TOPICS IN HIGH DIMENSIONAL STATISTICAL ESTIMATION AND INFERENCE

BY

RITWIK MITRA

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Statistics & Biostatistics

Written under the direction of

Professor Cun-Hui Zhang

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2015

ABSTRACT OF THE DISSERTATION

Topics In
High Dimensional Statistical Estimation And Inference

By RITWIK MITRA

Dissertation Director:

Professor Cun-Hui Zhang

This thesis deals with three problems. The first two of the problems are related in that they are concerned with estimation of correlation and precision matrix in spectral norm. These two problems are tackled in Chapters 2, 3. The third problem is the construction of chi-squared type test for groups of variables in high dimensional linear regression.

In Chapter 2, we study concentration in spectral norm of nonparametric estimates of correlation matrices. We study two nonparametric estimates of correlation matrices in Gaussian copula models and prove that when both the number of variables and sample size are large, the spectral error of the nonparametric estimators is of no greater order than that of the latent sample covariance matrix, at least when compared with some of the sharpest known error bounds for the later. As an application, we

establish the minimax optimal rate in the estimation of high-dimensional bandable correlation matrices via tapering off of these nonparametric estimators. An optimal convergence rate for sparse principal component analysis is also established.

In Chapter 3, we study the sparse precision matrix estimation procedure in the same Gaussian copula model as in Chapter 2. We employ the scaled Lasso procedure for inversion of nonparametric correlation matrix estimates based on Kendall's tau. We prove the optimal rate of convergence in estimation of sparse precision matrices under the weaker condition of bound on the spectral norm of the precision matrix.

Chapter 4 deals with confidence regions and approximate chi-squared tests for variable groups in high-dimensional linear regression. We develop a scaled group Lasso for efficient chi-squared-based statistical inference of variable groups. We prove that the proposed methods capture the benefit of group sparsity under proper conditions, for statistical inference of the noise level and variable groups, large and small. Oracle inequalities are provided for the scaled group Lasso in prediction and several estimation losses, and for the group Lasso as well in a weighted mixed loss. Some simulation results are also provided in support of the theory.

Acknowledgements

My deepest gratitude to my advisor Dr. Cun-Hui Zhang whose unending patience and support has guided me through out my Ph.D. days. His utmost dedication to research and unparalleled grasp on the subject are constant sources of inspiration and awe. On top of teaching me everything I know about research, he has also offered his help and advice whenever I have needed either. For all of that, Thank you.

My thanks to my dissertation committee members Dr. Dan Yang, Dr. David Tyler and Dr. Han Liu for agreeing to be a part of the committee, their busy schedules notwithstanding. My thanks also to Dr. Tyler, Dr. Sackrowitz, Dr. Gundy and my advisor Dr. Zhang for appearing in my Ph.D. qualifying oral exam committee.

My heartfelt thanks to our entire department of statistics and biostatistics and also to all the graduate students in our department. They have all been a big part of my graduate life. I would also especially like to thank our department chair Dr. Regina Liu for being such an amazing and helpful mentor to so many of us. Dr. John Kolassa, our graduate director, has also been very welcoming and forthcoming in offering his help and advice. I would also like to thank Dr. Kesar Singh whom I got to know for two years before his unfortunate and untimely demise. Dr. Singh advised me at length about my research options and touched me immensely with his loving persona in the very short time we had together.

None of it would be possible without the love and sacrifice of my family. No measure of gratitude can capture the debt I owe to my father Dr. Sukamal Mitra for his love and support. My family in Barasat: Ababa, Bumba, Bon, Priyanka Di have always been there for me. I remember Piu, Ma, Bordi for all they did all through

my life and miss them dearly. Last but in no way least, I thank my wife Priyam. Whenever I have felt down or dejected, she has always cheered me up and helped me see what truly matters. Thank you for being my rock.

Dedication

To Baba

and

everyone in Barasat

and

to my dearest wife Priyam

Table of Contents

Abstract ii

Acknowledgements iv

Dedication vi

List of Tables ix

List of Figures x

1. Introduction 1

2. Non-parametric Estimation of Correlation Matrices 5

2.1. Introduction 5

2.2. Background & Preliminary Results 10

2.2.1. Data Model and Notation 10

2.2.2. Nonparametric Estimation of Correlation Matrix 12

2.3. Expected Spectrum Error Bounds 13

2.4. Large Deviation Inequalities 21

2.5. Discussion 25

2.5.1. Tapering Estimate of Bandable Correlation Matrices 25

2.5.2. Principal Component Analysis 28

2.6. Proofs 29

3. Nonparametric Estimation of Sparse Precision Matrices 41

3.1. Introduction 41

3.2. Problem Setup & Main Results 44
3.2.1. Nonparametric Estimates of Correlation Matrices 45
3.2.2. Inversion of Nonparametric Matrices via Scaled Lasso 47
3.3. Scaled Lasso with Nonparametric Correlation Matrix Estimate 52
3.4. Proof of Main Theorem 63

4. Inference for Grouped Variables 66
4.1. Introduction 66
4.2. Group Inference 69
4.2.1. Working assumption based on strong group sparsity 70
4.2.2. Bias correction via relaxed projection 72
4.2.3. An optimization strategy 74
4.2.4. Feasibility of relaxed orthogonal projection for random designs 79
4.2.5. Finding feasible solutions 84
4.3. Mixed Norm Consistency Results 85
4.3.1. Assumptions for fixed design matrix 86
4.3.2. Mixed norm consistency for group Lasso 88
4.3.3. Scaled Group Lasso 91
4.4. Simulation Results 98
4.4.1. Asymptotic test statistic 100
Small group sizes 100
Large group sizes 101

Bibliography 103

List of Tables

4.1. Table summarizing simulation set up for two scenarios along with estimate of scale parameter after 100 replications along with its standard error. The true value of the scale parameter is $\sigma = 1$. 100

List of Figures

4.1. Normal QQ plot for the test statistic for $\widehat{\sigma}$ in (4.3.23) in Theorem 4.5 with $n = 1000, p = \{200, 2000\}, g = 2, s = 8$. The results are produced with 100 replications of the scaled group Lasso. The red dotted line is fitted through 1^{st} and 3^{rd} sample quantile. 99

4.2. Chi square Q-Q plot for the test statistic for $\widehat{\boldsymbol{\mu}}_{G_j}$ with $n = 1000, p = 200, g = 10, s = 40$. The theoretical quantiles were drawn from χ^2_4 random variable. The group being tested has size 4. 101

4.3. Normal QQ plot for the test statistic for $\widehat{\boldsymbol{\mu}}_{G_j}$ with $n = 1000, p = 200, g = 2, s = 40$. Here the group size of the test group is 20. 102

Chapter 1
Introduction

In this thesis we have tackled three problems in high dimensional statistical estimation and inference. The first problem relates to nonparametric estimation of correlation matrices. The second problem relates to utilizing such nonparametric estimates for estimation of precision matrices. The third problem concerns construction of asymptotic tests for groups of variables in high dimensional linear regression problems.

With the advent of modern technological revolution, large amounts of data have become easily available. With the influx of data, the need to extract useful insights from the data has also become paramount. One important problem in developing such insights, relates to the estimation of the correlation structure that exists between a given set of features based on a given sample. This problem is well understood when the underlying distribution of the data is Gaussian. Sample correlation matrices based on Pearson correlation coefficient perform well when number of features is small compared to the number of samples. Even when number of features are larger compared to the sample size, dimension reduction schemes such as sparse Principal Component Analysis (PCA) based on sample correlation matrix perform well. However, for more general non-Gaussian data, sample correlation matrix may fail to be consistent. Nonparametric estimates of correlation matrices based on Kendall's tau and Spearman's rho have been proposed for a more general Gaussian copula model. Accuracy of such nonparametric estimates in various matrix norms have been studied. Among several such choices of matrix norms, spectral norm is of particular interest due to its unique

relevance in understanding the inherent subspace spanned by the data. Dimension reduction procedures like PCA, sparse PCA etc. depend on accuracy of the correlation matrix estimate in spectral norm. In Chapter 2, we take up the study of convergence of nonparametric estimates of correlation matrices in spectral norm. Expected spectrum error bound and a general large deviation bound for the maximum spectral error of a collection of submatrices of a given dimension is established. These results prove that the nonparametric estimates of correlation matrices for the larger class of Gaussian copula models match the sharpest known rates of convergence in spectral norm of sample correlation matrices in Gaussian data models. These results open up the door for application of such nonparametric estimates of correlation matrices in any of a multitude of statistical problems for which accuracy in spectral norm is crucial. As an illustration we show two examples of sparse PCA and estimation of banded correlation matrices via tapering off, where our results establish minimax optimal rates of convergence of the plugged-in nonparametric correlation matrix estimates when the underlying data follows a Gaussian copula model.

Continuing upon our development in Chapter 2, we study the estimation of inverse correlation matrices also called precision matrices in Chapter 3. The precision matrix captures the partial correlation structure of the data. In particular for Gaussian graphical models, precision matrix has an important interpretation: if the off-diagonal element of the precision matrix corresponding to indices (i, j) are zero, then there is no edge between node i and node j. Nonparametric estimation of precision matrices in Gaussian copula models have been studied under the assumption of column-wise sparsity; the so called degree of the precision matrix or the associated graph. Optimal rates of convergence in spectral norm have been established under the assumption of matrix ℓ_1 norm bound on the precision matrix. In Chapter 3 we prove such an optimal rate of convergence under the weaker spectral norm bound on the precision matrix. In constructing such estimates, we employ the scaled Lasso procedure that

was developed in Sun and Zhang [2012a] and was applied for estimation of sparse precision matrices in Gaussian data models in Sun and Zhang [2013].

While Chapters 2, 3 deal with the connected problems of estimation of correlation and precision matrices for high dimensional data, Chapter 4, tackles a different aspect of high dimensional statistics.

Linear regression is one of the most widely used techniques in modeling data of any size. For high dimensional data with a large number of available features but only a small number of it significant, several regularization strategies such as Lasso, SCAD, MCP etc have been developed that have good selection, estimation and prediction power. However, research in developing tests of significance of such estimated variables is still on going. In Zhang and Zhang [2014], the authors developed a de-biasing procedure using relaxed projections for constructing tests for estimates based on Lasso and scaled Lasso. This method is good for constructing inference for individual parameters and these low-dimensional projection estimators for individual coefficients can be directly used to construct efficient confidence regions and p-values for a group of small number of parameters. However, when some inherent grouping in the variable set is available, it might be more prudent to consider the groups of variables together while building a model. Lasso penalty fails to account for the grouping effect and group based regularization procedures like group Lasso etc. are more appropriate for such cases. While group Lasso has been widely studied in terms of group selection, estimation and prediction, direct inferential procedures for groups of variables have not been studied. In Chapter 4, we develop a chi-squared based group inference procedure for groups of variables using the ideas of de-biasing and low dimensional projection in regression with group Lasso. Moreover, we also establish faster rates of convergence of the group Lasso based test statistic to the chi-squared distribution when the true set of parameters satisfy a strong group sparsity condition. This shows the benefit of group sparsity in constructing asymptotic tests for groups

of variables. Our construction provides asymptotic normality of the test statistic even when the group to be tested is large. This provides considerable advantage over current procedures and would enable one to possibly construct test for more general sparse additive models. As part of our construction of tests for grouped variables, we develop a scaled group Lasso procedure and establish oracle inequalities and optimal convergence rates in estimation and prediction. The scaled group Lasso procedure obviates the need for cross-validation in finding the tuning parameter for group Lasso and estimates the scale parameter iteratively as part of the optimization scheme. We also show the benefit of group sparsity in estimation of the scale parameter in terms of faster convergence rates.

Before we dive into these topics, we note that Chapter 2 and Chapter 4 have been developed in Mitra and Zhang [2014a] an d Mitra and Zhang [2014b] which are available on arXiv.

Chapter 2

Non-parametric Estimation of Correlation Matrices

2.1 Introduction

We consider n iid copies $\{\boldsymbol{X}_i : 1 \leq i \leq n\}$ of a d-dimensional Gaussian random vector $(X_1, \ldots, X_d)^T$. We define $\mathbf{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^T \in \mathbb{R}^{n \times d}$. We assume that $\boldsymbol{X}_i$'s are centered and marginally scaled, so that $\mathbb{E}\boldsymbol{X} = \mathbf{0}$ and the correlation matrix is given by $\mathbb{E}\boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ with 1 in the diagonal. In this paper, we work within a high-dimensional 'double asymptotic' setting where $d \wedge n \to \infty$. We assume that instead of $\boldsymbol{X}$, we only observe n iid copies $\boldsymbol{Y}_i, 1 \leq i \leq n$, of the transformed variables

$$(f_1(X_1), \cdots, f_d(X_d))^T$$

where f_i's are unknown but strictly increasing. This is a form of the copula model Sklar [1959] for the distribution of the data. Because $\boldsymbol{X}$ follows a Gaussian distribution, it is a formulation of the Gaussian copula, cf. Bickel et al. [1993] and references therein. A slightly different but equivalent formulation of the Gaussian copula has been referred to as the nonparanormal model Liu et al. [2009]. Let $\mathbf{Y} = (\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n)^T$. Our goal here is to estimate the latent correlation structure $\boldsymbol{\Sigma}$ using the observed data matrix $\mathbf{Y}$.

If we could observe the latent data matrix $\mathbf{X}$, an obvious choice as an estimator would be the sample correlation matrix given by $\widetilde{\boldsymbol{\Sigma}}^s = \mathbf{X}^T\mathbf{X}/n$. It is for this reason

that we refer to the latent $\widetilde{\mathbf{\Sigma}}^s$ as an oracle estimator. It is also clear that $\widetilde{\mathbf{\Sigma}}^s$ is a sufficient statistic for estimating $\mathbf{\Sigma}$ when $\mathbf{X}$ is known. As a consequence, any statistical procedure based on $\mathbf{\Sigma}$ could be summarily described as $g(\widetilde{\mathbf{\Sigma}}^s)$ for some function g. In this respect, $\widetilde{\mathbf{\Sigma}}^s$ possesses great utility as an ideal raw estimate that lends itself to further analysis as the need be.

However, as noted above, we do not observe $\mathbf{X}$ but unknown strictly monotone transformations of columns of it, $\mathbf{Y}$. Thus the sample correlation matrix based on $\mathbf{Y}$, i.e. $\mathbf{Y}^T\mathbf{Y}/n$, is in general inconsistent in estimating the latent correlation structure $\mathbf{\Sigma}$. Two candidate nonparametric estimators in such a scenario are considered in this paper: Kendall's tau developed in Kendall [1938] and Spearman's rank correlation coefficient, developed by Charles Spearman in 1904. These are two widely used nonparametric measures of association. Their properties in fixed dimension have been studied in Kendall [1938, 1948], Kruskal [1958] and many others. More recently, in high-dimensional scenarios, correlation matrix estimators based on these measures have been taken up for study in Liu et al. [2012a] and Xue and Zou [2012] among others.

For the rest of this paper, we call $\widehat{\mathbf{\Sigma}}^\tau$ the correlation matrix estimator based on Kendall's tau and call $\widehat{\mathbf{\Sigma}}^\rho$ the one based on Spearman's rho. It will be interesting to study whether for any statistical procedure, say $g(\widetilde{\mathbf{\Sigma}}^s)$, based on the raw estimate $\widetilde{\mathbf{\Sigma}}^s$, it is possible to provide justification for the use of $g(\widehat{\mathbf{\Sigma}}^\tau)$ or $g(\widehat{\mathbf{\Sigma}}^\rho)$ as a viable replacement. It is however cumbersome to study each individual procedure separately. On the other hand, if g is sufficiently smooth with respect to some matrix norm, it would suffice to study the accuracy of $\widehat{\mathbf{\Sigma}}^\tau$ and $\widehat{\mathbf{\Sigma}}^\rho$ as estimates of $\mathbf{\Sigma}$ in such norms.

A complete description of properties of $\widehat{\mathbf{\Sigma}}^\tau$ and $\widehat{\mathbf{\Sigma}}^\rho$ as estimators of large $\mathbf{\Sigma}$ necessitates the derivation of the distributions of these matrix estimators. It is well known that in the multivariate Gaussian model, $\widetilde{\mathbf{\Sigma}}^s$ follows a Wishart distribution Anderson [1958]. To the contrary, derivation of the distribution of $\widehat{\mathbf{\Sigma}}^\tau$ and $\widetilde{\mathbf{\Sigma}}^s$ seems at

the present moment intractable. On the other hand, analysis of these nonparametric estimators for each individual element of the correlation matrix has been taken upon before. Both Kendall's tau and Spearman's rho are specific instances of U-statistics with bounded kernels. In Hoeffding [1948], the asymptotic normality of these nonparametric estimators for an individual correlation was established. Furthermore, the celebrated Hoeffding [1963] inequality provides large deviation bounds for these estimators as U-statistics with bounded kernels. These results provide tools for studying the concentration of $\widehat{\mathbf{\Sigma}}^{\tau}$ and $\widehat{\mathbf{\Sigma}}^{\rho}$ in the matrix max norm and its applications Liu et al. [2012a], Xue and Zou [2012] and the corresponding Gaussian copula graphical model Liu et al. [2012b].

It is important to note that while estimation accuracy in one specific matrix norm could be more appropriate for a certain set of statistical problems, some other set of problems might require accuracy in a different matrix norm. In this paper we focus on the spectral norm, which is also understood as the ℓ_2 operator norm. Many statistical problems can be studied with error bounds in the spectral norm of estimated correlation matrices. A primary example is the principal component analysis (PCA) since the spectral norm is essential in studying the effects of matrix perturbation on eigenvalues and eigenvectors.

Before beginning the study of convergence of $\widehat{\mathbf{\Sigma}}^{\tau}$ and $\widehat{\mathbf{\Sigma}}^{\rho}$ in the spectral norm, it is worthwhile to note that convergence rate of the latent sample covariance matrix $\widetilde{\mathbf{\Sigma}}^{s}$ in the spectral norm has been studied widely and established in a multitude of literature. A detailed overview and further references can be found in Vershynin [2012] among others. For example, one could derive, from the concentration inequality in Theorem II.13 of Davidson and Szarek [2001], that for $\mathbf{X} \in \mathbb{R}^{n \times d}$ with iid $N(\mathbf{0}, \mathbf{\Sigma})$ rows,

$$\sqrt{\mathbb{E}\|\widetilde{\mathbf{\Sigma}}^{s} - \mathbf{\Sigma}\|_S^2} \leq \|\mathbf{\Sigma}\|_S \left(2\sqrt{2}\sqrt{d/n} + \sqrt{2}d/n + 6(d/n^3)^{1/4}\right), \tag{2.1.1}$$

so that the consistency of $\widetilde{\mathbf{\Sigma}}^s$ follows when $d/n \to 0$. Additionally, the concentration inequality also provides a uniform bound on the spectral error for any s-dimensional diagonal submatrix for larger d. Taking any integer $s < d$ and sets $A \subset \{1, \cdots, d\}$, we have by the union bound

$$\begin{aligned} & \max_{|A| \leq s} \|(\widetilde{\mathbf{\Sigma}}^s - \mathbf{\Sigma})_{A \times A}\|_S \Big/ \max_{|A| \leq s} \|\mathbf{\Sigma}_{A \times A}\|_S \\ \leq & \left(\sqrt{s/n} + \sqrt{2\{t + \log \tbinom{d}{s}\}/n}\right)\left(2 + \sqrt{s/n} + \sqrt{2\{t + \log \tbinom{d}{s}\}/n}\right) \end{aligned} \tag{2.1.2}$$

with at least probability $1 - 2e^{-t}$. These spectral error bounds are explicit and of sharp order for the latent sample correlation matrix estimate $\widetilde{\mathbf{\Sigma}}^s$. In this light, it is apt to ask whether $\widehat{\mathbf{\Sigma}}^\tau$ and $\widehat{\mathbf{\Sigma}}^\rho$ also submit similar error bounds.

In Han and Liu [2013] a rate of $\sqrt{d \log d/n}$ was established for $\widehat{\mathbf{\Sigma}}^\tau$ in a transelliptical family of distributions Liu et al. [2012b]. In a separate but simultaneous work in Wegkamp and Zhao [2013] the same rate was established for $\widehat{\mathbf{\Sigma}}^\tau$ in an elliptical copula correlation factor model, which can be also viewed as elliptical copula. In this paper, we provide non-asymptotic spectrum error bounds in the more restrictive Gaussian copula model for both $\widehat{\mathbf{\Sigma}}^\tau$ and $\widehat{\mathbf{\Sigma}}^\rho$ which improve the convergence rates of these existing error bounds. In particular, we establish in Theorem 2.1 expected spectral error bounds to match (2.1.1), and under mild conditions on the sample size, we establish in Theorem 2.2 and its corollaries large deviation bounds to match (2.1.2). These results establish that in the Gaussian copula model the nonparametric estimators $\widehat{\mathbf{\Sigma}}^\tau$ and $\widehat{\mathbf{\Sigma}}^\rho$ perform as well as the oracle raw estimator $\widetilde{\mathbf{\Sigma}}^s$ in terms of the order of the spectral error. Consequently, a methodology based on $\widetilde{\mathbf{\Sigma}}^s$ that hinges on a spectrum error bound can be performed with the same rate of convergence if $\widehat{\mathbf{\Sigma}}^\tau$ or $\widehat{\mathbf{\Sigma}}^\rho$ are used in lieu of the latent $\widetilde{\mathbf{\Sigma}}^s$.

We discuss two different statistical problems where our results could be applied. The first, a ripe problem for application of spectral error bounds, is the estimation

of a large bandable correlation matrix. For high-dimensional data, proper estimation of large bandable $\mathbf{\Sigma}$ involves implementation of various regularization strategies such as banding, tapering, thresholding etc. These procedures and their properties have been studied in Wu and Pourahmadi [2003], Bickel and Levina [2008a,b], Karoui [2008], Lam and Fan [2009], Cai and Liu [2011], Cai and Zhou [2012], and Cai and Yuan [2012]. In particular, Cai et al. [2010b] established the optimal minimax rate of convergence for a tapered version of $\widetilde{\mathbf{\Sigma}}^s$ for certain classes of unknown bandable $\mathbf{\Sigma}$. In Xue and Zou [2014], a tapering estimator based on the Spearman's rank correlation was studied for the same class of parameters in the Gaussian copula model. However, the question of whether the nonparametric estimator could attain the optimal rate, was not resolved in their paper. Our spectral error bounds imply that the optimal rate is attained if one substitutes $\widetilde{\mathbf{\Sigma}}^s$ with either $\widehat{\mathbf{\Sigma}}^\tau$ or $\widehat{\mathbf{\Sigma}}^\rho$.

The second application involves error bounds in the estimation of the leading eigenvector in PCA both with and without a sparsity assumption on the eigenvector. With the advent and increasing prevalence of high dimensional data, various limitations of traditional procedures had come to the fore. For instance, Johnstone and Lu [2009] showed that when $d/n \to c > 0$, the principal component of $\widetilde{\mathbf{\Sigma}}^s$ is inconsistent in estimating the leading eigenvector of the true correlation matrix. Several remedies to this problem have been proposed, all being different formulations under the auspice of a general sparse PCA paradigm. In sparse PCA, the eigenvectors corresponding to the largest eigenvalues are assumed to be sparse. A vast array of sparse PCA approaches has been proposed and studied in Jolliffe et al. [2003], Zou et al. [2006], d'Aspremont et al. [2007], Vu and Lei [2012], Ma [2013], and Cai et al. [2013] among others. For the elliptical copula family, Han and Liu [2013] established the optimal rate of convergence in sparse PCA with $\widehat{\mathbf{\Sigma}}^\tau$ under an additional sign sub-Gaussian condition. We will demonstrate that our spectral error bounds for the nonparametric estimators can be directly applied to study the convergence rates for the principle

component direction. In particular, for sparse PCA the minimax rate as described in Vu and Lei [2012] will be established without imposing the sign sub-Gaussian condition.

Our work is organized as follows. In Section 2.2 we describe the Gaussian copula model and the Kendall's tau and Spearman's rho estimators for the correlation matrix. In Section 2.3, we provide upper bounds for the expected spectral error for these two correlation-matrix estimators in Theorem 2.1 and outline our analytical strategy. In Section 2.4, we provides a general large deviation inequality in Theorem 2.2. In Section 2.5 we discuss two problems where our results on spectral norm concentration could be utilized. Some of the proofs are relegated to the final section.

2.2 Background & Preliminary Results

We describe the basic data model and define the nonparametric estimates of $\mathbf{\Sigma}$.

2.2.1 Data Model and Notation

We consider the Gaussian copula or multivariate nonparametric transformational model

$$(Y_1, \cdots, Y_d)^T = (f_1(X_1), \cdots, f_d(X_d))^T, \tag{2.2.1}$$

where $(X_1, \cdots, X_d)^T \in \mathbb{R}^d$ is a multivariate Gaussian random vector with marginal $N(0,1)$ distribution and f_j are unknown strictly increasing functions. We are interested in estimating the population correlation matrix of $(X_1, \cdots, X_d)^T$, denoted by

$$\mathbf{\Sigma} = \mathbb{E}(X_1, \cdots, X_d)^T(X_1, \cdots, X_d), \tag{2.2.2}$$

based on a sample of iid copies of $(Y_1, \cdots, Y_d)^T$. Since the f_j absorbs the location and scale of the individual X_j, it is natural to assume $\mathbb{E}X_j = 0$ and $\mathbb{E}X_j^2 = 1$ on the

marginal distribution.

The observations $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{id})^T$, $i = 1, \cdots, n$, are iid copies of $(Y_1, \cdots, Y_d)^T$. They can be written as

$$Y_{ij} = f_j(X_{ij}) \quad i = 1 \cdots, n \quad j = 1, \cdots, d, \tag{2.2.3}$$

where $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{id})^T \in \mathbb{R}^d$ are independent copies of $(X_1, \cdots, X_d)^T \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ in (2.2.1). We denote by $\mathbf{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^T \in \mathbb{R}^{n\times d}$ the matrix with rows $\boldsymbol{X}_i^T$ and quite similarly $\mathbf{Y} = (\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n)^T \in \mathbb{R}^{n\times d}$.

We use the following notation throughout the paper. For vectors $\boldsymbol{u} \in \mathbb{R}^d$, the ℓ_p norm is denoted by $\|\boldsymbol{u}\|_p = \left(\sum_{k=1}^d |u_k|^p\right)^{1/p}$, with $\|\boldsymbol{u}\|_\infty = \max_{1\le k\le d} |u_k|$ and $\|\boldsymbol{u}\|_0 = \#\{j : u_j \neq 0\}$. For matrices $\mathbf{A} = (A_{jk})_{d\times d} \in \mathbb{R}^{d\times d}$, the $\ell_p \to \ell_q$ operator norm is denoted by $\|\mathbf{A}\|_{(p,q)} = \max_{\|\boldsymbol{u}\|_p=1} \|\mathbf{A}\boldsymbol{u}\|_q$. The $\ell_2 \to \ell_2$ operator norm, known as the spectrum norm, is

$$\|\mathbf{A}\|_S = \|\mathbf{A}\|_{(2,2)} = \max_{\|\boldsymbol{u}\|_2=1} |\boldsymbol{u}^T\mathbf{A}\boldsymbol{u}|$$

The vectorized ℓ_∞ and Frobenius norms are denoted by

$$\|\mathbf{A}\|_{\max} = \max_{j,k} |A_{jk}|, \quad \|\mathbf{A}\|_F = \sqrt{\text{trace}\left(\mathbf{A}^T\mathbf{A}\right)}.$$

For symmetric matrices $\mathbf{A}$, the j^{th} eigenpair of $\mathbf{A}$ is denoted by $\lambda_j(\mathbf{A})$ and $\boldsymbol{\theta}_j(\mathbf{A})$, so that $\lambda_1(\mathbf{A}) = \|\mathbf{A}\|_S$ and $\boldsymbol{\theta}_1(\mathbf{A})$ is the leading eigenvector. In addition to $\mathbb{E}$ and $\mathbb{P}$, which denote the expectation and probability measure, we denote by $\mathbb{E}_n$ the average over iid copies of variables in (2.2.3). For example,

$$\mathbb{E}_n h(x_j, x_k) = n^{-1} \sum_{i=1}^n h(X_{ij}, X_{ik}).$$

The relation $a_n = \mathcal{O}(b_n)$ will imply $a_n \leq K b_n$ for some fixed constant $K > 0$. Finally we denote $\mathbb{S}^{d-1} = \{\boldsymbol{u} \in \mathbb{R}^d : \ \|\boldsymbol{u}\|_2 = 1\}$.

2.2.2 Nonparametric Estimation of Correlation Matrix

The approach we adopt in estimating the correlation matrix $\mathbf{\Sigma} = (\Sigma_{jk})$ in (2.2.2) is based on Kendall's tau (τ) or Spearman's correlation coefficient rho (ρ).

With the observations Y_{ij} in (2.2.3), Kendall's tau is defined as

$$\widehat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \leq i_1 < i_2 \leq n} \operatorname{sgn}(Y_{i_1 j} - Y_{i_2 j}) \operatorname{sgn}(Y_{i_1 k} - Y_{i_2 k}), \tag{2.2.4}$$

and Spearman's rho as

$$\widehat{\rho}_{jk} = \frac{\sum_{i=1}^n (r_{ij} - (n+1)/2)(r_{ik} - (n+1)/2)}{\sqrt{\sum_{i=1}^n (r_{ij} - (n+1)/2)^2 \sum_{i=1}^n (r_{ik} - (n+1)/2)^2}}, \tag{2.2.5}$$

where r_{ij} is the rank of Y_{ij} among $Y_{1j}, \cdots, Y_{nj}$. In matrix notation,

$$\widehat{\mathbf{T}} = (\widehat{\tau}_{jk})_{d \times d}, \quad \widehat{\mathbf{R}} = (\widehat{\rho}_{jk})_{d \times d}. \tag{2.2.6}$$

The population version of Kendall's tau is given by

$$\tau_{jk} = \mathbb{E} \operatorname{sgn}(Y_{1j} - Y_{2j}) \operatorname{sgn}(Y_{1k} - Y_{2k}), \tag{2.2.7}$$

while the population version of Spearman's rho is given by

$$\rho_{jk} = 3\, \mathbb{E} \operatorname{sgn}(Y_{1j} - Y_{2j}) \operatorname{sgn}(Y_{1k} - Y_{3k}). \tag{2.2.8}$$

In matrix notation, the population version of (2.2.6) is

$$\mathbf{T} = (\tau_{jk})_{d \times d}, \quad \mathbf{R} = (\rho_{jk})_{d \times d}. \tag{2.2.9}$$

Since f_j are strictly increasing functions, we have $\operatorname{sgn}(f_j(u) - f_j(v)) = \operatorname{sgn}(u - v)$. Thus, Kendall's tau, Spearman's rho and their population version are unchanged if the observed $\boldsymbol{Y} = (Y_{ij})_{n\times d}$ is replaced by the unobserved $\boldsymbol{X} = (X_{ij})_{n\times d}$ in their definition. Since X_j follows a standard normal distribution, we have, from Kendall [1948] and Kruskal [1958], that for $\Sigma_{jk} = \mathbb{E}X_jX_k$,

$$\Sigma_{jk} = \sin\left(\frac{\pi}{2}\tau_{jk}\right) = 2\sin\left(\frac{\pi}{6}\rho_{jk}\right). \tag{2.2.10}$$

This immediately leads to the following correlation matrix estimator by Kendall's tau,

$$\widehat{\boldsymbol{\Sigma}}^{\tau} = (\widehat{\Sigma}^{\tau}_{jk})_{d\times d}, \qquad \widehat{\Sigma}^{\tau}_{jk} = \sin\left(\frac{\pi}{2}\widehat{\tau}_{jk}\right). \tag{2.2.11}$$

In the same light we define the correlation matrix estimator by Spearman's rho as

$$\widehat{\boldsymbol{\Sigma}}^{\rho} = (\widehat{\Sigma}^{\rho}_{jk})_{d\times d}, \qquad \widehat{\Sigma}^{\rho}_{jk} = 2\sin\left(\frac{\pi}{6}\widehat{\rho}_{jk}\right). \tag{2.2.12}$$

The following proposition states a slightly different version of Theorem 2.3 of Wegkamp and Zhao [2013] and a direct application of their argument to Spearman's rho.

Proposition 2.1. *Both matrices* $\boldsymbol{T} - (2/\pi)\boldsymbol{\Sigma}$ *and* $\boldsymbol{R} - (3/\pi)\boldsymbol{\Sigma}$ *are nonnegative-definite,* $\|\boldsymbol{T} - (2/\pi)\boldsymbol{\Sigma}\|_S \le (1 - 2/\pi)\|\boldsymbol{\Sigma}\|_S$, *and* $\|\boldsymbol{R} - (3/\pi)\boldsymbol{\Sigma}\|_S \le (1 - 3/\pi)\|\boldsymbol{\Sigma}\|_S$. *Consequently,*

$$\|\boldsymbol{T}\|_S \vee \|\boldsymbol{R}\|_S \le \|\boldsymbol{\Sigma}\|_S. \tag{2.2.13}$$

2.3 Expected Spectrum Error Bounds

While Spearman's rho and Kendall's tau are structurally different, they can be represented neatly as U-statistics of a special type. In this section we develop bounds

for the expected spectrum norm of their error via a certain decomposition of such U-statistics. This decomposition also provides an outline of our analysis of the concentration of the spectrum norm and the sparse spectrum norm of the error in subsequent sections.

Given a sequence of n observations from a population in $\mathbb{R}^d$, a matrix U-statistic with order m and kernels $h_{jk}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ can be written as

$$\mathbf{U}_n = (U_{n;jk})_{d\times d} \tag{2.3.1}$$

with elements

$$U_{n;jk} = \frac{(n-m)!}{n!} \sum_{1\le i_1 \ne \cdots \ne i_m \le n} h_{jk}(\boldsymbol{X}_{i_1}, \boldsymbol{X}_{i_2}, \cdots, \boldsymbol{X}_{i_m}). \tag{2.3.2}$$

Assume that $h_{jk}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ are permutation symmetric and set

$$\overline{h}_{jk}(\boldsymbol{x}) = \mathbb{E}\left[h_{jk}(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_m)\middle|\boldsymbol{X}_1 = \boldsymbol{x}\right] - c_{jk} \tag{2.3.3}$$

with any constants c_{jk}. The Hoeffding decomposition of $\mathbf{U}_n$ can be written as

$$\mathbf{U}_n - \mathbb{E}\mathbf{U}_n = \sum_{\ell=1}^{m} \binom{m}{\ell} \boldsymbol{\Delta}_n^{(\ell)} \tag{2.3.4}$$

where $\boldsymbol{\Delta}_n^{(1)}$ is an average of iid random matrices with elements

$$\Delta_{n;jk}^{(1)} = (\mathbb{E}_n - \mathbb{E})\overline{h}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\left(\overline{h}_{jk}(\boldsymbol{X}_i) - \mathbb{E}\overline{h}_{jk}(\boldsymbol{X}_1)\right) \tag{2.3.5}$$

and $\boldsymbol{\Delta}_n^{(\ell)} = (\Delta_{n;jk}^{(\ell)})_{d\times d}$ are matrix U-statistics with completely degenerate kernels of order ℓ. We refer to Hoeffding [1948], Hájek et al. [1967], Hájek [1968], van der Vaart [2000] and Serfling [2009] for detailed exposition on the Hoeffding decomposition and

additional references.

Since the components of the Hoeffding decomposition are orthogonal,

$$\mathbb{E}\left(\sum_{\ell=2}^{m}\binom{m}{\ell}\Delta_{n;jk}^{(\ell)}\right)^2 = \sum_{\ell=2}^{m}\binom{n}{\ell}^{-1}\binom{m}{\ell}^3\mathbb{E}\left(\Delta_{m;jk}^{(\ell)}\right)^2$$
$$\leq \binom{n}{2}^{-1}\binom{m}{2}\mathrm{Var}\big(h_{jk}(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_m)\big).$$

A consequence of the above calculation of variance is

$$\mathbb{E}\left\|\mathbf{U}_n - \mathbb{E}\mathbf{U}_n - m\boldsymbol{\Delta}_n^{(1)}\right\|_F^2 \leq \frac{m(m-1)}{n(n-1)}\sum_{j=1}^{d}\sum_{k=1}^{d}\mathrm{Var}\big(h_{jk}(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_m)\big).$$

We note that Kendall's tau and Spearman's rho are U-statistics of order $m = 2$ and 3 respectively, both with kernels satisfying

$$h_{jj}(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_m) = 1 \text{ and } \|h_{jk}(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_m)\|_\infty \leq 1 \text{ for } j \neq k.$$

It follows that the high order terms of their Hoeffding decompositions are explicitly bounded by

$$\mathbb{E}\left\|\mathbf{U}_n - \mathbb{E}\mathbf{U}_n - m\boldsymbol{\Delta}_n^{(1)}\right\|_F^2 \leq \frac{m(m-1)d(d-1)}{n(n-1)}. \tag{2.3.6}$$

Now we consider the term $\boldsymbol{\Delta}_n^{(1)}$. It turns out that in the Gaussian copula model (2.2.3), the first order kernel for Kendall's tau can be written as

$$\overline{h}_{jk}(x_1,\ldots,x_d) = \begin{cases}\overline{h}(x_j,x_k,\Sigma_{jk}), & j \neq k\\ 1 & j = k\end{cases}$$

with $\overline{h}(x_j,x_k,0) = \overline{h}_0(x_j)\overline{h}_0(x_k)$, where $\overline{h}_0(x) = 2\Phi(x) - 1$, and that of Spearman's rho is of the same form. This motivates a further decomposition of $\boldsymbol{\Delta}_n^{(1)}$ as a sum of

$\boldsymbol{\Delta}_n^{(0)}$ and $\boldsymbol{\Delta}_n^{(1)} - \boldsymbol{\Delta}_n^{(0)}$, with

$$\begin{aligned} \boldsymbol{\Delta}_n^{(0)} &= \Big(\Delta_{n;jk}^{(0)}\Big)_{d\times d} = \Big((\mathbb{E}_n - \mathbb{E})\overline{h}_0(x_j)\overline{h}_0(x_k)\Big)_{d\times d}, \\ \boldsymbol{\Delta}_n^{(1)} - \boldsymbol{\Delta}_n^{(0)} &= \Big((\mathbb{E}_n - \mathbb{E})\big(\overline{h}(x_j, x_k, \Sigma_{jk}) - \overline{h}(x_j, x_k, 0)\big)\Big)_{d\times d} \end{aligned} \tag{2.3.7}$$

It follows from the definition of the population Spearman's rho in (2.2.8) that

$$\mathbb{E}\,\overline{h}(X_j, X_k, 0) = \mathbb{E}\,\overline{h}_0(X_j)\overline{h}_0(X_k) = \rho_{jk}/3, \quad \forall 1 \le j \le k \le d.$$

Thus, the $\boldsymbol{\Delta}_n^{(0)}$ in (2.3.7) can be written as the difference between the sample covariance matrix of $\overline{h}_0(\boldsymbol{X}) = (\overline{h}_0(X_{ij}))_{n\times d}$ and its expectation:

$$\boldsymbol{\Delta}_n^{(0)} = n^{-1}\overline{h}_0(\boldsymbol{X})^T\overline{h}_0(\boldsymbol{X}) - \mathbf{R}/3. \tag{2.3.8}$$

Moreover, we will prove that for both Kendall's tau and Spearman's rho

$$\Big|\overline{h}(x_j, x_k, \Sigma_{jk}) - \overline{h}(x_j, x_k, 0)\Big| \le C_1\Big|\Sigma_{jk}\Big|, \ j \ne k. \tag{2.3.9}$$

with $C_1 = 2/\pi + 1 \le 2$ for Kendall's tau and $C_1 \le 1 + \sqrt{8}/\pi \le 2$ for Spearman's rho. Thus, since $\mathrm{Var}(\overline{h}_0^2(X_{ij})) = \int_0^1((2x-1)^2 - 1/3)^2dx = 4/45$ on the diagonal of $\boldsymbol{\Delta}_n^{(1)} - \boldsymbol{\Delta}_n^{(0)}$ and $\boldsymbol{\Delta}_n^{(1)} - \boldsymbol{\Delta}_n^{(0)}$ is an average of iid matrices,

$$\mathbb{E}\Big\|\boldsymbol{\Delta}_n^{(1)} - \boldsymbol{\Delta}_n^{(0)}\Big\|_S^2 \le \mathbb{E}\Big\|\boldsymbol{\Delta}_n^{(1)} - \boldsymbol{\Delta}_n^{(0)}\Big\|_F^2 \le C_1^2\sum_{j\ne k}\frac{\Sigma_{jk}^2}{n} + \frac{4d}{45n}. \tag{2.3.10}$$

Let $\mathbf{U}_n$ be the matrix U-statistics of either Kendall's tau or Spearman's rho, $\mathbf{U}_n = \widehat{\mathbf{T}} = (\widehat{\tau}_{jk})_{d\times d}$ or $\mathbf{U}_n = \widehat{\mathbf{R}} = (\widehat{\rho}_{jk})_{d\times d}$ as in (2.2.6) respectively, and $\widehat{\boldsymbol{\Sigma}}$ the corresponding estimator of $\boldsymbol{\Sigma}$ in (2.2.11) and (2.2.12). It follows from the expansion

of the sine function in (2.2.11) and (2.2.12) that

$$(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})_{jk} \approx a_0(\mathbf{U}_n - \mathbb{E}\mathbf{U}_n)_{jk}, \tag{2.3.11}$$

with $a_0 = \pi/2$ for $\boldsymbol{U}_n = \widehat{\mathbf{T}}$ and $a_0 = \pi/3$ for $\boldsymbol{U}_n = \widehat{\mathbf{R}}$. Thus, the estimators $\widehat{\mathbf{\Sigma}}$ can be decomposed as

$$\begin{aligned} \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma} &= a_0\Big\{(\mathbf{U}_n - \mathbb{E}\mathbf{U}_n) - m\mathbf{\Delta}_n^{(1)}\Big\} + a_0 m\Big(\mathbf{\Delta}_n^{(1)} - \mathbf{\Delta}_n^{(0)}\Big) \\ &\quad + a_0 m\mathbf{\Delta}_n^{(0)} + \Big\{(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}) - a_0(\mathbf{U}_n - \mathbb{E}\mathbf{U}_n)\Big\}, \end{aligned} \tag{2.3.12}$$

where the first two terms are bounded by (2.3.6) and (2.3.10) respectively and the third term is explicitly expressed as the difference between a sample covariance matrix and its expectation in (2.3.7). Moreover, the fourth term can be bounded with a higher order expansion of $\sin(t)$ in (2.2.11) and (2.2.12). We note that the fourth term on the right-hand side of (2.3.12) is not needed if one is interested in studying $\widehat{\mathbf{T}} - \mathbf{T}$ or $\widehat{\mathbf{R}} - \mathbf{R}$ without the sine transformation. This analysis leads to the following theorem.

Theorem 2.1. *Let $\widehat{\boldsymbol{T}}$ and $\widehat{\boldsymbol{R}}$ be respectively the Kendall's tau and Spearman's rho matrices in (2.2.6), $\boldsymbol{T}$ and $\boldsymbol{R}$ be their population version in (2.2.9), and $\widehat{\mathbf{\Sigma}}^{\tau} = (\widehat{\Sigma}^{\tau}_{jk})_{d\times d}$ and $\widehat{\mathbf{\Sigma}}^{\rho} = (\widehat{\Sigma}^{\rho}_{jk})_{d\times d}$ be the corresponding estimators in (2.2.11) and (2.2.12) for the population correlation matrix $\mathbf{\Sigma}$ in the Gaussian copula model (2.2.1). Then, for certain numerical constant C_0 and both $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}^{\tau}$ and $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}^{\rho}$*

$$\begin{aligned} &\mathbb{E}\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_S + \mathbb{E}\|\widehat{\boldsymbol{T}} - \boldsymbol{T}\|_S + \mathbb{E}\|\widehat{\boldsymbol{R}} - \boldsymbol{R}\|_S \\ &\qquad \le C_0\|\mathbf{\Sigma}\|_S\Big(\sqrt{d/n} + d/n\Big). \end{aligned} \tag{2.3.13}$$

In particular, defining $n_2 = 2\lfloor n/2 \rfloor$ *(where* $\lfloor x \rfloor$ *is the integer part of* x*),*

$$\begin{aligned}
\mathbb{E}\|\widehat{\boldsymbol{T}} - \boldsymbol{T}\|_S &\leq \sqrt{2d(d-2n)_+/\{n(n-1)\} + 4(2/\pi+1)^2\|\boldsymbol{\Sigma}\|_F^2/n} \\
&\quad +10\|\boldsymbol{\Sigma}\|_S\Big(\sqrt{(d+1)/(3n)} + (d+1)/n\Big), \qquad (2.3.14) \\
\mathbb{E}\|\widehat{\boldsymbol{\Sigma}}^{\tau} - \boldsymbol{\Sigma}\|_S &\leq \frac{\pi}{2}\mathbb{E}\|\widehat{\boldsymbol{T}} - \boldsymbol{T}\|_S + \frac{\pi}{2}\sqrt{\frac{\|\boldsymbol{\Sigma}\|_F^2 - d}{n_2}} + \frac{\pi^2\sqrt{3}d}{8n_2},
\end{aligned}$$

for Kendall's tau, and for Spearman's rho, with $n_3 = 3\lfloor n/3 \rfloor$

$$\begin{aligned}
\mathbb{E}\|\widehat{\boldsymbol{R}} - \boldsymbol{R}\|_S &\leq \sqrt{6d(d-2n)/\{n(n-1)\} + 9(1+\sqrt{8}/\pi)^2\|\boldsymbol{\Sigma}\|_F^2/n} \\
&\quad +15\|\boldsymbol{\Sigma}\|_S\Big(\sqrt{(d+1)/(3n)} + (d+1)/n\Big) + \|\boldsymbol{\Sigma}\|_F/n, \qquad (2.3.15) \\
\mathbb{E}\|\widehat{\boldsymbol{\Sigma}}^{\rho} - \boldsymbol{\Sigma}\|_S &\leq \frac{\pi}{3}\mathbb{E}\|\widehat{\boldsymbol{R}} - \boldsymbol{R}\|_S + \frac{\pi}{9}\sqrt{\frac{\|\boldsymbol{\Sigma}\|_F^2 - d}{n_3}} + \frac{\pi^2\sqrt{3}d}{36n_3} + \frac{2\pi\|\boldsymbol{\Sigma}\|_F}{3n}.
\end{aligned}$$

Corollary 2.1. *If* $\|\boldsymbol{\Sigma}\|_S^2 d/n \to 0$*, then*

$$\mathbb{E}\|\widehat{\boldsymbol{T}} - \boldsymbol{T}\|_S + \mathbb{E}\|\widehat{\boldsymbol{\Sigma}}^{\tau} - \boldsymbol{\Sigma}\|_S + \mathbb{E}\|\widehat{\boldsymbol{R}} - \boldsymbol{R}\|_S + \mathbb{E}\|\widehat{\boldsymbol{\Sigma}}^{\rho} - \boldsymbol{\Sigma}\|_S \to 0.$$

Remark 2.1. Up to a numerical constant factor, Theorem 2.1 match the bound (2.1.1) for the expected spectral error of the oracle sample covariance matrix $\widetilde{\boldsymbol{\Sigma}}^s$. While Han and Liu [2013] and Wegkamp and Zhao [2013] focused on large deviation bound of the spectral error of $\|\widehat{\boldsymbol{\Sigma}}^{\tau} - \boldsymbol{\Sigma}\|_S$ in the elliptical copula model, a direct application of their results requires $\|\boldsymbol{\Sigma}\|_S d(\log d)/n \to 0$ for the convergence in spectrum norm. Although their results are of sharper order when $\|\boldsymbol{\Sigma}\|_S \gg \log d$, it seems that when $\|\boldsymbol{\Sigma}\|_S = \mathcal{O}(1)$, the extra logarithmic factor cannot be removed in their analysis based on the matrix Bernstein inequality Tropp [2011].

The proof of Theorem 2.1 requires a number of inequalities which provide key details of the analysis outlined above the statement of the theorem. These inequalities are crucial for our derivation of large deviation spectrum error bounds as well. We

state these inequalities in a sequence of lemmas below and defer their proofs to the Appendix.

Let $\varphi_\rho(x, y)$ be the bivariate normal density with mean zero, variance one, and correlation ρ. Define

$$\overline{h}(x, y, \rho) = \int\int \text{sgn}(x - u)\text{sgn}(y - v)\varphi_\rho(u, v)dudv. \tag{2.3.16}$$

Lemma 2.1. *Let $\overline{h}(x, y, \rho)$ be as in (2.3.16). Based on $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ with iid $N(0, \boldsymbol{\Sigma})$ rows, Kendall's $\widehat{\tau}_{jk}$ is a U-statistic of order 2 with a permutation symmetric kernel $h_{j,k}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfying $|h_{jk}(\boldsymbol{x}_1, \boldsymbol{x}_2)| = 1$ and*

$$\mathbb{E}\Big[h_{jk}(\boldsymbol{X}_1, \boldsymbol{X}_2)\Big|\boldsymbol{X}_1 = \boldsymbol{x}\Big] = \overline{h}(x_j, x_k, \Sigma_{jk}) \ \forall\ j \neq k. \tag{2.3.17}$$

With $g(x, y, \rho) = \overline{h}(x, y, \rho) - \overline{h}(x, y, 0)$ and $C_1 = 2/\pi + 1$,

$$\big|g(x, y, \rho)\big| \leq C_1|\rho|, \ \ \big|(\partial/\partial x)g(x, y, \rho)\big| \leq |\rho|. \tag{2.3.18}$$

Moreover, with $\overline{h}_0(x) = 2\Phi(x) - 1$ and ρ_{jk} in (2.2.8),

$$\overline{h}(x, y, 0) = \overline{h}_0(x)\overline{h}_0(y), \ \ \mathbb{E}\,\overline{h}(X_{ij}, X_{ik}, 0) = \rho_{jk}/3 \ \forall\ j, k. \tag{2.3.19}$$

Lemma 2.2. *Let $\overline{h}(x, y, \rho)$ be as in (2.3.16) and $C_1 = \sqrt{8}/\pi + 1$. Based on $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ with iid $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ rows, Spearman's $\widehat{\rho}_{jk}$ is a U-statistic of order 3 with a permutation*

symmetric kernel $h^{\rho}_{j,k}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$ *satisfying*

$$|\mathbb{E}\widehat{\rho}_{jk} - \rho_{jk}| \le |\rho_{jk}|/(n+1) \le |\Sigma_{jk}|/(n+1), \tag{2.3.20}$$

$$|h^{\rho}_{jk}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)| \le 1, \tag{2.3.21}$$

$$\left|\overline{h}^{\rho}(x, y, \rho) - \overline{h}^{\rho}(x, y, 0)\right| \le C_1|\rho|, \tag{2.3.22}$$

$$\left|(\partial/\partial x)\left\{\overline{h}^{\rho}(x, y, \rho) - \overline{h}^{\rho}(x, y, 0)\right\}\right| \le |\rho|, \tag{2.3.23}$$

$$(1 + 1/n)\overline{h}^{\rho}(x, y, 0) = \overline{h}(x, y, 0), \tag{2.3.24}$$

where $\overline{h}^{\rho}(x_j, x_k, \Sigma_{jk}) = \mathbb{E}\left[h^{\rho}_{jk}(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3)\middle|\boldsymbol{X}_1 = \boldsymbol{x}\right] - \tau_{jk}/(n+1)$.

Lemma 2.3. *Inequalities (2.3.6) and (2.3.10) hold with* $C_1 = 2/\pi + 1 \le 2$ *and* $m = 2$ *for Kendall's tau and* $C_1 \le 1 + \sqrt{8}/\pi \le 2$ *and* $m = 3$ *for Spearman's rho. Moreover, for both Kendall's tau and Spearman's rho,*

$$\mathbb{E}\|(\boldsymbol{U}_n - \mathbb{E}\,\boldsymbol{U}_n) - m\boldsymbol{\Delta}^{(0)}_n\|^2_F \le \frac{m(m-1)d(d-2n)_+}{n(n-1)} + C_1^2\frac{\|\boldsymbol{\Sigma}\|^2_F}{n/m^2}. \tag{2.3.25}$$

Lemma 2.4. *Let* $\boldsymbol{\Delta}^{(0)}_n$ *as in (2.3.7) and* $\boldsymbol{R} = (\rho_{jk})_{d\times d}$. *Then,*

$$\mathbb{E}\|\boldsymbol{\Delta}^{(0)}_n\|_S \le 5\|\boldsymbol{\Sigma}\|_S\left(\sqrt{(d+1)/(3n)} + (d+1)/n\right) \tag{2.3.26}$$

and with at least probability $1 - 2e^{-t^2}$,

$$\|\boldsymbol{\Delta}^{(0)}_n\|_S \;\le\; 5\|\boldsymbol{\Sigma}\|_S\left(\sqrt{(d + t^2/\pi)/(3n)} + (d + (t^2+1)/\pi)/n\right). \tag{2.3.27}$$

Lemma 2.5. *(i) Let* $\widehat{\boldsymbol{\Sigma}}^{\tau} = (\widehat{\Sigma}^{\tau}_{jk})_{d\times d}$ *be as in (2.2.11) and* $\boldsymbol{\Delta}^{\tau} = (\Delta^{\tau}_{jk})_{d\times d}$ *with* $\Delta^{\tau}_{jk} = \widehat{\tau}_{jk} - \tau_{jk}$. *Let* $n_2 = 2\lfloor n/2\rfloor$ *where* $\lfloor x\rfloor$ *is the integer part of* x. *Then,*

$$\sqrt{\mathbb{E}\|(\widehat{\boldsymbol{\Sigma}}^{\tau} - \boldsymbol{\Sigma}) - (\pi/2)\boldsymbol{\Delta}^{\tau}\|^2_F} \le \frac{\pi}{2}\sqrt{\frac{\|\boldsymbol{\Sigma}\|^2_F - d}{n_2}} + \frac{\pi^2\sqrt{3}d}{8n_2}. \tag{2.3.28}$$

(ii) Let $\widehat{\mathbf{\Sigma}}^{\rho} = (\widehat{\Sigma}^{\rho}_{jk})_{d\times d}$ *be as in (2.2.12) and* $\mathbf{\Delta}^{\rho} = (\Delta^{\rho}_{jk})_{d\times d}$ *with* $\Delta^{\rho}_{jk} = \widehat{\rho}_{jk} - \mathbb{E}\widehat{\rho}_{jk}$. *Let* $n_3 = 3\lfloor n/3 \rfloor$ *where* $\lfloor x \rfloor$ *is the integer part of* x. *Then,*

$$\sqrt{\mathbb{E}\|(\widehat{\mathbf{\Sigma}}^{\rho} - \mathbf{\Sigma}) - (\pi/2)\mathbf{\Delta}^{\rho}\|_F^2} \le \frac{\pi}{9}\sqrt{\frac{\|\mathbf{\Sigma}\|_F^2 - d}{n_3}} + \frac{\pi^2\sqrt{3}d}{36n_3} + \frac{\pi\sqrt{\|\mathbf{\Sigma}\|_F^2 - d}}{3(n+1)} \quad (2.3.29)$$

Proof of Theorem 2.1. Let $n_m = m\lfloor n/m \rfloor$. As in (2.3.12), for Kendall's tau,

$$\begin{aligned} \|\widehat{\mathbf{T}} - \mathbf{T}\|_S &\le \left\|(\mathbf{U}_n - \mathbb{E}\mathbf{U}_n) - 2\mathbf{\Delta}_n^{(0)}\right\|_F + 2\left\|\mathbf{\Delta}_n^{(0)}\right\|_S, \\ \|\widehat{\mathbf{\Sigma}}^{\tau} - \mathbf{\Sigma}\|_S &\le \left\|(\widehat{\mathbf{\Sigma}}^{\tau} - \mathbf{\Sigma}) - (\pi/2)(\mathbf{U}_n - \mathbb{E}\mathbf{U}_n)\right\|_F + (\pi/2)\left\|\widehat{\mathbf{T}} - \mathbf{T}\right\|_S. \end{aligned}$$

with $\mathbf{U}_n = \widehat{\mathbf{T}}$ and $\mathbb{E}\mathbf{U}_n = \mathbf{T}$. It follows from (2.3.25) of Lemma 2.3 with $m = 2$, (2.3.26) of Lemma 2.4 and (2.3.28) of Lemma 2.5 that the inequalities in (2.3.14) hold.

Similarly, for Spearman's rho,

$$\begin{aligned} \|\widehat{\mathbf{R}} - \mathbf{R}\|_S &\le \left\|(\mathbf{U}_n - \mathbb{E}\mathbf{U}_n) - 3\mathbf{\Delta}_n^{(0)}\right\|_F + 3\left\|\mathbf{\Delta}_n^{(0)}\right\|_S + \left\|\mathbb{E}\mathbf{U}_n - \mathbf{R}\right\|_F, \\ \|\widehat{\mathbf{\Sigma}}^{\rho} - \mathbf{\Sigma}\|_S &\le \left\|(\widehat{\mathbf{\Sigma}}^{\rho} - \mathbf{\Sigma}) - (\pi/3)(\widehat{\mathbf{R}} - \mathbf{R})\right\|_F + (\pi/3)\|\widehat{\mathbf{R}} - \mathbf{R}\|_S, \end{aligned}$$

with $\mathbf{U}_n = \widehat{\mathbf{R}}$ and $\|\mathbb{E}\mathbf{U}_n - \mathbf{R}\|_F = \|(\mathbb{E}\widehat{\rho}_{jk} - \rho_{jk})_{d\times d}\|_F \le \sqrt{\|\mathbf{\Sigma}\|_F^2 - d}/(n+1)$ by (2.3.20). Thus, (2.3.25), (2.3.26) and (2.3.29) yield the inequalities in (2.3.15). ■

2.4 Large Deviation Inequalities

While the upper bounds for the expected spectral error in Theorem 2.1 and Corollary 2.1 match (2.1.1) for the oracle sample covariance matrix, it is useful only when $d/n \to 0$ as is the case in many applications. For $d > n$, large deviation bounds for the sparse spectral norm of the form (2.1.2) is often used instead. In the present section we provide large deviation inequalities for both the spectral norm and the

sparse spectral norm of the error for Kendall's tau and Spearman's rho.

The main result for this section is a large deviation bound in the following theorem for the maximum spectral error in a collection of diagonal submatrices.

Theorem 2.2. *Let $\widehat{\boldsymbol{T}}$ and $\widehat{\boldsymbol{R}}$ be respectively the Kendall's tau and Spearman's rho matrices in (2.2.6), $\boldsymbol{T}$ and $\boldsymbol{R}$ be their population version in (2.2.9), and $\widehat{\boldsymbol{\Sigma}}^{\tau} = (\widehat{\Sigma}^{\tau}_{jk})_{d\times d}$ and $\widehat{\boldsymbol{\Sigma}}^{\rho} = (\widehat{\Sigma}^{\rho}_{jk})_{d\times d}$ be the corresponding estimators in (2.2.11) and (2.2.12) for the population correlation matrix $\boldsymbol{\Sigma}$ in the Gaussian copula model (2.2.1). Let $1 \le s \le d$, $m \ge 1$ and $\mathscr{A}_{s,m}$ be a collection of m subsets $A \subset \{1, 2, \cdots, d\}$ with $|A| \le s$. Then, there exists a certain numerical constant C such that for both $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}^{\tau}$ and $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}^{\rho}$,*

$$\begin{aligned} & \|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{A\times A}\|_S + \|(\widehat{\boldsymbol{T}} - \boldsymbol{T})_{A\times A}\|_S + \|(\widehat{\boldsymbol{R}} - \boldsymbol{R})_{A\times A}\|_S \\ \le\ & C\|\boldsymbol{\Sigma}_{A\times A}\|_S\Big(\sqrt{(s + t + \log m)/n} + (s + t + \log m)/n\Big) \\ & + C\|\boldsymbol{\Sigma}_{A\times A}\|_{(2,\infty)}\|\boldsymbol{\Sigma}_{A\times A}\|_S^{1/2}\sqrt{(t + \log m)/n} + Cs(\log d + t)/n \end{aligned} \tag{2.4.1}$$

simultaneously for all $A \in \mathscr{A}_{s,m}$ with at least probability $1 - e^{-t}$.

Corollary 2.2. *If $t + \log d \le \beta \max\big\{\log(ed/s), \sqrt{(n/s)(t/s + \log(ed/s))}\big\}$, then for both $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}^{\tau}$ and $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}^{\rho}$ and a certain numerical constant C,*

$$\begin{aligned} & \max_{|A|\le s} \frac{\|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{A\times A}\|_S + \|(\widehat{\boldsymbol{T}} - \boldsymbol{T})_{A\times A}\|_S + \|(\widehat{\boldsymbol{R}} - \boldsymbol{R})_{A\times A}\|_S}{\|\boldsymbol{\Sigma}_{A\times A}\|_S + \|\boldsymbol{\Sigma}_{A\times A}\|_S^{1/2}\|\boldsymbol{\Sigma}_{A\times A}\|_{(2,\infty)}} \\ \le\ & C(1+\beta)\Big(\sqrt{(t + s\log(ed/s))/n} + (t + s\log(ed/s))/n\Big) \end{aligned} \tag{2.4.2}$$

with at least probability $1 - e^{-t}$.

Remark 2.2. Corollary 2.2 illustrates that for $\max_{|A|\le s}\|\boldsymbol{\Sigma}_{A\times A}\|_S = \mathcal{O}(1)$ and under a mild condition on (n, d, s), Theorem 2.2 yields a sparse spectral error bound that matches (2.1.2) of the latent $\widetilde{\boldsymbol{\Sigma}}^s$. Note that $\|\boldsymbol{\Sigma}_{A\times A}\|_{(2,\infty)} \le \|\boldsymbol{\Sigma}_{A\times A}\|_S$. In comparison, the spectral error bounds in Han and Liu [2013] and Wegkamp and Zhao [2013], which apply to the elliptical copula family, leads to $\max_{|A|\le s}\|(\widehat{\boldsymbol{\Sigma}}^{\tau} - \boldsymbol{\Sigma})_{A\times A}\|_S =$

$\mathcal{O}(s\sqrt{(\log d)/n})$ by the union bound. Han and Liu [2013] provided a concentration inequality of order $\sqrt{s(\log d)/n}$ for $\widehat{\mathbf{\Sigma}}^{\tau}$ in the transelliptical family under an additional 'sign sub-Gaussian' condition. They also provide two examples of elliptical copulas that satisfy the sign sub-Gaussian condition. The first example is the case of elliptical copulas with the latent correlation $\mathbf{\Sigma}$ satisfying a compound symmetric structure (i.e. $\Sigma_{jk} = \rho$ for all $j \neq k$). The second example is the case when $\mathbf{\Sigma}$ has a diagonal block structure with each diagonal block having a compound symmetric structure. However, it is unclear if the sign sub-Gaussian condition is readily verifiable in general. Theorem 2.2 and Corollary 2.2 establish the concentration of the nonparametric estimates for the Gaussian copula model without the sign sub-Gaussianity condition, although the Gaussian copula family is smaller than the transelliptical family.

The corollary below states a simpler but slightly weaker version of Theorem 2.2 for $s = d$. It matches (2.1.2) for $s = d$ when $\|\mathbf{\Sigma}\|_S = \mathcal{O}(1)$ and $t + \log d = \mathcal{O}(\sqrt{n/d})$.

Corollary 2.3. *For a certain numerical constant C,*

$$\begin{aligned} \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_S \;\leq\; & C\|\mathbf{\Sigma}\|_S\Big(\sqrt{(t+d)/n} + (t+d)/n\Big) \\ & + C\|\mathbf{\Sigma}\|_S^{1/2}\|\mathbf{\Sigma}\|_{(2,\infty)}\sqrt{t/n} + C(t+\log d)d/n \end{aligned} \tag{2.4.3}$$

with at least probability $1 - e^{-t}$ for both $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}^{\tau}$ and $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}^{\rho}$.

The proof of Theorem 2.2 is carried out by establishing large deviation inequalities for the first two terms in the decomposition in (2.3.12), an application of Lemma 2.4 to the third, and an application of an inequality of Wegkamp and Zhao [2013] to the fourth.

Lemma 2.6. *Let us take $C_1 = 2/\pi + 1 \leq 2$ for Kendall's tau and $C_1 \leq 1 + \sqrt{8}/\pi \leq 2$ for Spearman's rho. For both Kendall's tau and Spearman's rho,*

$$\|\mathbf{\Delta}_n^{(1)} - \mathbf{\Delta}_n^{(0)}\|_S \leq \sqrt{\frac{C_1^2\|\mathbf{\Sigma}\|_F^2 - 2d}{n}} + 2\sqrt{2}\|\mathbf{\Sigma}\|_{(2,\infty)}\|\mathbf{\Sigma}\|_S^{1/2}\sqrt{\frac{t}{n}} \tag{2.4.4}$$

with at least probability $1 - e^{-t}$.

Lemma 2.7. *Let* $\boldsymbol{U}_n - \mathbb{E}\,\boldsymbol{U}_n - m\boldsymbol{\Delta}_n^{(1)}$ *be as in (2.3.12). Then, with probability at least* $1 - e^{-t}$, *for a certain constant* $C > 0$,

$$\max_{|A|\le s} \|\big(\boldsymbol{U}_n - \mathbb{E}\,\boldsymbol{U}_n - m\boldsymbol{\Delta}_n^{(1)}\big)_{A\times A}\|_S \le Cs(\log d + t)/n.$$

We state an inequality of Wegkamp and Zhao [2013] in Lemma 2.8 (i) below and its extension to Spearman's rho in Lemma 2.8 (ii).

Lemma 2.8. *(i) Let* $\widehat{\boldsymbol{\Sigma}}^{\tau} = (\widehat{\Sigma}^{\tau}_{jk})_{d\times d}$ *be as in (2.2.11) and* $\boldsymbol{\Delta}^{\tau} = (\Delta^{\tau}_{jk})_{d\times d}$ *with* $\Delta^{\tau}_{jk} = \widehat{\tau}_{jk} - \tau_{jk}$. *Let* $n_2 = 2\lfloor n/2\rfloor$ *where* $\lfloor x\rfloor$ *is the integer part of* x. *Then,*

$$\big\|(\widehat{\boldsymbol{\Sigma}}^{\tau} - \boldsymbol{\Sigma})_{A\times A}\big\|_S \le \pi\big\|(\widehat{\boldsymbol{T}} - \boldsymbol{T})_{A\times A}\big\|_S + \frac{s\pi^2}{8}\big\|\boldsymbol{\Delta}^{\tau}\big\|^2_{\max}, \tag{2.4.5}$$

with $\mathbb{P}\big\{\big\|\boldsymbol{\Delta}^{\tau}\big\|_{\max} > 2t\big\} \le d^2e^{-n_2t^2}$ *for all* $t > 0$.
(ii) Let $\widehat{\boldsymbol{\Sigma}}^{\rho} = (\widehat{\Sigma}^{\rho}_{jk})_{d\times d}$ *be as in (2.2.12) and* $\boldsymbol{\Delta}^{\rho} = (\Delta^{\rho}_{jk})_{d\times d}$ *with* $\Delta^{\rho}_{jk} = \widehat{\rho}_{jk} - \mathbb{E}\widehat{\rho}_{jk}$. *Let* $n_3 = 3\lfloor n/3\rfloor$ *where* $\lfloor x\rfloor$ *is the integer part of* x. *Then,*

$$\big\|(\widehat{\boldsymbol{\Sigma}}^{\rho} - \boldsymbol{\Sigma})_{A\times A}\big\|_S \le C_2\big\|(\widehat{\boldsymbol{T}} - \boldsymbol{T})_{A\times A}\big\|_S + \frac{s\pi^2}{36}\big\|\boldsymbol{\Delta}^{\rho}\big\|^2_{\max} + \frac{\pi s^{1/2}\|\boldsymbol{\Sigma}_{A\times A}\|_{(2,\infty)}}{3(n+1)} \tag{2.4.6}$$

with $C_2 = (\pi/3)(2 - \sqrt{1 - 1/4}) < 1.2$, *and* $\mathbb{P}\big\{\big\|\boldsymbol{\Delta}^{\rho}\big\|_{\max} > \sqrt{6}t\big\} \le d^2e^{-n_3t^2}$ *for all* $t > 0$.

Proof of Theorem 2.2. We consider only $\widehat{\boldsymbol{\Sigma}}^{\tau}$ as the case for $\widehat{\boldsymbol{\Sigma}}^{\rho}$ is nearly identical. It follows from Lemma 2.8 that

$$\big\|(\widehat{\boldsymbol{\Sigma}}^{\tau} - \boldsymbol{\Sigma})_{A\times A}\big\|_S \le \pi\big\|(\widehat{\mathbf{T}} - \mathbf{T})_{A\times A}\big\|_S + Cs(t + \log d)/n, \quad \forall\ |A| \le s \tag{2.4.7}$$

with at least probability $1 - e^{-t}$. As in the decomposition in (2.3.12),

$$\widehat{\mathbf{T}} - \mathbf{T} = \left\{ \mathbf{\Delta}^{\tau} - 2\mathbf{\Delta}_n^{(1)} \right\} + 2 \left\{ \mathbf{\Delta}_n^{(1)} - \mathbf{\Delta}_n^{(0)} \right\} + 2\mathbf{\Delta}_n^{(0)}. \tag{2.4.8}$$

It follows from Lemma 2.7 that with at least probability $1 - e^{-t}$,

$$\max_{|A| \le s} \left\| \left\{ \mathbf{\Delta}^{\tau} - 2\mathbf{\Delta}_n^{(1)} \right\}_{A \times A} \right\|_S \le Cs(\log d + t)/n. \tag{2.4.9}$$

By applying Lemma 2.6 to the m sub-matrices with the union bound,

$$\begin{aligned} \|(\mathbf{\Delta}_n^{(1)} - \mathbf{\Delta}_n^{(0)})_{A \times A}\|_S \le C\|\mathbf{\Sigma}_{A \times A}\|_F / \sqrt{n} \\ + C\|\mathbf{\Sigma}_{A \times A}\|_{(2,\infty)} \|\mathbf{\Sigma}_{A \times A}\|_S^{1/2} \sqrt{(t + \log m)/n}, \quad \forall\ A \in \mathscr{A}_{s,m}, \end{aligned} \tag{2.4.10}$$

with at least probability $1 - m\exp(-t - \log m) \ge 1 - e^{-t}$. Similarly, Lemma 2.4 yields

$$\begin{aligned} \|(\mathbf{\Delta}_n^{(0)})_{A \times A}\|_S \ &\le\ C\|\mathbf{\Sigma}_{A \times A}\|_S \sqrt{(s + t + \log m)/n} \\ &\quad + C\|\mathbf{\Sigma}_{A \times A}\|_S (s + t + \log m)/n, \quad \forall\ A \in \mathscr{A}_{s,m} \end{aligned} \tag{2.4.11}$$

with at least probability $1 - e^{-t}$. The first term in (2.4.10) is dominated by the first term in (2.4.11) due to $\|\mathbf{\Sigma}_{A \times A}\|_F \le \sqrt{s}\|\mathbf{\Sigma}_{A \times A}\|_S$. Thus, applying (2.4.9), (2.4.10) and (2.4.11) to (2.4.8) yields (2.4.1) via (2.4.7). ■

2.5 Discussion

We describe two applications of our concentration inequality in the $d > n$ case.

2.5.1 Tapering Estimate of Bandable Correlation Matrices

We consider the Gaussian copula model in (2.2.1). We assume that the correlation matrix has a bandable structure in that the off-diagonal elements fall off to zero as we

move further away from diagonal. There are several formulations of such bandability. As in Cai et al. [2010b], we consider the parameter class

$$\mathscr{F}_\alpha(M_0, M_1) = \left\{ \mathbf{\Sigma} : \max_j \sum_{|i-j|>k} |\Sigma_{ij}| \le M_0 k^{-\alpha} \ \forall k, \|\mathbf{\Sigma}\|_S \le M_1 \right\}. \tag{2.5.1}$$

We adopt the estimator of Cai et al. [2010b] and plug in $\widehat{\mathbf{\Sigma}}^\tau$ and $\widehat{\mathbf{\Sigma}}^\rho$:

$$\widehat{\mathbf{\Sigma}}_{(k)}^{\tau-taper} = (w_{ij}\widehat{\Sigma}_{ij}^\tau)_{d\times d} \quad \widehat{\mathbf{\Sigma}}_{(k)}^{\rho-taper} = (w_{ij}\widehat{\Sigma}_{ij}^\rho)_{d\times d} \tag{2.5.2}$$

where w_{ij}'s are defined as

$$w_{ij} = \begin{cases} 1 & \text{when } |i-j| \le k/2 \\ 2 - 2\dfrac{|i-j|}{k} & \text{when } k/2 < |i-j| < k \\ 0 & \text{otherwise} \end{cases}$$

The nonparametric tapering estimator $\widehat{\mathbf{\Sigma}}_{(k)}^{\rho-taper}$ has been considered previously in Xue and Zou [2014], where an error bound

$$\sup_{\mathbf{\Sigma}\in\mathscr{F}_\alpha(M_0,M_1)} \mathbb{E}_{\mathbf{\Sigma}} \left\| \widehat{\mathbf{\Sigma}}_{(k)}^{\rho-taper} - \mathbf{\Sigma} \right\|_S^2 \le C_{M_0,M_1} \left(\frac{k^2 \log d}{n} + k^{-2\alpha} \right)$$

was established using a generalization of McDiarmid's inequality, where $\mathbb{E}_{\mathbf{\Sigma}}$ is the expectation in the Gaussian copula model (2.2.1) with correlation $\mathbf{\Sigma}$ in (2.2.2), and C_{M_0,M_1} is a constant depending on M_0 and M_1 only. It was mentioned in their paper that the above error bound may not be sharp as some key concentration inequalities were not available for rank-based estimators. Such key concentration inequalities are provided in Theorem 2.2 as the rate-optimal error bound in the following theorem demonstrates.

Theorem 2.3. *Let $\mathbb{E}_{\mathbf{\Sigma}}$ be the expectation under which (2.2.1) and (2.2.2) hold. Consider the tapered estimators $\widehat{\mathbf{\Sigma}}_{(k)} = \widehat{\mathbf{\Sigma}}_{(k)}^{\tau-taper}$ or $\widehat{\mathbf{\Sigma}}_{(k)} = \widehat{\mathbf{\Sigma}}_{(k)}^{\rho-taper}$ given in (2.5.2). Then,*

$$\sup_{\mathbf{\Sigma}\in\mathscr{F}_\alpha(M_0,M_1)} \mathbb{E}_{\mathbf{\Sigma}}\left\|\widehat{\mathbf{\Sigma}}_{(k)} - \mathbf{\Sigma}\right\|_S^2 \le C_{M_0,M_1}\Big(\frac{k+\log d}{n} + \frac{k^2(\log d)^2}{n^2} + k^{-2\alpha}\Big) \quad (2.5.3)$$

for all $1 \le k \le n$, where C_{M_0,M_1} is a constant depending on M_0 and M_1 only. In particular, for $k = \min\left(n^{1/(2\alpha+1)}, d\right)$ and $\log d \le \beta n^{\alpha/(1+2\alpha)}$,

$$\sup_{\mathbf{\Sigma}\in\mathscr{F}_\alpha(M_0,M_1)} \mathbb{E}_{\mathbf{\Sigma}}\left\|\widehat{\mathbf{\Sigma}}_{(k)} - \mathbf{\Sigma}\right\|_S^2 \le C_{M_0,M_1}(1+\beta)\min\Big(n^{\frac{-2\alpha}{1+2\alpha}} + \frac{\log d}{n}, \frac{d}{n}\Big) \quad (2.5.4)$$

The rate-optimality of (2.5.4) was proved in Cai et al. [2010b] and a combination of their analysis and Theorem 2.2 proves Theorem 2.3. For $\mathbf{H} = (H_{ij})_{d\times d} = \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$,

$$(w_{ij}H_{ij})_{d\times d} = k^{-1}\sum_{\ell=1}^{d+2k-1} \mathbf{H}_{A_\ell\times A_\ell} - k^{-1}\sum_{\ell=1}^{d+k-1} \mathbf{H}_{B_\ell\times B_\ell}$$

where $A_\ell = \{1\vee(\ell-2k),\ldots,\ell\}$ for $1 \le \ell < p+2k$ and $B_\ell = \{1\vee(\ell-k),\ldots,\ell\}$ for $1 \le \ell < p+k$. Let $A_{d+2k+\ell-1} = B_\ell$. Since $\{\mathbf{H}_{A_{\ell+2jk}\times A_{\ell+2jk}}, \ell+2jk < d+2k\}$ are disjoint diagonal blocks for $\ell = 1,\ldots,2k$ and $\{\mathbf{H}_{A_{\ell+jk}\times A_{\ell+jk}}, \ell+jk \ge d+2k\}$ are disjoint diagonal blocks for $\ell = 1,\ldots,k$,

$$\left\|(w_{ij}\widehat{\Sigma}_{ij})_{d\times d} - \mathbf{\Sigma}\right\|_S \le \left\|\big((1-w_{ij})\Sigma_{ij}\big)_{d\times d}\right\|_S + 3\max_{\ell\le 2d+3k-2}\left\|\mathbf{H}_{A_\ell\times A_\ell}\right\|_S$$

with $|A_\ell| \le 2k$. Since $w_{ij} = 0$ for $|i-j| \le k$, the first term above is bounded by $M_0k^{-\alpha}$ in the class. It follows from Theorem 2.2 that the second term above is bounded by

$$\mathbb{E}_{\mathbf{\Sigma}}\max_{\ell\le 2d+3k-2}\left\|\mathbf{H}_{A_\ell\times A_\ell}\right\|_S^2 \le C_{M_0,M_1}\int_0^\infty \Big(\frac{k+t+\log d}{n} + \frac{k^2(\log d+t)^2}{n^2}\Big)e^{-t}dt,$$

which implies (2.5.3).

Although the estimator in (2.5.2) is not adaptive due to the requirement of k as an input, this example demonstrates the utility of our results when Kendall's tau and Spearman's rho are used in place of the oracle sample covariance matrix. Based on the availability of the latent sample covariance matrix $\widehat{\mathbf{\Sigma}}^s$, Cai and Yuan [2012] proposed a block thresholding estimator to achieve the optimal rate in (2.5.4) without the knowledge of α. An interesting problem is whether the same can be achieved using the Kendall's tau or Spearman's rho, as it seems to need a modification of Theorem 2.2 for off diagonal blocks of the error $\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$.

2.5.2 Principal Component Analysis

Theorem 2.1 immediately yields the following theorem via the Weyl [1912] and Davis and Kahan [1970] inequalities.

Theorem 2.4. *Consider the Gaussian copula model in (2.2.1). Let $\boldsymbol{P}_k$, $\widehat{\boldsymbol{P}}_k^\tau$ and $\widehat{\boldsymbol{P}}_k^\rho$ be the projections to the span of the k leading eigenvectors of $\mathbf{\Sigma}$, $\widehat{\mathbf{\Sigma}}^\tau$ and $\widehat{\mathbf{\Sigma}}^\rho$ respectively corresponding to their k largest eigenvalues. Let λ_j be the j-th largest eigenvalue of $\mathbf{\Sigma}$. Then, for a certain numerical constant C,*

$$\max\left(\mathbb{E}\left\|\widehat{\boldsymbol{P}}_k^\tau - \boldsymbol{P}_k\right\|_S, \mathbb{E}\left\|\widehat{\boldsymbol{P}}_k^\tau - \boldsymbol{P}_k\right\|_S\right) \le C\|\mathbf{\Sigma}\|_S(\sqrt{d/n} + d/n)/(\lambda_k - \lambda_{k+1}).$$

Now we consider the problem of estimating the direction of a sparse leading eigenvector. We illustrate the utility of our sparse spectral error bound in the sparse PCA problem by plugging in $\{\widehat{\mathbf{\Sigma}}^\tau, \widehat{\mathbf{\Sigma}}^\rho\}$ in place of $\widetilde{\mathbf{\Sigma}}^s$ in a formulation of Vu and Lei [2012]. In particular, we consider an integer $s < d$ to be an upper bound on the number of nonzero components of the principal eigenvector $\boldsymbol{\theta}_1$ of $\mathbf{\Sigma}$. The following describes the

sparse estimates of the principal eigenvector based on $\widehat{\boldsymbol{\Sigma}}^{\tau}$ and $\widehat{\boldsymbol{\Sigma}}^{\rho}$.

$$\widehat{\boldsymbol{\theta}}^{\tau}_{1;s} = \underset{\boldsymbol{v}\in\mathbb{S}^{d-1}:\|\boldsymbol{v}\|_0\le s}{\arg\max} \left|\boldsymbol{v}^T\widehat{\boldsymbol{\Sigma}}^{\tau}\boldsymbol{v}\right| \qquad \widehat{\boldsymbol{\theta}}^{\rho}_{1;s} = \underset{\boldsymbol{v}\in\mathbb{S}^{d-1}:\|\boldsymbol{v}\|_0\le s}{\arg\max} \left|\boldsymbol{v}^T\widehat{\boldsymbol{\Sigma}}^{\rho}\boldsymbol{v}\right| \tag{2.5.5}$$

The following theorem provides the rate of convergence for sparse PCA.

Theorem 2.5 (Sparse PCA)**.** *Consider the Gaussian copula model in (2.2.1). Let* $(\lambda_1, \boldsymbol{\theta}_1)$ *be the leading eigenpair of* $\boldsymbol{\Sigma}$ *with* $\|\boldsymbol{\theta}_1\|_0 \le s \to \infty$*. Let* λ_2 *be the second largest eigenvalue of* $\boldsymbol{\Sigma}$*. Let* $\widehat{\boldsymbol{\theta}}^{\tau}_{1;s}$ *and* $\widehat{\boldsymbol{\theta}}^{\rho}_{1;s}$ *be the estimate obtained by the optimization defined in (2.5.5). If* $t + \log d \le \beta\sqrt{(n/s)(t+\log(ed/s))}$*, then for both* $\widehat{\boldsymbol{\theta}}_{1;s} = \widehat{\boldsymbol{\theta}}^{\tau}_{1;s}$ *and* $\widehat{\boldsymbol{\theta}}_{1;s} = \widehat{\boldsymbol{\theta}}^{\rho}_{1;s}$ *and some numeric constant* $C > 0$*,*

$$\left|\sin\angle(\widehat{\boldsymbol{\theta}}_{1;s}, \boldsymbol{\theta}_1)\right| \le \frac{C(1+\beta)}{\lambda_1 - \lambda_2}\left(\|\boldsymbol{\Sigma}\|_S + \|\boldsymbol{\Sigma}\|_S^{1/2}\|\boldsymbol{\Sigma}\|_{(2,\infty)}\right)\sqrt{(t + s\log(ed/s))/n}$$

with probability at least $1 - e^{-t}$*.*

Theorem 2.5 follows from Corollary 2.2 by an application of a similar result from Wang et al. [2013]. We omit the proofs.

2.6 Proofs

Proof of Lemma 2.1. By (2.2.4), the kernel for Kendall's tau is

$$h_{j,k}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathrm{sgn}(x_{1j} - x_{2j})\mathrm{sgn}(x_{1k} - x_{2k}).$$

The definition of $\overline{h}(x, y, \rho)$ in (2.3.16) directly yields (2.3.17) and the first identity of (2.3.19). It remains to verify the properties of $g(x, y, \rho)$ in (2.3.18) and compute the expectation in (2.3.19).

We first prove the following inequality:

$$\max_y \left|\Phi(y) - \Phi(y\sqrt{1-\rho^2})\right| \le |\rho|/2, \ \forall -1 \le \rho \le 1. \tag{2.6.1}$$

For fixed ρ, the above maximum is attained, $(d/dy)\{\Phi(y) - \Phi(y\sqrt{1-\rho^2})\} = 0$, when $e^{-y^2/2} = \sqrt{1-\rho^2}e^{-y^2(1-\rho^2)/2}$ or equivalently $(1-\rho^2)e^{y^2\rho^2} = 1$. Let $y_\rho = \rho^{-1}\sqrt{-\log(1-\rho^2)}$ be the solution. Since the equality is attained in (2.6.1) at $\rho = 1$, (2.6.1) is a consequence of

$$\begin{aligned} & \frac{d}{d\rho}\frac{\Phi(y_\rho) - \Phi(y_\rho\sqrt{1-\rho^2})}{\rho} \\ = \ & \frac{\varphi(y_\rho\sqrt{1-\rho^2})}{\sqrt{1-\rho^2}} - \frac{\Phi(y_\rho) - \Phi(y_\rho\sqrt{1-\rho^2})}{\rho^2} \\ \ge \ & 0. \end{aligned} \tag{2.6.2}$$

By the monotonicity of the normal density $\varphi(t)$ in $|t|$,

$$\Phi(y_\rho) - \Phi(y_\rho\sqrt{1-\rho^2}) \le y_\rho\big(1-\sqrt{1-\rho^2}\big)\varphi(y_\rho\sqrt{1-\rho^2}).$$

Since $y_\rho\rho = \sqrt{-\log(1-\rho^2)} \le \sqrt{\rho^2/(1-\rho^2)}$, (2.6.2) follows from

$$y_\rho\big(1-\sqrt{1-\rho^2}\big) = \frac{y_\rho\rho^2}{1+\sqrt{1-\rho^2}} \le \frac{\rho^2}{\sqrt{1-\rho^2}}.$$

This completes the proof of (2.6.1).

The joint normal density can be factorized as $\varphi_\rho(u,v) = \varphi(u)\varphi_\rho(v|u)$ with the conditional density $\varphi_\rho(v|u) \sim N(\rho u, 1-\rho^2)$. By (2.3.16),

$$\begin{aligned} g(x,y,\rho) &= \int \mathrm{sgn}(x-u)\varphi(u)\Big\{\int \mathrm{sgn}(y-v)\{\varphi_\rho(v|u) - \varphi(v)\}dv\Big\}du \\ &= \int \mathrm{sgn}(x-u)\varphi(u)\Big\{2\int_{-\infty}^{y}\{\varphi_\rho(v|u) - \varphi(v)\}dv\Big\}du \\ &= 2\int \mathrm{sgn}(x-u)\varphi(u)\Big\{\Phi((y-\rho u)/\sqrt{1-\rho^2}) - \Phi(y)\Big\}du. \end{aligned} \tag{2.6.3}$$

This gives the first part of (2.3.18) since $|\Phi((y-\rho u)/\sqrt{1-\rho^2}) - \Phi(y-\rho u)| \le |\rho|/2$ by (2.6.1) and $|\Phi(y-\rho u) - \Phi(y)| \le |\rho u|/\sqrt{2\pi}$.

Similarly, since $\mathrm{sgn}(x-u) = 2I\{u \le x\} - 1$,

$$\begin{aligned}\frac{\partial}{\partial x} g(x,y,\rho) &= \frac{\partial}{\partial x} 4\int_{-\infty}^{x} \varphi(u)\Big\{\Phi((y-\rho u)/\sqrt{1-\rho^2}) - \Phi(y)\Big\}du \\ &= 4\varphi(x)\Big\{\Phi((y-\rho x)/\sqrt{1-\rho^2}) - \Phi(y)\Big\}.\end{aligned}$$

It follows that

$$\begin{aligned}\Big|\frac{\partial}{\partial x} g(x,y,\rho)\Big| &= 4\varphi(x)\Big|\Phi((y-\rho x)/\sqrt{1-\rho^2}) - \Phi(y)\Big| \\ &\le 4\varphi(x)\Big(\frac{|\rho x|}{\sqrt{2\pi}} + \frac{|\rho|}{2}\Big).\end{aligned}$$

This gives the second part of (2.3.18) due to

$$\max_{x>0} 4\varphi(x)(x/\sqrt{2\pi} + 1/2) \le 0.987 < 1.$$

For $j \ne k$, (2.2.8) gives

$$\mathbb{E}\overline{h}_0(X_{1j}, X_{1k}, 0) = \mathbb{E}\mathrm{sgn}(X_{1j} - X_{2j})\mathrm{sgn}(X_{1k} - X_{3k}) = \rho_{jk}/3.$$

Since $U = \Phi(X_1) \sim \mathrm{uniform}(0,1)$, $\int \overline{h}_0^2(x)\varphi(x)dx = 4\mathrm{Var}(U) = 1/3$. The second identity of (2.3.19) follows. ■

Proof of Lemma 2.2. We need to include the sample size n in the subscript. As in Hoeffding [1948], Spearman's rho can be written as

$$\widehat{\rho}_{n,jk} = \frac{n-2}{n+1} u_{n,jk} + \frac{3}{n+1}\widehat{\tau}_{n,jk} \tag{2.6.4}$$

where $u_{n,jk}$ is a U-statistic of order 3 with kernel

$$h^*_{jk}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) = 3\mathrm{sgn}(x_{1,j} - x_{2j})\mathrm{sgn}(x_{1k} - x_{3k}). \tag{2.6.5}$$

For $x \in [0, \pi/2]$, both $\sin x$ and $\sin x - 2\sin(x/3)$ are concave functions with $\sin x - 2\sin(x/3) = 0$ at the two endpoints, so that $\sin(2x/3) \le 2\sin(x/3) \le \sin x$. Thus, with $x = \pi|\rho_{jk}|/2$, (2.2.10) implies that

$$\mathrm{sgn}(\tau_{jk}) = \mathrm{sgn}(\rho_{jk}),\ (\pi/3)|\rho_{jk}| \le (\pi/2)|\tau_{jk}| \le (\pi/2)|\rho_{jk}|. \tag{2.6.6}$$

Since $\mathbb{E}u_{jk} = \rho_{jk}$, $|\mathbb{E}\widehat{\rho}_{jk} - \rho_{jk}| = 3|\rho_{jk} - \tau_{jk}|/(n+1) \le |\rho_{jk}|/(n+1)$. This gives (2.3.20) as $|\rho_{jk}| \le |\Sigma_{jk}|$ by the concavity of $\sin(t)$ in $(0, \pi/6)$. Since $u_{n,jk}$ and $\widehat{\tau}_{n,jk}$ are U-statistics with kernel independent of n, $\widehat{\rho}_{n,jk}$ is a U-statistic with kernel

$$h^{\rho}_{jk}(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3) = \frac{n-2}{n+1}u_{3,jk} + \frac{3}{n+1}\widehat{\tau}_{3,jk}. \tag{2.6.7}$$

Since $|u_{3,jk}| = |4\widehat{\rho}_{3,jk} - 3\widehat{\tau}_{3,jk}| \le 1$ always holds, (2.3.21) follows.

Let $\overline{g}(x, \rho) = \int \overline{h}(x, y, \rho)\varphi(y)dy$. It follows from (2.6.5) that

$$\mathbb{E}\Big[u_{3,jk}\Big|\boldsymbol{X}_1 = \boldsymbol{x}\Big] = \overline{h}(x_j, x_k, 0) + \overline{g}(x_j, \Sigma_{jk}) + \overline{g}(x_k, \Sigma_{jk}).$$

Similarly, $\mathbb{E}\big[3\widehat{\tau}_{3,jk}\big|\boldsymbol{X}_1 = \boldsymbol{x}\big] = 2\overline{h}(x_j, x_k, \Sigma_{jk}) + \tau_{jk}$. Thus, we may take

$$\begin{aligned}\overline{h}^{\rho}(x_j, x_k, \Sigma_{jk}) \ &= \ \frac{n-2}{n+1}\Big(\overline{h}(x_j, x_k, 0) + \overline{g}(x_j, \Sigma_{jk}) + \overline{g}(x_k, \Sigma_{jk})\Big) \\ &\quad + \frac{2}{n+1}\overline{h}(x_j, x_k, \Sigma_{jk})\end{aligned}$$

with $c_{jk} = \tau_{jk}/(n+1)$ in (2.3.3). Since $\overline{g}(x, 0) = \int \overline{h}(x, 0)\overline{h}(y, 0)\varphi(y)dy = 0$, (2.3.24)

holds. Moreover, with $g(x,y,\rho) = \overline{h}(x,y,\rho) - \overline{h}(x,y,0)$ as in (2.3.18),

$$\overline{h}^{\rho}(x,y,\rho) - \overline{h}^{\rho}(x,y,0) = \frac{n-2}{n+1}\Big(\overline{g}(x,\rho) + \overline{g}(y,\rho)\Big) + \frac{2}{n+1}g(x,y,\rho),$$

so that (2.3.22) and (2.3.23) are consequences of

$$\big|\overline{g}(x,\rho)\big| \le |\rho|\Big(\frac{\sqrt{2}}{\pi} + \frac{1}{2}\Big), \quad \Big|\frac{\partial}{\partial x}\overline{g}(x,\rho)\Big| \le |\rho|, \tag{2.6.8}$$

Since $\int \mathrm{sgn}(x-u)\varphi(x)dx = -\overline{h}_0(u)$, (2.6.3) and (2.6.1) yield

$$\begin{aligned}\Big|\overline{g}(y,\rho)\Big| &= \Big|2\int \overline{h}_0(u)\varphi(u)\Big\{\Phi((y-\rho u)/\sqrt{1-\rho^2}) - \Phi(y)\Big\}du\Big| \\ &\le 2\int \Big|\overline{h}_0(u)\big(\rho/2 + \rho u/\sqrt{2\pi}\big)\Big|\varphi(u)du\end{aligned}$$

Since $\int |\overline{h}_0(u)|\varphi(u)du = \int_0^1 |2x-1|dx = 1/2$ and

$$\int |\overline{h}_0(u)u|\varphi(u)du = -2\int_0^{\infty} \overline{h}_0(u)d\varphi(u) = 2\int \varphi^2(u)du = 1/\sqrt{\pi},$$

we have $|\overline{g}(y,\rho)| \le |\rho|(1/2 + \sqrt{2}/\pi)$. In addition, (2.3.18) yields

$$\Big|\frac{\partial}{\partial x}\overline{g}(x,\rho)\Big| \le \max_{x,y}\Big|\frac{\partial}{\partial x}g(x,y,\rho)\Big| \le |\rho|.$$

Hence, (2.6.8) holds and the proof is complete. ■

Proof of Lemma 2.3. By Lemmas 2.1 and 2.2, both Kendall's tau and Spearman's rho are U-statistics with kernel bounded by 1, so that (2.3.6) holds. By (2.3.18) and (2.3.22), (2.3.9) holds, so that (2.3.10) holds. Since completely degenerate U-statistics of order two or higher are orthogonal to U-statistics of order 1, (2.3.6) and (2.3.10)

yield

$$\begin{aligned} & \mathbb{E}\|(\mathbf{U}_n - \mathbb{E}\mathbf{U}_n) - m\boldsymbol{\Delta}_n^{(0)}\|_F^2 \\ \leq \; & \frac{m(m-1)d(d-1)}{n(n-1)} + m^2\Big(C_1^2 \sum_{j\neq k} \frac{\Sigma_{jk}^2}{n} + \frac{4d}{45n}\Big). \end{aligned}$$

Inequality (2.3.25) follows from $C_1^2 \geq 2 + 4/45$ and $\sum_{j\neq k} \Sigma_{jk}^2 = \|\boldsymbol{\Sigma}\|_F^2 - d$. ■

Proof of Lemma 2.4. Let N_ϵ be the largest number of ϵ-balls one can pack in the $(1+\epsilon)$-ball centered at the origin and $\{\boldsymbol{u}_{(j)}, j \leq N_\epsilon\}$ be the centers of such ϵ-balls in one of such configurations. From straight forward volume comparison we have $N_\epsilon \epsilon^d \leq (1+\epsilon)^d$. For each $\boldsymbol{u} \in \mathbb{S}^{d-1}$, $\|\boldsymbol{u} - \boldsymbol{u}_{(j)}\|_2 \leq 2\epsilon$ for some $j \leq N_\epsilon$, so that

$$\begin{aligned} \left|\boldsymbol{u}^T \boldsymbol{\Delta}_n^{(0)} \boldsymbol{u}\right| &\leq \left|\boldsymbol{u}_{(j)}^T \boldsymbol{\Delta}_n^{(0)} \boldsymbol{u}_{(j)}\right| + \left|(\boldsymbol{u} - \boldsymbol{u}_{(j)})^T \boldsymbol{\Delta}_n^{(0)} (\boldsymbol{u} + \boldsymbol{u}_{(j)})\right| \\ &\leq \left|\boldsymbol{u}_{(j)}^T \boldsymbol{\Delta}_n^{(0)} \boldsymbol{u}_{(j)}\right| + 2\epsilon(2+2\epsilon)\|\boldsymbol{\Delta}_n^{(0)}\|_S. \end{aligned}$$

It follows that

$$\|\boldsymbol{\Delta}_n^{(0)}\|_S \leq \sup_{j\leq N_\epsilon} \frac{\left|\boldsymbol{u}_{(j)}^T \boldsymbol{\Delta}_n^{(0)} \boldsymbol{u}_{(j)}\right|}{1 - 4\epsilon(1+\epsilon)}, \quad N_\epsilon \leq (1 + 1/\epsilon)^d. \tag{2.6.9}$$

Since $\boldsymbol{X}$ has iid $N(0, \boldsymbol{\Sigma})$ rows, it can be written as $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}$ with a standard normal matrix $\boldsymbol{Z} \in \mathbb{R}^{n\times d}$. Let $\overline{h}_0(\mathbf{X})$ be the $n \times d$ matrix with elements $\overline{h}_0(X_{ij}) = 2\Phi(X_{ij}) - 1$ and

$$f_{\boldsymbol{u}}(\boldsymbol{Z}) = \|\overline{h}_0\left(\boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}\right)\boldsymbol{u}\|_2/\sqrt{n}.$$

By (2.3.7), $\boldsymbol{\Delta}_n^{(0)}$ has elements $(\mathbb{E}_n - \mathbb{E})\overline{h}_0(x_j)\,\overline{h}_0(x_k)$ so that

$$\boldsymbol{u}^T \boldsymbol{\Delta}_n^{(0)} \boldsymbol{u} = f_{\boldsymbol{u}}^2(\boldsymbol{Z}) - \mathbb{E} f_{\boldsymbol{u}}^2(\boldsymbol{Z}). \tag{2.6.10}$$

Since $(d/dt)\Phi(t) \le 1/\sqrt{2\pi}$, for any $\mathbf{V}, \mathbf{W} \in \mathbb{R}^{n\times d}$ we have

$$|f_{\boldsymbol{u}}(\mathbf{V}) - f_{\boldsymbol{u}}(\mathbf{W})| \le \sqrt{\frac{2}{n\pi}}\|(\mathbf{V}-\mathbf{W})\boldsymbol{\Sigma}^{1/2}\|_F \le \sqrt{\frac{2\|\boldsymbol{\Sigma}\|_S}{n\pi}}\|\mathbf{V}-\mathbf{W}\|_F$$

Thus, the Lipschitz norm of $f_{\boldsymbol{u}}(\cdot)$ is bounded by $\sqrt{2\|\boldsymbol{\Sigma}\|_S/(n\pi)}$. By the Gaussian concentration inequality Borell [1975], we have

$$\mathbb{P}\Big\{\Big|f_{\boldsymbol{u}}(\boldsymbol{Z}) - \mathbb{E}f_{\boldsymbol{u}}(\boldsymbol{Z})\Big| > t\sqrt{2\|\boldsymbol{\Sigma}\|_S/(\pi n)}\Big\} \le 2e^{-t^2/2}. \tag{2.6.11}$$

It follows that

$$\mathbb{E}f_{\boldsymbol{u}}^2(\boldsymbol{Z}) - \Big(\mathbb{E}f_{\boldsymbol{u}}(\boldsymbol{Z})\Big)^2 = \mathrm{Var}\Big(f_{\boldsymbol{u}}(\mathbf{X})\Big) \le \frac{2\|\boldsymbol{\Sigma}\|_S}{\pi n}\int_0^\infty e^{-t^2/2}dt^2 = \frac{4\|\boldsymbol{\Sigma}\|_S}{\pi n}.$$

We note that $\mathbb{E}f_{\boldsymbol{u}}^2(\boldsymbol{Z}) = \boldsymbol{u}^T\mathbf{R}\boldsymbol{u}/3 \le \|\mathbf{R}\|_S/3$ as in (2.3.8), so that by (2.6.10)

$$\begin{aligned}\big|\boldsymbol{u}^T\boldsymbol{\Delta}_n^{(0)}\boldsymbol{u}\big| &\le \Big|f_{\boldsymbol{u}}^2(\mathbf{X}) - \Big(\mathbb{E}f_{\boldsymbol{u}}(\mathbf{X})\Big)^2\Big| + \frac{4\|\boldsymbol{\Sigma}\|_S}{\pi n}\\ &\le \Big(f_{\boldsymbol{u}}(\mathbf{X}) - \mathbb{E}f_{\boldsymbol{u}}(\mathbf{X})\Big)^2 + 2\Big(\|\mathbf{R}\|_S/3\Big)^{1/2}\Big|f_{\boldsymbol{u}}(\mathbf{X}) - \mathbb{E}f_{\boldsymbol{u}}(\mathbf{X})\Big| + \frac{4\|\boldsymbol{\Sigma}\|_S}{\pi n}.\end{aligned}$$

This inequality and (2.6.9) yield

$$\|\boldsymbol{\Delta}_n^{(0)}\|_S \le \frac{\zeta_n^2 + 2\Big(\|\mathbf{R}\|_S/3\Big)^{1/2}\zeta_n + 4\|\boldsymbol{\Sigma}\|_S/(\pi n)}{1 - 4\epsilon(1+\epsilon)} \tag{2.6.12}$$

with $\zeta_n = \max_{j\le(1+1/\epsilon)^d}\Big|f_{\boldsymbol{u}_{(j)}}(\mathbf{X}) - \mathbb{E}f_{\boldsymbol{u}_{(j)}}(\mathbf{X})\Big|$. It follows from (2.6.11) that

$$\mathbb{P}\Big\{\zeta_n > t\sqrt{2\|\boldsymbol{\Sigma}\|_S/(\pi n)}\Big\} \le 2(1+1/\epsilon)^d e^{-t^2/2}. \tag{2.6.13}$$

Let $x_* = 2(d\log(1+1/\epsilon)+\log 2)$. We have

$$\mathbb{E}\zeta_n^2 \le \frac{2\|\mathbf{\Sigma}\|_S}{\pi n}\int_0^\infty \min\Big\{2(1+1/\epsilon)^d e^{-t^2/2}, 1\Big\}dt^2 = \frac{2\|\mathbf{\Sigma}\|_S}{\pi n}\big(x_*+2\big)$$

Taking ϵ satisfying $\epsilon(1+\epsilon)=1/20$, we find $1/(1-4\epsilon(1+\epsilon))=5/4$ and $\log(1+1/\epsilon)\le\pi$, so that $x_*\le 2(\pi d+\log 2)$ and

$$\mathbb{E}\zeta_n^2 \le 4\|\mathbf{\Sigma}\|_S\big(d/n+(1+\log 2)/(\pi n)\big).$$

Combining this with (2.6.12), we have

$$\begin{aligned}\mathbb{E}\|\mathbf{\Delta}_n^{(0)}\|_S &\le (5/4)\big\{\mathbb{E}\zeta_n^2 + 2(\|\mathbf{R}\|_S/3)^{1/2}\mathbb{E}\zeta_n + 4\|\mathbf{\Sigma}\|_S/(\pi n)\big\}\\ &\le 5\|\mathbf{\Sigma}\|_S\big\{d/n+(2+\log 2)/(\pi n)\big\}\\ &\quad +5\big(\|\mathbf{\Sigma}\|_S\|\mathbf{R}\|_S/3\big)^{1/2}\big(d/n+(1+\log 2)/(\pi n)\big)^{1/2}.\end{aligned}$$

This yields (2.3.26) due to $2+\log 2\le\pi$ and $\|\mathbf{R}\|_S\le\|\mathbf{\Sigma}\|_S$. Moreover, by (2.6.13)

$$\mathbb{P}\Big\{\zeta_n > \sqrt{2\pi d+2t^2}\sqrt{2\|\mathbf{\Sigma}\|_S/(\pi n)}\Big\} \le 2e^{\pi d-(2\pi d+2t^2)/2} = 2e^{-t^2}$$

and outside this event (2.6.12) gives

$$\begin{aligned}\|\mathbf{\Delta}_n^{(0)}\|_S &\le 5\|\mathbf{\Sigma}\|_S(d/n+(t^2+1)/(\pi n))\\ &\quad +5\big(\|\mathbf{\Sigma}\|_S\|\mathbf{R}\|_S/3\big)^{1/2}\sqrt{d/n+t^2/(\pi n)}.\end{aligned}$$

This completes the proof due to $\|\mathbf{R}\|_S\le\|\mathbf{\Sigma}\|_S$. ■

Proof of Lemma 2.5. (i) Let $x=(\pi/2)\tau_{jk}$ and $y=(\pi/2)\Delta_{jk}^\tau$ so that $\widehat{\Sigma}_{jk}=\sin(x+y)$ and $\Sigma_{jk}=\sin x$. Because $\sin(x+y)-\sin x-y=(\cos x-1)y-\int_0^y(y-t)\sin(x+t)dt$,

$$\Big|\widehat{\Sigma}_{jk}^\tau-\Sigma_{jk}-(\pi/2)\Delta_{jk}^\tau\Big| \le \frac{2|xy|}{\pi}+\frac{y^2}{2} \le \frac{\pi}{2}\Big|\tau_{jk}\Delta_{jk}^\tau\Big|+\frac{\pi^2}{8}\Big|\Delta_{jk}^\tau\Big|^2.$$

Since $\widehat{\tau}_{jk}$ is a U-statistic of order $m = 2$ and a sign kernel in (2.2.4), the Hoeffding decoupling argument gives $\mathbb{E}(\Delta_{jk}^{\tau})^2 \le \mathbb{E}\big(2\,\mathrm{Bin}(n_2, p_{jk})/n_2 - 2p_{jk}\big)^2 \le 1/n_2$ and

$$\mathbb{E}(\Delta_{jk}^{\tau})^4 \le \mathbb{E}\Big(2\,\mathrm{Bin}(n_2, p_{jk})/n_2 - 2p_{jk}\Big)^4 \le 3/n_2^2,$$

where $p_{jk} = (1 + \tau_{jk})/2$. Since $\sum_{j\neq k} \tau_{jk}^2 \le \sum_{j\neq k} \Sigma_{jk}^2 = \|\mathbf{\Sigma}\|_F^2 - d$, we have

$$\sum_{j,k} \mathbb{E}\Big|\tau_{jk}\Delta_{jk}^{\tau}\Big|^2 \le \frac{\|\mathbf{\Sigma}\|_F^2 - d}{n_2}, \quad \sum_{j,k} \mathbb{E}\Big|\Delta_{jk}^{\tau}\Big|^4 \le \frac{3d^2}{n_2^2}.$$

Consequently, (2.3.28) holds.

(ii) Let $x = (\pi/6)\mathbb{E}\widehat{\rho}_{jk}$, $y = (\pi/6)\Delta_{jk}^{\rho}$ and $z = (\pi/6)(\mathbb{E}\widehat{\rho}_{jk} - \rho_{jk})$ so that $\widehat{\Sigma}_{jk} = 2\sin(x + y)$ and $\Sigma_{jk} = 2\sin(x - z)$. Due to $|z| \le (\pi/6)|\Sigma_{jk}|/(n + 1)$ by (2.3.20),

$$\begin{aligned}
\Big|\widehat{\Sigma}_{jk}^{\rho} - \Sigma_{jk} - \frac{\pi}{3}\Delta_{jk}^{\rho}\Big| &= 2\Big|\sin(x + y) - \sin(x - z) - y\Big| \\
&\le \frac{4|xy|}{\pi} + y^2 + 2|z| \\
&\le \frac{\pi}{9}\Big|\Sigma_{jk}\Delta_{jk}^{\rho}\Big| + \frac{\pi^2}{36}|\Delta_{jk}^{\rho}|^2 + \frac{\pi|\Sigma_{jk}|}{3(n+1)}.
\end{aligned}$$

Similar to part (i), (2.3.29) follows from $\mathbb{E}(\Delta_{jk}^{\tau})^2 \le 1/n_3$ and $\mathbb{E}(\Delta_{jk}^{\tau})^4 \le 3/n_3^2$. ■

Proof Of Lemma 2.6. We write

$$\mathbf{\Delta}_n^{(1)} - \mathbf{\Delta}_n^{(0)} = (\mathbb{E}_n - \mathbb{E})\mathbf{G} = n^{-1}\sum_{i=1}^{n} \mathbf{G}(\boldsymbol{X}_i) - \mathbb{E}\mathbf{G}(\boldsymbol{X}_1)$$

with $\mathbf{G}(\boldsymbol{x}) = \big(g_{jk}(\boldsymbol{x})\big)_{d\times d}$, where $g_{jk}(\boldsymbol{x}) = \overline{h}^{\rho}(x_j, x_k, \Sigma_{jk}) - \overline{h}^{\rho}(x_j, x_k, 0)$ for Kendall's tau and $g_{jk}(\boldsymbol{x}) = \overline{h}(x_j, x_k, \Sigma_{jk}) - \overline{h}(x_j, x_k, 0)$ for Spearman's rho. It follows from (2.3.18) and (2.3.23) that

$$|g_{jk}(\boldsymbol{y}) - g_{jk}(\boldsymbol{x})| \le |\Sigma_{jk}|\left\{|y_j - x_j| + |y_k - x_k|\right\}.$$

This inequality implies that for all d-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$\begin{aligned}
\|\mathbf{G}(\boldsymbol{x}) - \mathbf{G}(\boldsymbol{y})\|_S &\leq \max_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=1} \sum_{j=1}^{d}\sum_{k=1}^{d} |u_j u_k||g_{jk}(\boldsymbol{x}) - g_{jk}(\boldsymbol{y})| \\
&\leq \max_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=1} \sum_{j=1}^{d}\sum_{k=1}^{d} |u_j u_k \Sigma_{jk}| \left(|x_j - y_j| + |x_k - y_k|\right) \\
&\leq 2 \max_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=1} \sum_{j=1}^{d}\sum_{k=1}^{d} |u_j u_k \Sigma_{jk}(x_j - y_j)| \\
&\leq 2 \max_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=1} \sum_{j=1}^{d} |u_j (x_j - y_j)| \max_j \sum_k |u_k \Sigma_{jk}| \\
&\leq 2\|\boldsymbol{\Sigma}\|_{(2,\infty)} \|\boldsymbol{x} - \boldsymbol{y}\|_2.
\end{aligned}$$

Recall that $\mathbf{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^T \in \mathbb{R}^{n\times d}$ with iid $\boldsymbol{X}_i \sim N(0, \boldsymbol{\Sigma})$, so that the matrix $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$ has iid $N(0,1)$ entries. Since $\boldsymbol{X}_i$ are iid vectors, we may write $\mathbf{M}_G = \mathbb{E}\mathbf{G}(\boldsymbol{X}_1)$. Let $\boldsymbol{Z}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i$. Define a function $f : \mathbb{R}^{n\times d} \to \mathbb{R}$ by

$$f(\mathbf{Z}) = \|(\mathbb{E}_n - \mathbb{E})\mathbf{G}\|_S = \left\| \frac{1}{n}\sum_{i=1}^{n} \left\{ \mathbf{G}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}_i) - \mathbf{M}_G \right\} \right\|_S.$$

For matrices $\mathbf{V} = (\boldsymbol{V}_1, \cdots, \boldsymbol{V}_n)^T$ and $\mathbf{W} = (\boldsymbol{W}_1, \cdots, \boldsymbol{W}_n)^T$ in $\mathbb{R}^{n\times d}$, we have

$$\begin{aligned}
|f(\mathbf{V}) - f(\mathbf{W})| &= \left| \|\frac{1}{n}\sum_{i=1}^{n} \mathbf{G}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{V}_i) - \mathbf{M}_G\|_S - \|\frac{1}{n}\sum_{i=1}^{n} \mathbf{G}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{W}_i) - \mathbf{M}_G\|_S \right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{G}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{V}_i) - \mathbf{G}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{W}_i)\|_S \\
&\leq 2\|\boldsymbol{\Sigma}\|_{(2,\infty)} \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{V}_i - \boldsymbol{\Sigma}^{1/2}\boldsymbol{W}_i\|_2 \\
&\leq 2\frac{\|\boldsymbol{\Sigma}\|_{(2,\infty)}\|\boldsymbol{\Sigma}\|_S^{1/2}}{\sqrt{n}} \|\mathbf{V} - \mathbf{W}\|_F.
\end{aligned}$$

We have here a Lipschitz continuity in nd variables. An application of the concentration inequality for Lipschitz continuous functions yields that for any $t > 0$

$$\mathbb{P}\left(f(\mathbf{Z}) - \mathbb{E}f(\mathbf{Z}) > 2\|\mathbf{\Sigma}\|_{(2,\infty)}\|\mathbf{\Sigma}\|_S^{1/2}\frac{t}{\sqrt{n}} \right) \leq \exp\left\{-t^2/2\right\}$$

with $f(\mathbf{Z}) = \|(\mathbb{E}_n - \mathbb{E})\mathbf{G}\|_S = \|\mathbf{\Delta}_n^{(1)} - \mathbf{\Delta}_n^{(0)}\|_S$. From (2.3.10) it follows that

$$\mathbb{E}^2 f(\mathbf{Z}) \leq \mathbb{E}\|\mathbf{\Delta}_n^{(1)} - \mathbf{\Delta}_n^{(0)}\|_S^2 \leq C_1^2 \sum_{j\neq k} \frac{\Sigma_{jk}^2}{n} + \frac{4d}{45n} \leq \frac{C_1^2\|\mathbf{\Sigma}\|_F^2 - 2d}{n},$$

where $C_1 = 2/\pi + 1 \leq 2$ for Kendall's tau and $C_1 \leq 1 + \sqrt{8}/\pi \leq 2$ for Spearman's rho, with $C_1^2 \geq 2 + 4/45$. ■

Proof of Lemma 2.7. By Lemmas 2.1 and 2.2, $\left(\mathbf{U}_n - \mathbb{E}\mathbf{U}_n\right)_{jk}$ are U-statistics of order m and their kernels are uniformly bounded by 1, where $m = 2$ for Kendall's tau and $m = 3$ for Spearman's rho. Let $\mathbf{D} = (D_{jk})_{d\times d}$ with $D_{jk} = \left(\mathbf{U}_n - \mathbb{E}\mathbf{U}_n - m\mathbf{\Delta}_n^{(1)}\right)_{jk}$. Since $m\mathbf{\Delta}_n^{(1)}$ is the first order Hoeffding decomposition of $\left(\mathbf{U}_n - \mathbb{E}\mathbf{U}_n\right)_{jk}$, D_{jk} is second order degenerate. Thus, by Arcones and Gine [1993], $\mathbb{P}\left\{\left|D_{jk}\right| > Ct/n\right\} \leq 4e^{-t}$ for a certain numerical constant C. This gives $\mathbb{P}\Big\{\|\mathbf{D}\|_{\max} > Ct/n\Big\} \leq 4d^2e^{-t}$. Because $\max_{|A|\leq s}\|\mathbf{D}_{A\times A}\|_S \leq s\|\mathbf{D}\|_{\max}$, choosing $t = s(2\log 2d + t)$ completes the proof. ■

Proof of Lemma 2.8. We prove part (ii) only as part (i) can be found in Wegkamp and Zhao [2013]. Let $x = (\pi/6)\mathbb{E}\widehat{\rho}_{jk}$, $y = (\pi/6)\Delta_{jk}^{\rho}$ and $z = (\pi/6)(\mathbb{E}\widehat{\rho}_{jk} - \rho_{jk})$, so that $\widehat{\Sigma}_{jk} = 2\sin(x+y)$ and $\Sigma_{jk} = 2\sin(x-z)$. By (2.3.20),

$$\begin{aligned}
& \left|\widehat{\Sigma}_{jk}^{\rho} - \Sigma_{jk} - \cos((\pi/6)\rho_{jk})(\pi/3)\Delta_{jk}^{\rho}\right| \\
= \;& 2\Big|\sin(x+y) - \sin(x-z) - y\cos(x-z)\Big| \\
\leq \;& 2|z| + y^2 \\
\leq \;& \frac{\pi^2}{36}|\Delta_{jk}^{\rho}|^2 + \frac{\pi|\Sigma_{jk}|}{3(n+1)}.
\end{aligned}$$

We have $\|(|\Delta^{\rho}_{jk}|^2)_{A\times A}\|_S \leq s\left\|\mathbf{\Delta}^{\rho}\right\|^2_{\max}$ and $\|(|\Sigma_{jk}|)_{A\times A}\|_S \leq \sqrt{s}\|\mathbf{\Sigma}\|_{(2,\infty)}$. The tail probability bound for $\left\|\mathbf{\Delta}^{\rho}\right\|_{\max}$ follows by applying the union bound to the Hoeffding [1963] inequality. As in Wegkamp and Zhao [2013], due to $\cos((\pi/6)\rho_{jk}) = \sqrt{1-\Sigma^2_{jk}/4}$,

$$\left\|\left(\cos((\pi/6)\rho_{jk})\Delta^{\rho}_{jk}\right)_{A\times A}\right\|_S \leq \sum_{m=0}^{\infty}\left|\binom{1/2}{m}\right| 4^{-m}\left\|\mathbf{\Delta}^{\rho}\right\|_S.$$

This completes the proof as $\sum_{m=0}^{\infty}\left|\binom{1/2}{m}\right| 4^{-m} = 2-\sqrt{1-1/4}$. ■

Chapter 3

Nonparametric Estimation of Sparse Precision Matrices

3.1 Introduction

We consider n iid copies $\{\boldsymbol{X}_i : 1 \leq i \leq n\}$ of a p-dimensional Gaussian random vector $\boldsymbol{X}$. We define $\mathbf{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^T \in \mathbb{R}^{n \times p}$. We assume $\boldsymbol{X}_i$'s are centered and scaled, so that $\mathbb{E}\boldsymbol{X} = \mathbf{0}$ and the correlation matrix is given by $\mathbb{E}\boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$. Given the population correlation matrix $\boldsymbol{\Sigma}$, an important problem is the estimation of the matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ also called the precision matrix. Suppose that instead of $\boldsymbol{X} = (X_1, \cdots, X_p)^T$ we only observe n iid copies $(\boldsymbol{Y}_i, 1 \leq i \leq n)$ of the transformed variable

$$\boldsymbol{Y} = (f_1(X_1), \cdots, f_p(X_p))^T. \tag{3.1.1}$$

Here f_i's are unknown but strictly increasing. Let $\mathbf{Y} \in \mathbb{R}^{n \times p}$ be the new data matrix with each row being n independent copies of $\boldsymbol{Y}$. This model has been referred to as a multivariate Gaussian transformational family in Mitra and Zhang [2014a]. The objective is to estimate the latent precision matrix $\boldsymbol{\Theta}$ from $\mathbf{Y}$.

If $\mathbf{X}$ were observable, a direct inversion of sample correlation matrix $\widehat{\boldsymbol{\Sigma}}^s = \mathbf{X}^T\mathbf{X}/n$ estimated from the data would be a straight forward solution especially for $p < n$ scenarios. However, the modern regime of statistical problems often involve high dimensional scenarios where $p > n \to \infty$. In such cases, $\widehat{\boldsymbol{\Sigma}}^s$ is not invertible and

additional sparsity assumptions are usually imposed on the target $\mathbf{\Theta}$ for estimation via regularization. An example of such assumptions are, an upper bound s on the number of non-zero off diagonal entries of $\mathbf{\Theta}$. Another example is an upper bound d on the maximum number of nonzero entries in any column of $\mathbf{\Theta}$; usually referred to as the degree of the matrix $\mathbf{\Theta}$.

Assuming observable $\mathbf{X}$, several regularization strategies have been developed that aim to efficiently estimate $\mathbf{\Theta}$ under such sparsity conditions. A detailed theoretical study of the convergence of such regularized estimates in different matrix norms have been done. Yuan and Lin [2007] has developed the graphical Lasso (gLasso) procedure based on Lasso (Tibshirani [1996b]) penalization of off-diagonal entries. The model selection properties of gLasso has been studied in Ravikumar et al. [2008]. In Rothman et al. [2008], a Sparse Permutation Invariant Covariance Estimation (SPICE) procedure was proposed that is identical to the gLasso formulation. A convergence rate of $\sqrt{s(\log p)/n}$ in the spectral norm is established for such estimators. See also Ravikumar et al. [2011]. Lam and Fan [2009] have studied the gLasso formulation under concave penalties. Also Banerjee et al. [2008], Friedman et al. [2008] have proposed similar formulations for estimation of $\mathbf{\Theta}$ where all elements of $\mathbf{\Theta}$ are penalized. In Yuan [2010] and Cai et al. [2011], the authors established a much faster rate of convergence of $d\sqrt{(\log p)/n}$ using Dantzig selector (Candes and Tao [2007]) based estimates. In particular the CLIME estimator proposed by Cai et al. [2011] used Dantzig selector based estimate for each column of $\mathbf{\Theta}$ separately. Their results required that the matrix ℓ_1 norm of $\mathbf{\Theta}$, denoted by $\|\mathbf{\Theta}\|_1$ be bounded. Similar procedures based on Lasso have also been studied. In particular Meinshausen and Bühlmann [2006] proposed a neighborhood selection method based on Lasso for each node of a Gaussian graphical model. See also Yang and Kolaczyk [2010], Rocha et al. [2008] etc. In Sun and Zhang [2013], the authors used the scaled Lasso procedure developed in Sun and Zhang [2012a] to estimate each column of $\mathbf{\Theta}$. In contrast to

related work in this problem, their approach does not require a cross validation procedure for estimating the tuning parameter. This scaled Lasso procedure yields a convergence rate of $d\sqrt{(\log p)/n}$ under the boundedness assumption on the spectral norm $\|\mathbf{\Theta}\|_S$ of the precision matrix $\mathbf{\Theta}$. More recently Zhang and Zou [2014] introduced a new convex loss function called D-trace loss and obtained estimates of $\mathbf{\Theta}$ under ℓ_1 penalization. They provided algorithms for optimization and established $d\sqrt{(\log p)/n}$ convergence rate in spectral norm for sub-Gaussian distributions under irrepresentability conditions.

Considerable research has also been directed towards the study of estimation of correlation and precision matrices under more general distribution families as in (3.1.1). When only $\mathbf{Y}$ is observable (instead of $\mathbf{X}$), the estimation of the latent correlation $\mathbf{\Sigma}$ is done through nonparametric estimates. Spectral norm consistency of nonparametric estimates of the latent correlation matrix $\mathbf{\Sigma}$ has been studied in detail for elliptical copula families in Han and Liu [2013],Wegkamp and Zhao [2013] for Kendall's tau, and for Gaussian copula models in Mitra and Zhang [2014a] for both Kendall's tau and Spearman's rho based nonparametric estimates. In Liu et al. [2009] a nonparanormal family was defined that relaxes the Gaussianity assumption; it is a reformulation of the Gaussian copula model and a slight variant of (3.1.1). Liu et al. [2009] established a convergence rate of $\sqrt{(s \log p \log^2 n)/\sqrt{n}}$ for estimation of $\mathbf{\Theta}$ in spectral norm. In this work the authors used gLasso procedure on sample correlation matrix constructed via estimation of unknown copula function. In Liu et al. [2012a], Xue and Zou [2012], the authors proposed nonparametric estimates of $\mathbf{\Sigma}$ based on Kendall's tau and Spearman's rho and proposed its use in estimation $\mathbf{\Theta}$ via gLasso, CLIME, neighborhood Lasso and neighborhood Dantzig selector etc. Assuming bound on $\|\mathbf{\Theta}\|_1$, Liu et al. [2012a] established a $d\sqrt{(\log p)/n}$ convergence rate for such estimates. In related work, Liu and Wang [2012] developed a calibrated procedure (TIGER) for estimation of Gaussian graphical models where they used

square root Lasso (See Belloni et al. [2011]) . Liu et al. [2012b] developed a rank based CLIME method for more general elliptical models with correlation matrices based on Kendall's tau. Zhao and Liu [2014] developed a calibrated estimation procedure for precision matrices for elliptical models. All these works require bound on matrix ℓ_1 norm of the matrix $\mathbf{\Theta}$. More recently, Barber and Kolar [2015] developed a Lasso based procedure for confidence intervals for individual elements of the precision matrix based on Kendall's tau estimate of correlation matrix in transelliptical copula models.

We apply the scaled Lasso based procedure described in Sun and Zhang [2013] to the nonparametric estimates of correlation estimates and establish the $d\sqrt{(\log p)/n}$ convergence rate in spectral norm under the weaker condition of bound on $\|\mathbf{\Theta}\|_S$. In particular, Theorem 3.1 gives the detailed convergence rate and related assumptions. Our work is organized into two sections. In Section 3.2 we describe the scaled Lasso procedure for nonparametric estimates of precision matrices and provide convergence rates. In Section 3.3 we describe the scaled Lasso procedure based on nonparametric correlation matrix estimates and provide probability inequalities necessary for consistency of scaled Lasso estimates. The proof of Theorem 3.1 is similar to the one in Sun and Zhang [2013]. We nonetheless provide a proof for the sake of completion and relegate it to the final section.

3.2 Problem Setup & Main Results

We use the following notations. For vectors $\boldsymbol{u} \in \mathbb{R}^p$, the ℓ_p norm is denoted by $\|\boldsymbol{u}\|_p = \left(\sum_{k=1}^d |u_k|^p\right)^{1/p}$, with $\|\boldsymbol{u}\|_\infty = \max_{1\le k\le d}|u_k|$ and $\|\boldsymbol{u}\|_0 = \#\{j : u_j \neq 0\}$. For matrices $\mathbf{A} = (A_{jk})_{p\times p} \in \mathbb{R}^{p\times p}$, the $\ell_p \to \ell_q$ operator norm is denoted by $\|\mathbf{A}\|_{p,q} = \max_{\|\boldsymbol{u}\|_p=1} \|\mathbf{A}\boldsymbol{u}\|_q$. The $\ell_2 \to \ell_2$ operator norm, known as the spectrum norm, is $\|\mathbf{A}\|_S = \|\mathbf{A}\|_{2,2} = \max_{\|\boldsymbol{u}\|_2=1} |\boldsymbol{u}^T\mathbf{A}\boldsymbol{u}|$. The vectorized ℓ_∞ and Frobenius norms are denoted by $\|\mathbf{A}\|_{\max} = \max_{j,k}|A_{jk}|$ and $\|\mathbf{A}\|_F = \sqrt{\text{trace}\left(\mathbf{A}^T\mathbf{A}\right)}$. For symmetric matrices $\mathbf{A}$, the j^{th} eigenvalue of $\mathbf{A}$ is denoted by $\lambda_j(\mathbf{A})$ so that $\lambda_1(\mathbf{A}) = \|\mathbf{A}\|_S$.

For $A, B \subset \{1, \cdots, p\}$, the matrix $\mathbf{A}_{A,B} \in \mathbb{R}^{|A|\times|B|}$ is constructed from $\mathbf{A}$ with corresponding rows and columns as indexed in A, B. The matrix $\mathbf{A}_{-A,-B} \in \mathbb{R}^{(p-|A|)\times(p-|B|)}$ will denote the matrix $\mathbf{A}$ with indices in A, B not chosen. For any matrix $\mathbf{A}$, $\mathbf{A}_{i\bullet}, \mathbf{A}_{\bullet j}$ will denote the i^{th} row and j^{th} column of $\mathbf{A}$ resp. In all subsequent discussion, $\mathbb{E}$ and $\mathbb{P}$ denote the expectation and probability measure. Give a function f and an iid sample $\{X_i\}_{i=1}^n$, we denote by $\mathbb{E}_n f(X) = 1/n \sum_{i=1}^n f(X_i)$.

Finally the asymptotic relation $a_n = \mathcal{O}(b_n)$ will imply $a_n \leq K b_n$ for some fixed constant $K > 0$. The notations $a \vee b$ and $a \wedge b$ will denote respectively the maximum and minimum of a and b.

3.2.1 Nonparametric Estimates of Correlation Matrices

We consider the multivariate nonparametric transformation family as described in Mitra and Zhang [2014a]. Let $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{ip})$ with $Y_{ij} = f_j(X_{ij})$. Here the univariate functions f_j's are continuous, monotone and unknown. The objective is to estimate the inverse of the correlation matrix $\boldsymbol{\Sigma} = \mathbb{E}\boldsymbol{X}\boldsymbol{X}^T$, denoted by $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ using the observations $\mathbf{Y} = (Y_{ij})_{n\times p}$ only. To this end, we use the nonparametric estimates of correlation matrices as follows. The population versions of Kendall' tau and Spearman's rho are given by

$$\tau_{jk} = \mathbb{E}\text{sgn}(Y_{i_1 j} - Y_{i_2 j})\text{sgn}(Y_{i_1 k} - Y_{i_2 k}), \quad \rho_{jk} = 3\mathbb{E}\text{sgn}(Y_{i_1 j} - Y_{i_2 j})\text{sgn}(Y_{i_1 k} - Y_{i_3 k}). \tag{3.2.1}$$

The sample Kendall's tau and Spearman's rho estimates of dependence between $\mathbf{Y}_{\bullet j}$ and $\mathbf{Y}_{\bullet k}$ are given by

$$\widehat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \le i_1 < i_2 \le n} \text{sgn}(Y_{i_1 j} - Y_{i_2 j}) \text{sgn}(Y_{i_1 k} - Y_{i_2 k}), \tag{3.2.2}$$

$$\widehat{\rho}_{jk} = \frac{\sum_{i=1}^{n} (r_{ij} - (n+1)/2)(r_{ik} - (n+1)/2)}{\sqrt{\sum_{i=1}^{n} (r_{ij} - (n+1)/2)^2} \sqrt{\sum_{i=1}^{n} (r_{ik} - (n+1)/2)^2}}, \tag{3.2.3}$$

where r_{ij} denote the rank of Y_{ij} among $\mathbf{Y}_{\bullet j}$. We will denote by $\mathbf{T}, \widehat{\mathbf{T}} \in \mathbb{R}^{p \times p}$, the population and sample version of Kendall's tau matrix and similarly for $\mathbf{R}, \widehat{\mathbf{R}}$. Kendall [1948] and Kruskal [1958] provided a recipe for constructing the correlation matrix $\mathbf{\Sigma}$ using Kendall's tau and Spearman's rho measures of association. Since f_j's are strictly increasing, it follows from their results that,

$$\Sigma_{jk} = \sin(\frac{\pi}{2} \tau_{jk}) = 2 \sin(\frac{\pi}{6} \rho_{jk}). \tag{3.2.4}$$

Using these identities, nonparametric estimates of $\mathbf{\Sigma}$ based on τ and $\widehat{\rho}$ has been proposed and analyzed. Let us define,

$$[\widehat{\mathbf{\Sigma}}^{\tau}]_{j,k} = \begin{cases} \sin(\frac{\pi}{2} \widehat{\tau}_{jk}) & j \neq k \\ 1 & j = k \end{cases}, \qquad [\widehat{\mathbf{\Sigma}}^{\rho}]_{j,k} = \begin{cases} 2 \sin(\frac{\pi}{6} \widehat{\rho}_{jk}) & j \neq k \\ 1 & j = k \end{cases}. \tag{3.2.5}$$

See Liu et al. [2012a], Xue and Zou [2012], Han and Liu [2013], Wegkamp and Zhao [2013], Mitra and Zhang [2014a],,Barber and Kolar [2015] etc. In particular Mitra and Zhang [2014a] showed that, under the Gaussian copula model in (3.1.1), for any set A with $|A| \subset \{1, \cdots, p\}$ with $|A| \le m < p$ and $0 < \epsilon < 1$ and some fixed constant

$C > 0$,

$$\max_{A:|A|\leq m} \|(\widehat{\mathbf{\Sigma}}^{\tau} - \mathbf{\Sigma})_{A\times A}\|_S \leq C\|\mathbf{\Sigma}\|_S(\|\mathbf{\Sigma}\|_S^{1/2} + 1)\sqrt{m\log(ep/\epsilon)/n}$$
$$= f(\mathbf{\Sigma}, m, n, p, \epsilon), \text{ say,} \tag{3.2.6}$$

with probability at least $1 - \epsilon$. The same error rate is also true for $\widehat{\mathbf{\Sigma}}^{\rho}$. These rates match corresponding results for sample correlation matrix $\widehat{\mathbf{\Sigma}}^{s}$ (given $\mathbf{X}$ is observable) as provided in Davidson and Szarek [2001]. See also Vershynin [2012] etc.

3.2.2 Inversion of Nonparametric Matrices via Scaled Lasso

We now describe the scaled Lasso procedure (Sun and Zhang [2013]) for estimation of sparse precision matrix by some $\widehat{\mathbf{\Theta}}$ so that $\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Sigma}} \approx \mathbf{I}_p$ where $\widehat{\mathbf{\Sigma}} = \{\widehat{\mathbf{\Sigma}}^{\tau}, \widehat{\mathbf{\Sigma}}^{\rho}\}$ is constructed based on data coming from a Gaussian copula model in (3.1.1). For ease of discussion we only deal with $\widehat{\mathbf{\Sigma}}^{\tau}$ in the following while noting that the same results hold for $\widehat{\mathbf{\Sigma}}^{\rho}$.

Let $\mathbf{\Theta}^* = (\Theta^*_{ij})p \times p$ be the positive definite target matrix satisfying $\mathbf{\Sigma}\mathbf{\Theta}^* = \mathbf{I}_p$. We define the degree of the matrix $\mathbf{\Theta}^*$ as

$$\deg(\mathbf{\Theta}^*) = \max_j |S_j| + 1 = d, \tag{3.2.7}$$

where $S_j = \{i \neq j : \Theta^*_{ij} \neq 0\}$; number of off-diagonal nonzero elements in the j^{th} column. Scaled Lasso is used to estimate $\mathbf{\Theta}^*$ column by column under the sparsity condition (3.2.7) . As mentioned in Introduction, one advantage of employing a scaled Lasso based estimation procedure is due to the relaxation of the condition of bound in ℓ_1 norm of $\mathbf{\Theta}^*$. We impose the following spectral norm bound condition on $\mathbf{\Sigma}$ and

$\boldsymbol{\Theta}^*$, namely

$$\|\boldsymbol{\Sigma}\|_S \vee \|\boldsymbol{\Theta}^*\|_S = \mathcal{O}(1). \tag{3.2.8}$$

Before we begin our description of the scaled Lasso procedure, let us define the following quantities. Let $\boldsymbol{\beta} \in \mathbb{R}^{p\times p}$ and

$$\sigma_j^2 = (\boldsymbol{\Theta}^*_{jj})^{-1} \quad \boldsymbol{\beta}_{\bullet j} = -\boldsymbol{\Theta}^*_{\bullet j}(\boldsymbol{\Theta}^*_{jj})^{-1}. \tag{3.2.9}$$

Clearly then $\text{diag}(\boldsymbol{\Theta}^*) = \text{diag}(\sigma_j^{-2}, 1 \le j \le p)$ and $\boldsymbol{\Theta}^* = \boldsymbol{\beta} \cdot \text{diag}(\boldsymbol{\Theta}^*)$. The scaled Lasso based inversion is based on the idea of solving the following problem.

$$\{\widehat{\boldsymbol{\beta}}_{\bullet j}, \widehat{\sigma}_j\} = \arg\min_{\boldsymbol{b},\sigma} \left\{ \frac{\boldsymbol{b}^T \widehat{\boldsymbol{\Sigma}}^\tau \boldsymbol{b}}{2\sigma} + \frac{\sigma}{2} + \lambda_0 \sum_{j=1}^{p} |b_k| \right\}, \quad \text{where } b_j = -1. \tag{3.2.10}$$

The penalty function (3.2.10) was defined in Sun and Zhang [2013] based on the theory of scaled Lasso developed in Sun and Zhang [2012a]. The only difference is that since we are working with a correlation matrix $\boldsymbol{\Sigma}$, the scaling parameter $\Sigma_{kk}^{-1/2} = 1$ and is omitted in the penalty term. The tuning parameter is given by

$$\lambda_0 = A\sqrt{2(\log p^2)/n\epsilon},$$

where $A > 1$ and $0 < \epsilon < 1$ are fixed constants. We also consider the non-scaled Lasso problem as given by,

$$\widehat{\boldsymbol{\beta}}^{Lasso}_{\bullet j} = \arg\min_{\boldsymbol{b}} \left\{ \frac{\boldsymbol{b}^T \widehat{\boldsymbol{\Sigma}}^\tau \boldsymbol{b}}{2} + \lambda \sum_{j=1}^{p} |b_k| \right\}, \quad \text{where } b_j = -1. \tag{3.2.11}$$

The optimization problem in (3.2.10), (3.2.11) are non- convex unless $\widehat{\boldsymbol{\Sigma}}^\tau$ (or $\widehat{\boldsymbol{\Sigma}}^\rho$) are positive semi definite. While the matrices $\widehat{\mathbf{T}}, \widehat{\mathbf{R}}$ are positive semi definite, the matrices

$\widehat{\mathbf{\Sigma}}^{\tau}, \widehat{\mathbf{\Sigma}}^{\rho}$ are indefinite. However, according to theories developed in Zhang and Huang [2008], Zhang [2010], bound on minimum and maximum eigenvalues of submatrices of a particular order of $\mathbf{\Sigma}$ is sufficient for theoretical guarantees on consistency etc. of the estimator in (3.2.11). Let us assume the

$$\text{Sparse Minimum Eigenvalue Condition:} \qquad \lambda_{\min}(\mathbf{\Sigma}_{A,A}) \geq c_* + f(\mathbf{\Sigma}, m, n, p, \epsilon), \tag{3.2.12}$$

for all sets $A \subset \{1, \cdots, p\}$ with $|A| = m \leq d$ and some small fixed constant $c_* > 0$. Under the condition that $d\sqrt{(\log p)/n} = o(1)$, the condition (3.2.12) implies that $\lambda_{\min}(\mathbf{\Sigma}_{A,A}) = c_* + o(1)$. The sparse minimum eigenvalue condition (3.2.12) ensures that

$$\lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\tau}_{A,A}) > c_* > 0 \text{ for all } A \subset \{1, \cdots, p\} \text{ with } |A| \leq m < d$$

via (3.2.6) and Weyl's inequality. The sparsity of Lasso estimator as in (3.2.11) has been developed in Zhang and Huang [2008] under a sparse Riesz condition (SRC). In light of assumptions (3.2.8) and (3.2.12), it follows from their theory that $\|\widehat{\boldsymbol{\beta}}^{Lasso}_{\bullet j}\|_0 = \mathcal{O}(d)$. See also Zhang [2010] where variable selection consistency and sign consistency of parameter estimates for minimax concave penalty has been established under SRC.

The optimization problem in (3.2.11) submits the following KKT conditions.

$$\begin{cases} \widehat{\mathbf{\Sigma}}^{\tau}_{k\bullet}\widehat{\boldsymbol{\beta}}^{Lasso}_{\bullet j}(\lambda) = -\lambda \text{sgn}(\widehat{\beta}^{Lasso}_{k,j}(\lambda)) & \text{if } \widehat{\beta}^{Lasso}_{kj} \neq 0 \\ \widehat{\mathbf{\Sigma}}^{\tau}_{k\bullet}\widehat{\boldsymbol{\beta}}^{Lasso}_{\bullet j}(\lambda) \in \lambda[-1, 1] & \text{if } \widehat{\beta}^{Lasso}_{kj} = 0 \end{cases} \tag{3.2.13}$$

with $\widehat{\beta}_{jj}^{Lasso}(\lambda) = -1$. The final solution for (3.2.10) is given by the iterative scheme

$$\begin{aligned} \widehat{\sigma}_j^2 &\leftarrow [\widehat{\boldsymbol{\beta}}_{\bullet j}^{Lasso}]^T \widehat{\boldsymbol{\Sigma}}^\tau \widehat{\boldsymbol{\beta}}_{\bullet j}^{Lasso}, \\ \lambda' &\leftarrow \widehat{\sigma}_j^2 \lambda_0, \\ \widehat{\boldsymbol{\beta}}_{\bullet j}^{Lasso} &\leftarrow \widehat{\boldsymbol{\beta}}_{\bullet j}^{Lasso}(\lambda'). \end{aligned} \tag{3.2.14}$$

The final estimate is then given by

$$\mathrm{diag}(\widetilde{\boldsymbol{\Theta}}^\tau) = \mathrm{diag}(\widehat{\sigma}_j^2, 1 \le j \le p), \qquad \widetilde{\boldsymbol{\Theta}}^\tau = -\widehat{\boldsymbol{\beta}} \cdot \mathrm{diag}(\widetilde{\boldsymbol{\Theta}}^\tau). \tag{3.2.15}$$

The final step of the estimation process involves a symmetrization procedure using the optimization

$$\widehat{\boldsymbol{\Theta}}^\tau = \underset{\mathbf{M}\,:\,\mathbf{M}=\mathbf{M}^T}{\arg\min} \|\mathbf{M} - \widetilde{\boldsymbol{\Theta}}^\tau\|_1 = \max_{1\le j\le p} \sum_{i=1}^{p} |[\mathbf{M} - \widetilde{\boldsymbol{\Theta}}^\tau]_{ij}|, \tag{3.2.16}$$

where $\widetilde{\boldsymbol{\Theta}}^\tau$ is obtained via (3.2.14) and (3.2.15). Before we proceed to bound the error in estimation of $\boldsymbol{\Theta}^*$ via nonparametric estimate $\widetilde{\boldsymbol{\Theta}}^\tau$, we first define the following quantity. This symmetrization step can be solved efficiently via linear programming; see Yuan [2010]. We state our main theorem concerning the convergence of $\widehat{\boldsymbol{\Theta}}$ to $\boldsymbol{\Theta}^*$. Let us use the notation $\rho^* = \|\boldsymbol{\Sigma}\|_S$ and $\rho_* = \lambda_{\min}(\boldsymbol{\Sigma})$.

Theorem 3.1. *Let $0 < \epsilon < 1$. Also assume that $d\sqrt{(\log(p/\epsilon))/n} < 1$. Additionally let $(\rho*/\rho_* + 1/2)(\sqrt{\rho^*} + 1)\sqrt{(d\log(p/\epsilon))/n} < a$ for some small constant $a > 0$. Let $\boldsymbol{\Sigma}$ satisfies the minimum eigenvalue condition in (3.2.12). Also assume that $\|\boldsymbol{\Sigma}\|_S \vee \|\boldsymbol{\Theta}^*\|_S = \mathcal{O}(1)$. Then*

$$\|\widehat{\boldsymbol{\Theta}}^\tau - \boldsymbol{\Theta}^*\|_S \le \|\widehat{\boldsymbol{\Theta}}^\tau - \boldsymbol{\Theta}^*\|_1 \le C_1' \|\boldsymbol{\Theta}^*\|_1 \frac{d\log p}{n} + C_2' \left\{\max_j \sqrt{\Theta_{jj}^*}\right\} d\sqrt{\frac{\log p}{n}}$$

with probability at least $1-\epsilon$ *and* $C_1', C_2' > 0$ *are fixed constants.*

It is easy to see that under the stronger condition that $\|\mathbf{\Theta}^*\|_1 = \mathcal{O}(1)$, convergence rate for $\|\widehat{\mathbf{\Theta}}^\tau - \mathbf{\Theta}^*\|_S$ matches the convergence rate established for $\widehat{\mathbf{\Sigma}}^\tau$ in Liu et al. [2012a] under ℓ_1 bound. Also, since $\|\mathbf{\Theta}^*\|_1 = \max_j \sum_{i=1}^p |\Theta_{ij}^*| \leq d \max_{i,j} |\Theta_{ij}^*| \leq d \max_j |\Theta_{jj}^*|$, the same convergence rate is retained under the boundedness of $\|\mathbf{\Theta}^*\|_S$ and uniform upper bound on the diagonal elements of $\mathbf{\Theta}^*$. Also note that the condition $(\rho*/\rho_* + 1/2)(\sqrt{\rho^*}+1)\sqrt{(d\log(p/\epsilon))/n}$ is small if $\max\{\rho^*\sqrt{\rho^*}, \sqrt{\rho^*}\}/\sqrt{d}$ is small which is true for large enough d and $\|\mathbf{\Sigma}\|_S$ bounded.

The results on the consistency of $\widehat{\mathbf{\Theta}}^\tau$ as described in Theorem 3.1 hinges on the corresponding consistency results for scaled Lasso estimates. Oracle inequalities for the Lasso has been derived under the so called sign restricted cone invertibility conditions (SCIF) in Ye and Zhang [2010]. Oracle inequalities for scaled Lasso was derived under those cone conditions in Sun and Zhang [2012a]. Let us write,

$$\widehat{\mathbf{\Sigma}}^\tau = \begin{bmatrix} \widehat{\Sigma}_{j,j}^\tau & \widehat{\mathbf{\Sigma}}_{j,-j}^\tau \\ \widehat{\mathbf{\Sigma}}_{-j,j}^\tau & \widehat{\mathbf{\Sigma}}_{-j,-j}^\tau \end{bmatrix}. \tag{3.2.17}$$

Define,

$$(\sigma_j^*)^2 = \boldsymbol{\beta}_{\bullet j}^T \widehat{\mathbf{\Sigma}}^\tau \boldsymbol{\beta}_{\bullet j} \quad \textit{and} \quad z_{(j)}^* = \|\widehat{\mathbf{\Sigma}}_{-j,j}^\tau - \widehat{\mathbf{\Sigma}}_{-j,-j}^\tau \boldsymbol{\beta}_{-j,j}\|_\infty / \sigma_j^*.$$

Also let us assume that $\mathrm{SCIF}_1(\xi, S, \widehat{\mathbf{\Sigma}}_{-j,-j}^\tau)$ (See Section 3.3 for definition) is bounded away from 0. In light of the theory developed in Sun and Zhang [2012a] and Sun and Zhang [2013], the scaled Lasso results for the nonparametric matrix estimate $\widehat{\mathbf{\Sigma}}^\tau$ follows exactly the same way, based on probability bounds for the quantities σ_j^*/σ_j and $z_{(j)}^*$. We provide the following results which follows from Theorem 3.2, 3.3 in Section 3.3

Corollary 3.1 (Corollary to Theorem 3.2)**.** *Let* $0 < \epsilon < 1$ *and* $d\sqrt{\log(p/\epsilon)/n} \leq 1$.

Then for n large enough and $\|\mathbf{\Sigma}\|_S \vee \|\mathbf{\Theta}^*\|_S = \mathcal{O}(1)$, with probability at least $1-\epsilon$

$$\max_{1\le j\le p} |\left(\sigma_j^*/\sigma_j\right)^2 - 1| = \mathcal{O}\left(\sqrt{\frac{\log(p/\epsilon)}{n}}\right). \tag{3.2.18}$$

Corollary 3.2 (Corollary to Theorem 3.3)**.** *Let $0<\epsilon<1$ and $d\sqrt{\log(p/\epsilon)/n} < 1$. Assume that $\|\mathbf{\Sigma}\|_S \vee \|\mathbf{\Theta}^*\|_S = \mathcal{O}(1)$. Then with probability at least $1-\epsilon$*

$$\max\left\{\max_{1\le j\le p}\frac{\|\widehat{\mathbf{\Sigma}}^{\tau}_{-j,j} - \widehat{\mathbf{\Sigma}}^{\tau}_{-j,-j}\boldsymbol{\beta}_{-j,j}\|_\infty}{\sigma_j^*}, \max_{1\le j\le p}\|\widehat{\mathbf{\Sigma}}^{\tau}_{-j,j} - \widehat{\mathbf{\Sigma}}^{\tau}_{-j,-j}\boldsymbol{\beta}_{-j,j}\|_\infty\right\}$$
$$= \mathcal{O}\left(\sqrt{\frac{\log(p^2/\epsilon)}{n}}\right).$$

From Corollaries 3.1, 3.2, it follows using the proofs in Sun and Zhang [2012a] that

$$|\widehat{\sigma}_j/\sigma_j^* - 1| = \mathcal{O}_{\mathbb{P}}\left(\frac{d\log p}{n}\right), \quad \|\widehat{\boldsymbol{\beta}}_{-j,j} - \boldsymbol{\beta}_{-j,j}\|_1/\sigma_j^* = \mathcal{O}_{\mathbb{P}}\left(d\sqrt{\frac{\log p}{n}}\right). \tag{3.2.19}$$

Using these results, the proof of Theorem 3.1 follows using the same argument as in Sun and Zhang [2013] with slight changes as appropriate. We provide a detailed proof in Section 3.4 nonetheless for the sake of completion.

3.3 Scaled Lasso with Nonparametric Correlation Matrix Estimate

We define the cone,

$$\mathscr{C}(\xi, S) = \left\{\boldsymbol{u}\in\mathbb{R}^{p-1} : \|\boldsymbol{u}_{S^c}\|_1 \le \xi\|\boldsymbol{u}_S\|_1\right\} \tag{3.3.1}$$

and the corresponding sign restricted cone as

$$\mathscr{C}_{-}(\xi, S) = \left\{ \boldsymbol{u} \in \mathscr{C}(\xi, S) : u_j \boldsymbol{\Sigma}_{j\bullet} \boldsymbol{u} \leq 0 \ \forall j \notin S \right\}. \tag{3.3.2}$$

Let us consider the sign restricted cone invertibility condition given by

$$\text{SCIF}_1(\xi, S, \boldsymbol{\Sigma}) = \inf \left\{ \frac{|S| \|\boldsymbol{\Sigma}\boldsymbol{u}\|_\infty}{\|\boldsymbol{u}\|_1} : \boldsymbol{u} \in \mathscr{C}_{-}(\xi, S) \right\} > 0. \tag{3.3.3}$$

The SCIF_1 is the most general condition and is weaker than restricted eigenvalue (RE) condition (Bickel et al. [2009]) and compatibility condition (van de Geer [2007]). Oracle inequalities for Lasso has been studied in Ye and Zhang [2010] under SCIF. Oracle inequalities for grouped variables under a group based SCIF has been studied in Mitra and Zhang [2014b].

We consider the optimization problem (3.2.11). Note that pre-multiplying $\boldsymbol{w}_{-j} - \widehat{\boldsymbol{\beta}}_{-j,j}$ to the KKT conditions in (3.2.13), it follows that

$$(\boldsymbol{w}_{-j} - \widehat{\boldsymbol{\beta}}_{-j,j})(\widehat{\boldsymbol{\Sigma}}^{\tau}_{-j,-j}\widehat{\boldsymbol{\beta}}_{-j,j} - \widehat{\Sigma}^{\tau}_{-j,j}) = \lambda(\|\widehat{\boldsymbol{\beta}}_{-j,j}\|_1 - \|\boldsymbol{w}_{-j}\|_1),$$

which can be rearranged to obtain,

$$\begin{aligned}
&(\boldsymbol{w}_{-j} - \widehat{\boldsymbol{\beta}}_{-j,j})\widehat{\boldsymbol{\Sigma}}^{\tau}_{-j,-j}\widehat{\boldsymbol{\beta}}_{-j,j}(\widehat{\boldsymbol{\beta}}_{-j,j} - \boldsymbol{\beta}_{-j,j}) \\
&\qquad \leq \lambda(\|\widehat{\boldsymbol{\beta}}_{-j,j}\|_1 - \|\boldsymbol{w}_{-j}\|_1) + \|\boldsymbol{w}_{-j} - \widehat{\boldsymbol{\beta}}_{-j,j}\|_1 \|\widehat{\boldsymbol{\Sigma}}^{\tau}_{-j,j} - \widehat{\boldsymbol{\Sigma}}^{\tau}_{-j,-j}\boldsymbol{\beta}_{-j,j}\|_\infty \\
&\qquad \leq \lambda(\|\widehat{\boldsymbol{\beta}}_{-j,j}\|_1 - \|\boldsymbol{w}_{-j}\|_1) + \sigma_j^* z^*_{(j)} \|\boldsymbol{w}_{-j} - \widehat{\boldsymbol{\beta}}_{-j,j}\|_1.
\end{aligned} \tag{3.3.4}$$

Equation (3.3.4) is analogous to the basic inequality which is the starting point for the analysis of scaled Lasso estimators as described in Sun and Zhang [2012a]. It is easy to check that the consistency results in Sun and Zhang [2012a], especially Theorem 1 and Theorem 2 (except the asymptotic normality) follow mutatis mutandis and we

have the convergence result as in (3.2.19). We only need to check that probability bounds on σ_j^*/σ_j and $z_{(j)}^*$ submit similar rates. The following Theorems 3.2, 3.3 provide explicit rates.

Theorem 3.2. *Let us consider the true precision matrix* $\mathbf{\Theta}^*$. *Let us define the degree of the precision matrix as* $\deg(\mathbf{\Theta}^*) = \max_j |\{l : \Theta_{lj}^* \neq 0\}| = d$. *Take any* $0 < \epsilon < 1/10$. *Then with probability at least* $1 - 10\epsilon$,

$$\begin{aligned}&\left|\left(\sigma_j^*/\sigma_j\right)^2 - 1\right| \\ &\leq C\frac{\|\mathbf{\Theta}^*_{\bullet j}\|_2}{\Theta^*_{jj}}\left\{\frac{d\log(d/\epsilon)}{n} + (\|\mathbf{\Sigma}\|_S + \|\mathbf{\Sigma}\|_S^{3/2})\sqrt{\frac{\log(1/\epsilon)}{n}} + \|\mathbf{\Sigma}\|_S\frac{\log(e/\epsilon)}{n}\right\}, \quad (3.3.5)\end{aligned}$$

where $C > 0$ *is a fixed numeric constant.*

Proof of Theorem 3.2. Define $A_j \subset \{1, \cdots, p\}$ such that $|A_j| = |\{l : \Theta_{lj}^* \neq 0\}| \leq \deg(\mathbf{\Theta}^*) \equiv d \quad \forall j$. In the following we omit the superscript $*$ in $\mathbf{\Theta}^*$ for ease of notation. Note that,

$$\begin{aligned}\left(\sigma_j^*/\sigma_j\right)^2 - 1 &= \frac{1}{\Theta_{jj}}\mathbf{\Theta}_{\bullet j}^T\widehat{\mathbf{\Sigma}}^\tau\mathbf{\Theta}_{\bullet j} - 1 \\ &= \frac{1}{\Theta_{jj}}\left\{\mathbf{\Theta}_{\bullet j}^T(\widehat{\mathbf{\Sigma}}^\tau - \mathbf{\Sigma})\mathbf{\Theta}_{\bullet j} + \mathbf{\Theta}_{\bullet j}^T\mathbf{\Sigma}\mathbf{\Theta}_{\bullet j}\right\} - 1 \\ &= \frac{1}{\Theta_{jj}}\left\{\mathbf{\Theta}_{\bullet j}^T(\widehat{\mathbf{\Sigma}}^\tau - \mathbf{\Sigma})\mathbf{\Theta}_{\bullet j} + \Theta_{jj}^2\boldsymbol{\beta}_{\bullet j}^T\mathbf{\Sigma}\boldsymbol{\beta}_{\bullet j}\right\} - 1 \\ &= \frac{1}{\Theta_{jj}}\left\{\mathbf{\Theta}_{\bullet j}^T(\widehat{\mathbf{\Sigma}}^\tau - \mathbf{\Sigma})\mathbf{\Theta}_{\bullet j} + \Theta_{jj}\right\} - 1 \\ &= \{\mathbf{\Theta}_{\bullet j}^T(\widehat{\mathbf{\Sigma}}^\tau - \mathbf{\Sigma})\mathbf{\Theta}_{\bullet j}\}/\Theta_{jj}.\end{aligned}$$

Now note that as in Wegkamp and Zhao [2013], using Taylor expansion,

$$\begin{aligned}|\mathbf{\Theta}_{\bullet j}^T(\widehat{\mathbf{\Sigma}}^\tau - \mathbf{\Sigma})\mathbf{\Theta}_{\bullet j}| \leq \frac{\pi}{2}|\mathbf{\Theta}_{\bullet j}^T\left\{\cos((\pi/2)\mathbf{T}) \circ (\widehat{\mathbf{T}} - \mathbf{T})\right\}\mathbf{\Theta}_{\bullet j}| \\ + \frac{\pi^2}{8}|\mathbf{\Theta}_{\bullet j}^T\left\{\sin((\pi/2)\bar{\mathbf{T}}) \circ (\widehat{\mathbf{T}} - \mathbf{T}) \circ (\widehat{\mathbf{T}} - \mathbf{T})\right\}\mathbf{\Theta}_{\bullet j}|,\end{aligned}$$

where $\bar{\mathbf{T}}$ is such that $[\bar{\mathbf{T}}]_{j,k}$ lies in the interval between $[\widehat{\mathbf{T}}]_{j,k}$ and $[\mathbf{T}]_{j,k}$. Now using properties of Hadamard product, we have

$$|\boldsymbol{\Theta}_{\bullet j}^T \left\{ \sin((\pi/2)\bar{\mathbf{T}}) \circ (\widehat{\mathbf{T}} - \mathbf{T}) \circ (\widehat{\mathbf{T}} - \mathbf{T}) \right\} \boldsymbol{\Theta}_{\bullet j}| \leq d \|\boldsymbol{\Theta}_{\bullet j}\|_2^2 \|(\widehat{\mathbf{T}} - \mathbf{T})_{A_j, A_j}\|_{\max}^2.$$

Now using [Barber and Kolar, 2015, Lemma C.1] $\cos((\pi/2)\mathbf{T}) = \sum_r t_r \boldsymbol{a}_r \boldsymbol{b}_r^T$ with $\|\boldsymbol{a}_r\|_\infty, \|\boldsymbol{b}_r\|_\infty \leq 1$ and $\sum_r t_r = 4$. Thus again from [Barber and Kolar, 2015, Lemma D.2]

$$\begin{aligned} |\boldsymbol{\Theta}_{\bullet j}^T \left\{ \cos((\pi/2)\mathbf{T}) \circ (\widehat{\mathbf{T}} - \mathbf{T}) \right\} \boldsymbol{\Theta}_{\bullet j}| &\leq \sum_r t_r |(\boldsymbol{\Theta}_{\bullet j} \circ \boldsymbol{a}_r)^T (\widehat{\mathbf{T}} - \mathbf{T})(\boldsymbol{\Theta}_{\bullet j} \circ \boldsymbol{b}_r)| \\ &\leq 4 |\boldsymbol{\Theta}_{\bullet j}^T (\widehat{\mathbf{T}} - \mathbf{T}) \boldsymbol{\Theta}_{\bullet j}|. \end{aligned}$$

We thus have,

$$|\boldsymbol{\Theta}_{\bullet j}^T (\widehat{\boldsymbol{\Sigma}}^\tau - \boldsymbol{\Sigma}) \boldsymbol{\Theta}_{\bullet j}| \leq d \frac{\pi^2}{8} \|\boldsymbol{\Theta}_{\bullet j}\|_2^2 \|(\widehat{\mathbf{T}} - \mathbf{T})_{A_j, A_j}\|_{\max}^2 + 2\pi |\boldsymbol{\Theta}_{\bullet j}^T (\widehat{\mathbf{T}} - \mathbf{T}) \boldsymbol{\Theta}_{\bullet j}|. \quad (3.3.6)$$

Using Hoeffding inequality, for any $t > 0$ and some fixed small constant $c_1 > 0$,

$$\mathbb{P}\left[\|(\widehat{\mathbf{T}} - \mathbf{T})_{A_j, A_j}\|_{\max}^2 \leq c_1 \frac{\log d + t}{n} \right] > 1 - 2e^{-t}. \quad (3.3.7)$$

Now we use the decomposition from Mitra and Zhang [2014a], namely

$$\widehat{\mathbf{T}} - \mathbf{T} = 2\boldsymbol{\Delta}_n^{(0)} + 2(\boldsymbol{\Delta}_n^{(1)} - \boldsymbol{\Delta}_n^{(0)}) + (\widehat{\mathbf{T}} - \mathbf{T} - 2\boldsymbol{\Delta}_n^{(1)}), \quad (3.3.8)$$

where,

$$\begin{aligned} [\boldsymbol{\Delta}_n^{(0)}]_{j,k} &= (\mathbb{E}_n - \mathbb{E}) \overline{h}_0(X_j) \overline{h}_0(X_k), \\ [\boldsymbol{\Delta}_n^{(1)} - \boldsymbol{\Delta}_n^{(0)}]_{j,k} &= (\mathbb{E}_n - \mathbb{E})(\overline{h}(X_j, X_k, \Sigma_{jk}) - \overline{h}_0(X_j) \overline{h}_0(X_k)). \end{aligned}$$

Here $\overline{h}(X_{i_1j}, X_{i_1k}, \Sigma_{jk}) = \mathbb{E}\left\{\text{sgn}(X_{i_1j} - X_{i_2j})\text{sgn}(X_{i_1k} - X_{i_2k})|\boldsymbol{X}_{i_1\bullet}\right\}$ and $\overline{h}_0(X) = 2\Phi(X) - 1$. We now control each term one by one. First note that from (proof of) [Mitra and Zhang, 2014a, Lemma 4], for any $t > 0$,

$$\mathbb{P}\left[|\boldsymbol{\Theta}_{\bullet j}^T \boldsymbol{\Delta}_n^{(0)} \boldsymbol{\Theta}_{\bullet j}| \leq \|\boldsymbol{\Theta}_{\bullet j}\|_2^2 \left\{\sqrt{8}\|\boldsymbol{\Sigma}\|_S \frac{t}{n\pi} + \frac{4}{\sqrt{3}}\sqrt{\|\mathbf{R}\|_S \|\boldsymbol{\Sigma}\|_S}\sqrt{\frac{t}{n\pi}} + 4\frac{\|\boldsymbol{\Sigma}\|_S}{n\pi}\right\}\right]$$
$$> 1 - 2e^{-t}. \tag{3.3.9}$$

Now note that from [Mitra and Zhang, 2014a, Lemm 6], it follows by Gaussian concentration of Lipschitz continuous function that, for any $t > 0$

$$|\boldsymbol{\Theta}_{\bullet j}^T(\boldsymbol{\Delta}_n^{(1)} - \boldsymbol{\Delta}_n^{(0)})\boldsymbol{\Theta}_{\bullet j} - M| \leq 2\|\boldsymbol{\Theta}_{\bullet j}\|_2^2 \|\boldsymbol{\Sigma}\|_{2,\infty} \|\boldsymbol{\Sigma}\|_S^{1/2} \frac{t}{\sqrt{n}}$$

with probability at least $1 - 2e^{-t^2/2}$. Here $M = \mathbb{E}\boldsymbol{\Theta}_{\bullet j}^T(\boldsymbol{\Delta}_n^{(1)} - \boldsymbol{\Delta}_n^{(0)})\boldsymbol{\Theta}_{\bullet j}$. Note that

$$\begin{aligned}
&\mathbb{E}\boldsymbol{\Theta}_{\bullet j}^T(\boldsymbol{\Delta}_n^{(1)} - \boldsymbol{\Delta}_n^{(0)})\boldsymbol{\Theta}_{\bullet j} \\
&= \sum_{l,m} [\boldsymbol{\Theta}]_{l,j}[\boldsymbol{\Theta}]_{m,j}\mathbb{E}[\boldsymbol{\Delta}_n^{(1)} - \boldsymbol{\Delta}_n^{(0)}]_{l,m} \\
&= \sum_{l,m} [\boldsymbol{\Theta}]_{lj}[\boldsymbol{\Theta}]_{mj}\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[g(X_{il}, X_{im}, \Sigma_{lm}) - \mathbb{E}g(X_{il}, X_{im}, \Sigma_{lm})\right]\right\} \\
&= 0,
\end{aligned}$$

so that for any $t > 0$,

$$\mathbb{P}\left[|\boldsymbol{\Theta}_{\bullet j}^T(\boldsymbol{\Delta}_n^{(1)} - \boldsymbol{\Delta}_n^{(0)})\boldsymbol{\Theta}_{\bullet j}| \leq \sqrt{8}\|\boldsymbol{\Theta}_{\bullet j}\|_2^2 \|\boldsymbol{\Sigma}\|_{2,\infty} \|\boldsymbol{\Sigma}\|_S^{1/2}\sqrt{\frac{t}{n}}\right] > 1 - 2e^{-t}. \tag{3.3.10}$$

Finally note that, the elements of the matrix $(\widehat{\mathbf{T}} - \mathbf{T} - 2\boldsymbol{\Delta}_n^{(1)})$ are degenerate U-statistic. Let us write,

$$|\boldsymbol{\Theta}_{\bullet j}^T(\widehat{\mathbf{T}} - \mathbf{T} - \boldsymbol{\Delta}_n^{(1)})\boldsymbol{\Theta}_{\bullet j}| \leq d\|\boldsymbol{\Theta}_{\bullet j}\|_2^2\|(\widehat{\mathbf{T}} - \mathbf{T} - 2\boldsymbol{\Delta}_n^{(1)})_{A_j,A_j}\|_{\max}.$$

Now by using exponential inequality for degenerate U-statistic we have, for any $t > 0$ we have $\mathbb{P}\{\|(\widehat{\mathbf{T}} - \mathbf{T} - 2\boldsymbol{\Delta}_n^{(1)})_{A_j,A_j}\|_{\max}) \leq c_2(\log d + t)/n\} > 1 - e^{-t}$. Thus,

$$\mathbb{P}\left[|\boldsymbol{\Theta}_{\bullet j}^T(\widehat{\mathbf{T}} - \mathbf{T} - \boldsymbol{\Delta}_n^{(1)})\boldsymbol{\Theta}_{\bullet j}| \leq c_2\|\boldsymbol{\Theta}_{\bullet j}\|_2^2\frac{d\log d + dt}{n}\right] > 1 - 4e^{-t}. \tag{3.3.11}$$

Now using Equations (3.3.6), (3.3.7), (3.3.8), (3.3.9), (3.3.10), (3.3.11) and putting $e^{-t} = \epsilon > 0$ the final theorem follows. Note that we have also used the inequalities $\|\mathbf{R}\|_S \leq \|\boldsymbol{\Sigma}\|_S$ (see [Mitra and Zhang, 2014a, Proposition 1]) and $\|\boldsymbol{\Sigma}\|_{2,\infty} \leq \|\boldsymbol{\Sigma}\|_S$. ■

Theorem 3.3. *Let us consider the true precision matrix $\boldsymbol{\Theta}^*$. Define the degree of the precision matrix as*

$$\deg(\boldsymbol{\Theta}^*) = \max_j |\{l : \Theta_{lj}^* \neq 0\}| = d.$$

Let $0 < \epsilon < 1/5$. Then with probability at least $1 - 5\epsilon$,

$$\|\widehat{\boldsymbol{\Sigma}}_{-j,j}^{\tau} - \widehat{\boldsymbol{\Sigma}}_{-j,-j}^{\tau}\boldsymbol{\beta}_{-j,j}\|_\infty \leq C\left\{(\|\boldsymbol{\Sigma}\|_S\|\boldsymbol{\Theta}_{\bullet j}^*\|_2^2 + \|\boldsymbol{\Theta}_{\bullet j}^*\|_2\|\boldsymbol{\Sigma}\|_S + 1)\sqrt{\frac{\log(p/\epsilon)}{n}}\right.$$
$$\left. + \|\boldsymbol{\Theta}_{\bullet j}^*\|_2\left(\frac{\sqrt{d}\log(dp/\epsilon)}{n} + \frac{d\log(dp/\epsilon)}{n}\right)\right\},$$

where $C > 0$ is a fixed constant.

In the following we will prove Theorem 3.3. First note that,

$$\begin{aligned}\|\widehat{\mathbf{\Sigma}}^{\tau}_{-j,j} - \widehat{\mathbf{\Sigma}}^{\tau}_{-j,-j}\boldsymbol{\beta}_{-j,j}\|_\infty &= \frac{\|\widehat{\mathbf{\Sigma}}^{\tau}_{-j,j}\Theta^*_{j,j} + \widehat{\mathbf{\Sigma}}^{\tau}_{-j,-j}\mathbf{\Theta}^*_{-j,j}\|_\infty}{\Theta^*_{j,j}} \\ &= (\Theta^*_{j,j})^{-1}\|(\widehat{\mathbf{\Sigma}}^{\tau}_{-j,\bullet} - \mathbf{\Sigma}_{-j,\bullet})\mathbf{\Theta}^*_{\bullet,j}\|_\infty \\ &= (\Theta^*_{j,j})^{-1}\max_{k\neq j}\sum_{l=1}^{p}(\widehat{\Sigma}^{\tau}_{kl} - \Sigma_{kl})\Theta^*_{lj}.\end{aligned}$$

Now to prove Theorem 3.3, we consider the decomposition of $\widehat{\mathbf{\Sigma}}^{\tau} - \mathbf{\Sigma}$ as described in Mitra and Zhang [2014a],

$$\begin{aligned}\widehat{\mathbf{\Sigma}}^{\tau} - \mathbf{\Sigma} = (\widehat{\mathbf{\Sigma}}^{\tau} - \mathbf{\Sigma}) - \frac{\pi}{2}(\widehat{\mathbf{T}} - \mathbf{T}) + \frac{\pi}{2}\left\{(\widehat{\mathbf{T}} - \mathbf{T}) - 2\mathbf{\Delta}^{(1)}_n\right\} \\ + \pi\left(\mathbf{\Delta}^{(1)}_n - \mathbf{\Delta}^{(0)}_n\right) + \pi\mathbf{\Delta}^{(0)}_n. \qquad (3.3.12)\end{aligned}$$

The following lemmas control each term in the decomposition term by term.

Lemma 3.1. *Take $0 < \epsilon < 1$. Then with probability at least $1 - \epsilon$, we have*

$$\|[\mathbf{\Delta}^{(0)}_n]_{-j,\bullet}\mathbf{\Theta}^*_{\bullet,j}\|_\infty \leq (4\|\mathbf{\Sigma}\|_S\|\mathbf{\Theta}^*_{\bullet,j}\|_2^2 + 1)\sqrt{\frac{\log(2p/\epsilon)}{n}}.$$

Proof of Lemma 3.1. Note that

$$\|[\mathbf{\Delta}^{(0)}_n]_{-j,\bullet}\mathbf{\Theta}^*_{\bullet,j}\|_\infty = \max_{k\neq j}|\sum_l[\mathbf{\Delta}^{(0)}_n]_{k,l}\Theta^*_{lj}| = \frac{1}{n}\sum_{i=1}^{n}\left\{\overline{h}_0(X_{ik})\sum_l\overline{h}_0(X_{il})\,\Theta^*_{lj} - M\right\},$$

where $M = \mathbb{E}\overline{h}_0(X_{ik})\sum_l\overline{h}_0(X_{il})\,\Theta^*_{lj}$. Let us write $W_i = \overline{h}_0(X_{ik})$, $V_i = \sum_l\overline{h}_0(X_{il})\,\Theta^*_{lj}$ so that we can write

$$\|[\mathbf{\Delta}^{(0)}_n]_{-j,\bullet}\mathbf{\Theta}^*_{\bullet,j}\|_\infty = \frac{1}{n}\sum_i W_iV_i - \mathbb{E}W_iV_i.$$

Recall that $\overline{h}_0(X) = 2\Phi(X) - 1$. Note that $|W_i| \leq 1$ for all i. Now we use the

symmetrization technique. (see Van Der Vaart and Wellner [1996]). Let (W_i', V_i') be iid copies of (W_i, V_i) for all i. Let $\{\varepsilon_i\}_1^n$ be a sequence of Rademacher random variables so that $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$ and are independent of $\{W_i, V_i\}_1^n$ Then, for all $\lambda \in \mathbb{R}$

$$\begin{aligned}
\mathbb{E}\exp\{\lambda \sum_{i=1}^n W_i V_i - \mathbb{E}W_i V_i\} &= \mathbb{E}\exp\{\lambda \sum_{i=1}^n W_i V_i - \mathbb{E}W_i' V_i'\} \\
&= \mathbb{E}\exp\left\{\lambda \mathbb{E}\left(\sum_{i=1}^n (W_i V_i - W_i' V_i') \middle| W_i V_i\right)\right\} \\
&\leq \mathbb{E}\exp\left\{\lambda\left(\sum_{i=1}^n (W_i V_i - W_i' V_i')\right)\right\} \\
&= \mathbb{E}\exp\left\{\lambda\left(\sum_{i=1}^n \varepsilon_i (W_i V_i - W_i' V_i')\right)\right\} \\
&= \prod_{i\leq n} \mathbb{E}\exp\left\{\lambda \varepsilon_i W_i V_i\right\} \mathbb{E}\exp\left\{-\lambda \varepsilon_i W_i V_i\right\} \\
&\leq \prod_{i\leq n} \mathbb{E}\exp\left\{\lambda^2 V_i^2\right\}.
\end{aligned}$$

Now using Chernoff bound for any $\lambda > 0$,

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_i (W_i V_i - \mathbb{E}W_i V_i) > t\right) &\leq e^{-n\lambda t}\, \mathbb{E}\exp\left\{\lambda \sum_i (W_i V_i - \mathbb{E}W_i V_i)\right\} \\
&\leq e^{-n\lambda t}\, \mathbb{E}\exp\left\{\lambda^2 \sum_i V_i^2\right\}.
\end{aligned}$$

Now note that $V_i = \sum_l \overline{h}_0(X_{il})\, \Theta^*_{lj}$ is Lipschitz continuous with Lipschitz constant $\sqrt{2/\pi}\|\mathbf{\Theta}^*_{\bullet j}\|_2$ and $\mathbb{E}V_i = 0$, so that using Gaussian concentration of Lipschitz functions (see Borell [1975]),

$$\mathbb{P}\left(|V_i| > \sqrt{2/\pi}\|\mathbf{\Sigma}\|_S^{1/2}\|\mathbf{\Theta}^*_{\bullet j}\|_2 t\right) \leq 2e^{-t^2/2},$$

so that $\mathbb{E}\exp(tV_i) \leq \exp\{(\|\mathbf{\Sigma}\|_S\|\mathbf{\Theta}^*_{\bullet j}\|_2^2 t^2)/\pi\}$. Let $c = (\|\mathbf{\Sigma}\|_S\|\mathbf{\Theta}^*_{\bullet j}\|_2^2)/\pi$. Now using

[Han and Liu, 2013, Lemma A.1], it follows that

$$\mathbb{E}\exp\left(\frac{V_i}{\sqrt{12c}}\right)^2 \leq 2.$$

Let us write $\lambda_0^2 = 1/(12c)$ and choose $\lambda = \sqrt{(1/n)\log(2p/\epsilon)}$. Note that $\lambda < \lambda_0$ for n large enough. Now take $t = C\lambda$, where choice of C will be specified below. We have,

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_i (W_iV_i - \mathbb{E}W_iV_i) > C\lambda\right) &\leq e^{-Cn\lambda^2}\left[\mathbb{E}\exp\left\{\lambda^2 V_i^2\right\}\right]^n \\ &\leq e^{-Cn\lambda^2}\left[\mathbb{E}\exp\left\{\lambda_0^2 V_i^2\right\}\right]^{n\lambda^2/\lambda_0^2} \\ &\leq e^{-n\lambda^2(C-\log 2/\lambda_0^2)}.\end{aligned}$$

Here in the second inequality, we have used the fact that since $x \to x^\alpha$ is concave for $\alpha < 1$, and applied Jensen's inequality. Taking $C = (12\log 2/\pi)\|\mathbf{\Sigma}\|_S\|\mathbf{\Theta}^*_{\bullet j}\|_2^2 + 1$, we have,

$$\mathbb{P}\left(\sum_l [\mathbf{\Delta}_n^{(0)}]_{k,l}\Theta^*_{lj} > C\sqrt{\frac{\log(2p/\epsilon)}{n}}\right) \leq \frac{\epsilon}{2p}.$$

The other direction also follows similarly using Chernoff bound. Finally the lemma follows using union bound and the fact that $(12\log 2)/\pi \leq 4$. ■

Lemma 3.2. *Take $0 < \epsilon < 1/2$. Then with probability at least $1 - 2\epsilon$, we have*

$$\|[\mathbf{\Delta}_n^{(1)} - \mathbf{\Delta}_n^{(0)}]_{-j,\bullet}\mathbf{\Theta}^*_{\bullet,j}\|_\infty \leq \|\mathbf{\Theta}^*_{\bullet j}\|_2(\|\mathbf{\Sigma}\|_{2,\infty} + 1)\sqrt{\frac{\log(p/\epsilon)}{n}}.$$

Proof of Lemma 3.2. Let us write $g_{kl}(X_{ik}, X_{il}) = \overline{h}_{kl}(X_{ik}, X_{il}, \Sigma_{kl}) - \overline{h}_0(X_{ik})\,\overline{h}_0(X_{il})$ and let $M_{kl} = \mathbb{E}g_{kl}(X_{ik}, X_{il})$.

$$\sum_l [\mathbf{\Delta}_n^{(1)} - \mathbf{\Delta}_n^{(0)}]_{k,l}\Theta^*_{lj} = \frac{1}{n}\sum_{i=1}^n \sum_l (g_{kl}(X_{ik}, X_{il}) - M_{kl})\Theta^*_{lj}.$$

Using [Mitra and Zhang, 2014a, Lemma 6], note that, with $\mathbf{X} = (X_{ik}), \mathbf{Y} = (Y_{ik}) \in$

$\mathbb{R}^{n\times p}$

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{l}(g_{kl}(X_{ik},X_{il})-M_{kl})\Theta^*_{lj}-\frac{1}{n}\sum_{i=1}^{n}\sum_{l}(g_{kl}(Y_{ik},Y_{il})-M_{kl})\Theta^*_{lj}\right|\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{l}|X_{ik}-Y_{ik}||\Sigma_{kl}||\Theta^*_{lj}|+\frac{1}{n}\sum_{i=1}^{n}\sum_{l}|X_{il}-Y_{il}||\Sigma_{kl}||\Theta^*_{lj}|\\
&\leq \frac{1}{\sqrt{n}}\|\mathbf{\Sigma}_{\bullet k}\|_2\|\mathbf{\Theta}^*_{\bullet j}\|_2\|\mathbf{X}_{\bullet k}-\mathbf{Y}_{\bullet k}\|_2+\frac{1}{\sqrt{n}}\|\mathbf{\Theta}^*_{\bullet j}\|_2\|\mathbf{X}-\mathbf{Y}\|_F\\
&\leq \frac{1}{\sqrt{n}}\|\mathbf{\Theta}^*_{\bullet j}\|_2(\|\mathbf{\Sigma}_{\bullet k}\|_2+1)\|\mathbf{X}-\mathbf{Y}\|_F.
\end{aligned}$$

Thus by Gaussian concentration of Lipschitz continuity we have,

$$\mathbb{P}\left(|\sum_{l}[\mathbf{\Delta}^{(1)}_n-\mathbf{\Delta}^{(0)}_n]_{k,l}\Theta^*_{lj}|\leq\|\mathbf{\Theta}^*_{\bullet j}\|_2(\|\mathbf{\Sigma}_{\bullet k}\|_2+1)\sqrt{\frac{t}{n}}\right)>2e^{-t/2}.$$

We have used the fact that $\mathbb{E}\sum_l[\mathbf{\Delta}^{(1)}_n-\mathbf{\Delta}^{(0)}_n]_{k,l}\Theta^*_{lj}=0$. Thus the final Lemma follows by union bound and taking $e^{-t/2}=\epsilon$. ■

Now consider the second order degenerate term in the U-statistics representation of Kendall's tau.

Lemma 3.3. *For any $\epsilon>0$ with probability at least $1-\epsilon$, we have*

$$\|[(\widehat{\boldsymbol{T}}-\boldsymbol{T})-2\mathbf{\Delta}^{(1)}_n]_{-j,\bullet}\mathbf{\Theta}^*_{\bullet,j}\|_\infty\leq C\|\mathbf{\Theta}^*_{\bullet j}\|_2\left\{d\log(2d/\epsilon)/n\right\},$$

where $C>0$ are fixed constants.

Proof of Lemma 3.3. For brevity let us write$(\widehat{\mathbf{T}}-\mathbf{T})-2\mathbf{\Delta}^{(1)}_n\equiv\mathbf{D}=(D_{jk})_{p\times p}$. We have $\|\mathbf{D}_{-j,\bullet}\mathbf{\Theta}^*_{\bullet,j}\|_\infty=\max_{k\neq j}|\sum_{l=1}^{p}D_{kl}\Theta^*_{lj}|\leq\max_{k\neq j}\|\mathbf{D}_{A_k,A_k}\|_S\|\mathbf{\Theta}^*_{\bullet j}\|_2$. Now note that $\|\mathbf{D}_{A_k,A_k}\|_S\leq d\|\mathbf{D}_{A_k,A_k}\|_{\max}$. From exponential inequality for degenerate U-statistics it is clear that $\mathbb{P}(\|\mathbf{D}_{A_k,A_k}\|_{\max}>Ct/n)\leq 4d^2e^{-t}$. putting $4d^2e^{-t}=\epsilon$ we have the result. ■

The final lemma handles the Taylor's expansion term.

Lemma 3.4. *For $\epsilon > 0$, with probability at least $1 - \epsilon$,*

$$\begin{aligned}&\|[(\widehat{\boldsymbol{\Sigma}}^{\tau} - \boldsymbol{\Sigma}) - \frac{\pi}{2}(\widehat{\boldsymbol{T}} - \boldsymbol{T})]_{-j,\bullet}\boldsymbol{\Theta}^*_{\bullet,j}\|_\infty \\ &\leq \|\boldsymbol{\Theta}^*_{\bullet j}\|_2 \left\{ \frac{\pi}{2}\|\boldsymbol{T}\|_{2,\infty}\sqrt{\frac{\log(dp/\epsilon)}{n_2}} + \frac{\pi^2}{8}\frac{\sqrt{d}\log(dp/\epsilon)}{n_2} \right\}.\end{aligned}$$

Proof of Lemma 3.4. Let us denote $(\widehat{\boldsymbol{\Sigma}}^{\tau} - \boldsymbol{\Sigma}) - \frac{\pi}{2}(\widehat{\mathbf{T}} - \mathbf{T}) = \mathbf{M} = (M_{jk})_{p\times p}$. As before, $\|\mathbf{M}_{-j,\bullet}\boldsymbol{\Theta}^*_{\bullet j}\|_\infty = \max_{k\neq j}|\sum_{l=1}^p M_{kl}\Theta^*_{lj}|$. Define $A_j \subset \{1,\cdots,p\}$ such that $|A_j| = |\{l : \Theta^*_{lj} \neq 0\}| \leq \deg(\boldsymbol{\Theta}^*) \equiv d\ \forall j$. By Wegkamp and Zhao [2013], Lemma 5 in Mitra and Zhang [2014a], we have

$$|M_{kl}| \leq \pi/2|\tau_{kl}(\widehat{\tau}_{kl} - \tau_{kl})| + \pi^2/8|\widehat{\tau}_{kl} - \tau_{kl}|^2,$$

which implies that

$$\begin{aligned}\max_{k\neq j}|\sum_{l=1}^p M_{kl}\Theta^*_{lj}| \leq \frac{\pi}{2}\|\boldsymbol{\Theta}^*_{\bullet j}\|_2\|\mathbf{T}\|_{2,\infty}\max_{k\neq j, l\in A_k}|\widehat{\tau}_{kl} - \tau_{kl}| \\ + \sqrt{d}\frac{\pi^2}{8}\|\boldsymbol{\Theta}^*_{\bullet j}\|_2\max_{k\neq j, l\in A_k}|\widehat{\tau}_{kl} - \tau_{kl}|^2.\end{aligned}$$

From Hoeffding's inequality it follows that $\mathbb{P}(\max_{k\neq j, l\in A_k}|\widehat{\tau}_{kl} - \tau_{kl}| > t) \leq e^{-n_2t^2 + \log dp}$. Setting $e^{-n_2t^2+\log dp} = \epsilon$ yields the result. ■

Proof of Theorem 3.3 follows.

Proof of Theorem 3.3. The theorem follows directly from Lemmas 3.1, 3.2, 3.3, 3.4. Final result follows using the fact that $\|\mathbf{T}\|_S \leq \|\boldsymbol{\Sigma}\|_S$. Also, we use the fact that for any symmetric matrix $\mathbf{A}$, $\|\mathbf{A}\|_{2,\infty} \leq \|\mathbf{A}\|_S$. ■

3.4 Proof of Main Theorem

Proof of Theorem 3.1. Let us consider the following definitions using notations from Sun and Zhang [2013]. Let $A, B \subset \{1, \cdots, p\}$ with $A \cap B = \emptyset$. Also let $|A| = a, |B| \leq b$. Define

$$\delta_a^{\pm}(\widehat{\mathbf{\Sigma}}^{\tau}) = \max_{A, \boldsymbol{u}: \|\boldsymbol{u}\|_2 = 1} \pm\{\|\widehat{\mathbf{\Sigma}}^{\tau}_{A\times A}\boldsymbol{u}\|_2 - 1\}, \quad \theta^2_{a,b}(\widehat{\mathbf{\Sigma}}^{\tau}) = \max_{A,B,\boldsymbol{u},\boldsymbol{v}: \|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1} \boldsymbol{v}^T \widehat{\mathbf{\Sigma}}^{\tau}_{A,B}\boldsymbol{u}/n. \tag{3.4.1}$$

Note that, from (3.2.6), we have

$$\max_{A:|A|\leq m} \|(\widehat{\mathbf{\Sigma}}^{\tau} - \mathbf{\Sigma})_{A\times A}\|_S \leq C\|\mathbf{\Sigma}\|_S(\|\mathbf{\Sigma}\|_S^{1/2} + 1)\sqrt{m\log(ep/\epsilon)/n} = \rho^*(\sqrt{\rho^*} + 1)c, \text{ say,} \tag{3.4.2}$$

with probability at least $1 - \epsilon$ for n large enough and some fixed constant $C > 0$. In (3.4.2), we have used the definition $c = C_2\sqrt{m\log(ep/\epsilon)/n}$ and $\rho^* = \|\mathbf{\Sigma}\|_S$. It is to be noted that in general $\widehat{\mathbf{\Sigma}}^{\tau}$ may not be positive semi-definite and thus a similar result for $1 - \delta_m^-(\widehat{\mathbf{\Sigma}}^{\tau})$ is not available. Now note that using shifting inequality from Ye and Zhang [2010], Cai et al. [2010a], for $l \geq d$, and setting $m = 4l$,

$$\begin{aligned}
\text{SCIF}_1(\xi, S_j; \widehat{\mathbf{\Sigma}}^{\tau}_{-j,-j}) &\geq \frac{1}{1+\xi}\left\{1 - \delta^-_{|S_j|+l}(\widehat{\mathbf{\Sigma}}^{\tau}_{-j,-j}) - \xi\sqrt{\frac{|S_j|}{4l}}\theta^{(2)}_{4l,4l+|S_+j|}(\widehat{\mathbf{\Sigma}}^{\tau}_{-j,-j})\right\} \\
&\geq \frac{1}{1+\xi}\left\{1 - \delta^-_{4l}(\widehat{\mathbf{\Sigma}}^{\tau}) - \xi\sqrt{\frac{d}{4l}}(1 + \delta^+_{4l}(\widehat{\mathbf{\Sigma}}^{\tau}))\right\} \\
&\geq \frac{1}{1+\xi}\left\{\min_{A:|A|\leq 4l}\lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\tau}_{A\times A}) - \xi\sqrt{\frac{d}{4l}}\max_{A:|A|\leq 4l}\|\widehat{\mathbf{\Sigma}}^{\tau}\|_S\right\} \\
&\geq \frac{1}{1+\xi}\left\{\rho_* - \rho^*(\sqrt{\rho^*} + 1)c - \xi\sqrt{\frac{d}{4l}}(\rho^* + \rho^*(\sqrt{\rho^*} + 1)c)\right\}.
\end{aligned}$$

The last line of the string of inequalities above follows from Weyl's inequality and (3.4.2) as follows. First note that $\lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\tau}_{A\times A}) \geq \lambda_{\min}(\mathbf{\Sigma}_{A\times A}) - \|(\widehat{\mathbf{\Sigma}}^{\tau} - \mathbf{\Sigma})_{A\times A}\|_S$, so

that taking minimum over all sets A such that $|A| \leq 4l$

$$\begin{aligned}\min_{A:|A|\leq 4l} \lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\tau}_{A\times A}) &\geq \min_{A:|A|\leq 4l} \lambda_{\min}(\mathbf{\Sigma}_{A\times A}) - \max_{A:|A|\leq 4l} \|(\widehat{\mathbf{\Sigma}}^{\tau} - \mathbf{\Sigma})_{A\times A}\|_S \\ &\geq \rho_* - \rho^*(\sqrt{\rho^*}+1)c.\end{aligned}$$

The bound on $\max_{A:|A|\leq 4l} \|\widehat{\mathbf{\Sigma}}^{\tau}\|_S$ follows using triangle inequality and (3.4.2). Now take $l = \xi^2 d(\rho^*/\rho_*)^2 > d$ so that

$$\text{SCIF}_1(\xi, S_j; \widehat{\mathbf{\Sigma}}^{\tau}_{-j,-j}) \geq \frac{\rho_*}{1+\xi}\left\{\frac{1}{2} - (\frac{\rho*}{\rho_*} + \frac{1}{2})(\sqrt{\rho^*}+1)c\right\}.$$

Thus it is clear that under the condition of boundedness of $\rho^* = \|\mathbf{\Sigma}\|_S$, the quantity $\text{SCIF}_1(\xi, S_j; \widehat{\mathbf{\Sigma}}^{\tau}_{-j,-j})$ is bounded from below for all j. Moreover, note that under the condition of the theorem, the quantity $(\rho*/\rho_* + 1/2)(\sqrt{\rho^*}+1)c$ is small. Now we verify the convergence rate for $\widetilde{\mathbf{\Theta}}^{\tau}$ in estimating $\mathbf{\Theta}^*$. Following Sun and Zhang [2013], we have

$$\begin{aligned}&\|\widetilde{\mathbf{\Theta}}^{\tau}_{\bullet j} - \mathbf{\Theta}^*_{\bullet j}\|_1 \\ &= \|\left\{-(\mathbf{0}, \widehat{\beta}_{j,j})\widetilde{\Theta}_{jj} - (\boldsymbol{\beta}_{-j,j}, 0)\widetilde{\Theta}_{jj}\right\} - \mathbf{\Theta}^*_{\bullet j} + \left\{-(\widehat{\boldsymbol{\beta}}_{-j,j}, 0) + (\boldsymbol{\beta}_{-j,j}, 0)\right\}\widetilde{\Theta}_{jj}\|_1 \\ &\leq \|\widetilde{\Theta}_{jj}(-\boldsymbol{\beta}_{-j,j} - \boldsymbol{\beta}_{j,j}) - \mathbf{\Theta}^*_{\bullet j}\|_1 + |\widetilde{\Theta}_{jj}|\ \|\widehat{\boldsymbol{\beta}}_{-j,j} - \boldsymbol{\beta}_{-j,j}\|_1 \\ &\leq \|\mathbf{\Theta}^*_{\bullet j}\|_1(\widetilde{\Theta}_{jj}/\Theta^*_{jj} - 1) + |\widetilde{\Theta}_{jj}|\ \|\widehat{\boldsymbol{\beta}}_{-j,j} - \boldsymbol{\beta}_{-j,j}\|_1 \\ &\leq \|\mathbf{\Theta}_{\bullet j}\|_1 \left|(\widehat{\sigma}_j/\sigma_j)^{-2} - 1\right| + (\widehat{\sigma}_j/\sigma^*)^{-2}(\sigma^*_j/\sigma_j)^{-1}\frac{\|\widehat{\boldsymbol{\beta}}_{-j,j} - \boldsymbol{\beta}_{-j,j}\|_1}{\sigma_j\sigma^*_j} \\ &\leq \|\mathbf{\Theta}_{\bullet j}\|_1 C_1\frac{d\log p}{n} + \sigma_j^{-1}\left\{1 + C_2\frac{d\log p}{n}\right\}\sqrt{1 + C_3\sqrt{\frac{\log p}{n}}}\left\{C_4 d\sqrt{\frac{\log p}{n}}\right\}.\end{aligned}$$

Here we have used the fact that $\beta_{jj} = \widehat{\beta}_{jj} = 1$ and that

$$
\begin{aligned}
&|(\sigma_j/\widehat{\sigma}_j)^2 - 1| \\
&= |(\sigma_j/\sigma_j^*)^2(\sigma_j^*/\widehat{\sigma}_j)^2 - 1| \\
&\leq |(\sigma_j/\sigma_j^*)^2 - 1||(\sigma_j^*/\widehat{\sigma}_j)^2 - 1| + |(\sigma_j/\sigma_j^*)^2 - 1| + |(\sigma_j^*/\widehat{\sigma}_j)^2 - 1| \\
&= \mathcal{O}_{\mathbb{P}}(\sqrt{\log p/n})\ \mathcal{O}_{\mathbb{P}}(d\log p/n) + \mathcal{O}_{\mathbb{P}}(\sqrt{\log p/n}) + \mathcal{O}_{\mathbb{P}}(d\log p/n) \quad \text{for all } j \\
&= \mathcal{O}_{\mathbb{P}}(d\log p/n) \quad \text{for all } j.
\end{aligned}
$$

Thus it follows directly that

$$
\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_1 \leq C_1'\|\boldsymbol{\Theta}^*\|_1 \frac{d\log p}{n} + \frac{C_2'}{\sigma_j} d\sqrt{\frac{\log p}{n}}.
$$

Now from (3.2.16) the final statement of the theorem for $\widehat{\boldsymbol{\Theta}}^{\tau}$ follows. ■

Chapter 4
Inference for Grouped Variables

4.1 Introduction

We consider the linear regression model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4.1.1}$$

where $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n \times p}$ is a design matrix, $\boldsymbol{y} \in \mathbb{R}^n$ is a response vector, $\boldsymbol{\varepsilon} \sim \mathsf{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ with an unknown noise level σ, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ is a vector of unknown regression coefficients. We are interested in making statistical inference about a group of coefficients $\boldsymbol{\beta}_G = (\beta_j, j \in G)^T$. For small p, the F-distribution, which is approximately chi-square with proper normalization, provides classical confidence regions for $\boldsymbol{\beta}_G$ and p-values for testing $\boldsymbol{\beta}_G$. We want to construct approximate versions of such procedures for potentially large groups in high-dimensional models where p is large, possibly much larger than n.

For individual regression coefficients, Zhang and Zhang [2014] proposed a low-dimensional projection estimator (LDPE) for regular statistical inference at the parametric $n^{-1/2}$ rate under proper conditions. Their results provide

$$\sqrt{n}\big(\widehat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}_G\big) = \mathsf{N}_{|G|}\big(\mathbf{0}, \sigma^2 \mathbf{V}_{G,G}\big) + \mathrm{Rem}_G \tag{4.1.2}$$

along with known covariance structure $\mathbf{V}_{G,G}$ and sufficient conditions for the asymptotic normality, $\mathrm{Rem}_G = o(1)$, when the group size $|G|$ is bounded. For random

designs, the above covariance structure matches the Fisher information in the least favorable sub-model in a general context as described in Zhang [2011], and a proof of the asymptotic efficiency of the LDPE was provided in van de Geer et al. [2014]. Earlier, Sun and Zhang [2012a] proved the consistency and efficiency of a scaled Lasso estimate of the noise level σ. However, the analysis of the LDPE, which guarantees $\|\text{Rem}_G\|_\infty \lesssim \|\boldsymbol{\beta}\|_0(\log p)/\sqrt{n}$, does not directly imply sharp error bound for the ℓ_2- or equivalently chi-square-based group inference for large groups. As $\text{Var}(\chi_{|G|}) \approx 1/2$, the trivial bound $\|\text{Rem}_G\|_2 \lesssim |G|^{1/2}\|\boldsymbol{\beta}\|_0(\log p)/\sqrt{n}$ yields an extra $\sqrt{|G|}$ factor. Thus, the group inference problem is unsolved when one is unwilling to impose the condition $|G|^{1/2}\|\boldsymbol{\beta}\|_0(\log p)/\sqrt{n} \to 0$. Our goal is to construct $\widehat{\boldsymbol{\beta}}_G$ satisfying $\|\text{Rem}_G\|_2 = o(1)$ in an expansion of form (4.1.2) with moderately large $|G|$. The impact of such a result is certainly beyond F- or chi-square-type statistical inference.

Our approach is based on the natural idea that group sparsity can be exploited in statistical inference of variable groups. To this end, we propose to use an estimated efficient score matrix to correct the bias of a scaled group Lasso estimator. This combines and extends the ideas of the group Lasso Yuan and Lin [2006], scaled Lasso and LDPE, and will be shown to captures the benefit of group sparsity in both high-dimensional estimation as in Huang and Zhang [2010] and in bias correction.

The type of statistical inference under consideration here is regular in the sense that it does not require model selection consistency, and that it attains asymptotic efficiency in the sense of Fisher information without being super-efficient. A characterization of such inference is that it does not require a uniform signal strength condition on informative features, e.g. a lower bound on the non-zero $|\beta_j|$ above an inflated noise level due to model uncertainly adjustment, known as the "beta-min" condition. Many attempts have been made to assess the model selected by high dimensional regularizers; For example, some early work was done in Knight and Fu

[2000], sample splitting was considered in Wasserman and Roeder [2009] and Meinshausen et al. [2009], and subsampling was considered in Meinshausen and Bühlmann [2010] and Shah and Samworth [2013]. Leeb and Potscher [2006] proved that the sampling distribution of statistics based on selected models is not estimable. Berk et al. [2010] and Laber and Murphy [2011] proposed conservative approaches. Alternative approaches were proposed in Lockhart et al. [2014] and Meinshausen [2014].

The basic idea of Zhang and Zhang [2014] and Zhang [2011] is to correct the bias of high-dimensional regularized estimators by projecting its residual to a direction close to that of the efficient score. Such bias correction, which has been called de-biasing, is parallel to correcting the bias of nonparametric estimators in semiparametric inference Bickel et al. [1993]. Bühlmann [2013] adopted a similar approach to correct the bias of ridge regression. van de Geer et al. [2014] considered an extension to generalized linear model. Javanmard and Montanari [2014] obtained sharper results for Gaussian designs. Belloni et al. [2014] considered estimation of treatment effects with a large number of controls. Sun and Zhang [2012b], Ren et al. [2013] and Jankova and van de Geer [2014] considered extensions to graphical models and precision matrix estimation.

Since our proposed method relies upon group regularized initial estimator, in the following we provide a brief discussion of the literature on the topic. The group Lasso Yuan and Lin [2006] can be defined as

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\omega}) = \arg\min_{\boldsymbol{\beta}} \mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}), \quad \mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}) = \frac{\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2n} + \sum_{j=1}^{M} \omega_j \|\boldsymbol{\beta}_{G_j}\|_2, \tag{4.1.3}$$

where $\{G_j, , 1 \le j \le M\}$ forms a partition of the index set $\{1, \ldots, p\}$ of variables. It is worthwhile to note that when the group effects are being regularized, the choice of basis $\mathbf{X}_{G_j} = (\boldsymbol{x}_k, k \in G_j)$ within the group may not play a prominent role, so that the design is often "pre-normalized" to satisfy $\mathbf{X}_{G_j}^T \mathbf{X}_{G_j}/n = \mathbf{I}_{G_j \times G_j}$ as in Yuan

and Lin [2006]. The group Lasso and its variants have been studied in Bach [2008], Koltchinskii and Yuan [2008], Obozinski et al. [2008], Nardi and Rinaldo [2008], Liu and Zhang [2009], Huang and Zhang [2010], and Lounici et al. [2011] among many others. Huang and Zhang [2010] characterized the benefit of group Lasso in ℓ_2 estimation, versus the Lasso Tibshirani [1996a], under the assumption of strong group sparsity; see (4.2.1). Huang et al. [2009] and Breheny and Huang [2011] developed methodologies for concave group and bilevel regularization. We refer to Bühlmann and van de Geer [2011] and Huang et al. [2012] for further discussion and additional references. More recently, In Bunea et al. [2014], the authors developed a square root group Lasso procedure based on square root Lasso , developed in Belloni et al. [2011], that achieves optimal estimation properties with a tuning sequence that bypasses the need to estimate the scale parameter of the noise.

This paper is organized as follows. In Section 4.2, we describe the main results of the paper on statistical inference of variable groups. In Section 4.3, we study a scaled group Lasso needed for the construction in Section 4.2. In Section 4.4, we present some simulation results to demonstrate the feasibility and performance of the proposed methods.

4.2 Group Inference

We present our results in five subsections. In Subsection 4.2.1 describes our working assumption on the availability of a certain initial estimates of $\boldsymbol{\beta}$ and σ. The working assumption is based on the existing literature on group Lasso and will be verified in Section 4.3 under proper conditions. In Subsection 4.2.2 develops bias correction formulations as extension from statistical inference of real parameters. Subsection 4.2.3 provides optimization strategies (see equations (4.2.20) and (4.2.23)) for construction of inference procedures for groups of variables. Subsection 4.2.4 provides sufficient conditions (Theorem 4.3) under which a feasible solution to the optimization problem

(4.2.20) is available. Subsection 4.2.5 discusses strategies for finding feasible solutions.

We use the following notation throughout the paper. For vectors $\boldsymbol{u} \in \mathbb{R}^d$, the ℓ_p norm is denoted by $\|\boldsymbol{u}\|_p = (\sum_{k=1}^d |u_k|^p)^{1/p}$, with $\|\boldsymbol{u}\|_\infty = \max_{1\le k\le d} |u_k|$ and $\|\boldsymbol{u}\|_0 = \#\{j : u_j \neq 0\}$. For matrix $\mathbf{A} = (A_{jk})_{d_1\times d_2} \in \mathbb{R}^{d_1\times d_2}$, the spectrum norm is denoted by $\|\mathbf{A}\|_S = \max_{\|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1} \boldsymbol{u}^T\mathbf{A}\boldsymbol{v}$, the Frobenius norm by $\|\mathbf{A}\|_F = \{\text{trace}(\mathbf{A}^T\mathbf{A})\}^{1/2}$, and the nuclear norm by $\|\mathbf{A}\|_N = \max_{\|\mathbf{B}\|_S=1} \text{trace}(\mathbf{B}^T\mathbf{A})$. Given $A \subset \{1, \cdots, p\}$, for any vector $\boldsymbol{u} \in \mathbb{R}^p$, $\boldsymbol{u}_A \in \mathbb{R}^{|A|}$ denotes a vector with corresponding components from $\boldsymbol{u}$, $\mathbf{X}_A \in \mathbb{R}^{n\times|A|}$ denotes the sub-matrix of $\mathbf{X}$ with corresponding columns as indicated by the set A, $\mathbf{X}_{-A}$ denotes the sub-matrix of $\mathbf{X}$ with column indices belonging to the complement of A, and $\mathcal{R}(\mathbf{X}_A)$ denotes the column space spanned by columns of $\mathbf{X}_A$. Additionally, $\mathbb{E}$ and $\mathbb{P}$, denote the expectation and probability measure and $\xrightarrow{\text{D}}$ the convergence in distribution. Finally, $\boldsymbol{\beta}^*$ denotes the true regression coefficient vector.

4.2.1 Working assumption based on strong group sparsity

We assume an inherent and pre-specified non overlapping group structure of the feature set. Put precisely, assume that $\{1, \cdots, p\} = \cup_{j=1}^M G_j$ such that $G_j \cap G_k = \varnothing$. Define $d_j = |G_j|$ for all j so that $\sum_{j=1}^M d_j = p$. For any index set $T \subset \{1, \cdots, M\}$, we define $G_T = \cup_{j\in T} G_j$. In the following, we allow the quantities n, p, M, d_j's etc. to all grow to infinity.

In light of this group structure, further results on consistency of group regularized estimators of $\boldsymbol{\beta}^*$ will be based on a weighted mixed $\ell_{(2,1)}$, defined as $\sum_{j=1}^M \omega_j \|\boldsymbol{u}_{G_j}\|_2$ for $\boldsymbol{u} = (\boldsymbol{u}_{G_j}; 1 \le j \le M) \in \mathbb{R}^p$ with $\boldsymbol{u}_{G_j} \in \mathbb{R}^{|G_j|}$, where $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_M) \in \mathbb{R}^M$ with $\omega_j > 0$ for all j. This norm will be used both as penalty and as a key loss function. Weighted mixture norm of this type provides suitable description of the complexity of the unknown $\boldsymbol{\beta}$ when the following strong group sparsity holds Huang and Zhang [2010].

Strong group sparsity: *With the given group structure* $\{G_j, j = 1, \ldots, M\}$ *as*

a partition of $\{1, \ldots, p\}$*, there exists a group-index set,* $S^* \subset \{1, \cdots, M\}$*, such that*

$$|S^*| \leq g, \quad |G_{S^*}| \leq s, \quad supp(\boldsymbol{\beta}^*) \subset G_{S^*} = \cup_{j \in S^*} G_j. \tag{4.2.1}$$

In this case, we say that the true coefficient vector $\boldsymbol{\beta}^*$ *is* (g, s) *strongly group sparse with group support* S^**.*

Under the strong group sparsity assumption, various error bounds for group regularized methods have been established in the literature as we reviewed in the introduction. With the support of the existing results and our own in Section 4.3, we make the following working assumption.

Working assumption: *Suppose that we have estimators* $\widehat{\boldsymbol{\beta}}^{(init)}$ *and* $\widehat{\sigma}$ *satisfying*

$$\left| \frac{\widehat{\sigma}}{\sigma^*} - 1 \right| + \frac{1}{n^{1/2}} \sum_{j=1}^{M} \omega_j \left\| \boldsymbol{X}_{G_j} \widehat{\boldsymbol{\beta}}^{(init)}_{G_j} - \boldsymbol{X}_{G_j} \boldsymbol{\beta}^*_{G_j} \right\|_2 = \mathcal{O}_{\mathbb{P}} \left(\frac{s + g \log M}{n} \right), \tag{4.2.2}$$

where $\omega_j \propto \sqrt{|G_j|/n} + \sqrt{(2/n) \log M}$*,* $\sigma^* = \|\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{y}\|_2/\sqrt{n}$ *is an oracle estimate of the noise level* σ*, and* G_j*,* s *and* g *are as in (4.2.1).*

As we will prove in Section 4.3, the error bound for $\widehat{\boldsymbol{\beta}}^{(init)}$ in (4.2.2) is attainable under proper conditions on the design matrix if the group Lasso is used with a consistent estimate of σ, and the error bounds for both $\widehat{\boldsymbol{\beta}}^{(init)}$ and $\widehat{\sigma}$ in (4.2.2) are attainable if a scaled group Lasso is used. See Corollaries 4.1 and 4.2. The working assumption exhibits the benefit of strong group sparsity, since a reasonable working assumption under the ℓ_0 sparsity condition $\|\widehat{\boldsymbol{\beta}}\|_0 \leq s$ would be

$$\left| \frac{\widehat{\sigma}}{\sigma^*} - 1 \right| + \left(\frac{\log p}{n} \right)^{1/2} \|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}_{\mathbb{P}} \left(\frac{s \log p}{n} \right). \tag{4.2.3}$$

Although error bounds in (4.2.2) and (4.2.3) do not dominate each other due to different interpretation of s when $\text{supp}(\boldsymbol{\beta}^*) \neq G_{S^*}$, the right-hand side of (4.2.2) is of smaller order when s is of the same order in both settings and $g \ll s$.

4.2.2 Bias correction via relaxed projection

Given a regularized initial estimator $\widehat{\boldsymbol{\beta}}^{(init)}$ of the regression coefficient vector, Zhang and Zhang [2014] proposed to use a relaxed projection to correct the bias of $\widehat{\beta}_j^{(init)}$ via

$$\widehat{\beta}_j = \widehat{\beta}_j^{(init)} + \frac{\boldsymbol{z}_j^T(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(init)})}{\boldsymbol{z}_j^T\boldsymbol{x}_j}, \tag{4.2.4}$$

where $\boldsymbol{z}_j$ is designed to be nearly orthogonal to all $\boldsymbol{x}_k, k \neq j$. For the estimation of $\boldsymbol{\beta}_G$, a formal vectorization of (4.2.4) is

$$\widehat{\boldsymbol{\beta}}_G = \widehat{\boldsymbol{\beta}}_G^{(init)} + (\mathbf{Z}_G^T\mathbf{X}_G)^\dagger\mathbf{Z}_G^T(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(init)}), \tag{4.2.5}$$

where $\mathbf{Z}_G$ is an $n \times |G|$ matrix and $\mathbf{A}^\dagger$ denotes Moore-Penrose pseudo inverse of a matrix $\mathbf{A}$. The problem is to choose $\widehat{\boldsymbol{\beta}}^{(init)}$ and $\mathbf{Z}_G$.

Zhang and Zhang [2014] proposed two choices of $\boldsymbol{z}_j$ to match ℓ_1 regularized initial estimators $\widehat{\boldsymbol{\beta}}^{(init)}$, which naturally controls $\|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}^*\|_1$. The first proposal is a point $\boldsymbol{z}_j \approx \boldsymbol{z}_j^o$ in the Lasso path in the regression of $\boldsymbol{x}_j$ against $\mathbf{X}_{-j} = (\boldsymbol{x}_k, k \neq j)$:

$$\boldsymbol{x}_j = \mathbf{X}_{-j}\boldsymbol{\gamma}_{-j} + \boldsymbol{z}_j^o. \tag{4.2.6}$$

The Karush-Kuhn-Tucker (KKT) conditions for $\boldsymbol{z}_j$ automatically controls $\|\boldsymbol{z}_j^T\mathbf{X}_{-j}\|_\infty$, and thus

$$\left|\widehat{\beta}_j - \beta_j^* - \frac{\boldsymbol{z}_j^T\boldsymbol{\varepsilon}}{\boldsymbol{z}_j^T\boldsymbol{x}_j}\right| = \left|\frac{\boldsymbol{z}_j^T\mathbf{X}_{-j}(\widehat{\boldsymbol{\beta}}_{-j}^{(init)} - \boldsymbol{\beta}_{-j}^*)}{\boldsymbol{z}_j^T\boldsymbol{x}_j}\right| \leq \frac{\|\boldsymbol{z}_j^T\mathbf{X}_{-j}\|_\infty\|\widehat{\boldsymbol{\beta}}_{-j}^{(init)} - \boldsymbol{\beta}_{-j}^*\|_1}{|\boldsymbol{z}_j^T\boldsymbol{x}_j|}$$

in an ℓ_∞-ℓ_1 split. The second proposal of Zhang and Zhang [2014], closely related to the first one and given in the discussion section of their paper, is a constrained

variance minimization scheme

$$\boldsymbol{z}_j = \arg\min_{\boldsymbol{z}} \Big\{ \|\boldsymbol{z}\|_2^2 : |\boldsymbol{z}^T \boldsymbol{x}_j/n| = 1, \max_{k\neq j} |\boldsymbol{z}^T \boldsymbol{x}_k/n| \le \lambda_j' \Big\}. \tag{4.2.7}$$

While the Lasso with penalty level λ_j provides a feasible solution $\boldsymbol{z}_j/\lambda_j' = (\boldsymbol{x}_j - \mathbf{X}_{-j}\widehat{\boldsymbol{\gamma}}_{-j})/\lambda_j$ for (4.2.7), an advantage of (4.2.7) is a guaranteed bias bound

$$\left| \widehat{\beta}_j - \beta_j^* - \frac{\boldsymbol{z}_j^T \boldsymbol{\varepsilon}}{n} \right| \le \frac{\lambda_j' \|\widehat{\boldsymbol{\beta}}_{-j}^{(init)} - \boldsymbol{\beta}_{-j}^*\|_1}{n}$$

whenever the optimization problem is feasible. For Gaussian designs, such feasibility of $\boldsymbol{z} = n\boldsymbol{z}_j^o/\boldsymbol{x}_j^T\boldsymbol{z}_j^o$ follows from an application of the union bound Javanmard and Montanari [2014].

The algebraic extension of the above proposals is straightforward. Write

$$\mathbf{X}_G = \mathbf{X}_{-G}\boldsymbol{\Gamma}_{-G,G} + \mathbf{Z}_G^o. \tag{4.2.8}$$

We may directly approximate $\mathbf{Z}_G^o$ via a regularized multivariate regression in (4.2.8) or mimic properties of $\mathbf{Z}_G^o$ with a regularized optimization scheme. The question is to make a right choice of the regularization to match a proper initial estimator of $\boldsymbol{\beta}$. One possibility is to use an ℓ_1 regularized estimate of $\boldsymbol{\Gamma}_{-G,j}$ in the univariate regression of $\boldsymbol{x}_j$ against $\mathbf{X}_{-G}$ for all individual $j \in G$. This has been considered in van de Geer [2014]. However, the advantage of such a scheme is unclear compared with directly using $(\widehat{\beta}_j, j \in G)^T$ with the $\widehat{\beta}_j$ in (4.2.4). It is worthwhile to mention that the central limit theorem for (4.2.4) came with large deviation bounds to justify Bonferroni adjustments Zhang and Zhang [2014], so that (4.2.4) and its variations can be used to test $H_0 : \boldsymbol{\beta}_G^* = \mathbf{0}$ versus an alternative hypothesis on $\|\boldsymbol{\beta}_G^*\|_\infty$, especially when an ℓ_1 regularized $\widehat{\boldsymbol{\beta}}^{(init)}$ is used van de Geer et al. [2014]. However, we are interested in extensions of traditional F- or chi-square-type tests for ℓ_2 alternatives

and to take advantage of group sparsity of $\boldsymbol{\beta}^*$.

4.2.3 An optimization strategy

In this subsection we propose a multivariate extension of (4.2.7) to match the group structure and weights in our working assumption (4.2.2).

We write (4.2.5) in terms projections so that the resulting optimization scheme will be rotation and scale free within the subspaces under consideration. As our goal in essence is to construct inferential procedure for $\mathbf{X}_G\boldsymbol{\beta}_G$, we rewrite the regression problem (4.1.1) as follows:

$$\boldsymbol{y} = \mathbf{X}_G\boldsymbol{\beta}_G + \sum_{G_k \not\subseteq G} \mathbf{X}_{G_k \setminus G}\boldsymbol{\beta}_{G_k \setminus G} + \boldsymbol{\varepsilon} = \boldsymbol{\mu}_G + \sum_{G_k \not\subseteq G} \boldsymbol{\mu}_{G_k \setminus G} + \boldsymbol{\varepsilon}. \tag{4.2.9}$$

Here and in the sequel, the following notation is used. For any $A \subset \{1, \ldots, p\}$, $\boldsymbol{\mu}_A = \mathbf{X}_A\boldsymbol{\beta}_A$ and $\mathbf{Q}_A$ is the orthogonal projection to $\mathcal{R}(\mathbf{X}_A)$, the column space of $\mathbf{X}_A$, i.e.

$$\mathbf{Q}_A = \mathbf{X}_A(\mathbf{X}_A^T\mathbf{X}_A)^{-1}\mathbf{X}_A^T. \tag{4.2.10}$$

By $\mathbf{Q}_A^\perp$, we denote orthogonal projection into $\mathcal{R}^\perp(\mathbf{X}_A)$. In the simplest case where the variable group of interest matches the group sparsity in the following sense:

$$\mathbf{X}_G\boldsymbol{\beta}_G^* = \sum_{G_k \cap G \neq \emptyset} \mathbf{X}_{G_k}\boldsymbol{\beta}_{G_k}^*, \tag{4.2.11}$$

e.g. $G = G_{j_0}$ for some j_0, (4.2.9) becomes

$$\boldsymbol{y} = \boldsymbol{\mu}_G + \sum_{G_k \cap G = \emptyset} \boldsymbol{\mu}_{G_k} + \boldsymbol{\varepsilon}.$$

Let $\mathbf{P}_G$ be an orthogonal projection matrix close to $\mathbf{Q}_G$ in certain distance and approximately orthogonal to $\mathbf{Q}_{G_k \setminus G}$ for all k with $G_k \not\subseteq G$. We write (4.2.5) in terms of projections as

$$\widehat{\boldsymbol{\beta}}_G = (\mathbf{P}_G \mathbf{X}_G)^\dagger \mathbf{P}_G \left(\boldsymbol{y} - \sum_{G_k \not\subseteq G} \widehat{\boldsymbol{\mu}}^{(init)}_{G_k \setminus G} \right), \quad \text{when } \operatorname{rank}(\mathbf{P}_G \mathbf{X}_G) = |G|, \tag{4.2.12}$$

$$\widehat{\boldsymbol{\mu}}_G = (\mathbf{P}_G \mathbf{Q}_G)^\dagger \mathbf{P}_G \left(\boldsymbol{y} - \sum_{G_k \not\subseteq G} \widehat{\boldsymbol{\mu}}^{(init)}_{G_k \setminus G} \right), \quad \text{when } \|\mathbf{P}_G \mathbf{Q}_G^\perp\|_S < 1, \tag{4.2.13}$$

where $\widehat{\boldsymbol{\mu}}^{(init)}_A = \mathbf{X}_A \boldsymbol{\beta}^{(init)}_A$ with an initial estimator $\widehat{\boldsymbol{\beta}}^{(init)}$. We note that $\|\mathbf{P}_G \mathbf{Q}_G^\perp\|_S < 1$ iff $\operatorname{rank}(\mathbf{P}_G \mathbf{X}_G) = \operatorname{rank}(\mathbf{X}_G)$, so that the condition in (4.2.13) is slightly weaker than the condition in (4.2.12). Moreover, $\|\mathbf{P}_G \mathbf{Q}_G^\perp\|_S = \|\mathbf{P}_G - \mathbf{Q}_G\|_S = \cos\theta_{\min}$ where $\theta_{\min}$ is the minimum principle angle between subspaces $\mathcal{R}(\mathbf{P}_G)$ and $\mathcal{R}^\perp(\mathbf{X}_G)$. Thus, $\|\mathbf{P}_G \mathbf{Q}_G^\perp\|_S = 1$ iff the two subspaces have a nontrivial intersection.

Given $\widehat{\sigma}$ an estimate of the noise level, we test the hypothesis $H_0 : \boldsymbol{\beta}_G = \mathbf{0}$ with the following statistic:

$$T_G = \frac{1}{\widehat{\sigma}} \left\| \mathbf{P}_G \left(\boldsymbol{y} - \sum_{G_k \not\subseteq G} \widehat{\boldsymbol{\mu}}^{(init)}_{G_k \setminus G} \right) \right\|_2 . \tag{4.2.14}$$

A test of this form can be easily converted into elliptical confidence regions for linear mappings of $\boldsymbol{\beta}_G$ in usual way.

Let $\mathbf{P}_G = \mathbf{Z}_G (\mathbf{Z}_G^T \mathbf{Z}_G)^\dagger \mathbf{Z}_G^T$ and assume $\operatorname{rank}(\mathbf{Z}_G^T \mathbf{X}_G) = |G|$. We show that (4.2.12) and (4.2.13) are consistent with (4.2.5) as follows. Since both $\mathbf{Z}_G$ and $\mathbf{X}_G$ are $n \times |G|$ matrices, we have $\operatorname{rank}(\mathbf{X}_G) = \operatorname{rank}(\mathbf{Z}_G) = |G| \leq n$, so that $\operatorname{rank}(\mathbf{P}_G \mathbf{Q}_G) = \operatorname{rank}(\mathbf{P}_G \mathbf{X}_G) = |G|$. It follows that $\mathbf{P}_G \mathbf{X}_G (\mathbf{P}_G \mathbf{X}_G)^\dagger \mathbf{P}_G = \mathbf{P}_G$. As $\mathbf{Z}_G^T \mathbf{X}_G$ is a $|G| \times |G|$ invertible matrix, $\mathbf{P}_G \mathbf{X}_G (\mathbf{Z}_G^T \mathbf{X}_G)^{-1} \mathbf{Z}_G^T = \mathbf{P}_G$. Since $\operatorname{rank}(\mathbf{P}_G \mathbf{X}_G) = |G|$, we are allowed to cancel $\mathbf{P}_G \mathbf{X}_G$ to obtain $(\mathbf{P}_G \mathbf{X}_G)^\dagger \mathbf{P}_G = (\mathbf{Z}_G^T \mathbf{X}_G)^\dagger \mathbf{Z}_G^T$. This provides the consistency between (4.2.12) and (4.2.5). Furthermore, since $\mathbf{X}_G = \mathbf{Q}_G \mathbf{X}_G =$

$(\mathbf{P}_G\mathbf{Q}_G)^\dagger\mathbf{P}_G\mathbf{X}_G$, we also have $\mathbf{X}_G\widehat{\boldsymbol{\beta}}_G = \widehat{\boldsymbol{\mu}}_G$ for the consistency of (4.2.13).

Let $\mathbf{Q}$ be the projection to $\mathcal{R}(\mathbf{X})$. In the low-dimensional case of $\text{rank}(\mathbf{X}) = p < n$, we may set $\mathbf{P}_G = \mathbf{Q}\prod_{G_k \not\subseteq G}\mathbf{Q}^{\perp}_{G_k\setminus G}$, so that (4.2.12) is the least squares estimator of $\boldsymbol{\beta}_G$ and $T_G^2/|G|$ is the F-statistic for testing $H_0 : \boldsymbol{\beta}_G = 0$ when $\widehat{\sigma}$ is the degree adjusted estimate of noise level based on the residuals of the least squares estimator. Of course, we need to relax the requirement of the orthogonality condition $\mathbf{P}_G\mathbf{Q}_{G_k\setminus G} = 0$ for all $G_k \not\subseteq G$ in the high-dimensional case.

To find the proper relaxation, we first inspect the deviation of (4.2.12), (4.2.13) and (4.2.14) from the low-dimensional regression theory. Let $\boldsymbol{\beta}^*$ be the true $\boldsymbol{\beta}$ and $\boldsymbol{\mu}^*_A = \mathbf{X}_A\boldsymbol{\beta}^*_A$ for all $A \subset \{1,\ldots,p\}$. It follows immediately from (4.2.12), (4.2.13) and (4.2.14) that

$$\widehat{\boldsymbol{\beta}}_G = \boldsymbol{\beta}^*_G + (\mathbf{P}_G\mathbf{X}_G)^\dagger\Big(\mathbf{P}_G\boldsymbol{\varepsilon} - \text{Rem}_G\Big), \quad \text{when } \text{rank}(\mathbf{P}_G\mathbf{X}_G) = |G|, \tag{4.2.15}$$

$$\widehat{\boldsymbol{\mu}}_G = \boldsymbol{\mu}^*_G + (\mathbf{P}_G\mathbf{Q}_G)^\dagger\Big(\mathbf{P}_G\boldsymbol{\varepsilon} - \text{Rem}_G\Big), \quad \text{when } \|\mathbf{P}_G\mathbf{Q}^{\perp}_G\|_S < 1, \tag{4.2.16}$$

with a remainder term

$$\text{Rem}_G = \sum_{G_k\not\subseteq G}\mathbf{P}_G\Big(\widehat{\boldsymbol{\mu}}^{(init)}_{G_k\setminus G} - \boldsymbol{\mu}^*_{G_k\setminus G}\Big) = \sum_{G_k\not\subseteq G}\Big(\mathbf{P}_G\mathbf{Q}_{G_k\setminus G}\Big)\Big(\widehat{\boldsymbol{\mu}}^{(init)}_{G_k\setminus G} - \boldsymbol{\mu}^*_{G_k\setminus G}\Big).$$

Moreover, when $\boldsymbol{\beta}^*_G = \mathbf{0}$,

$$\left|T_G - \frac{\|\boldsymbol{P}_G\boldsymbol{\varepsilon}\|_2}{\sigma}\right| \le \frac{\big\|\text{Rem}_G\big\|_2}{\widehat{\sigma}} + \left|\frac{\sigma}{\widehat{\sigma}} - 1\right|\frac{\|\boldsymbol{P}_G\boldsymbol{\varepsilon}\|_2}{\sigma}. \tag{4.2.17}$$

As $\mathbf{P}_G$ is an orthogonal projection matrix depending on $\mathbf{X}$ only, $\mathbf{P}_G\boldsymbol{\varepsilon}/\sigma$ is a standard normal vector living in the image of $\mathbf{P}_G$ and $\|\mathbf{P}_G\boldsymbol{\varepsilon}/\sigma\|_2^2$ has the chi-square distribution with $\text{rank}(\mathbf{P}_G)$ degrees of freedom. Thus, chi-square based inference can be carried out using the projection estimators in (4.2.12) and (4.2.13) and test statistic T_G in

(4.2.14) under proper conditions on $\|\text{Rem}_G\|_2$ and $\widehat{\sigma}$. For example,

$$\begin{cases} \sup_t \left|\mathbb{P}\Big\{\|(\mathbf{P}_G\mathbf{X}_G)(\widehat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}^*_G)\|_2 \le \widehat{\sigma}t\Big\} - \mathbb{P}\Big\{\chi^2_{|G|} \le t\Big\}\right| \to 0, \\ \sup_t \left|\mathbb{P}\Big\{\|(\mathbf{P}_G\mathbf{Q}_G)(\widehat{\boldsymbol{\mu}}_G - \boldsymbol{\mu}^*_G)\|_2 \le \widehat{\sigma}t\Big\} - \mathbb{P}\Big\{\chi^2_{|G|} \le t\Big\}\right| \to 0, \\ \boldsymbol{\mu}^*_G = \mathbf{0} \;\Rightarrow\; \sup_t \left|\mathbb{P}\Big\{T^2_G \le t\Big\} - \mathbb{P}\Big\{\chi^2_{|G|} \le t\Big\}\right| \to 0, \end{cases} \tag{4.2.18}$$

under the conditions $\|\text{Rem}_G\|_2 \to 0$, $\text{rank}(\mathbf{P}_G) = |G|$ and $|G|^{1/2}(\widehat{\sigma}/\sigma - 1) \to 0$.

We still need to find an upper bound for $\|\text{Rem}_G\|_2$. To this end we use (4.2.2) to obtain

$$\begin{aligned} \|\text{Rem}_G\|_2 &\le \left(\max_{G_k \not\subseteq G} M_k \omega_k^{-1} \|\mathbf{P}_G\mathbf{Q}_{G_k}\|_S\right) \sum_{G_k \not\subseteq G} \omega_k \|\widehat{\boldsymbol{\mu}}^{(init)}_{G_k} - \boldsymbol{\mu}_{G_k}\|_2 \\ &= \mathcal{O}_{\mathbb{P}}\left(\frac{s + g\log M}{n^{1/2}}\right)\left(\max_{G_k \not\subseteq G} M_k \omega_k^{-1} \|\mathbf{P}_G\mathbf{Q}_{G_k}\|_S\right), \end{aligned} \tag{4.2.19}$$

where $M_k = \max_{\|\mathbf{X}_{G_k}\boldsymbol{u}_{G_k}\|_2=1} \|\mathbf{X}_{G_k\setminus G}\boldsymbol{u}_{G_k\setminus G}\|_2$. We note that $M_k = 1$ when $\mathbf{X}^T_{G_k}\mathbf{X}_{G_k}/n = \mathbf{I}_{d_k\times d_k}$. The error bound in (4.2.19) motivates the following extension of (4.2.7):

$$\mathbf{P}_G = \arg\min_{\mathbf{P}} \Big\{\|\mathbf{P}\mathbf{Q}^{\perp}_G\|_S : \mathbf{P} = \mathbf{P}^2 = \mathbf{P}^T,\ \|\mathbf{P}_G\mathbf{Q}_{G_k\setminus G}\|_S \le \omega'_k\ \forall\ G_k \not\subseteq G\Big\}. \tag{4.2.20}$$

We say that $\mathbf{P}_G$ is a feasible solution of (4.2.20) if it satisfies all the constraints. We summarize the above analysis in the following theorem.

Theorem 4.1. *Let $\widehat{\boldsymbol{\beta}}_G$ be given by (4.2.12) and T_G by (4.2.14) with a feasible solution $\boldsymbol{P}_G$ of (4.2.20) with rank$(\boldsymbol{P}_G) = |G|$. Suppose that (4.2.2) holds for $\widehat{\boldsymbol{\beta}}^{(init)}$ and $\widehat{\sigma}$, and*

$$\frac{|G|}{n} \to 0, \quad \frac{s + g\log M}{n^{1/2}}\left(\frac{|G|^{1/2}}{n^{1/2}} + \max_{G_k \not\subseteq G} M_k \frac{\omega'_k}{\omega_k}\right) \to 0, \tag{4.2.21}$$

with the M_k in (4.2.19). Then, (4.2.18) holds. In particular, with $\|Rem_G\|_2 = o_{\mathbb{P}}(1)$,

$$(\boldsymbol{P}_G\boldsymbol{X}_G)\big(\widehat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}^*_G\big) = \mathsf{N}(\mathbf{0}, \sigma^2\boldsymbol{P}_G) + Rem_G. \tag{4.2.22}$$

Remark 4.1. The optimization problem (4.2.20) also provides geometric insights. As we have mentioned earlier, the quantity $\|\mathbf{P}_G\mathbf{Q}_G^{\perp}\|_S$, which equals $\|\mathbf{P}_G - \mathbf{Q}_G\|_S$, is the so-called 'gap' between the subspaces spanned by $\mathbf{P}_G$ and $\mathbf{Q}_G$, which we try to minimize. This minimization is done subject to upper-bounds on $\|\mathbf{P}_G\mathbf{Q}_{G_k\setminus G}\|_S$. When $p < n$ and $\omega_k' = 0$, $\boldsymbol{P}_G$ in (4.2.20) is the projection to the orthogonal complement of $\sum_{G_k\not\subseteq G}\mathcal{R}(\mathbf{X}_{G_k\setminus G})$ in $\mathcal{R}(\mathbf{X})$, or equivalently the linear space $\big(\prod_{G_k\not\subseteq G}\mathbf{Q}_{G_k\setminus G}^{\perp}\big)\mathcal{R}(\mathbf{X})$.

Proof of Theorem 4.1. It follows from (4.2.19) and the feasibility of $\mathbf{P}_G$ in (4.2.20) that

$$\|\mathrm{Rem}_G\|_2 = o_{\mathbb{P}}(1)$$

in (4.2.15), (4.2.16) under condition (4.2.21) and (4.2.17). In addition, (4.2.2) and (4.2.21) imply

$$\left|\frac{\sigma}{\widehat{\sigma}} - 1\right| = o_{\mathbb{P}}(|G|^{-1/2}) + O_{\mathbb{P}}(n^{-1/2}) = o_{\mathbb{P}}(|G|^{-1/2}),$$

so that by (4.2.17)

$$\left|T_G - \frac{\|\boldsymbol{P}_G\boldsymbol{\varepsilon}\|_2}{\sigma}\right| \le o_{\mathbb{P}}(1) + o_{\mathbb{P}}(1)\frac{\|\boldsymbol{P}_G\boldsymbol{\varepsilon}\|_2}{\sigma|G|^{1/2}} = o_{\mathbb{P}}(1).$$

The conclusions follow immediately. ■

A modification of (4.2.20), which removes the factors M_k in condition (4.2.21), is to write

$$\boldsymbol{y} = \widetilde{\mathbf{X}}_G\boldsymbol{\beta}_G + \sum_{G_k\not\subseteq G}\mathbf{Q}_{G_k\setminus G}\boldsymbol{\mu}_{G_k} + \boldsymbol{\varepsilon},$$

where $\widetilde{\mathbf{X}}_G$ is a $n \times |G|$ matrix defined by $\widetilde{\mathbf{X}}_G\boldsymbol{v}_G = \sum_{k=1}^{M}\big(\mathbf{Q}_{G_k\setminus G}^{\perp}\mathbf{X}_{G_k\cap G}\big)\boldsymbol{v}_{G\cap G_k}$. We note that $\widetilde{\mathbf{X}}_G = \mathbf{X}_G$ when $\mathbf{X}_{G_k}^T\mathbf{X}_{G_k}/n = \mathbf{I}_{G_k\times G_k}$ for all k with $0 < |G_k \setminus G| < |G_k|$.

Let $\widetilde{\mathbf{Q}}_G$ be the projection to the column space of $\widetilde{\mathbf{X}}_G$. The optimization scheme and statistical methods are changed accordingly as follows:

$$\begin{aligned}
\mathbf{P}_G &= \arg\min_{\mathbf{P}} \left\{ \|\mathbf{P}\widetilde{\mathbf{Q}}_G^{\perp}\|_S : \mathbf{P} = \mathbf{P}^2 = \mathbf{P}^T,\ \|\mathbf{P}_G \mathbf{Q}_{G_k \setminus G}\|_S \le \omega_k' \ \forall\ k \right\}, \\
\widehat{\boldsymbol{\beta}}_G &= (\mathbf{P}_G \widetilde{\mathbf{X}}_G)^{\dagger} \mathbf{P}_G \left(\boldsymbol{y} - \sum_{G_k \not\subseteq G} \mathbf{Q}_{G_k \setminus G} \widehat{\boldsymbol{\mu}}_{G_k}^{(init)} \right), \text{ when } \operatorname{rank}(\mathbf{P}_G \widetilde{\mathbf{X}}_G) = |G|, \\
T_G &= \frac{1}{\widehat{\sigma}} \left\| \mathbf{P}_G \left(\boldsymbol{y} - \sum_{G_k \not\subseteq G} \mathbf{Q}_{G_k \setminus G} \widehat{\boldsymbol{\mu}}_{G_k}^{(init)} \right) \right\|_2 .
\end{aligned} \tag{4.2.23}$$

With $\{\mathbf{X}_G, \mathbf{Q}_G\}$ replaced by $\{\widetilde{\mathbf{X}}_G, \widetilde{\mathbf{Q}}_G\}$, our analysis yields the following theorem.

Theorem 4.2. *Let $\boldsymbol{P}_G$, $\widehat{\boldsymbol{\beta}}_G$ and T_G be given by (4.2.23). Suppose that (4.2.2) holds and*

$$\frac{|G|}{n} \to 0, \quad \frac{s + g \log M}{n^{1/2}} \left(\frac{|G|^{1/2}}{n^{1/2}} + \max_{G_k \setminus G \neq \emptyset} \frac{\omega_k'}{\omega_k} \right) \to 0. \tag{4.2.24}$$

Then, (4.2.18) and (4.2.22) hold with $\{\boldsymbol{X}_G, \boldsymbol{Q}_G\}$ replaced by $\{\widetilde{\boldsymbol{X}}_G, \widetilde{\boldsymbol{Q}}_G\}$.

The optimization problems in (4.2.20) and (4.2.23) are still somewhat abstract for the moment, although our theorems only require feasible solutions. In the following we prove the feasibility of $\mathbf{P}_G$ in (4.2.20) for Gaussian designs and describe penalized regression methods to find feasible solutions of (4.2.20) and (4.2.23).

4.2.4 Feasibility of relaxed orthogonal projection for random designs

Let $\boldsymbol{e}_i$ be the i-th canonical unit vector of $\mathbb{R}^n$. Throughout this subsection, we assume that the matrix $\mathbf{X}$ has iid subGaussian rows $\boldsymbol{e}_i^T \mathbf{X}$ satisfying $\mathbb{E}\mathbf{X} = \mathbf{0}$, $\mathbb{E}(\mathbf{X}^T\mathbf{X}/n) = \boldsymbol{\Sigma}$

with a positive-definite $\mathbf{\Sigma}$, and that for certain constant $v_0 > 1$

$$\sup_{\boldsymbol{b}\neq\mathbf{0}} \mathbb{E}\exp\left(\frac{(\boldsymbol{e}_i^T\mathbf{X}\boldsymbol{b})^2}{v_0\boldsymbol{b}^T\mathbf{\Sigma}\boldsymbol{b}} + \frac{1}{v_0}\right) \leq 2. \tag{4.2.25}$$

where Let $\mathbf{\Gamma}_{-G,G} = \mathbf{\Sigma}_{-G,-G}^{-1}\mathbf{\Sigma}_{-G,G}$. We write the regression model (4.2.8) as

$$\mathbf{X}_G = \mathbf{X}_{-G}\mathbf{\Gamma}_{-G,G} + \mathbf{Z}_G^o = \sum_{k=1}^{M}\mathbf{X}_{G_k\setminus G}\mathbf{\Gamma}_{G_k\setminus G,G} + \mathbf{Z}_G^o. \tag{4.2.26}$$

Let $\mathbf{P}_G^o$ be the orthogonal projection to the column space of $\mathbf{Z}_G^o$,

$$\mathbf{P}_G^o = \mathbf{Z}_G^o\Big((\mathbf{Z}_G^o)^T\mathbf{Z}_G^o\Big)^\dagger(\mathbf{Z}_G^o)^T. \tag{4.2.27}$$

We use the following lemma to evaluate $\mathbf{P}_G^o$. The inequality is well known; See for example Vershynin [2012], and for Gaussian $\mathbf{X}$ the supplementary material for Ma [2013].

Lemma 4.1. *Let $\boldsymbol{B}_k$ be matrices of p rows and rank r_k. Let $\boldsymbol{P}_k$ be the projection to the range of $\boldsymbol{X}\boldsymbol{B}_k$ and $\mathbf{\Omega}_{1,2} = ((\boldsymbol{B}_1^T\mathbf{\Sigma}\boldsymbol{B}_1)^\dagger)^{1/2}\boldsymbol{B}_1^T\mathbf{\Sigma}\boldsymbol{B}_2((\boldsymbol{B}_2^T\mathbf{\Sigma}\boldsymbol{B}_2)^\dagger)^{1/2}$. Let $r = rank(\mathbf{\Omega}_{1,2})$ and $1 \geq \lambda_1 \geq \cdots \geq \lambda_r > 0$ be the nonzero singular values of $\mathbf{\Omega}_{1,2}$. Define $\lambda_{\min} = \lambda_r I\{r = r_1 = r_2\}$. Then, there exists a numerical constant $C_0 > 1$ such that when $C_0 v_0\sqrt{t/n + (r_1+r_2)/n} < \epsilon_0 < 1$,*

$$\mathbb{P}\left\{\|((\boldsymbol{B}_1^T\mathbf{\Sigma}\boldsymbol{B}_1)^\dagger)^{1/2}\boldsymbol{B}_1^T(\boldsymbol{X}^T\boldsymbol{X}/n)\boldsymbol{B}_2((\boldsymbol{B}_2^T\mathbf{\Sigma}\boldsymbol{B}_2)^\dagger)^{1/2} - \mathbf{\Omega}_{1,2}\|_S \leq \epsilon_0\right\} \geq 1 - e^{-t}. \tag{4.2.28}$$

and

$$\mathbb{P}\left\{\|\boldsymbol{P}_1\boldsymbol{P}_2\|_S \leq \frac{\lambda_1(1+\epsilon_0)}{1-\epsilon_0}, \|\boldsymbol{P}_1\boldsymbol{P}_2^\perp\|_S^2 \leq 1 - \left(\frac{\lambda_{\min}(1-\epsilon_0)}{1+\epsilon_0}\right)^2\right\} \geq 1 - e^{-t}. \tag{4.2.29}$$

Moreover, $\lambda_1 < 1$ iff $rank(\boldsymbol{B}_1, \boldsymbol{B}_2) = r_1 + r_2$ and $\lambda_{\min} > 0$ iff $rank(\boldsymbol{B}_1^T\boldsymbol{B}_2) = r_1 = r_2$.

Proof of Lemma 4.1. Let $\boldsymbol{u}_j, 1 \le j \le r_k$, be the eigenvectors of $\mathbf{B}_k^T \boldsymbol{\Sigma} \mathbf{B}_k$ corresponding to positive eigenvalues and $\mathbf{U}_k = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{r_k})$. Let $\mathbf{Z}_k = \mathbf{X}\mathbf{B}_k((\mathbf{B}_k^T \boldsymbol{\Sigma} \mathbf{B}_k)^\dagger)^{1/2}\mathbf{U}_k \in \mathbb{R}^{n \times r_k}$. We have $\mathbb{E}\mathbf{Z}_k = \mathbf{0}$, $\mathbb{E}(\mathbf{Z}_k^T \mathbf{Z}_k/n) = \mathbf{I}_{r_k \times r_k}$, $\mathbb{E}(\mathbf{Z}_1^T \mathbf{Z}_2/n) = \mathbf{U}_1^T \boldsymbol{\Omega}_{1,2} \mathbf{U}_2$, and

$$\sup_{\|\boldsymbol{b}\|_2 \le 1} \mathbb{E} \exp\left(\frac{(\boldsymbol{e}_i^T \mathbf{Z}_k \boldsymbol{b})^2}{v_0} + \frac{1}{v_0} \right) \le 2, \ k = 1, 2.$$

Moreover, $\mathbf{P}_k = \mathbf{Z}_k(\mathbf{Z}_k^T \mathbf{Z}_k)^\dagger \mathbf{Z}_k^T$ and $\|\mathbf{U}_1^T \boldsymbol{\Omega}_{1,2} \mathbf{U}_2\|_S = \|\boldsymbol{\Omega}_{1,2}\|_S \le 1$.

For $1 \le j \le k \le 2$ and any vectors $\boldsymbol{v}_k \in \mathbb{R}^{r_k}$ with $\|\boldsymbol{v}_k\|_2 = 1$,

$$\boldsymbol{v}_j^T \Big(\mathbf{Z}_j^T \mathbf{Z}_k/n - \mathbb{E}\mathbf{Z}_j^T \mathbf{Z}_k/n \Big) \boldsymbol{v}_k = \frac{1}{n} \sum_{i=1}^n \Big\{ (\boldsymbol{e}_i^T \mathbf{Z}_j \boldsymbol{v}_j)(\boldsymbol{e}_i^T \mathbf{Z}_k \boldsymbol{v}_k) - \boldsymbol{v}_j^T \mathbb{E}(\mathbf{Z}_j^T \mathbf{Z}_k/n) \boldsymbol{v}_k \Big\}$$

is an average of iid variables with

$$\begin{aligned} & \mathbb{E} \exp\left(\frac{(\boldsymbol{e}_i^T \mathbf{Z}_j \boldsymbol{v}_j)(\boldsymbol{e}_i^T \mathbf{Z}_k \boldsymbol{v}_k) - \boldsymbol{v}_j^T \mathbb{E}(\mathbf{Z}_j^T \mathbf{Z}_k/n) \boldsymbol{v}_k}{v_0} \right) \\ \le & \left\{ \prod_{k=1}^2 \sqrt{\mathbb{E} \exp\left((\boldsymbol{e}_i^T \mathbf{Z}_k \boldsymbol{v}_k)^2/v_0 \right)} \right\} e^{1/v_0} \\ \le & \ 2. \end{aligned}$$

Since the size of an ϵ-net of the unit ball in $\mathbb{R}^{r_k}$ is bounded by $(1+2/\epsilon)^{r_k}$, the Bernstein inequality implies that for $r^* = r_1 + r_2$ and a certain numerical constant C_0,

$$\mathbb{P}\Big\{ \|\mathbf{Z}_j^T \mathbf{Z}_k/n - \mathbb{E}(\mathbf{Z}_j^T \mathbf{Z}_k/n)\|_S > C_0 v_0 \max\left(\sqrt{t/n + r^*/n}, t/n + r^*/n \right) \Big\} \le e^{-t}/3.$$

This yields (4.2.28) as $\|\mathbf{U}_1^T \boldsymbol{\Delta} \mathbf{U}_2\|_S = \|\boldsymbol{\Delta}\|_S$ for all $\boldsymbol{\Delta}$ of proper dimension.

Suppose $\mathrm{rank}(\mathbf{P}_k) = r_k$. Let $r_0 = \mathrm{rank}(\mathbf{P}_1 \mathbf{P}_2)$ and $1 \ge \widehat{\lambda}_1 \ge \cdots \ge \widehat{\lambda}_{r_0} > 0$ be the (nonzero) singular values of $\mathbf{P}_1 \mathbf{P}_2$. We have $\|\mathbf{P}_1 \mathbf{P}_2\|_S = \widehat{\lambda}_1$ and $\|\mathbf{P}_1 \mathbf{P}_2^\perp\|_S = \|\mathbf{P}_1 - \mathbf{P}_2\|_S = \sqrt{1 - \widehat{\lambda}_{\min}^2}$ with $\widehat{\lambda}_{\min} = \widehat{\lambda}_{r_0} I\{r_0 = r_1 = r_2\}$. By definition,

$$\mathbf{P}_1 \mathbf{P}_2 = \mathbf{Z}_1 (\mathbf{Z}_1^T \mathbf{Z}_1)^{-1} \mathbf{Z}_1^T \mathbf{Z}_2 (\mathbf{Z}_2^T \mathbf{Z}_2)^{-1} \mathbf{Z}_2^T.$$

Since $(\mathbf{Z}_k^T\mathbf{Z}_k)^{-1/2}\mathbf{Z}_k^T$ are unitary maps from the range of $\mathbf{P}_k$ to $\mathbb{R}^{r_k}$, the singular values of $\mathbf{P}_1\mathbf{P}_2$ is the same as those of

$$(\mathbf{Z}_1^T\mathbf{Z}_1)^{-1/2}\mathbf{Z}_1^T\mathbf{Z}_2(\mathbf{Z}_2^T\mathbf{Z}_2)^{-1/2}.$$

Now suppose that $\|\mathbf{Z}_j^T\mathbf{Z}_k/n - \mathbb{E}(\mathbf{Z}_j^T\mathbf{Z}_k/n)\|_S \le C_0v_0\sqrt{t/n + r/n} \le \epsilon_0 < 1$ for $1 \le j \le k \le 2$. Recall that $1 \ge \lambda_1 \ge \cdots \ge \lambda_r > 0$ are the nonzero singular values of $\mathbf{\Omega}_{1,2}$ and $\lambda_{\min} = \lambda_r I\{r = r_1 = r_2\}$. As $\mathbb{E}(\mathbf{Z}_k^T\mathbf{Z}_k/n) = \mathbf{I}_{r_k\times r_k}$, we have $\mathrm{rank}(\mathbf{P}_k) = r_k$. Moreover, as $\mathbb{E}(\mathbf{Z}_1^T\mathbf{Z}_2/n) = \mathbf{U}_1^T\mathbf{\Omega}_{1,2}\mathbf{U}_2$ with unitary maps $\mathbf{U}_1$ and $\mathbf{U}_2$, the Weyl inequality implies that

$$\widehat{\lambda}_1 \le \frac{\lambda_1(1+\epsilon_0)}{1-\epsilon_0}, \quad \widehat{\lambda}_{\min} \ge \frac{\lambda_{\min}(1-\epsilon_0)}{1+\epsilon_0}.$$

Thus, (4.2.29) holds. As the conditions for $\lambda_1 < 1$ and $\lambda_{\min} > 0$ follow from the positive-definiteness of $\mathbf{\Sigma}$, the proof is complete. ■

As we have discussed below (4.2.14), when $\mathbf{P}_G^o$ is used, (4.2.12) has the interpretation as

$$\widehat{\boldsymbol{\beta}}_G \;=\; (\mathbf{P}_G^o\mathbf{X}_G)^\dagger\mathbf{P}_G^o\left(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}_{-G}^{(init)}\right) = \left((\mathbf{Z}_G^o)^T\mathbf{X}_G\right)^\dagger(\mathbf{Z}_G^o)^T\left(\boldsymbol{y} - \mathbf{X}_{-G}\widehat{\boldsymbol{\beta}}_{-G}^{(init)}\right).$$

Theorem 4.3. *Suppose the subGaussian condition (4.2.25) holds with* $0 < c_* \le eigen(\mathbf{\Sigma}) \le c^*$ *and fixed* $\{v_0, c_*, c^*\}$.
Let $\omega_k' = \xi n^{-1/2}\left(\sqrt{|G| + |G_k \setminus G|} + \sqrt{\log(M/\delta)}\right)$, $\lambda_{\min}$ *be the smallest eigenvalue of* $\{\mathbf{\Sigma}_{G,G}^{-1/2}(\mathbf{\Sigma}^{-1})_{G,G}\mathbf{\Sigma}_{G,G}^{-1/2}\}^{1/2}$, $\xi n^{-1/2}\left(\sqrt{|G|} + \sqrt{\log(M/\delta)}\right) \le \eta_n$, *and* $a_n = \lambda_{\min}(1-\eta_n)/(1+\eta_n)$. *Then, there exist numerical constants* $\epsilon_0 \in (0,1)$ *and* $\xi_0 < \infty$ *such that*

when $\xi \geq \xi_0 v_0$ *and* $\eta_n \leq \epsilon_0$,

$$\mathbb{P}\left\{ \begin{array}{l} \textit{(4.2.20) has a feasible solution } \boldsymbol{P}_G \textit{ with} \\ \textit{rank}(\boldsymbol{P}_G \boldsymbol{X}_G) = |G| \textit{ and } \|\boldsymbol{P}_G \boldsymbol{Q}_G^{\perp}\|_S \leq \sqrt{1-a_n^2} \end{array} \right\} \geq 1-\delta. \qquad (4.2.30)$$

Let $\widehat{\boldsymbol{\beta}}_G$ *and* T_G *be as in (4.2.12) and (4.2.13). Suppose that (4.2.2) holds for* $\widehat{\boldsymbol{\beta}}^{(init)}$ *and* $\widehat{\sigma}$, *and*

$$\frac{|G|}{n} \to 0, \ \max_{G_k \setminus G \neq \emptyset} \frac{|G_k|}{n} \to 0, \ \frac{s + g \log M}{n^{1/2}} \left(\frac{|G|^{1/2}}{n^{1/2}} + \max_{G_k \setminus G \neq \emptyset} \frac{\omega_k'}{\omega_k} \right) \to 0. \qquad (4.2.31)$$

Then, (4.2.18) and (4.2.22) hold with $\|Rem_G\|_2 = o_{\mathbb{P}}(1)$.

Proof of Theorem 4.3. By (4.2.27), $\mathbf{P}_G^o$ is the orthogonal projection to the range of $\mathbf{Z}_G^o = \mathbf{X}\mathbf{B}_G^o$ with $\mathbf{B}_G^o = (\mathbf{\Sigma}^{-1})_{*,G}(\mathbf{\Sigma}^{-1})_{G,G}^{-1}$. By definition, $\mathbf{Q}_{G_k \setminus G}$ is the projection to the range of $\mathbf{X}_{G_k \setminus G} = \mathbf{X}\mathbf{B}_{G_k \setminus G}$ and $\mathbf{Q}_G$ to the range of $\mathbf{X}_G = \mathbf{X}\mathbf{B}_G$, where $\mathbf{B}_{G_k \setminus G}$ and $\mathbf{B}_G$ are 0-1 diagonal matrices projecting to the indicated spaces. Define $\mathbf{\Omega} = \mathbf{\Sigma}_{G,G}^{-1/2}\{(\mathbf{\Sigma}^{-1})_{G,G}\}^{1/2}$. We have $\mathbf{B}_{G_k \setminus G}^T \mathbf{\Sigma} \mathbf{B}_G^o = \mathbf{\Sigma}_{G_k \setminus G,*}\mathbf{B}_G^o = 0$, $\mathbf{B}_G^T \mathbf{\Sigma} \mathbf{B}_G^o = \mathbf{\Sigma}_{G,*}\mathbf{B}_G^o = (\mathbf{\Sigma}^{-1})_{G,G} = (\mathbf{B}_G^o)^T \mathbf{\Sigma} \mathbf{B}_G^o$ and

$$(\mathbf{B}_G^T \mathbf{\Sigma} \mathbf{B}_G)^{-1/2} \mathbf{B}_G^T \mathbf{\Sigma} \mathbf{B}_G^o \big((\mathbf{B}_G^o)^T \mathbf{\Sigma} \mathbf{B}_G^o \big)^{-1/2} = \mathbf{\Sigma}_{G,G}^{-1/2}\{(\mathbf{\Sigma}^{-1})_{G,G}\}^{1/2} = \mathbf{\Omega} \in \mathbb{R}^{|G| \times |G|}.$$

Moreover, $\mathbf{\Omega} = \mathbf{\Sigma}_{G,G}^{-1/2}\{(\mathbf{\Sigma}^{-1})_{G,G}\}^{1/2}$ is a $|G| \times |G|$ matrix of rank $|G|$ and the smallest singular value of $\mathbf{\Omega}$ is $\lambda_{\min}$. Thus, by (4.2.29) of Lemma 4.1 and the definition of ω_k' and a_n,

$$\mathbb{P}\Big\{ \|\boldsymbol{P}_G \mathbf{Q}_{G_k \setminus G}\|_S \leq \omega_k' \ \forall k \leq M, \ \|\boldsymbol{P}_G \mathbf{Q}_G^{\perp}\|_S \leq \sqrt{1-a_n^2} \Big\} \geq 1-\delta.$$

This yields (4.2.30). It remains to proof $\max_{G_k \setminus G \neq \emptyset} M_k = O_{\mathbb{P}}(1)$ in view of Theorem 4.1. To this end, we notice that due to the condition $|G_k| + g \log M \ll n$, (4.2.28) of

Lemma 4.1 with $\mathbf{B}_1 = \mathbf{B}_2$ implies $\|\mathbf{X}_A^T\mathbf{X}_A/n - \mathbf{\Sigma}_{A,A}\|_S = o_{\mathbb{P}}(1)$ for both $A = G_k$ and $A = G_k \setminus G$ and all k with $G_k \setminus G \neq \emptyset$, so that $\max_{G_k \setminus G \neq \emptyset} M_k = o_{\mathbb{P}}(1) + O(1)$. ■

4.2.5 Finding feasible solutions

While (4.2.30) of Theorems 4.3 guarantees a feasible solution of (4.2.20), we discuss here penalized multivariate regression methods for finding feasible solutions of (4.2.20) and (4.2.23). As the only difference between (4.2.20) and (4.2.23) is the respective use of $\mathbf{X}_G$ and $\widetilde{\mathbf{X}}_G$. We provide formulas here only for (4.2.20), with the understanding that formulas for (4.2.23) can be generated in the same way with $\mathbf{X}_G$ replaced by $\widetilde{\mathbf{X}}_G$.

In view (4.2.26), a general formulation of the penalized multivariate regression is

$$\widehat{\mathbf{\Gamma}}_{-G,G} = \underset{\mathbf{\Gamma}_{-G,G}}{\arg\min} \left\{ \frac{1}{2n} \left\| \mathbf{X}_G - \sum_{G_k \not\subseteq G} \mathbf{X}_{G_k \setminus G} \mathbf{\Gamma}_{G_k \setminus G, G} \right\|_F^2 + R(\mathbf{\Gamma}_{-G,G}) \right\}, \quad (4.2.32)$$

where $\|\cdot\|_F$ is the Frobenius norm and $R(\mathbf{\Gamma}_{-G,G})$ is a penalty function. Define

$$\mathbf{Z}_G = \mathbf{X}_G - \sum_{G_k \not\subseteq G} \mathbf{X}_{G_k \setminus G} \widehat{\mathbf{\Gamma}}_{G_k \setminus G, G}, \quad \mathbf{P}_G = \mathbf{Z}_G (\mathbf{Z}_G^T \mathbf{Z}_G)^{-1} \mathbf{Z}_G^T. \quad (4.2.33)$$

Our main interest is to find a feasible solutions of (4.2.20) and (4.2.23), not to estimate $\mathbf{\Gamma}_{-G,G}$.

The following weighted group nuclear penalty matches the dual of the constraint in (4.2.20) and (4.2.23):

$$R(\mathbf{\Gamma}_{-G,G}) = \sum_{G_k \not\subseteq G} \frac{\xi \omega_k''}{n^{1/2}} \left\| \mathbf{X}_{G_k \setminus G} \mathbf{\Gamma}_{G_k \setminus G, G} \right\|_N. \quad (4.2.34)$$

It follows from the KKT conditions for (4.2.32) and (4.2.34) that

$$\left\| \mathbf{Q}_{G_k \setminus G} \mathbf{Z}_G / \sqrt{n} \right\|_S \leq \xi \omega_k''. \quad (4.2.35)$$

If we set $\omega_k'' = \omega_k$ in (4.2.34), then conditions (4.2.21) and (4.2.24) become

$$\xi \|(\mathbf{Z}_G^T \mathbf{Z}_G/n)^{-1/2}\|_S \frac{s + g \log M}{n^{1/2}} \to 0, \tag{4.2.36}$$

provided $\max_{G_k \not\subseteq G} M_k = O(1)$ in the case of Theorem 4.1. When the group sizes are not too large, one may even consider to replace the weighted group nuclear penalty with a weighted group Frobenius penalty

$$R(\mathbf{\Gamma}_{-G,G}) = \sum_{G_k \not\subseteq G} \frac{\xi \omega_k''}{n^{1/2}} \left\| \mathbf{X}_{G_k \setminus G} \mathbf{\Gamma}_{G_k \setminus G, G} \right\|_F$$

as this can be conveniently computed using the group Lasso software.

Remark 4.2. Compared with existing sample size condition $n^{1/2} \gg \|\boldsymbol{\beta}\|_0 \log p$ for statistical inference of a univariate parameter at $n^{-1/2}$ rate, the sample size conditions in (4.2.21), (4.2.24) (4.2.31) and (4.2.36) clearly demonstrate the benefit of group sparsity as in Huang and Zhang [2010]. Moreover, the extra factor $\sqrt{|G|}$ is removed in a number of scenarios even in case of large group sizes. For example $|G| \lesssim \min_{G_k \not\subseteq G}\{|G_k| + \log(M/\delta)\}$ in (4.2.24) and (4.2.31), or $\xi \|(\mathbf{Z}_G^T \mathbf{Z}_G/n)^{-1/2}\|_S \ll |G|^{1/2}$ in (4.2.36).

4.3 Mixed Norm Consistency Results

Using the group sparsity of the regression coefficient vector and sparse eigenvalue conditions on the design matrix, Huang and Zhang [2010] provided ℓ_2 oracle inequalities to show the benefits of the group Lasso over the Lasso. In this section we provide similar results on mixed weighted norms for both the group Lasso and the scaled group Lasso under different conditions on the design.

4.3.1 Assumptions for fixed design matrix

In the Lasso problem, performance bounds of the estimator are derived based on various conditions on the design matrix, for example, restricted isometry property Candes and Tao [2005], the compatibility condition van de Geer [2007], the sparse Riesz condition Zhang and Huang [2008], the restricted eigenvalue condition Bickel et al. [2009], Koltchinskii [2009], and cone invertibility conditions Ye and Zhang [2010]. van de Geer and Bühlmann [2009] showed that the compatibility condition is weaker than the restricted eigenvalues condition for the prediction and ℓ_1 loss, while Ye and Zhang [2010] showed that both conditions can be weakened by cone invertibility conditions. In the following, we define grouped versions of such conditions.

Let us first define a group wise mixed norm cone for $T \subset \{1 \cdots, M\}$ and $\xi \geq 0$ as

$$\mathscr{C}^{(G)}(\xi, \boldsymbol{\omega}, T) = \left\{ \boldsymbol{u} : \sum_{j \in T^c} \omega_j \|\boldsymbol{u}_{G_j}\|_2 \leq \xi \sum_{j \in T} \omega_j \|\boldsymbol{u}_{G_j}\|_2 \neq 0 \right\}. \tag{4.3.1}$$

Following Nardi and Rinaldo [2008] and Lounici et al. [2011], the restricted eigenvalue (RE) is defined as

$$\mathrm{RE}^{(G)}(\xi, \boldsymbol{\omega}, T) = \inf_{\boldsymbol{u}} \left\{ \frac{\|\mathbf{X}\boldsymbol{u}\|_2}{\sqrt{n}\|\boldsymbol{u}_{G_T}\|_2} : \boldsymbol{u} \in \mathscr{C}^{(G)}(\xi, \boldsymbol{\omega}, T) \right\}. \tag{4.3.2}$$

For the weighted $\ell_{2,1}$ norm, the group-wise compatibility constant (CC) can be defined as

$$\mathrm{CC}^{(G)}(\xi, \boldsymbol{\omega}, T) = \inf_{\boldsymbol{u}} \left\{ \frac{\|\mathbf{X}\boldsymbol{u}\|_2 \sqrt{\sum_{j \in T} \omega_j^2}}{\sqrt{n} \sum_{j \in T} \omega_j \|\boldsymbol{u}_{G_j}\|_2} : \boldsymbol{u} \in \mathscr{C}^{(G)}(\xi, \boldsymbol{\omega}, T) \right\}. \tag{4.3.3}$$

We also introduce the notion of group wise cone invertibility factor and extend it to sign-restricted cone invertibility factor. The cone invertibility factor (CIF) is defined

as

$$\mathrm{CIF}_1^{(G)}(\xi, \boldsymbol{\omega}, T) = \inf_{\boldsymbol{u} \in \mathscr{C}^{(G)}(\xi, \boldsymbol{\omega}, T)} \frac{\max_j \left[\omega_j^{-1} \|\mathbf{X}_{G_j}^T \mathbf{X}\boldsymbol{u}\|_2\right] \sum_{j \in T} \omega_j^2}{n \sum_{j \in T} \omega_j \|\boldsymbol{u}_{G_j}\|_2}. \tag{4.3.4}$$

Now we define the sign-restricted cone as

$$\mathscr{C}_-^{(G)}(\xi, \boldsymbol{\omega}, T) = \left\{ \boldsymbol{u} : \boldsymbol{u} \in \mathscr{C}^{(G)}(\xi, \boldsymbol{\omega}, T), \quad \boldsymbol{u}_{G_j}^T \mathbf{X}_{G_j}^T \mathbf{X}\boldsymbol{u} \leq 0 \ \forall j \in T^c \right\}, \tag{4.3.5}$$

and the group-wise sign-restricted cone invertibility factor (SCIF) as

$$\mathrm{SCIF}_1^{(G)}(\xi, \boldsymbol{\omega}, T) = \inf_{\boldsymbol{u} \in \mathscr{C}_-^{(G)}(\xi, \boldsymbol{\omega}, T)} \frac{\max_j \left[\omega_j^{-1} \|\mathbf{X}_{G_j}^T \mathbf{X}\boldsymbol{u}\|_2\right] \sum_{j \in T} \omega_j^2}{n \sum_{j \in T} \omega_j \|\boldsymbol{u}_{G_j}\|_2}. \tag{4.3.6}$$

It follows from $\|\mathbf{X}\boldsymbol{u}\|_2^2 / \max_j(\omega_j^{-1}\|\mathbf{X}_{G_j}^T \mathbf{X}\boldsymbol{u}\|_2) \leq \sum_j \omega_j \|\boldsymbol{u}_{G_j}\|_2 \leq (1+\xi) \sum_{j \in T} \omega_j \|\boldsymbol{u}_{G_j}\|_2$ and the Cauchy-Schwarz inequality that

$$(\mathrm{RE}^{(G)}(\xi, \boldsymbol{\omega}, T))^2 \leq (\mathrm{CC}^{(G)}(\xi, \boldsymbol{\omega}, T))^2 \leq (\xi + 1)\ \mathrm{CIF}_1^{(G)}(\xi, \boldsymbol{\omega}, T). \tag{4.3.7}$$

Moreover, the SCIF is always no smaller than the CIF. Thus, following (4.3.7), the restricted eigenvalue condition $\mathrm{RE}^{(G)}(\xi, \boldsymbol{\omega}, T) > \kappa_0$ implies that all the other quantities are bounded from below by κ_0. In the following we derive the mixed norm consistency results for the non-scaled group Lasso problem in Theorem 4.4 and extend it to the scaled group Lasso in Theorem 4.5. We establish these results under the weakest assumption on the SCIF.

The SCIF in (4.3.6) will be used to derive oracle inequalities for the prediction and weighted $\ell_{2,1}$ loss. For the ℓ_2 loss, we define the SCIF as

$$\mathrm{SCIF}_2^{(G)}(\xi, \boldsymbol{\omega}, T) = \inf_{\boldsymbol{u} \in \mathscr{C}_-^{(G)}(\xi, \boldsymbol{\omega}, T)} \frac{\max_j \left[\omega_j^{-1} \|\mathbf{X}_{G_j}^T \mathbf{X}\boldsymbol{u}\|_2\right] (\sum_{j \in T} \omega_j^2)^{1/2}}{n\|\boldsymbol{u}\|_2/(1+\xi)}. \tag{4.3.8}$$

We may also use the ℓ_2 version of the CIF, denoted by $\mathrm{CIF}_2^{(G)}(\xi, \boldsymbol{\omega}, T)$ and defined by replacing the sign-restricted cone $\mathscr{C}_-^{(G)}(\xi, \boldsymbol{\omega}, T)$ with the cone in (4.3.1). It follows from a shifting inequality Cai et al. [2010a], Ye and Zhang [2010] that

$$\omega_{\min}\sqrt{s}\|\boldsymbol{u}\|_2 \sum_{j\in S^*} \omega_j \|\boldsymbol{u}_{G_j}\|_2 \;\le\; 3\Big(\sum_{j\in S^*} \omega_j^2\Big) \max_{|T|\le s} \|\boldsymbol{u}_{G_T}\|_2^2$$

for $\boldsymbol{u} \in \mathscr{C}^{(G)}(3, \boldsymbol{\omega}, S^*)$ and $|S^*| \le s$, where $\omega_{\min} = \min_{1\le j\le M} \omega_j$. Thus,

$$\frac{\big(\sum_{j\in S^*} \omega_j^2\big)^{1/2}}{\mathrm{SCIF}_2^{(G)}(3, \boldsymbol{\omega}, S^*)} \le \frac{\big(\sum_{j\in S^*} \omega_j^2\big)^{1/2}}{\mathrm{CIF}_2^{(G)}(3, \boldsymbol{\omega}, S^*)} \le \frac{3\sum_{j\in S^*} \omega_j^2/(\omega_{\min}\sqrt{s})}{\min_{|T|\le s}(\mathrm{RE}^{(G)}(3, \boldsymbol{\omega}, T))^2}.$$

Again, the cone invertibility factors provide error bounds of sharper form than (4.3.2), in view of Theorem 4.4 below and Theorem 3.1 of Lounici et al. [2011].

4.3.2 Mixed norm consistency for group Lasso

Theorem 4.4. *Let $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{\omega})$ be a solution of (4.1.3) with data $(\boldsymbol{X}, \boldsymbol{y})$ and $\boldsymbol{\beta}^*$ be a vector with $\mathrm{supp}(\boldsymbol{\beta}^*) \subset G_{S^*}$ for some $S^* \subset \{1, \cdots, M\}$. Let $\xi > 1$ and define*

$$\mathcal{E} = \left\{ \max_{1\le j\le M} \frac{\|\boldsymbol{X}_{G_j}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*)\|_2}{\omega_j n} \le \frac{\xi - 1}{\xi + 1} \right\}. \tag{4.3.9}$$

Then in the event $\mathcal{E}$, we have

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n \le \frac{\{2\xi/(\xi+1)\}^2 \sum_{j\in S^*} \omega_j^2}{\mathrm{SCIF}_1^{(G)}(\xi, \boldsymbol{\omega}, S^*)}, \tag{4.3.10}$$

and

$$\left\{ \sum_{j=1}^M \omega_j^2 \left(\frac{\|\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{\beta}^*_{G_j}\|_2}{\omega_j} \right)^q \right\}^{1/q} \le \frac{2\xi\big(\sum_{j\in S^*} \omega_j^2\big)^{1/q}}{\mathrm{SCIF}_q^{(G)}(\xi, \boldsymbol{\omega}, S^*)}, \quad q = 1, 2. \tag{4.3.11}$$

Moreover, if the regression model in (4.1.1) holds with Gaussian error and a design matrix $\boldsymbol{X}$ *satisfying* $\max_{j\le M}\|\boldsymbol{X}_{G_j}/\sqrt{n}\|_S \le 1$, *then*

$$\mathbb{P}(\mathcal{E}) > 1-\delta, \tag{4.3.12}$$

when $\omega_j \ge A\sigma\left\{\sqrt{d_j/n}+\sqrt{(2/n)\log(M/\delta)}\right\}$ *for some* $0<\delta<1$ *and* $A \ge (\xi+1)/(\xi-1)$.

Remark 4.3. From Theorem 4.4, $\max\{\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2^2, \sum_{j=1}^M \omega_j\|\widehat{\boldsymbol{\beta}}_{G_j}-\boldsymbol{\beta}^*_{G_j}\|_2\} = \mathcal{O}((s+g\log M)/n)$ when the SCIF can be treated as constant. This shows the benefit of the group Lasso compared with the Lasso as in Huang and Zhang [2010]. The same convergence rate can be derived from the ℓ_2 consistency result in Huang and Zhang [2010]. Their result however, is derived under a sparse eigenvalue condition on the design matrix $\mathbf{X}$.

Proof of Theorem 4.4. The KKT conditions for the group Lasso asserts that

$$\begin{aligned} &\frac{1}{n}\mathbf{X}_{G_j}^T(\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) = \omega_j\widehat{\boldsymbol{\beta}}_{G_j}/\|\widehat{\boldsymbol{\beta}}_{G_j}\|_2, \quad \widehat{\boldsymbol{\beta}}_{G_j}\neq\mathbf{0},\\ &\frac{1}{n}\|\mathbf{X}_{G_j}^T(\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})\|_2 \le \omega_j, \qquad\qquad\quad \widehat{\boldsymbol{\beta}}_{G_j}=\mathbf{0}. \end{aligned} \tag{4.3.13}$$

It follows that in the event $\mathcal{E}$

$$\omega_j^{-1}\|\mathbf{X}_{G_j}^T(\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}^*)\|_2/n \le 1+\|\mathbf{X}_{G_j}^T\boldsymbol{\varepsilon}\|_2/(n\omega_j) \le 2\xi/(\xi+1). \tag{4.3.14}$$

Now take any $\boldsymbol{w}\in\mathbb{R}^p$. Pre-multiplying by $(\widehat{\boldsymbol{\beta}}_{G_j}-\boldsymbol{w}_{G_j})^T$ on both sides in (4.3.13), we have

$$(\widehat{\boldsymbol{\beta}}-\boldsymbol{w})^T\mathbf{X}^T(\boldsymbol{\varepsilon}-\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*))/n \ge \sum_{j=1}^M \omega_j\|\widehat{\boldsymbol{\beta}}_{G_j}\|_2 - \sum_{j=1}^M \omega_j\|\boldsymbol{w}_{G_j}\|_2.$$

Rearranging we get,

$$\begin{aligned}
&\frac{1}{n}(\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*)^T(\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{w}) + \sum_{j\notin S^*}\omega_j\|\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{w}_{G_j}\|_2 \\
\leq &\sum_{j\notin S^*}\omega_j\|\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{w}_{G_j}\|_2 - \sum_{j=1}^{M}\omega_j\|\widehat{\boldsymbol{\beta}}_{G_j}\|_2 + \sum_{j=1}^{M}\omega_j\|\boldsymbol{w}_{G_j}\|_2 + (\widehat{\boldsymbol{\beta}} - \boldsymbol{w})^T\mathbf{X}^T\boldsymbol{\varepsilon}/n \\
\leq &\sum_{j\in S^*}\omega_j\|\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{w}_{G_j}\|_2 + 2\sum_{j\notin S^*}\omega_j\|\boldsymbol{w}_{G_j}\|_2 + \sum_{j=1}^{M}\|\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{w}_{G_j}\|_2\|\mathbf{X}_{G_j}^T\boldsymbol{\varepsilon}\|_2/n \\
\leq &\sum_{j\in S^*}\omega_j\|\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{w}_{G_j}\|_2 + 2\sum_{j\notin S^*}\omega_j\|\boldsymbol{w}_{G_j}\|_2 + \frac{\xi-1}{\xi+1}\sum_{j=1}^{M}\omega_j\|\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{w}_{G_j}\|_2.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\frac{1}{n}(\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*)^T(\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{w}) + \frac{2}{\xi+1}\sum_{j\notin S^*}\omega_j\|\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{w}_{G_j}\|_2 \\
\leq &\frac{2\xi}{\xi+1}\sum_{j\in S^*}\omega_j\|\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{w}_{G_j}\|_2 + 2\sum_{j\notin S^*}\omega_j\|\boldsymbol{w}_{G_j}\|_2.
\end{aligned}$$

Putting $\boldsymbol{w} = \boldsymbol{\beta}^*$ and $\boldsymbol{h} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, we have

$$(1+\xi)\|\mathbf{X}\boldsymbol{h}\|_2^2/n + 2\sum_{j\notin S^*}\omega_j\|\boldsymbol{h}_{G_j}\|_2 \leq 2\xi\sum_{j\in S^*}\omega_j\|\boldsymbol{h}_{G_j}\|_2,$$

whence it follows that $\boldsymbol{h} \in \mathscr{C}^{(G)}(\xi, \boldsymbol{\omega}, S^*)$. Moreover, from KKT conditions (4.3.13), pre-multiplying both sides by $\boldsymbol{h}_{G_j}$ for $j \notin S^*$, we have in the event $\mathcal{E}$,

$$\boldsymbol{h}_{G_j}\mathbf{X}_{G_j}^T\mathbf{X}\boldsymbol{h}/n \leq \|\boldsymbol{h}_{G_j}\|_2\left(\|\mathbf{X}_{G_j}^T\boldsymbol{\varepsilon}\|_2/n - \omega_j\right) \leq 0.$$

Hence $\boldsymbol{h} \in \mathscr{C}_{-}^{(G)}(\xi, \boldsymbol{\omega}, S^*)$. Consequently, by (4.3.6) and (4.3.14),

$$\frac{(1+\xi)\|\mathbf{X}\boldsymbol{h}\|_2^2/n}{2\xi \sum_{j\in S^*} \omega_j^2} \leq \frac{\sum_{j\in S^*} \omega_j \|\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{\beta}^*_{G_j}\|_2}{\sum_{j\in S^*} \omega_j^2} \leq \frac{\max_j \omega_j^{-1} \|\mathbf{X}_{G_j}^T \mathbf{X}\boldsymbol{h}\|_2}{n\, \mathrm{SCIF}_1^{(G)}(\xi, \boldsymbol{\omega}, S^*)} \leq \frac{2\xi/(\xi+1)}{\mathrm{SCIF}_1^{(G)}(\xi, \boldsymbol{\omega}, S^*)}.$$

The bound for the weighted $\ell_{2,1}$ loss follows as $\sum_{j=1}^M \omega_j \|\boldsymbol{h}_{G_j}\|_2 \leq (1+\xi) \sum_{j\in S^*} \omega_j \|\boldsymbol{h}_{G_j}\|_2$. The proof for the ℓ_2 loss is nearly identical and thus omitted.

Finally, we prove (4.3.12). As $\boldsymbol{\varepsilon} \sim \mathsf{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, it follows from the Gaussian concentration inequality that for any $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\|\mathbf{X}_{G_j}^T \boldsymbol{\varepsilon}\|_2 / (\sigma \|\mathbf{X}_{G_j}\|_S) \leq \|\boldsymbol{\varepsilon}/\sigma\|_2 \leq \sqrt{n} \left\{ \sqrt{d_j} + \sqrt{2\log(1/\delta)} \right\}.$$

The result in (4.3.12) follows by an application of union bound. ■

4.3.3 Scaled Group Lasso

In the optimization problem (4.1.3), scale-invariance considerations have not been taken into account. Usually the individual penalty level ω_j's could be chosen proportional to the scale σ as a remedy. This issue has been discussed and studied, as pertaining to the Lasso problem, in several literature. See Huber [2011], Städler et al. [2010], Antoniadis [2010], Sun and Zhang [2010], Belloni et al. [2011], Sun and Zhang [2012a], Sun and Zhang [2013] and many more. In Bunea et al. [2014], the authors extended the idea of square root Lasso in Belloni et al. [2011] to group Lasso problems. They developed the so-called group square root Lasso (GSRL) procedure that bypasses the need to estimate the unknown scale parameter σ. In our development, we follow the recipe prescribed in Sun and Zhang [2012a] which provides an iterative scheme for estimation of the scale parameter σ and thereby that of ω_j's and thus obviates the need for cross validation.

Following Antoniadis [2010] we define an optimization problem,

$$(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}) = \arg\min_{\boldsymbol{\beta},\sigma} \ \mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}, \sigma), \tag{4.3.15}$$

$$\text{where} \ \ \mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}, \sigma) = \frac{\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2n\sigma} + \frac{(1-a)\sigma}{2} + \sum_{j=1}^{M} \omega_j \|\boldsymbol{\beta}_{G_j}\|_2. \tag{4.3.16}$$

Following Sun and Zhang [2010] we define an iterative algorithm for the estimation of $\{\boldsymbol{\beta}, \sigma\}$,

$$\begin{aligned} \widehat{\sigma}^{(k+1)} &\leftarrow \|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(k)}\|_2 / \sqrt{(1-a)n}, \\ \boldsymbol{\omega}' &\leftarrow \widehat{\sigma}^{(k+1)} \boldsymbol{\omega}, \\ \widehat{\boldsymbol{\beta}}^{(k+1)} &\leftarrow \arg\min_{\boldsymbol{\beta}} \mathcal{L}_{\boldsymbol{\omega}'}(\boldsymbol{\beta}), \end{aligned} \tag{4.3.17}$$

where $\mathcal{L}_{\boldsymbol{\omega}'}(\boldsymbol{\beta})$ was as defined in (4.1.3). Due to the convexity of the joint loss function $\mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}, \sigma)$, the solution of (4.3.15) and the limit of (4.3.17) give the same estimator, which we call scaled group Lasso. The constant $a \geq 0$ provides control over the degrees of freedom adjustments. In practice, for scaled group Lasso in the $p > n$ setting, we take $a = 0$ for all subsequent discussions. It is clear that that with $a = 0$ and $\boldsymbol{\omega}' = \widehat{\sigma}\boldsymbol{\omega}$, one has $\widehat{\sigma}\mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}, \widehat{\sigma}) = \mathcal{L}_{\boldsymbol{\omega}'}(\boldsymbol{\beta}) + \widehat{\sigma}^2/2$. The algorithm in (4.3.17) suggests a profile optimization approach. The following lemma is similar to Proposition 1 in Sun and Zhang [2012a] and characterizes the solution via partial derivative of the profile objective.

Lemma 4.2. *Let $\widehat{\boldsymbol{\beta}}(\boldsymbol{\omega})$ denote a solution of the optimization problem in (4.1.3). Then, $\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega})$ is a minimizer of $\mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}, \sigma)$ in (4.3.16) for given σ, and the profile loss function $\mathcal{L}_{\boldsymbol{\omega}}(\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega}), \sigma)$ is convex and continuously differentiable in σ with*

$$\frac{\partial}{\partial\sigma}\mathcal{L}_{\boldsymbol{\omega}}(\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega}), \sigma) = \frac{1}{2} - \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega})\|_2^2}{2n\sigma^2}. \tag{4.3.18}$$

Moreover, the algorithm in (4.3.17) converges to a minimizer $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma})$ *in (4.3.15) satisfying* $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\widehat{\sigma}\boldsymbol{\omega})$*, and the estimator* $\widehat{\boldsymbol{\beta}}$ *and* $\widehat{\sigma}$ *are scale equivariant in* $\boldsymbol{y}$*.*

Proof of Lemma 4.2. For $\eta \geq 0$ define

$$\mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}, \sigma, \eta) = \frac{\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \sum_{j=1}^{M} \omega_j \|\boldsymbol{\beta}_{G_j}\|_2^{1+\eta} + \frac{\eta\sigma^2}{2}$$

and $\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega}, \eta) = \arg\min_{\boldsymbol{\beta}} \ \mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}, \sigma, \eta)$. As $\mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}, \sigma, \eta)$ is convex in $(\boldsymbol{\beta}, \sigma)$, the profile loss $\mathcal{L}_{\boldsymbol{\omega}}(\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega}, \eta), \sigma, \eta)$ is convex in σ for all $\eta \geq 0$. Note that for $\eta > 0$

$$\begin{aligned} & \frac{\partial}{\partial\sigma}\mathcal{L}_{\boldsymbol{\omega}}(\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega}, \eta), \sigma, \eta) \\ = & \left\{ \frac{\partial}{\partial\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\theta}, \sigma, \eta)\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega}, \eta)} \right\}^T \frac{\partial\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega}, \eta)}{\partial\sigma} + \frac{\partial}{\partial t}\mathcal{L}_{\boldsymbol{\omega}}(\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega}), t, \eta)\Big|_{t=\sigma} \\ = & 1/2 - \|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega}, \eta)\|_2^2/(2n\sigma^2) + \eta\sigma \end{aligned}$$

as all derivatives involved are continuous. Moreover, as $\mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}, \sigma) = \mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}, \sigma, 0)$ is strictly convex in $\mathbf{X}\boldsymbol{\beta}$,

$$\lim_{\eta\to 0+} \frac{\partial}{\partial\sigma}\mathcal{L}_{\boldsymbol{\omega}}(\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega}, \eta), \sigma, \eta) \to 1/2 - \|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega})\|_2^2/(2n\sigma^2).$$

Consequently,

$$\begin{aligned} \mathcal{L}_{\boldsymbol{\omega}}(\widehat{\boldsymbol{\beta}}(\sigma_2\boldsymbol{\omega}), \sigma_2) - \mathcal{L}_{\boldsymbol{\omega}}(\widehat{\boldsymbol{\beta}}(\sigma_1\boldsymbol{\omega}), \sigma_1) &= \lim_{\eta\to 0+} \int_{\sigma_1}^{\sigma_2} \left\{ \frac{\partial}{\partial\sigma}\mathcal{L}_{\boldsymbol{\omega}}(\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega}, \eta), \sigma, \eta) \right\} d\sigma \\ &= \int_{\sigma_1}^{\sigma_2} \left\{ 1/2 - \|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega})\|_2^2/(2n\sigma^2) \right\} d\sigma. \end{aligned}$$

All other claims follow from the joint convexity of $\mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}, \sigma)$ and the strict convexity of the loss function in $\mathbf{X}\boldsymbol{\beta}$. ■

We now present the consistency theorem for scaled group Lasso which extends

Theorem 4.4 by providing convergence results for the estimate of scale. Define

$$\mu(\boldsymbol{\omega},\xi)=\frac{2\xi\sum_{j\in S^*}\omega_j^2}{\mathrm{SCIF}_1^{(G)}(\xi,\boldsymbol{\omega},S^*)},\quad \tau_-=\frac{2\mu(\boldsymbol{\omega},\xi)(\xi-1)}{\xi+1},\quad \tau_+=\frac{\tau_-}{2}+\mu(\boldsymbol{\omega},\xi).$$

Let $m_{d,n}$ be the median of the beta$(d/2, n/2-d/2)$ distribution and define

$$\omega_{*,j}\geq\sqrt{m_{d_j,n}}+\sqrt{\frac{2\log(M/\delta)}{(n\vee 2)-3/2}},\ A_*=\frac{(\xi+1)/(\xi-1)}{\sqrt{1-2\mu(\boldsymbol{\omega}_*,\xi)(\xi+1)/(\xi-1)}},$$

where $\boldsymbol{\omega}_*$ is the vector with elements $\omega_{*,j}$. We will show that $\sqrt{m_{d_j,n}}\leq(d_j/n)^{1/2}+n^{-1/2}$ in the proof of the following theorem.

Theorem 4.5. *Let $\{\widehat{\boldsymbol{\beta}},\widehat{\sigma}\}$ be a solution of the optimization problem (4.3.16) with data $(\boldsymbol{X},\boldsymbol{y})$ and $\boldsymbol{\beta}^*$ be a vector with $supp(\boldsymbol{\beta}^*)\subset G_{S^*}$ for some $S^*\subset\{1,\cdots,M\}$. Let $\xi>1$.*

(i) Suppose $\mathrm{SCIF}_1^{(G)}(\xi,\boldsymbol{\omega},S^*)>0$ in (4.3.6) and $\tau_+<1$. Define the following event

$$\mathcal{E}=\left\{\max_{1\leq j\leq M}\frac{\|\boldsymbol{X}_{G_j}^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^*)\|_2}{\omega_j n\sigma^*/\sqrt{1+\tau_-}}<\frac{\xi-1}{\xi+1}\right\},\tag{4.3.19}$$

where $\sigma^*=\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^*\|_2/\sqrt{n}$ is the oracle noise level. Then in the event $\mathcal{E}$, we have

$$\frac{\sigma^*}{\sqrt{1+\tau_-}}\leq\widehat{\sigma}\leq\frac{\sigma^*}{\sqrt{1-\tau_+}},\tag{4.3.20}$$

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n\leq\frac{(\sigma^*)^2\{2\xi/(\xi+1)\}^2\sum_{j\in S^*}\omega_j^2}{(1-\tau_+)\mathrm{SCIF}_1^{(G)}(\xi,\boldsymbol{\omega},S^*)},\tag{4.3.21}$$

and

$$\left\{\sum_{j=1}^M\omega_j^2\left(\frac{\|\widehat{\boldsymbol{\beta}}_{G_j}-\boldsymbol{\beta}^*_{G_j}\|_2}{\omega_j}\right)^q\right\}^{1/q}\leq\frac{2\sigma^*\xi\big(\sum_{j\in S^*}\omega_j^2\big)^{1/q}}{\sqrt{1-\tau_+}\mathrm{SCIF}_q^{(G)}(\xi,\boldsymbol{\omega},S^*)},\quad q=1,2.\tag{4.3.22}$$

(ii) Suppose the regression model in (4.1.1) holds with Gaussian error and a design matrix satisfying $\max_{j\le M}\|\boldsymbol{X}_{G_j}/\sqrt{n}\|_S \le 1$. *If* $\sqrt{n}\mu(\boldsymbol{\omega},\xi)\to 0$, *then*

$$\sqrt{n}\,(\widehat{\sigma}/\sigma - 1) \xrightarrow{\mathrm{D}} \mathsf{N}(0, 1/2). \tag{4.3.23}$$

Moreover, if $\omega_j = A\omega_{*,j}$ *with* $A \ge A_*$, *then*

$$\mathbb{P}(\mathcal{E}) \ge 1 - \delta. \tag{4.3.24}$$

Corollary 4.1. *Consider the setup of Theorem 4.5 (ii). Assume that the design matrix* $\boldsymbol{X}$ *satisfies the following sign restricted cone invertibility condition:*

$$\mathrm{SCIF}_1^{(G)}(\xi, S^*) > c > 0 \text{ for some fixed } c > 0.$$

Let $0 < \delta < 1$ *be a fixed small constant and take*

$$\omega_j = A\left\{\sqrt{d_j/n} + \sqrt{(2/n)\log(M/\delta)}\right\} \text{ with a constant } A > (\xi+1)/(\xi-1).$$

Then, for a certain fixed constant $C > 0$ *and with probability at least* $1-\delta$

$$\begin{aligned}\max\Bigg\{&\Big|1 - \frac{\widehat{\sigma}}{\sigma^*}\Big|, \frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2}{\sigma^2}, \frac{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2}{\sigma^2}, \sum_{j=1}^M \frac{\|\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{\beta}^*_{G_j}\|_2}{\sigma/\omega_j},\\ &\sum_{j=1}^M \frac{\|\boldsymbol{X}_{G_j}(\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{\beta}^*_{G_j})\|_2}{n^{1/2}\sigma/\omega_j}\Bigg\}\\ &\le C\left\{|G_{S^*}| + |S^*|\log(M/\delta)\right\}/n.\end{aligned} \tag{4.3.25}$$

Corollary 4.1 touches upon the mixed prediction loss $\sum_{j=1}^M \omega_j\|\mathbf{X}_{G_j}\widehat{\boldsymbol{\beta}}_{G_j} - \mathbf{X}_{G_j}\boldsymbol{\beta}^*_{G_j}\|_2$

the first time in this section. The reason for this omission is two fold. Firstly,

$$\left\{\sum_{j=1}^{M}\omega_j^2\left(\frac{\|\mathbf{X}_{G_j}\widehat{\boldsymbol{\beta}}_{G_j}-\mathbf{X}_{G_j}\boldsymbol{\beta}^*_{G_j}\|_2}{n^{1/2}\omega_j}\right)^q\right\}^{1/q}$$
$$\le \max_{j\le M}\left\|\frac{\mathbf{X}_{G_j}}{\sqrt{n}}\right\|_S\left\{\sum_{j=1}^{M}\omega_j^2\left(\frac{\|\widehat{\boldsymbol{\beta}}_{G_j}-\boldsymbol{\beta}^*_{G_j}\|_2}{\omega_j}\right)^q\right\}^{1/q}$$

so that (4.3.11) and (4.3.22) automatically generate the corresponding bounds for the mixed prediction error under the respective conditions. Secondly, upper bounds for the mixed prediction loss can be obtained by reparametrization within the given group structure. The following corollary provides details of such reparametrization in the case of scaled group Lasso.

Corollary 4.2. *Let* $\boldsymbol{X}_{G_j}=\boldsymbol{U}_{G_j}\boldsymbol{\Lambda}_{G_j}\boldsymbol{V}_{G_j}^T$ *be the SVD of* $\boldsymbol{X}_{G_j}$ *with* $\boldsymbol{\Lambda}_{G_j}\in\mathbb{R}^{|G_j|\times|G_j|}$. *Define* $\boldsymbol{b}$ *by* $\boldsymbol{b}_{G_j}=\boldsymbol{\Lambda}_{G_j}\boldsymbol{V}_{G_j}^T\boldsymbol{\beta}_{G_j}$ *and* $\boldsymbol{U}$ *by* $\boldsymbol{U}\boldsymbol{b}=\sum_{j=1}^{M}\boldsymbol{U}_{G_j}\boldsymbol{b}_{G_j}$. *Then,*

$$\begin{aligned}\left\{\sum_{j=1}^{M}\omega_j^2\left(\frac{\|\boldsymbol{X}_{G_j}\widehat{\boldsymbol{\beta}}_{G_j}-\boldsymbol{X}_{G_j}\boldsymbol{\beta}^*_{G_j}\|_2}{\omega_j}\right)^q\right\}^{1/q} &= \left\{\sum_{j=1}^{M}\omega_j^2\left(\frac{\|\widehat{\boldsymbol{b}}_{G_j}-\boldsymbol{b}^*_{G_j}\|_2}{\omega_j}\right)^q\right\}^{1/q}\\ &\le \frac{2\sigma^*\xi\big(\sum_{j\in S^*}\omega_j^2\big)^{1/q}}{\sqrt{1-\tau_+}\,\mathrm{SCIF}_q^{(G)}(\xi,\boldsymbol{\omega},S^*)},\quad q=1,2,\end{aligned}$$

when the conditions for (4.3.22), including the definition of the estimator and the SCIF, hold with $\boldsymbol{X}$, $\boldsymbol{\beta}$ *and* $\boldsymbol{\beta}^*$ *replaced by* $\boldsymbol{U}$, $\boldsymbol{b}$ *and* $\boldsymbol{b}^*$ *respectively.*

Remark 4.4. Corollary 4.1 can be viewed as an extension of the main results of Huang and Zhang [2010] to the scaled group Lasso although here the regularity condition of the design is of a weaker form and smaller penalty levels are allowed.

Proof of Theorem 4.5. We follow the proof in Sun and Zhang [2012a]. Let $t\ge\sigma^*/\sqrt{1+\tau_-}$ and $\boldsymbol{h}_{G_j}=\widehat{\boldsymbol{\beta}}_{G_j}(t\boldsymbol{\omega})-\boldsymbol{\beta}^*_{G_j}$. As the oracle noise level is defined as $(\sigma^*)^2=\|\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}^*\|_2^2/n$, we have

$$(\sigma^*)^2-\|\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}(t\boldsymbol{\omega})\|_2^2/n=(\mathbf{X}\boldsymbol{h})^T(2\boldsymbol{\varepsilon}-\mathbf{X}\boldsymbol{h})/n=(\mathbf{X}\boldsymbol{h})^T(\boldsymbol{\varepsilon}+\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}(t\boldsymbol{\omega}))/n \quad (4.3.26)$$

Suppose $\mathcal{E}$ happens so that $\|\mathbf{X}_{G_j}^T\boldsymbol{\varepsilon}\|_2/n \leq t\omega_j(\xi-1)/(\xi+1)$. It follows that

$$\left|(\mathbf{X}\boldsymbol{h})^T\boldsymbol{\varepsilon}/n\right| = \left|\sum_{j=1}^M \boldsymbol{h}_{G_j}^T\mathbf{X}_{G_j}^T\boldsymbol{\varepsilon}/n\right| \leq \frac{\xi-1}{\xi+1}\sum_{j=1}^M t\omega_j\|\boldsymbol{h}_{G_j}\|_2.$$

Moreover, the KKT condition implies

$$\left|\boldsymbol{h}^T\mathbf{X}^T(\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}(t\boldsymbol{\omega}))/n\right| = \left|\sum_{j=1}^M \boldsymbol{h}_{G_j}^T\mathbf{X}_{G_j}^T(\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}(t\boldsymbol{\omega}))/n\right| \leq \sum_{j=1}^M t\omega_j\|\boldsymbol{h}_{G_j}\|_2.$$

As $(\mathbf{X}\boldsymbol{h})^T(2\boldsymbol{\varepsilon}-\mathbf{X}\boldsymbol{h})/n \leq 2(\mathbf{X}\boldsymbol{h})^T\boldsymbol{\varepsilon}/n$, inserting these inequalities to (4.3.26) yields

$$-\left(\frac{\xi-1}{\xi+1}+1\right)\sum_{j=1}^M t\omega_j\|\boldsymbol{h}_{G_j}\|_2 \leq \sigma^{*2} - \|\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}(t\boldsymbol{\omega})\|_2^2/n \leq 2\frac{\xi-1}{\xi+1}\sum_{j=1}^M t\omega_j\|\boldsymbol{h}_{G_j}\|_2.$$

A rescaled version $\widehat{\boldsymbol{\beta}}(t\boldsymbol{\omega})$ can be written as

$$\frac{\widehat{\boldsymbol{\beta}}(t\boldsymbol{\omega})}{t} = \arg\min_{\boldsymbol{b}}\left\{\frac{\|\boldsymbol{y}/t-\mathbf{X}\boldsymbol{b}\|_2^2}{2n} + \sum_{j=1}^M \omega_j\|\boldsymbol{b}_{G_j}\|_2\right\}$$

as the group Lasso estimator with target $\boldsymbol{\beta}^*/t$ and noise vector $\boldsymbol{\varepsilon}/t$. As $t \geq \sigma^*/\sqrt{1+\tau_-}$, the condition of Theorem 4.4 is satisfied with the rescaled noise $\boldsymbol{\varepsilon}/t$, so that

$$t^{-1}\sum_{j=1}^M \omega_j\|\boldsymbol{h}_{G_j}\|_2 = \sum_{j=1}^M \omega_j\|\widehat{\boldsymbol{\beta}}_{G_j}(t\boldsymbol{\omega})/t - \boldsymbol{\beta}_{G_j}^*/t\|_2 < \mu(\boldsymbol{\omega},\xi).$$

As $\tau_- = 2\mu(\boldsymbol{\omega},\xi)(\xi-1)/(\xi+1)$ and $\tau_+ = \mu(\boldsymbol{\omega},\xi)\{(\xi-1)/(\xi+1)+1\}$, we have

$$-\tau_+t^2 = -\left(\frac{\xi-1}{\xi+1}+1\right)t^2\mu(\boldsymbol{\omega},\xi) < \sigma^{*2} - \|\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}(t\boldsymbol{\omega})\|_2^2/n < 2\frac{\xi-1}{\xi+1}t^2\mu(\boldsymbol{\omega},\xi) = \tau_-t^2.$$

The upper bound above for $t = \sigma^*/\sqrt{1+\tau_-}$ implies

$$t^2 - \|y-\mathbf{X}\widehat{\boldsymbol{\beta}}(t\boldsymbol{\omega})\|_2^2/n < t^2 - (\sigma^*)^2 + \tau_-t^2 = 0,$$

so that $\widehat{\sigma} > t = \sigma^*/\sqrt{1+\tau_-}$ by Lemma 4.2. Similarly, the lower bound yields $\widehat{\sigma} \geq \sigma^*/\sqrt{1-\tau_+}$.

As $\widehat{\sigma} \geq \sigma^*/\sqrt{1+\tau_-}$, the error bounds in Theorem 4.4 holds for $\{\boldsymbol{y}/\widehat{\sigma}, \boldsymbol{\beta}^*/\widehat{\sigma}, \widehat{\boldsymbol{\beta}}/\widehat{\sigma}\}$, which implies (4.3.21) and (4.3.22) due to $\widehat{\sigma} \leq \sigma^*/\sqrt{1-\tau_+}$. When (4.1.1) holds with Gaussian error, $|\widehat{\sigma}/\sigma^* - 1| = o_P(\mu(\boldsymbol{\omega}, \xi)) = o_P(n^{-1/2})$ by (4.3.20) and the condition in $\mu(\boldsymbol{\omega}, \xi)$, so that (4.3.23) follows from the central limit theorem for $\sigma^*/\sigma \sim \chi_n/\sqrt{n}$.

Let $\boldsymbol{u}^* = \boldsymbol{\varepsilon}/\|\boldsymbol{\varepsilon}\|_2$, $\boldsymbol{Q}_{G_j}$ be the orthogonal projection to the range of $\mathbf{X}_{G_j}$ and $f(\boldsymbol{u}^*) = \|\boldsymbol{Q}_{G_j}\boldsymbol{u}^*\|_2$. As $f(\boldsymbol{u}^*) = 1$ for $n = 1$, we assume $n \geq 2$ without loss of generality. The vector $\boldsymbol{u}^*$ is uniformly distributed in the sphere $\mathbb{S}^{n-1}$ and $f(\boldsymbol{u}^*)$ is a unit Lipschitz function of $\boldsymbol{u}^*$ with median $\sqrt{m_{d_j,n}}$. As $\sigma^* = \|\boldsymbol{\varepsilon}\|_2/\sqrt{n}$, $\|\mathbf{X}_{G_j}^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^*)/(n\sigma^*)\|_2 \leq f(\boldsymbol{u}^*)$ when $\|\mathbf{X}_{G_j}/\sqrt{n}\|_S \leq 1$. In this case and for $t > 0$ and $n \geq 2$,

$$\mathbb{P}\left\{\|\boldsymbol{Q}_{G_j}\boldsymbol{u}^*\|_2 \geq \sqrt{m_{d_j,n}} + \frac{t}{\sqrt{n-3/2}}\right\} \leq e^{(4n-6)^{-2}}\mathbb{P}\Big\{\mathsf{N}(0,1) > t\Big\} \leq e^{-t^2/2}$$

by the Lévy concentration inequality as in Lemma 17 of Sun and Zhang [2013]. Thus, $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$ by the union bound when $(\xi-1)\omega_j/\{(\xi+1)\sqrt{1+\tau_-}\} \geq \omega_{*,j}$. Now, consider $\omega_j = A\omega_{*,j}$. Let $\tau_* = 2\mu(\boldsymbol{\omega}_*, \xi)(\xi-1)/(\xi+1)$. It follows from (4.3.1) and (4.3.6) that $\mu(\boldsymbol{\omega}, \xi) = A^2\mu(\boldsymbol{\omega}^*, \xi)$, so that $\tau_- = A^2\tau_*$. Consequently,

$$\frac{(\xi-1)\omega_j}{(\xi+1)\sqrt{1+\tau_-}\omega_j^*} = \frac{(\xi-1)A}{(\xi+1)\sqrt{1+A^2\tau_-^*}} \geq 1$$

if and only if $A \geq \{(\xi+1)/(\xi-1)\}/\{1-\{(\xi+1)/(\xi-1)\}^2\tau_-^*\}^{1/2} = A_*$. Finally, we note that $\sqrt{m_{d_j,n}} \leq \mathbb{E}f(\boldsymbol{u}^*) + e^{(4n-6)^{-2}}\mathbb{E}|\mathsf{N}(0, 1/(n-3/2))|/2 \leq (d_j/n)^{1/2} + n^{-1/2}$. ■

4.4 Simulation Results

We provide a few simulation results for our theories developed in Sections 4.2.1 and 4.3. As a prelude, in the following we first show the performance of scaled group

Lasso procedure in a simulation experiment. We consider a two simulation designs with $(n = 1000, p = 200)$ and $(n = 1000, p = 2000)$ design matrices with the elements of the design matrix generated independently from $\mathsf{N}(0, 1)$. We assume that the true parameter $\boldsymbol{\beta}^*$ has an inherent grouping with total set of p parameters divided into groups of size $d_j = 4$. In the design $(n = 1000, p = 200)$ we have total number of groups $M = 50$ and in $(n = 1000, p = 2000)$, $M = 500$. For both scenarios, the true parameter $\boldsymbol{\beta}^*$ is assumed to be $(g = 2, s = 8)$ strong group sparse with its non-zero coefficients in $\{-1, 1\}$. Both simulation designs have a $\mathsf{N}(0, \sigma^2)$ error added to the true regression model $\mathbf{X}\boldsymbol{\beta}^*$ with $\sigma = 1$. We also assume that the design matrix is group wise orthogonalized in the sense of $\mathbf{X}_{G_j}^T \mathbf{X}_{G_j}/n = \mathbf{I}_{G_j \times G_j}$, $j = 1, \ldots, M$.

In estimation of σ we employ the scaled group Lasso procedure as shown in 4.3.17. The groupwise penalty factors ω_j's are chosen to equal to $\lambda\sqrt{d_j}$ for some fixed $\lambda > 0$. The implementation of group Lasso procedure is via the R package `gglasso`.

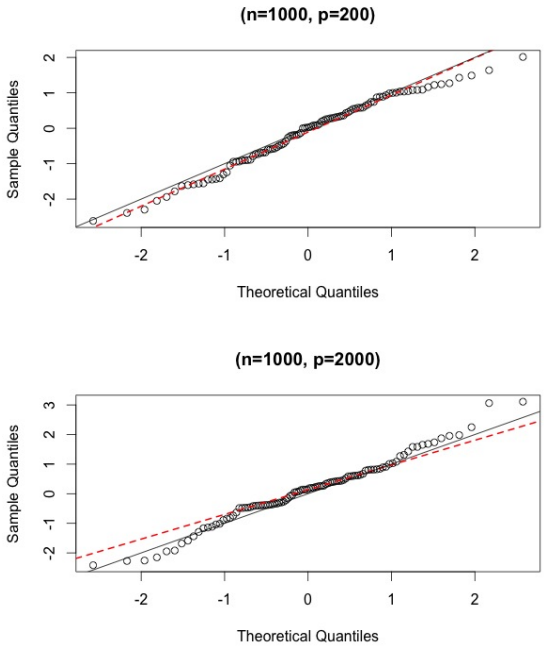

Figure 4.1: Normal QQ plot for the test statistic for $\widehat{\sigma}$ in (4.3.23) in Theorem 4.5 with $n = 1000, p = \{200, 2000\}, g = 2, s = 8$. The results are produced with 100 replications of the scaled group Lasso. The red dotted line is fitted through 1^{st} and 3^{rd} sample quantile.

n	p	# of Groups M	g	s	$\widehat{\sigma}(\text{SE}(\widehat{\sigma}))$
1000	200	50	2	8	0.997 (0.02)
1000	2000	500	2	8	1.002(0.02)

Table 4.1: Table summarizing simulation set up for two scenarios along with estimate of scale parameter after 100 replications along with its standard error. The true value of the scale parameter is $\sigma = 1$.

Table 4.1 summarizes the two design setups and estimates of scale parameter. In the design setup with $(n = 1000, p = 200)$, the estimate of $\widehat{\sigma}$ averaged over a 100 replications is 0.997 with a standard deviation of 0.02. In the design setup with $(n = 1000, p = 2000)$, the estimate of $\widehat{\sigma}$ averaged over a 100 replications is 1.0002

with a standard deviation of 0.02. Additionally Figure 4.1 shows the Gaussian Q-Q plots of the test statistic $\sqrt{2n}\,(\widehat{\sigma}/\sigma - 1)$.

4.4.1 Asymptotic test statistic

We also seek the empirical validation of the asymptotic convergence of the group $\boldsymbol{\beta}_{G_j}$ as described in our theoretical results. For bias correction we take the penalty function in (4.2.32) to be the Frobenius norm and apply group Lasso based optimization. We also consider a new simulation design similar as before with $(n = 1000, p = 200)$ and $\sigma = 1$. We will consider two different schemes for empirical analysis for asymptotic convergence.

Small group sizes

The true parameter $\boldsymbol{\beta}^*$ is simulated to be $(s = 40, g = 10)$ strong group sparse with its nonzero values in the interval [2,3]. More specifically, $\boldsymbol{\beta}^*$ is grouped into groups of sizes $d_j = 4$ for all j. We construct the test statistic of $\boldsymbol{\mu}_{G_j}$ as in (4.2.14) for one of the nonzero groups. Figure 4.2 provides χ^2_4 based Q-Q plot for the sample quantiles of our test statistic.

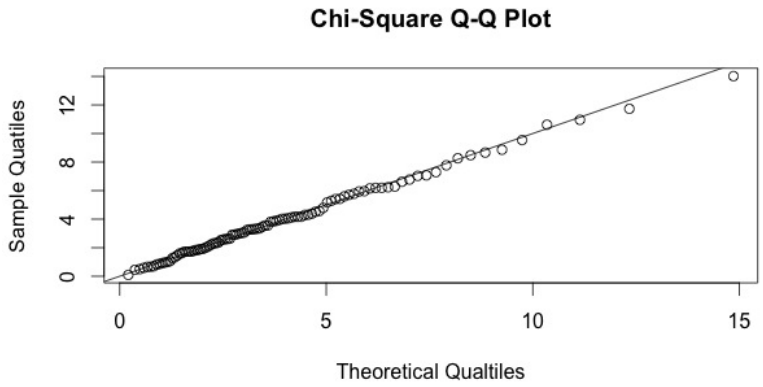

Figure 4.2: Chi square Q-Q plot for the test statistic for $\widehat{\boldsymbol{\mu}}_{G_j}$ with $n = 1000, p = 200, g = 10, s = 40$. The theoretical quantiles were drawn from χ^2_4 random variable. The group being tested has size 4.

Large group sizes

The true parameter $\boldsymbol{\beta}^*$ is simulated to be $(s = 40, g = 2)$ strong group sparse with its nonzero values between [2,3]. More specifically, $\boldsymbol{\beta}^*$ is grouped into 20 groups each of sizes $d_j = 20$ for all j. We let the sparsity of the true parameter $\boldsymbol{\beta}^*$ to be $s = 40$ contained within 2 separate groups. Again, we construct the test statistic of $\boldsymbol{\mu}_{G_j}$ as in (4.2.14) for one of the nonzero groups. Figure 4.3 shows the Q-Q plot for this group's test statistic. As the figure suggests, for large group sizes asymptotic normality of the group test statistic is empirically supported.

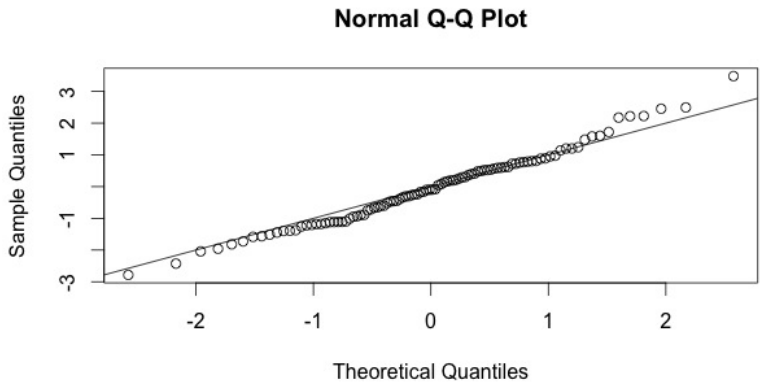

Figure 4.3: Normal QQ plot for the test statistic for $\widehat{\boldsymbol{\mu}}_{G_j}$ with $n = 1000, p = 200, g = 2, s = 40$. Here the group size of the test group is 20.

Bibliography

T. W. Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958. 6

A. Antoniadis. Comments on: ℓ_1-penalization for mixture regression models. *Test*, 19(2):257–258, 2010. 91

M. A. Arcones and E. Gine. Limit theorems for u-processes. *The Annals of Probability*, pages 1494–1542, 1993. 39

F. R. Bach. Consistency of the Group Lasso and Multiple Kernel Learning. *The Journal of Machine Learning Research*, 9:1179–1225, jun 2008. 69

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008. 42

R. F. Barber and M. Kolar. Rocket: Robust confidence intervals via kendall's tau for transelliptical graphical models. *arXiv preprint arXiv:1502.07641*, 2015. 44, 46, 55

A. Belloni, V. Chernozhukov, and L. Wang. Square-root Lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011. 44, 69, 91

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81 (2):608–650, 2014. 68

R. Berk, L. Brown, and L. Zhao. Statistical inference after model selection. *Journal of Quantitative Criminology*, 26:217–236, 2010. 68

P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227, 2008a. 9

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008b. 9

P. J. Bickel, J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993. 5, 68

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009. 53, 86

C. Borell. The brunn-minkowski inequality in gauss space. *Inventiones Mathematicae*, 30(2):207–216, 1975. 35, 59

P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.*, 5: 232–253, 2011. 69

P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242, sep 2013. 68

P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011. 69

F. Bunea, J. Lederer, and Y. She. The group square-root lasso: Theoretical properties and fast algorithms. *Information Theory, IEEE Transactions on*, 60(2):1313–1325, 2014. 69, 91

T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494), 2011. 9

T. Cai and H. Zhou. Minimax estimation of large covariance matrices under ℓ_1 norm. *Statistica Sinica*, 22(4):1319, 2012. 9

T. Cai, L. Wang, and G. Xu. Shifting inequality and recovery of sparse signals. *IEEE Transactions on Signal Processing*, 58:1300–1308, 2010a. 63, 88

T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106 (494):594–607, 2011. 42

T. T. Cai and M. Yuan. Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics*, 40(4):2014–2042, aug 2012. 9, 28

T. T. Cai, C.-H. Zhang, and H. H. Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010b. 9, 26, 27

T. T. Cai, Z. Ma, and Y. Wu. Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, dec 2013. 9

E. Candes and T. Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007. 42

E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. on Information Theory*, 51:4203–4215, 2005. 86

A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM review*, 49(3):434–448, 2007. 9

K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1:317–366, 2001. 7, 47

C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, mar 1970. 28

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008. 42

J. Hájek. Asymptotic normality of simple linear rank statistics under alternatives. *The Annals of Mathematical Statistics*, pages 325–346, 1968. 14

J. Hájek, Z. Šidák, and P. K. Sen. *Theory of rank tests*. Academic press New York, 1967. 14

F. Han and H. Liu. Optimal Rates of Convergence for Latent Generalized Correlation Matrix Estimation in Transelliptical Distribution. *arXiv preprint arXiv:1305.6916*, page 34, may 2013. 8, 9, 18, 22, 23, 43, 46, 60

W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. 7, 14, 31

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963. 7, 40

J. Huang and T. Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38 (4):1978–2004, aug 2010. 67, 69, 70, 85, 89, 96

J. Huang, S. Ma, H. Xie, and C.-H. Zhang. A group bridge approach for variable selection. *Biometrika*, 96:339–355, 2009. 69

J. Huang, P. Breheny, and S. Ma. A selective review of group selection in high-dimensional models. *Statistical Science*, 27(4):481–499, 2012. 69

P. J. Huber. *Robust statistics*. Springer, 2011. 91

J. Jankova and S. van de Geer. Confidence intervals for high-dimensional inverse covariance estimation. *arXiv preprint arXiv:1403.6752*, 2014. 68

A. Javanmard and A. Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60:6522 – 6554, July 2014. 68, 73

I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104 (486), 2009. 9

I. T. Jolliffe, N. T. Trendafilov, and M. Uddin. A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003. 9

N. E. Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, pages 2717–2756, 2008. 9

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. 6

M. G. Kendall. Rank correlation methods. 1948. 6, 13, 46

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378, oct 2000. 67

V. Koltchinskii. The dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15: 799–828, 2009. 86

V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In *Proceedings of COLT*, 2008. 69

W. H. Kruskal. Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–861, 1958. 6, 13, 46

E. Laber and S. Murphy. Adaptive confidence intervals for the test error in classification (with discussion). *Journal of the American Statistical Association*, 106: 904–913, 2011. 68

C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37(6B):4254, 2009. 9, 42

H. Leeb and B. M. Potscher. Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34:2554–2591, 2006. 68

H. Liu and L. Wang. Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. *arXiv preprint arXiv:1209.2437*, 2012. 43

H. Liu and J. Zhang. Estimation consistency of the group lasso and its applications. *Journal of Machine Learning Research-Proceedings Track*, 5:376–383, 2009. 69

H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, 10:2295–2328, 2009. 5, 43

H. Liu, F. Han, M. Yuan, J. Lafferty, L. Wasserman, et al. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012a. 6, 7, 43, 46, 51

H. Liu, F. Han, and C.-h. Zhang. Transelliptical graphical models. In *Advances in Neural Information Processing Systems*, pages 809–817, 2012b. 7, 8, 44

R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, apr 2014. 68

K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 2011. 69, 86, 88

Z. Ma. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013. 9, 80

N. Meinshausen. Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2014. 68

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006. 42

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010. 68

N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488), 2009. 68

R. Mitra and C.-H. Zhang. Multivariate analysis of nonparametric estimates of large correlation matrices. *arXiv preprint arXiv:1403.6195*, 2014a. 4, 41, 43, 45, 46, 55, 56, 57, 58, 60, 62

R. Mitra and C.-H. Zhang. The benefit of group sparsity in group inference with de-biased scaled group lasso. *arXiv preprint arXiv:1412.4170*, 2014b. 4, 53

Y. Nardi and A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633, 2008. 69, 86

G. Obozinski, M. J. Wainwright, and M. I. Jordan. Union support recovery in high-dimensional multivariate regression. In *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on*, pages 21–26. IEEE, 2008. 69

P. Ravikumar, M. J. Wainwright, G. Raskutti, B. Yu, et al. High-dimensional covariance estimation by minimizing ℓ_1 penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011. 42

P. K. Ravikumar, G. Raskutti, M. J. Wainwright, and B. Yu. Model selection in gaussian graphical models: High-dimensional consistency of ℓ_1-regularized mle. In *Advances in Neural Information Processing Systems*, pages 1329–1336, 2008. 42

Z. Ren, T. Sun, C.-H. Zhang, and H. H. Zhou. Asymptotic normality and optimalities in estimation of large Gaussian graphical model. *arXiv preprint arXiv:1309.6024*, 2013. 68

G. V. Rocha, P. Zhao, and B. Yu. A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice). *arXiv preprint arXiv:0807.3734*, 2008. 42

A. J. Rothman, P. J. Bickel, E. Levina, J. Zhu, et al. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008. 42

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*, volume 162. Wiley-Interscience, 2009. 14

R. D. Shah and R. J. Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):55–80, 2013. 68

A. Sklar. Fonctions de répartition à n dimensions e leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8(3):229–231, 1959. 5

N. Städler, P. Bühlmann, and S. Geer. ℓ_1-penalization for mixture regression models. *TEST*, 19(2):209–256, jun 2010. 91

T. Sun and C.-H. Zhang. Comments on: ℓ_1-penalization for mixture regression models. *Test*, 19(2):270–275, 2010. 91, 92

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012a. 3, 42, 48, 51, 52, 53, 67, 91, 92, 96

T. Sun and C.-H. Zhang. Comments on: Optimal rates of convergence for sparse covariance matrix estimation. *Statistica Sinica*, 22:1354–1358, 2012b. 68

T. Sun and C.-H. Zhang. Sparse matrix inversion with scaled lasso. *Journal of Machine Learning Research*, 14:3385–3418, 2013. 3, 42, 44, 47, 48, 51, 52, 63, 64, 91, 98

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288, 1996a. 69

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58:267–288, 1996b. 42

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, aug 2011. 18

S. van de Geer. The deterministic lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich, 2007. 53, 86

S. van de Geer. Worst possible sub-directions in high-dimensional models. *Contributions in infinite-dimensional statistics and related topics*, page 131, 2014. 73

S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009. 86

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 06 2014. 67, 68, 73

A. van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge university press, 2000. 14

A. W. Van Der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996. 59

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed sensing*, pages 210–268, 2012. 7, 47, 80

V. Q. Vu and J. Lei. Minimax rates of estimation for sparse pca in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, volume 22, 2012. 9, 10, 28

Z. Wang, F. Han, and H. Liu. Sparse principal component analysis for high dimensional multivariate time series. *Journal of Machine Learning Research (AISTATS Track)*, 2013. 29

L. Wasserman and K. Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009. 68

M. Wegkamp and Y. Zhao. Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *arXiv preprint arXiv:1305.6526*, may 2013. 8, 13, 18, 22, 23, 24, 39, 40, 43, 46, 54, 62

H. Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen. *Mathematische Annalen*, 71:441–479, 1912. 28

W. B. Wu and M. Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844, 2003. 9

L. Xue and H. Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571, 2012. 6, 7, 43, 46

L. Xue and H. Zou. Rank-based tapering estimation of bandable correlation matrices. *Statistica Sinica*, 24:83–100, 2014. 9, 26

S. Yang and E. D. Kolaczyk. Target detection via network filtering. *Information Theory, IEEE Transactions on*, 56(5):2502–2515, 2010. 42

F. Ye and C.-H. Zhang. Rate Minimaxity of the Lasso and Dantzig Selector for the lq Loss in lr Balls. *The Journal of Machine Learning Research*, 11:3519–3540, 2010. 51, 53, 63, 86, 88

M. Yuan. High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research*, 11:2261–2286, 2010. 42, 50

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

M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007. 42

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, pages 894–942, 2010. 49

C.-H. Zhang. Statistical inference for high-dimensional data. In *Mathematisches Forschungsinstitut Oberwolfach: Very High Dimensional Semiparametric Models, Report No. 48/2011*, pages 28–31, 2011. 67, 68

C.-H. Zhang and J. Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008. 49, 86

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, jan 2014. 3, 66, 68, 72, 73

T. Zhang and H. Zou. Sparse precision matrix estimation via lasso penalized d-trace loss. *Biometrika*, page ast059, 2014. 43

T. Zhao and H. Liu. Calibrated precision matrix estimation for high-dimensional

elliptical distributions. *IEEE TRANSACTIONS ON INFORMATION THEORY*, 60(12):7874–7887, 2014. 44

H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006. 9