Mean structure	
$\mu \in \mathbb{R}^m$	Vector of grand means of each gene
$\gamma \in \mathbb{R}^m$	Vector of mean differences for each gene
$\nu = \frac{1}{2}\begin{bmatrix} \frac{1}{n_1}1_{n_1}^T & \frac{1}{n_2}1_{n_2}^T \end{bmatrix}^T \in \mathbb{R}^n$	Inner product with ν computes global mean
Outcome of model selection step	
$J_0 \subset \{1, 2, \ldots, m\}$	Indices selected for group centering
$J_1 \subset \{1, 2, \ldots, m\}$	Indices selected for global centering
Sizes of gene subsets	
$m_0 = \lvert J_0 \rvert$	Number of group centered genes
$m_1 = \lvert J_1 \rvert$	Number of globally centered genes
Projection matrices	
$P_1 = 1_n \nu^T$	Projection matrix that performs global centering
P_2 (as in (3.81))	Projection matrix that performs group centering
Sample covariance matrices	
$S(B, J_0, J_1) = \frac{m_1}{m} S_1(B) + \frac{m_0}{m} S_2(B)$	Model selection sample covariance matrix
$S_1(B, J_1) = \frac{1}{m_1} \sum_{j \in J_1} (I - P_1) X_j X_j^T (I - P_1)$	Globally centered sample covariance matrix
$S_2(B, J_0) = \frac{1}{m_0} \sum_{j \in J_0} (I - P_2) X_j X_j^T (I - P_2)$	Group centered sample covariance matrix
Decomposition of $S(B, J_0, J_1)$	
$S_{\mathrm{I}} = S(B, J_0, J_1) - \mathbb{E}\left[S(B, J_0, J_1)\right]$	Bias
$S_{\mathrm{II}} = \frac{1}{m}(I - P_1) M_{J_1} M_{J_1}^T (I - P_1)$	False negatives (deterministic)
$S_{\mathrm{III}} = \frac{1}{m}(I - P_1) M_{J_1} \varepsilon^T (I - P_1)$	False negatives (random)
$S_{\mathrm{IV}} = m^{-1}(I - P_2)\varepsilon_{J_0}\varepsilon_{J_0}^T(I - P_2) +$ $m^{-1}(I - P_1)\varepsilon_{J_1}\varepsilon_{J_1}^T(I - P_1)$	True negatives

3.6.2 Two-Group Model and Centering

We begin by introducing some relevant notation for the two-group model and centering. Define the group membership vector $\delta_n \in \mathbb{R}^n$ as

$$\delta_n := \begin{bmatrix} 1_{n_1}^T & -1_{n_2}^T \end{bmatrix}^T \in \mathbb{R}^n. \tag{3.72}$$

In the two-group model, the mean matrix M can be expressed as

$$M = 1_n \mu^T + (1/2)\delta_n \gamma^T, \tag{3.73}$$

where $\mu \in \mathbb{R}^m$ is a vector of grand means, and $\gamma \in \mathbb{R}^m$ is the vector of mean differences. According to (3.73), the (i, j)th entry of M can be expressed as

$$m_{ij} = \begin{cases} \mu_j + \gamma_j/2 & \text{if sample } i \text{ is in group one} \\ \mu_j - \gamma_j/2 & \text{if sample } i \text{ is in group two.} \end{cases} \tag{3.74}$$

Define the vector $\nu \in \mathbb{R}^n$ as

$$\nu = \frac{1}{2}\begin{bmatrix} \frac{1}{n_1} 1_{n_1}^T & \frac{1}{n_2} 1_{n_2}^T \end{bmatrix}^T \in \mathbb{R}^n, \tag{3.75}$$

so that for the jth column of the data matrix $X_j \in \mathbb{R}^n$,

$$\mathbb{E}\left(\nu^T X_j\right) = \frac{1}{2}\mathbb{E}\left(\frac{1}{n_1}\sum_{k=1}^{n_1} X_{jk} + \frac{1}{n_2}\sum_{k=n_1+1}^{n} X_{jk}\right) = \mu_j. \tag{3.76}$$

Note that

$$\nu^T 1_n = (1/2)(1+1) = 1, \quad \text{and} \quad \nu^T \delta_n = (1/2)(1-1) = 0. \tag{3.77}$$

Next we define a projection matrix that performs global centering. Define the non-

orthogonal projection matrix

$$P_1 := 1_n \nu^T \in \mathbb{R}^{n \times n}. \tag{3.78}$$

Applying the projection matrix to the mean matrix yields

$$P_1 M = 1_n \nu^T \left(1_n \mu^T + (1/2)\delta_n \gamma^T\right) = 1_n \mu^T + (1/2)(\nu^T \delta_n) 1_n \gamma^T = 1_n \mu^T, \tag{3.79}$$

with residuals

$$(I - P_1)M = M - P_1 M = M - 1_n \mu^T = (1/2)\delta_n \gamma^T. \tag{3.80}$$

Define

$$P_2 = \begin{bmatrix} n_1^{-1} 1_{n_1} 1_{n_1}^T & \\ & n_2^{-1} 1_{n_2} 1_{n_2}^T \end{bmatrix}. \tag{3.81}$$

Note that $P_2 1_n = 1_n$ and $P_2 \delta_n = \delta_n$, so

$$P_2 M = P_2 1_n \mu^T + (1/2) P_2 \delta_n \gamma^T = 1_n \mu^T + (1/2)\delta_n \gamma^T = M, \tag{3.82}$$

and therefore $(I - P_2)M = 0$.

Define

$$\breve{B} = (I - P_1)B(I - P_1) = \left(\breve{b}_{ij}\right) \tag{3.83}$$

$$\widetilde{B} = (I - P_2)B(I - P_2) = \left(\widetilde{b}_{ij}\right) \tag{3.84}$$

$$\breve{\breve{B}} = (I - P_1)B(I - P_2) = \left(\breve{\breve{b}}_{ij}\right). \tag{3.85}$$

Let $\breve{b}_{\max}$, $\widetilde{b}_{\max}$, and $\breve{\breve{b}}_{\max}$ denote the maximum diagonal entries of $\breve{B}$, $\widetilde{B}$, and $\breve{\breve{B}}$, respectively.

3.6.3 Model Selection Centering

For a subset $J \subset \{1, \ldots, m\}$, let X_J denote the submatrix of X consisting of columns indexed by J. For the fixed sets of genes J_0 and J_1, define the sample covariance

$$S(B, J_0, J_1) = m^{-1} \sum_{k \in J_0} (I - P_2) X_k X_k^T (I - P_2)^T + m^{-1} \sum_{k \in J_1} (I - P_1) X_k X_k^T (I - P_1)^T =: \mathrm{I} + \mathrm{II}. \tag{3.86}$$

Note that $\mathbb{E}\left[S(B, J_0, J_1)\right] = B^\sharp$, with

$$B^\sharp = \frac{\mathrm{tr}\,(A_{J_0})}{m}(I - P_2)B(I - P_2) + \frac{\mathrm{tr}\,(A_{J_1})}{m}(I - P_1)B(I - P_1). \tag{3.87}$$

Define the sample correlation matrix,

$$\widehat{\Gamma}_{ij}(B) = \frac{(S(B, J_0, J_1))_{ij}}{\sqrt{(S(B, J_0, J_1))_{ii}(S(B, J_0, J_1))_{jj}}}. \tag{3.88}$$

The baseline Gemini estimators *Zhou* (2014a) are then defined as follows, using a pair of penalized estimators for the correlation matrices $\rho(A) = (a_{ij}/\sqrt{a_{ii}a_{jj}})$ and $\rho(B) = (b_{ij}/\sqrt{b_{ii}b_{jj}})$:

$$\widehat{A}_\rho = \underset{A_\rho > 0}{\arg\min} \left\{ \mathrm{tr}\left(\widehat{\Gamma}(A)A_\rho^{-1}\right) + \log|A_\rho| + \lambda_B |A_\rho^{-1}|_{1,\mathrm{off}} \right\}, \tag{3.89a}$$

$$\widehat{B}_\rho = \underset{B_\rho > 0}{\arg\min} \left\{ \mathrm{tr}\left(\widehat{\Gamma}(B)B_\rho^{-1}\right) + \log|B_\rho| + \lambda_A |B_\rho^{-1}|_{1,\mathrm{off}} \right\}. \tag{3.89b}$$

We will focus on $\widehat{B}_\rho$ using the input as defined in (3.88).

The proof proceeds as follows. Lemma III.22, the equivalent of Proposition III.16 for Algorithm 1, establishes entry-wise convergence rates of the sample covariance matrix for fixed sets of group and globally centered genes. We use this to prove Theorem III.21 below in Section 3.6.4 and to prove Theorem II.4 in Section 3.6.5.

3.6.4 Convergence for fixed gene sets

We first state a standalone result, Theorem III.21, which provides rates of convergence when $S(B, J_0, J_1)$ as in (3.86) is calculated using fixed sets of group centered and globally centered genes, J_0 and J_1, respectively. This result shows how the algorithm used in the preliminary step to choose which genes to group center can be decoupled from the rest of the estimation procedure. The proof is presented below in Section 3.6.4.2.

Theorem III.21. *Suppose that (A1'), (A2'), and (A3) hold. Let J_0 and J_1 denote sets such that $J_0 \cap J_1 = \varnothing$ and $J_0 \cup J_1 = \{1, \ldots, m\}$. Let $m_0 = |J_0|$ and $m_1 = |J_1|$ denote the sizes of the sets. Let $\tau_{global} > 0$ satisfy*

$$\max_{j \in J_1} |\gamma_j| \leqslant \tau_{global}, \tag{3.90}$$

for $\tau_{global} = C\sqrt{\log(m)}\|(D^T B^{-1} D)^{-1}\|_2^{1/2} \asymp \sqrt{\frac{\log(m)}{n}}$.

Consider the data as generated from model (3.73) with $\varepsilon = B^{1/2} Z A^{1/2}$, where $A \in \mathbb{R}^{m \times m}$ and $B \in \mathbb{R}^{n \times n}$ are positive definite matrices, and Z is an $n \times m$ random matrix as defined in Theorem II.1. Let λ_A denote the penalty parameter for estimating B. Suppose the penalty parameter λ_A in (3.89b) satisfies

$$\lambda_A \geqslant C'' \left[C_A K \frac{\log^{1/2}(m \vee n)}{\sqrt{m}} + \frac{\|B\|_1}{n_{\min}} \right]. \tag{3.91}$$

where C'' is an absolute constant.

Suppose the number of off-diagonal entries of B^{-1} satisfies

$$|B^{-1}|_{0,\text{off}} \leqslant \min\left(m, n \log(m)\right). \tag{3.92}$$

(I) *Let* $\mathcal{E}_4(J_0, J_1)$ *be the event such that*

$$\left\| \mathrm{tr}\,(A) \left(\widehat{W}_2 \widehat{B}_\rho \widehat{W}_2 \right)^{-1} - B^{-1} \right\|_2 \leqslant \frac{C' \lambda_A \sqrt{|B^{-1}|_{0,\mathrm{off}} \vee 1}}{b_{\min} \varphi^2_{\min}(\rho(B))}. \tag{3.93}$$

Then $P(\mathcal{E}_4(J_0, J_1)) \geqslant 1 - C/m^d$.

(II) *With probability at least* $1 - C'/m^d$*, for all* j*,*

$$\|\widehat{\beta}_j(\widehat{B}^{-1}) - \beta_j^*\|_2 \leqslant C_1 \lambda_A \sqrt{\frac{n_{ratio}\,(|B^{-1}|_{0,off} \vee 1)}{n_{\min}}} + C_2 \sqrt{\log(m)} \|(D^T B^{-1} D)^{-1}\|_2^{1/2}. \tag{3.94}$$

3.6.4.1 Decomposition of sample covariance matrix

The error in the sample covariance $S(B, J_0, J_1)$ can be decomposed as

$$S(B, J_0, J_1) - B = \left[B^\sharp - B \right] + \left[S(B, J_0, J_1) - B^\sharp \right], \tag{3.95}$$

where the first term corresponds to bias and the second term to variance. We now further decompose the variance term. The first term of $S(B, J_0, J_1)$ in (3.86) can be decomposed as,

$$\begin{aligned}
\mathrm{I} &= m^{-1}(I - P_2) X_{J_0} X_{J_0}^T (I - P_2) \\
&= m^{-1}(I - P_2)(M_{J_0} + \varepsilon_{J_0})(M_{J_0} + \varepsilon_{J_0})^T (I - P_2) \\
&= m^{-1}(I - P_2)\varepsilon_{J_0}\varepsilon_{J_0}^T (I - P_2) + m^{-1}(I - P_2) M_{J_0} \varepsilon_{J_0}^T (I - P_2) \\
&\quad + m^{-1}(I - P_2)\varepsilon_{J_0} M_{J_0}^T (I - P_2) + m^{-1}(I - P_2) M_{J_0} M_{J_0}^T (I - P_2),
\end{aligned} \tag{3.96}$$

and the second term can be decomposed analogously, as

$$\begin{aligned}\mathrm{II} = {} & m^{-1}(I-P_1)\varepsilon_{J_1}\varepsilon_{J_1}^T(I-P_1) + m^{-1}(I-P_1)M_{J_1}\varepsilon_{J_1}^T(I-P_1) \\ & + m^{-1}(I-P_1)\varepsilon_{J_1}M_{J_1}^T(I-P_1) + m^{-1}(I-P_1)M_{J_1}M_{J_1}^T(I-P_1). \qquad (3.97)\end{aligned}$$

By the above decompositions, it follows that $S(B, J_0, J_1)$ can be expressed as

$$S(B, J_0, J_1) = S_{\mathrm{II}} + S_{\mathrm{III}} + S_{\mathrm{III}}^T + S_{\mathrm{IV}}, \qquad (3.98)$$

with

$$S_{\mathrm{II}} = m^{-1}(I-P_2)M_{J_0}M_{J_0}^T(I-P_2) + m^{-1}(I-P_1)M_{J_1}M_{J_1}^T(I-P_1). \qquad (3.99)$$

$$S_{\mathrm{III}} = m^{-1}(I-P_2)M_{J_0}\varepsilon_{J_0}^T(I-P_2) + m^{-1}(I-P_1)M_{J_1}\varepsilon_{J_1}^T(I-P_1) \qquad (3.100)$$

$$S_{\mathrm{IV}} = m^{-1}(I-P_2)\varepsilon_{J_0}\varepsilon_{J_0}^T(I-P_2) + m^{-1}(I-P_1)\varepsilon_{J_1}\varepsilon_{J_1}^T(I-P_1) \qquad (3.101)$$

For each of S_{II}, S_{III}, and S_{IV}, the first term comes from (3.96) and the second term comes from (3.97).

The terms S_{II} and S_{III} can be simplified, as follows. Because $(I-P_2)M_{J_0} = 0$, it follows that the first term of S_{II} is zero:

$$m^{-1}(I-P_2)M_{J_0}M_{J_0}^T(I-P_2) = 0.$$

and the first term of S_{III} is also zero,

$$m^{-1}(I-P_2)M_{J_0}\varepsilon_{J_0}^T(I-P_2) = 0,$$

Therefore the terms S_{II} and S_{III} are equal to

$$S_{\mathrm{II}} = m^{-1}(I - P_1)M_{J_1}M_{J_1}^T(I - P_1), \tag{3.102}$$

$$S_{\mathrm{III}} = m^{-1}(I - P_1)M_{J_1}\varepsilon_{J_1}^T(I - P_1). \tag{3.103}$$

Let $S_{\mathrm{I}} = B^\sharp - B$. We have thus decomposed the error in the sample covariance as

$$S(B, J_0, J_1) - B = \underbrace{S_{\mathrm{I}}}_{\text{bias}} + \underbrace{\left[\left(S_{\mathrm{IV}} - B^\sharp\right) + S_{\mathrm{III}} + S_{\mathrm{II}}\right]}_{\text{variance}}. \tag{3.104}$$

In Lemma III.23, we provide an error bound for each term in the decomposition (3.104).

We next state Lemma III.22, which establishes the maximum of entry-wise errors for estimating B using the sample covariance for fixed gene sets as defined in (3.104). Lemma III.22 is used in the proof of Theorem III.21. Following, we state Lemma III.23, which is used in the proof of Lemma III.22.

Lemma III.22. *Suppose the conditions of Theorem III.21 hold. Let $\mathcal{E}_6(J_0, J_1)$ denote the event*

$$\mathcal{E}_6(J_0, J_1) = \left\{ \|S(B, J_0, J_1) - B\|_\infty \leqslant C_A K \frac{\log^{1/2}(m \vee n)}{\sqrt{m}} + \frac{\|B\|_1}{n_{\min}} \right\}. \tag{3.105}$$

Then $\mathcal{E}_6(J_0, J_1)$ holds with probability at least $1 - \frac{8}{(m \vee n)^2}$.

Lemma III.23. *Let the model selection-based sample covariance $S(B, J_0, J_1)$ be as defined in* (3.86)*, where J_1 and J_0 are fixed sets of variables that are globally centered,*

and group centered, respectively. Let $m_0 = |J_0|$ *and* $m_1 = |J_1|$*. Define the rates*

$$r_1 = \frac{3\|B\|_1}{n_{\min}}, \tag{3.106}$$

$$r_2 = (4m)^{-1}\|\gamma_{J_1}\|_2^2, \tag{3.107}$$

$$r_3 = C_3 d^{1/2} K^2 \log^{1/2}(m) m^{-1} \left(\gamma_{J_1}^T A_{J_1}\gamma_{J_1}\right)^{1/2} \breve{b}_{\max}^{1/2}, \tag{3.108}$$

$$r_4 = C_4 d^{1/2} K \log^{1/2}(m) m^{-1} \|A\|_F \|B\|_2. \tag{3.109}$$

(I) *Deterministically,*

$$\left\|B^\sharp - B\right\|_\infty \leqslant r_1 \quad \text{and} \quad \|S_{\mathrm{II}}\|_\infty \leqslant r_2. \tag{3.110}$$

(II) *Define the events*

$$\mathcal{E}_{\mathrm{I}} = \left\{\left\|S_{\mathrm{IV}} - B^\sharp\right\|_\infty \leqslant r_4\right\} \quad \text{and} \quad \mathcal{E}_{\mathrm{II}} = \left\{\|S_{\mathrm{III}}\|_\infty \leqslant r_3\right\}. \tag{3.111}$$

Then $\mathcal{E}_{\mathrm{I}}$ *and* $\mathcal{E}_{\mathrm{II}}$ *occur with probability at least* $1 - 2/m^d$*.*

Lemmas III.22 and III.23 are proved in Section 3.7. We analyze term S_{I} in Section 3.7.2, term S_{II} in Section 3.7.3, term S_{III} in Section 3.7.4, and term S_{IV} in Section 3.7.5.

3.6.4.2 Proof of Theorem III.21

Let us first define the event $\mathcal{E}_{\mathrm{global}}$, that is, the GLS error based on the true B^{-1} is small:

$$\mathcal{E}_{\mathrm{global}} = \left\{\left\|\widehat{\gamma}(B^{-1}) - \gamma\right\|_\infty < \sqrt{\log(m)}\|(D^T B^{-1} D)^{-1}\|_2^{1/2}\right\}. \tag{3.112}$$

Let $\mathcal{E}_4(J_0, J_1)$ be defined as in (3.93), denoting small operator norm error in esti-

mating B^{-1}:

$$\mathcal{E}_4(J_0, J_1) = \left\{ \left\| \mathrm{tr}\,(A) \left(\widehat{W}_2 \widehat{B}_\rho \widehat{W}_2 \right)^{-1} - B^{-1} \right\|_2 \leqslant \frac{C' \lambda_A \sqrt{|B^{-1}|_{0,\mathrm{off}} \vee 1}}{b_{\min} \varphi^2_{\min}(\rho(B))} \right\}. \quad (3.113)$$

Note that $\mathcal{E}_4(J_0, J_1)$ holds deterministically under event $\mathcal{E}_6(J_0, J_1)$ as defined in (3.105) of Lemma III.22.

Define the event bounding the perturbation in mean estimation due to error in estimating B^{-1}:

$$\mathcal{E}_5(J_0, J_1) = \left\{ \left\| \widehat{\gamma}(\widehat{B}^{-1}) - \widehat{\gamma}(B^{-1}) \right\|_\infty < C n_{\min}^{-1/2} \left\| \widehat{B}^{-1} - B^{-1} \right\|_2 \right\}. \quad (3.114)$$

Conditional on a fixed matrix $\widehat{B}^{-1}$ that satisfies $\mathcal{E}_4(J_0, J_1)$, event $\mathcal{E}_5(J_0, J_1)$ holds with probability at least $1 - C/m^d$, by Lemma III.6 (used in the proof of Theorem II.1).

The overall rate of convergence follows by applying the union bound to the events $\mathcal{E}_{\mathrm{global}} \cap \mathcal{E}_4(J_0, J_1) \cap \mathcal{E}_5(J_0, J_1)$, as follows:

$$\begin{aligned}
&P(\mathcal{E}^c_{\mathrm{global}} \cup \mathcal{E}_4(J_0, J_1)^c \cup \mathcal{E}_5(J_0, J_1)^c) \\
&\leqslant P(\mathcal{E}^c_{\mathrm{global}}) + P(\mathcal{E}_4(J_0, J_1)^c) + P(\mathcal{E}_5(J_0, J_1)^c \mid \mathcal{E}_4(J_0, J_1)) P(\mathcal{E}_4(J_0, J_1)) \\
&\qquad + P(\mathcal{E}_5(J_0, J_1)^c \mid \mathcal{E}_4(J_0, J_1)^c) P(\mathcal{E}_4(J_0, J_1)^c) \\
&\leqslant P(\mathcal{E}^c_{\mathrm{global}}) + P(\mathcal{E}_4(J_0, J_1)^c) + P(\mathcal{E}_4(J_0, J_1)^c) + P(\mathcal{E}_5(J_0, J_1)^c \mid \mathcal{E}_4(J_0, J_1)) \\
&= P(\mathcal{E}^c_{\mathrm{global}}) + 2P(\mathcal{E}_4(J_0, J_1)^c) + P(\mathcal{E}_5(J_0, J_1)^c \mid \mathcal{E}_4(J_0, J_1)),
\end{aligned}$$

where $P(\mathcal{E}^c_{\mathrm{global}})$ and $P(\mathcal{E}_5(J_0, J_1)^c \mid \mathcal{E}_4(J_0, J_1))$ are bounded in Theorem II.1, and $P(\mathcal{E}_4(J_0, J_1)^c)$ has high probability under Lemma III.22.

3.6.5 Proof of Theorem II.4

Let $\widehat{\gamma}^{\text{init}}$ denote the output from Algorithm 1. By our choice of the threshold parameter τ_{init} as in (2.16), that is,

$$\tau_{\text{init}} = C\left(\frac{\log^{1/2}(m)}{\sqrt{m}} + \frac{\|B\|_1}{n_{\min}}\right)\sqrt{\frac{n_{\text{ratio}}\left(|B^{-1}|_{0,\text{off}} \vee 1\right)}{n_{\min}}} + C\sqrt{\log(m)}\|(D^T B^{-1} D)^{-1}\|_2^{1/2},$$

we have a partition $(\widetilde{J}_0, \widetilde{J}_1)$ such that $\widetilde{J}_0$ is the set of variables selected for group centering and $\widetilde{J}_1$ is the set of variables selected for global centering. The partition results in a sample covariance matrix $S(B, \widetilde{J}_0, \widetilde{J}_1)$ as defined in (3.86). Define the event that the Algorithm 1 estimate $\widehat{\gamma}^{\text{init}}$ is close to γ in the sense that

$$\mathcal{E}_{A1} = \left\{\left\|\widehat{\gamma}^{\text{init}} - \gamma\right\|_\infty < \tau_{\text{init}}\right\}. \tag{3.115}$$

Note that the event $\mathcal{E}_{A1}$ implies that the false negatives have small true mean differences. That is, on event $\mathcal{E}_{A1}$, by the triangle inequality,

$$\left\|\gamma_{\widetilde{J}_1}\right\|_\infty \leqslant \left\|\gamma_{\widetilde{J}_1} - \widehat{\gamma}^{\text{init}}_{\widetilde{J}_1}\right\|_\infty + \left\|\widehat{\gamma}^{\text{init}}_{\widetilde{J}_1}\right\|_\infty \leqslant \tau_{\text{init}} + \tau_{\text{init}} = 2\tau_{\text{init}}, \tag{3.116}$$

where $\left\|\widehat{\gamma}^{\text{init}}_{\widetilde{J}_1}\right\|_\infty < \tau_{\text{init}}$ by definition of $\mathcal{E}_{A1}$, and $\left\|\gamma_{\widetilde{J}_1} - \widehat{\gamma}^{\text{init}}_{\widetilde{J}_1}\right\|_\infty < \tau_{\text{init}}$ by definition of the thresholding set $\widetilde{J}_1$.

Under the assumptions of Theorem III.21, $\tau_{\text{init}} \leqslant \tau_{\text{global}}$ with τ_{global} as defined in (3.90), so condition (3.90) of Theorem III.21 is satisfied. Under the conditions of Theorem III.21, event $\mathcal{E}_6(J_0, J_1)$ as defined in Lemma III.22 holds with high probability; that is, the entrywise error in the sample covariance matrix is small.

Let $\mathcal{E}_B$ denote event (2.28) in Theorem II.4. In view of Theorem III.9 and Lemma

III.10, event $\mathcal{E}_B$ holds on $\mathcal{E}_6(J_0, J_1)$. Hence

$$\begin{aligned} P\left(\mathcal{E}_B^c\right) &= P\left(\mathcal{E}_6(J_0,J_1)^c \mid \mathcal{E}_{A1}\right) P\left(\mathcal{E}_{A1}\right) + P\left(\mathcal{E}_6(J_0,J_1)^c \mid \mathcal{E}_{A1}^c\right) P\left(\mathcal{E}_{A1}^c\right) \\ &\leqslant P\left(\mathcal{E}_6(J_0,J_1)^c \mid \mathcal{E}_{A1}\right) + P\left(\mathcal{E}_{A1}^c\right) \\ &\leqslant 2/m^d + 2/m^d, \end{aligned}$$

where the first term is bounded in Lemma III.22 and the second in Theorem II.3.

Recall the event $\mathcal{E}_{\text{global}}$ as defined in (3.112). Event (2.29) in Theorem II.4 holds under the intersection of events $\mathcal{E}_{\text{global}} \cap \mathcal{E}_5(\widetilde{J}_0, \widetilde{J}_1) \cap \mathcal{E}_B \cap \mathcal{E}_{A1}$. Hence the probability of (2.29) can be bounded as follows:

$$\begin{aligned} &P(\mathcal{E}_{\text{global}}^c \cup \mathcal{E}_5(\widetilde{J}_0, \widetilde{J}_1)^c \cup \mathcal{E}_B^c \cup \mathcal{E}_{A1}^c) \\ &\leqslant P(\mathcal{E}_{\text{global}}^c) + P(\mathcal{E}_B^c) + P(\mathcal{E}_5(\widetilde{J}_0, \widetilde{J}_1)^c \mid \mathcal{E}_B) P(\mathcal{E}_B) \\ &\quad + P(\mathcal{E}_5(\widetilde{J}_0, \widetilde{J}_1)^c \mid \mathcal{E}_B^c) P(\mathcal{E}_B^c) + P\left(\mathcal{E}_{A1}^c\right) \\ &\leqslant P(\mathcal{E}_{\text{global}}^c) + P(\mathcal{E}_B^c) + P(\mathcal{E}_B^c) + P(\mathcal{E}_5(\widetilde{J}_0, \widetilde{J}_1)^c \mid \mathcal{E}_B) + P\left(\mathcal{E}_{A1}^c\right) \\ &= P(\mathcal{E}_{\text{global}}^c) + 2P(\mathcal{E}_B^c) + P(\mathcal{E}_5(\widetilde{J}_0, \widetilde{J}_1)^c \mid \mathcal{E}_B) + P\left(\mathcal{E}_{A1}^c\right), \end{aligned}$$

where $P(\mathcal{E}_{\text{global}}^c)$ and $P(\mathcal{E}_5(\widetilde{J}_0, \widetilde{J}_1)^c \mid \mathcal{E}_B)$ are bounded in Theorem 1, $P(\mathcal{E}_B^c)$ is bounded above, and $P\left(\mathcal{E}_{A1}^c\right)$ is bounded in Theorem II.3.

3.7 Proof of Lemmas III.22 and III.23

We first prove Lemma III.22 in Section 3.7.1. The rest of the section contains the proof of Lemma III.23, where part I is proved in Sections 3.7.2 and 3.7.3 and part II in Sections 3.7.4 and 3.7.5.

3.7.1 Proof of Lemma III.22

The entrywise error in the sample covariance matrix (3.86) can be decomposed as

$$\|S(B, J_0, J_1) - B\|_\infty \leqslant \left\|S(B, J_0, J_1) - B^\sharp\right\|_\infty + \left\|B^\sharp - B\right\|_\infty \tag{3.117}$$

$$\leqslant \left\|S_{\mathrm{IV}} - B^\sharp\right\|_\infty + 2\|S_{\mathrm{III}}\|_\infty + \|S_{\mathrm{II}}\|_\infty + \left\|B^\sharp - B\right\|_\infty. \tag{3.118}$$

Let $r_{n,m} = r_1 + r_2 + 2r_3 + r_4$. By parts I and II of Lemma III.23,

$$\begin{aligned}
&P\left(\|S(B, J_0, J_1) - B\|_\infty \geqslant r_{n,m}\right) \\
&\leqslant P\left(\left\|S_{\mathrm{IV}} - B^\sharp\right\|_\infty + 2\|S_{\mathrm{III}}\|_\infty + \|S_{\mathrm{II}}\|_\infty + \left\|B^\sharp - B\right\|_\infty \geqslant r_{n,m}\right) \quad \text{(by (3.118))} \\
&\leqslant P\left(\left\|S_{\mathrm{IV}} - B^\sharp\right\|_\infty + 2\|S_{\mathrm{III}}\|_\infty + r_2 + r_1 \geqslant r_{n,m}\right) \quad \text{(by (3.110))} \\
&= P\left(\left\|S_{\mathrm{IV}} - B^\sharp\right\|_\infty + 2\|S_{\mathrm{III}}\|_\infty \geqslant r_4 + 2r_3\right) \\
&\leqslant P\left(\left\|S_{\mathrm{IV}} - B^\sharp\right\|_\infty \geqslant r_4\right) + P\left(2\|S_{\mathrm{III}}\|_\infty \geqslant 2r_3\right) \quad \text{(by (3.111))} \\
&\leqslant \frac{2}{m^d} + \frac{2}{m^d} = \frac{4}{m^d}.
\end{aligned}$$

We show that under the assumptions of Theorem III.21, the entrywise error in terms S_{II} and S_{III} is $O\left(C_A\sqrt{\frac{\log(m)}{m}}\right)$. Recall that the entrywise rates of convergence of S_{II} and S_{III} are stated in equations (3.107) and (3.108), respectively. Let $s = |\mathrm{supp}(\gamma)|$ denote the sparsity of γ. Let $m_{01} = |\mathrm{supp}\left(\gamma_{J_1}\right)|$ denote the number of false negatives.

First, we express the entrywise rate of convergence of S_{II} in terms of τ_{global}. By (3.90), $\|\gamma_{J_1}\|_\infty \leqslant \tau_{\mathrm{global}}$, which implies that $\|\gamma_{J_1}\|_2^2 \leqslant m_{01}\tau_{\mathrm{global}}^2 \leqslant s\tau_{\mathrm{global}}^2$, where the last inequality holds because $m_{01} \leqslant s$ by definition. Therefore,

$$r_2 = (4m)^{-1}\|\gamma_{J_1}\|_2^2 \leqslant \frac{s\tau_{\mathrm{global}}^2}{4m} \leqslant C\frac{s\log(m)}{4nm}\|B\|_2, \tag{3.119}$$

where the last step holds because $\tau_{\mathrm{global}} = C\sqrt{\log(m)}\|(D^TB^{-1}D)^{-1}\|_2^{1/2} \asymp \sqrt{\frac{\log(m)}{n}}\|B\|_2^{1/2}$ by assumption. Applying (A3) to the right-hand side of (3.119) implies that $r_2 =$

$O\left(C_A\sqrt{\frac{\log(m)}{m}}\right)$.

Next, consider term S_{III}. First note that

$$\gamma_{J_1}^T A_{J_1}\gamma_{J_1} \leqslant \|\gamma_{J_1}\|_2^2 \|A_{J_1}\|_2 \leqslant m_{01}\tau_{\text{global}}^2 \|A_{J_1}\|_2 , \tag{3.120}$$

where the last inequality holds by (3.90). This implies that r_3 is on the order

$$\begin{aligned}\frac{\log^{1/2}(m)}{m}\left(\breve{b}_{\max}\gamma_{J_1}^T A_{J_1}\gamma_{J_1}\right)^{1/2} &\leqslant \breve{b}_{\max}^{1/2}\|A_{J_1}\|_2^{1/2}\left(\frac{\log^{1/2}(m)m_{01}^{1/2}}{m}\right)\tau_{\text{global}} \\ &\leqslant C\frac{\log(m)}{\sqrt{n}}\frac{\sqrt{s}}{m}\|A_{J_1}\|_2^{1/2}\|B\|_2^{1/2}\breve{b}_{\max}^{1/2}, \end{aligned}\tag{3.121}$$

where the last inequality holds because $m_{01} \leqslant s \leqslant m$ and $\tau_{\text{global}} \asymp \sqrt{\frac{\log(m)}{n}}\|B\|_2^{1/2}$. Under (A2'), the right-hand side of (3.121) satisfies

$$\frac{\log(m)}{\sqrt{n}}\frac{\sqrt{s}}{m}\|A_{J_1}\|_2^{1/2}\|B\|_2^{1/2}\breve{b}_{\max}^{1/2} \leqslant \sqrt{\log(m)}\frac{\sqrt{s}}{m}C_A\frac{\|A_{J_1}\|_2^{1/2}}{\|A\|_2^{1/2}} \leqslant C_A\sqrt{\frac{\log(m)}{m}}, \tag{3.122}$$

where the last inequality holds because $s \leqslant m$.

3.7.2 Proof of part I of Lemma III.23, term I

We bound the entrywise bias,

$$\begin{aligned}\left\|B^\sharp - B\right\|_{\max} &= \left\|\frac{\operatorname{tr}(A_{J_0})}{m}\widetilde{B} + \frac{\operatorname{tr}(A_{J_1})}{m}\breve{B} - B\right\|_{\max} \\ &\leqslant \frac{\operatorname{tr}(A_{J_0})}{m}\left\|\widetilde{B} - B\right\|_{\max} + \frac{\operatorname{tr}(A_{J_1})}{m}\left\|\breve{B} - B\right\|_{\max}. \end{aligned}\tag{3.123}$$

Note that

$$\begin{aligned}\left\|\breve{B} - B\right\|_{\max} &= \|(I-P_1)B(I-P_1) - B\|_{\max} = \|P_1BP_1 - P_1B - BP_1\|_{\max} \\ &\leqslant \|P_1BP_1\|_{\max} + \|P_1B\|_{\max} + \|BP_1\|_{\max}. \end{aligned}\tag{3.124}$$

We bound the first term of (3.124) as follows:

$$\left|(P_1 B P_1)_{ij}\right| \leqslant \left\|p_i^{(1)}\right\|_2 \left\|p_j^{(1)}\right\|_2 \|B\|_2 \leqslant \frac{\|B\|_2}{n_{\min}}.$$

For the second term of (3.124),

$$(P_1 B)_{ij} = \left|b_i^T p_j^{(1)}\right| \leqslant \|b_i\|_1 \left\|p_j^{(1)}\right\|_\infty \leqslant \|B\|_1 \left\|p_j^{(1)}\right\|_\infty \leqslant \frac{\|B\|_1}{n_{\min}},$$

where $\left\|p_j^{(1)}\right\|_\infty \leqslant \frac{1}{n_{\min}}$ by the definition of P_1 in (3.78). We have shown $\|BP_1\|_{\max} \leqslant \frac{\|B\|_1}{n_{\min}}$. Likewise, $\|BP_1\|_{\max} \leqslant \frac{\|B\|_1}{n_{\min}}$. Therefore,

$$\left\|\breve{B} - B\right\|_{\max} \leqslant 3\frac{\|B\|_1}{n_{\min}}. \tag{3.125}$$

Because the projection matrix P_2 satisfies $\left\|p_j^{(2)}\right\|_\infty \leqslant \frac{1}{n_{\min}}$, an analogous proof shows that

$$\left\|\widetilde{B} - B\right\|_{\max} \leqslant \frac{3\|B\|_1}{n_{\min}}. \tag{3.126}$$

Substituting (3.125) and (3.126) into (3.123) yields

$$\begin{aligned}
\left\|B^\sharp - B\right\|_{\max} &\leqslant \frac{\operatorname{tr}(A_{J_0})}{m}\left\|\breve{B} - B\right\|_{\max} + \frac{\operatorname{tr}(A_{J_1})}{m}\left\|\widetilde{B} - B\right\|_{\max} \\
&\leqslant \left(\frac{\operatorname{tr}(A_{J_0})}{m} + \frac{\operatorname{tr}(A_{J_1})}{m}\right)\frac{3\|B\|_1}{n_{\min}} \\
&= \frac{\operatorname{tr}(A)}{m}\frac{3\|B\|_1}{n_{\min}} \\
&= \frac{3\|B\|_1}{n_{\min}}.
\end{aligned} \tag{3.127}$$

3.7.3 Proof of part I of Lemma III.23, term II

In this section we prove a deterministic entrywise bound on S_{II}. By (3.80), it follows that

$$(I - P_1)M_{J_1}M_{J_1}^T(I - P_1) = (1/4)\left\|\gamma_{J_1}\right\|_2^2 \delta_n \delta_n^T,$$

which implies

$$\left\|(I - P_1)M_{J_1}M_{J_1}^T(I - P_1)\right\|_\infty = \left\|(1/4)\left\|\gamma_{J_1}\right\|_2^2 \delta_n \delta_n^T\right\|_\infty = (1/4)\left\|\gamma_{J_1}\right\|_2^2.$$

Therefore S_{II} satisfies the maximum entrywise bound

$$\|S_{\mathrm{II}}\|_\infty = \left\|m^{-1}(I - P_1)M_{J_1}M_{J_1}^T(I - P_1)\right\|_\infty = \left\|(4m)^{-1}\left\|\gamma_{J_1}\right\|_2^2 \delta_n \delta_n^T\right\|_\infty = (4m)^{-1}\left\|\gamma_{J_1}\right\|_2^2,$$

so

$$\|S_{\mathrm{II}}\|_\infty = r_2.$$

Note that if J_1 is chosen so that $\|\gamma_{J_1}\|_\infty \leqslant \tau$, then $\|\gamma_{J_1}\|_2^2 \leqslant m_{01}\tau^2$, where m_{01} is the number of false negatives, so

$$\frac{\|\gamma_1\|_2^2}{4m} \leqslant \frac{m_{01}}{4m}\tau^2 \leqslant \frac{\tau^2}{4}. \tag{3.128}$$

which implies that the entrywise rate of convergence of S_{II} is $O(\tau^2)$.

3.7.4 Proof of part II of Lemma III.23, term III

Let p_i denote the ith column of P_1^T, for $i = 1, \ldots, n$. Let m_k denote the kth column of M. Let ε_k denote the kth column of ε. The term S_{III} can be expressed as

$$\begin{aligned}
(S_{\text{III}})_{ij} &= m^{-1}(e_i - p_i)^T M_{J_1} \varepsilon_{J_1}^T (e_j - p_j) \\
&= m^{-1} \operatorname{tr}\left(\varepsilon_{J_1}^T (e_j - p_j)(e_i - p_i)^T M_{J_1}\right) \\
&= m^{-1} \sum_{k \in J_1} \varepsilon_k^T (e_j - p_j)(e_i - p_i)^T m_k \\
&= m^{-1} \operatorname{vec}\{\varepsilon_{J_1}\}^T \left(I_{m_1} \otimes (e_j - p_j)(e_i - p_i)^T\right) \operatorname{vec}\{M_{J_1}\} \\
&= m^{-1} \operatorname{vec}\{Z\}^T \left(A_{J_1}^{1/2} \otimes B^{1/2}(e_j - p_j)(e_i - p_i)^T\right) \operatorname{vec}\{M_{J_1}\} \\
&= \operatorname{vec}\{Z\}^T \psi_{ij},
\end{aligned}$$

where

$$\psi_{ij} := m^{-1} \left(A_{J_1}^{1/2} \otimes B^{1/2}(e_j - p_j)(e_i - p_i)^T\right) \operatorname{vec}\{M_{J_1}\}. \tag{3.129}$$

The squared Euclidean norm of ψ_{ij} is

$$\begin{aligned}
\|\psi_{ij}\|_2^2 &= \operatorname{vec}\{M_{J_1}\}^T \left(A_{J_1} \otimes (e_i - p_i)(e_j - p_j)^T B (e_j - p_j)(e_i - p_i)^T\right) \operatorname{vec}\{M_{J_1}\} / m^2 \\
&= \operatorname{vec}\{M_{J_1}\}^T \left(A_{J_1} \otimes \breve{b}_{jj}(e_i - p_i)(e_i - p_i)^T\right) \operatorname{vec}\{M_{J_1}\} / m^2 \\
&= \breve{b}_{jj} \sum_{k \in J_1} \sum_{\ell \in J_1} a_{k\ell} m_k^T (e_i - p_i)(e_i - p_i)^T m_\ell / m^2 \\
&= \breve{b}_{jj} \sum_{k \in J_1} \sum_{\ell \in J_1} a_{k\ell} (\delta_n)_i \gamma_k (\delta_n)_i \gamma_\ell / \left(4m^2\right) \\
&= \breve{b}_{jj} \sum_{k \in J_1} \sum_{\ell \in J_1} a_{k\ell} \gamma_k \gamma_\ell / \left(4m^2\right) \\
&= \breve{b}_{jj} \gamma_{J_1}^T A_{J_1} \gamma_{J_1} / \left(4m^2\right).
\end{aligned} \tag{3.130}$$

By the Hanson-Wright inequality (Theorem 2.1),

$$\mathbb{P}\left(\left|\mathrm{vec}\left\{Z\right\}^T \psi_{ij} - \left\|\psi_{ij}\right\|_2\right| > d^{1/2}K^2\sqrt{\log(m)}\left\|\psi_{ij}\right\|_2\right) \leqslant 2\exp\left\{-d\log(m)\right\} = 2/m^d. \tag{3.131}$$

Therefore

$$\begin{aligned}\mathbb{P}\left(\left|(S_{\mathrm{III}})_{ij}\right| > \left(1 + d^{1/2}K^2\sqrt{\log(m)}\right)\left\|\psi_{ij}\right\|_2\right) &= \mathbb{P}\left(\left|\mathrm{vec}\left\{Z\right\}^T \psi_{ij}\right| > \left\|\psi_{ij}\right\|_2 + d^{1/2}K^2\sqrt{\log(m)}\left\|\psi_{ij}\right\|_2\right. \\ &\leqslant \mathbb{P}\left(\left|\mathrm{vec}\left\{Z\right\}^T \psi_{ij} - \left\|\psi_{ij}\right\|_2\right| > d^{1/2}K^2\sqrt{\log(m)}\left\|\psi_{ij}\right\|_2\right. \\ &\leqslant 2/m^d,\end{aligned}$$

where the last step follows from (3.131). By (3.130), it follows that

$$\left(1 + d^{1/2}K^2\sqrt{\log(m)}\right)\left\|\psi_{ij}\right\|_2 \leqslant r_3, \tag{3.132}$$

so

$$\mathbb{P}\left(\left|(S_{\mathrm{III}})_{ij}\right| > r_3\right) \leqslant \mathbb{P}\left(\left|(S_{\mathrm{III}})_{ij}\right| > \left(1 + d^{1/2}K^2\sqrt{\log(m)}\right)\left\|\psi_{ij}\right\|_2\right) \leqslant 2/m^d, \tag{3.133}$$

by (3.132). By the union bound,

$$\mathbb{P}\left(\left\|S_{\mathrm{III}}\right\|_\infty > r_3\right) \leqslant \sum_{i=1}^{m}\sum_{j=1}^{m}\mathbb{P}\left(\left|(S_{\mathrm{III}})_{ij}\right| > r_3\right) \leqslant 2/m^{d-2}.$$

3.7.5 Proof of part II of Lemma III.23, term IV

We now analyze term S_{IV}. To do so, we express S_{IV} as a quadratic form in order to apply the Hanson-Wright inequality.

Let $p_i^{(1)}$ denote the ith column of P_1^T. Let $p_i^{(2)}$ denote the ith column of P_2^T. Define

$$H_{\text{group}}^{ij} = I_{m_0} \otimes \left(e_j - p_j^{(2)}\right)\left(e_j - p_j^{(2)}\right)^T \qquad \text{and} \qquad H_{\text{global}}^{ij} = I_{m_1} \otimes \left(e_j - p_j^{(1)}\right)\left(e_j - p_j^{(1)}\right)^T, \tag{3.134}$$

and let

$$H^{ij}(J_0, J_1) = \begin{bmatrix} H_{\text{group}}^{ij} & \\ & H_{\text{global}}^{ij} \end{bmatrix}, \tag{3.135}$$

where $H_{\text{group}}^{ij} \in \mathbb{R}^{m_0 n \times m_0 n}$, $H_{\text{global}}^{ij} \in \mathbb{R}^{m_1 n \times m_1 n}$, and $H^{ij}(J_0, J_1) \in \mathbb{R}^{mn \times mn}$. Recall that

$$S_{\text{IV}} = m^{-1}(I - P_2)\varepsilon_{J_0}\varepsilon_{J_0}^T(I - P_2) + m^{-1}(I - P_1)\varepsilon_{J_1}\varepsilon_{J_1}^T(I - P_1).$$

The second term of S_{IV} can be expressed as a quadratic form, as follows (where ε_k denotes the kth column of $\varepsilon \in \mathbb{R}^{n \times m}$):

$$\begin{aligned} m^{-1}(I - P_1)\varepsilon_{J_1}\varepsilon_{J_1}^T(I - P_1) &= m^{-1} \sum_{k \in J_1} \left(e_i - p_i^{(1)}\right)^T \varepsilon_k \varepsilon_k^T \left(e_j - p_j^{(1)}\right) \\ &= m^{-1} \sum_{k \in J_1} \operatorname{tr}\left(\left(e_i - p_i^{(1)}\right)^T \varepsilon_k \varepsilon_k^T \left(e_j - p_j^{(1)}\right)\right) \\ &= m^{-1} \sum_{k \in J_1} \varepsilon_k^T \left(e_j - p_j^{(1)}\right)\left(e_i - p_i^{(1)}\right)^T \varepsilon_k \\ &= m^{-1} \operatorname{vec}\{\varepsilon_{J_1}\}^T \left(I_{m_1} \otimes \left(e_j - p_j^{(1)}\right)\left(e_i - p_i^{(1)}\right)^T\right) \operatorname{vec}\{\varepsilon_{J_1}\}^T \\ &= m^{-1} \operatorname{vec}\{\varepsilon_{J_1}\}^T H_{\text{global}}^{ij} \operatorname{vec}\{\varepsilon_{J_1}\}^T. \end{aligned} \tag{3.136}$$

Analogously, the first term of S_{IV} can be expressed as a quadratic form:

$$\begin{aligned} m^{-1}(I - P_2)\varepsilon_{J_0}\varepsilon_{J_0}^T(I - P_2) &= m^{-1} \sum_{k \in J_0} \left(e_i - p_i^{(2)}\right)^T \varepsilon_k \varepsilon_k^T \left(e_j - p_j^{(2)}\right) \\ &= m^{-1} \operatorname{vec}\{\varepsilon_{J_0}\}^T H_{\text{group}}^{ij} \operatorname{vec}\{\varepsilon_{J_0}\}^T. \end{aligned} \tag{3.137}$$

We now express S_{IV} as a quadratic form. Let $\pi(X)$ denote the matrix X with

reordered columns:

$$\pi(X) = \begin{bmatrix} X_{J_0} & X_{J_1} \end{bmatrix} \quad \text{and} \quad \pi(A) = \mathrm{Cov}\left(\mathrm{vec}\left\{\pi(X)\right\}\right). \tag{3.138}$$

Then by (3.136) and (3.137),

$$\begin{aligned}
(S_{\mathrm{IV}})_{ij} &= m^{-1}\,\mathrm{vec}\left\{\varepsilon_{J_0}\right\}^T H^{ij}_{\mathrm{group}}\,\mathrm{vec}\left\{\varepsilon_{J_0}\right\}^T + m^{-1}\,\mathrm{vec}\left\{\varepsilon_{J_1}\right\}^T H^{ij}_{\mathrm{global}}\,\mathrm{vec}\left\{\varepsilon_{J_1}\right\}^T \\
&= m^{-1}\,\mathrm{vec}\left\{\pi\left(\varepsilon\right)\right\}^T H^{ij}(J_0, J_1)\,\mathrm{vec}\left\{\pi\left(\varepsilon\right)\right\} \\
&= m^{-1}\,\mathrm{vec}\left\{Z\right\}^T \left(\left(\pi(A)^{1/2} \otimes B^{1/2}\right) H^{ij}(J_0, J_1) \left(\pi(A)^{1/2} \otimes B^{1/2}\right)\right) \mathrm{vec}\left\{Z\right\},
\end{aligned}$$

where the last step holds by decorrelation, with $Z \in \mathbb{R}^{n \times m}$ as a random matrix with independent subgaussian entries.

Note that the (i, j)th entry of S_{IV} can be expressed as

$$\left(S_{\mathrm{IV}}\right)_{ij} = \mathrm{vec}\left\{Z\right\}^T \Phi_{i,j}\,\mathrm{vec}\left\{Z\right\}, \tag{3.139}$$

with

$$\Phi_{i,j} = m^{-1}\left(\pi(A)^{1/2} \otimes B^{1/2}\right) H^{ij}(J_0, J_1) \left(\pi(A)^{1/2} \otimes B^{1/2}\right). \tag{3.140}$$

Having expressed $\left(S_{\mathrm{IV}}\right)_{ij}$ as a quadratic form in (3.139), we find the trace and Frobenius norm of $\Phi_{i,j}$, then apply the Hanson-Wright inequality. First we find the trace of $\Phi_{i,j}$. Let

$$\mathcal{I}_0 = \begin{bmatrix} I_{m_0 \times m_0} & 0_{m_0 \times m_1} \\ 0_{m_1 \times m_0} & 0_{m_1 \times m_1} \end{bmatrix} \quad \text{and} \quad \mathcal{I}_1 = \begin{bmatrix} 0_{m_0 \times m_0} & 0_{m_0 \times m_1} \\ 0_{m_1 \times m_0} & I_{m_1 \times m_1} \end{bmatrix}. \tag{3.141}$$

Note that $H^{ij}(J_0, J_1)$ can be written as a sum of Kronecker products,

$$H^{ij}(J_0, J_1) = \mathcal{I}_0 \otimes \left(e_j - p_j^{(2)}\right)\left(e_i - p_i^{(2)}\right)^T + \mathcal{I}_1 \otimes \left(e_j - p_j^{(1)}\right)\left(e_i - p_i^{(1)}\right)^T, \tag{3.142}$$

hence (3.140) can be expressed as

$$m^{-1}\left(\pi(A)^{1/2}\otimes B^{1/2}\right)\left(\mathcal{I}_0\otimes\left(e_j-p_j^{(2)}\right)\left(e_i-p_i^{(2)}\right)^T\right)\left(\pi(A)^{1/2}\otimes B^{1/2}\right) \tag{3.143}$$

$$+\,m^{-1}\left(\pi(A)^{1/2}\otimes B^{1/2}\right)\left(\mathcal{I}_1\otimes\left(e_j-p_j^{(1)}\right)\left(e_i-p_i^{(1)}\right)^T\right)\left(\pi(A)^{1/2}\otimes B^{1/2}\right). \tag{3.144}$$

The trace of the term (3.143) is

$$\begin{aligned}
&m^{-1}\operatorname{tr}\left(\left(\pi(A)^{1/2}\otimes B^{1/2}\right)\left(\mathcal{I}_0\otimes\left(e_j-p_j^{(2)}\right)\left(e_i-p_i^{(2)}\right)^T\right)\left(\pi(A)^{1/2}\otimes B^{1/2}\right)\right)\\
&=m^{-1}\operatorname{tr}\left(\pi(A)^{1/2}\mathcal{I}_0\pi(A)^{1/2}\otimes B^{1/2}\left(e_j-p_j^{(2)}\right)\left(e_i-p_i^{(2)}\right)^T B^{1/2}\right)\\
&=m^{-1}\operatorname{tr}\left(\pi(A)^{1/2}\mathcal{I}_0\pi(A)^{1/2}\right)\operatorname{tr}\left(B^{1/2}\left(e_j-p_j^{(2)}\right)\left(e_i-p_i^{(2)}\right)^T B^{1/2}\right)\\
&=m^{-1}\operatorname{tr}\left(\mathcal{I}_0\pi(A)\right)\left(\left(e_i-p_i^{(2)}\right)^T B\left(e_j-p_j^{(2)}\right)\right)\\
&=m^{-1}\operatorname{tr}\left(A_{J_0}\right)\left[(I-P_2)B(I-P_2))\right]_{ij}\\
&=m^{-1}\operatorname{tr}\left(A_{J_0}\right)\widetilde{b}_{ij}.
\end{aligned}$$

Analogously, the trace of the term (3.144) is

$$\begin{aligned}
&m^{-1}\operatorname{tr}\left(\left(\pi(A)^{1/2}\otimes B^{1/2}\right)\left(\mathcal{I}_1\otimes\left(e_j-p_j^{(1)}\right)\left(e_i-p_i^{(1)}\right)^T\right)\left(\pi(A)^{1/2}\otimes B^{1/2}\right)\right)\\
&=m^{-1}\operatorname{tr}\left(A_{J_1}\right)\left[(I-P_1)B(I-P_1))\right]_{ij}\\
&=m^{-1}\operatorname{tr}\left(A_{J_1}\right)\breve{b}_{ij}.
\end{aligned}$$

Let $b_{ij}^\sharp$ denote the (i,j)th entry of $B^\sharp$ defined in (3.87). We have shown that the trace of $\Phi_{i,j}$ (as defined in (3.140)) is

$$\operatorname{tr}\left(\Phi_{i,j}\right)=m^{-1}\operatorname{tr}\left(A_{J_0}\right)\widetilde{b}_{ij}+m^{-1}\operatorname{tr}\left(A_{J_1}\right)\breve{b}_{ij}=b_{ij}^\sharp. \tag{3.145}$$

Next, we find the Frobenius norm of $\Phi_{i,j}$. For convenience, define

$$\mathcal{A}_0 = \pi(A)^{1/2}\mathcal{I}_0\pi(A)^{1/2} \quad \text{and} \quad \mathcal{A}_1 = \pi(A)^{1/2}\mathcal{I}_1\pi(A)^{1/2} \tag{3.146}$$

$$\mathcal{B}_{2,ij} = B^{1/2}\left(e_j - p_j^{(2)}\right)\left(e_i - p_i^{(2)}\right)^T B^{1/2} \quad \text{and} \quad \mathcal{B}_{1,ij} = B^{1/2}\left(e_j - p_j^{(1)}\right)\left(e_i - p_i^{(1)}\right)^T B^{1/2}. \tag{3.147}$$

Then

$$\begin{aligned}
\|\Phi_{i,j}\|_F^2 &= \left\|m^{-1}\left(\pi(A)^{1/2}\otimes B^{1/2}\right)H^{ij}(J_0,J_1)\left(\pi(A)^{1/2}\otimes B^{1/2}\right)\right\|_F^2 \\
&= m^{-2}\left\|\mathcal{A}_0\otimes\mathcal{B}_{2,ij} + \mathcal{A}_1\otimes\mathcal{B}_{1,ij}\right\|_F^2 \\
&= m^{-2}\operatorname{tr}\left(\left(\mathcal{A}_0\otimes\mathcal{B}_{2,ij} + \mathcal{A}_1\mathcal{B}_{1,ij}\right)^T\left(\mathcal{A}_0\otimes\mathcal{B}_{2,ij} + \mathcal{A}_1\otimes\mathcal{B}_{1,ij}\right)\right) \\
&= m^{-2}\operatorname{tr}\left(\mathcal{A}_0^T\mathcal{A}_0\otimes\mathcal{B}_{2,ij}^T\mathcal{B}_{2,ij}\right) + m^{-2}\operatorname{tr}\left(\mathcal{A}_1^T\mathcal{A}_1\otimes\mathcal{B}_{1,ij}^T\mathcal{B}_{1,ij}\right) \\
&\quad + m^{-2}\operatorname{tr}\left(\mathcal{A}_0^T\mathcal{A}_1\otimes\mathcal{B}_{2,ij}^T\mathcal{B}_{1,ij}\right) + m^{-2}\operatorname{tr}\left(\mathcal{A}_1^T\mathcal{A}_0\otimes\mathcal{B}_{1,ij}^T\mathcal{B}_{2,ij}\right).
\end{aligned} \tag{3.148}$$

We now find the traces of each of the terms in (3.148). First, note that

$$\operatorname{tr}\left(\mathcal{A}_0^T\mathcal{A}_0\right) = \operatorname{tr}\left(\mathcal{I}_0\pi(A)\mathcal{I}_0\pi(A)\right) = \operatorname{tr}\left(A_{J_0}^2\right) = \|A_{J_0}\|_F^2. \tag{3.149}$$

Analogously,

$$\operatorname{tr}\left(\mathcal{A}_1^T\mathcal{A}_1\right) = \|A_{J_1}\|_F^2. \tag{3.150}$$

For the cross-term, let $A_{J_0J_1}$ denote the $m_0\times m_1$ submatrix of $\pi(A)$ given by columns

of A in J_0 and rows of A in J_1. Then

$$\begin{aligned}\operatorname{tr}\left(\mathcal{A}_0^T\mathcal{A}_1\right) &= \operatorname{tr}\left(\mathcal{I}_0\pi(A)\mathcal{I}_1\pi(A)\right)\\ &= \operatorname{tr}\left(\begin{bmatrix} 0_{m_0\times m_0} & A_{J_0J_1}\\ 0_{m_1\times m_0} & 0_{m_1\times m_1}\end{bmatrix}\pi(A)\right)\\ &= \operatorname{tr}\left(A_{J_0J_1}^T A_{J_0J_1}\right)\\ &= \|A_{J_0J_1}\|_F^2\,. \end{aligned} \tag{3.151}$$

Next,

$$\begin{aligned}\operatorname{tr}\left(\mathcal{B}_{1,ij}^T\mathcal{B}_{1,ij}\right) &= \operatorname{tr}\left(B^{1/2}\left(e_i-p_i^{(1)}\right)\left(e_j-p_j^{(1)}\right)^T B\left(e_j-p_j^{(1)}\right)\left(e_i-p_i^{(1)}\right)^T B^{1/2}\right)\\ &= \left(\left(e_j-p_j^{(1)}\right)^T B\left(e_j-p_j^{(1)}\right)\right)\left(\left(e_i-p_i^{(1)}\right)^T B\left(e_i-p_i^{(1)}\right)\right)\\ &= \breve{b}_{jj}\breve{b}_{ii}. \end{aligned} \tag{3.152}$$

Analogously,

$$\begin{aligned}\operatorname{tr}\left(\mathcal{B}_{2,ij}^T\mathcal{B}_{2,ij}\right) &= \left(\left(e_j-p_j^{(2)}\right)^T B\left(e_j-p_j^{(2)}\right)\right)\left(\left(e_i-p_i^{(2)}\right)^T B\left(e_i-p_i^{(2)}\right)\right)\\ &= \widetilde{b}_{jj}\widetilde{b}_{ii}. \end{aligned} \tag{3.153}$$

The cross-terms yield

$$\operatorname{tr}\left(\mathcal{B}_{1,ij}^T\mathcal{B}_{2,ij}\right) = \left(\left(e_j-p_j^{(1)}\right)^T B\left(e_j-p_j^{(2)}\right)\right)\left(\left(e_i-p_i^{(2)}\right)^T B\left(e_i-p_i^{(1)}\right)\right) = \breve{b}_{ii}\breve{b}_{jj}. \tag{3.154}$$

The squared Frobenius norm of $\Phi_{i,j}$ is

$$\begin{aligned}\|\Phi_{i,j}\|_F^2 &= \frac{1}{m^2}\left(\|A_{J_0}\|_F^2 \breve{b}_{ii}\breve{b}_{jj} + \|A_{J_1}\|_F^2 \widetilde{b}_{ii}\widetilde{b}_{jj} + 2\|A_{J_0,J_1}\|_F^2 \breve{b}_{ii}\breve{b}_{jj}\right) \\ &\leqslant \frac{1}{m^2} C\left(\|A_{J_0}\|_F^2 + \|A_{J_1}\|_F^2 + 2\|A_{J_0 J_1}\|_F^2\right)\|B\|_2^2 \\ &= C\frac{1}{m^2}\|A\|_F^2\|B\|_2^2 .\end{aligned}$$

We now apply the Hanson-Wright inequality,

$$\begin{aligned}\mathbb{P}\left(\left|(S_{\mathrm{I}})_{ij} - b_{ij}^{\sharp}\right| > r_4\right) &= \mathbb{P}\left(\left|\mathrm{vec}\{Z\}^T \Phi_{i,j}\,\mathrm{vec}\{Z\} - \mathrm{tr}\left(\Phi_{i,j}\right)\right| > r_4\right) \\ &\leqslant 2\exp\left(-c\min\left\{d\log(m), d^{1/2}\sqrt{\log(m)}\frac{\|\Phi_{i,j}\|_F}{\|\Phi_{i,j}\|_2}\right\}\right) \\ &\leqslant 2\max\left(m^{-d}, \exp\left(d^{1/2}\sqrt{\log(m)}r^{1/2}(\Phi_{i,j})\right)\right).\end{aligned}$$

The first step holds by (3.139) and (3.145).

CHAPTER IV

Matrix-variate modeling of pitch curves in linguistics research

This chapter is joint work with my advisors Kerby Shedden and Shuheng Zhou.

Phonetics is the branch of linguistics that considers the production and perception of speech sounds. Large volumes of speech data from human volunteers can be readily collected for analysis. One common type of phonetic data that is of interest to linguistics researchers is "pitch curve" data, in which the frequency of voiced sounds is quantified at high temporal resolution. These curves can be seen as a form of functional data, in that the pitch varies smoothly in time. Pitch curves are relevant for addressing a variety of research questions in psycholinguistics, including questions related to language change and the relationship between subtle acoustical variations in speech and people's perception of it. Such analyses may involve contrasting pitch curves within and between subjects, words, word categories, and populations of speakers.

Studies involving pitch curves require substantial data pre-processing, for example, to segment the speech into words or word fragments by identifying consonant boundaries of word fragments. Here we consider collections of pitch curves that have

been pre-processed into vectors of 19 pitch measures within a word. The pitch measurements are equally spaced in time within a word, but not necessarily between words, in order to accommodate differing word durations and variation in people's rates of speech.

Research questions in linguistics may focus on language change within a population over time, heterogeneity in speech patterns within a population at a particular point in time, and relationships between production and perception of speech. Research studies in this area tend to involve large numbers of recordings per subject, since once a subject is recruited to the study it is relatively easy and inexpensive to record them speaking many words. On the other hand, logistical and cost constraints may limit the number of different subjects in a study. Thus, the design of linguistics studies resembles that of many studies in cognitive psychology and neuroscience, in that there are relatively few subjects, with many trials per subject. Traditional statistical methods for repeated measures data, such as hierarchical random effects regression have been widely applied in this field (*Baayen et al.*, 2008; *Clark*, 1973; *Quené and Van den Bergh*, 2008; *Aston et al.*, 2010). Here we consider recently-developed statistical approaches for analyzing matrix variate data as potential tools for use by researchers in this area. In particular, we consider the covariance matrices and graph structures among words, and among time points within an utterance of a word. We argue that understanding conditional independence structures will allow researchers to gain insights into group-wise differences in speech perception and production, and learn about inter-individual variation in speech production and processing.

A key issue is that we require an overarching model to define how within-index associations (e.g. associations among words) can be integrated into an overall covariance structure for the data. Previous researchers have proposed Kronecker product-based and sum-based approaches for doing this. For example, the Gemini approach (Zhou 2014), considered and extended earlier in chapters 2 and 3 of this thesis, is a product-

based approach to covariance modeling. The Terralasso (*Greenwald et al.*, 2017) and other recently proposed approaches (*Park et al.*, 2017) are sum-based.

Here we analyze data consisting of pitch curves in the Afrikaans language collected from 23 female native speakers of the language (*Coetzee et al.*, 2018). Each subject uttered each of 93 distinct Afrikaans words four times (four trials). The order of the 93×4 word presentations was randomized. The speaker's pitch is measured at 19 time points. The original purpose of the study was to gain an understanding of perception/production associations among Afrikaans speakers, and to consider this in the context of intergenerational language change. Here, we focus instead on relating acoustical similarity as inferred through the pitch curve data to pre-defined word attributes.

4.1 Introduction to pitch curve data

4.1.1 Phonetics terminology

In phonetics, consonants can be grouped based on the physical mechanism of their pronunciation; such categories include labial, alveolar, nasal, and fricative consonants (*Laver*, 1994). Labial consonants (e.g. b and p) are pronounced with the lips; alveolar consonants (e.g. t and d) are pronounced with the tongue behind the teeth; fricatives (e.g. v and f) are pronounced with partial of obstruction of the air; nasal consonants (e.g. m and n) are pronounced with air passing through the nose.

Voicing refers to whether the vocal folds (also called vocal cords) vibrate during pronunciation. For a "voiced" pronunciation the vocal cords vibrate, whereas for a "voiceless" pronunciation they do not. While some consonants are voiced, the vast majority of pitch curve data is based on vowels, which are always voiced. Typically voiced consonants in Afrikaans include b, d, w, v, m, and n. Typically voiceless consonants in Afrikaans include p, t, f, and, and k. The International Phonetic

Alphabet (IPA) represents sounds across multiple languages. The words selected for the Afrikaans data (*Coetzee et al.*, 2018) contain five IPA vowels.

4.1.2 Voicing and pitch in linguistics research

Voicing in linguistics refers to whether the vocal folds vibrate during an utterance (*Ladefoged and Disner*, 2012). Prior research in linguistics has found that vowel pitch after voiceless consonants is higher on average than after voiced consonants (*House and Fairbanks*, 1953). As demonstrated by subsequent research, this finding holds in multiple languages, including French and Italian (*Kirby and Ladd*, 2016).

Linguists have performed studies to investigate the reason for this phenomenon. *Hanson* (2009) compared pitch after voiced and voiceless consonants to pitch after nasal consonants, treating nasal consonants as a reference point. Nasal consonants were chosen as a reference for physiological reasons, in particular airflow through the nose does not disrupt pitch (*Hanson*, 2009).

Hanson (2009) examined English syllables, spoken in either a high pitch context or a low pitch context (where the test syllable was embedded in a sentence, and the researcher demonstrated how to pronounce the sentence with high pitch or low pitch). The study found that in a high pitch context, vowel pitch after a voiceless consonant is higher than after a nasal during the initial 100 ms of the vowel; by contrast, vowel pitch after a voiced consonant is comparable to that after a nasal consonant.

We analyze data on vowel pitch after voiced and voiceless consonants in Afrikaans. The purpose of the study by *Coetzee et al.* (2018) is to assess whether speakers of Afrikaans speak with raised vowel pitch after voiceless consonants, compared with voiced consonants, and whether listeners utilize this pitch difference to aid in perception of the initial consonant.

Note that in our analysis, the covariance estimation is not driven by the mean pitch level, because (as discussed in Section 4.2), we remove mean structure through

trial residualization (for each speaker, word, and time point, we subtract out the mean pitch over four utterances of the word).

4.1.3 Preliminary exploration of pitch curve data

To provide a basic illustration of pitch curve data, we display the mean pitch curves, averaged over speaker, trial, and word within initial consonant, in Figure 4.1. This graph demonstrates that mean vowel pitch curves depend on the initial consonant. In Figure 4.2, we display the first utterance of the word "met" for each of the speakers. These are individual raw pitch curves that have not been centered or averaged. There is substantial heterogeneity among speakers pronouncing a given word. Among other characteristics, we see that some speakers have higher voices, and others have lower voices.

When we analyze the pitch curve data, we first remove several sources of variation that are of secondary interest. For example, most people have stable speaking pitches (e.g. based on age and gender). Also, it is desirable to remove the stable (population averaged) pitch trajectory of a word, so that we analyze trial variation. We take this a step further and remove stable pitch curve features at the speaker $\times$ word level, so that we focus on variation present in individual utterances. Specifically, we center the data using trial residualization, subtracting from each individual pitch curve the corresponding point-wise mean pitch curve over each subject $\times$ word, taken over the four trials. To illustrate, in Figure 4.3, we display four trials, centered by first removing the speaker $\times$ word mean, and then averaging these residuals over speakers and words within an initial consonant type.

4.2 Matrix-variate models for pitch curve data

In the linguistics study considered here, the raw data can be represented as an array with four indices, corresponding respectively to speaker, word, trial, and time

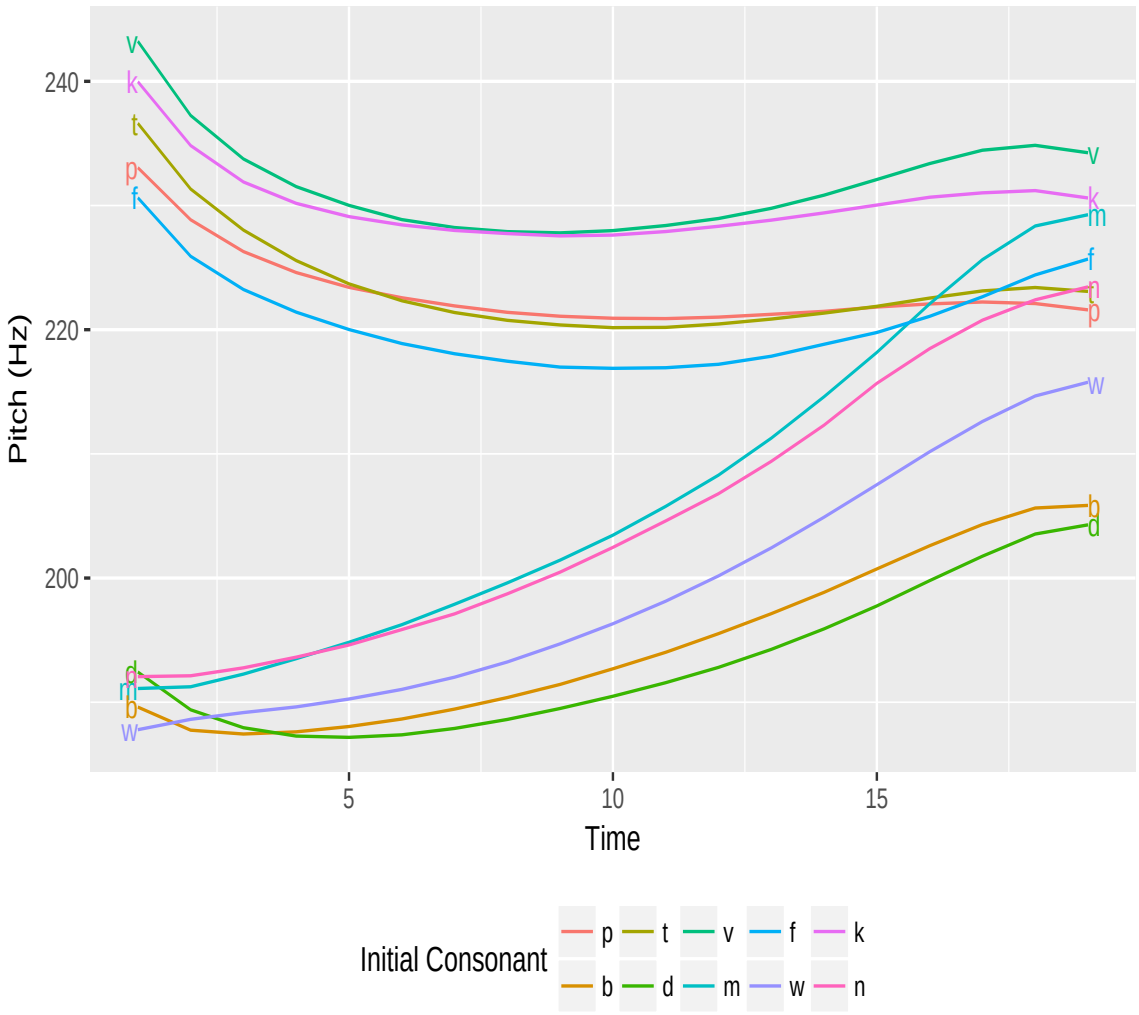

Figure 4.1: This figure displays pitch curves, averaged over speaker, trial, and word, for each initial consonant. The consonants p, t, f, and k are typically voiceless, whereas the consonants b, d, m, w, v, and n are typically voiced. This figure is related to Figure 6 in *Coetzee et al.* (2018), which displays pitch curves for older and younger speakers, for words starting with b, d, m, and n. As discussed in *Coetzee et al.* (2018), vowel pitch is higher on average after voiceless consonants than after voiced consonants.

point. Let $X_{i,j,r,t}$ denote the pitch measurement for speaker i, word j, trial r, and time t. Let n_s, n_w, n_r, and n_t denote the number of speakers, words, trials, and time points, respectively. We describe a matrix-variate model that captures word-word and time-time correlations, treating the trials as replicates nested within speakers by

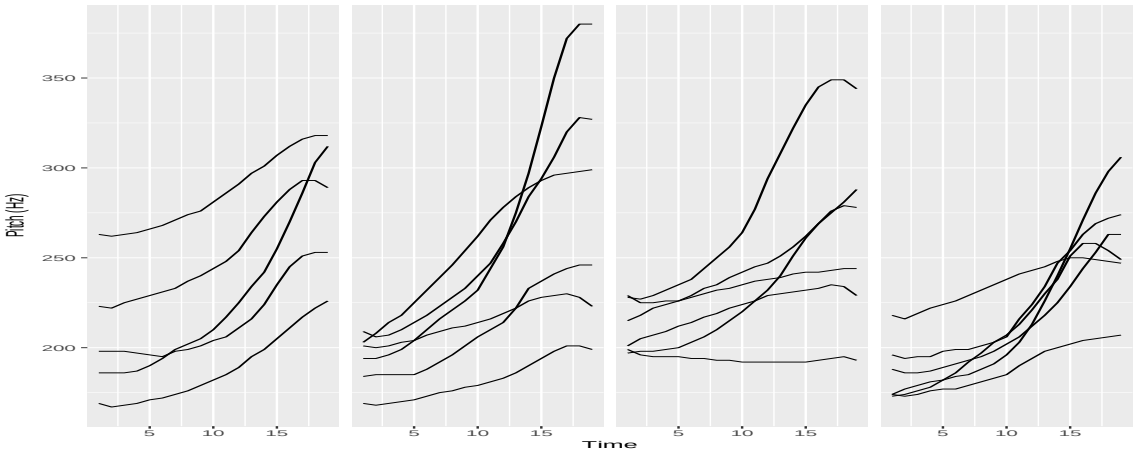

Figure 4.2: Pitch curves for the 23 speakers are displayed in four panels, for the word "met." For ease of visualization, the pitch curves for the speakers are displayed in four panels.

words. We assume that for each speaker i, a common mean matrix $M(i) \in \mathbb{R}^{n_w \times n_t}$ is shared across the four trials. Let $X(i,r) \in \mathbb{R}^{n_w \times n_t}$ denote the data for speaker i, trial r. Under our assumption,

$$X(i,r) - \frac{1}{n_r} \sum_{r=1}^{n_r} X(i,r) \tag{4.1}$$

has expected value zero.

For $r = 1, \ldots, n_r$ let $X(i,r) \in \mathbb{R}^{n_w \times n_t}$ denote speaker i's data for trial r. Adopting the Gemini approach, we consider the matrix-variate model

$$\mathrm{Cov}(\mathrm{vec}(X(i,r))) = A \otimes B, \tag{4.2}$$

where A is a time-time covariance matrix and B is a word-word covariance matrix. We will use estimation procedures with known properties to recover A and B from the data, then use the corresponding estimated graph structures to explore word-word and time-time associations.

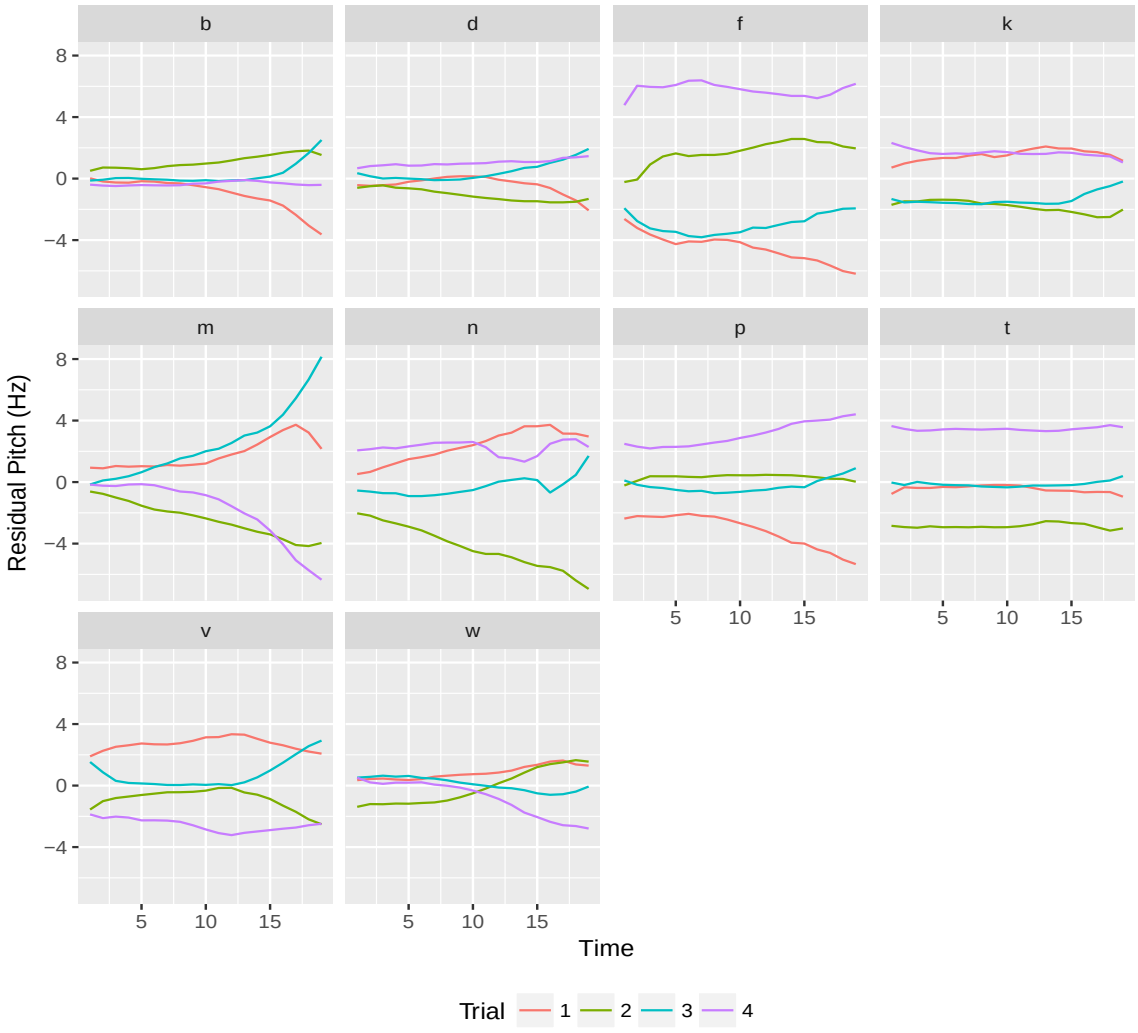

Figure 4.3: Pitch curves for each of the four trials, averaged over speaker and word for each initial consonant (with a separate panel shown for each initial consonant). The trials are centered as in (4.1).

4.2.1 Model-based centering

Since most data have mean as well as covariance structures, it will usually be necessary to remove the mean structure before or in parallel with covariance estimation. One natural two-stage approach is to use a flexible regression model to capture mean effects, and then proceed by estimating the covariance structure based on the residuals from the regression model fit. We found that when using a 30 degree of freedom regression model fit with least squares, having terms for age, voicing condition, and four b-splines for time, along with all pairwise interactions among these terms, the Gram matrices based on words were approximately low-rank (Figures 4.4 and 4.5). This suggests that the mean structure was not successfully removed. We therefore adopted the centering approach described above in (4.1), which yielded well-conditioned word $\times$ word Gram matrices.

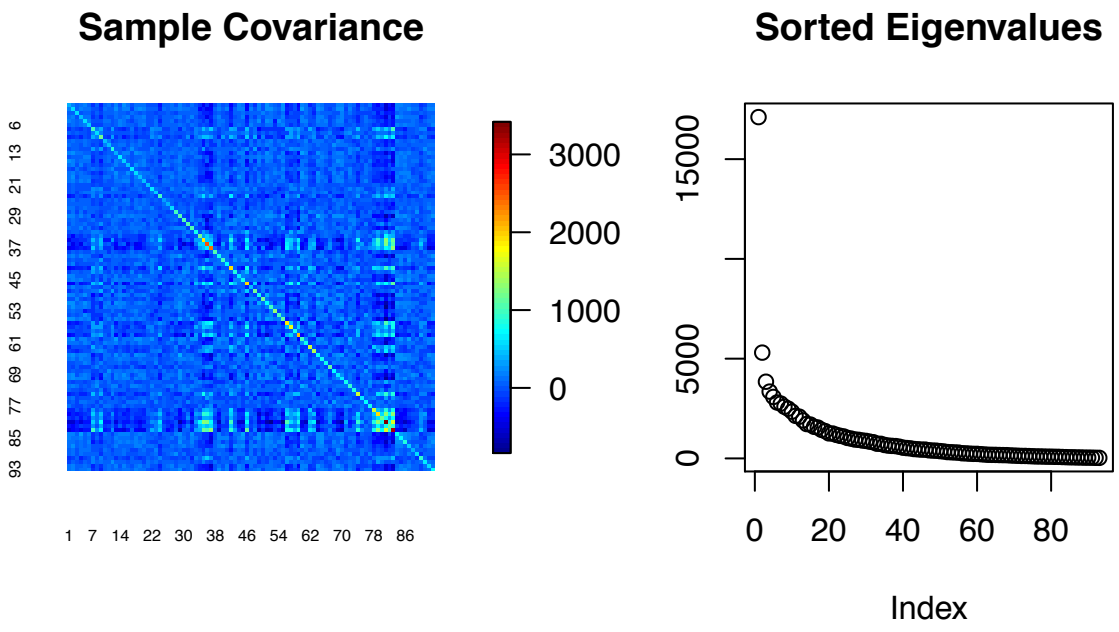

Figure 4.4: Heatmap of sample covariance matrix and sorted eigenvalues when the data is centered using a regression model including age, word voicing condition, and four basis splines to capture the effect of time.

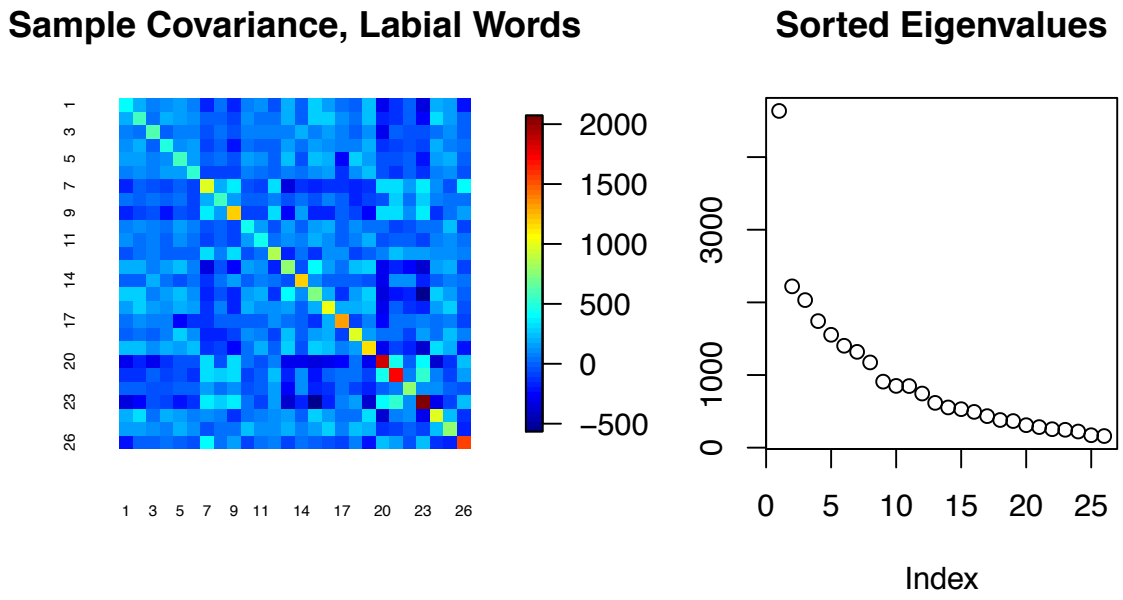

Figure 4.5: Heatmap of sample covariance matrix and sorted eigenvalues for labial words when the data is centered using a regression model including age, word voicing condition, and four basis splines to capture the effect of time.

4.2.2 Connections between trial differencing and trial centering

Due to the functional nature of the data, the pitch curves primarily occupy a subset of the Euclidean space $\mathbb{R}^{19}$ in which the pitch curve vectors lie. As a result, time-time Gram matrices calculated from raw data tend to have very large condition numbers. In particular, there is often a large dominating eigenvector reflecting variation in the typical pitches of different speakers' voices, e.g. with females and younger speakers tending to have higher pitched voices compared to males and older speakers. In addition, each word has a characteristic pitch curve common to all speakers that is of secondary interest here. We thus sought to remove these sources of variation that are unimportant to our aims. There are several ways to do this, including model-based approaches. We focus on a "trial-based" centering approach that removes the local mean for a given speaker uttering a given word.

The Afrikaans data set consists of four replicates, so there are multiple possible ways to take trial differences (e.g. trial 2 minus trial 1, trial 3 minus trial 2, etc.). We show that the for a particular combination of trial differences defined below, the trial differences can be expressed in terms of trial residualization (i.e. centering by subtracting out the mean over the trials). Trial residualization also removes the mean pitch level.

Define the matrices $D(1), D(2), D(3) \in \mathbb{R}^{n_t \times n_w n_s}$

$$
\begin{aligned}
D(1) &= X(2) + X(3) - X(1) - X(4) \\
D(2) &= X(3) + X(4) - X(1) - X(2) \\
D(3) &= X(1) + X(3) - X(2) - X(4),
\end{aligned}
$$

where $X(r) \in \mathbb{R}^{n_t \times n_w n_s}$ is the data for the rth trial of all speakers, $r = 1, \ldots, 4$.

Let

$$
S_1 = D(1)D(1)^T + D(2)D(2)^T + D(3)D(3)^T. \tag{4.3}
$$

We show that S_1 can be expressed in terms of trial-centered data. Note that the Gram matrices can be expressed as

$$\begin{aligned} D(1)D(1)^T = {} & X(2)X(2)^T + X(2)X(3)^T - X(2)X(1)^T - X(2)X(4)^T \\ & + X(3)X(2)^T + X(3)X(3)^T - X(3)X(1)^T - X(3)X(4)^T \\ & - X(1)X(2)^T - X(1)X(3)^T + X(1)X(1)^T + X(1)X(4)^T \\ & - X(4)X(2)^T - X(4)X(3)^T + X(4)X(1)^T + X(4)X(4)^T, \end{aligned}$$

$$\begin{aligned} D(2)D(2)^T = {} & X(3)X(3)^T + X(3)X(4)^T - X(3)X(1)^T - X(3)X(2)^T \\ & + X(4)X(3)^T + X(4)X(4)^T - X(4)X(1)^T - X(4)X(2)^T \\ & - X(1)X(3)^T - X(1)X(4)^T + X(1)X(1)^T + X(1)X(2)^T \\ & - X(2)X(3)^T - X(2)X(4)^T + X(2)X(1)^T + X(2)X(2)^T, \end{aligned}$$

and

$$\begin{aligned} D(3)D(3)^T = {} & X(1)X(1)^T + X(1)X(3)^T - X(1)X(2)^T - X(1)X(4)^T \\ & + X(3)X(1)^T + X(3)X(3)^T - X(3)X(2)^T - X(3)X(4)^T \\ & - X(2)X(1)^T - X(2)X(3)^T + X(2)X(2)^T + X(2)X(4)^T \\ & - X(4)X(1)^T - X(4)X(3)^T + X(4)X(2)^T + X(4)X(4)^T. \end{aligned}$$

Summing the Gram matrices and cancelling terms yields the expression

$$\begin{aligned} S_1 &= D(1)D(1)^T + D(2)D(2)^T + D(3)D(3)^T \\ &= 3\sum_{r=1}^{4} X(r)X(r)^T - \sum_{1 \leqslant r,\ell \leqslant 4, r \neq \ell} X(r)X(\ell)^T. \end{aligned} \tag{4.4}$$

Let

$$\overline{X} = \frac{1}{4}\sum_{r=1}^{4} X(r). \tag{4.5}$$

Then

$$\begin{aligned}
\sum_{r=1}^{4}(X(r)-\overline{X})(X(r)-\overline{X})^T &= \sum_{r=1}^{4} X(r)X(r)^T - 4\overline{X}(\overline{X}^T) \\
&= \sum_{r=1}^{4} X(r)X(r)^T - \frac{1}{4}\left(\sum_{r=1}^{4} X(r)X(r)^T + \sum_{1\leqslant r,\ell\leqslant 4, r\neq\ell} X(r)X(\ell)^T\right) \\
&= \frac{3}{4}\sum_{r=1}^{4} X(r)X(r)^T - \frac{1}{4}\sum_{1\leqslant r,\ell\leqslant 4, r\neq\ell} X(r)X(\ell)^T,
\end{aligned}$$

which is proportional to (4.4).

4.3 Covariance and precision matrices for time points and words

Our goal is to quantify the dependencies among words and among time points, using methods that target the conditional correlations between two words given all other words, or between two time points given all other time points. For matrix variate data that are dependent along only one axis, the Glasso is a widely-used approach for doing this. If there may be dependencies along both axes, and if the covariance matrix of the vectorized random matrix is a Kronecker product of factors corresponding to rows and to columns, Zhou (2014) showed that the Glasso can be applied separately to the row and column Gram matrices, but using a different regularization parameter to account for the additional dependence. Her work also showed that when replicates are present, less regularization is required compared to the setting with a single realization. Furthermore, Zhou (2014) proposed a three-step penalized algorithm in which the estimated precision matrix along one axis is used to decorrelate the data along the other axis, improving accuracy over baseline Gemini

estimators under specified conditions.

We treat the 20 subjects and four trials in the Afrikaans study as $20 \times 4 = 80$ independent random arrays with a common covariance structure. Each such array is a $n_t \times n_w$ matrix which has been centered over the trials as discussed above in (4.1). We then apply the Glasso method with a range of regularization parameters, separately to the word and time Gram matrices. This approach gives us graph structures among the words and among the time points.

Let $X(i,r) \in \mathbb{R}^{n_w \times n_t}$ denote the data for speaker i, trial r. Let

$$\overline{X}(i) = \frac{1}{n_r} \sum_{r=1}^{n_r} X(i,r) \tag{4.6}$$

denote the average over trials for speaker i. To estimate the covariance matrices, we calculate the word-word sample covariance matrix as

$$\frac{1}{n_s}\frac{1}{n_r} \sum_{i=1}^{n_s} \sum_{r=1}^{n_r} \left(X(i,r) - \overline{X}(i)\right) \left(X(i,r) - \overline{X}(i)\right)^T \in \mathbb{R}^{n_w \times n_w} \tag{4.7}$$

and the time-time sample covariance as

$$\frac{1}{n_s}\frac{1}{n_r} \sum_{i=1}^{n_s} \sum_{r=1}^{n_r} \left(X(i,r) - \overline{X}(i)\right)^T \left(X(i,r) - \overline{X}(i)\right) \in \mathbb{R}^{n_t \times n_t}. \tag{4.8}$$

Note that in this formulation, speakers and trials are taken as replicates, so each Gram matrix is an average of $n_s \cdot n_t$ Gram matrices.

As noted above, here we are working with 4-index data (speaker, time, word, replicate), but we wish to describe the population in terms of covariance and precision matrices. We can form a Gram matrix, say for words, by matricizing the 4-way tensor into a $n_w \times (n_t \cdot n_s \cdot n_r)$ matrix, then forming the $n_w \times n_w$ Gram matrix. Alternatively, we can think of the data as consisting of n_s replications of a 3-way tensor, in which case the word Gram matrix would result from matricizing the data to obtain n_s distinct

$n_w \times (n_t \cdot n_r)$ matrices. In the Gemini approach, when replicates are available, their Gram matrices are summed, so these two approaches lead to the same overall Gram matrix. However, the theoretical regularization level differs depending on whether the data are modeled as having independent replicates. Here, we treat speakers as independent replicates, and regularize accordingly.

4.3.1 Glasso regularization

The inverse covariance graphs are estimated using graphical lasso. For the time-time inverse correlation matrix, the penalty is

$$\lambda = \sqrt{\frac{\log(n_w)}{n_s \cdot n_r \cdot n_w}}, \tag{4.9}$$

where n_w is the number of words, n_s is the number of speakers, and n_r is the number of replicates. For the word-word inverse correlation matrix, the penalty is

$$\lambda = \sqrt{\frac{\log(n_w)}{n_s \cdot n_r \cdot n_{t,\text{eff}}}}, \tag{4.10}$$

where n_w is the number of words, and the denominator is the product of the number of people, trials, and effective time points per utterance. Note that the effective time points per utterance is smaller than 19, because the pitch curves are smooth curves, so adjacent points are dependent. Due to the stretched time scale over short vowels versus the long vowels, we believe that $n_{t,\text{eff}}$ for the short vowels is smaller than that for the long ones; hence we recommend using larger penalty when we interpret the graphs over short vowels. In future work, when we run cross-validation, we will assess whether larger penalties are selected for the short vowels.

4.3.2 Time-time and word-word correlation and covariance

Since the pitch curves are smooth, strong local correlations along the time axis are expected. The time-time dependence structure is informative in that it provides a characterization of the variance function of the pitch curves as a function of time, and reveals the extent to which local dependencies decay.

Figure 4.6 displays sample covariance, sample correlation, Glasso covariance, Glasso inverse covariance, Glasso correlation, and Glasso inverse correlation for the labial words. Glasso is run using a penalty five times that of of the theoretical value. In Section A.0.1 of the Appendix, analogous figures are shown for the other word groups.

The time-time covariance matrices for each word group (labial, alveolar, nasal, vf) indicate that the variance increases over time; that is, the pitch exhibits greater variability at the end of the word utterance than at the beginning. This indicates that speech may be more constrained at the beginning of a word token than at the end.

The correlation matrices are approximately banded, and essentially all pairwise correlations are above 0.5. In some cases the correlations decay faster at the end of the utterance than at the beginning.

The diagonal entries of the inverse covariance matrix reflect the residual variances of each time point when regressed on the other other time points; a small diagonal entry corresponds to large residual variance. For each of the word groups, the diagonal entries of the precision matrix are decreasing in time, also consistent with the early portion of the utterance being more constrained and predictable than the later portion of the utterance. Unless one has a strong conviction that the time-time covariance matrix (to be estimated) is nonstationary, it is worth trying to use it decorrelate the data along the time coordinate, so as to increase the accuracy in estimating the Pearson correlation coefficients between and among words, (c.f. Chapter 5, on future work). In Table 4.1, we report metrics related to the Glasso estimate of the time-time

Word Group	Avg. node degree	# edges	$\mathrm{tr}(B)/\|B\|_F$	$\|B\|_2$
All words (93 words)	9.3	88	4.05	5.0
Labial (26 words)	9.5	90	4.05	5.0
Alveolar (30 words)	9.8	93	4.05	5.0
Nasal words (20 words)	9.8	93	4.07	5.0
vf words (17 words)	8.4	80	4.03	5.1

Table 4.1: Metrics related to estimate of time-time correlation matrix.

Penalty	Avg. node degree	# edges	$\mathrm{tr}(A)/\|A\|_F$	$\|A\|_2$	$\kappa(A)$
0.1	27.94	1299	8.89	2.83	19.41
0.16	18.95	881	9.07	2.29	11.29
0.26	8.69	404	9.34	1.77	4.48
0.36	2.71	126	9.56	1.39	2.14
0.46	0.6	28	9.63	1.18	1.39

Table 4.2: Metrics related to estimate of word-word correlation matrix

correlation matrix. Based on the estimated effective sample using all words ($n_r = 3$, $n_w = 93$, $n_s = 20$), using the identity matrix for the word-word covariance, the theoretical penalty is $\sqrt{\log(93)/(20 * 3 * 93)} = 0.03$. In practice, due to dependence on the other axis, one should use a larger penalty when estimating the time-time inverse covariance.

4.3.3 Metrics for word-word inverse correlation estimates

We report metrics of the estimated correlation matrix for all words, using a sequence of Glasso penalty parameters in Table 4.2. Based on the estimated effective sample using all words ($n_r = 3$, $n_{t,\text{eff}} = 3$, $n_s = 20$), using the identity matrix for the time-time covariance, the theoretical penalty is $\sqrt{\log(93)/(20 * 3 * 3)} = 0.16$.

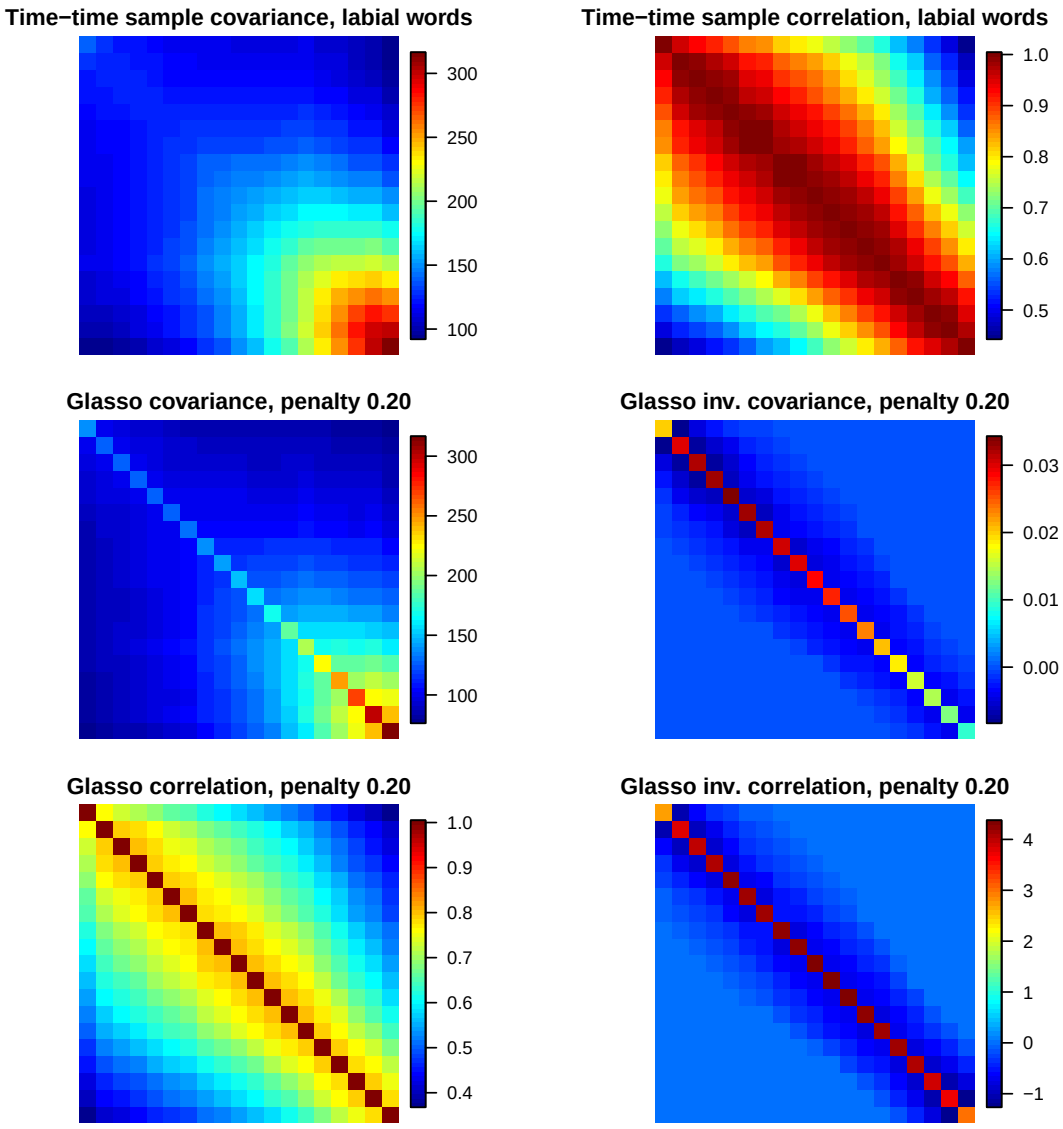

Figure 4.6: Time-time sample covariance (top left), sample correlation (top right), Glasso covariance (middle left), Glasso inverse covariance (middle right), Glasso correlation (bottom left), and Glasso inverse correlation (bottom right), for words beginning with a labial consonant. The sample covariance is calculated as in (4.8), and the Glasso penalty parameter is chosen as five times the value of (4.9).

4.3.4 Analyzing edges related to long and short vowels

We illustrate that for words with long vowels, edges are driven by the vowel, whereas for short vowels, this phenomenon does not seem to be apparent.

Figure 4.7 displays the estimated inverse covariance graph for words with long vowels, using nodewise regression with a penalty of 0.16 and threshold of 0.08. Figure 4.8 displays an analogous plot estimated with Glasso with penalty 0.32 and threshold of zero, and Figure 4.9 displays the analogous plot with a penalty of 0.39. We see the presence of several strong within-vowel group edges.

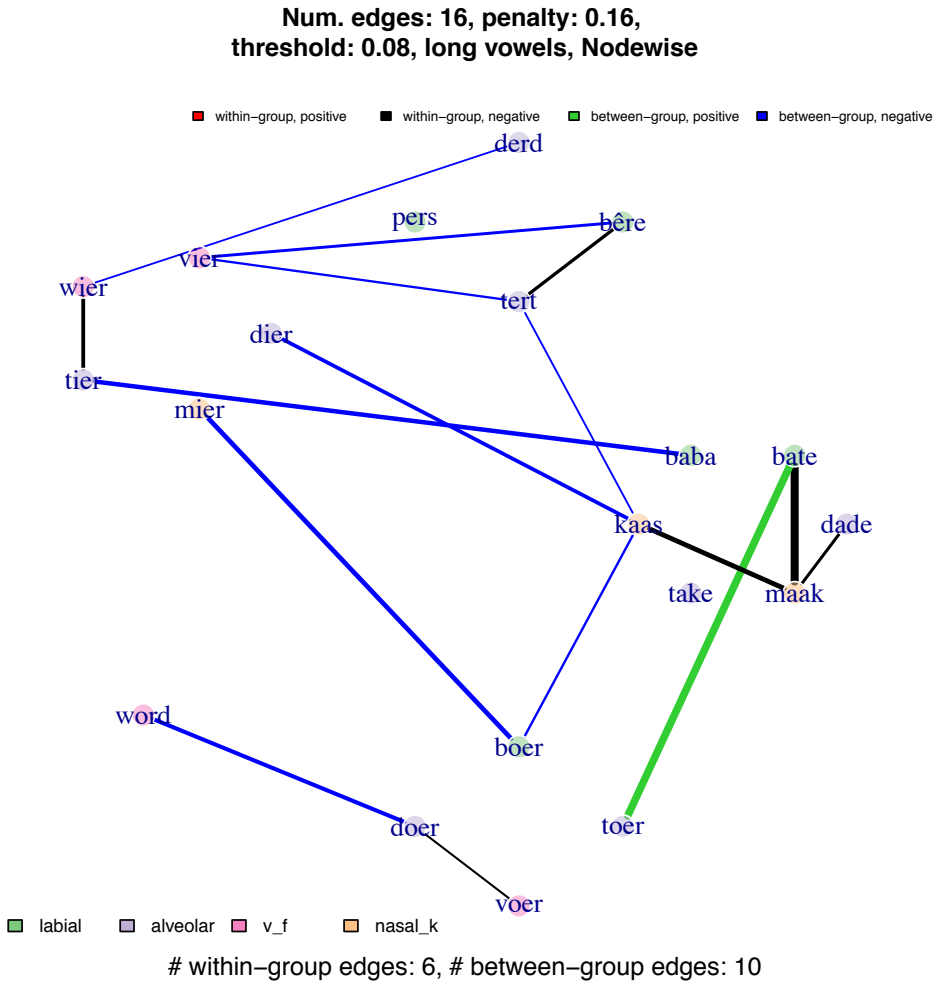

Figure 4.7: Inverse correlation edge graph for words with long vowels. Based on the estimated effective sample ($n_r = 3$, $n_{t,\text{eff}} = 3$ or 4, $n_s = 20$) and the theoretical guidance from *Zhou* (2014a), we believe the theoretical penalty should be in the range of $[0.11, 0.13]$; in future work, we aim to make this rigorous. The words are organized by vowel, with each circle of words sharing a common vowel ("word" is the only word with a long "o" vowel; in Afrikaans, it means "become").

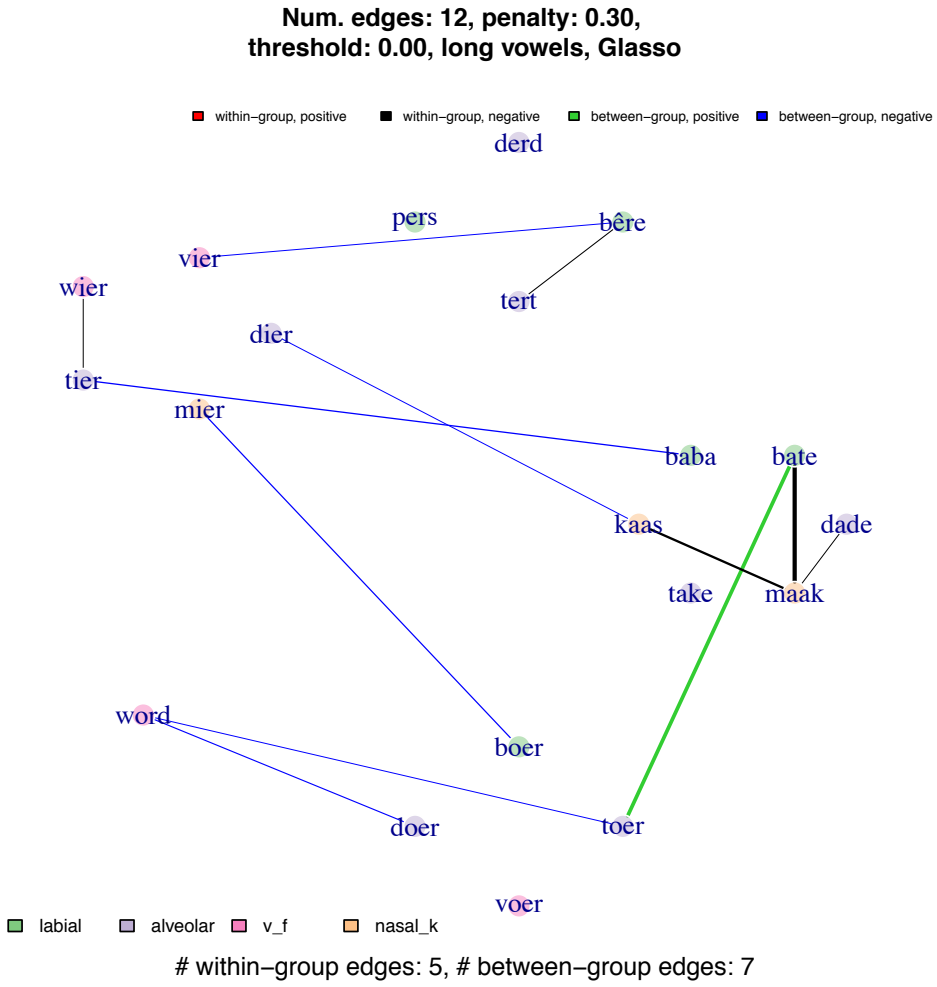

Figure 4.8: Inverse correlation edge graph for words with long vowels. Based on the estimated effective sample ($n_r = 3$, $n_{t,\text{eff}} = 3$ or 4, $n_s = 20$) and the theoretical guidance from *Zhou* (2014a), we believe the theoretical penalty should be in the range of $[0.11, 0.13]$; in future work, we aim to make this rigorous. The words are organized by vowel, with each circle of words sharing a common vowel ("word" is the only word with a long "o" vowel; in Afrikaans, it means "become").

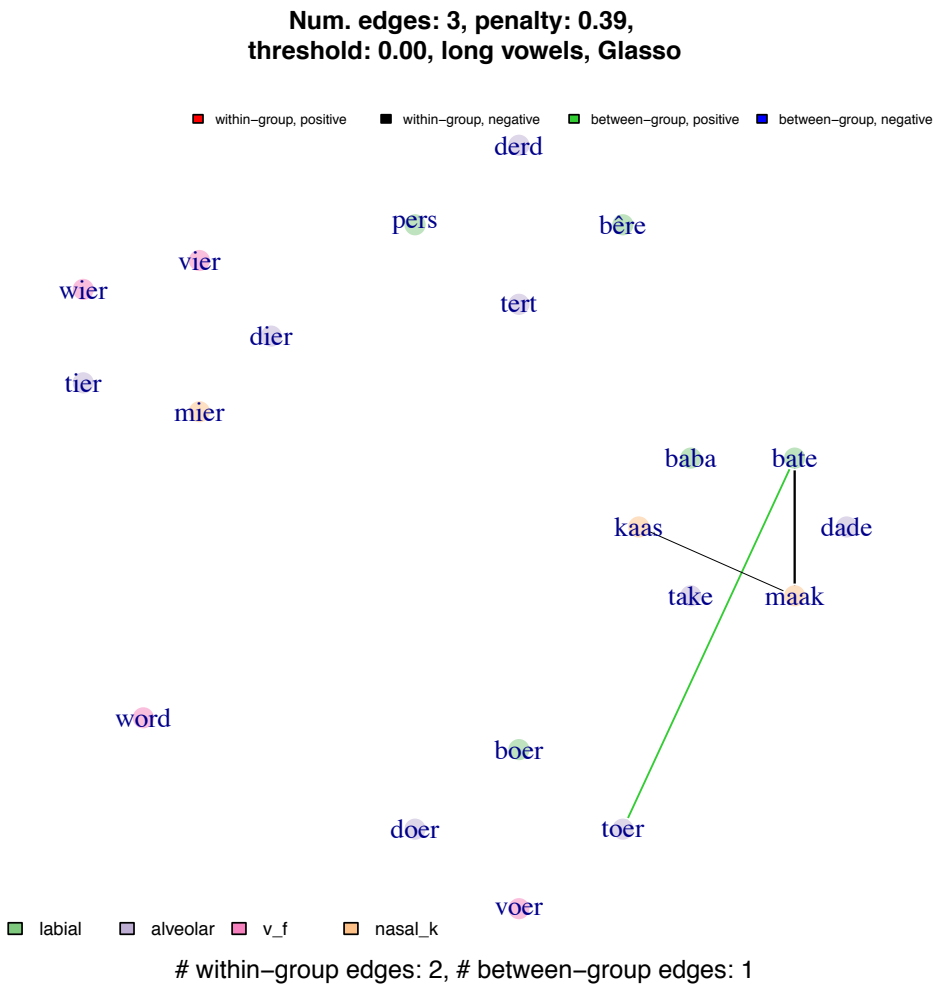

Figure 4.9: Inverse correlation edge graph, estimated by Glasso, for words with long vowels. The words are organized by vowel, with each circle of words sharing a common vowel ("word" is the only word with a long "o" vowel; in Afrikaans, it means "become").

Figure 4.10 displays a bar chart of the fraction of edges present among each pair of long vowels. The edges are estimated using a sequence of penalty parameters for Glasso and nodewise regression. Note that when the penalty is zero, the Glasso estimate reduces to the inverse sample correlation, which is a fully dense matrix, so the fraction of edges is equal to one. Figure 4.12 is the analogous display for short vowels. For long vowels, at higher penalty ($\lambda = 0.3$), the fraction of within-vowel edges is larger than the fraction of between-vowel edges.

Among the long vowels, as we increase the penalty, the fraction of edges decreases more rapidly for some vowel pairs than for others. For word pairs that have larger Pearson correlation but smaller penalized inverse correlation, the words are marginally correlated, but not conditionally correlated given the other words; that is, the relationship between those words is explained by other words. As seen in Figure 4.10, the long vowel pairs "a"-"a" and "a"-"u" persist to a penalty of 0.4. For short vowels, by contrast, as seen in Figure 4.12, the edges appear to be uniformly distributed among vowel pairs.

For each pair of long vowels, Figure 4.11 displays the average absolute values of the Pearson correlation entries among edges. Note that the edges are obtained via the precision matrix, but the average is taken using entries of the sample correlation matrix. For example, let $E(A, A)$ denote the set of edges between words with a long "a" vowel, and let $|E(A, A)|$ denote the number of edges between words with long "a" vowels. Then we calculate

$$\frac{1}{|E(A,A)|} \sum_{(i,j) \in E(A,A)} |S_{ij}|. \tag{4.11}$$

Figure 4.13 displays the analogous plot for long vowels.

Note that as the penalty increases, the number of edges decreases, so the average Pearson correlation is taken over the stronger entries that remain. At the highest

Word One	Word Two	Pearson Correlation
bate	maak	0.50
kaas	maak	0.41
baba	tier	0.37
bate	maak	0.50
bate	toer	−0.48
boer	kaas	0.28
boer	mier	0.36
bêre	tert	0.35
bêre	vier	0.33
bate	kaas	0.22
dade	maak	0.32
derd	wier	0.27
dier	kaas	0.35
doer	voer	0.26
doer	word	0.36
kaas	tert	0.28
tert	vier	0.30
wier	tier	0.35

Table 4.3: Word-word Pearson correlations.

penalty shown, three edges remain: bate-maak, maak-kaas, and bate-toer. Pearson correlations between word pairs with strong edges are shown in Table 4.3.

Figure 4.14 displays the trial residual pitch curves for maak and kaas. For multiple speakers, the variability increases towards the end of the word, flaring out over time. The Pearson correlation between two words is high if corresponding utterances within speakers predominantly have the same sign (e.g. if the first utterance of maak is positive for the same time points as the first utterance of kaas, the second utterance of maak is positive for the same time points as the second utterance of kaas, etc., and if this pattern holds across speakers). Analgously, Figure 4.15 shows the trial residual pitch curves for bate and maak. Figure 4.16 shows the trial residual pitch curves for bate and toer.

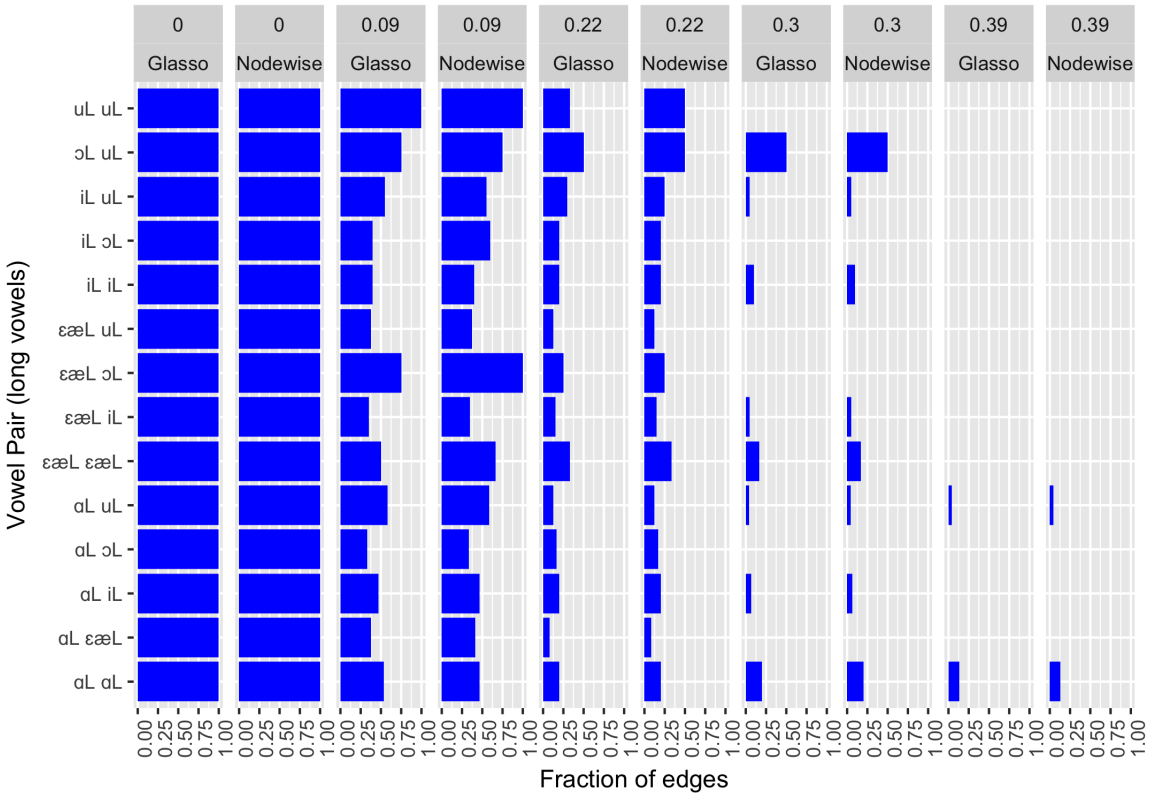

Figure 4.10: Bar chart of fraction of edges for long vowels, estimated using Glasso and nodewise regression. For certain penalty parameters, the cross-links between some pairs of long vowels disappear. For example, the *ɛæ-o* vowel pairs have many edges at smaller penalty parameters, but no edges at a penalty of 0.3.

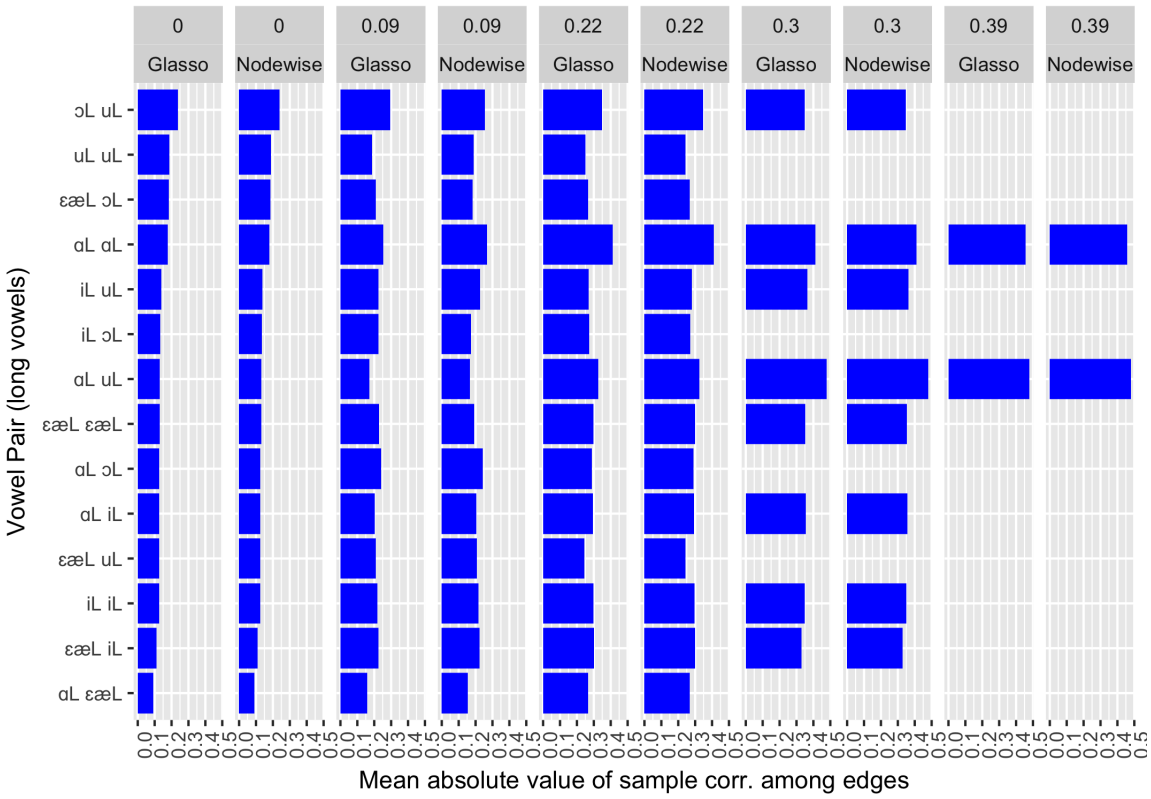

Figure 4.11: Bar chart of average sample correlation among edges for long vowels, estimated using Glasso and nodewise regression.

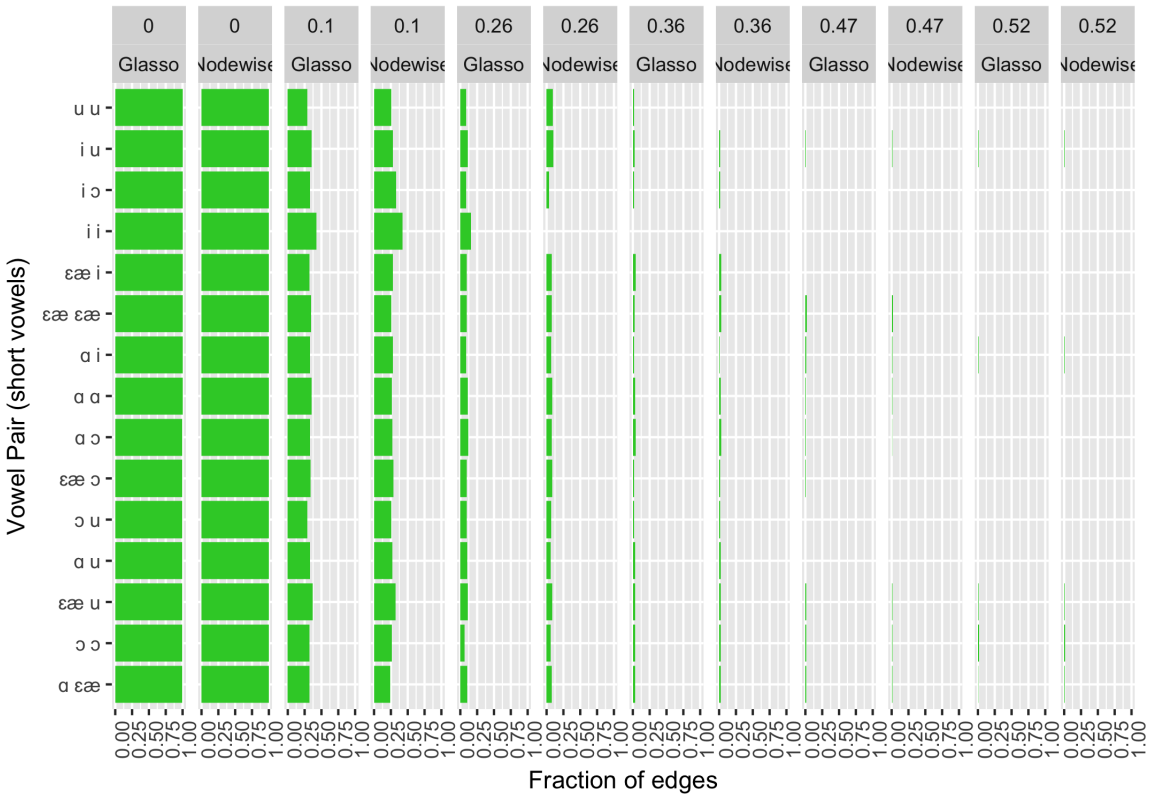

Figure 4.12: Bar chart of fraction of edges for short vowels, estimated using Glasso and nodewise regression.

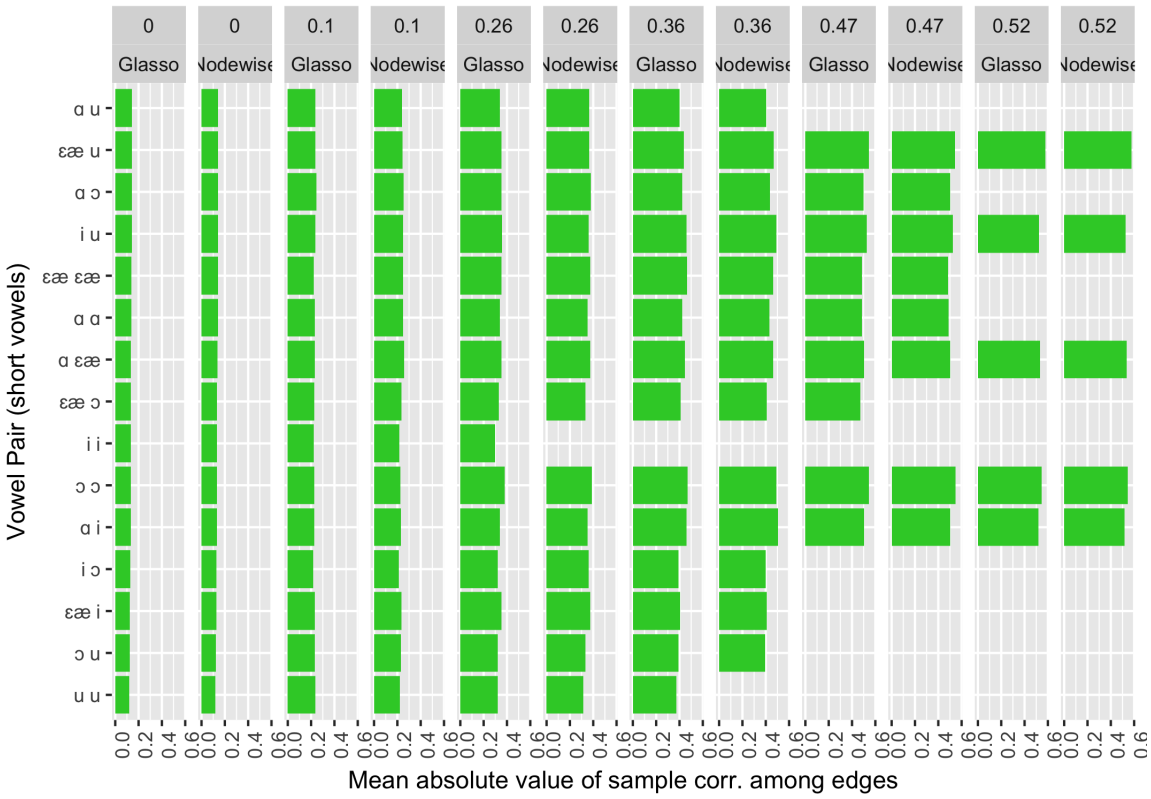

Figure 4.13: Bar chart of average sample correlation among edges for short vowels, estimated using Glasso and nodewise regression.

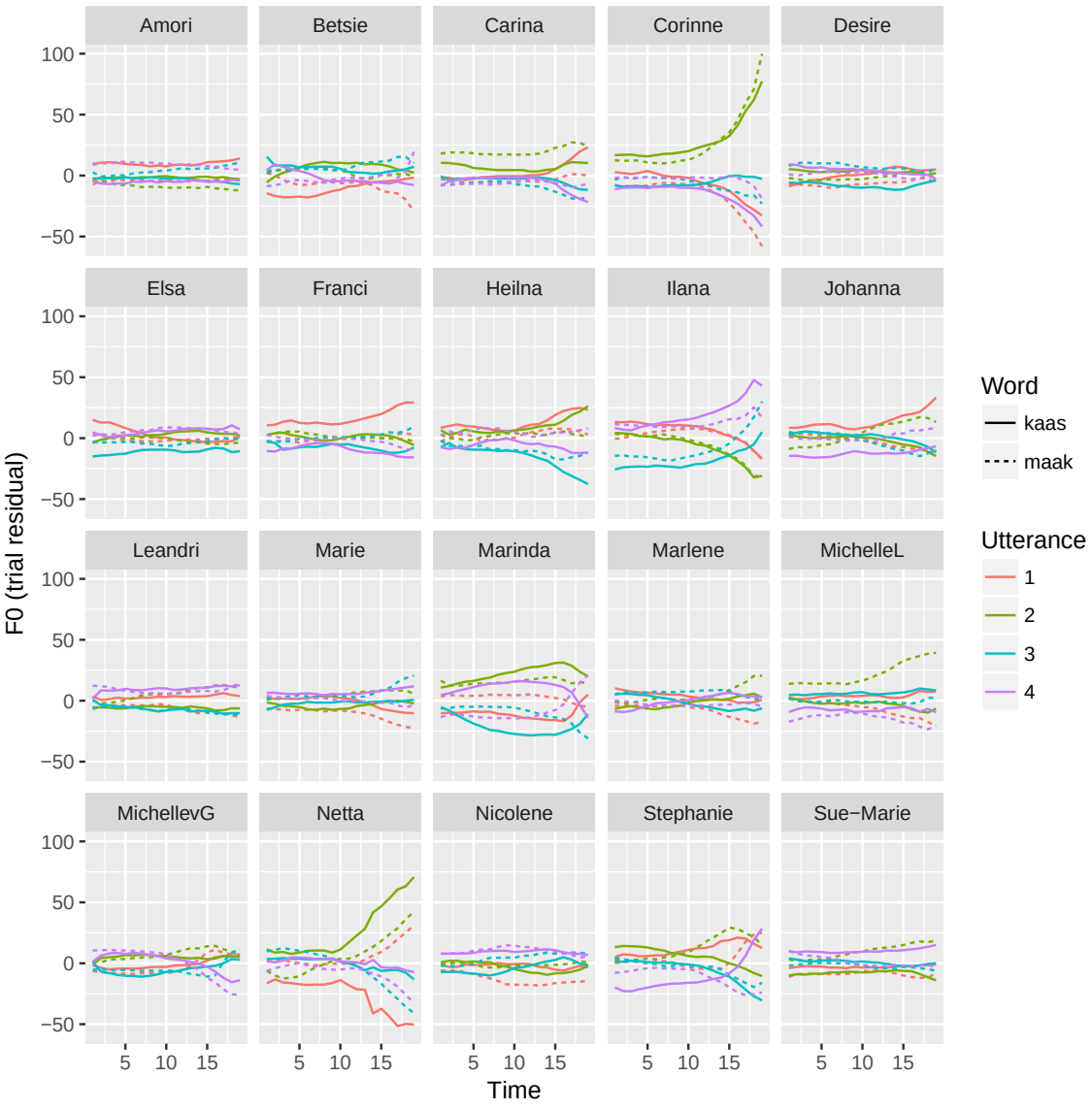

Figure 4.14: Trial residual pitch curves for the words maak and kaas.

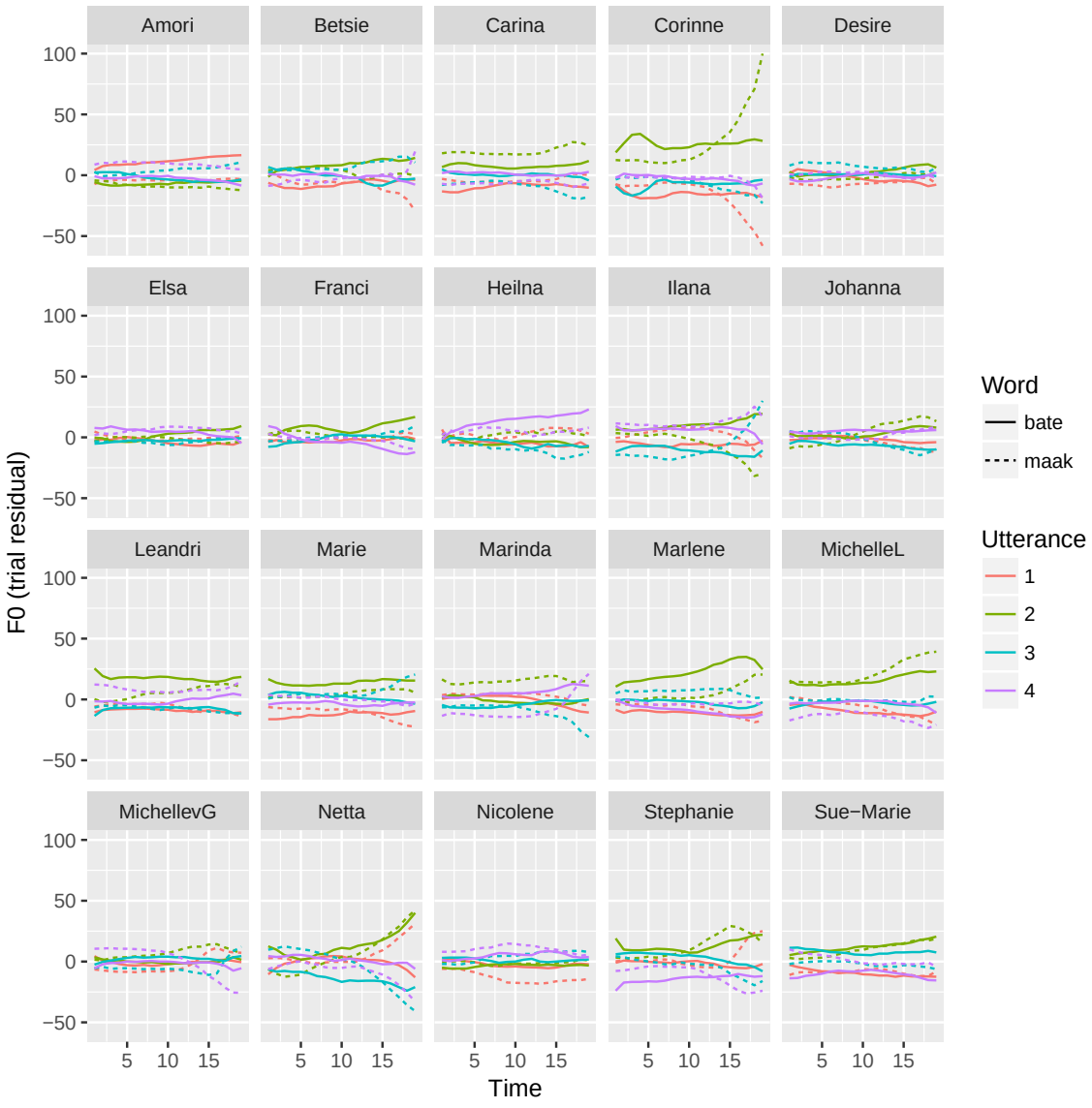

Figure 4.15: Trial residual pitch curves for the words bate and maak.

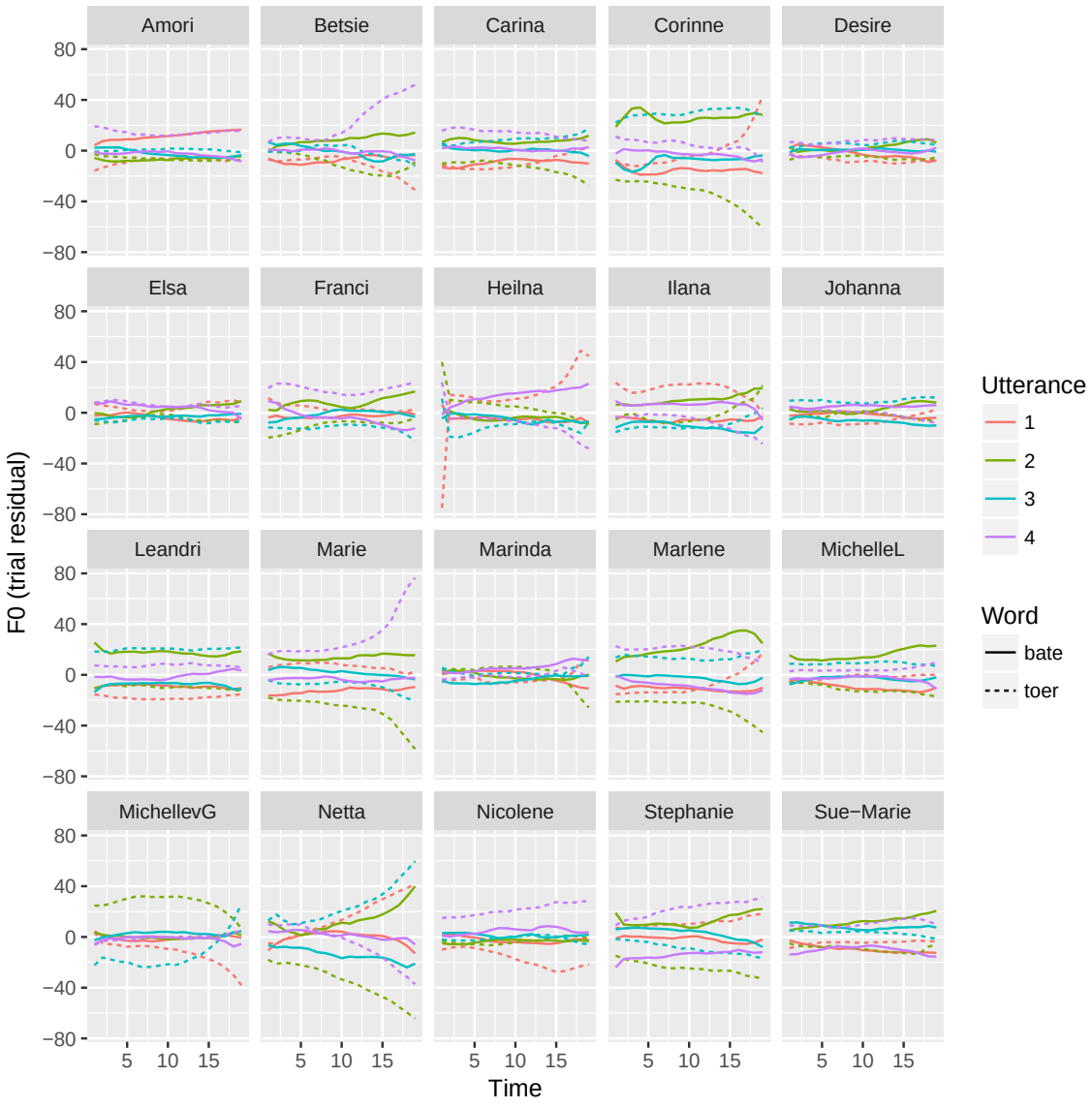

Figure 4.16: Trial residual pitch curves for the words bate and toer.

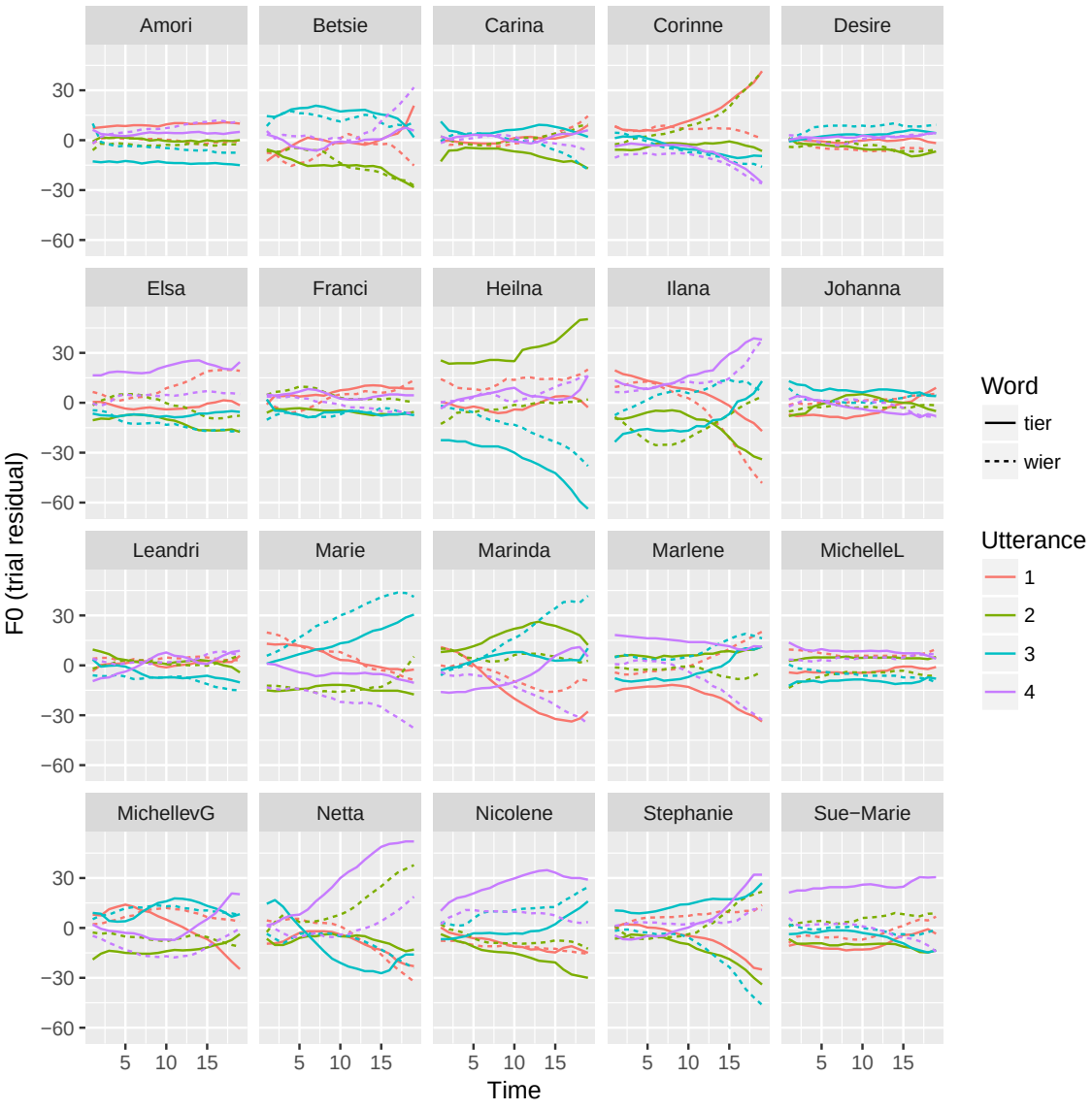

Figure 4.17: Trial residual pitch curves for the words wier and tier.

4.4 Visualization of edges

In Figure 4.18, we display the inverse correlation graph for all words, organized by initial consonant, with the top row of circles corresponding to voiced consonants, and the bottom row corresponding to voiceless consonants. Almost all of the edges are between group rather than within group; that is, almost all edges are between words starting with different consonants. Table 4.4 displays the Pearson correlations for word pairs that have edges in Figure 4.18.

In Figure 4.19, we present a high-level summary of this edge graph, by aggregating words with the same consonant into "supernodes." Two supernodes are connected if there is an edge in Figure 4.18 between two words with the corresponding consonants, estimated by both Glasso and nodewise regression. This diagram holds for a particular choice of penalty and threshold. We show that similar patterns hold if we perturb the penalty, and also if we use nodewise regression instead of Glasso. In Figure 4.20, we display a an edge graph analogous to Figure 4.18, but with a smaller penalty (0.32). In Figure 4.21 we display the Glasso edge graph for penalty 0.32 with threshold 0.1. In Figures 4.22, 4.23, and 4.24, we display nodewise regression graphs for three choices of penalty parameter (0.32, 0.37, and 0.43), with threshold 0.08. The graphs illustrate that nodewise regression estimates a similar graph structure to Glasso.

In Figure 4.20 we compare the edges for Glasso and nodewise regression; both methods are run with a penalty of 0.32 and a threshold around 0.1 (0.1 for Glasso, 0.08 for nodewise regession). At a similar level of penalty and thresholding, the Glasso graph is denser than the nodewise graph. In Figure 4.18, we show an analogous graph, with a larger penalty of 0.37.

Word One	Word Two	Pearson Correlation
kop	tor	0.53
nog	wond	0.56
den	pen	0.49
baba	ken	−0.49
bate	maak	0.50
bate	tas	0.52
bate	toer	−0.48
berg	mier	0.48
bied	das	0.49
boet	kies	0.60
bot	pars	0.50
dare	baba	0.48
dare	tas	0.49
doer	pen	0.50
kat	met	−0.48
ken	tand	0.48
kerk	piek	0.45
koet	met	0.51
met	vier	0.52
met	wat	0.53
nek	was	0.51
nek	woed	0.58
padd	pond	0.46
term	vier	0.60

Table 4.4: Word-word pearson correlations for words with edges in Figure 4.18.

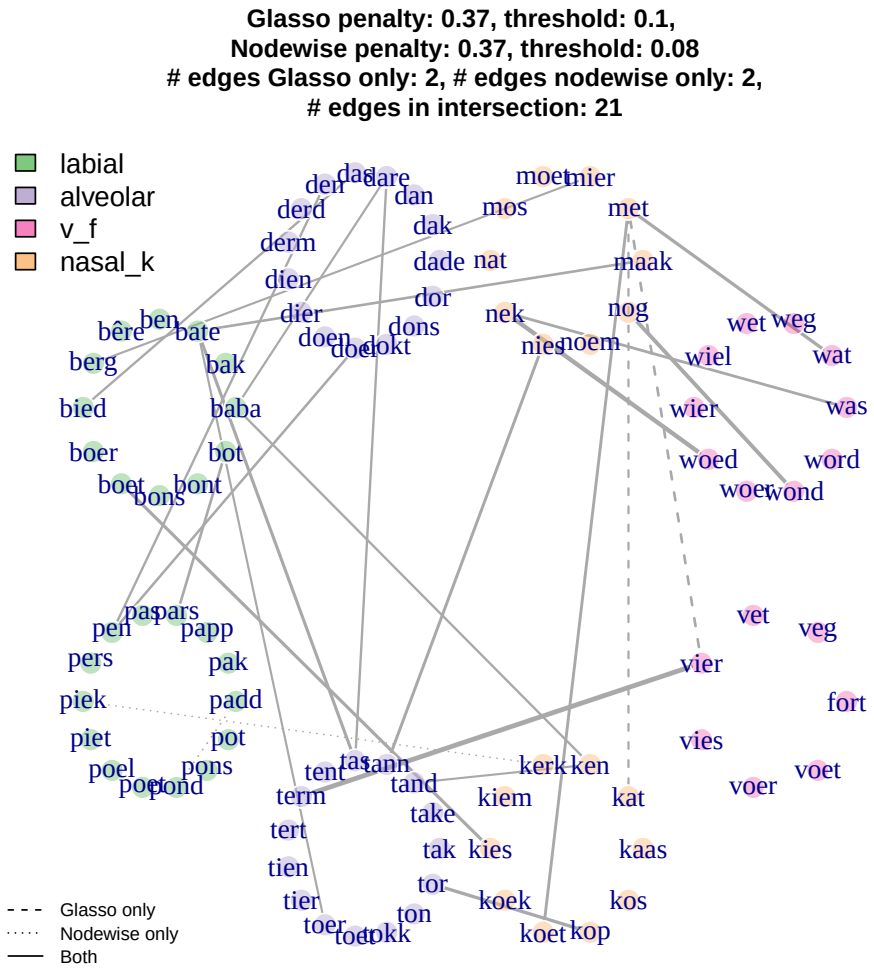

Figure 4.18: Inverse covariance graph of all words, comparing Glasso edges with nodewise regression edges. The Glasso penalty is 0.37, followed by a threshold of 0.1, and the nodewise regression penalty is 0.37, followed by a threshold of 0.08. The words are organized by initial consonant, with typically voiced consonants in the top row and typically voiceless consonants in the bottom row.

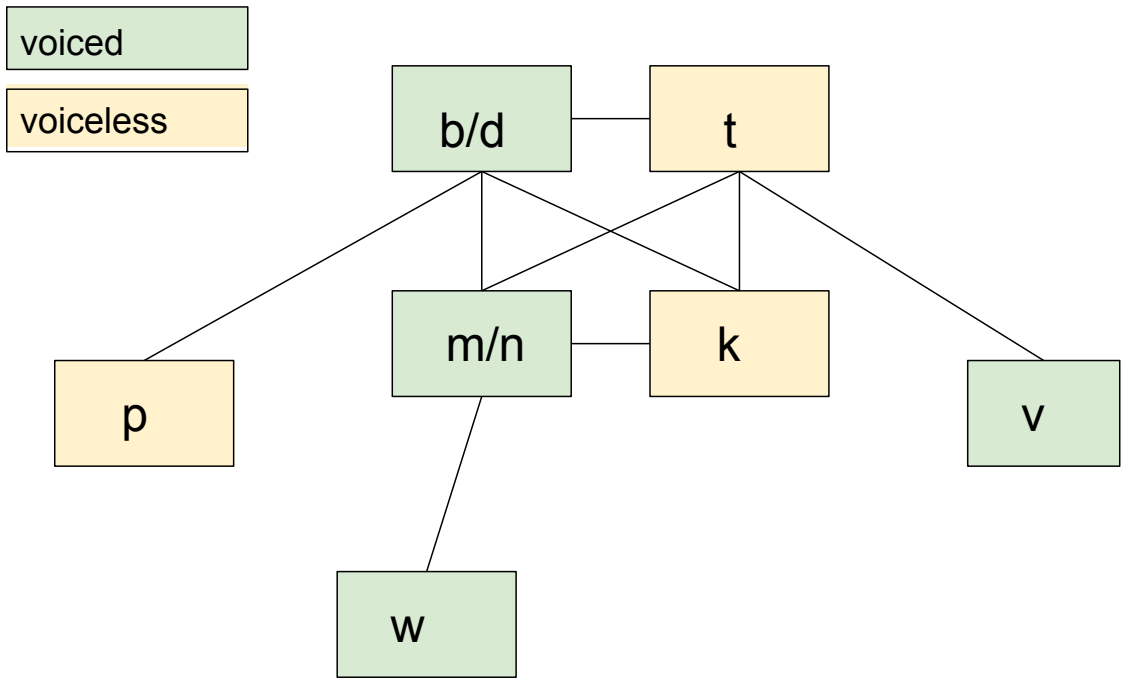

Figure 4.19: Diagram displaying connectivity among consonants, providing a higher-level representation of Figure 4.18 by combining nodes within a consonant type into "supernodes." Two nodes are connected in this diagram if there is an edge between words with the corresponding initial consonants in Figure 4.18.

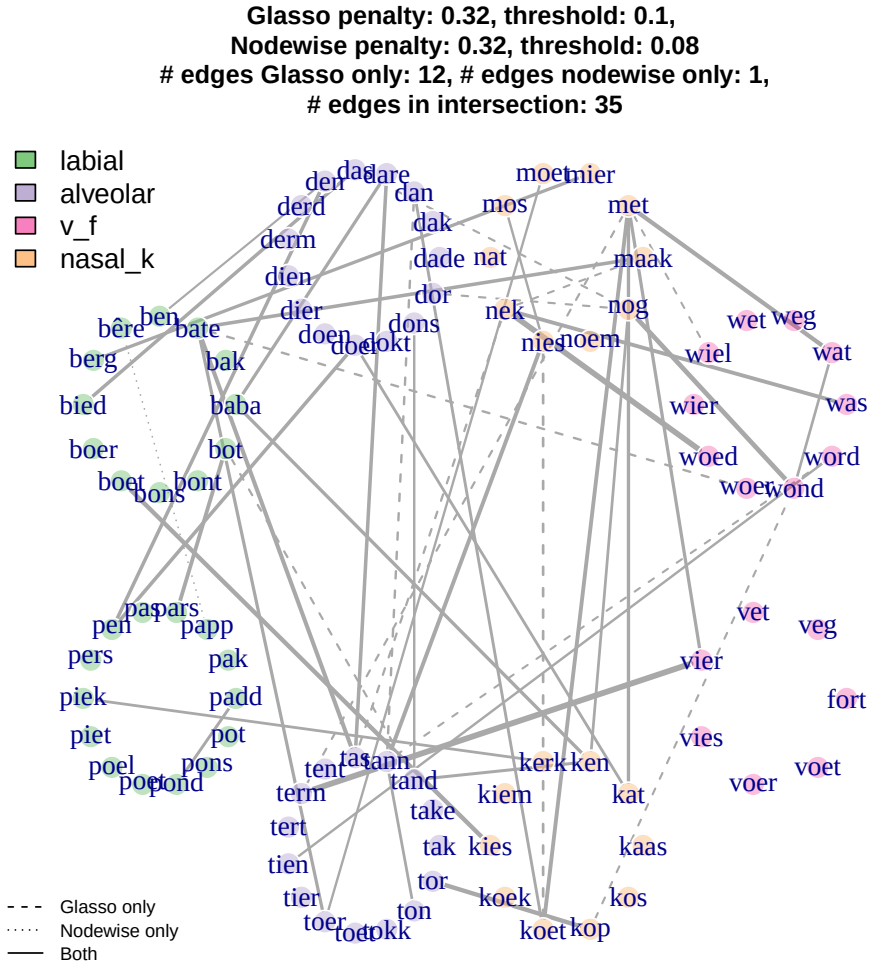

Figure 4.20: Inverse covariance graph of all words, comparing Glasso edges with nodewise regression edges. The Glasso penalty is 0.32, followed by a threshold of 0.1, and the nodewise regression penalty is 0.32, followed by a threshold of 0.08. The words are organized by initial consonant, with typically voiced consonants in the top row and typically voiceless consonants in the bottom row.

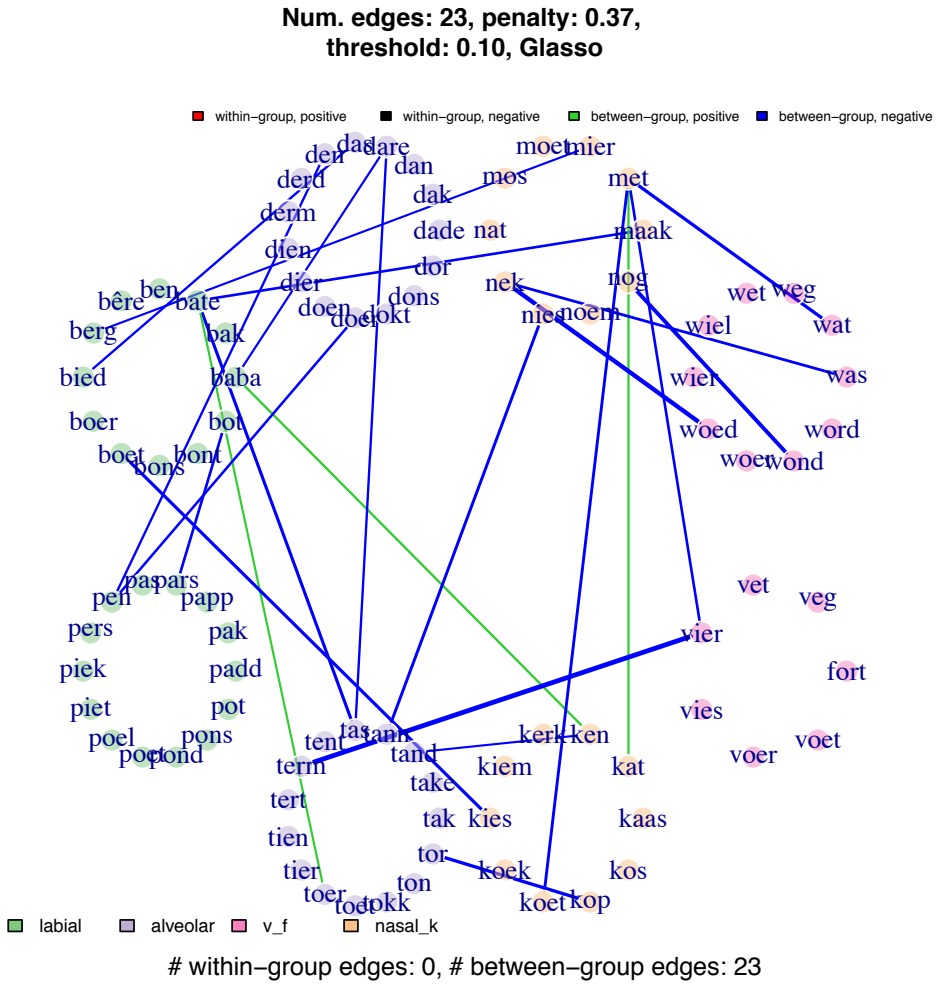

Figure 4.21: Inverse covariance graph of all words, estimated using Glasso, organized by initial consonant, with typically voiced consonants in the top row and typically voiceless consonants in the bottom row.

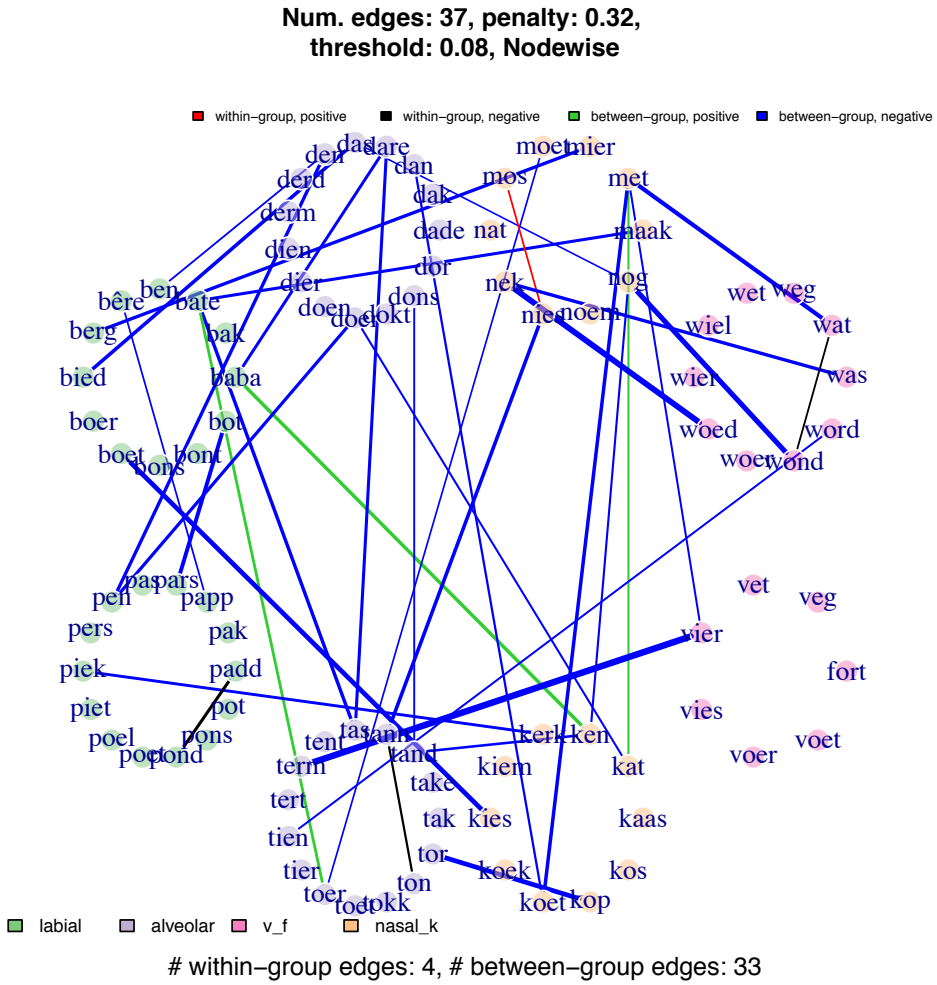

Figure 4.22: Inverse covariance graph of all words, estimated using nodewise regression, organized by initial consonant, with typically voiced consonants in the top row and typically voiceless consonants in the bottom row.

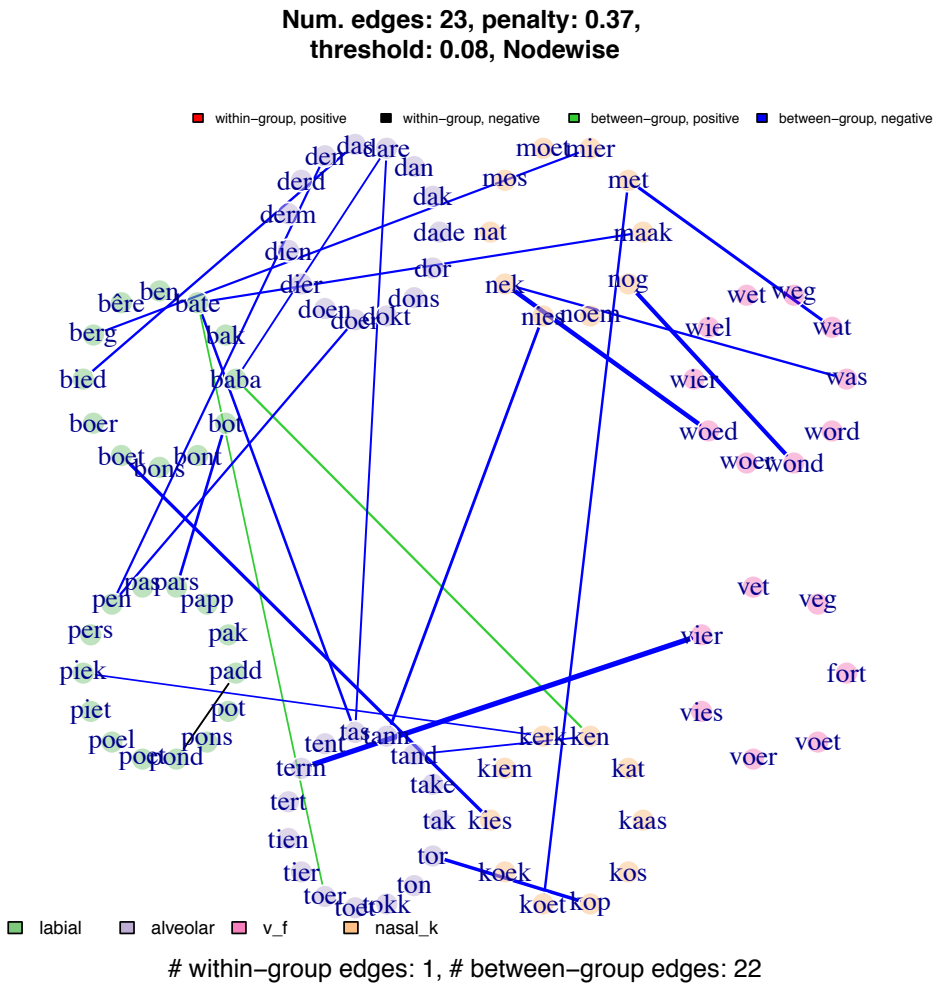

Figure 4.23: Inverse covariance graph of all words, estimated using nodewise regression, organized by initial consonant, with typically voiced consonants in the top row and typically voiceless consonants in the bottom row.

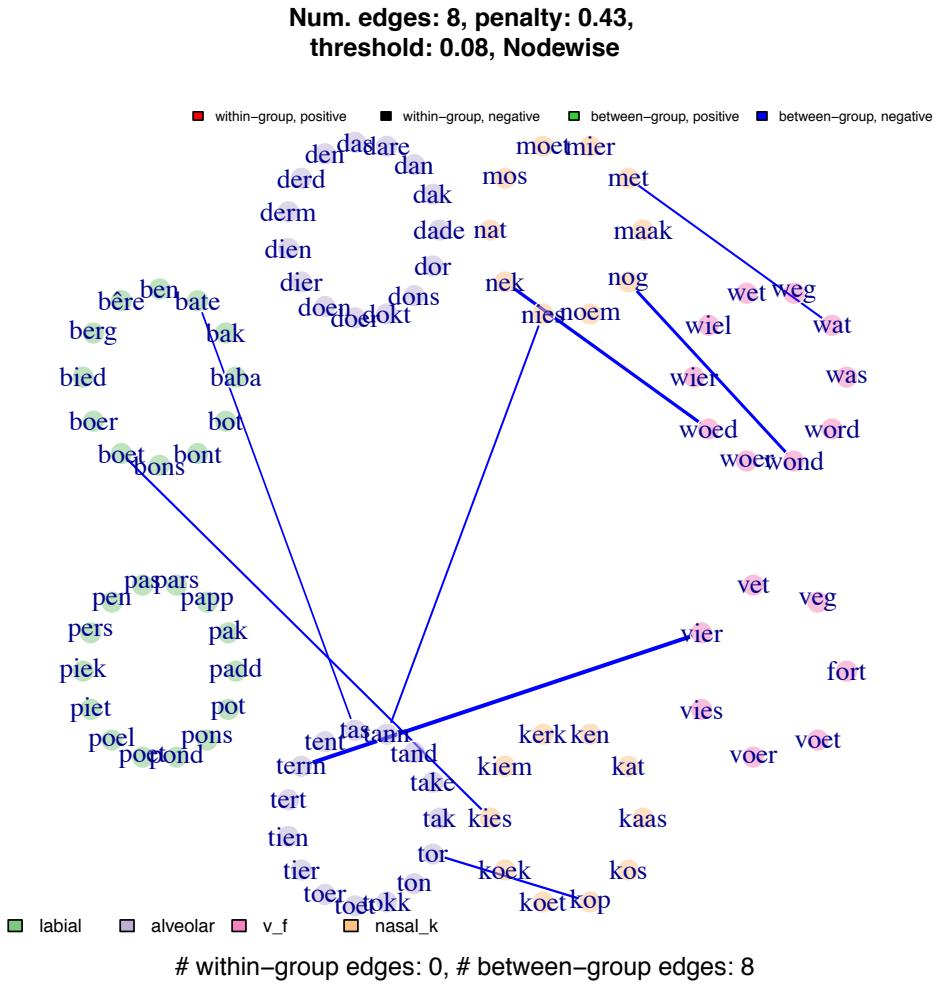

Figure 4.24: Inverse covariance graph of all words, estimated using nodewise regression, organized by initial consonant, with typically voiced consonants in the top row and typically voiceless consonants in the bottom row.

4.4.1 Labial and alveolar words

In Figures 4.25, 4.26, and 4.27 we show the inverse covariance graph estimated using a sequence of Glasso penalty parameters, with a threshold of 0.1. For small penalty values, words of all four initial consonants (b, d, p, t) are densely connected. As the penalty increases the edges between words beginning with p and t drop off.

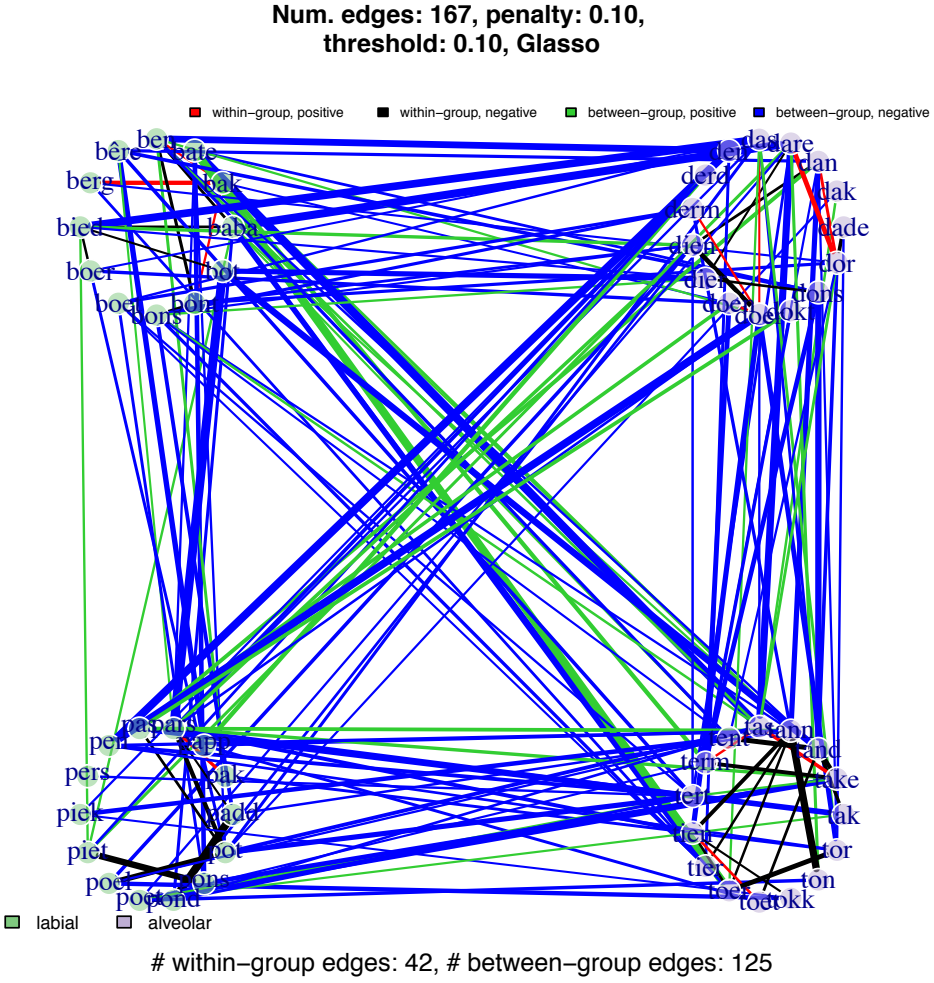

Figure 4.25: Inverse covariance graph of labial and alveolar words Glasso with a peanlty of 0.1 and a threshold of 0.1.

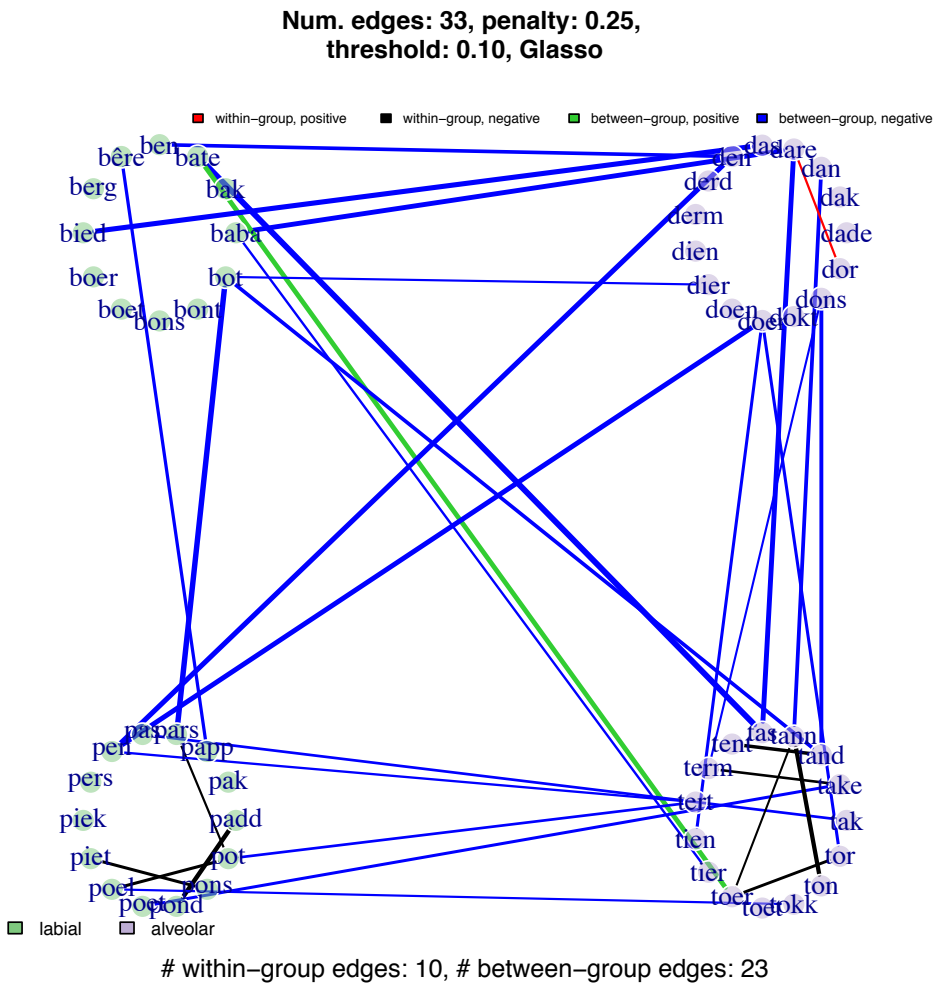

Figure 4.26: Inverse covariance graph of labial and alveolar words Glasso with a peanlty of 0.25 and a threshold of 0.1.

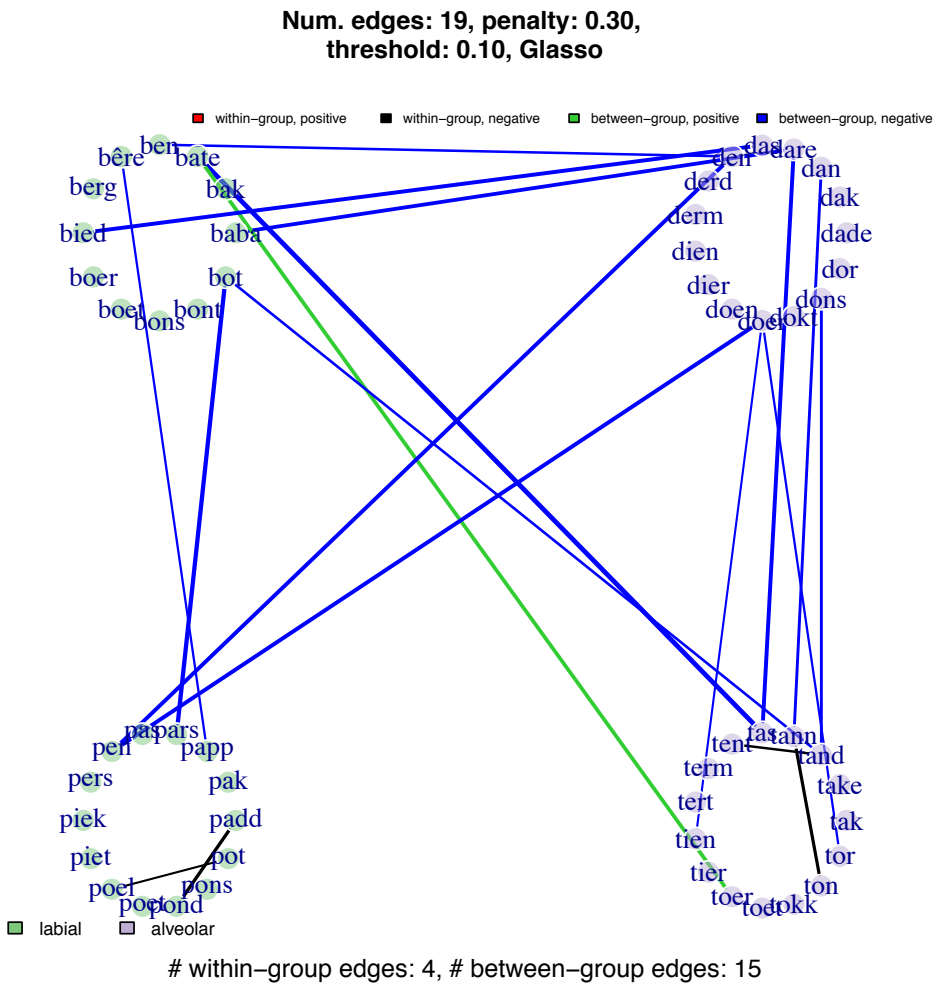

Figure 4.27: Inverse covariance graph of labial and alveolar words Glasso with a peanlty of 0.3 and a threshold of 0.1.

4.4.2 Initial consonant connectivities

Figure 4.28 displays a bar chart of the fraction of edges between each pair of initial consonants, for a sequence of Glasso penalty parameters. When counting edges, the "m" and "n" are treated as a single consonant, as are the consonants "v" and "f". We see that even at a penatly of 0.43, edges persist between "mn" words and "w" words.

Figure 4.29 displays the mean Pearson correlation among edges in the for each consonant pair.

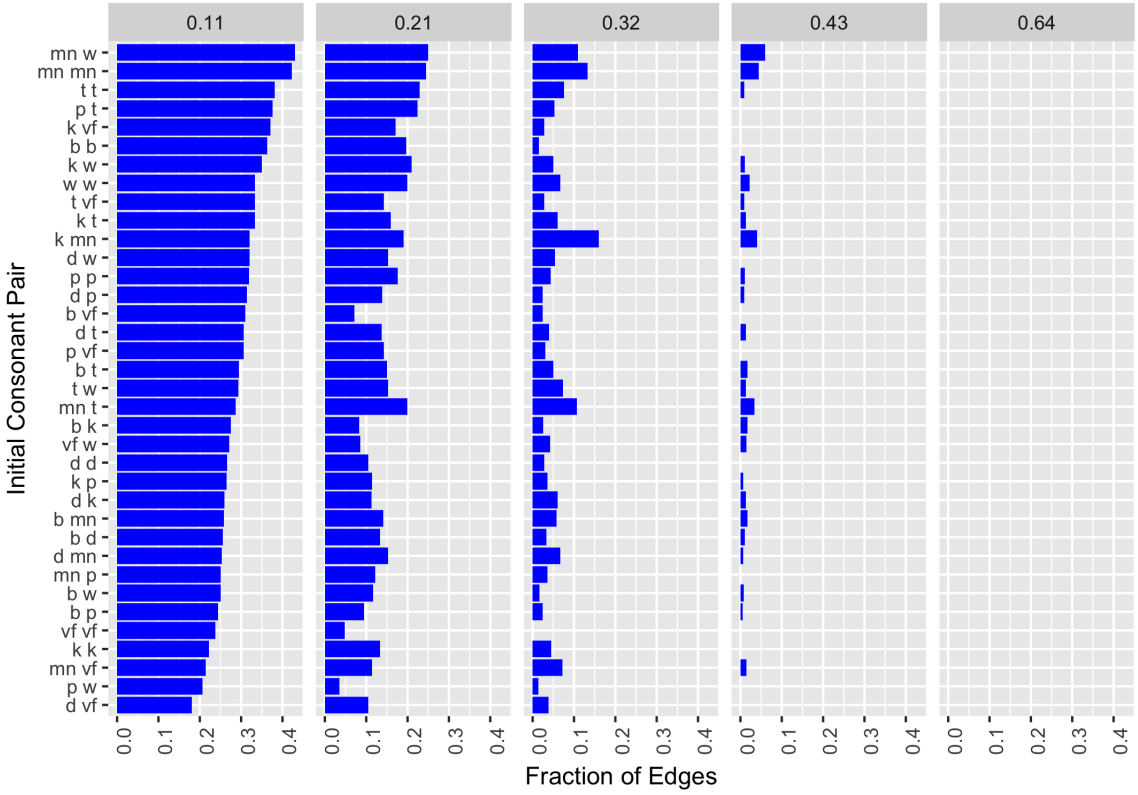

Figure 4.28: Fraction of edges between each pair of initial consonants as we vary the Glasso penalty.

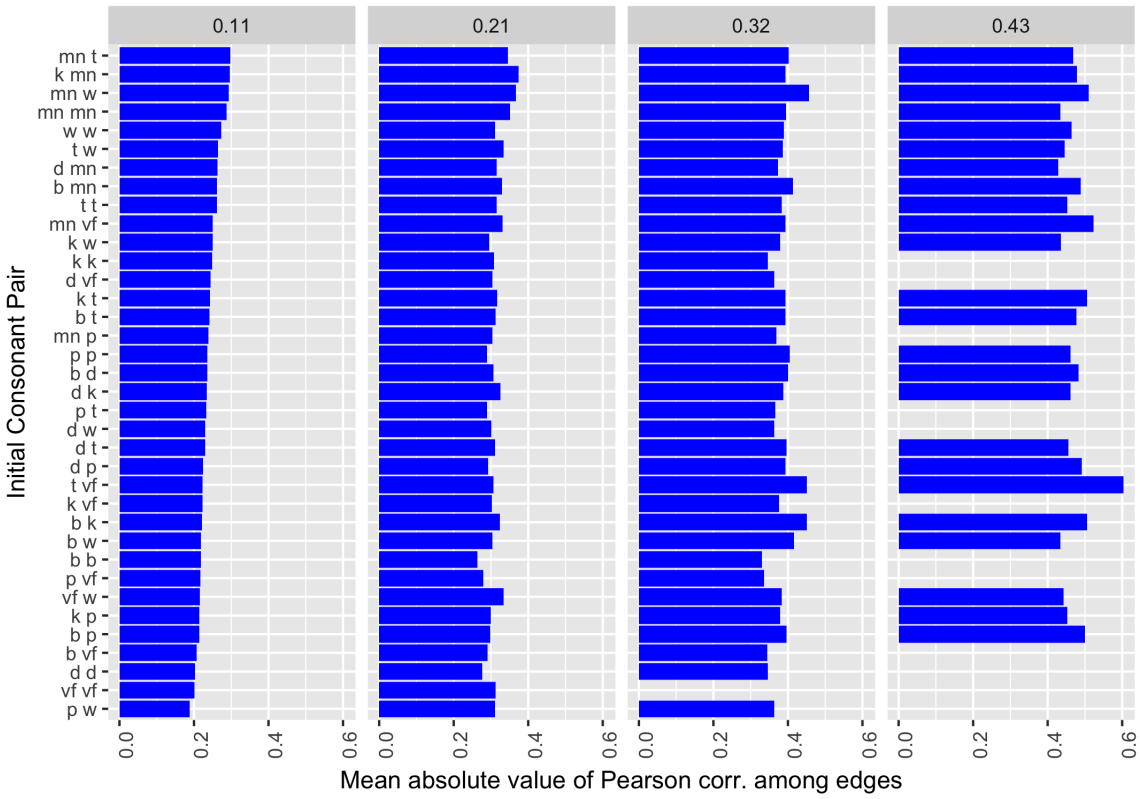

Figure 4.29: Mean absolute value of Pearson correlation among edges between each pair of initial consonants.

4.4.3 Comparing Glasso and nodewise regression graphs for pairs of word groups

We display inverse correlation graphs between each pair of word groups (labial, alveolar, nasal, and vf). Glasso and nodewise regression were run on all the words; in the following figures, we visualize subgraphs of the full graph. The line type indicates whether the edge appears in both the Glasso and nodewise regression graphs or in just one of the two. Both methods are run with a penalty of 0.32 and threshold of 0.16. We see that the edges are similar between the methods, but with more edges for Glasso than nodewise regression. In Section A.0.3 of the Appendix, we display analogous plots with a penalty of 0.26 and threshold 0.08.

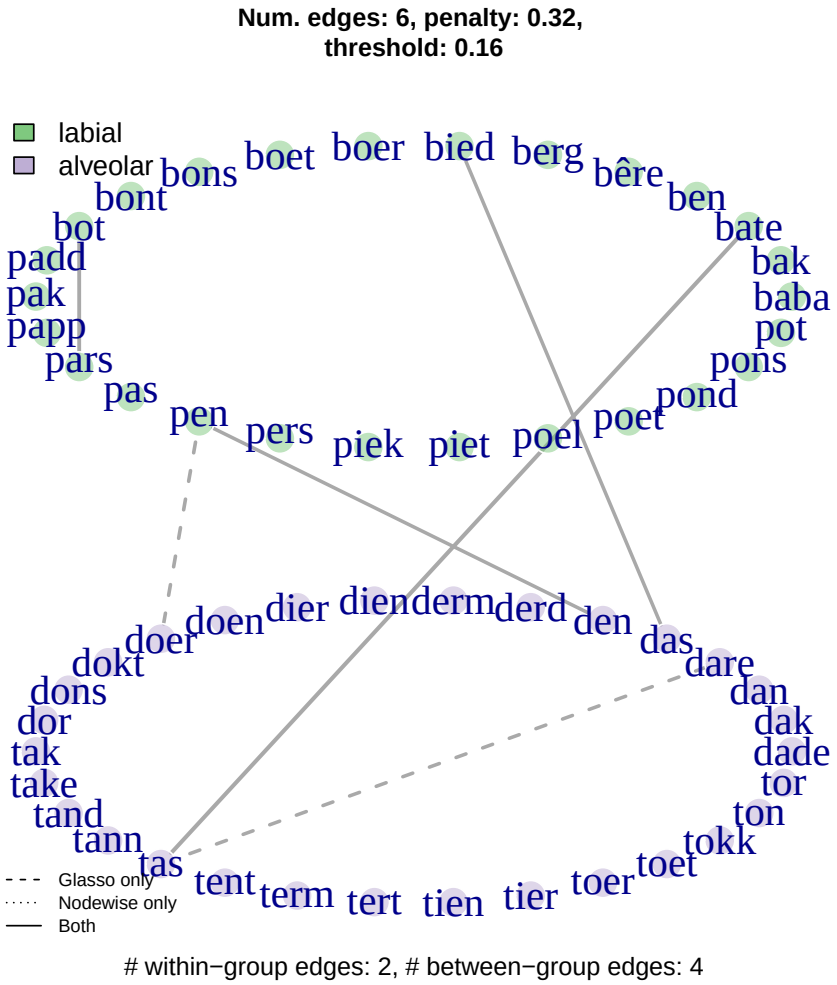

Figure 4.30: Inverse covariance graph of labial and alveolar words. This graph displays a subgraph of a graph for all 93 words, estimated using Glasso and nodewise regression with thresholding.

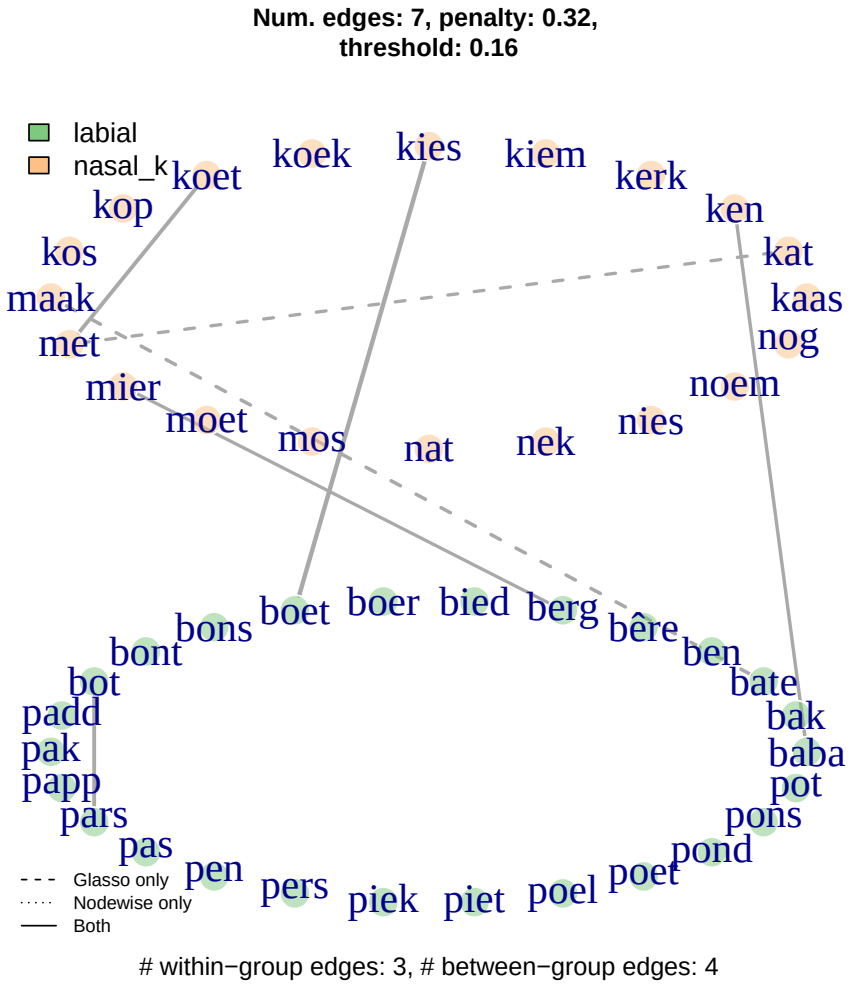

Figure 4.31: Inverse covariance graph of labial and nasal words. This graph displays a subgraph of a graph for all 93 words, estimated using Glasso and nodewise regression with thresholding.

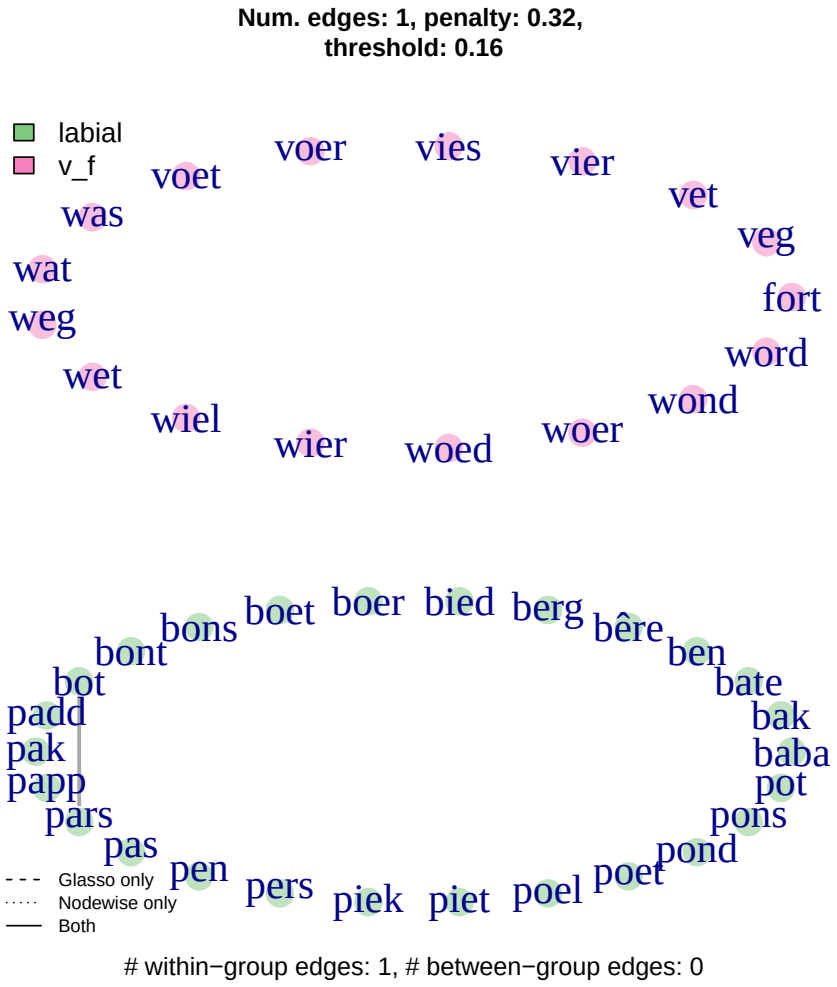

Figure 4.32: Inverse covariance graph of labial and vf words. This graph displays a subgraph of a graph for all 93 words, estimated using Glasso and nodewise regression with thresholding.

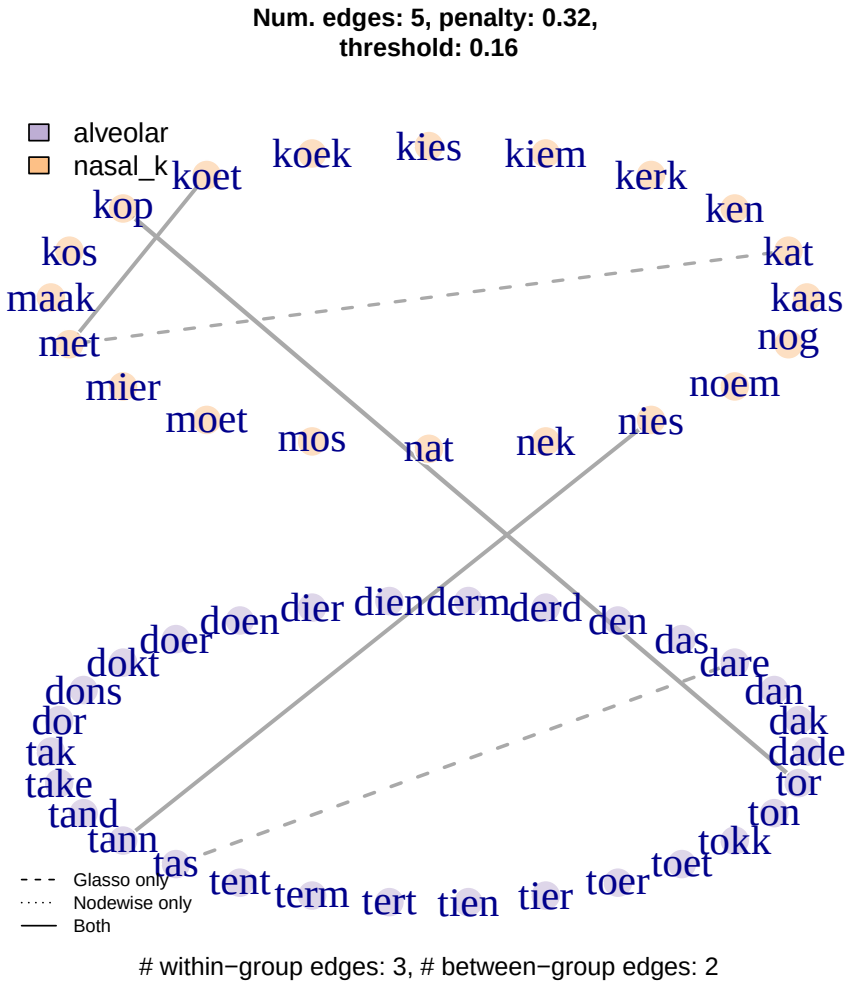

Figure 4.33: Inverse covariance graph of alveolar and nasal words. This graph displays a subgraph of a graph for all 93 words, estimated using Glasso and nodewise regression with thresholding.

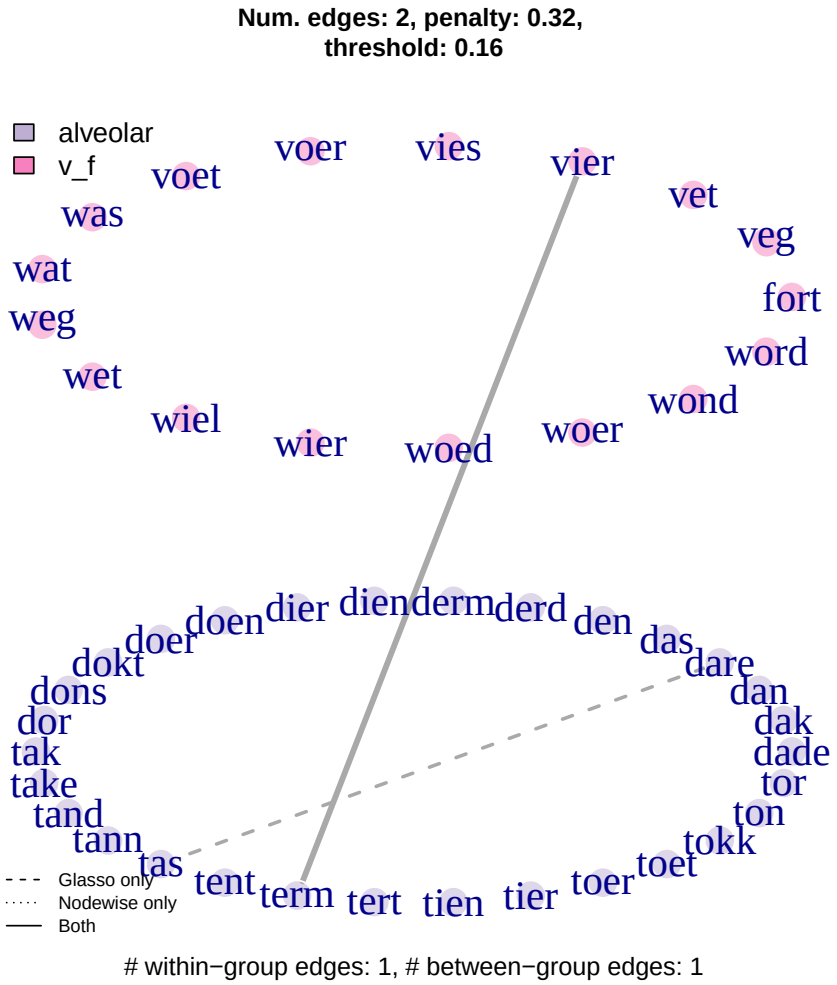

Figure 4.34: Inverse covariance graph of alveolar and vf words. This graph displays a subgraph of a graph for all 93 words, estimated using Glasso and nodewise regression with thresholding.

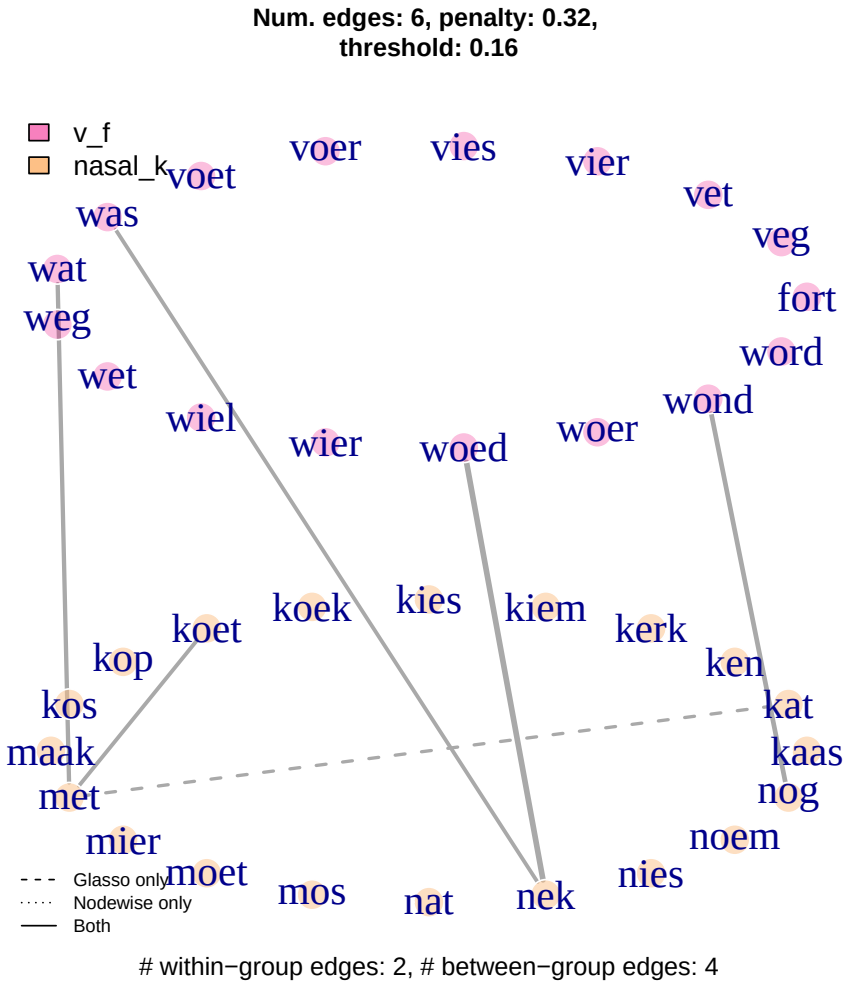

Figure 4.35: Inverse covariance graph of nasal and vf. This graph displays a subgraph of a graph for all 93 words, estimated using Glasso and nodewise regression with thresholding.

4.4.4 Comparison of time inverse covariance graphs for each pair of word groups

For each pair of word groups, we compare the time-time inverse correlation graphs, by taking intersections and set differences. We threshold each graph down to 70 edges. In each graph, nodes are connected to approximately five nearest neighbors on each side. The time-time edges are similar among the word groups; that is, most of the nodes are in the intersections of the graphs. This suggests that we can consider using a combined time-time inverse covariance matrix pooling over the words to decorrelate along the time axis, potentially improving the word-word covariance estimates, discussed in *Zhou* (2014a).

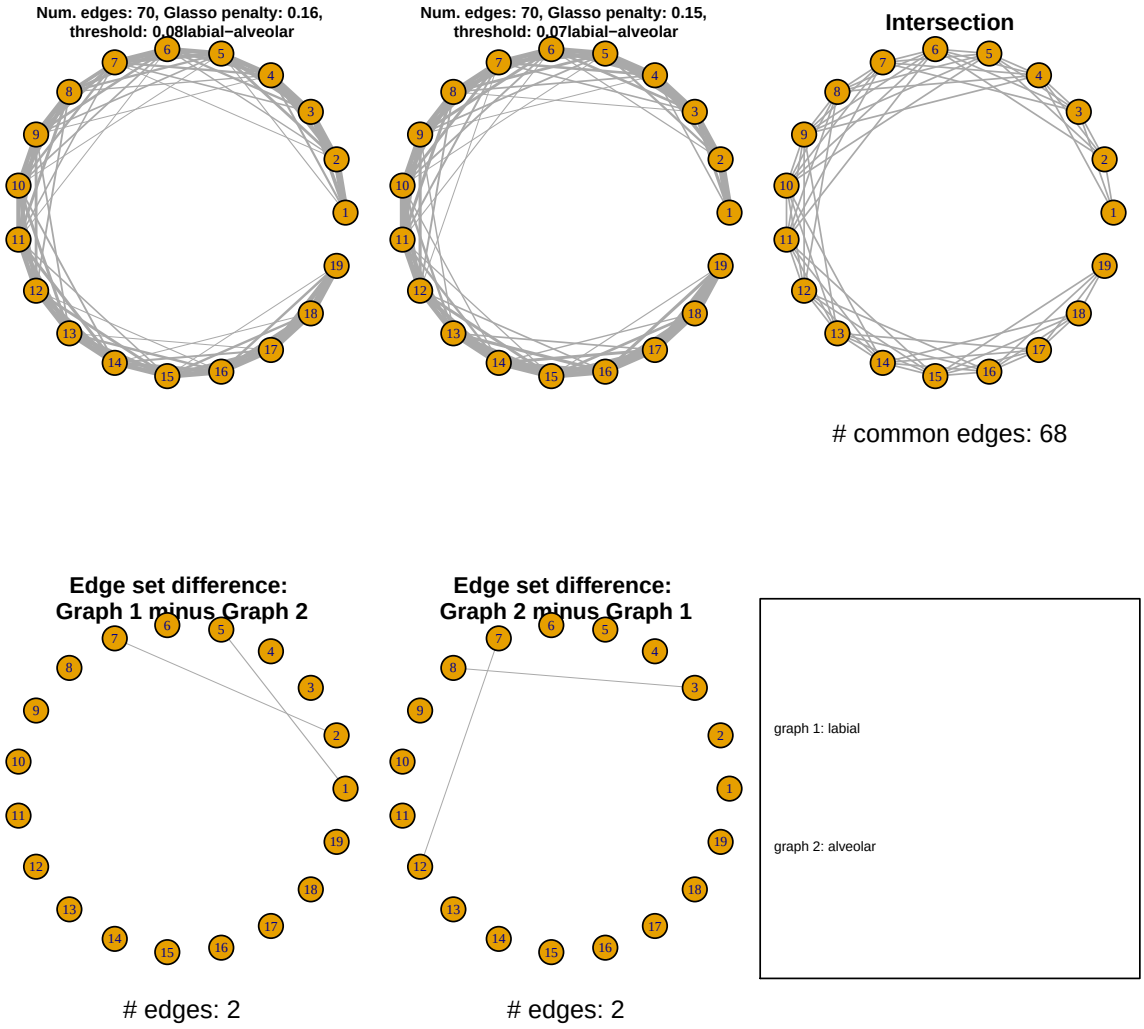

Figure 4.36: Time-time inverse covariance graphs for labial and alveolar words, as well as graph intersection and set differences. The inverse correlation matrices are thresholded so that 70 edges remain in each word group.

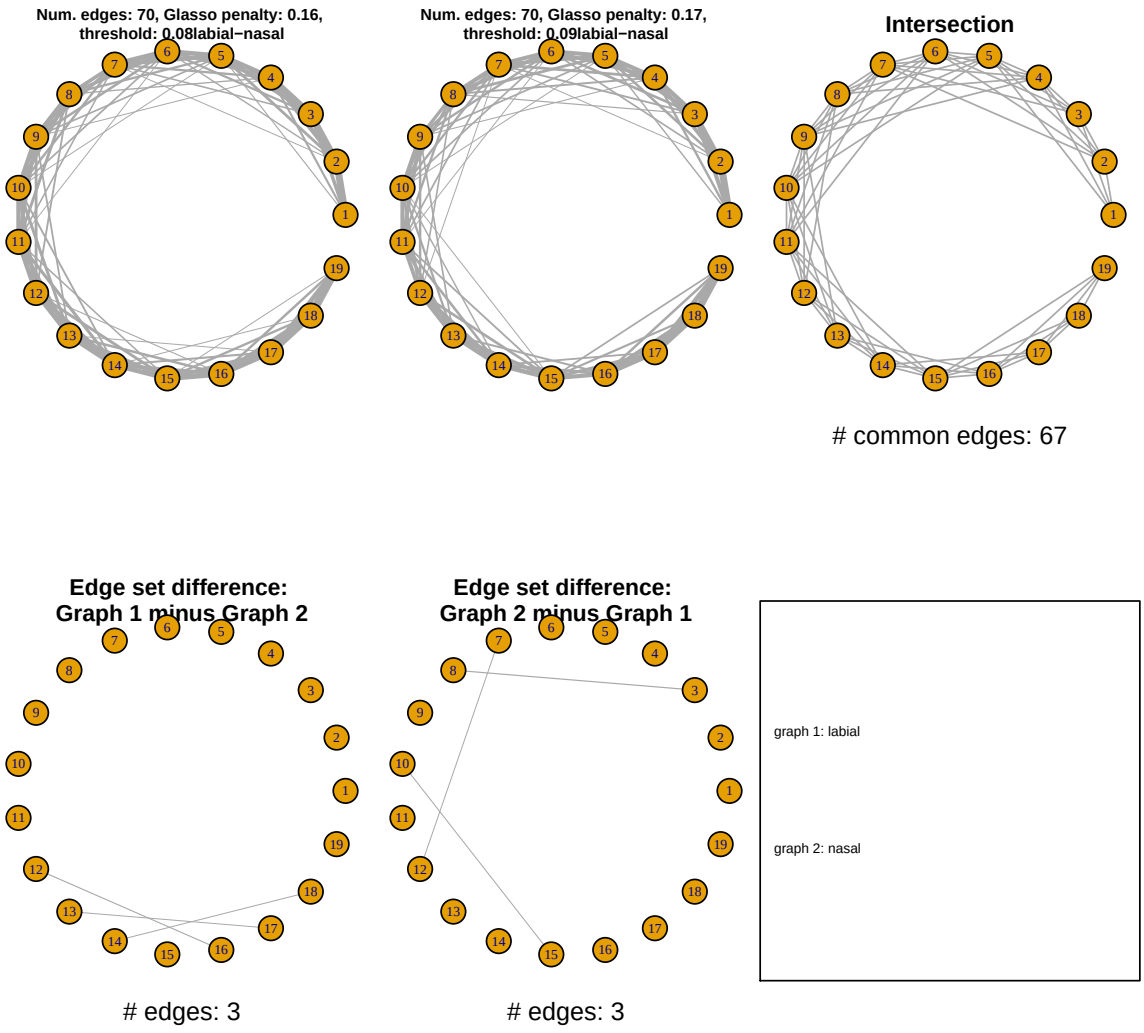

Figure 4.37: Time-time inverse covariance graphs for labial and nasal words, as well as graph intersection and set differences.

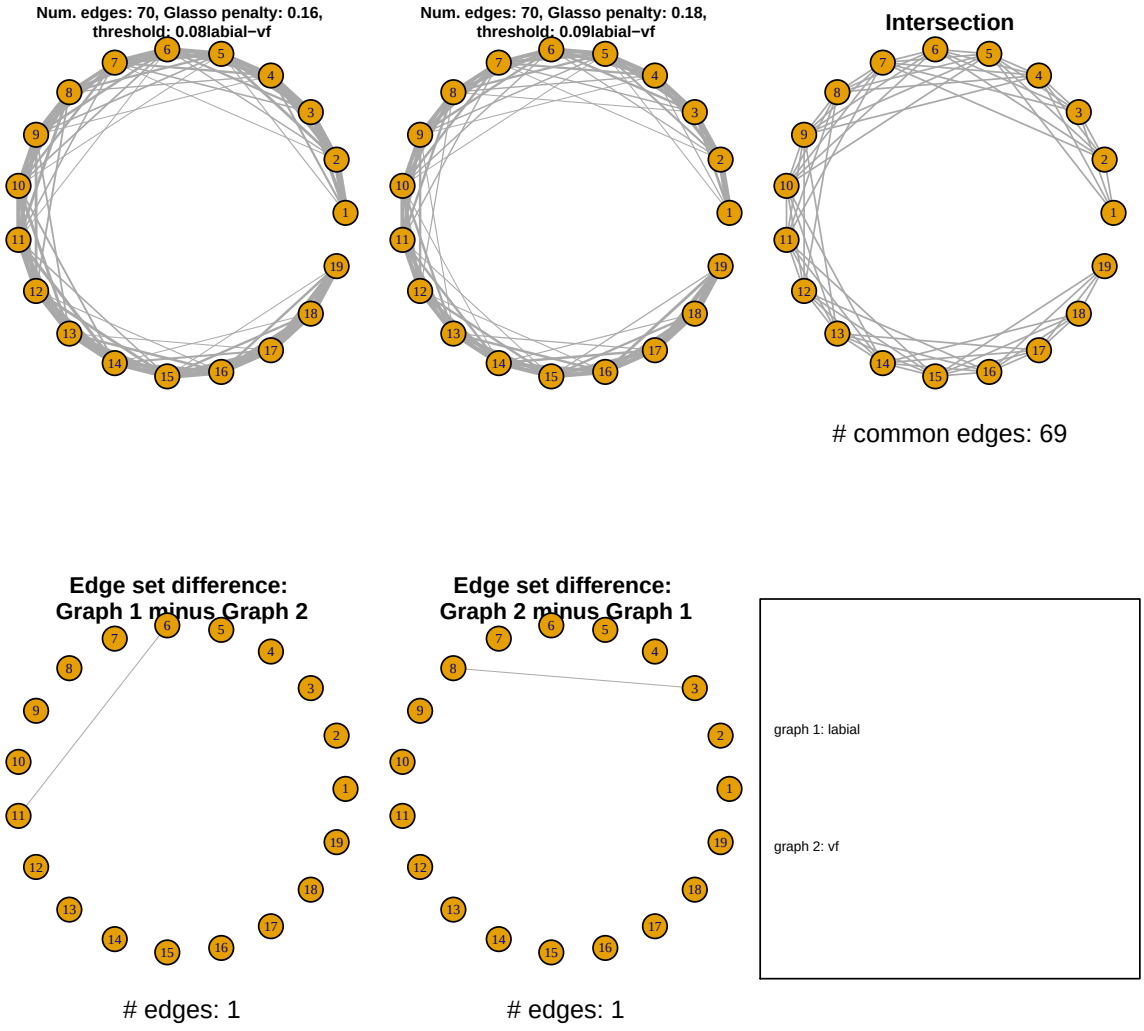

Figure 4.38: Time-time inverse covariance graphs for labial and vf words, as well as graph intersection and set differences.

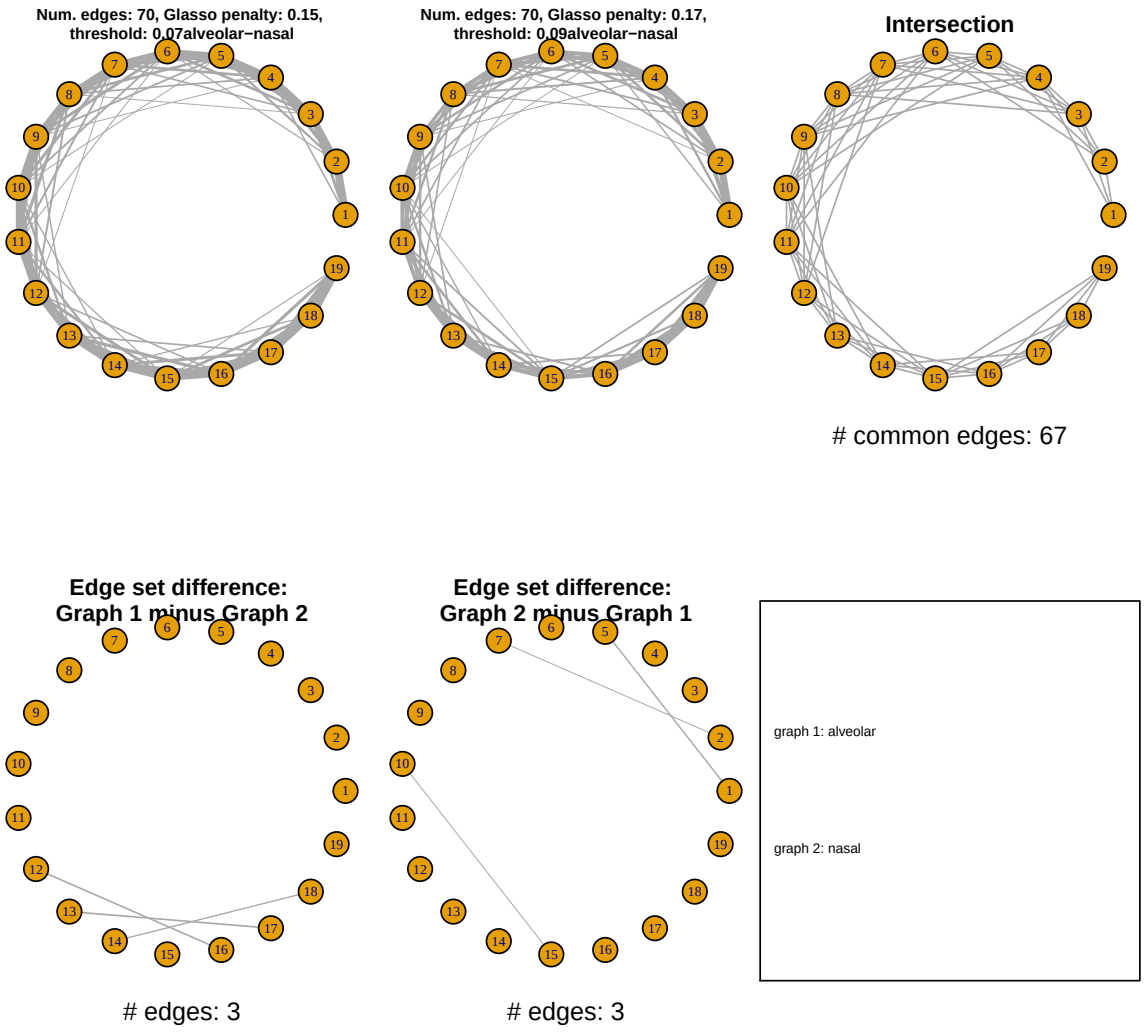

Figure 4.39: Time-time inverse covariance graphs for alveolar and nasal words, as well as graph intersection and set differences.

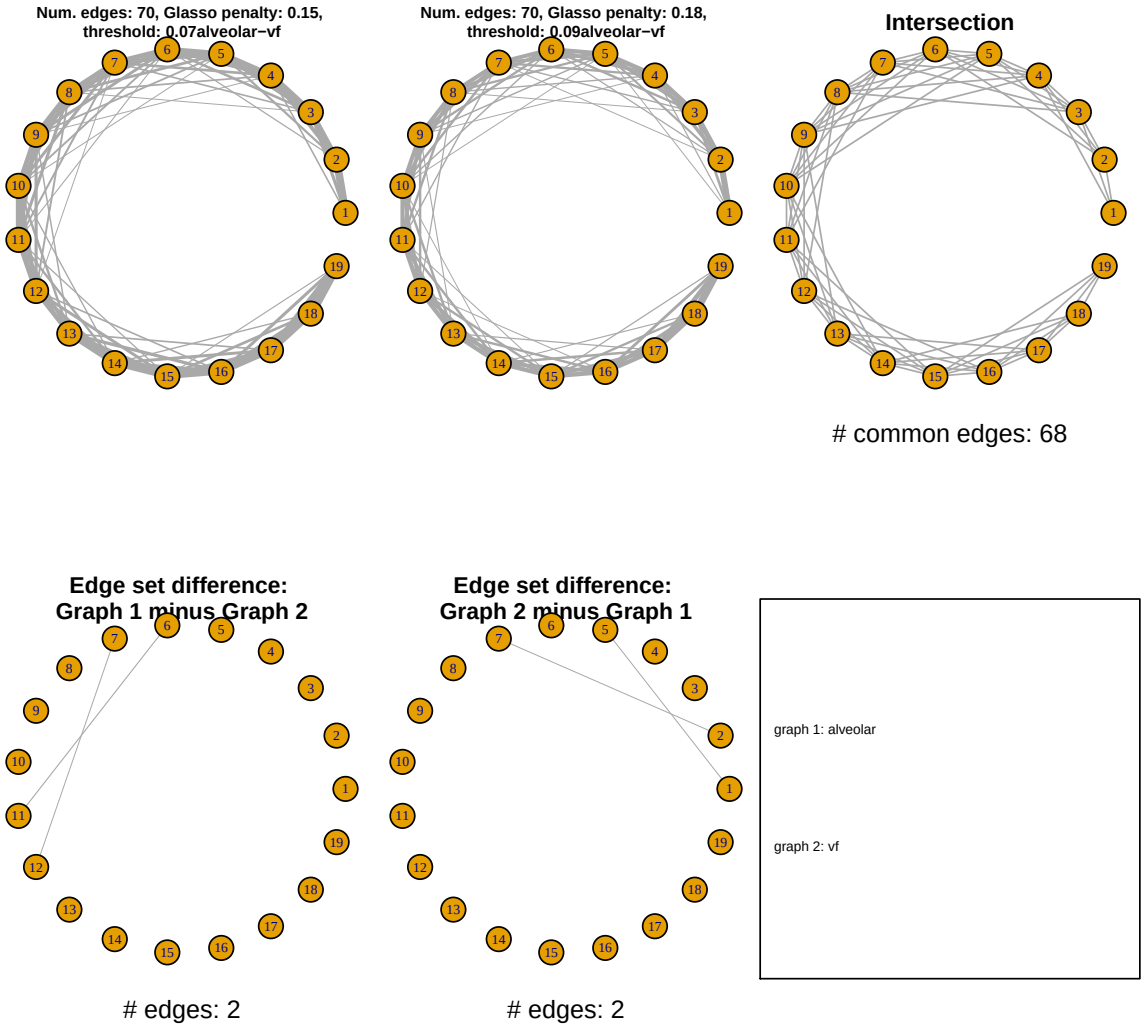

Figure 4.40: Time-time inverse covariance graphs for alveolar and vf words, as well as graph intersection and set differences.

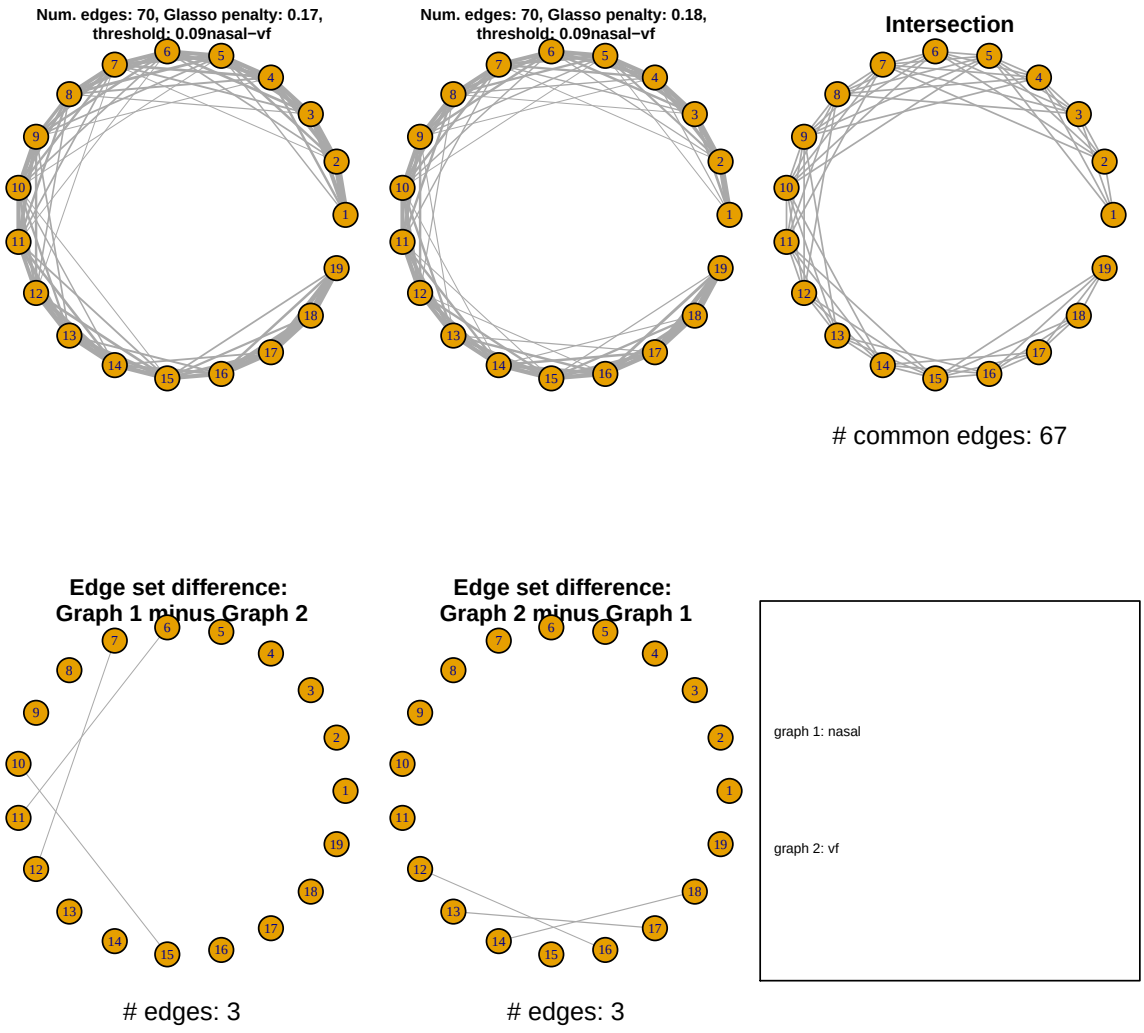

Figure 4.41: Time-time inverse covariance graphs for nasal and vf words, as well as graph intersection and set differences.

4.5 Conclusion

In this chapter we analyzed multi-indexed data containing trial replicates. We used the trial replicates to center the data, removing speaker-by-word means at each time point. We found that among words with long vowels, the vowel appears to be associated with the presence of word-word edges. We also found more between-consonant edges than within-consonant edges. In future work, we will investigate hypothesis testing of the edges to assess their validity, as well as cross-validation to select the penalty; in addition, we will examine whether the patterns we found hold in pitch curve data sets in other languages.

CHAPTER V

Future Work

We now discuss directions for future work in analyzing pitch curve data.

5.0.1 Decorrelation along the time axis

One direction for future work is to use the three-step algorithm proposed in *Zhou* (2014a) to decorrelate the data along the time axis in order to obtain more accurate word-word covariance estimates. The decorrelation can be performed either using a single time-time matrix across all words, or to separately estimate time-time matrices for subsets of the words. Some subsets of the words have time-time covariance matrices that are closer to stationary, so we can pool those words together and decorrelate using a common time-time inverse covariance matrix.

5.0.2 Cross-validation

Another direction for future work is to perform cross-validation to validate the choice of penalty parameter.

We consider a cross-validation procedure to select word and time penalty parameters, making use of the trial replicates.

1. Partition the people into pairs.

2. For each pair of people, withhold that pair, and estimate word-word and time-time precision matrices using the remaining people, sweeping out time and word penalties $\{(\lambda_i, \nu_i)\}$, with $\nu_i = k\lambda_i$. Run cross-validation for values of k equal to $1, 1.5, 2, 3$, and 6.

3. To evaluate the likelihood of the test set data, use the data matrix resulting from trial differencing and person averaging (of the test set pair of people). As discussed in Section 4.2.2, trial residualization can be expressed in terms of three trial differencing schemes:

   - (2 - 1) + (3 - 4)
   - (3 - 1) + (4 - 2)
   - (1 - 2) + (3 - 4).

   Run cross-validation three separate times, once using each type of trial difference when calculating the likelihood of the test set.

When calculating the likelihood of test set data under the estimated parameters from the training set, do the following:

1. Let A denote the time-time covariance matrix, and let B denote the word-word covariance matrix. The matrix-variate normal likelihood is

$$p(X \mid A, B) = \frac{\exp\left(-\frac{1}{2}\mathrm{tr}\left[A^{-1}X^T B^{-1}X\right]\right)}{(2\pi)^{n_2 n_4/2}|A|^{n_2/2}|B|^{n_4/2}}. \tag{5.1}$$

2. When calculating the likelihood of the test set data, we use the unpenalized likelihood.

$$\log(p(X(1), X(2), X(3), X(4) \mid A, B)) = \tag{5.2}$$

$$-\frac{1}{2}\sum_{r=1}^{4}\operatorname{tr}\left[A^{-1}(X(r)-\overline{X})^T B^{-1}(X(r)-\overline{X})\right]-\frac{n_p}{2}\log|A|-\frac{n_t}{2}\log|B| \tag{5.3}$$

5.0.3 Permutation tests and hypothesis testing

Another direction for future work hypothesis testing to validate edges. Some word groups exhibit more long range temporal correlations than others. The following permutation procedure can be used to assess whether the longer-range edges are due to the word groups or due to chance.

For $k = 1, \ldots, K$,

1. Let word group 1 consist of half the labial words and half the alveolar words, selected randomly. Let word group 2 consist of the remaining labial and alveolar words.

2. Estimate inverse correlation matrices $\widehat{B}^{-1}[k,1]$ and $\widehat{B}^{-1}[k,2]$ using each of the two word groups, respectively.

Average the precision matrices over the permutations:

$$\widehat{B}^{-1}[1]=\frac{1}{K}\sum_{k=1}^{K}\widehat{B}^{-1}[k,1] \quad \text{and} \quad \widehat{B}^{-1}[2]=\frac{1}{K}\sum_{k=1}^{K}\widehat{B}^{-1}[k,2]. \tag{5.4}$$

We obtain graphs from $\widehat{B}^{-1}[1]$ and $\widehat{B}^{-1}[2]$ by thresholding so that each graph has 75 edges. We then compare the edges using intersection and set differences.

5.0.4 Other matrix-variate models

Another direction for future work is to fit Kronecker sum models for the covariance or inverse covariance matrix. A related problem is model selection, in particular assessing whether Kronecker sum or Kronecker product models better fit the data.

5.0.5 Assessing the reasons for edges between word groups

Another direction for future research is to assess linguistic mechanisms that underlie the edges, and to assess whether the word-word and time-time patterns we found in the Afrikaans data set also appear in other languages.

APPENDIX

APPENDIX A

Additional Figures

A.0.1 Time-time covariance, correlation, inverse covariance, and inverse correlation

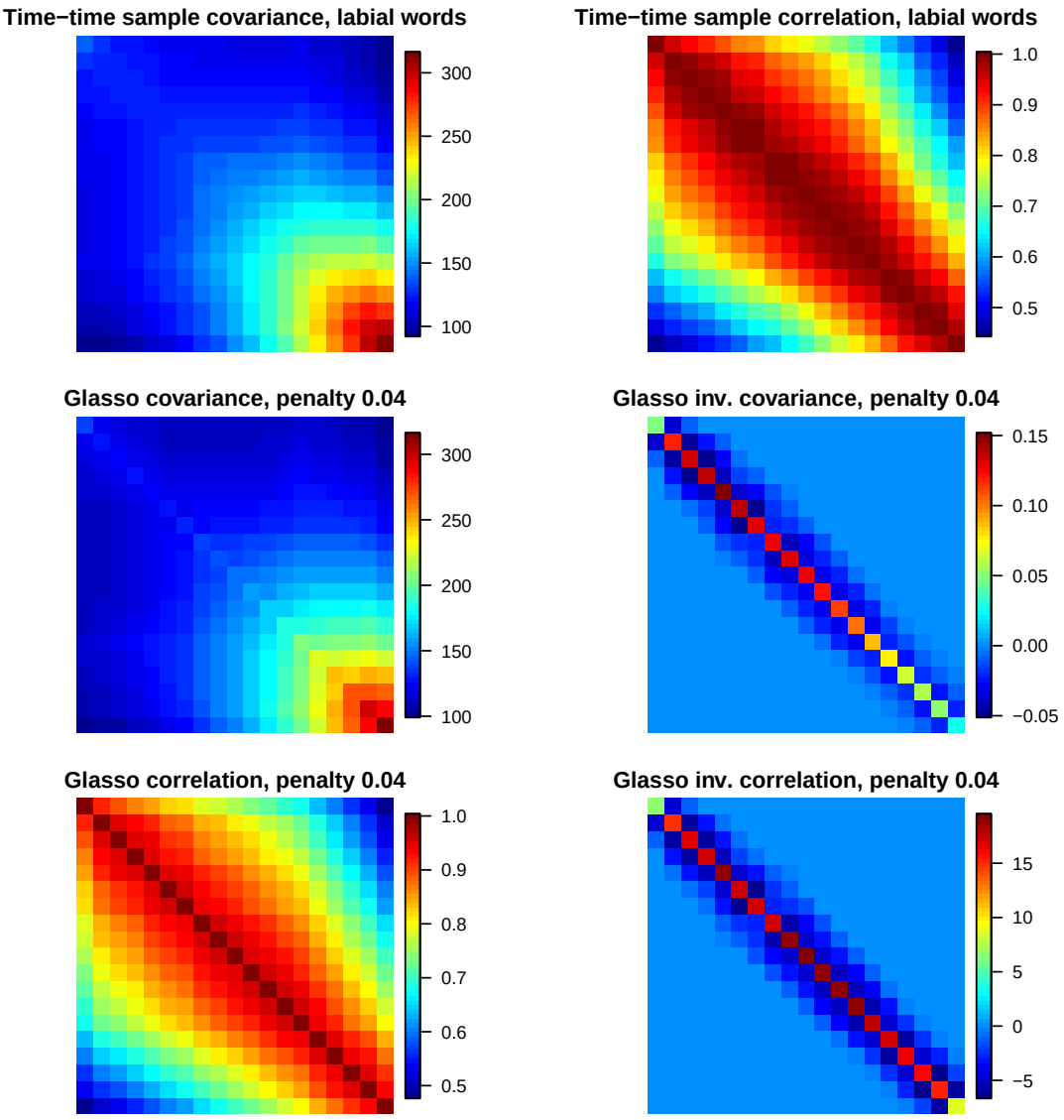

Figure A.1: Time-time sample covariance (top left), sample correlation (top right), Glasso covariance (middle left), Glasso inverse covariance (middle right), Glasso correlation (bottom left), and Glasso inverse correlation (bottom right), for words beginning with a labial consonant. The sample covariance is calculated as in (4.8), and the Glasso penalty parameter is chosen as in (4.9).

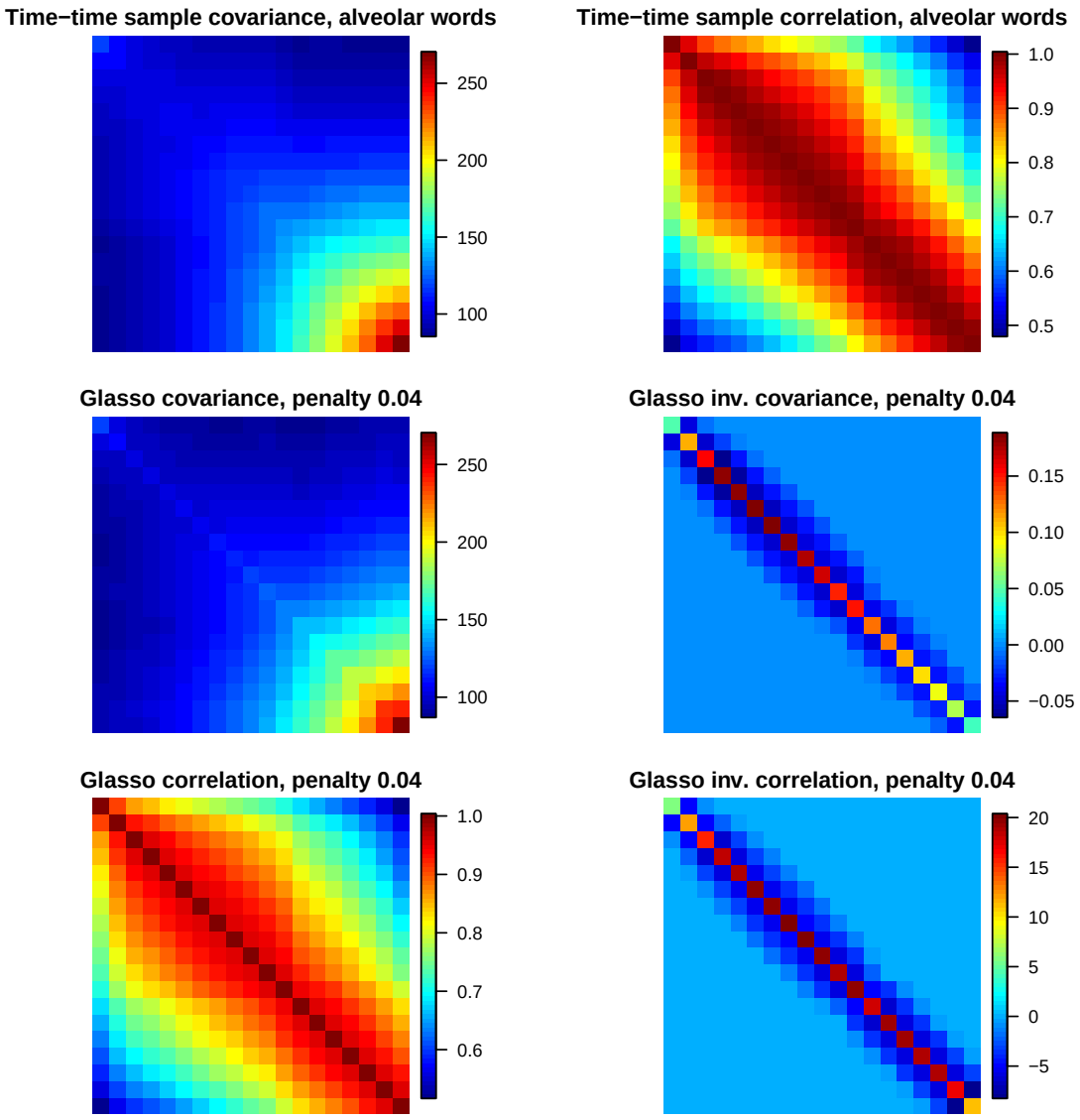

Figure A.2: Time-time sample covariance (top left), sample correlation (top right), Glasso covariance (middle left), Glasso inverse covariance (middle right), Glasso correlation (bottom left), and Glasso inverse correlation (bottom right), for words beginning with an alveolar consonant. The sample covariance is calculated as in (4.8), and the Glasso penalty parameter is chosen as in (4.9).

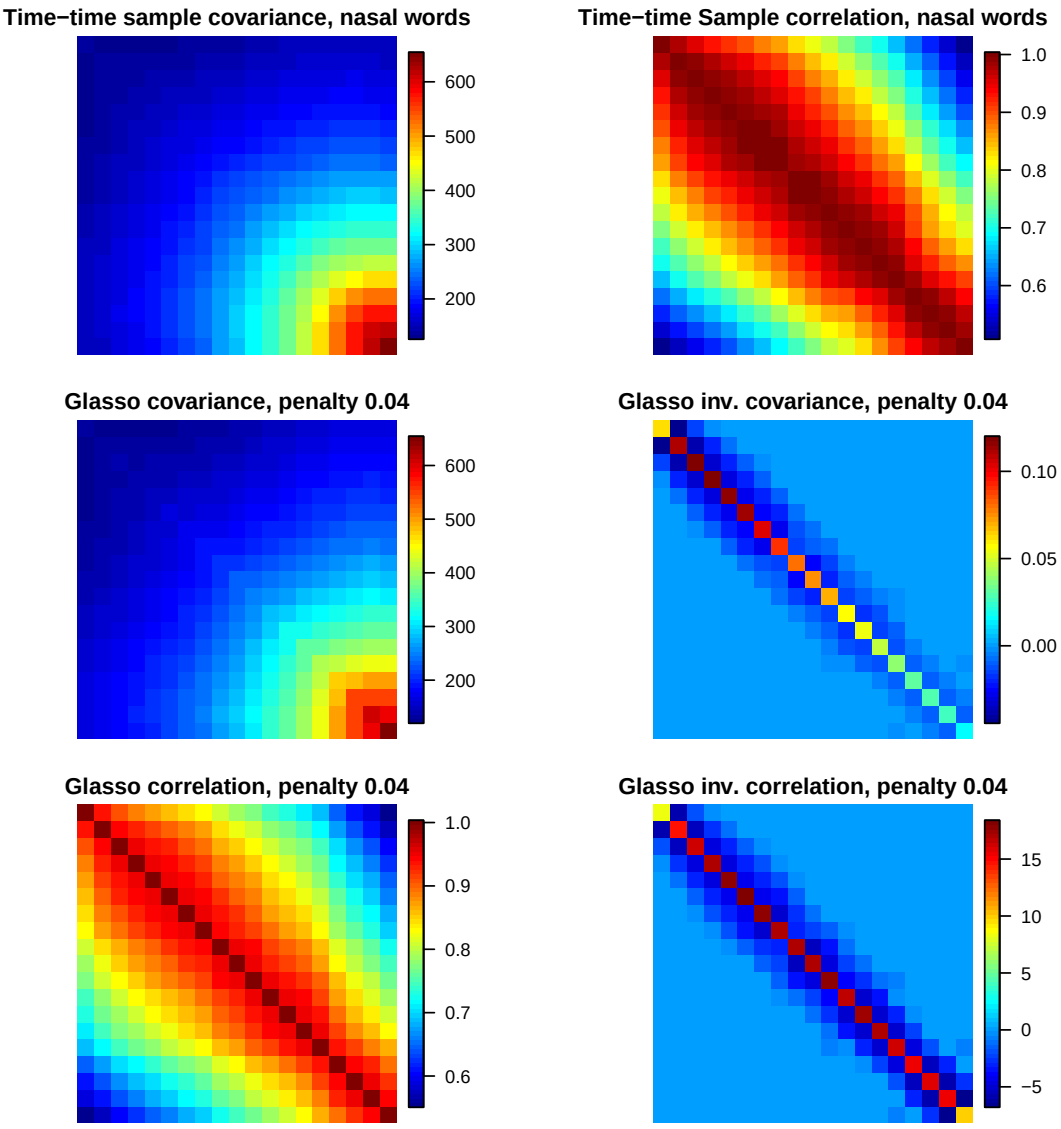

Figure A.3: Time-time sample covariance (top left), sample correlation (top right), Glasso covariance (middle left), Glasso inverse covariance (middle right), Glasso correlation (bottom left), and Glasso inverse correlation (bottom right), for words beginning with a nasal consonant. The sample covariance is calculated as in (4.8), and the Glasso penalty parameter is chosen as in (4.9).

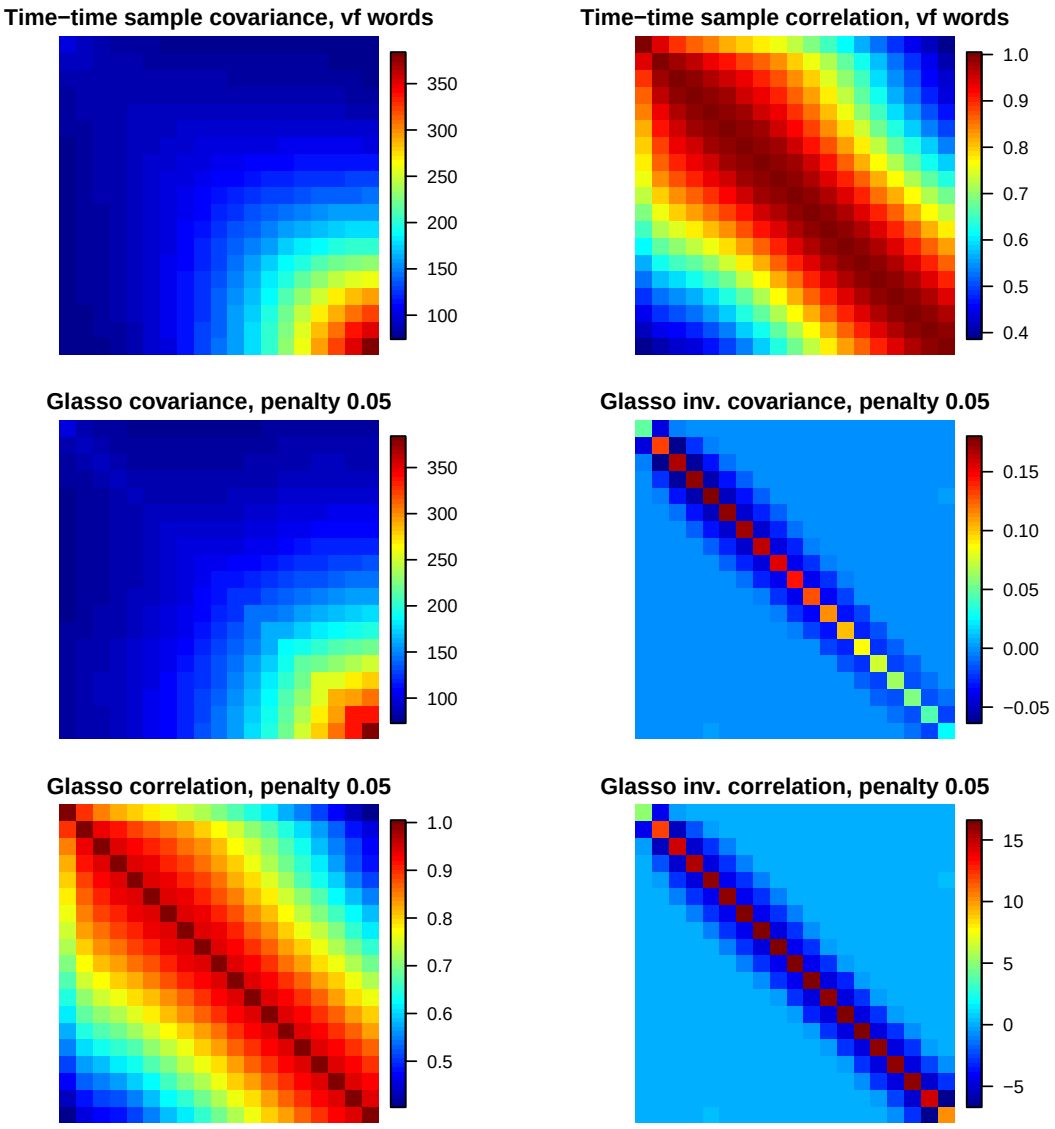

Figure A.4: Time-time sample covariance (top left), sample correlation (top right), Glasso covariance (middle left), Glasso inverse covariance (middle right), Glasso correlation (bottom left), and Glasso inverse correlation (bottom right), for words beginning with a vf consonant. The sample covariance is calculated as in (4.8), and the Glasso penalty parameter is chosen as in (4.9).

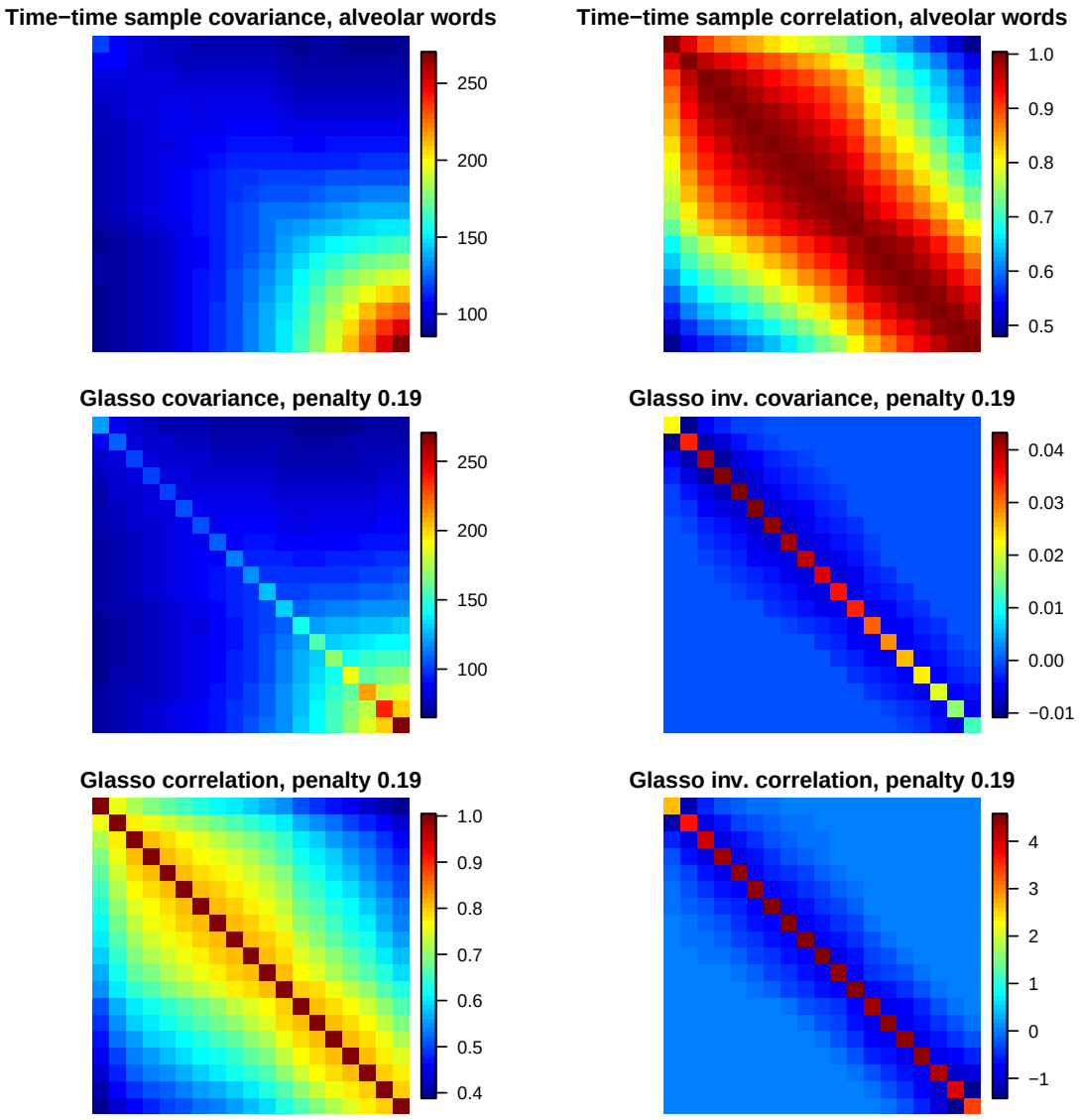

Figure A.5: Time-time sample covariance (top left), sample correlation (top right), Glasso covariance (middle left), Glasso inverse covariance (middle right), Glasso correlation (bottom left), and Glasso inverse correlation (bottom right), for words beginning with an alveolar consonant. The sample covariance is calculated as in (4.8), and the Glasso penalty parameter is chosen as five times the value of (4.9).

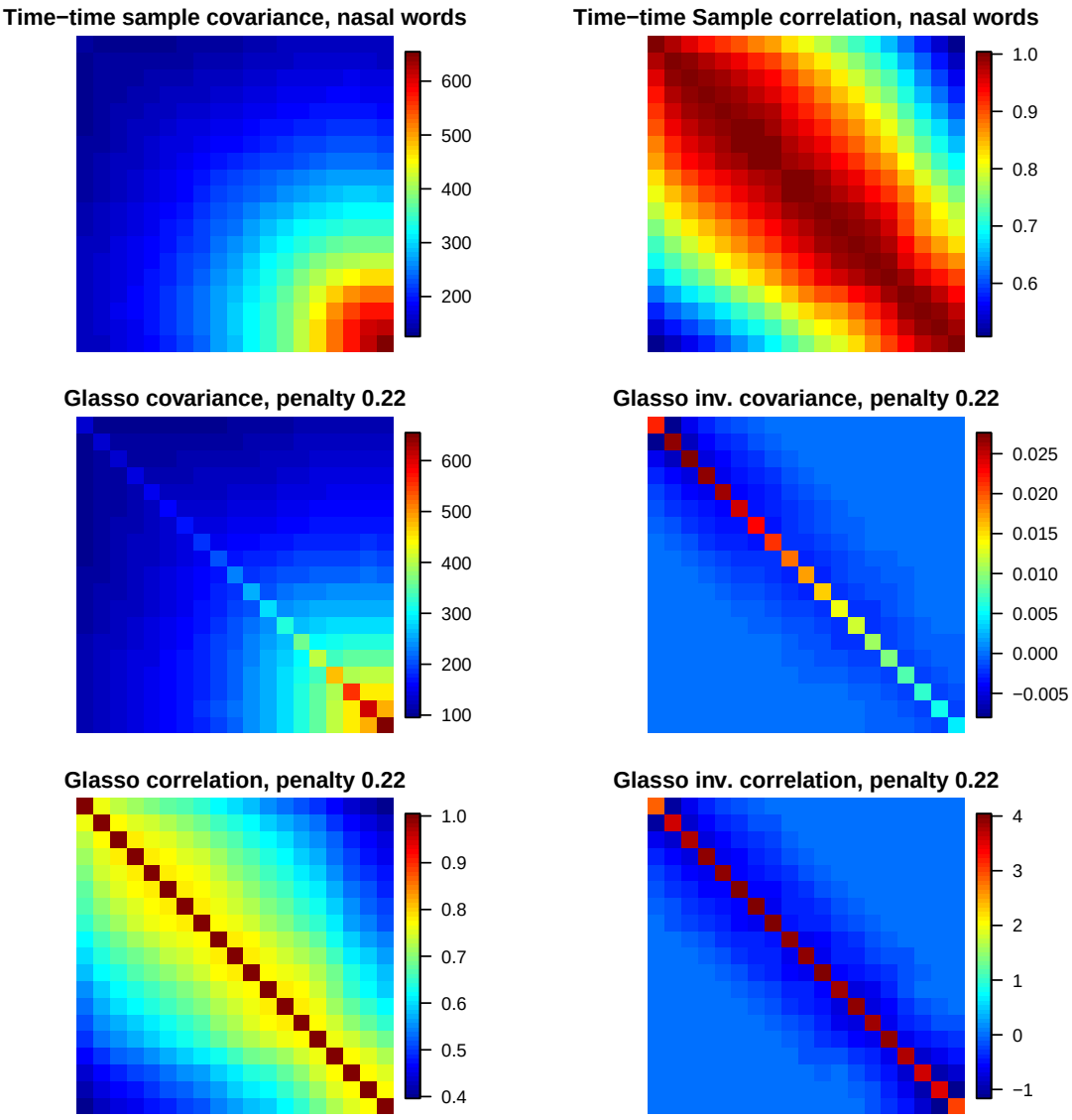

Figure A.6: Time-time sample covariance (top left), sample correlation (top right), Glasso covariance (middle left), Glasso inverse covariance (middle right), Glasso correlation (bottom left), and Glasso inverse correlation (bottom right), for words beginning with a nasal consonant. The sample covariance is calculated as in (4.8), and the Glasso penalty parameter is chosen as five times the value of (4.9).

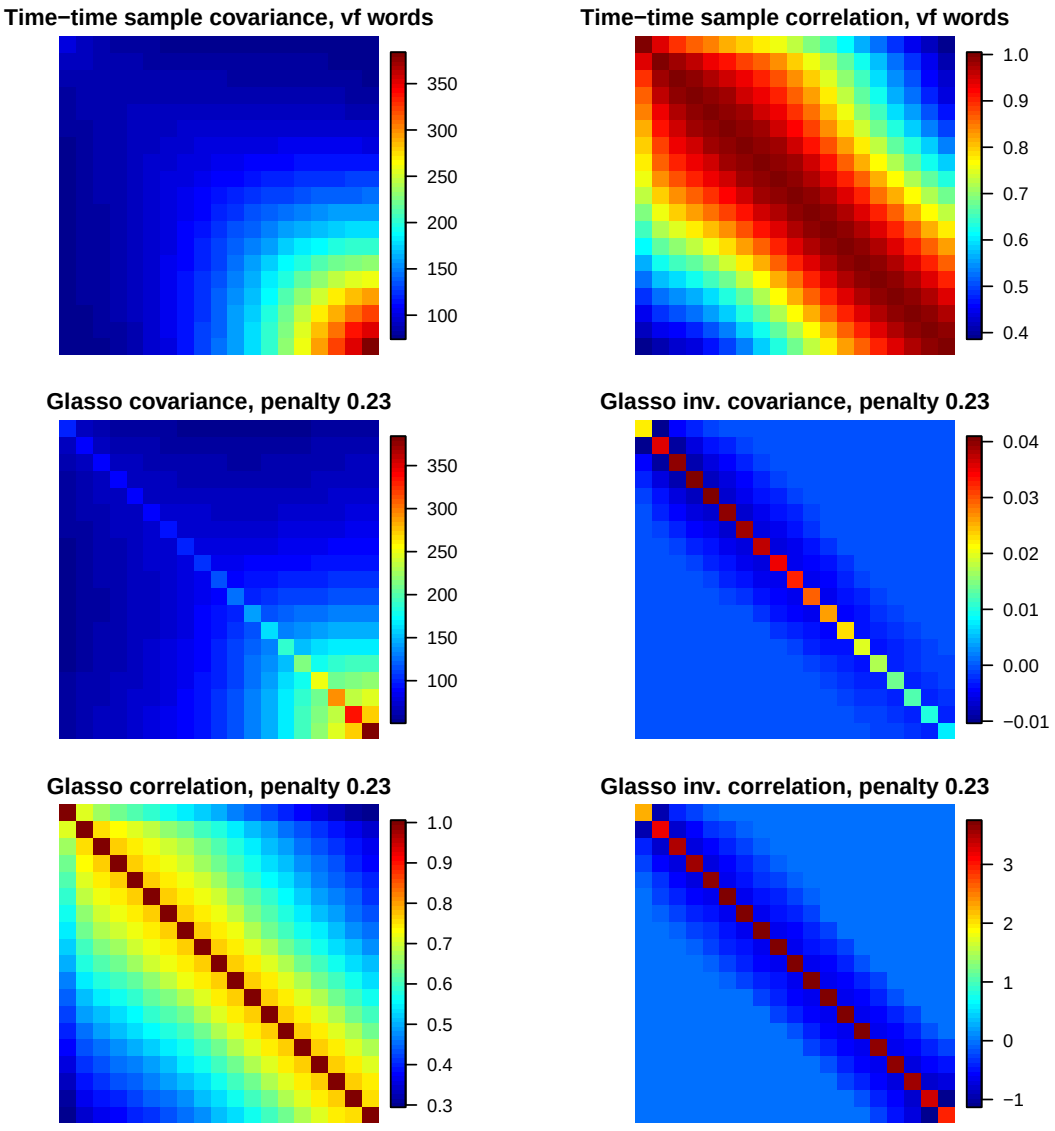

Figure A.7: Time-time sample covariance (top left), sample correlation (top right), Glasso covariance (middle left), Glasso inverse covariance (middle right), Glasso correlation (bottom left), and Glasso inverse correlation (bottom right), for words beginning with a vf consonant. The sample covariance is calculated as in (4.8), and the Glasso penalty parameter is chosen as five times the value of (4.9).

A.0.2 Word-word sample correlation and covariance heatmaps, and Glasso covariance, inverse covariance, correlation, and inverse correlation

The words are, we use an alphabetic ordering, which has the effect of grouping them by initial consonant. In the graphs, the words are also sorted alphabetically.

For each word group, there are strong edges that survive penalization and thresholding.

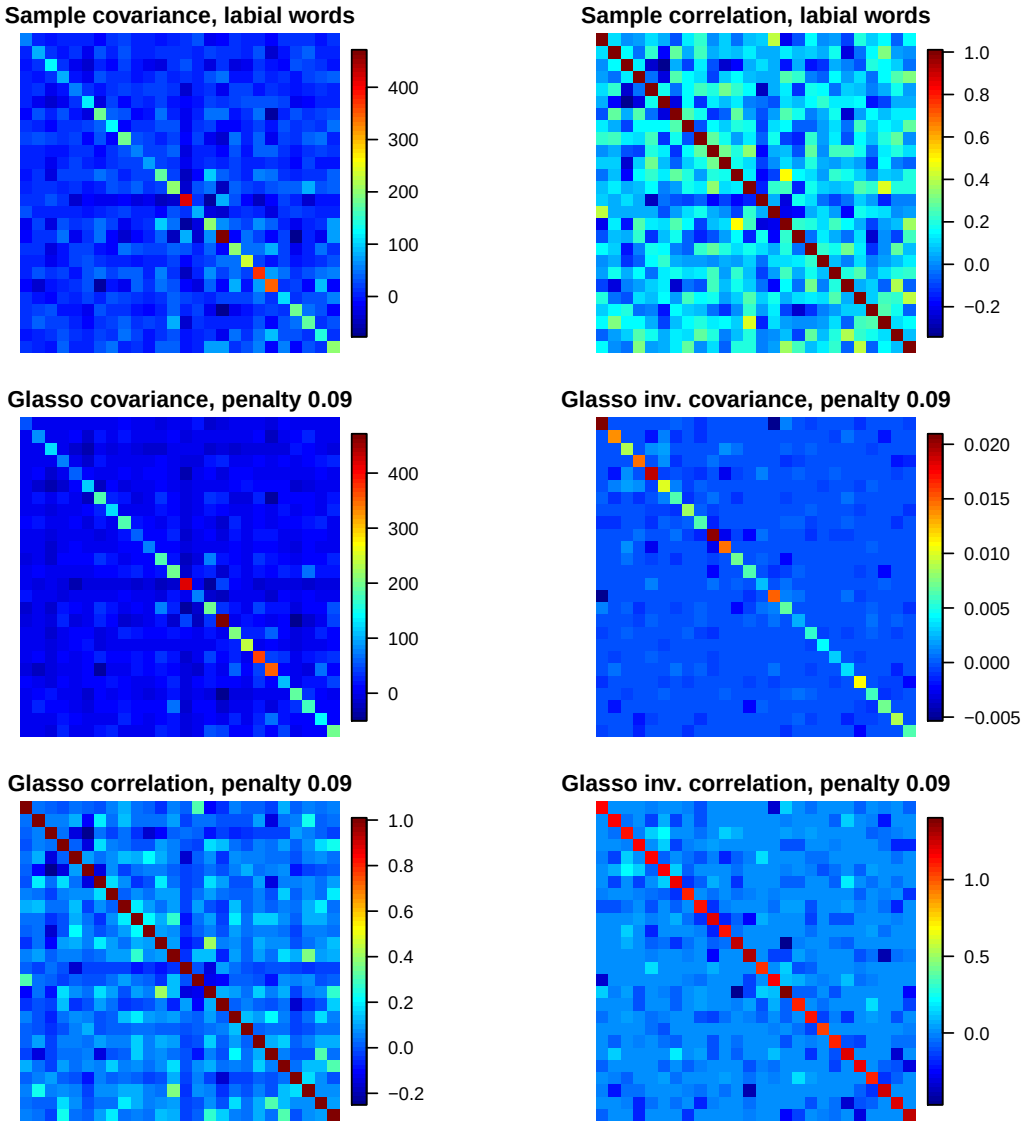

Figure A.8: Glasso covariance, inverse covariance, correlation, and inverse correlation for labial words. The top row of plots displays the estimated covariance and correlation matrices. The bottom row displays the estimated inverse covariance and inverse correlation matrices.

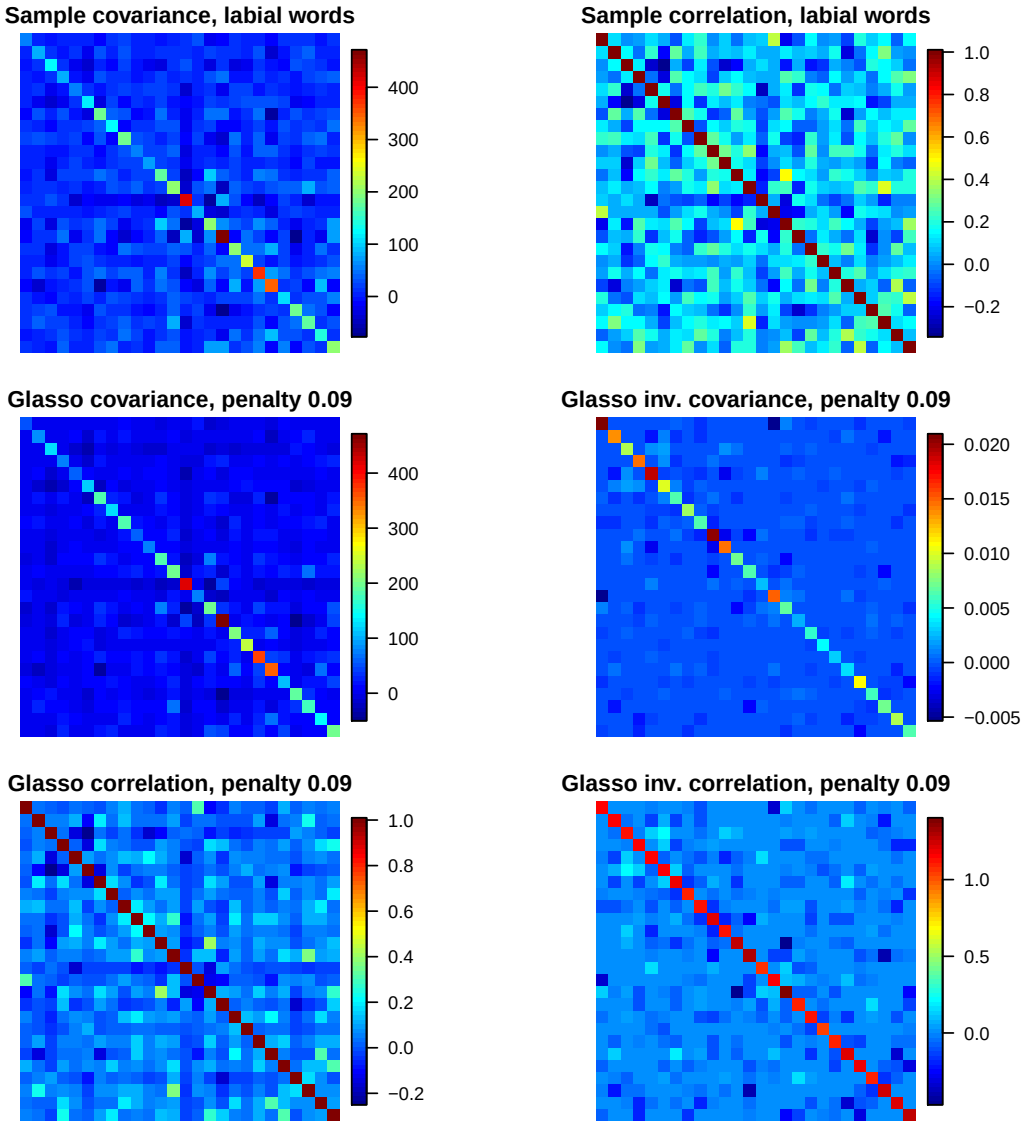

Figure A.9: Glasso covariance, inverse covariance, correlation, and inverse correlation for labial words. The top row of plots displays the estimated covariance and correlation matrices. The bottom row displays the estimated inverse covariance and inverse correlation matrices.

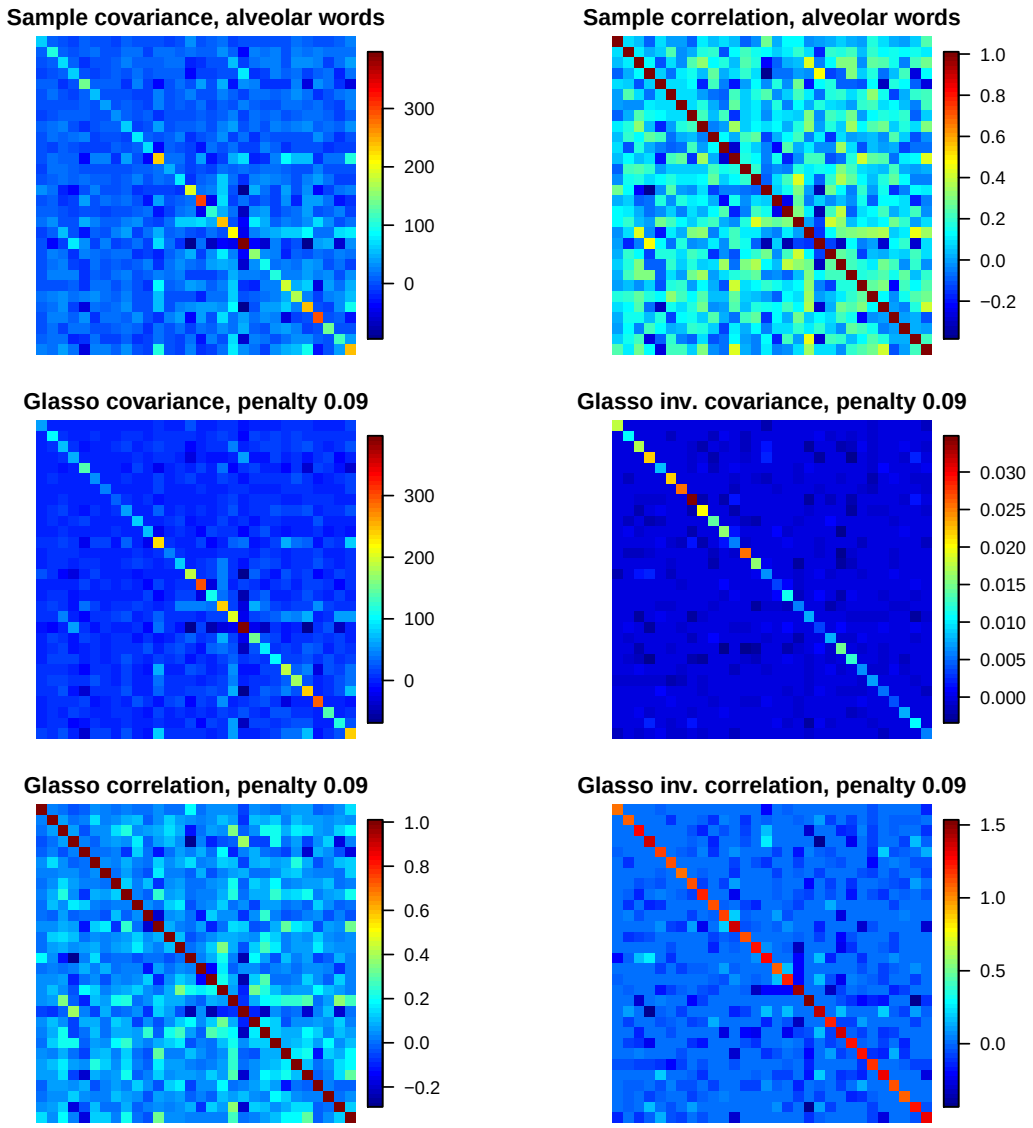

Figure A.10: Glasso covariance, inverse covariance, correlation, and inverse correlation for labial words. The top row of plots displays the estimated covariance and correlation matrices. The bottom row displays the estimated inverse covariance and inverse correlation matrices.

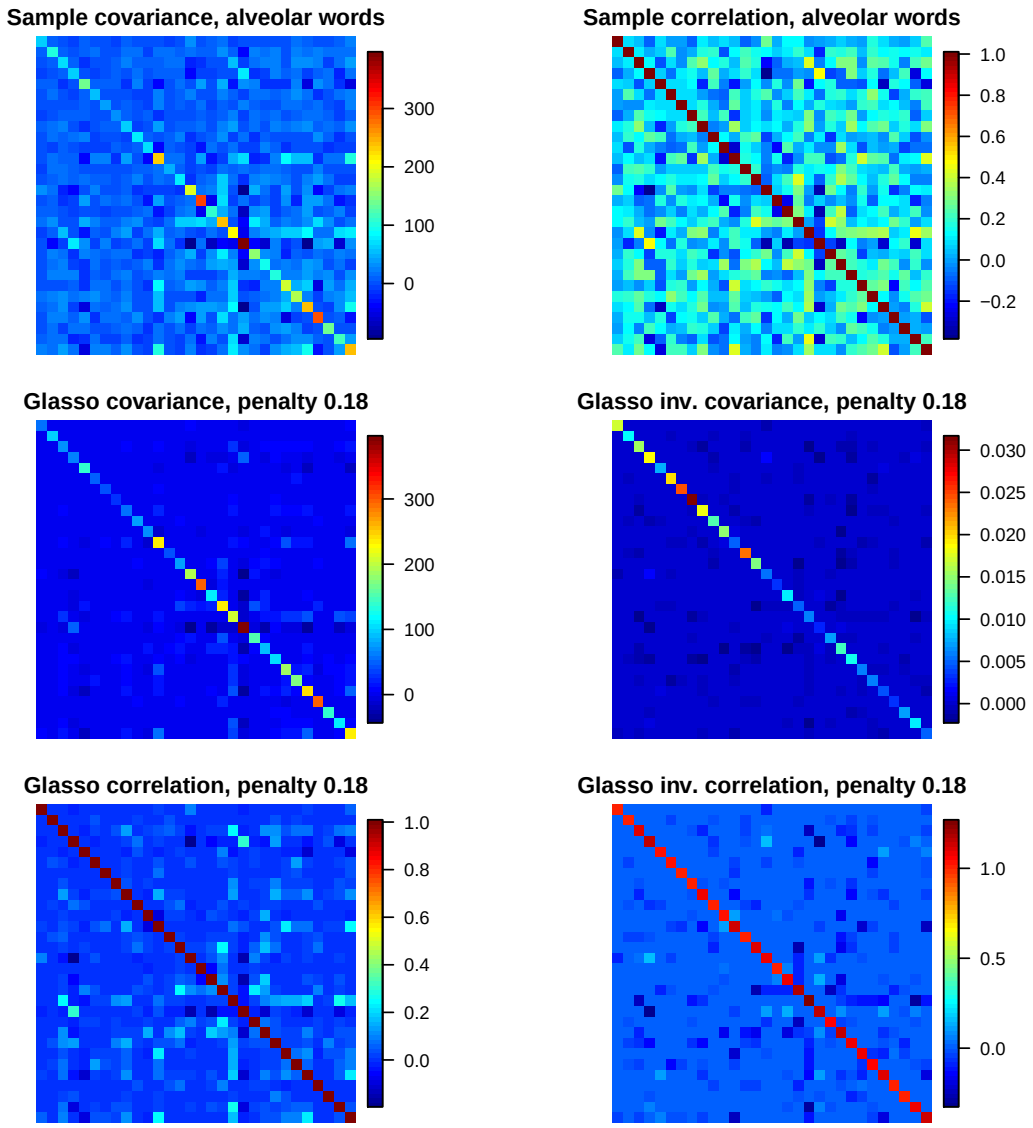

Figure A.11: Glasso covariance, inverse covariance, correlation, and inverse correlation for labial words. The top row of plots displays the estimated covariance and correlation matrices. The bottom row displays the estimated inverse covariance and inverse correlation matrices.

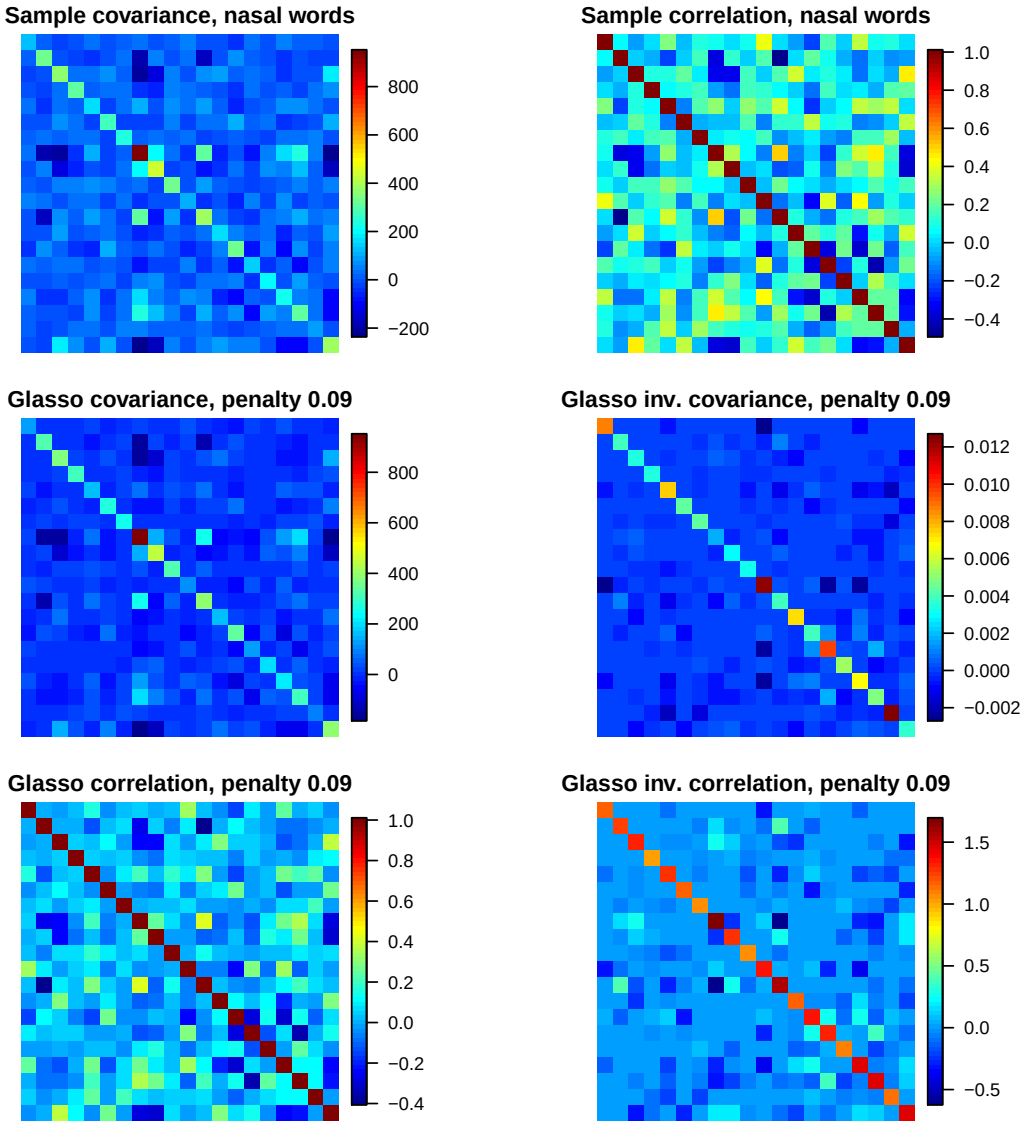

Figure A.12: Glasso covariance, inverse covariance, correlation, and inverse correlation for labial words. The top row of plots displays the estimated covariance and correlation matrices. The bottom row displays the estimated inverse covariance and inverse correlation matrices.

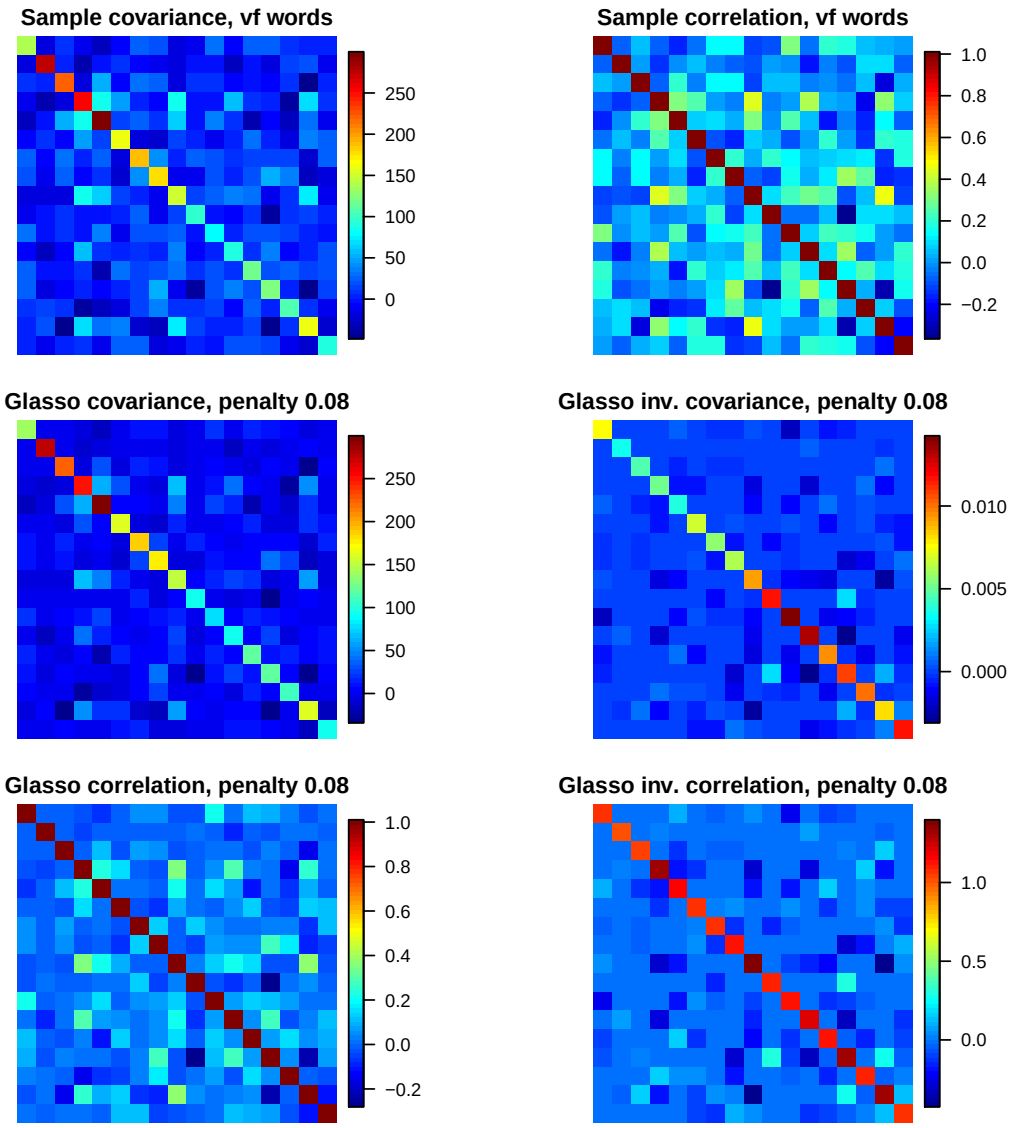

Figure A.13: Glasso covariance, inverse covariance, correlation, and inverse correlation for labial words. The top row of plots displays the estimated covariance and correlation matrices. The bottom row displays the estimated inverse covariance and inverse correlation matrices.

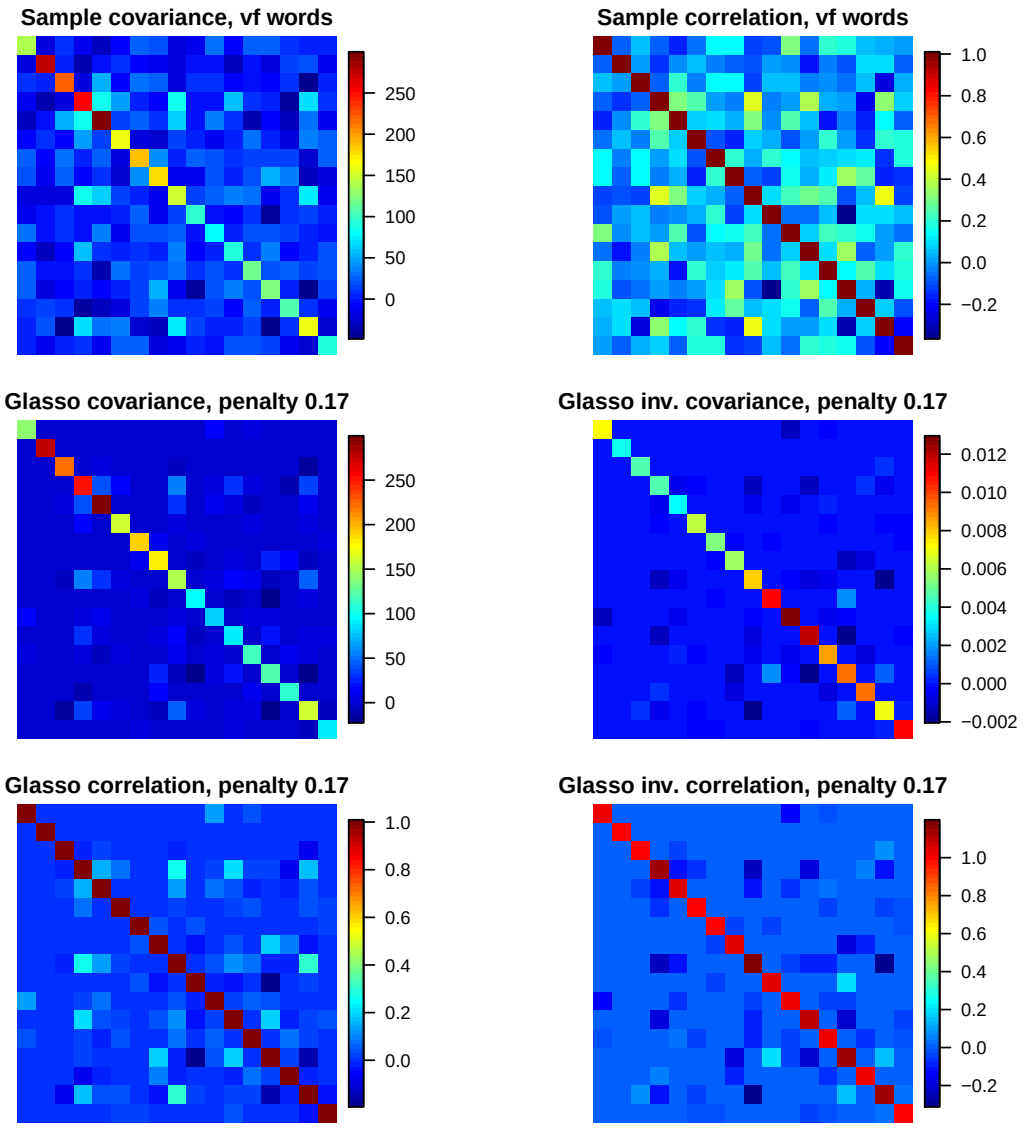

Figure A.14: Glasso covariance, inverse covariance, correlation, and inverse correlation for labial words. The top row of plots displays the estimated covariance and correlation matrices. The bottom row displays the estimated inverse covariance and inverse correlation matrices.

A.0.3 Edge graphs comparing Glasso and nodewise regression, for each pair of word groups (labial, alveolar, nasal, vf)

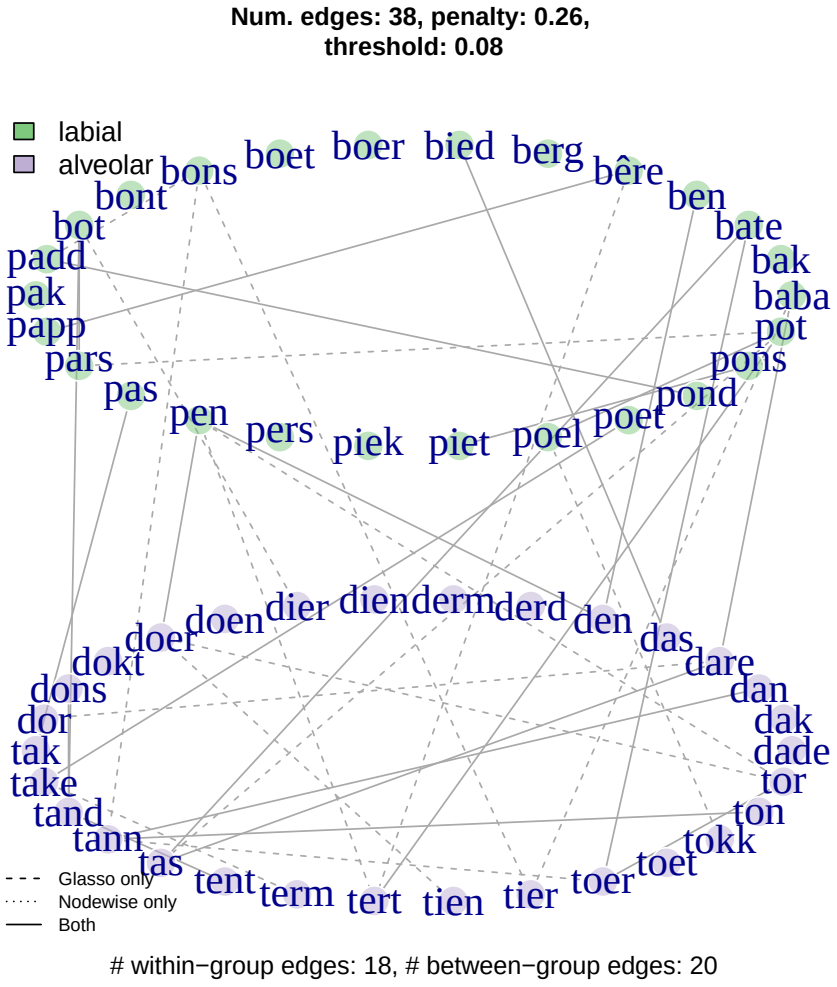

Figure A.15: Inverse covariance graph of labial and alveolar words. This graph displays a subgraph of a graph for all 93 words, estimated using Glasso and nodewise regression with thresholding.

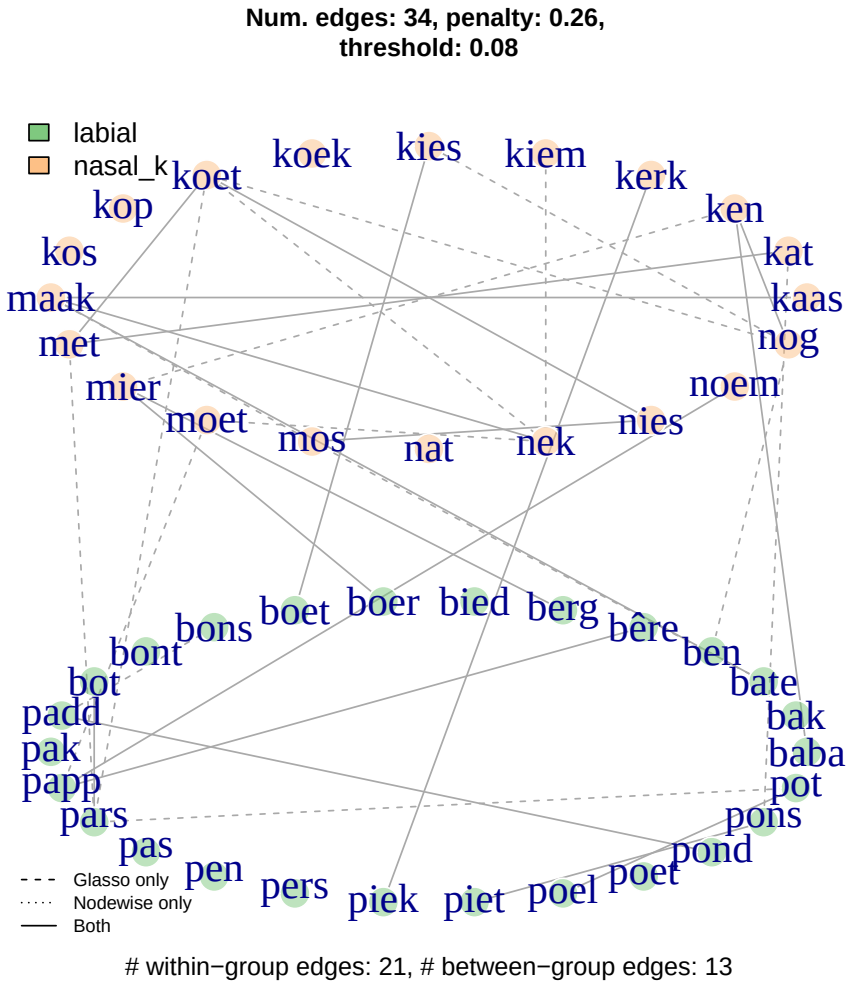

Figure A.16: Inverse covariance graph of labial and nasal words. This graph displays a subgraph of a graph for all 93 words, estimated using Glasso and nodewise regression with thresholding.

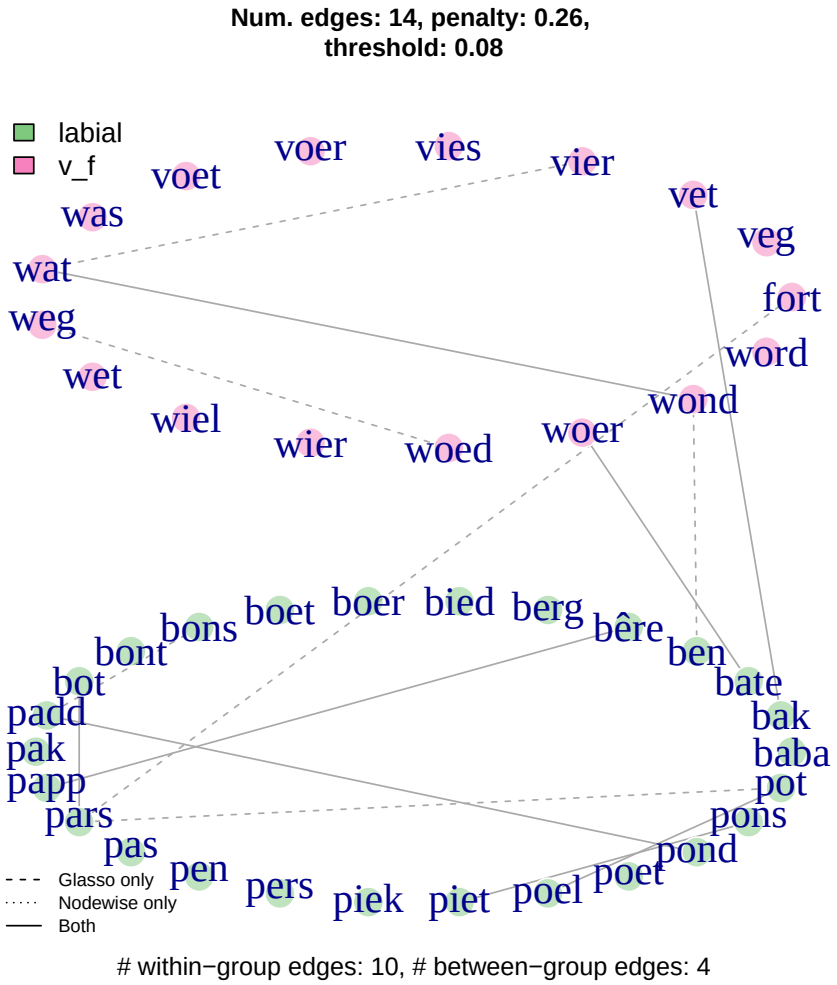

Figure A.17: Inverse covariance graph of labial and vf words. This graph displays a subgraph of a graph for all 93 words, estimated using Glasso and nodewise regression with thresholding.

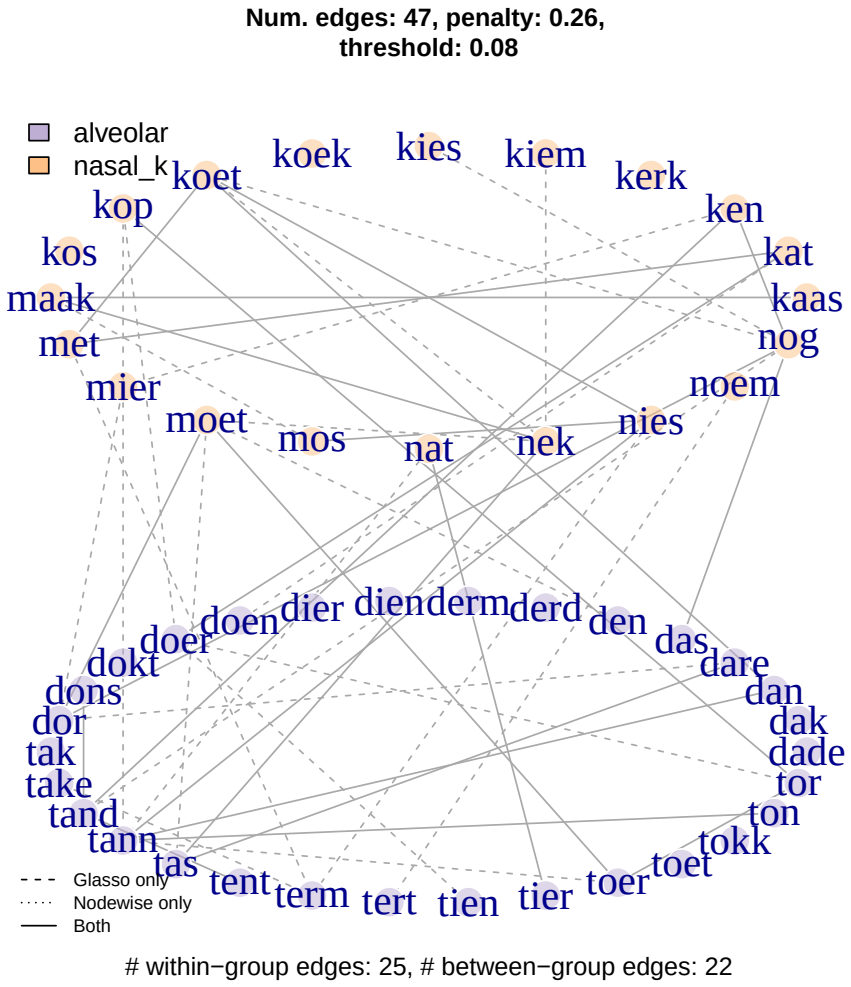

Figure A.18: Inverse covariance graph of alveolar and nasal words. This graph displays a subgraph of a graph for all 93 words, estimated using Glasso and nodewise regression with thresholding.

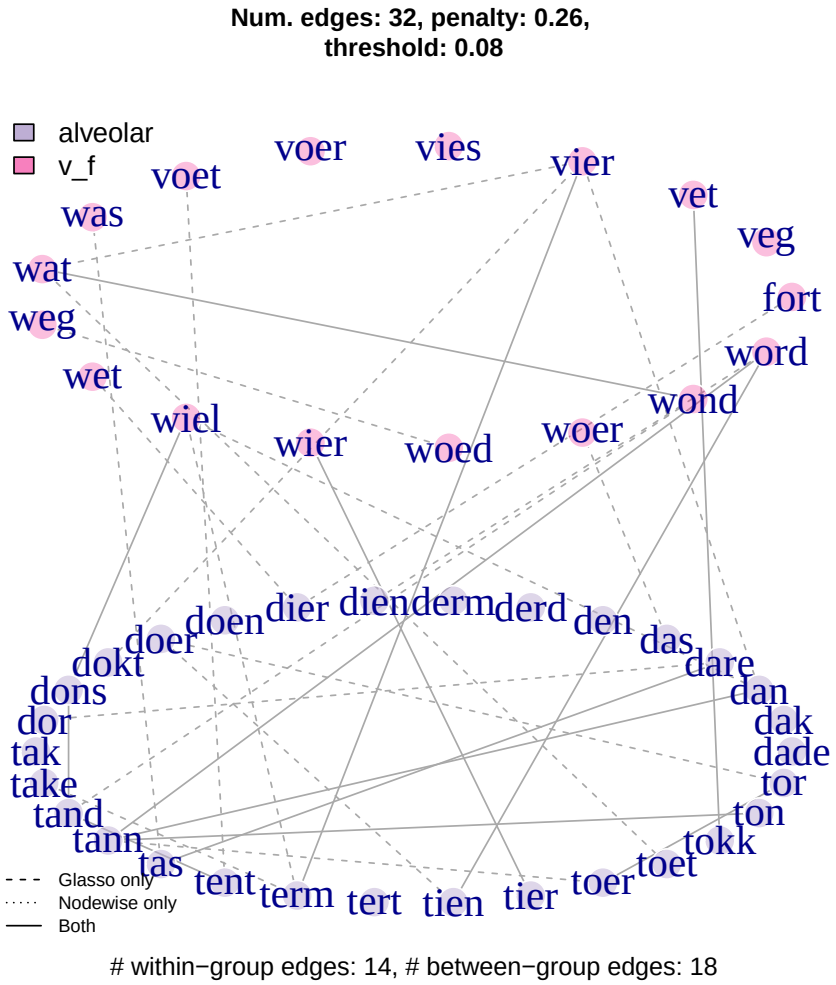

Figure A.19: Inverse covariance graph of alveolar and vf words. This graph displays a subgraph of a graph for all 93 words, estimated using Glasso and nodewise regression with thresholding.

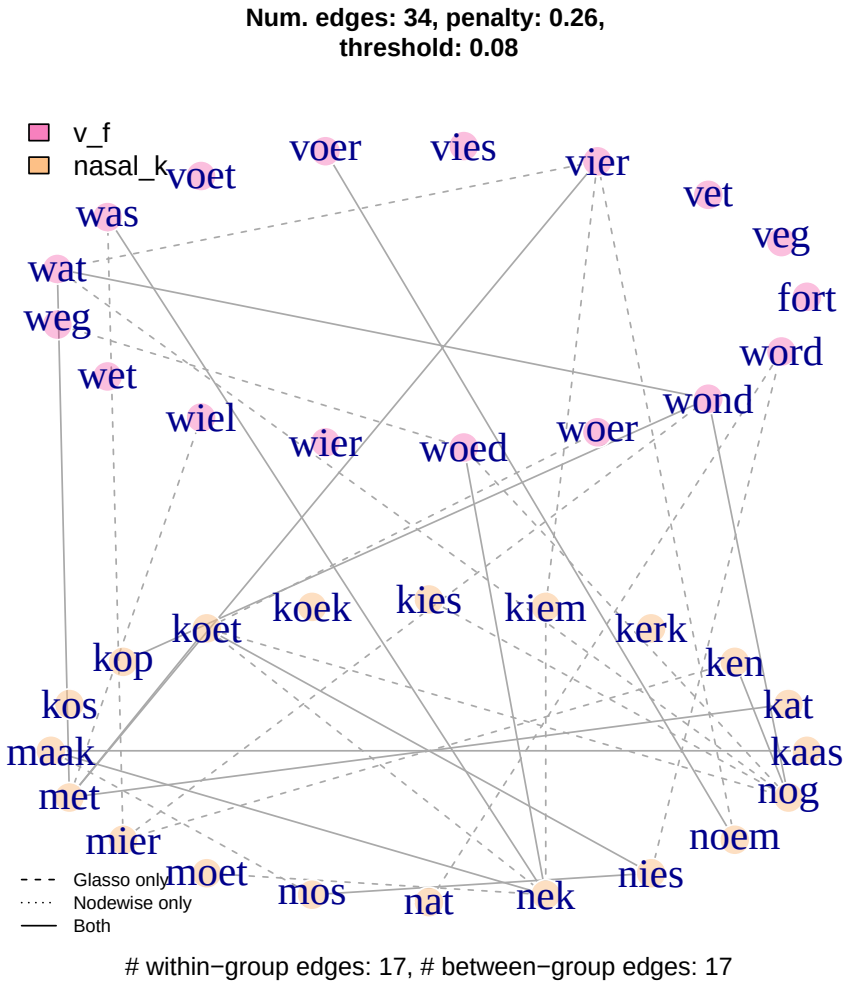

Figure A.20: Inverse covariance graph of nasal and vf words. This graph displays a subgraph of a graph for all 93 words, estimated using Glasso and nodewise regression with thresholding.

BIBLIOGRAPHY

BIBLIOGRAPHY

Aitken, A. C. (1936), IV.–on least squares and linear combination of observations, *Proceedings of the Royal Society of Edinburgh*, *55*, 42–48.

Allen, G. I., and R. Tibshirani (2012), Inference with transposable data: modelling the effects of row and column correlations, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(4), 721–743.

Aston, J. A., J.-M. Chiou, and J. P. Evans (2010), Linguistic pitch analysis using functional principal component mixed effect models, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *59*(2), 297–317.

Baayen, R. H., D. J. Davidson, and D. M. Bates (2008), Mixed-effects modeling with crossed random effects for subjects and items, *Journal of memory and language*, *59*(4), 390–412.

Bai, Z., and H. Saranadasa (1996), Effect of high dimension: by an example of a two sample problem, *Statistica Sinica*, pp. 311–329.

Banerjee, O., L. E. Ghaoui, and A. d'Aspremont (2008), Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data, *Journal of Machine Learning Research*, *9*, 485–516.

Benjamini, Y., and Y. Hochberg (1995), Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300.

Benjamini, Y., and D. Yekutieli (2001), The control of the false discovery rate in multiple testing under dependency, *Annals of statistics*, pp. 1165–1188.

Cai, T., X. Jessie Jeng, and J. Jin (2011), Optimal detection of heterogeneous and heteroscedastic mixtures, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(5), 629–662.

Cai, T. T., and Y. Xia (2014), High-dimensional sparse manova, *Journal of Multivariate Analysis*, *131*, 174–196.

Cai, T. T., H. Li, W. Liu, and J. Xie (2016), Joint estimation of multiple high-dimensional precision matrices, *Statistica Sinica*, *26*(2), 445.

Chen, S. X., Y.-L. Qin, et al. (2010), A two-sample test for high-dimensional data with applications to gene-set testing, *The Annals of Statistics*, *38*(2), 808–835.

Chen, S. X., J. Li, and P. S. Zhong (2014), Two-sample tests for high dimensional means with thresholding and data transformation, *arXiv preprint arXiv:1410.2848*.

Clark, H. H. (1973), The language-as-fixed-effect fallacy: A critique of language statistics in psychological research, *Journal of verbal learning and verbal behavior*, *12*(4), 335–359.

Coetzee, A. W., P. S. Beddor, K. Shedden, W. Styler, and D. Wissing (2018), Plosive voicing in afrikaans: Differential cue weighting and tonogenesis, *Journal of Phonetics*, *66*, 185–216.

Dempster, A. P. (1972), Covariance selection, *Biometrics*, pp. 157–175.

Devlin, B., and K. Roeder (1999), Genomic control for association studies, *Biometrics*, *55*(4), 997–1004.

Dutilleul, P. (1999), The mle algorithm for the matrix normal distribution, *Journal of Statistical Computation and Simulation*, *64*(2), 105–123.

Edgar, R., M. Domrachev, and A. E. Lash (2002), Gene expression omnibus: Ncbi gene expression and hybridization array data repository, *Nucleic acids research*, *30*(1), 207–210.

Efron, B. (2005), Large-scale simultaneous hypothesis testing: the choice of a null hypothesis, *Journal of the American Statistical Association*, *99*, 96–104.

Efron, B. (2007), Correlation and large-scale simultaneous significance testing, *Journal of the American Statistical Association*, *102*(477).

Efron, B. (2009), Are a set of microarrays independent of each other?, *Ann. App. Statist.*, *3*(3), 922–942.

Efron, B. (2010), *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, vol. 1, Cambridge University Press.

Fan, J., Y. Feng, and Y. Wu (2009), Network exploration via the adaptive LASSO and SCAD penalties, *The Annals of Applied Statistics*, *3*, 521–541.

Friedman, J., T. Hastie, and R. Tibshirani (2008), Sparse inverse covariance estimation with the graphical Lasso, *Biostatistics*, *9*(3), 432–441, doi: 10.1093/biostatistics/kxm045.

Greenwald, K., S. Zhou, and A. Hero (2017), Tensor graphical lasso, *arXiv preprint arXiv:1705.03983*.

Hall, P., J. Jin, et al. (2010), Innovated higher criticism for detecting sparse signals in correlated noise, *The Annals of Statistics*, *38*(3), 1686–1732.

Hanson, H. M. (2009), Effects of obstruent consonants on fundamental frequency at vowel onset in english, *The Journal of the Acoustical Society of America*, *125*(1), 425–441.

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

Hornstein, M., R. Fan, K. Shedden, and S. Zhou (2018), Joint mean and covariance estimation with unreplicated matrix-variate data, *Journal of the American Statistical Association*, (just-accepted).

House, A. S., and G. Fairbanks (1953), The influence of consonant environment upon the secondary acoustical characteristics of vowels, *The Journal of the Acoustical Society of America*, *25*(1), 105–113.

Kalaitzis, A., J. Lafferty, N. Lawrence, and S. Zhou (2013), The bigraphical lasso, in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1229–1237.

Kirby, J. P., and D. R. Ladd (2016), Effects of obstruent voicing on vowel f 0: Evidence from "true voicing" languages, *The Journal of the Acoustical Society of America*, *140*(4), 2400–2411.

Kruskal, W. (1988), Miracles and statistics: The casual assumption of independence, *Journal of the American Statistical Association*, *83*(404), 929–940.

Ladefoged, P., and S. F. Disner (2012), *Vowels and consonants*, John Wiley & Sons.

Lam, C., and J. Fan (2009), Sparsistency and rates of convergence in large covariance matrices estimation, *Annals of Statistics*, *37*(6B), 4254–4278.

Laver, J. (1994), *Principles of phonetics*, Cambridge University Press.

Leek, J. T., R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry (2010), Tackling the widespread and critical impact of batch effects in high-throughput data, *Nature Reviews Genetics*, *11*(10), 733–739.

Lepage, P., et al. (2011), Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis, *Gastroenterology*, *141*(1), 227–236.

Li, J., and P.-S. Zhong (2014), A rate optimal procedure for sparse signal recovery under dependence, *arXiv preprint arXiv:1410.2839*.

Loh, P., and M. Wainwright (2012), High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity, *The Annals of Statistics*, *40*(3), 1637–1664.

Meinshausen, N., and P. Bühlmann (2006), High dimensional graphs and variable selection with the Lasso, *Annals of Statistics*, *34*(3), 1436–1462.

Owen, A. B. (2005), Variance of the number of false discoveries, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(3), 411–426.

Park, S., K. Shedden, and S. Zhou (2017), Non-separable covariance models for spatio-temporal data, with applications to neural encoding analysis, *arXiv preprint arXiv:1705.05265*.

Peng, J., P. Wang, N. Zhou, and J. Zhu (2009), Partial correlation estimation by joint sparse regression models, *Journal of the American Statistical Association*, *104*, 735–746.

Quené, H., and H. Van den Bergh (2008), Examples of mixed-effects modeling with crossed random effects and with binomial data, *Journal of Memory and Language*, *59*(4), 413–425.

Ravikumar, P., M. Wainwright, G. Raskutti, and B. Yu (2011), High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence, *Electronic Journal of Statistics*, *4*, 935–980.

Ren, Z., T. Sun, C.-H. Zhang, and H. H. Zhou (2015), Asymptotic normality and optimalities in estimation of large gaussian graphical model, *Annals of Statistics*, *43*(3), 991–1026.

Rothman, A., P. Bickel, E. Levina, and J. Zhu (2008), Sparse permutation invariant covariance estimation, *Electronic Journal of Statistics*, *2*, 494–515.

Storey, J. D. (2003), The positive false discovery rate: a Bayesian interpretation and the q-value, *Annals of statistics*, pp. 2013–2035.

Sugden, L. A., M. R. Tackett, Y. A. Savva, W. A. Thompson, and C. E. Lawrence (2013), Assessing the validity and reproducibility of genome-scale predictions, *Bioinformatics*, *29*(22), 2844–2851.

Sun, Y., N. R. Zhang, and A. B. Owen (2012), Multiple hypothesis testing adjusted for latent variables, with an application to the agemap gene expression data, *The Annals of Applied Statistics*, pp. 1664–1688.

Tan, K. M., and D. M. Witten (2014), Sparse biclustering of transposable data, *Journal of Computational and Graphical Statistics*, *23*(4), 985–1008.

Tsiligkaridis, T., A. Hero, and S. Zhou (2013), On convergence of kronecker graphical lasso algorithms, *IEEE Transactions on Signal Processing*, *61*, 1743 – 1755.

Wang, J., Q. Zhao, T. Hastie, and A. B. Owen (2015), Confounder adjustment in multiple hypothesis testing, *arXiv preprint arXiv:1508.04178*.

Weichsel, P. (1962), The kronecker product of graphs, *Proc. Amer. Math. Soc.*, *13*, 47–52.

Werner, K., M. Jansson, and P. Stoica (2008), On estimation of covariance matrices with kronecker product structure, *IEEE Transactions on Signal Processing*, *56*(2), 478 – 491.

Yuan, M. (2010), High dimensional inverse covariance matrix estimation via linear programming, *Journal of Machine Learning Research*, *11*, 2261–2286.

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

Zhou, S. (2010), Thresholded lasso for high dimensional variable selection and statistical estimation, *Tech. rep.*, Department of Statistics, University of Michigan.

Zhou, S. (2014a), Gemini: Graph estimation with matrix variate normal instances, *Annals of Statistics*, *42*(2), 532–562.

Zhou, S. (2014b), Supplement to "Gemini: Graph estimation with matrix variate normal instances", *Annals of Statistics*, doi:10.1214/13-AOS1187SUPP, dOI:10.1214/13-AOS1187SUPP.

Zhou, S., J. Lafferty, and L. Wasserman (2010), Time varying undirected graphs, *Machine Learning*, *80*(2–3), 295–319.

Zhou, S., P. Rütimann, M. Xu, and P. Bühlmann (2011), High-dimensional covariance estimation based on Gaussian graphical models, *Journal of Machine Learning Research*, *12*, 2975–3026.