High Dimensional Correlation Networks and Their Applications

by

Hamed Firouzi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in The University of Michigan
2015

Doctoral Committee:

Professor Alfred O. Hero III, Chair
Assistant Professor Laura K. Balzano
Professor Roman Vershynin
Professor Ji Zhu

To my parents

ACKNOWLEDGEMENTS

This thesis would not have been possible without the support of my advisor, committee members, parents, mentors, colleagues, and friends.

Specifically, I would like to thank my advisor, Professor Alfred Hero, who has given me all the support and guidance I needed as a doctoral student. I am grateful to have had an advisor who trusted my ability and respected my decisions. I have truly enjoyed learning from his bright and creative mind, his excellent critical thinking skills, his deep insights, and his vast knowledge, throughout my PhD. I am forever indebted to him for the priceless things I have learned from him.

I also owe my committee member, Professor Roman Vershynin, Professor Laura Balzano, and Professor Ji Zhu a great deal of thanks for their time and advice which is reflected in this thesis. My sincere thanks also goes to Professor Bala Rajaratnam whose constant encouragement, insightful questions, and valuable feedback has made a significant impact on my research.

I would also like to thank the excellent professors from whom I have had the chance to learn, in particular, Professors Raj Nadakuditi, Clayton Scott, Jeffery Fessler, Silvio Savarese, Kim Winick, Marina Epelman, Quentin Stout, Nejat Seyhun, Erhan Bayraktar, Reza Kamaly, Jussi Keppo, Brian Thelen, Alexander Barvinok, Mark Rudelson, Selim Esedoglu, Ralf Spatzier, Jinho Baik, Daniel Burns, Mattias Jonsson, David Kaufman, Bin Nan, Shyamala Nagaraj, Tyler Shumway, Dennis Capozza, Steve Slezak, and Marc Lipson.

I would like to show my gratitude to Shelly Feldkamp, Becky Turanksi, Rachel

Antoun, Gail Carr, Beth Lawson, and Karen Liska for all of their help regarding the administrative matters.

I thank my fellow labmates in Michigan including Zhaoshi, Mark, Dennis, Alex, Dae-Yon, Yu-Hui, Joel, Tianpei, Pin-Yu, Taposh, Kevin, Kristjan, Brendan, Greg, Joyce, Sung-Jin, and Arnau, for the valuable discussions and the constant support.

Living in Ann Arbor has been a unique and joyful experience because of Aria, Mahmood, Hamed, Amirreza, Mohammadreza, Ali, Nima, Javid, Curtis, Christina, Dimitris, Reza, Hassan, Kamran, Kaveh, Mary, Mahta, Arash, and many other wonderful friends. I would like to specially thank Sarah for all of her help and support in preparation of this thesis.

Finally, I would like to thank my parents, my sister, and my brother, for their never ending love and tell them from the bottom of my heart that without you, this dissertation would never have been possible.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

High Dimensional Correlation Networks And Their Applications

by

Hamed Firouzi

Chair: Alfred O. Hero III

Analysis of interactions between variables in a large data set has recently attracted special attention in the context of high dimensional multivariate statistical analysis. Variable interactions play a role in many inference tasks, such as, classification, clustering, estimation, and prediction. This thesis focuses on the discovery of correlation and partial correlation structures as well as their applications in high dimensional data analysis and inference. The thesis considers problems of screening *correlation and partial correlation networks* by thresholding the sample correlation or the sample partial correlation matrix. The selection of the threshold is guided by our high dimensional asymptotic theory for screening such networks. Scalable methods of edge and hub screening are developed for applications in spatio-temporal analysis of time series, variable selection for linear prediction, and support recovery. The proposed methods are specifically designed for very high dimensional data with limited number of samples. Moreover, the correlation screening theory developed in this thesis provides high dimensional family-wise error rates on false discoveries.

CHAPTER I

Introduction

We live in an era of information explosion. Currently the term "Big Data" is used in many different contexts ranging from genomics to social networks to finance. Probably the most obvious reason for such information explosion is technology. Nowadays, sensors and gadgets digitize lots of information that was previously unavailable. As a result, methods for analysis of large data sets have become essential for extracting information from such data sets. In many domains the data comes in the form of a fat matrix with many columns (variables) but few rows (samples) from which information about variable inter-relations is to be extracted. Such information is valuable for performing inference tasks such as classification, clustering, estimation, and prediction. This thesis develops new methods of analysis for extracting such information in the high dimensional, sample starved regime. Specifically, we develop methods for extracting information about high dimensional correlation and partial correlation matrices from few samples.

The correlation or partial correlation network associated with a vector of variables can be specifically useful when the covariance or inverse covariance are sparse matrices. In this case, the nodes of the network correspond to the column indices (variables) and the edge locations correspond to locations within the matrix where there is a non-zero entry. The objective of edge discovery is to recover the locations

of these edges from limited samples of the variables. The objective of hub discovery is to recover nodes in the graph that have a large number of edges. While there are many ways to estimate these edge or hub locations, this thesis focuses on a simple and scalable thresholding method. This method constructs an estimated correlation or partial correlation network by thresholding the sample correlation or sample partial correlation matrix, respectively.

Motivated by practical problems such as model selection and spatio-temporal analysis of time series, we generalize the recently proposed correlation edge and hub discovery framework (*Hero and Rajaratnam*, 2011, 2012) in several ways. These frameworks are illustrated for real life data sets. Specifically, the thesis addresses four different extensions of previous edge and hub screening work. Each chapter describing these problems and their solutions have the following common structure. In each chapter we introduce a correlation network inference problem inspired by a practical application. We then propose asymptotic results associated with edge or hub screening in the defined correlation network. Afterwards we develop a scalable algorithm for overcoming the high dimensional computational complexity problem.

1.1 Overview of hub screening methods

Consider the problem of screening for variables that have significant correlations in a large and fat data matrix. Examples of such data sets are gene expression arrays, multimedia databases and multivariate financial time series. A correlation network between the variables in a data set can be formed by thresholding the absolute value of the entries of the sample correlation matrix, obtained by the outer product (along the fat dimension) of the data matrix. The thresholded sample correlation matrix yields a sparse adjacency matrix defining the sample correlation network (graph). Two nodes (vertices) are connected with an edge if the absolute value of the sample correlation coefficient between their corresponding variables is greater than a fixed

threshold (See Fig. 1.2). The Correlation screening method introduced in (*Hero and Rajaratnam*, 2011) addresses the following problem:

Problem A. (Global correlation screening). Assume that the ensemble correlation matrix is sparse and that the non-sparse components correspond to variables that are highly correlated with other variables. The problem is to discover the highly correlated variables with low false positive rate in cases where the number of samples may be significantly smaller than the total number of variables. A related problem is to assign statistical significance to the discovered variables.

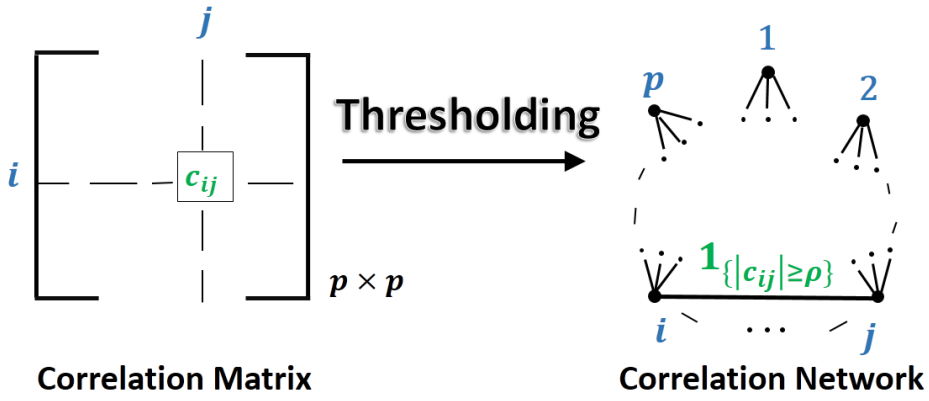

Figure 1.1: We define a sample correlation network by thresholding the magnitudes of the entries of the sample correlation matrix.

In cases where the data comes from an underlying Markovian structure, for example a Gaussian graphical model (*Lauritzen*, 1996), discovering the partial correlation structures may be of interest. For such cases constructing the sample partial correlation network by thresholding the sample partial correlation matrix can be useful. The partial correlation screening problem was introduced in (*Hero and Rajaratnam*, 2012):

Problem B. (Global partial correlation screening). Assume that the ensemble par-

tial correlation matrix is sparse and that the non-sparse components correspond to variables that are highly partially correlated with other variables. The problem is to discover the highly partially correlated variables with low false positive rate in cases where the number of samples may be significantly smaller than the total number of variables. A related problem is to assign statistical significance to the discovered variables.

Mathematically, these problems are introduced as follows. There is a random vector of dimension p, denoted by $\mathbf{X}$, from which $n < p$ samples are available. Assume that the distribution of $\mathbf{X}$ is in general family of elliptically contoured distributions with mean $\boldsymbol{\mu}$ and non-singular $p \times p$ dispersion matrix $\boldsymbol{\Sigma}$:

$$f_{\mathbf{X}}(\mathbf{x}) = g\left((\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{1.1}$$

in which g is a non-negative integrable function. Examples of such distributions include the multivariate Gaussian and the multivariate student-t. Let $0 \leq \rho \leq 1$ be a fixed correlation or partial correlation threshold. A correlation or a partial correlation network can be constructed by thresholding the magnitude of the entries of the sample correlation matrix or sample partial correlation matrix, respectively defined below

$$\mathbf{R} = \mathbf{D}_{\mathbf{S}}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}_{\mathbf{S}}^{-\frac{1}{2}}, \tag{1.2}$$

and

$$\mathbf{P} = \mathbf{D}_{\mathbf{R}^\dagger}^{-\frac{1}{2}}\mathbf{R}^\dagger, \mathbf{D}_{\mathbf{R}^\dagger}^{-\frac{1}{2}} \tag{1.3}$$

where $\mathbf{R}^\dagger$ is the Moore-Penrose pseudo-inverse of $\mathbf{R}$, $\mathbf{D}_{\mathbf{A}}$ represents the diagonal matrix that is obtained by zeroing out all but diagonal entries of the matrix $\mathbf{A}$, and

the $p \times p$ sample covariance matrix $\mathbf{S}$ is defined as:

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_{(i)} - \overline{\mathbf{X}})^T (\mathbf{X}_{(i)} - \overline{\mathbf{X}}), \tag{1.4}$$

in which $\mathbf{X}_{(i)}$ is the ith row of the $n \times p$ data matrix $\mathbb{X}$, and $\overline{\mathbf{X}}$ is the vector average of all n rows of $\mathbb{X}$. For simplicity, we use the general term *hub* for a node that is highly correlated with many other variables in a correlation network, or is highly partially correlated with other variables in a partial correlation network.

Under a null hypothesis that the true covariance matrix $\mathbf{\Sigma}$ is block sparse of degree k (i.e. by re-arranging rows and columns all of the non-zero non-diagonal entries can be collected in a $k \times k$ block) correlation and hub screening methods of (*Hero and Rajaratnam*, 2011, 2012) develop asymptotic family-wise false discovery rates to assign p-value to nodes of a (partial) correlation network for being a hub of certain node degree δ, in a regime where the number samples n is fixed and the number of variables p goes to infinity (which we also refer to as high dimensional regime, low sample regime, sample starving regime, or purely high dimensional regime). This allows screening for hubs in a (partial) correlation network given a false positive rate.

1.2 Local hub screening

The original theory of (partial) correlation screening assigns p-values to nodes of a (partial) correlation network for being a hub of certain degree. A key relationship behind the procedure of p-value assignment is the following Poisson-type equation:

$$\mathbb{P}(N_{\rho,\delta} > 0) \to 1 - \exp\left(-\mathbb{E}(N_{\rho,\delta} > 0)\right), \tag{1.5}$$

in which $N_{\rho,\delta}$ is the number of hubs of degree at least δ in a (partial) correlation network constructed using (partial) correlation threshold ρ. An important point

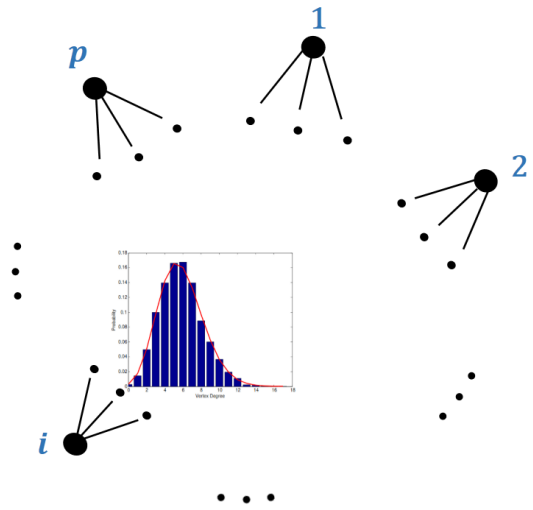

Figure 1.2: Under certain conditions, as $p \to \infty$ and $\rho \to 1$, the degree of a vertex in a (partial) correlation network is approximately a Poisson random variable.

to mention is that (1.5) holds in specific regimes where n is fixed, $p \to \infty$ and $\rho \to 1$. Since $N_{\rho,\delta}$ is a global characteristic of the (partial) correlation network, the convergence region (in terms of n, p and ρ) can be relatively small. In other words the convergence rates in (1.5) can be relatively slow (compared to some local limits) resulting in an untrusted procedure for assigning p-values as the approximation error for calculating the p-value may be of the same order as the p-value itself.

Problem 1. *(Local (partial) correlation screening).* Can we come up with a procedure for assigning p-values to the nodes of a (partial) correlation network for being hubs, which only incorporates local characteristics of the network instead of the total number of hubs $N_{\rho,\delta}$?

We show in Chapter II that under milder conditions (compared to the global Poisson-type limit of (1.5)), the degree of a node in a (partial) correlation network converges to a Poisson random variable, in total variation distance. Using the new Poisson limits, we propose a procedure for assigning p-values to the event that a node of the (partial) correlation network has abnormally high vertex degree under the block sparse covariance null hypothesis. The high dimensional convergence rate of

the proposed local hub screening procedure is faster than the original global procedure of (*Hero and Rajaratnam*, 2012) which achieves the global Poisson-type limit (1.5). Specifically, we will see that for fixed n the convergence rates associated with the local limits are at least a factor of p faster than those of the global limits. It is worth mentioning that the theory developed in Chapter II allows performing both family-wise false discovery rate and false discovery rate control on the number of hubs.

1.3 Spectral correlation hub screening of multivariate time series

Next we focus on hub screening in the context of stationary multivariate time series. Spatio-temporal correlation analysis of multivariate time series is important in applications such as wireless sensor networks, computer networks, neuroimaging and finance (*Vuran et al.*, 2004; *Paffenroth et al.*, 2013; *Friston et al.*, 2011; *Zhang et al.*, 2003; *Tsay*, 2005). Hub screening in the context of spatio-temporal analysis can be useful in identifying the *important* parameters of the spatio-temporal models aimed for various purposes such as reducing the complexity, performing sensitivity analysis, and optimal allocation of resources.

Assume that N consecutive time samples of a p-variate time series is available. Spatio-temporal analysis is computationally challenging in situations where the number of time series p is large. A naive approach is to treat the time series as a sequence of N independent identically distributed p-dimensional vectors. However, such analysis completely ignores the temporal correlations in the time series. At the other extreme, another approach is to consider all correlations between any two time instants. However, this approach would entail the estimation of an $Np \times Np$ correlation matrix which can be computationally costly as well as statistically unstable.

We consider a more general approach where we look at the stationary (partial) correlations over time within a time window of given size $n \leq N$. Following this idea, we divide the time series in to $m = N/n$ windows of n consecutive samples. However, instead of estimating the temporal correlations directly, we perform analysis on the Discrete Fourier Transforms (DFT) of the time series. In Chapter III we focus on the following problem.

Problem 2. *(Global spectral correlation screening).* Given a number N of consecutive time samples for a stationary p-variate time series, the problem is to discover hubs in the complex-valued spectral correlation matrix of the time series, for large values of p and N.

In Chapter III we show that for stationary, jointly Gaussian time series under the mild condition of absolute summability of the auto- and cross-correlation functions, different Fourier components (frequencies) become asymptotically independent of each other as the DFT length n increases. This property of stationary Gaussian processes allows us to focus on the $p \times p$ correlations at each frequency separately without having to consider correlations between different frequencies. Moreover, spectral analysis isolates correlations at specific frequencies or timescales, potentially leading to greater insight.

The spectral approach reduces the detection of hub time series to the independent detection of hubs at each frequency. However, in exchange for achieving spectral resolution, the sample size is reduced by the factor n, from N to $m = N/n$. To confidently detect hubs in this high-dimensional, low sample regime (large p, small m), as well as to accommodate complex-valued DFTs, we develop a method that we call complex-valued (partial) correlation screening. This is a generalization of the correlation and partial correlation screening method of (*Hero and Rajaratnam*, 2011, 2012) to complex-valued random variables. Using the proposed theory we develop a method for assigning p-values to the nodes in the (partial) correlation network

of the DFT components of the p time series. To make aggregate inferences based on all frequencies, straightforward procedures for multiple inference can be used as described in Chapter III.

1.4 Variable selection and prediction in high dimensional linear regression using hub screening

Consider the problem of under-determined multivariate regression in which a set of high dimensional training data $\{Y_i, X_{i1}, ..., X_{ip}\}_{i=1}^n$ is given and a linear estimate of the response variable Y_i, $1 \leq i \leq n < p$, is desired:

$$Y_i = a_1 X_{i1} + \ldots + a_p X_{ip} + \epsilon_i, \ 1 \leq i \leq n, \tag{1.6}$$

where X_{ij} is the ith sample of independent variable (also referred to as predictor variable or regressor variable) X_j, Y_i is the ith sample of dependent variable Y (also referred to as response variable), a_j is the regression coefficients corresponding to X_j, and ϵ_i is the residual, $1 \leq i \leq n, 1 \leq j \leq p$. There are many applications in which the number of predictor variables p is larger than the number of samples n. Such applications arise in text processing of internet documents, gene expression array analysis, combinatorial chemistry, and other areas (*Guyon and Elisseeff*, 2003). In this $p > n$ situation, training a linear predictor becomes difficult due to rank deficient normal equations, overfitting errors, and high computation complexity. We consider the following problem in the context of hub screening.

Problem 3. *(Two stage global screening for prediction).* Assuming a sparse ground truth linear model, the problem is to design a scalable algorithm that identifies the true support set and an accurate regression function, in the high dimensional regression setting of the model (1.6), under a budget constraint on the total number of variables sampled over the two stages of the procedure.

In chapter IV we present a general adaptive procedure for budget-limited predictor design in high dimension called two-stage Sampling, Prediction and Adaptive Regression via Correlation Screening (SPARCS). More specifically, assume that the cost of acquiring the full set of variables $\mathbf{X} = [X_1, \ldots, X_p]$ increases linearly in its dimension. SPARCS breaks the data collection into two stages in order to achieve an optimal tradeoff between sampling cost and predictor performance. In the first stage we collect a few (n) expensive samples $\{Y_i, X_{i1}, ..., X_{ip}\}_{i=1}^{n}$, at the full dimension $p \gg n$ of $\mathbf{X}$, winnowing the number of variables down to a smaller dimension $l < p$ using some form of variable selection. In the second stage we collect a larger number $(t - n)$ of cheaper samples of the l variables that passed the screening of the first stage. After the second stage, a low dimensional predictor is constructed by solving the regression problem using all t samples of the selected variables.

Unlike the global and local correlation and hub screening methods considered in (*Hero and Rajaratnam*, 2011, 2012) and in Chapter II, in the first stage of SPARCS, we screen for connectivity in a bipartite graph between the predictor variables $\{X_1, \ldots, X_p\}$ and the response variable Y. An edge exists in the bipartite graph between independent variable X_j and response variable Y if the thresholded min-norm least squares regression coefficient vector $\mathbf{B} = [b_1, \ldots, b_p]$ has a non-zero jth entry. When the jth entry of this thresholded vector is zero the jth independent variable is thrown out. Using this idea, in Chapter IV we propose a scalable algorithm called predictive correlation screening (PCS) for performing variable selection at the first stage of SPARCS procedure. We show that two-stage SPARCS predictor that uses PCS in the first stage outperforms the two-stage predictors which use state of the art variable selection methods such as LASSO (*Tibshirani*, 1996) and SIS (*Fan and Lv*, 2008).

The SPARCS method proposed in Chapter IV considers the case of scalar response variable. The results for the general case where the response is a vector are presented in the Appendix A.

1.5 Covariance and inverse covariance support recovery via correlation and partial correlation thresholding

Finally, we consider the problem of (inverse) covariance support recovery. Discovering the structure of a high-dimensional covariance matrix or its inverse (also referred to as (inverse) covariance support recovery) is an attractive problem which is useful in various contexts. In the context of covariance estimation, discovering the structure of the covariance matrix or the inverse covariance matrix can be the first stage of a two-stage esimator of the covriance matrix or its inverse. The second stage of such two-stage procedure is to estimate the non-zero entries of the (inverse) covariance matrix given the support recovered at the first stage. In the context of graphical models, inverse covariance support recovery can be used to tackle the problem of learning the structure of graphical models. It is well known that the zeros in the inverse covariance matrix of multivariate normal distribution imply the absence of an edge in the corresponding graphical model (*Bishop et al.*, 2006). Discovering such structure is of interest in many applications such as social networks, epidemiology, and finance.

Motivated by above applications, we focus on the following problem in the context of correlation screening.

Problem 4. *(Global screening for support recovery).* Assuming a sparse population (inverse) covariance matrix, the problem is to design a scalable algorithm that accurately detects the non-zero entries of the high dimensional population (inverse) covariance matrix.

In Chapter V we use correlation and partial correlation edge screening for the purpose of covariance and inverse covariance support recovery. It is well known that in high-dimensional regimes, where the number of samples n is relatively small compared to the number of variables p, the sample covariance matrix performs poorly as an

estimator of the population covariance matrix. In Chapter V we show that despite poor estimation performance, thresholding the sample (partial) correlation coefficients can perform well in discovering the true structure (i.e., the detection of non-zero entries) of the population (inverse) covariance matrix. We generalize the support recovery results presented in Chapter IV for the problem of covariance structure discovery. More specifically, we show that in a purely high-dimensional regime where n is fixed and p goes to infinity, under certain conditions, the total number of edges in a (partial) correlation network converges to a Poisson random variable. Using the proposed Poisson asymptotic result we introduce an algorithm for discovering the edges of a (partial) correlation network at a specified false discovery rate. We show that, under the assumption of elliptically contoured distribution, such structure discovery method only requires $n = \Theta(\log p)$ samples to recover the true structure with probability converging to one.

1.6 List of relevant publications

Book chapters:

H. Firouzi, D. Wei, and A.O. Hero, "Spectral Correlation Hub Screening of Multivariate Time Series," *Excursions in Harmonic Anlaysis*, Eds. R. Balan, M. Begue, J. J. Benedetto, W. Czaja and K. Okoudjou, Springer 2014.

Journal publications:

H. Firouzi, B. Rajaratnam, and A.O. Hero, "Two-stage Sampling, Prediction and Adaptive Regression via Correlation Screening (SPARCS)", to be submitted to *IEEE Transactions on Information Theory.*

Conference publications:

H. Firouzi, B. Rajaratnam, and A.O. Hero, "Predictive Correlation Screening: Appli-

cation to Two-stage Predictor Design in High Dimension," *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013.

H. Firouzi, and A.O. Hero, "Local hub screening in sparse correlation graphs," *SPIE Optical Engineering+ Applications*. International Society for Optics and Photonics, 2013.

H. Firouzi, D. Wei, and A.O. Hero. "Spatio-Temporal Analysis of Gaussian WSS Processes via Complex Correlation and Partial Correlation Screening," *Global Conference on Signal and Information Processing (GlobalSIP)*, IEEE, 2013.

H. Firouzi, B. Rajaratnam, and A.O. Hero. "Two-stage variable selection for molecular prediction of disease," *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 5th International Workshop on. IEEE, 2013.

CHAPTER II

Local hub screening

2.1 Introduction

In this chapter we present a method called *local hub screening* for detecting hubs in a sparse correlation or partial correlation network over p nodes. The proposed method is related to the hub screening method (*Hero and Rajaratnam*, 2012) where a Poisson-type limit is used to specify p-values on the number of spurious hub nodes found in the network. In this chapter we also establish Poisson limits. However, instead of being on the global number of hub nodes found, here the Poisson limit applies to the node degree found at an individual node. This allows us to define asymptotic p-values that are local to each node. We will see that the convergence rates for proposed local hub screening method are at least a factor of p faster than those of global correlation and hub screening. We illustrate the proposed method for a connectomics application on a to fMRI brain activation dataset.

Identifying hubs in correlation and partial correlation networks is an important problem which arises in applications such as gene expression analysis, sensor networks and information theoretic imaging (*Friedman et al.*, 2008; *Wiesel et al.*, 2010; *Liu et al.*, 2012). Similar to the previous work (*Hero and Rajaratnam*, 2011, 2012) we consider the problem of detection of hubs of high correlation in large scale networks. Hub screening methods (*Hero and Rajaratnam*, 2012, 2011) attempt to identify the

hubs in a (partial) correlation network by hard thresholding the magnitude of the entries of the sample (partial) correlation matrix.

Correlation and hub screening methods use a U-score representation of the sample (partial) correlation matrix to obtain an asymptotic expression for the expected number of hubs as the number of nodes p becomes large while the number of samples n is fixed. When the (partial) correlation matrix is sparse, the asymptotic expression for the expected number of hubs, does not depend on the underlying joint distribution of the variables. Furthermore, the probability that there exists at least one hub of degree greater than a given integer δ converges to the probability that a Poisson random variable exceeds zero, where the rate of the Poisson variable is the expected number of hubs of degree $> \delta$. Computing this probability enables assignment of p-values to different nodes of the network. Since the expected number of hubs in a correlation network is a general property of the network (not a local one), the approximate p-value obtained for a specific node depends on all nodes in the network.

It has been shown that (*Hero and Rajaratnam*, 2011, 2012) there is an abrupt phase transition in the number of non-zero entries of the thresholded sample (partial) correlation matrix as a function of the correlation threshold. Such a phase transition can be observed even in cases where there is no actual correlation between the nodes in the network. The value of the critical threshold ρ_c where the phase transition occurs, can be found as a function of n, p and the underlying distribution of the data. When the sample (partial) correlation matrix is thresholded with correlation threshold $\rho < \rho_c$, there will be many false discoveries. The situation becomes worse when the number of nodes p is significantly larger than the number of samples n ($n \ll p$). Indeed, in such a high dimensional regime, the value of the critical threshold ρ_c approaches 1 and almost all hub discoveries are false alarms.

In this chapter we introduce a new (partial) correlation screening method, called local hub screening, that relies on a Poisson limit for the degree distribution of the

degree of any node in the network. We give a theorem which shows that, for specific regimes of n, p and ρ, the degree of a specific vertex in the thresholded (partial) correlation graph is approximately a Poisson random variable. We also provide an expression for the rate of the Poisson processes corresponding to different vertices. The Poisson approximation allows us to assign p-value on observed hub degree to each node of the network. We show that the rate of convergence to the Poisson limit is at least a factor of p faster than the rates that govern previous hub screening methods. As a result, the local hub screening method proposed in this chapter can be applied to a wider range of p and n with higher accuracy. Unlike previous correlation and hub screening methods, the p-values assigned to a specific node depend only on the local dependencies of that node.

The rest of this chapter is organized as follows. Section 2.2 provides the necessary preliminaries. In Sec. 2.3 we introduce our theory for local hub screening method. Also we present a numerical example which validates the theoretical predictions. Finally, in Sec. 2.4 we propose our local hub screening method for assigning p-values.

2.2 Preliminaries and notations

Assume $\mathbf{X} = [X_1, \ldots, X_p]$ and is a random vector, from which n observations are available. We represent the $n \times p$ data matrix as $\mathbb{X}$. Throughout this chapter, we assume that the vector $\mathbf{X}$ has an elliptically contoured density with mean $\boldsymbol{\mu}_x$ and non-singular $p \times p$ covariance matrix $\boldsymbol{\Sigma}_x$, i.e. the probability density function is of the form $f_{\mathbf{X}}(\mathbf{x}) = g\left((\mathbf{x} - \boldsymbol{\mu}_x)^T {\boldsymbol{\Sigma}_x}^{-1} (\mathbf{x} - \boldsymbol{\mu}_x)\right)$, in which g is a non-negative integrable function. The correlation matrix and partial correlation matrix are defined as $\boldsymbol{\Gamma} = \mathbf{D}_{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \boldsymbol{\Sigma} \mathbf{D}_{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$ and $\boldsymbol{\Omega} = \mathbf{D}_{\boldsymbol{\Sigma}^{-1}}^{-\frac{1}{2}} \boldsymbol{\Sigma}^{-1} \mathbf{D}_{\boldsymbol{\Sigma}^{-1}}^{-\frac{1}{2}}$, respectively, where for a matrix $\mathbf{A}$, $\mathbf{D}_{\mathbf{A}}$ represents the diagonal matrix that is obtained by zeroing out all but diagonal entries of $\mathbf{A}$.

The $p \times p$ sample covariance matrix $\mathbf{S}$ for data $\mathbb{X}$ is defined as:

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_{(i)} - \overline{\mathbf{X}})^T (\mathbf{X}_{(i)} - \overline{\mathbf{X}}), \tag{2.1}$$

where $\mathbf{X}_{(i)}$ is the ith row of data matrix $\mathbb{X}$, and $\overline{\mathbf{X}}$ is the vector average of all n rows of $\mathbb{X}$. The $p \times p$ sample correlation and sample partial correlation matrices are then defined as, $\mathbf{R} = \mathbf{D}_{\mathbf{S}}^{-\frac{1}{2}} \mathbf{S} \mathbf{D}_{\mathbf{S}}^{-\frac{1}{2}}$ and $\mathbf{P} = \mathbf{D}_{\mathbf{R}^\dagger}^{-\frac{1}{2}} \mathbf{R}^\dagger \mathbf{D}_{\mathbf{R}^\dagger}^{-\frac{1}{2}}$, respectively, where $\mathbf{R}^\dagger$ is the Moore-Penrose pseudo-inverse of $\mathbf{R}$.

Our theory for local hub screening is based on the U-scores representation of the correlation and partial correlation matrices. It can be shown that there exist a $(n-1) \times p$ matrix $\mathbb{U}_{\mathbf{R}}$ with unit norm columns, such that the following representation holds (*Hero and Rajaratnam*, 2012):

$$\mathbf{R} = \mathbb{U}_{\mathbf{R}}^T \mathbb{U}_{\mathbf{R}}. \tag{2.2}$$

Based on Lemma 1 of the hub screening work (*Hero and Rajaratnam*, 2012) we have:

$$\mathbf{R}^\dagger = \mathbb{U}_{\mathbf{R}}^T (\mathbb{U}_{\mathbf{R}} \mathbb{U}_{\mathbf{R}}^T)^{-2} \mathbb{U}_{\mathbf{R}}^T. \tag{2.3}$$

Hence by defining $\mathbb{U}_{\mathbf{P}} = (\mathbb{U}_{\mathbf{R}} \mathbb{U}_{\mathbf{R}}^T)^{-1} \mathbb{U}_{\mathbf{R}} \mathbf{D}^{-\frac{1}{2}}_{\mathbb{U}_{\mathbf{R}}^T (\mathbb{U}_{\mathbf{R}} \mathbb{U}_{\mathbf{R}}^T)^{-2} \mathbb{U}_{\mathbf{R}}}$ we have the following representation of the sample partial correlation matrix:

$$\mathbf{P} = \mathbb{U}_{\mathbf{P}}^T \mathbb{U}_{\mathbf{P}}, \tag{2.4}$$

where $\mathbb{U}_{\mathbf{P}}$ is a $(n-1) \times p$ matrix with unit-norm columns.

The following will be necessary for Sec. 2.3. We denote the $(n-2)$-dimensional unit sphere in $\mathbb{R}^{n-1}$ and its surface area by S_{n-2} and a_n, respectively. Assume that $\mathbf{U}, \mathbf{V}$ are two independent and uniformly distributed random vectors on S_{n-2}. For

a threshold $\rho \in [0,1]$, let $r = \sqrt{2(1-\rho)}$. P_0 is then defined as the probability that either $\|\mathbf{U}-\mathbf{V}\|_2 \leq r$ or $\|\mathbf{U}+\mathbf{V}\|_2 \leq r$. P_0 can be computed using the formula for the area of spherical caps on S_{n-2}:

$$P_0 = I_{1-\rho^2}(\frac{n-2}{2}, \frac{1}{2}), \tag{2.5}$$

where $I_x(a,b)$ is the regularized incomplete beta function with parameters a and b. For arbitrary joint density $f_{\mathbf{U}_1,\ldots,\mathbf{U}_p}(\mathbf{u}_1,\ldots,\mathbf{u}_p)$ defined on the Cartesian product $S_{n-2}^p = S_{n-2} \times \cdots \times S_{n-2}$, define

$$\overline{f_{\mathbf{U}_i,\mathbf{U}_{*-i}}}(\mathbf{u},\mathbf{v}) = \frac{1}{p-1}\sum_{j\neq i,j=1}^{p}\frac{1}{2}(f_{\mathbf{U}_i,\mathbf{U}_j}(\mathbf{u},\mathbf{v}) + f_{\mathbf{U}_i,\mathbf{U}_j}(\mathbf{u},-\mathbf{v})). \tag{2.6}$$

Let k represent an upper bound on the number of non-zero entries in any row of covariance matrix $\mathbf{\Sigma}_x$. Define the dependency coefficient $\Delta_{i,p,n,k}$ as

$$\Delta_{i,p,n,k} = \max_{j\neq i}\left\|(f_{\mathbf{U}_i,\mathbf{U}_j|\mathbf{U}_{A_k(i,j)}} - f_{\mathbf{U}_i,\mathbf{U}_j})/f_{\mathbf{U}_i,\mathbf{U}_j}\right\|_\infty, \tag{2.7}$$

in which $A_k(i,j)$ is defined as the complement of the union of the sets of indices of the k nearest neighbors of nodes i and j in the correlation graph associated with $\mathbf{\Sigma}_x$.

The function J of the joint density $f_{\mathbf{U},\mathbf{V}}(\mathbf{u},\mathbf{v})$ is defined as:

$$J(f_{\mathbf{U},\mathbf{V}}) = a_n \int_{S_{n-2}} f_{\mathbf{U},\mathbf{V}}(\mathbf{u},\mathbf{u})d\mathbf{u}. \tag{2.8}$$

There are several intuitive interpretations for $J(f_{\mathbf{U},\mathbf{V}})$ (see for example (*Hero and Rajaratnam*, 2011)). Simple calculations show that when $\mathbf{U}$ and $\mathbf{V}$ are independent and uniform over S_{n-2}, $J(f_{\mathbf{U},\mathbf{V}}) = 1$. As we will see later, $J(\overline{f_{\mathbf{U}_i,\mathbf{U}_{*-i}}})$ will play an important role in the expressions for average vertex degrees.

Finally, the total variation distance between the probability distributions of two

integer valued random variables M and N is defined as

$$d_{TV}(M,N) = \sup_{A\subset\mathbb{Z}} |\mathbb{P}(M\in A) - \mathbb{P}(N\in A)|, \tag{2.9}$$

where $\mathbb{Z}$ is the set of integer numbers.

2.3 Local Hub Screening

2.3.1 Asymptotic hub degree distribution

We define the generic matrix notation $\mathbf{\Phi} = [\Phi_{ij}]_{i,j=1}^{p}$ to denote either the sample correlation matrix $\mathbf{R}$ or the sample partial correlation matrix $\mathbf{P}$. Correspondingly, we define $\mathbb{U} = [\mathbf{U}_1, \cdots, \mathbf{U}_p]$ as the generic notation for the U-score representation of matrix $\mathbf{\Phi}$, i.e.:

$$\mathbf{\Phi} = \mathbb{U}^T\mathbb{U}. \tag{2.10}$$

For $\rho \in [0,1]$, we define the *(partial) correlation graph* $\mathcal{G}_\rho(\mathbf{\Phi})$ as follows. The vertices of $\mathcal{G}_\rho(\mathbf{\Phi})$ are $v_1, \cdots, v_p$ which correspond to $\mathbf{U}_1, \cdots, \mathbf{U}_p$ respectively. For $1 \le i, j \le p$, v_i and v_j are connected in $\mathcal{G}_\rho(\mathbf{\Phi})$ if the magnitude of the sample (partial) correlation coefficient between X_i and X_j is at least ρ, i.e. $|\Phi_{ij}| = |\mathbf{U}_i^T\mathbf{U}_j| \ge \rho$. Denote the degree of v_i by d_i. For a positive integer δ, a vertex of $\mathcal{G}_\rho(\mathbf{\Phi})$ is called a hub if $d_i \ge \delta$.

The following theorem shows that under certain conditions, the degree d_i of vertex v_i is approximately a Poisson random variable. This theorem also provides an approximate expression for the mean of d_i.

Theorem II.1. *Let* $\mathbb{U} = [\mathbf{U}_1, \cdots, \mathbf{U}_p]$ *be a* $(n-1)\times p$ *random matrix with* $\mathbf{U}_i \in S_{n-2}$ *where* $n \ge 3$ *is a fixed integer. Assume that the joint density of any subset of* $\mathbf{U}_i$*'s is*

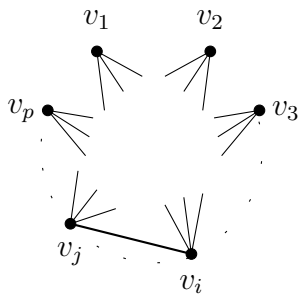

Figure 2.1: Local hub screening thresholds the sample correlation or partial correlation matrix, denoted by the matrix $\mathbf{\Phi}$ in (5.4) to find variables X_i that are highly correlated with other variables. This is equivalent to finding hubs in a graph $\mathcal{G}_\rho(\mathbf{\Phi})$ with p vertices $v_1, \cdots, v_p$. For $1 \leq i, j \leq p$, v_i is connected to v_j in $\mathcal{G}_\rho(\mathbf{\Phi})$ if $|\Phi_{ij}| \geq \rho$.

bounded and differentiable. Then:

$$|\mathbb{E}[d_i] - \Lambda_{i,p,n,\rho}| \leq 2(p-1)P_0 a_n \sqrt{2(1-\rho)}\dot{M}_1, \tag{2.11}$$

in which

$$\Lambda_{i,p,n,\rho} = (p-1)P_0 J(\overline{f_{\mathbf{U}_i,\mathbf{U}_{*-i}}}), \tag{2.12}$$

and

$$\dot{M}_1 = \max_{j \neq i} \sup_{\mathbf{u},\mathbf{v} \in S_{n-2}} \|\nabla_{\mathbf{v}} f_{\mathbf{U}_j|\mathbf{U}_i}(\mathbf{v}|\mathbf{u})\|_2. \tag{2.13}$$

Furthermore, let N_i be a Poisson random variable with rate $\mathbb{E}[d_i]$. Then:

$$d_{TV}(d_i, N_i) \leq (p-1)k^2 P_0^2 a_n^2((M_1^i)^2 + M_2^i) + (p-1)P_0 a_n \Delta_{i,p,n,k}, \tag{2.14}$$

where

$$M_1^i = \max_{j \neq i} \sup_{\mathbf{u},\mathbf{v} \in S_{n-2}} \|f_{\mathbf{U}_j|\mathbf{U}_i}(\mathbf{v}|\mathbf{u})\|_2, \tag{2.15}$$

and

$$M_2^i = \max_{j,k\neq i} \sup_{\mathbf{u},\mathbf{v},\mathbf{w}\in S_{n-2}} \|f_{\mathbf{U}_j,\mathbf{U}_k|\mathbf{U}_i}(\mathbf{v},\mathbf{w}|\mathbf{u})\|_2. \tag{2.16}$$

Proof. For $\mathbf{U} \in S_{n-2}$, let $A(r,\mathbf{U})$ be the union of two anti-polar caps in S_{n-2} of radius $r = \sqrt{2(1-\rho)}$ centered at $\mathbf{U}$ and $-\mathbf{U}$. Moreover let ϕ_{ij} be the indicator of $\mathbf{U}_j \in A(r,\mathbf{U}_i)$, i.e., the event that the magnitude sample (partial) correlation between the ith and jth variable exceeds ρ. For each $1 \leq i \leq p$, we have the following representation for the vertex degree d_i:

$$d_i = \sum_{j\neq i, j=1}^{p} \phi_{ij}. \tag{2.17}$$

Therefore,

$$\mathbb{E}[d_i] = \sum_{j\neq i, j=1}^{p} \mathbb{E}[\phi_{ij}]. \tag{2.18}$$

We have

$$\mathbb{E}[\phi_{ij}] = \int_{S_{n-2}} d\mathbf{u} \int_{A(r,\mathbf{u})} d\mathbf{v} f_{\mathbf{U}_i,\mathbf{U}_j}(\mathbf{u},\mathbf{v}). \tag{2.19}$$

Hence, using mean value theorem we have:

$$|\mathbb{E}[\phi_{ij}] - P_0\left(J\left(\frac{1}{2}(f_{\mathbf{U}_i,\mathbf{U}_j}(\mathbf{u},\mathbf{v}) + f_{\mathbf{U}_i,\mathbf{U}_j}(\mathbf{u},-\mathbf{v}))\right)\right)| \leq 2P_0 a_n \sqrt{2(1-\rho)}\dot{M}_1. \tag{2.20}$$

Summing over j will then conclude the relation (2.11).

Next we prove the relation (2.14). We use Chen-Stein method (*Arratia et al.*, 1990). Define the index set $B(i,j) = \{(l,m) : l \in \mathcal{N}_k(i), m \in \mathcal{N}_k(j)\}$, where $\mathcal{N}_k(i)$ is the set of indices of the k-nearest neighbors of $\mathbf{U}_i$. Note that $|B(i,j)| \leq k^2$. Using

Theorem 1 of (*Arratia et al.*, 1990), we have:

$$2 \max_A |\mathbb{P}(d_i \in A) - \mathbb{P}(N_i \in A)| \leq b_1 + b_2 + b_3, \tag{2.21}$$

where

$$b_1 = \sum_{j \neq i, j=1}^{p} \sum_{(l,m) \in B(i,j)} \mathbb{E}[\phi_{ij}]\mathbb{E}[\phi_{lm}], \tag{2.22}$$

$$b_2 = \sum_{j \neq i, j=1}^{p} \sum_{(l,m) \in B(i,j)} \mathbb{E}[\phi_{ij}\phi_{lm}], \tag{2.23}$$

and, for $p_{ij} = E[\phi_{ij}]$,

$$b_3 = \sum_{j \neq i, j=1}^{p} \mathbb{E}\left[\mathbb{E}[\phi_{ij} - p_{ij} | \{\phi_{lm} : (l,m) \notin B(i,j) \cup \{(i,j)\}\}]\right]. \tag{2.24}$$

Note that we have

$$\mathbb{E}[\phi_{ij}] = \int_{S_{n-2}} d\mathbf{u} \int_{A(r,\mathbf{u})} d\mathbf{v} f_{\mathbf{U}_i, \mathbf{U}_j}(\mathbf{u}, \mathbf{v}) \leq P_0 a_n M_1^i, \tag{2.25}$$

and

$$\begin{aligned} \mathbb{E}\left[\phi_{ij}\phi_{il}\right] &= \int_{S_{n-2}} d\mathbf{u} \int_{A(r,\mathbf{u})} d\mathbf{v}_1 \int_{A(r,\mathbf{u})} d\mathbf{v}_2 f_{\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_l}(\mathbf{u}, \mathbf{v}_1, \mathbf{v}_2) & (2.26) \\ &\leq P_0^2 a_n^2 M_2^i. & (2.27) \end{aligned}$$

Applying the bound (2.25) to the summand of b_1 we obtain

$$b_1 \leq (p-1)k^2 P_0^2 a_n^2 (M_1^i)^2. \tag{2.28}$$

Likewise, the bound (2.27) applied to b_2 gives

$$b_2 \leq (p-1)k^2 P_0^2 a_n^2 M_2^i. \tag{2.29}$$

Furthermore, with $A_k(i,j) = \mathcal{N}_k(i) \cup \mathcal{N}_k(j) - \{i,j\}$ we have

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}[\phi_{ij} - p_{ij} | \{\phi_{lm} : (l,m) \notin B(i,j) \cup \{(i,j)\}\}]\right] = \mathbb{E}\left[\mathbb{E}[\phi_{ij} - p_{ij} | \mathbf{U}_{A_k(i,j)}]\right] \\
&= \int\limits_{S_{n-2}^{|A_k(i,j)|}} d\mathbf{u}_{A_k(i,j)} \int\limits_{S_{n-2}} d\mathbf{u}_i \int\limits_{A(r,\mathbf{u}_i)} d\mathbf{u}_j \\
&\left(\frac{f_{\mathbf{U}_i,\mathbf{U}_j|\mathbf{U}_{A_k(i,j)}}(\mathbf{u}_i,\mathbf{u}_j|\mathbf{u}_{A_k(i,j)}) - f_{\mathbf{U}_i,\mathbf{U}_j}(\mathbf{u}_i,\mathbf{u}_j)}{f_{\mathbf{U}_i,\mathbf{U}_j}(\mathbf{u}_i,\mathbf{u}_j)} \right) f_{\mathbf{U}_i,\mathbf{U}_j}(\mathbf{u}_i,\mathbf{u}_j) f_{\mathbf{U}_{A_k(i,j)}}(\mathbf{u}_{A_k(i,j)}) \\
&\quad \leq P_0 a_n \Delta_{i,p,n,k}.
\end{aligned} \tag{2.30}$$

This yields

$$b_3 \leq (p-1) P_0 a_n \Delta_{i,p,n,k}. \tag{2.31}$$

Hence, combining bounds (2.28), (2.29), (2.31) along with the inequality (2.21) gives the bound (2.14). □

Comparing the bounds in equations (2.11) and (2.14) with those of Proposition 1 in (*Hero and Rajaratnam*, 2012) shows that the rates of convergence of (2.11) and (2.14) in Theorem II.1 converge to 0, p times faster.

When the rows of data matrix $\mathbb{X}$ are independent (i.e. when the samples are independent) and $\mathbf{\Sigma}_x$ is diagonal, U-scores $\mathbf{U}_1, \cdots, \mathbf{U}_p$ are uniform on S_{n-2}. Moreover if the random variables $\mathbf{X}_1, \cdots, \mathbf{X}_p$ are independent, then $\mathbf{U}_1, \cdots, \mathbf{U}_p$ are independent. Under this independence assumption $J(\overline{f_{\mathbf{U}_i,\mathbf{U}_{*-i}}}) = 1$. In this case Poisson limits do not depend on the possibly unknown underlying marginal distribution of the U-scores. Using similar arguments as in hub screening (*Hero and Rajaratnam*, 2012) it can be

shown that if $\mathbf{\Sigma}_x$ is block sparse of degree k, we have:

$$J(\overline{f_{\mathbf{U}_i,\mathbf{U}_{*-i}}}) = 1 + O(k/p) \tag{2.32}$$

and

$$\Delta_{i,p,n,k} = \begin{cases} 0, & \mathbf{\Phi} = \mathbf{R} \\ O\left(k/p\right), & \mathbf{\Phi} = \mathbf{P} \end{cases} \tag{2.33}$$

Later we will see that under the assumption of block sparsity of $\mathbf{\Sigma}_x$, we can use the Poisson limit introduced in Theorem II.1 to assign p-values to vertices of $\mathcal{G}_\rho(\mathbf{\Phi})$ for being hubs.

2.3.2 A numerical example

To illustrate the accuracy of the expressions given in Theorem II.1 for the mean and the distribution of the degree of a specific vertex in a correlation graph, we performed a simple numerical simulation. We generated $n = 100$ independent samples of $p = 5000$ independent and identically distributed (i.i.d) standard normal random variables, constructed the U-scores, constructed the correlation graph $\mathcal{G}_\rho$ and applied the local hub screening.

The value of the correlation threshold was set to $\rho = 0.32$. Figure 2.2 shows the normalized histogram of the degree of the vertex v_{1000} in the graph $\mathcal{G}_\rho$. The histogram was obtained by performing $N = 10^4$ simulations. The solid red line shows a Poisson distribution with rate $\Lambda_{1000,5000,100,0.32}$ given in (2.12). Since both the samples and the variables are independent, $J(\overline{f_{\mathbf{U}_{1000},\mathbf{U}_{*-1000}}}) = 1$ and $\Lambda_{1000,5000,100,0.32} = (p-1)P_0 = (5000-1) \times (1.1722 \times 10^{-3}) = 5.8599$. In the figure the dashed green line corresponds to a Poisson distribution with rate equal to 5.8704, the empirical mean degree.

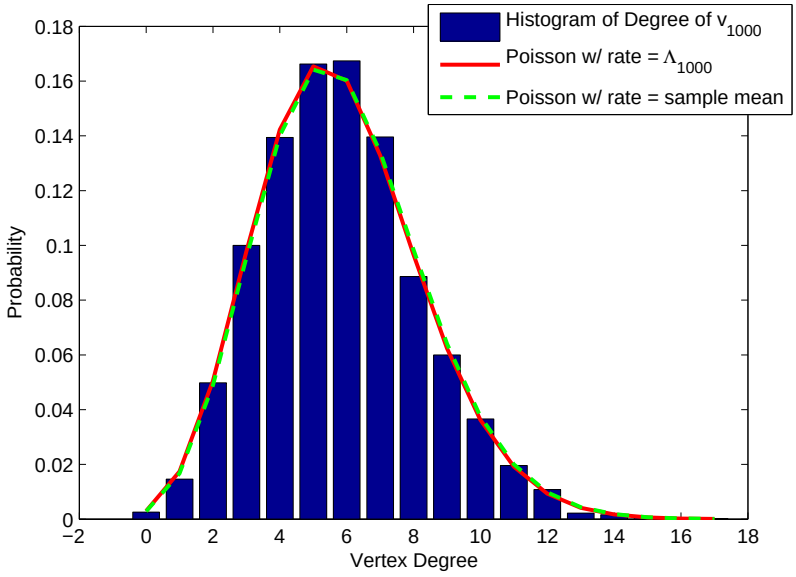

Figure 2.2: A numerical example which confirms the validity of expressions in Theorem II.1. Here $n = 100, p = 5000$ and $\rho = 0.32$.

2.4 Application

2.4.1 Assigning p-values to hubs

Under the null hypothesis of block sparse covariance matrix $\mathbf{\Sigma}_x$ the result of Theorem II.1 along with the approximations (2.32) and (2.33) can be used to assign p-values to the observed degrees $\{d_i\}_{i=1}^p$ of nodes $v_1, \cdots, v_p$. The procedure for assigning p-values is as follows:

1. Choose an initial threshold ρ^*.
2. Select a value $\delta \in \{1, \cdots, \max_{1\leq i\leq p} d_i\}$, where d_i's are the vertex degrees in $\mathcal{G}_{\rho^*}(\mathbf{\Phi})$.
3. For each $1 \leq i \leq p$ let $\rho_\delta(i)$ be the δ-th largest element of $\{|\Phi_{ij}|, j \neq i, 1 \ \leq j \leq p\}$.
4. Approximate the p-value corresponding to vertex v_i as

$$pv_\delta(i) = 1 - F_{\Lambda_{i,p,n,\rho_\delta(i)}}(\delta - 1), \tag{2.34}$$

in which $F_\Lambda(\delta-1)$ is the cumulative distribution function of a Poisson random variable with rate Λ computed at $\delta-1$, i.e. $F_\Lambda(\delta-1)=e^{-\Lambda}\sum_{l=0}^{\delta-1}\Lambda^l/l!$.

The above procedure is similar to the procedures introduced in correlation and predictive correlation screening (*Hero and Rajaratnam*, 2011, 2012; *Firouzi et al.*, 2013) with the difference that here the local Poisson rates $\Lambda_{i,p,n,\rho}$ are used to approximate the p-values whereas correlation screening methods use the global rates for the avergae number of hubs in the (partial) correlation graphs. The advantage of using procedure above comes from the fact that the error bounds for the convergence of the local Poisson rates are at least p times faster than the error bounds for the convergence of the global rates (*Hero and Rajaratnam*, 2011, 2012; *Firouzi et al.*, 2013). This leads to a larger convergence region in terms of p, n and ρ for the rates introduced in Theorem II.1. Therefore, local hub screening applies to a wider range of operating conditions.

2.4.2 Phase transition threshold

The average degree of vertex v_i in $\mathcal{G}_\rho(\mathbf{\Phi})$ exhibits a phase transition as a function of the correlation threshold ρ (see Fig. 2.3). For a given n there is a *critical threshold* ρ_c such that as $\rho \downarrow \rho_c$ the average degree of vertex v_i in the graph $\mathcal{G}_\rho(\mathbf{\Phi})$ is small and increases very slowly. As ρ continues to decrease to values below ρ_c, the average degree of vertex v_i increases rapidly. The rapidity of the phase transition depends on the value of n. For large values of n the phase transition is more evident. We define the critical threshold ρ_c to be the point where $d\mathbb{E}[d_i]/d\rho=-(p-1)$. An approximate value for the critical threshold can be obtained using the approximation (2.11):

$$\rho_c \approx \sqrt{1-c_n}, \tag{2.35}$$

where $c_n = (2J(\overline{f_{\mathbf{U}_i,\mathbf{U}_{*-i}}}))^{-2/(n-4)}$. The value of ρ_c depends on p only through the quantity $J(\overline{f_{\mathbf{U}_i,\mathbf{U}_{*-i}}})$. Therefore, in the cases where the approximation (2.32) is valid the value of the critical threshold does not depend on p nor does it depend on the distribution of the data.

Generally the value of the initial threshold ρ^* will be application dependent, reflecting the minimal correlation that is scientifically significant. In cases where a minimal threshold is not specified by the experimenter, the critical phase transition threshold ρ_c can be used as ρ^*. This ensures that the full range of statistically significant hub correlations is covered in the local hub screening process.

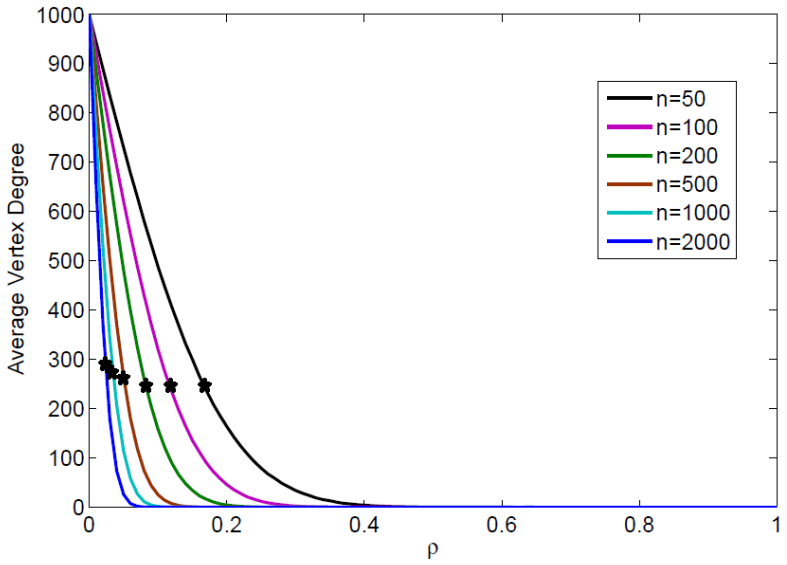

Figure 2.3: Average vertex degree as a function of correlation threshold ρ. The average is obtained for a specific vertex by performing 10^4 experiments. The plots correspond to $n = 2000, 1000, 500, 200, 100, 50$ from left to right, respectively. The samples are draws of $p = 1000$ i.i.d. standard normal random variables. As we see there is a phase transition in the mean vertex degree as a function of ρ. The phase transition becomes sharper as n grows. The critical phase transition threshold ρ_c obtained from (2.35) is shown on the plots using black stars. The values for the critical threshold can be found in Table 2.1

n	2000	1000	500	200	100	50
ρ_c	0.0263	0.0373	0.0528	0.0840	0.1197	0.1723

Table 2.1: The value of critical threshold ρ_c obtained from formula (2.35) for different values of n. The predicted ρ_c approximates the phase transition thresholds in Fig. 2.3.

2.4.3 Application to Connectomics

We illustrate the proposed procedure on a fMRI dataset to assign p-values to different seeds in human brain connectome for being a hub. Studies show that detection of hubs plays a key role in the field of connectomics and can provide insights into the structure of human brain. (*Bullmore and Sporns*, 2009; *He and Evans*, 2010).

In this experiment, the dataset consists of 30 human subjects from which 17 are diagnosed with attention deficit hyperactivity disorder (ADHD). For each subject a number of n samples (which varies between 78 to 340 for different subjects), are used to construct the sample correlation matrix between the resting state blood-oxygen-level dependent (BOLD) signals of $p = 1166$ seeds in the brain.

We applied the procedure described in Sec. 2.4.1 to assign p-values to vertices of the correlation graphs constructed by thresholding the correlation matrices corresponding to each subject. Figure 2.4 shows the *waterfall plots* of p-values corresponding to the 30 different subjects. For a fixed δ, the waterfall plot corresponding to each subject is obtained by linearly interpolating the pairs $\{(\rho_\delta(i), \log\log(1 - pv_\delta(i))^{-1})\}_{i=1}^{p}$ which are ordered based on the absolute values of their first components (i.e., the quantities $|\rho_\delta(i)|$). The initial threshold is chosen to be $\rho^* = 0.86$ which is well beyond the critical thresholds for different subjects. Note that since the number of samples n is different for each subject, the statistical significance obtained by (2.34) using a specific value of $\rho_\delta(i)$ is different for each subject. For this reason the waterfall plots for different subjects do not intersect. The results are shown for $\delta = 1, 2, 3, 4$. We can see that as δ becomes larger there are less discoveries since more seeds fails

to pass the degree threshold. Also, despite the fact that there are fewer healthy subjects (13 out of 30), the healthy subjects tend to be more persistent in appearing in waterfall plots for larger values of δ.

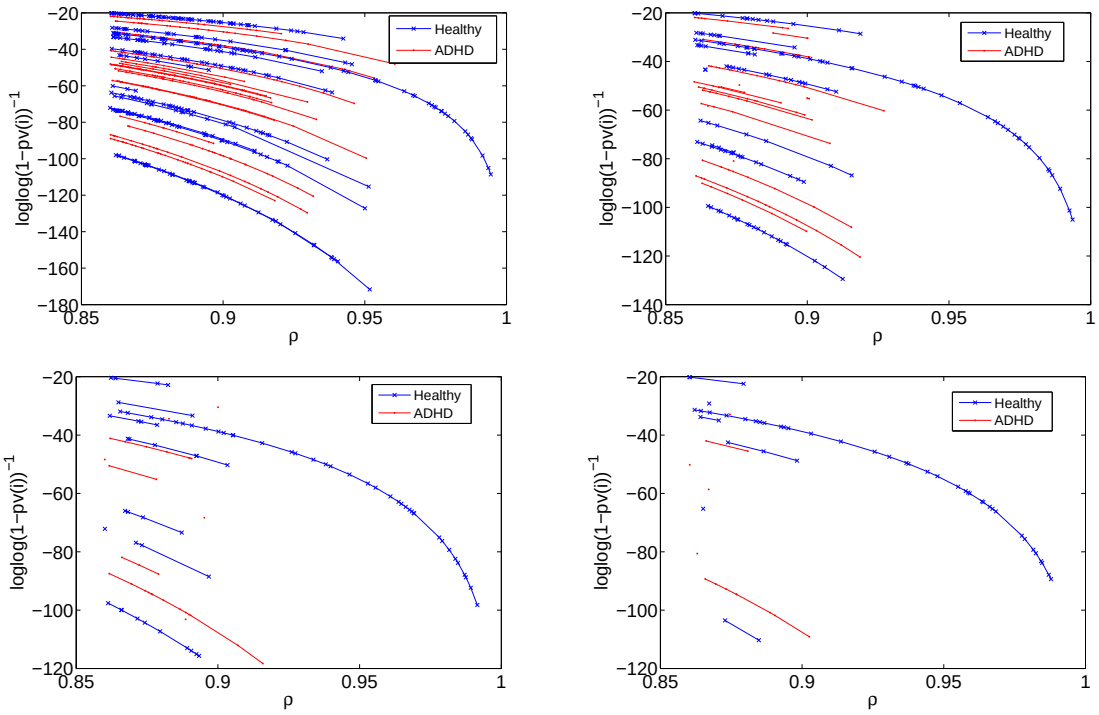

Figure 2.4: Waterfall plots of p-values for a fMRI dataset plotted in terms of $\log\log(1-pv_\delta(i))^{-1}$. The seeds plotted correspond to vertices with degree at least δ in the correlation graph with initial threshold $\rho^* = 0.86$. Upper left, upper right, lower left, and lower right plots correspond to $\delta = 1, 2, 3,$ and 4, respectively.

2.5 Conclusion

We introduced local hub screening for detecting hubs in correlation and partial correlation graphs. Local hub screening assigns p-values to vertices under the null hypothesis of that the covariance matrix is block sparse. The procedure for assigning p-values is justified by a Poisson limit theory of distribution of vertex degrees. We presented a numerical example to confirm the accuracy of our theoretical predictions.

CHAPTER III

Spectral correlation hub screening of multivariate time series

3.1 Introduction

This chapter discusses correlation analysis of stationary multivariate Gaussian time series in the spectral or Fourier domain. While bearing similarities to the local hub screening method of Chapter II, the goal here is to identify the hubs in the correlation network corresponding to the time series, i.e., those times series variables that are highly correlated with a specified number of other time series variables. We show that the Fourier components of the time series at different frequencies are asymptotically statistically independent. This property permits independent correlation analysis at each frequency, alleviating computational and statistical challenges of high-dimensional time series. To detect correlation hubs at each frequency, an existing correlation screening method is extended to the complex-valued variables to accommodate complex-valued Fourier components. We characterize the number of hub discoveries at specified correlation and degree thresholds in the regime of increasing dimension and fixed sample size. The theory specifies appropriate thresholds to apply to sample correlation matrices to detect hubs and also allows statistical significance to be attributed to hub discoveries. Numerical results illustrate the accuracy

of the theory and the usefulness of the proposed spectral framework.

Correlation analysis of multivariate time series is important in many applications such as wireless sensor networks, computer networks, neuroimaging, and finance (*Vuran et al.*, 2004; *Paffenroth et al.*, 2013; *Friston et al.*, 2011; *Zhang et al.*, 2003; *Tsay*, 2005). This chapter focuses on the problem of detecting *hubs* in the time series, variables that have a high degree of interaction with other variables as measured by correlation or partial correlation. Detection of hubs can lead to reduced computational and/or sampling costs. For example in wireless sensor networks, the identification of hub nodes can be useful for reducing power usage and adding or removing sensors from the network (*Stanley et al.*, 2012; *Li et al.*, 2008). Hub detection can also give new insights about underlying structure in the dataset. In neuroimaging for instance, studies have consistently shown the existence of highly connected hubs in brain graphs (connectomes) (*Bullmore and Sporns*, 2009). In finance, a hub might indicate a vulnerable financial instrument or a sector whose collapse could have a major effect on the market (*Hero and Rajaratnam*, 2012).

Correlation analysis becomes challenging for multivariate time series when the dimension p of the time series, i.e. the number of scalar time series, and the number of time samples N are large (*Zhang et al.*, 2003). A naive approach is to treat the time series as a set of independent samples of a p-dimensional random vector and estimate the associated covariance or correlation matrix, but this approach completely ignores temporal correlations as it only considers dependences at the same time instant and not between different time instants. The work in (*Chen et al.*, 2013) accounts for temporal correlations by quantifying their effect on convergence rates in covariance and precision matrix estimation; however, only correlations at the same time instant are estimated. A more general approach is to consider all correlations between any two time instants of any two series within a window of $n \leq N$ consecutive samples, where the previous case corresponds to $n = 1$. However, in general this would entail the

estimation of an $np \times np$ correlation matrix from a reduced sample of size $m = N/n$, which can be computationally costly as well as statistically unstable.

In this chapter, we propose *spectral* correlation analysis to overcome the issues discussed above. As before, the time series are divided into m temporal segments of n consecutive samples, but instead of estimating temporal correlations directly, the method performs analysis on the Discrete Fourier Transforms (DFT) of the time series. We prove in Theorem III.1 that for stationary, jointly Gaussian time series under the mild condition of absolute summability of the auto- and cross-correlation functions, different Fourier components (frequencies) become asymptotically independent of each other as the DFT length n increases. This property of stationary Gaussian processes allows us to focus on the $p \times p$ correlations at each frequency separately without having to consider correlations between different frequencies. Moreover, spectral analysis isolates correlations at specific frequencies or timescales, potentially leading to greater insight. To make aggregate inferences based on all frequencies, straightforward procedures for multiple inference can be used as described in Section 3.5.

The spectral approach reduces the detection of hub time series to the independent detection of hubs at each frequency. However, in exchange for achieving spectral resolution, the sample size is reduced by the factor n, from N to $m = N/n$. To confidently detect hubs in this high-dimensional, low-sample regime (large p, small m), as well as to accommodate complex-valued DFTs, we develop a method that we call *complex-valued (partial) correlation screening*. This is a generalization of the correlation and partial correlation screening method of (*Hero and Rajaratnam*, 2011, 2012; *Firouzi et al.*, 2013) to complex-valued random variables. For each frequency, the method computes the sample (partial) correlation matrix of the DFT components of the p time series. Highly correlated variables (hubs) are then identified by thresholding the sample correlation matrix at a level ρ and screening for rows (or columns) with a specified number δ of non-zero entries.

We characterize the behavior of complex-valued correlation screening in the high-dimensional regime of large p and fixed sample size m. Specifically, Theorem III.5 and Corollary III.6 give asymptotic expressions in the limit $p \to \infty$ for the mean number of hubs detected at thresholds ρ, δ and the probability of discovering at least one such hub. Bounds on the rates of convergence are also provided. These results show that the number of hub discoveries undergoes a phase transition as ρ decreases from 1, from almost no discoveries to the maximum number, p. An expression (3.33) for the critical threshold $\rho_{c,\delta}$ is derived to guide the selection of ρ under different settings of p, m, and δ. Furthermore, given a null hypothesis that the population correlation matrix is sufficiently sparse, the expressions in Corollary III.6 become independent of the underlying probability distribution and can thus be easily evaluated. This allows the statistical significance of a hub discovery to be quantified, specifically in the form of a p-value under the null hypothesis. We note that our results on complex-valued correlation screening apply more generally than to spectral correlation analysis and thus may be of independent interest.

The remainder of the chapter is organized as follows. Section 3.3 presents notation and definitions for multivariate time series and establishes the asymptotic independence of spectral components. Section 3.4 describes complex-valued correlation screening and characterizes its properties in terms of numbers of hub discoveries and phase transitions. Section 3.5 discusses the application of complex-valued correlation screening to the spectra of multivariate time series. Finally, Sec. 3.6 illustrates the applicability of the proposed framework through simulation analysis.

3.2 Preliminaries and notation

A triplet $(\Omega, \mathcal{F}, \mathbb{P})$ represents a probability space with sample space Ω, σ-algebra of events $\mathcal{F}$, and probability measure $\mathbb{P}$. For an event $A \in \mathcal{F}$, $\mathbb{P}(A)$ represents the probability of A. Scalar random variables and their realizations are denoted with

upper case and lower case letters, respectively. Random vectors and their realizations are denoted with bold upper case and bold lower case letters. The expectation operator is denoted as $\mathbb{E}$. For a random variable X, the cumulative probability distribution (cdf) of X is defined as $F_X(x) = \mathbb{P}(X \leq x)$. For an absolutely continuous cdf $F_X(.)$ the probability density function (pdf) is defined as $f_X(x) = dF_X(x)/dx$. The cdf and pdf are defined similarly for random vectors. Moreover, we follow the definitions in (*Durrett*, 2010) for conditional probabilities, conditional expectations and conditional densities.

For a complex number $z = a + b\sqrt{-1} \in \mathbb{C}$, $\Re(z) = a$ and $\Im(z) = b$ represent the real and imaginary parts of z, respectively. A complex-valued random variable is composed of two real-valued random variables as its real and imaginary parts. A complex-valued Gaussian variable has real and imaginary parts that are Gaussian. A complex-valued (Gaussian) random vector is a vector whose entries are complex-valued (Gaussian) random variables. The covariance of a p-dimensional complex-valued random vector $\mathbf{Y}$ and a q-dimensional complex-valued random vector $\mathbf{Z}$ is a $p \times q$ matrix defined as

$$\text{cov}(\mathbf{Y}, \mathbf{Z}) = \mathbb{E}\left[(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])^H\right],$$

where ^{H} denotes the Hermitian transpose. We write $\text{cov}(\mathbf{Y})$ for $\text{cov}(\mathbf{Y}, \mathbf{Y})$ and $\text{var}(Y) = \text{cov}(Y, Y)$ for the variance of a scalar random variable Y. The correlation coefficient between random variables Y and Z is defined as

$$\text{cor}(Y, Z) = \frac{\text{cov}(Y, Z)}{\sqrt{\text{var}(Y)\text{var}(Z)}}.$$

Matrices are also denoted by bold upper case letters. In most cases the distinction between matrices and random vectors will be clear from the context. For a matrix $\mathbf{A}$ we represent the (i, j)th entry of $\mathbf{A}$ by a_{ij}. Also $\mathbf{D_A}$ represents the diagonal matrix

that is obtained by zeroing out all but the diagonal entries of $\mathbf{A}$.

3.3 Spectral representation of multivariate time series

3.3.1 Definitions

Let $\mathbf{X}(k) = [X^{(1)}(k), X^{(2)}(k), \cdots X^{(p)}(k)]$, $k \in \mathbb{Z}$, be a multivariate time series with time index k. We assume that the time series $X^{(1)}, X^{(2)}, \cdots X^{(p)}$ are second-order stationary random processes, i.e.:

$$\mathbb{E}[X^{(i)}(k)] = \mathbb{E}[X^{(i)}(k+\Delta)] \tag{3.1}$$

and

$$\text{cov}[X^{(i)}(k), X^{(j)}(l)] = \text{cov}[X^{(i)}(k+\Delta), X^{(j)}(l+\Delta)] \tag{3.2}$$

for any integer time shift Δ.

For $1 \leq i \leq p$, let $\mathbf{X}^{(i)} = [X^{(i)}(k), \cdots, X^{(i)}(k+n-1)]$ denote any vector of n consecutive samples of time series $X^{(i)}$. The n-point Discrete Fourier Transform (DFT) of $\mathbf{X}^{(i)}$ is denoted by $\mathbf{Y}^{(i)} = [Y^{(i)}(0), \cdots, Y^{(i)}(n-1)]$ and defined by

$$\mathbf{Y}^{(i)} = \mathbf{W}\mathbf{X}^{(i)},\ 1 \leq i \leq p$$

in which $\mathbf{W}$ is the DFT matrix:

$$\mathbf{W} = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & \omega & \cdots & \omega^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{n-1} & \cdots & \omega^{(n-1)^2} \end{bmatrix},$$

where $\omega = e^{-2\pi\sqrt{-1}/n}$.

We denote the $n \times n$ population covariance matrix of $\mathbf{X}^{(i)}$ as $\mathbf{C}^{(i,i)} = [c_{kl}^{(i,i)}]_{1\leq k,l\leq n}$ and the $n \times n$ population cross covariance matrix between $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(j)}$ as $\mathbf{C}^{(i,j)} = [c_{kl}^{(i,j)}]_{1\leq k,l\leq n}$ for $i \neq j$. The translation invariance properties (3.1) and (3.2) imply that $\mathbf{C}^{(i,i)}$ and $\mathbf{C}^{(i,j)}$ are Toeplitz matrices. Therefore $c_{kl}^{(i,i)}$ and $c_{kl}^{(i,j)}$ depend on k and l only through the quantity $k-l$. Representing the (k,l)th entry of a Toeplitz matrix $\mathbf{T}$ by $t(k-l)$, we write

$$c_{kl}^{(i,i)} = c^{(i,i)}(k-l) \text{ and } c_{kl}^{(i,j)} = c^{(i,j)}(k-l),$$

where $k-l$ takes values from $1-n$ to $n-1$. In addition, $\mathbf{C}^{(i,i)}$ is symmetric.

3.3.2 Asymptotic independence of spectral components

The following theorem states that for stationary time series, DFT components at different spectral indices (i.e. frequencies) are asymptotically uncorrelated under the condition that the auto-covariance and cross-covariance functions are absolutely summable. This theorem follows directly from the spectral theory of large Toeplitz matrices, see, for example, (*Grenander and Szegő*, 1958) and (*Gray*, 2006). However, for the benefit of the reader we give a self contained proof of the theorem.

Theorem III.1. *Assume* $\lim_{n\to\infty} \sum_{t=0}^{n-1} |c^{(i,j)}(t)| = M^{(i,j)} < \infty$ *for all* $1 \leq i,j \leq p$. *Define* $\mathrm{err}^{(i,j)}(n) = M^{(i,j)} - \sum_{m'=0}^{n-1} |c^{(i,j)}(m')|$ *and* $\mathrm{avg}^{(i,j)}(n) = \frac{1}{n}\sum_{m'=0}^{n-1} \mathrm{err}^{(i,j)}(m')$. *Then for* $k \neq l$, *we have:*

$$\mathrm{cor}\left(Y^{(i)}(k), Y^{(j)}(l)\right) = O(\max\{1/n, \mathrm{avg}^{(i,j)}(n)\}).$$

In other words $Y^{(i)}(k)$ *and* $Y^{(j)}(l)$ *are asymptotically uncorrelated as* $n \to \infty$.

Proof. Without loss of generality we assume that the time series have zero mean (i.e.

$\mathbb{E}[X^{(i)}(k)] = 0, 1 \leq i \leq p, 0 \leq k \leq n-1)$. We first establish a representation of $\mathbb{E}[Z^{(i)}(k)Z^{(j)}(l)^*]$ for general linear functionals:

$$Z^{(i)}(k) = \sum_{m'=0}^{n-1} g_k(m')X^{(i)}(m'),$$

in which $g_k(.)$ is an arbitrary complex sequence for $0 \leq k \leq n-1$. We have:

$$\begin{aligned}
& \mathbb{E}[Z^{(i)}(k)Z^{(j)}(l)^*] \\
= \; & \mathbb{E}\left[\left(\sum_{m'=0}^{n-1} g_k(m')X^{(i)}(m')\right)\left(\sum_{n'=0}^{n-1} g_l(n')X^{(j)}(n')\right)^*\right] \\
= \; & \sum_{m'=0}^{n-1} g_k(m') \sum_{n'=0}^{n-1} g_l(n')^* \mathbb{E}[X^{(i)}(m')X^{(j)}(n')^*] \\
= \; & \sum_{m'=0}^{n-1} g_k(m') \sum_{n'=0}^{n-1} g_l(n')^* c_{m'n'}^{(i,j)} \qquad (3.3)
\end{aligned}$$

Now for a Toeplitz matrix $\mathbf{T}$, define the circulant matrix $\mathbf{D_T}$ as:

$$\mathbf{D_T} = \begin{bmatrix}
t(0) & t(-1)+t(n-1) & \cdots & t(1-n)+t(1) \\
t(1)+t(1-n) & t(0) & \cdots & t(2-n)+t(2) \\
\vdots & \vdots & \ddots & \vdots \\
t(n-2)+t(-2) & t(n-3)+t(-3) & \cdots & t(-1)+t(n-1) \\
t(n-1)+t(-1) & t(n-2)+t(-2) & \cdots & t(0)
\end{bmatrix}$$

We can write:

$$\mathbf{C}^{(i,j)} = \mathbf{D}_{\mathbf{C}^{(i,j)}} + \mathbf{E}^{(i,j)}$$

for some Toeplitz matrix $\mathbf{E}^{(i,j)}$. Thus $c^{(i,j)}(m'-n') = d^{(i,j)}(m'-n') + e^{(i,j)}(m'-n')$ where $d^{(i,j)}(m'-n')$ and $e^{(i,j)}(m'-n')$ are the (m',n') entries of $\mathbf{D}_{\mathbf{C}^{(i,j)}}$ and $\mathbf{E}^{(i,j)}$,

respectively. Therefore, (3.3) can be written as:

$$\sum_{m'=0}^{n-1} g_k(m') \sum_{n'=0}^{n-1} g_l(n')^* d^{(i,j)}(m'-n') + \sum_{m'=0}^{n-1}\sum_{n'=0}^{n-1} g_k(m') g_l(n')^* e^{(i,j)}(m'-n')$$

The first term can be written as:

$$\sum_{m'=0}^{n-1} g_k(m') \left(g_l^* \circledast d^{(i,j)}\right)(m') = \sum_{m'=0}^{n-1} g_k(m') v_l^{(i,j)}(m')$$

where we have recognized $v_l^{(i,j)}(m') = g_l^* \circledast d^{(i,j)}$ as the circular convolution of $g_l^*(.)$ and $d^{(i,j)}(.)$ (*Oppenheim et al.*, 1989). Let $G_k(.)$ and $D^{(i,j)}(.)$ be the the DFT of $g_k(.)$ and $d^{(i,j)}(.)$, respectively. By Plancherel's theorem (*Conway*, 1990) we have:

$$\begin{aligned}
\sum_{m'=0}^{n-1} g_k(m') v_l^{(i,j)}(m') &= \sum_{m'=0}^{n-1} g_k(m') \left(v_l^{(i,j)}(m')^*\right)^* \\
&= \sum_{m'=0}^{n-1} G_k(m') \left(G_l(m') D^{(i,j)}(-m')^*\right)^* \\
&= \sum_{m'=0}^{n-1} G_k(m') G_l(m')^* D^{(i,j)}(-m'). \qquad (3.4)
\end{aligned}$$

Now let $g_k(m') = \omega^{km'}/\sqrt{n}$ for $0 \le k, m' \le n-1$. For this choice of $g_k(.)$ we have $G_k(m') = 0$ for all $m' \neq n-k$ and $G_k(n-k) = 1$. Hence for $k \neq l$ the quantity (3.4) becomes 0. Therefore using the representation $\mathbf{E}^{(i,j)} = \mathbf{C}^{(i,j)} - \mathbf{D}_{\mathbf{C}^{(i,j)}}$ we have:

$$\begin{aligned}
|\mathrm{cov}\left(Y^{(i)}(k), Y^{(j)}(l)\right)| &= |\mathbb{E}[Y^{(i)}(k) Y^{(j)}(l)^*]| \\
&= |\sum_{m'=0}^{n-1}\sum_{n'=0}^{n-1} g_k(m') g_l(n')^* e^{(i,j)}(m'-n')| \\
&\le \frac{1}{n}\sum_{m'=0}^{n-1}\sum_{n'=0}^{n-1} |e^{(i,j)}(m'-n')| \\
&= \frac{2}{n}\sum_{m'=0}^{n-1} m' |c^{(i,j)}(m')|, \qquad (3.5)
\end{aligned}$$

in which the last equation is due to the fact that $|c^{(i,j)}(-m')| = |c^{(i,j)}(m')|$.

Now using (3.4) and (3.5) we obtain expressions for var $\left(Y^{(i)}(k)\right)$ and var $\left(Y^{(j)}(l)\right)$. Letting $j = i$ and $l = k$ in (3.4) and (3.5) gives:

$$\begin{aligned}
& \text{var}\left(Y^{(i)}(k)\right) = \text{cov}\left(Y^{(i)}(k), Y^{(i)}(k)\right) \\
= & \sum_{m'=0}^{n-1} G_k(m')G_k(m')^* D^{(i,i)}(-m') + \sum_{m'=0}^{n-1}\sum_{n'=0}^{n-1} g_k(m')g_k(n')^* e^{(i,i)}(m'-n') \\
= & \; n.\frac{1}{\sqrt{n}}.\frac{1}{\sqrt{n}} D^{(i,i)}(k) + \sum_{m'=0}^{n-1}\sum_{n'=0}^{n-1} g_k(m')g_k(n')^* e^{(i,i)}(m'-n') \\
= & \; D^{(i,i)}(k) + \sum_{m'=0}^{n-1}\sum_{n'=0}^{n-1} g_k(m')g_k(n')^* e^{(i,i)}(m'-n'), \qquad (3.6)
\end{aligned}$$

in which the magnitude of the summation term is bounded as:

$$\begin{aligned}
& |\sum_{m'=0}^{n-1}\sum_{n'=0}^{n-1} g_k(m')g_k(n')^* e^{(i,i)}(m'-n')| \\
\leq & \; \frac{1}{n}\sum_{m'=0}^{n-1}\sum_{n'=0}^{n-1} |e^{(i,i)}(m'-n')| \\
= & \; \frac{2}{n}\sum_{m'=0}^{n-1} m'|c^{(i,i)}(m')|. \qquad (3.7)
\end{aligned}$$

Similarly:

$$\text{var}\left(Y^{(j)}(l)\right) = D^{(j,j)}(l) + \sum_{m'=0}^{n-1}\sum_{n'=0}^{n-1} g_l(m')g_l(n')^* e^{(j,j)}(m'-n'), \qquad (3.8)$$

in which

$$\begin{aligned}
& |\sum_{m'=0}^{n-1}\sum_{n'=0}^{n-1} g_l(m')g_l(n')^* e^{(j,j)}(m'-n')| \\
\leq & \; \frac{2}{n}\sum_{m'=0}^{n-1} m'|c^{(j,j)}(m')|. \qquad (3.9)
\end{aligned}$$

To complete the proof the following lemma is needed.

Lemma III.2. *If* $\{a_{m'}\}_{m'=0}^{\infty}$ *is a sequence of non-negative numbers such that* $\sum_{m'=0}^{\infty} a_{m'} = M < \infty$. *Define* $\mathrm{err}(n) = M - \sum_{m'=0}^{n-1} a_{m'}$ *and* $\mathrm{avg}(n) = \frac{1}{n}\sum_{m'=0}^{n-1} \mathrm{err}(m')$. *Then* $|\frac{1}{n}\sum_{m'=0}^{n-1} m' a_{m'}| \leq M/n + \mathrm{err}(n) + \mathrm{avg}(n)$.

Proof. Let $S_0 = 0$ and for $n \geq 1$ define $S_n = \sum_{m'=0}^{n-1} a_{m'}$. We have:

$$\sum_{m'=0}^{n-1} m a_{m'} = (n-1)S_n - (S_0 + S_1 + \ldots + S_{n-1}).$$

Therefore:

$$\frac{1}{n}\sum_{m'=0}^{n-1} m' a_{m'} = \frac{n-1}{n} S_{n-1} - \frac{1}{n}\sum_{m'=0}^{n-1} S_{m'}.$$

Since $M - M/n - \mathrm{err}(n) \leq \frac{n-1}{n} S_{n-1} \leq M$ and $M - \mathrm{avg}(n) \leq \frac{1}{n}\sum_{m'=0}^{n-1} S_{m'} \leq M$, using the triangle inequality the result follows. □

Now let $a_{m'} = |c^{(i,j)}(m')|$. By assumption $\lim_{n\to\infty} \sum_{m'=0}^{n-1} a_{m'} = M^{(i,j)} < \infty$. Therefore, Lemma III.2 along with (3.5) concludes:

$$\mathrm{cov}\left(Y^{(i)}(k), Y^{(j)}(l)\right) = O(\max\{1/n, \mathrm{err}^{(\mathrm{i,j})}(n), \mathrm{avg}^{(\mathrm{i,j})}(n)\}). \tag{3.10}$$

$\mathrm{err}^{(i,j)}(n)$ is a decreasing decreasing function of n. Therefore $\mathrm{avg}^{(i,j)}(n) \geq \mathrm{err}^{(i,j)}(n)$, for $n \geq 1$. Hence:

$$\mathrm{cov}\left(Y^{(i)}(k), Y^{(j)}(l)\right) = O(\max\{1/n, \mathrm{avg}^{(i,j)}(n)\}).$$

Similarly using Lemma III.2 along with (3.6), (3.7), (3.8) and (3.9) we obtain:

$$|\mathrm{var}\left(Y^{(i)}(k)\right) - D^{(i,i)}(k)| = O(\max\{1/n, \mathrm{avg}^{(i,i)}(n)\}), \tag{3.11}$$

and

$$|\text{var}\left(Y^{(j)}(l)\right) - D^{(j,j)}(l)| = O(\max\{1/n, \text{avg}^{(j,j)}(n)\}). \tag{3.12}$$

Using the definition

$$\text{cor}\left(Y^{(i)}(k), Y^{(j)}(l)\right) = \frac{\text{cov}\left(Y^{(i)}(k), Y^{(j)}(l)\right)}{\sqrt{\text{var}\left(Y^{(i)}(k)\right)}\sqrt{\text{var}\left(Y^{(j)}(l)\right)}},$$

and the fact that as $n \to \infty$, $D^{(i,i)}(k)$ and $D^{(j,j)}(l)$ converge to constants $\mathbf{C}^{(i,i)}(k)$ and $\mathbf{C}^{(j,j)}(l)$, respectively, equations (3.10), (3.11) and (3.12) conclude:

$$\text{cor}\left(Y^{(i)}(k), Y^{(j)}(l)\right) = O(\max\{1/n, \text{avg}^{(i,j)}(n)\}).$$

□

As an example we apply Theorem III.1 to a scalar auto-regressive (AR) process $X(k)$ specified by

$$X(k) = \sum_{l=1}^{L} \varphi_l X(k-l) + \varepsilon(k),$$

in which φ_l are real-valued coefficients and $\varepsilon(.)$ is a stationary process with no temporal correlation. The auto-covariance function of an AR process can be written as (*Hamilton*, 1994):

$$c(t) = \sum_{l=1}^{L} \alpha_l r_l^{|t|},$$

in which $r_1, \ldots, r_l$ are the roots of the polynomial $\beta(x) = x^L - \sum_{l=1}^{L} \varphi_l x^{L-l}$. It is known that for a stationary AR process, $|r_l| < 1$ for all $1 \leq l \leq L$ (*Hamilton*, 1994).

Therefore, using the definition of err(.) we have:

$$\begin{aligned}\mathrm{err}(n) &= \sum_{t=n}^{\infty} |c(t)| = \sum_{t=n}^{\infty} |\sum_{l=1}^{L} \alpha_l r_l^t| \leq \sum_{l=1}^{L} |\alpha_l| \sum_{t=n}^{\infty} |r_l|^t \\ &= \sum_{l=1}^{L} |\alpha_l| \frac{|r_l|^n}{1-|r_l|} \leq C\zeta^n,\end{aligned}$$

in which $C = \sum_{l=1}^{L} |\alpha_l|/(1-|r_l|)$ and $\zeta = \max_{1\leq l\leq L} |r_l| < 1$. Hence:

$$\mathrm{avg}(n) = \frac{1}{n}\sum_{m'=0}^{n-1} \mathrm{err}(m') \leq \frac{1}{n}\sum_{m'=0}^{n-1} C\zeta^{m'} \leq \frac{C}{n(1-\zeta)}.$$

Therefore, Theorem III.1 concludes:

$$\mathrm{cor}\,(Y(k), Y(l)) = O(1/n),\ k \neq l,$$

where $Y(.)$ represents the n-point DFT of the AR process $X(.)$.

In the sequel, we assume that the time series $\mathbf{X}$ is multivariate Gaussian, i.e., $X^{(1)}, \ldots, X^{(p)}$ are jointly Gaussian processes. It follows that the DFT components $Y^{(i)}(k)$ are jointly (complex) Gaussian as linear functionals of $\mathbf{X}$. Theorem III.1 then immediately implies asymptotic independence of DFT components through a well-known property of jointly Gaussian random variables.

Corollary III.3. *Assume that the time series* $\mathbf{X}$ *is multivariate Gaussian. Under the absolute summability conditions in Theorem III.1, the DFT components* $Y^{(i)}(k)$ *and* $Y^{(j)}(l)$ *are asymptotically independent for* $k \neq l$ *and* $n \to \infty$.

Corollary III.3 implies that for large n, correlation analysis of the time series $\mathbf{X}$ can be done independently on each frequency in the spectral domain. This reduces the problem of screening for hub time series to screening for hub variables among the p DFT components at a given frequency. A procedure for the latter problem and a corresponding theory are described next.

3.4 Complex-valued correlation hub screening

This section discusses complex-valued correlation hub screening, a generalization of real-valued correlation screening in (*Hero and Rajaratnam*, 2011, 2012), for identifying highly correlated components of a complex-valued random vector from its sample values. The method is applied to multivariate time series in Section 3.5 to discover correlation hubs among the spectral components at each frequency. Sections 3.4.1 and 3.4.2 describe the underlying statistical model and the screening procedure. Sections 3.4.3 and 3.4.4 provide background on the U-score representation of correlation matrices and associated definitions and properties. Section 3.4.5 contains the main theoretical result characterizing the number of hub discoveries in the high-dimensional regime, while Section 3.4.6 elaborates on the phenomenon of phase transitions in the number of discoveries.

3.4.1 Statistical model

We use the generic notation $\mathbf{Z} = [Z_1, Z_2, \cdots, Z_p]^T$ in this section to refer to a complex-valued random vector. The mean of $\mathbf{Z}$ is denoted as $\boldsymbol{\mu}$ and its $p \times p$ non-singular covariance matrix is denoted as $\boldsymbol{\Sigma}$. We assume that the vector $\mathbf{Z}$ follows a complex elliptically contoured distribution with pdf $f_{\mathbf{Z}}(\mathbf{z}) = g\left((\mathbf{z}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})\right)$, in which $g : \mathbb{R}^{\geq 0} \to \mathbb{R}^{>0}$ is an integrable and strictly decreasing function (*Micheas et al.*, 2006). This assumption generalizes the Gaussian assumption made in Section 3.3 as the Gaussian distribution is one example of an elliptically contoured distribution.

In correlation hub screening, the quantities of interest are the correlation matrix and partial correlation matrix associated with $\mathbf{Z}$. These are defined as $\boldsymbol{\Gamma} = \mathbf{D}_{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \boldsymbol{\Sigma} \mathbf{D}_{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$ and $\boldsymbol{\Omega} = \mathbf{D}_{\boldsymbol{\Sigma}^{-1}}^{-\frac{1}{2}} \boldsymbol{\Sigma}^{-1} \mathbf{D}_{\boldsymbol{\Sigma}^{-1}}^{-\frac{1}{2}}$, respectively. Note that $\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}$ are normalized matrices with unit diagonals.

3.4.2 Screening procedure

The goal of correlation hub screening is to identify highly correlated components of the random vector $\mathbf{Z}$ from its sample realizations. Assume that m samples $\mathbf{z}_1, \ldots, \mathbf{z}_m \in \mathbb{R}^p$ of $\mathbf{Z}$ are available. To simplify the development of the theory, the samples are assumed to be independent and identically distributed (i.i.d.) although the theory also applies to dependent samples.

We compute sample correlation and partial correlation matrices from the samples $\mathbf{z}_1, \ldots, \mathbf{z}_m$ as surrogates for the unknown population correlation matrices $\mathbf{\Gamma}$ and $\mathbf{\Omega}$ in Section 3.4.1. First define the $p \times p$ sample covariance matrix $\mathbf{S}$ as $\mathbf{S} = \frac{1}{m-1} \sum_{i=1}^{m} (\mathbf{z}_i - \overline{\mathbf{z}})(\mathbf{z}_i - \overline{\mathbf{z}})^H$, where $\overline{\mathbf{z}}$ is the sample mean, the average of $\mathbf{z}_1, \ldots, \mathbf{z}_m$. The sample correlation and sample partial correlation matrices are then defined as $\mathbf{R} = \mathbf{D}_{\mathbf{S}}^{-\frac{1}{2}} \mathbf{S} \mathbf{D}_{\mathbf{S}}^{-\frac{1}{2}}$ and $\mathbf{P} = \mathbf{D}_{\mathbf{R}^\dagger}^{-\frac{1}{2}} \mathbf{R}^\dagger \mathbf{D}_{\mathbf{R}^\dagger}^{-\frac{1}{2}}$, respectively, where $\mathbf{R}^\dagger$ is the Moore-Penrose pseudo-inverse of $\mathbf{R}$.

Correlation hubs are screened by applying thresholds to the sample (partial) correlation matrix. A variable Z_i is declared a hub screening discovery at degree level $\delta \in \{1, 2, \ldots\}$ and threshold level $\rho \in [0, 1]$ if

$$|\{j : j \neq i, |\psi_{ij}| \geq \rho\}| \geq \delta,$$

where $\mathbf{\Psi} = \mathbf{R}$ for correlation screening and $\mathbf{\Psi} = \mathbf{P}$ for partial correlation screening. We denote by $N_{\delta,\rho} \in \{0, \ldots, p\}$ the total number of hub screening discoveries at levels δ, ρ.

Correlation hub screening can also be interpreted in terms of the *(partial) correlation graph* $\mathcal{G}_\rho(\mathbf{\Psi})$, depicted in Fig. 5.1 and defined as follows. The vertices of $\mathcal{G}_\rho(\mathbf{\Psi})$ are $v_1, \cdots, v_p$ which correspond to $Z_1, \cdots, Z_p$, respectively. For $1 \leq i, j \leq p$, v_i and v_j are connected by an edge in $\mathcal{G}_\rho(\mathbf{\Psi})$ if the magnitude of the sample (partial) correlation coefficient between Z_i and Z_j is at least ρ. A vertex of $\mathcal{G}_\rho(\mathbf{\Psi})$ is called a

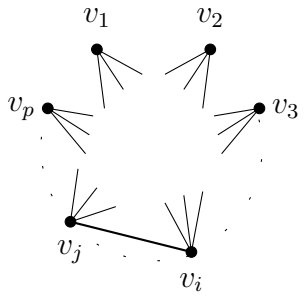

Figure 3.1: Complex-valued (partial) correlation hub screening thresholds the sample correlation or partial correlation matrix, denoted generically by the matrix $\mathbf{\Psi}$, to find variables Z_i that are highly correlated with other variables. This is equivalent to finding hubs in a graph $\mathcal{G}_\rho(\mathbf{\Psi})$ with p vertices $v_1, \cdots, v_p$. For $1 \leq i, j \leq p$, v_i is connected to v_j in $\mathcal{G}_\rho(\mathbf{\Psi})$ if $|\psi_{ij}| \geq \rho$.

δ-hub if its degree, the number of incident edges, is at least δ. Then the number of discoveries $N_{\delta,\rho}$ defined earlier is the number of δ-hubs in the graph $\mathcal{G}_\rho(\mathbf{\Psi})$.

3.4.3 U-score representation of correlation matrices

Our theory for complex-valued correlation screening is based on the U-score representation of the sample correlation and partial correlation matrices. Similarly to the real case (*Hero and Rajaratnam*, 2012), it can be shown that there exists an $(m-1) \times p$ complex-valued matrix $\mathbb{U}_{\mathbf{R}}$ with unit-norm columns $\mathbf{u}_{\mathbf{R}}^{(i)} \in \mathbb{C}^{m-1}$ such that the following representation holds:

$$\mathbf{R} = \mathbb{U}_{\mathbf{R}}^H \mathbb{U}_{\mathbf{R}}. \tag{3.13}$$

Similar to Lemma 1 in (*Hero and Rajaratnam*, 2012) it is straightforward to show that:

$$\mathbf{R}^\dagger = \mathbb{U}_{\mathbf{R}}^H (\mathbb{U}_{\mathbf{R}} \mathbb{U}_{\mathbf{R}}^H)^{-2} \mathbb{U}_{\mathbf{R}}.$$

Hence by defining $\mathbb{U}_{\mathbf{P}} = (\mathbb{U}_{\mathbf{R}}\mathbb{U}_{\mathbf{R}}^H)^{-1}\mathbb{U}_{\mathbf{R}}\mathbf{D}_{\mathbb{U}_{\mathbf{R}}^H(\mathbb{U}_{\mathbf{R}}\mathbb{U}_{\mathbf{R}}^H)^{-2}\mathbb{U}_{\mathbf{R}}}^{-\frac{1}{2}}$ we have the representation:

$$\mathbf{P} = \mathbb{U}_{\mathbf{P}}^H\mathbb{U}_{\mathbf{P}}, \tag{3.14}$$

where the $(m-1) \times p$ matrix $\mathbb{U}_{\mathbf{P}}$ has unit-norm columns $\mathbf{u}_{\mathbf{P}}^{(i)} \in \mathbb{C}^{m-1}$.

3.4.4 Properties of U-scores

The U-score factorizations in (3.13) and (3.14) show that sample (partial) correlation matrices can be represented in terms of unit vectors in $\mathbb{C}^{m-1}$. This subsection presents definitions and properties related to U-scores that will be used in Section 3.4.5.

We denote the unit spheres in $\mathbb{R}^{m-1}$ and $\mathbb{C}^{m-1}$ as S_{m-1} and T_{m-1}, respectively. The surface areas of S_{m-1} and T_{m-1} are denoted as a_{m-1} and b_{m-1} respectively. Define the interleaving function $h : \mathbb{R}^{2m-2} \to \mathbb{C}^{m-1}$ as below:

$$\begin{aligned} &h([x_1, x_2, \cdots, x_{2m-2}]^T) = \\ &[x_1 + x_2\sqrt{-1}, x_3 + x_4\sqrt{-1}, \cdots, x_{2m-3} + x_{2m-2}\sqrt{-1}]^T. \end{aligned}$$

Note that $h(.)$ is a one-to-one and onto function and it maps S_{2m-2} to T_{m-1}.

For a fixed vector $\mathbf{u} \in T_{m-1}$ and a threshold $0 \leq \rho \leq 1$ define the spherical cap in T_{m-1}:

$$A_\rho(\mathbf{u}) = \{\mathbf{y} : \mathbf{y} \in T_{m-1}, |\mathbf{y}^H\mathbf{u}| \geq \rho\}.$$

Also define P_0 as the probability that a random point $\mathbf{Y}$ that is uniformly distributed on T_{m-1} falls into $A_\rho(\mathbf{u})$. Below we give a simple expression for P_0 as a function of ρ and m.

Lemma III.4. *Let* $\mathbf{Y}$ *be an* $(m-1)$*-dimensional complex-valued random vector that*

is uniformly distributed over T_{m-1}. *We have* $P_0 = \mathbb{P}(\mathbf{Y} \in A_\rho(\mathbf{u})) = (1-\rho^2)^{m-2}$.

Proof. Without loss of generality we assume $\mathbf{u} = [1, 0, \cdots, 0]^T$. We have:

$$P_0 = \mathbb{P}(|Y_1| \geq \rho) = \mathbb{P}(\Re(Y_1)^2 + \Im(Y_1)^2 \geq \rho^2).$$

Since $\mathbf{Y}$ is uniform on T_{m-1}, we can write $\mathbf{Y} = \mathbf{X}/\|\mathbf{X}\|_2$, in which $\mathbf{X}$ is complex-valued random vector whose entries are i.i.d. complex-valued Gaussian variables with mean 0 and variance 1. Thus:

$$\begin{aligned} P_0 &= \mathbb{P}\left(\left(\Re(X_1)^2 + \Im(X_1^2)\right)/\|\mathbf{X}\|_2^2 \geq \rho^2\right) \\ &= \mathbb{P}\left((1-\rho^2)\left(\Re(X_1)^2 + \Im(X_1)^2\right) \geq \rho^2 \sum_{k=2}^{m-1} \Re(X_k)^2 + \Im(X_k)^2\right). \end{aligned}$$

Define $V_1 = \Re(X_1)^2 + \Im(X_1)^2$ and $V_2 = \sum_{k=2}^{m-1} \Re(X_k)^2 + \Im(X_k)^2$. V_1 and V_2 are independent and have chi-squared distributions with 2 and $2(m-2)$ degrees of freedom, respectively (*Simon*, 2007). Therefore,

$$\begin{aligned} P_0 &= \int_0^\infty \int_{\rho^2 v_2/(1-\rho^2)}^\infty \chi^2_2(v_1)\chi^2_{2(m-2)}(v_2)dv_1dv_2 \\ &= \int_0^\infty \chi^2_{2(m-2)}(v_2) \int_{\rho^2 v_2/(1-\rho^2)}^\infty \frac{1}{2}e^{-v_1/2}dv_1dv_2 \\ &= \int_0^\infty \frac{1}{2^{m-2}\Gamma(m-2)} v_2^{m-3} e^{-v_2/2} e^{-\frac{\rho^2}{2(1-\rho^2)}v_2} dv_2 \\ &= \frac{1}{\Gamma(m-2)}(1-\rho^2)^{m-2} \int_0^\infty x^{m-3}e^{-x}dx \\ &= \frac{1}{\Gamma(m-2)}(1-\rho^2)^{m-2}\Gamma(m-2) = (1-\rho^2)^{m-2}, \end{aligned}$$

in which we have made a change of variable $x = \frac{v_2}{2(1-\rho^2)}$. □

Under the assumption that the joint pdf of $\mathbf{Z}$ exists, the p columns of the U-

score matrix have joint pdf $f_{\mathbf{U}_1,\ldots,\mathbf{U}_p}(\mathbf{u}_1,\ldots,\mathbf{u}_p)$ on $T_{m-1}^p = \times_{i=1}^p T_{m-1}$. The following $(\delta+1)$-fold average of the joint pdf will play a significant role in Section 3.4.5. This $(\delta+1)$-fold average is defined as:

$$\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*\delta+1}}}(\mathbf{u}_1,\ldots,\mathbf{u}_{\delta+1}) = \frac{1}{(2\pi)^{\delta+1}p\binom{p-1}{\delta}} \times$$
$$\sum_{1\le i_1<\cdots<i_\delta\le p, i_{\delta+1}\notin\{i_1,\cdots,i_\delta\}} \int_0^{2\pi}\int_0^{2\pi}\cdots\int_0^{2\pi}$$
$$f_{\mathbf{U}_{i_1},\ldots,\mathbf{U}_{i_\delta},\mathbf{U}_{i_{\delta+1}}}(e^{\sqrt{-1}\theta_1}\mathbf{u}_1,\ldots,e^{\sqrt{-1}\theta_\delta}\mathbf{u}_\delta, e^{\sqrt{-1}\theta}\mathbf{u}_{\delta+1})\; d\theta_1\cdots d\theta_\delta\; d\theta.$$

Also for a joint pdf $f_{\mathbf{U}_1,\ldots,\mathbf{U}_{\delta+1}}(\mathbf{u}_1,\ldots,\mathbf{u}_{\delta+1})$ on $T_{m-1}^{\delta+1}$ define

$$J(f_{\mathbf{U}_1,\ldots,\mathbf{U}_{\delta+1}}) = a_{2m-2}^{\delta}\int_{S_{2m-2}} f_{\mathbf{U}_1,\ldots,\mathbf{U}_{\delta+1}}(h(\mathbf{u}),\ldots,h(\mathbf{u}))d\mathbf{u}.$$

Note that $J(f_{\mathbf{U}_1,\ldots,\mathbf{U}_{\delta+1}})$ is proportional to the integral of $f_{\mathbf{U}_1,\ldots,\mathbf{U}_{\delta+1}}$ over the manifold $\mathbf{u}_1 = \ldots = \mathbf{u}_{\delta+1}$. The quantity $J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}})$ is key in determining the asymptotic average number of hubs in a complex-valued correlation network. This will be described in more detail in Sec. 3.4.5.

Let $\vec{i} = (i_0, i_1, \ldots, i_\delta)$ be a set of distinct indices, i.e., $1 \le i_0 \le p, 1 \le i_1 < \ldots < i_\delta \le p$ and $i_1,\ldots,i_\delta \ne i_0$. For a U-score matrix $\mathbb{U}$ define the dependency coefficient between the columns $\mathbf{U}_{\vec{i}} = \{\mathbf{U}_{i_0}, \mathbf{U}_{i_1},\ldots,\mathbf{U}_{i_\delta}\}$ and their complementary k-NN (k-nearest neighbor) set $A_k(\vec{i})$ defined in (3.29) and Fig. 3.2 as

$$\Delta_{p,m,k,\delta}(\vec{i}) = \left\| (f_{\mathbf{U}_{\vec{i}}|\mathbf{U}_{A_k(\vec{i})}} - f_{\mathbf{U}_{\vec{i}}})/f_{\mathbf{U}_{\vec{i}}} \right\|_\infty,$$

where $\|\cdot\|_\infty$ denotes the supremum norm. The average of these coefficients is defined

as:

$$\|\Delta_{p,m,k,\delta}\|_1 = \frac{1}{p\binom{p-1}{\delta}} \sum_{i_0=1}^{p} \sum_{\substack{i_1,\ldots,i_\delta \neq i_0 \\ 1 \leq i_1 < \ldots < i_\delta \leq p}} \Delta_{p,m,k,\delta}(\vec{i}). \tag{3.15}$$

3.4.5 Number of hub discoveries in the high-dimensional limit

We now present the main theoretical result on complex-valued correlation screening. The following theorem gives asymptotic expressions for the mean number of δ-hubs and the probability of discovery of at least one δ-hub in the graph $\mathcal{G}_\rho(\mathbf{\Psi})$. It also gives bounds on the rates of convergence to these approximations as the dimension p increases and $\rho \to 1$. We use $\mathbb{U} = [\mathbf{U}_1, \cdots, \mathbf{U}_p]$ as a generic notation for the U-score representation of the sample (partial) correlation matrix. The asymptotic expression for the mean $\mathbb{E}[N_{\delta,\rho}]$ is denoted by Λ and is given by:

$$\Lambda = p\binom{p-1}{\delta} P_0^\delta J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}}). \tag{3.16}$$

Define $\eta_{p,\delta}$ as:

$$\eta_{p,\delta} = p^{1/\delta}(p-1)P_0 = p^{1/\delta}(p-1)(1-\rho^2)^{(m-2)}, \tag{3.17}$$

where the last equation is due to Lemma III.4. The parameter k below represents an upper bound on the true hub degree, i.e. the number of non-zero entries in any row of the population covariance matrix $\mathbf{\Sigma}$. Also let $\varphi(\delta)$ be the function that takes values $\varphi(\delta) = 2$ for $\delta = 1$ and $\varphi(\delta) = 1$ for $\delta > 1$.

Theorem III.5. *Let $\mathbb{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_p]$ be a $(m-1)\times p$ random matrix with $\mathbf{U}_i \in T_{m-1}$ where $m > 2$. Let $\delta \geq 1$ be a fixed integer. Assume the joint pdf of any subset of the*

$\mathbf{U}_i$*'s is bounded and differentiable. Then, with* Λ *defined in (3.16),*

$$|\mathbb{E}[N_{\delta,\rho}] - \Lambda| \leq O\left(\eta_{p,\delta}^{\delta} \max\left\{\eta_{p,\delta} p^{-1/\delta}, (1-\rho)^{1/2}\right\}\right). \tag{3.18}$$

Furthermore, let $N_{\delta,\rho}^*$ *be a Poisson distributed random variable with rate* $\mathbb{E}[N_{\delta,\rho}^*] = \Lambda/\varphi(\delta)$*. If* $(p-1)P_0 \leq 1$*, then*

$$\begin{aligned}
&\left|\mathbb{P}(N_{\delta,\rho} > 0) - \mathbb{P}(N_{\delta,\rho}^* > 0)\right| \leq \\
&\begin{cases} O\left(\eta_{p,\delta}^{\delta} \max\left\{\eta_{p,\delta}^{\delta} (k/p)^{\delta+1}, Q_{p,k,\delta}, \|\Delta_{p,m,k,\delta}\|_1, p^{-1/\delta}, (1-\rho)^{1/2}\right\}\right), & \delta > 1 \\ O\left(\eta_{p,1} \max\left\{\eta_{p,1} (k/p)^2, \|\Delta_{p,m,k,1}\|_1, p^{-1}, (1-\rho)^{1/2}\right\}\right), & \delta = 1 \end{cases},
\end{aligned} \tag{3.19}$$

with $Q_{p,k,\delta} = \eta_{p,\delta}\left(k/p^{1/\delta}\right)^{\delta+1}$ *and* $\|\Delta_{p,m,k,\delta}\|_1$ *defined in* (3.15)*.*

Proof. The proof is similar to the proof of proposition 1 in (*Hero and Rajaratnam*, 2012). First we prove (3.18). Let $\phi_i = I(d_i \geq \delta)$ be the indicator of the event that $d_i \geq \delta$, in which d_i represents the degree of the vertex v_i in the graph $\mathcal{G}_\rho(\mathbf{\Psi})$. We have $N_{\delta,\rho} = \sum_{i=1}^{p} \phi_i$. With ϕ_{ij} being the indicator of the presence of an edge in $\mathcal{G}_\rho(\mathbf{\Psi})$ between vertices v_i and v_j we have the relation:

$$\phi_i \;=\; \sum_{l=\delta}^{p-1} \sum_{\vec{k} \in \check{\mathcal{C}}_i(p-1,l)} \prod_{j=1}^{l} \phi_{ik_j} \prod_{q=l+1}^{p-1} (1 - \phi_{ik_q}) \tag{3.20}$$

where we have defined the index vector $\vec{k} = (k_1, \ldots, k_{p-1})$ and the set

$$\check{\mathcal{C}}_i(p-1, l) =$$

$$\{\vec{k} : k_1 < \ldots < k_l, k_{l+1} < \ldots < k_{p-1}\; k_j \in \{1, \ldots, p\} - \{i\}, k_j \neq k_{j'}\}.$$

The inner summation in (3.20) simply sums over the set of distinct indices not equal

to i that index all $\binom{p-1}{l}$ different types of products of the form: $\prod_{j=1}^{l}\phi_{ik_j}\prod_{q=l+1}^{p-1}(1-\phi_{ik_q})$. Subtracting $\sum_{\vec{k}\in\check{\mathcal{C}}_i(p-1,\delta)}\prod_{j=1}^{\delta}\phi_{ik_j}$ from both sides of (3.20)

$$\begin{aligned}
&\phi_i - \sum_{\vec{k}\in\check{\mathcal{C}}_i(p-1,\delta)}\prod_{j=1}^{\delta}\phi_{ik_j} \\
&\quad = \sum_{l=\delta+1}^{p-1}\sum_{\vec{k}\in\check{\mathcal{C}}_i(p-1,l)}\prod_{j=1}^{l}\phi_{ik_j}\prod_{q=l+1}^{p-1}(1-\phi_{ik_q}) \\
&\qquad + \sum_{\vec{k}\in\check{\mathcal{C}}_i(p-1,l)}\sum_{q=\delta+1}^{p-1}(-1)^{q-\delta} \\
&\qquad\qquad \sum_{k'_{\delta+1}<\ldots<k'_q,\{k'_{\delta+1},\ldots,k'_q\}\subset\{k_{\delta+1},\ldots,k_{p-1}\}}\prod_{j=1}^{l}\phi_{ik_j}\prod_{s=\delta+1}^{q}\phi_{ik'_s} \qquad (3.21)
\end{aligned}$$

in which we have used the expansion

$$\prod_{q=\delta+1}^{p-1}(1-\phi_{ik_q}) = 1+\sum_{q=\delta+1}^{p-1}(-1)^{q-\delta}\sum_{k'_{\delta+1}<\ldots<k'_q,\{k'_{\delta+1},\ldots,k'_q\}\subset\{k_{\delta+1},\ldots,k_{p-1}\}}\prod_{s=\delta+1}^{q}\phi_{ik'_s}.$$

The following simple asymptotic representation will be useful in the sequel. For any $i_1,\ldots,i_k\in\{1,\ldots,p\}$, $i_1\neq\cdots\neq i_k\neq i$, $k\in\{1,\ldots,p-1\}$,

$$\begin{aligned}
\mathbb{E}\left[\prod_{j=1}^{k}\phi_{ii_j}\right] &= \int_{S_{2m-2}}\int_{h^{-1}(A_\rho(\mathbf{v}))}\cdots\int_{h^{-1}(A_\rho(\mathbf{v}))} \\
&\qquad f_{\mathbf{U}_{i_1},\ldots,\mathbf{U}_{i_k},\mathbf{U}_i}(h(\mathbf{v}_1),\cdots,h(\mathbf{v}_k),h(\mathbf{v}))\;d\mathbf{v}_1\cdots d\mathbf{v}_k\;d\mathbf{v} \\
&\leq P_0^k a_{2m-2}^k M_{k|1} \qquad (3.22)
\end{aligned}$$

where $P_0, A_\rho(\mathbf{u})$ and the function $h(.)$ are defined in Sec. 3.4.4. Moreover

$$M_{k|1} = \max_{i_1\neq\cdots\neq i_{k+1}}\left\|f_{\mathbf{U}_{i_1},\ldots,\mathbf{U}_{i_k}|\mathbf{U}_{i_{k+1}}}\right\|_\infty.$$

The following simple generalization of (3.22) to arbitrary product indices ϕ_{ij} will also

be needed

$$\mathbb{E}\left[\prod_{l=1}^{q}\phi_{i_l j_l}\right] \leq P_0^q a_{2m-2}^q M_{|Q|}, \tag{3.23}$$

where Q =unique($\{i_l, j_l\}_{l=1}^q$) is the set of unique indices among the distinct pairs $\{(i_l, j_l)\}_{l=1}^q$ and $M_{|Q|}$ is a bound on the joint pdf of $\mathbf{U}_Q$.

Define the random variable

$$\theta_i = \binom{p-1}{\delta}^{-1} \sum_{\vec{k}\in\check{\mathcal{C}}_i(p-1,\delta)} \prod_{j=1}^{\delta}\phi_{ik_j}.$$

We show below that for sufficiently large p

$$\left|\mathbb{E}[\phi_i] - \binom{p-1}{\delta}\mathbb{E}[\theta_i]\right| \;\leq\; \gamma_{p,\delta}((p-1)P_0)^{\delta+1}, \tag{3.24}$$

where $\gamma_{p,\delta} = \max_{\delta+1\leq l<p}\{a_{2m-2}^l M_{l|1}\}\left(e - \sum_{l=0}^{\delta}\frac{1}{l!}\right)(1+(\delta!)^{-1})$ and $M_{l|1}$ is a least upper bound on any l-dimensional joint pdf of the variables $\{\mathbf{U}_i\}_{j\neq i}^p$ conditioned on $\mathbf{U}_i$.

To show inequality (3.24) take expectations of (3.21) and apply the bound (3.22) to obtain

$$\begin{aligned}
&\left|\mathbb{E}[\phi_i] - \binom{p-1}{\delta}\mathbb{E}[\theta_i]\right| \leq \\
&\left|\sum_{l=\delta+1}^{p-1}\binom{p-1}{l}P_0^l a_{2m-2}^l M_{l|1} + \binom{p-1}{\delta}\sum_{l=1}^{p-1-\delta}\binom{p-1-\delta}{l}P_0^{\delta+l}a_{2m-2}^{\delta+l}M_{\delta+l|1}\right| \\
&\leq A(1+(\delta!)^{-1}),
\end{aligned} \tag{3.25}$$

where

$$A = \sum_{l=\delta+1}^{p-1}\binom{p-1}{l}((p-1)P_0)^l a_{2m-2}^l M_{l|1}.$$

The line (3.25) follows from the identity $\binom{p-1-\delta}{l}\binom{p-1}{\delta} = \binom{p-1}{l+\delta}\binom{l+\delta}{l}$ and a change of

index in the second summation on the previous line. Since $(p-1)P_0 < 1$

$$\begin{aligned} |A| &\leq \max_{\delta+1\leq l<p}\{a_{2m-2}^l M_{l|1}\} \sum_{l=\delta+1}^{p-1} \binom{p-1}{l} ((p-1)P_0)^l \\ &\leq \max_{\delta+1\leq l<p}\{a_{2m-2}^l M_{l|1}\} \left(e - \sum_{l=0}^{\delta} \frac{1}{l!}\right) ((p-1)P_0)^{\delta+1}. \end{aligned}$$

Application of the mean value theorem to the integral representation (3.22) yields

$$\left|\mathbb{E}[\theta_i] - P_0^\delta J(\overline{f_{\mathbf{U}_{*1-i},\ldots,\mathbf{U}_{*\delta-i},\mathbf{U}_i}})\right| \leq \tilde{\gamma}_{p,\delta}((p-1)P_0)^\delta r, \tag{3.26}$$

where

$$\begin{aligned} \overline{f_{\mathbf{U}_{*1-i},\ldots,\mathbf{U}_{*\delta-i},\mathbf{U}_i}}(\mathbf{u}_1, &\quad \ldots, \mathbf{u}_{\delta+1}) = \\ \frac{1}{(2\pi)^\delta \binom{p-1}{\delta}} &\quad \sum_{\substack{1\leq i_1<\cdots<i_\delta\leq p \\ i\notin\{i_1,\cdots,i_\delta\}}} \int_0^{2\pi}\cdots\int_0^{2\pi} \\ f_{\mathbf{U}_{i_1},\ldots,\mathbf{U}_{i_\delta},\mathbf{U}_i}(e^{\sqrt{-1}\theta_1}\mathbf{u}_1, &\quad \ldots, e^{\sqrt{-1}\theta_\delta}\mathbf{u}_\delta, \mathbf{u}_{\delta+1})\; d\theta_1\cdots d\theta_\delta, \end{aligned}$$

$r = \sqrt{2(1-\rho)}$, $\tilde{\gamma}_{p,\delta} = 2a_{2m-2}^{\delta+1}\dot{M}_{\delta+1|1}/\delta!$ and $\dot{M}_{\delta+1|1}$ is a bound on the norm of the gradient

$$\nabla_{\mathbf{u}_{i_1},\ldots,\mathbf{u}_{i_\delta}} \overline{f_{\mathbf{U}_{*1-i},\ldots,\mathbf{U}_{*\delta-i}|\mathbf{U}_i}(\mathbf{u}_{i_1},\ldots,\mathbf{u}_{i_\delta}|\mathbf{u}_i)}.$$

Combining (3.24)-(3.26) and the relation $r = O((1-\rho)^{1/2})$,

$$\begin{aligned} &\left|\mathbb{E}[\phi_i] - \binom{p-1}{\delta} P_0^\delta J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}})\right| \\ \leq\; & O\left(((p-1)P_0)^\delta \max\left\{(p-1)P_0, (1-\rho)^{1/2}\right\}\right). \end{aligned}$$

Summing over i and recalling the definitions (3.16) and (3.17) of Λ and $\eta_{p,\delta}$,

$$\begin{aligned} |\mathbb{E}[N_{\delta,\rho}] - \Lambda| &\leq O\left(p((p-1)P_0)^\delta \max\left\{(p-1)P_0, (1-\rho)^{1/2}\right\}\right) \\ &= O\left(\eta_{p,\delta}^\delta \max\left\{\eta_{p,\delta} p^{-1/\delta}, (1-\rho)^{1/2}\right\}\right). \end{aligned}$$

This establishes the bound (3.18).

Next we prove the bound (3.19) by using the Chen-Stein method (*Arratia et al.*, 1990). Define:

$$\tilde{N}_{\delta,\rho} = \frac{1}{\varphi(\delta)} \sum_{i_0=1}^{p} \sum_{1 \leq i_1 < \ldots < i_\delta \leq p} \prod_{j=1}^{\delta} \phi_{i_0 i_j}, \tag{3.27}$$

Where the second sum is over the indices $1 \leq i_1 < \ldots < i_\delta \leq p$ such that $i_j \neq i_0, 1 \leq j \leq \delta$. For $\vec{i} \stackrel{\text{def}}{=} (i_0, i_1, \ldots, i_\delta)$ define the index set $B_{\vec{i}} = B_{i_0,i_1,\ldots,i_\delta} = \{(j_0, j_1, \ldots, j_\delta) : j_l \in \mathcal{N}_k(i_l) \cup \{i_l\}, l = 0, \ldots, \delta\} \cap \mathcal{C}^<$ where $\mathcal{C}^< = \{(j_0, \ldots, j_\delta) : 1 \leq j_0 \leq p, 1 \leq j_1 < \cdots < j_\delta \leq p, j_l \neq j_0, 1 \leq l \leq \delta\}$. These index the distinct sets of points $\mathbf{U}_{\vec{i}} = \{\mathbf{U}_{i_0}, \mathbf{U}_{i_1}, \ldots, \mathbf{U}_{i_\delta}\}$ and their respective k-NN's. Note that $|B_{\vec{i}}| \leq k^{\delta+1}$. Identifying $\tilde{N}_{\delta,\rho} = \sum_{\vec{i} \in \mathcal{C}^<} \prod_{l=1}^{\delta} \phi_{i_0 i_l}$ and $N^*_{\delta,\rho}$ a Poisson distributed random variable with rate $\mathbb{E}[\tilde{N}_{\delta,\rho}]$, the Chen-Stein bound (*Arratia et al.*, 1990, Theorem 1) is

$$2 \max_A |\mathbb{P}(\tilde{N}_{\delta,\rho} \in A) - \mathbb{P}(N^*_{\delta,\rho} \in A)| \leq b_1 + b_2 + b_3, \tag{3.28}$$

where

$$b_1 = \sum_{\vec{i} \in \mathcal{C}^<} \sum_{\vec{j} \in B_{\vec{i}}} \mathbb{E}\left[\prod_{l=1}^{\delta} \phi_{i_0 i_l}\right] \mathbb{E}\left[\prod_{q=1}^{\delta} \phi_{j_0 j_q}\right],$$

$$b_2 = \sum_{\vec{i} \in \mathcal{C}^<} \sum_{\vec{j} \in B_{\vec{i}-\{\vec{i}\}}} \mathbb{E}\left[\prod_{l=1}^{\delta} \phi_{i_0 i_l} \prod_{q=1}^{\delta} \phi_{j_0 j_q}\right],$$

and, for $p_{\vec{i}} = \mathbb{E}[\prod_{l=1}^{\delta} \phi_{i_0 i_l}]$,

$$b_3 = \sum_{\vec{i} \in \mathcal{C}^<} \mathbb{E}\left[\mathbb{E}\left[\prod_{l=1}^{\delta} \phi_{i_0 i_l} - p_{\vec{i}} \middle| \phi_{\vec{j}} : \vec{j} \notin B_{\vec{i}}\right]\right].$$

Over the range of indices in the sum of b_1 $\mathbb{E}[\prod_{l=1}^{\delta} \phi_{i i_l}]$ is of order $O(P_0^{\delta})$, by (3.23), and therefore

$$b_1 \leq O\left(p^{\delta+1} k^{\delta+1} P_0^{2\delta}\right) = O\left(\eta_{p,\delta}^{2\delta} (k/p)^{\delta+1}\right),$$

which follows from definition (3.17). More care is needed to bound b_2 due to the repetition of characteristic functions ϕ_{ij}. Since $\vec{i} \neq \vec{j}$, $\mathbb{E}[\prod_{l=1}^{\delta} \phi_{i_0 i_l} \prod_{q=1}^{\delta} \phi_{j_0 j_q}]$ is a multiplication of at least $\delta + 1$ different characteristic functions, hence by (3.23),

$$\mathbb{E}[\prod_{l=1}^{\delta} \phi_{i_0 i_l} \prod_{q=1}^{\delta} \phi_{j_0 j_q}] = O\left(P_0^{\delta+1}\right).$$

Therefore, we conclude that

$$b_2 \;\; \leq \;\; O\left(p^{\delta+1} k^{\delta+1} P_0^{\delta+1}\right).$$

Next we bound the term b_3 in (3.28). The set

$$A_k(\vec{i}) = B_{\vec{i}}^c - \{\vec{i}\} \tag{3.29}$$

indexes the complementary k-NN of $\mathbf{U}_{\vec{i}}$ (see Fig. 3.2) so that, using the representation

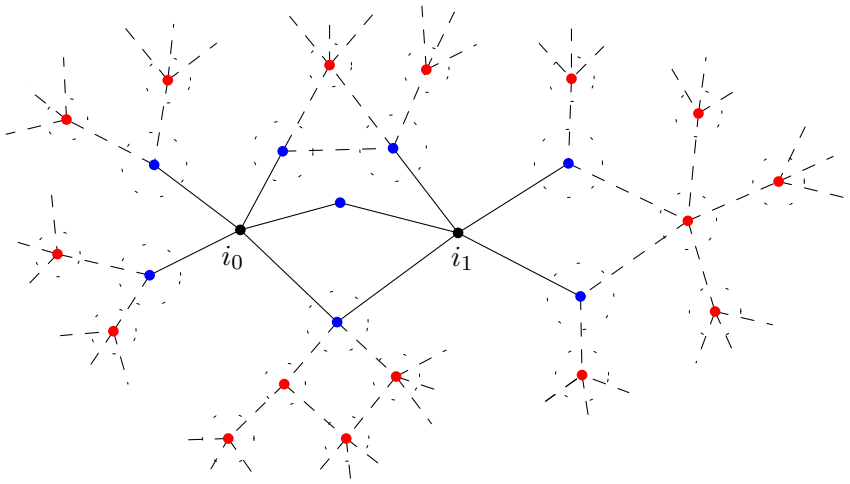

Figure 3.2: The complementary k-NN set $A_k(\vec{i})$ illustrated for $\delta = 1$ and $k = 5$. Here we have $\vec{i} = (i_0, i_1)$. The vertices i_0, i_1 and their k-NNs are depicted in black and blue respectively. The complement of the union of $\{i_0, i_1\}$ and its k-NNs is the complementary k-NN set $A_k(\vec{i})$ and is depicted in red.

(3.23),

$$
\begin{aligned}
b_3 &= \sum_{\vec{i}\in\mathcal{C}^<} \mathbb{E}\left[\mathbb{E}\left[\prod_{l=1}^{\delta}\phi_{i_0 i_l} - p_{\vec{i}}\,\middle|\,\mathbf{U}_{A_k(\vec{i})}\right]\right] \\
&= \sum_{\vec{i}\in\mathcal{C}^<} \int_{S_{2m-2}^{|A_k(\vec{i})|}} d\mathbf{u}_{A_k(\vec{i})} \left(\prod_{l=1}^{\delta}\int_{S_{2m-2}} d\mathbf{u}_{i_0} \int_{A(r,\mathbf{u}_{i_0})} d\mathbf{u}_{i_l}\right) \\
&\quad \left(\frac{f_{\mathbf{U}_{\vec{i}}|\mathbf{U}_{A_k}}(\mathbf{u}_{\vec{i}}|\mathbf{u}_{A_k(\vec{i})}) - f_{\mathbf{U}_{\vec{i}}}(\mathbf{u}_{\vec{i}})}{f_{\mathbf{U}_{\vec{i}}}(\mathbf{u}_{\vec{i}})}\right) f_{\mathbf{U}_{\vec{i}}}(\mathbf{u}_{\vec{i}}) f_{\mathbf{U}_{A_k(\vec{i})}}(\mathbf{u}_{A_k(\vec{i})}) \\
&\leq O\left(p^{\delta+1}P_0^{\delta}\|\Delta_{p,m,k,\delta}\|_1\right) = O\left(\eta_{p,\delta}^{\delta}\|\Delta_{p,m,k,\delta}\|_1\right).
\end{aligned}
$$

Note that by definition of $\tilde{N}_{\delta,\rho}$ we have $\tilde{N}_{\delta,\rho} > 0$ if and only if $N_{\delta,\rho} > 0$. This yields:

$$
\begin{aligned}
&|\mathbb{P}(N_{\delta,\rho} > 0) - (1 - \exp(-\Lambda))| \leq \left|\mathbb{P}(\tilde{N}_{\delta,\rho} > 0) - \mathbb{P}(N_{\delta,\rho} > 0)\right| + \\
&\left|\mathbb{P}(\tilde{N}_{\delta,\rho} > 0) - \left(1 - \exp(-\mathbb{E}[\tilde{N}_{\delta,\rho}])\right)\right| + \left|\exp(-\mathbb{E}[\tilde{N}_{\delta,\rho}]) - \exp(-\Lambda)\right| \\
\leq\ & b_1 + b_2 + b_3 + O\left(\left|\mathbb{E}[\tilde{N}_{\delta,\rho}] - \Lambda\right|\right) \qquad (3.30)
\end{aligned}
$$

Combining the above inequalities on b_1, b_2 and b_3 yields the first three terms in the argument of the "max" on the right side of (3.19).

It remains to bound the term $|\mathbb{E}[\tilde{N}_{\delta,\rho}]-\Lambda|$. Application of the mean value theorem to the multiple integral (3.23) gives

$$\left|\mathbb{E}\left[\prod_{l=1}^{\delta}\phi_{ii_l}\right]-P_0^{\delta}J\left(f_{\mathbf{U}_{i_1},\ldots,\mathbf{U}_{i_\delta},\mathbf{U}_i}\right)\right| \;\leq\; O\left(P_0^{\delta}r\right).$$

Applying relation (3.27) yields

$$\left|\mathbb{E}[\tilde{N}_{\delta,\rho}]-p\binom{p-1}{\delta}P_0^{\delta}J\left(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}}\right)\right| \;\leq\; O\left(p^{\delta+1}P_0^{\delta}r\right)=O\left(\eta_{p,\delta}^{\delta}r\right).$$

Combine this with (3.30) to obtain the bound (3.19). This completes the proof of Theorem III.5. □

An immediate consequence of Theorem III.5 is the following result, similar to Proposition 2 in (*Hero and Rajaratnam*, 2012), which provides asymptotic expressions for the mean number of δ-hubs and the probability of the event $N_{\delta,\rho}>0$ as p goes to ∞ and ρ converges to 1 at a prescribed rate.

Corollary III.6. *Let $\rho_p\in[0,1]$ be a sequence converging to one as $p\to\infty$ such that $\eta_{p,\delta}=p^{1/\delta}(p-1)(1-\rho_p^2)^{(m-2)}\to e_{m,\delta}\in(0,\infty)$. Then*

$$\lim_{p\to\infty}\mathbb{E}[N_{\delta,\rho_p}]=\Lambda_\infty=e_{m,\delta}^{\delta}/\delta!\;\lim_{p\to\infty}J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}}). \tag{3.31}$$

Assume that $k=o(p^{1/\delta})$ and that for the weak dependency coefficient $\|\Delta_{p,m,k,\delta}\|_1$, defined via (3.15)*, we have $\lim_{p\to\infty}\|\Delta_{p,m,k,\delta}\|_1=0$. Then*

$$\mathbb{P}(N_{\delta,\rho_p}>0)\to 1-\exp(-\Lambda_\infty/\varphi(\delta)). \tag{3.32}$$

Corollary III.6 shows that in the limit $p \to \infty$, the number of detected hubs depends on the true population correlations only through the quantity $J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}})$. In some cases $J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}})$ can be evaluated explicitly. Similar to the argument in (*Hero and Rajaratnam*, 2012), it can be shown that if the population covariance matrix $\mathbf{\Sigma}$ is sparse in the sense that its non-zero off-diagonal entries can be arranged into a $k \times k$ submatrix by reordering rows and columns, then

$$J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}}) = 1 + O(k/p).$$

Hence, if $k = o(p)$ as $p \to \infty$, the quantity $J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}})$ converges to 1. If $\mathbf{\Sigma}$ is diagonal, then $J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}}) = 1$ exactly. In such cases, the quantity Λ_∞ in Corollary III.6 does not depend on the unknown underlying distribution of the U-scores. As a result, the expected number of δ-hubs in $\mathcal{G}_\rho(\mathbf{\Psi})$ and the probability of discovery of at least one δ-hub do not depend on the underlying distribution. We will see in Sec. 3.5 that this result is useful in assigning statistical significance levels to vertices of the graph $\mathcal{G}_\rho(\mathbf{\Psi})$.

3.4.6 Phase transitions and critical threshold

It can be seen from Theorem III.5 and Corollary III.6 that the number of δ-hub discoveries exhibits a phase transition in the high-dimensional regime where the number of variables p can be very large relative to the number of samples m. Specifically, assume that the population covariance matrix $\mathbf{\Sigma}$ is block-sparse as in Section 3.4.5. Then as the correlation threshold ρ is reduced, the number of δ-hub discoveries abruptly increases to the maximum, p. Conversely as ρ increases, the number of discoveries quickly approaches zero. Similarly, the family-wise error rate (i.e. the probability of discovering at least one δ-hub in a graph with no true hubs) exhibits a phase transition as a function of ρ. Figure 3.3 shows the family-wise error rate

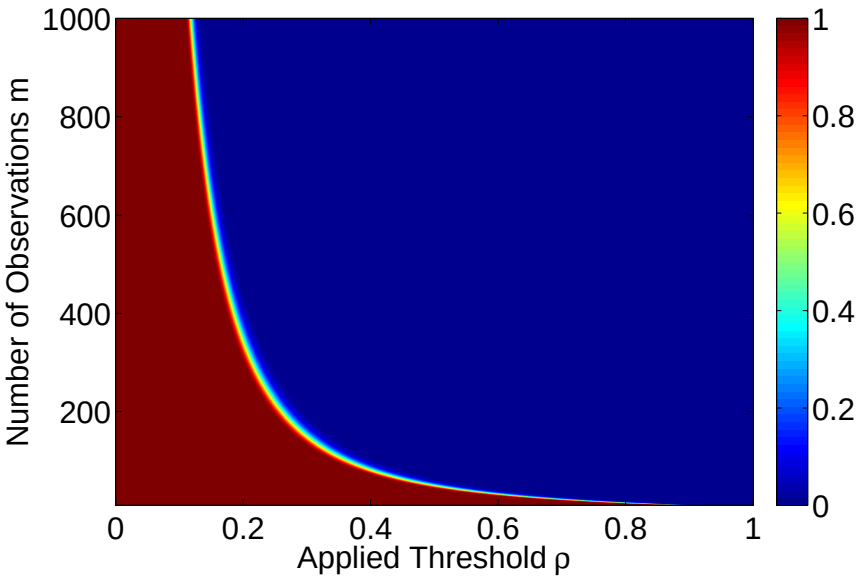

Figure 3.3: Family-wise error rate as a function of correlation threshold ρ and number of samples m for $p = 1000, \delta = 1$. The phase transition phenomenon is clearly observable in the plot.

obtained via expression (3.32) for $\delta = 1$ and $p = 1000$, as a function of ρ and the number of samples m. It is seen that for a fixed value of m there is a sharp transition in the family-wise error rate as a function of ρ.

The phase transition phenomenon motivates the definition of a critical threshold $\rho_{c,\delta}$ as the threshold ρ satisfying the following slope condition:

$$\partial \mathbb{E}[N_{\delta,\rho}]/\partial \rho = -p.$$

Using (3.16) the solution of the above equation can be approximated via the expression below:

$$\rho_{c,\delta} = \sqrt{1 - (c_{m,\delta}(p-1))^{-2\delta/(\delta(2m-3)-2)}}, \tag{3.33}$$

where $c_{m,\delta} = b_{m-1}\delta J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}})$. The screening threshold ρ should be chosen greater than $\rho_{c,\delta}$ to prevent excessively large numbers of false positives. Note that

the critical threshold $\rho_{c,\delta}$ also does not depend on the underlying distribution of the U-scores when the covariance matrix $\mathbf{\Sigma}$ is block-sparse.

Expression (3.33) is similar to the expression obtained in (*Hero and Rajaratnam*, 2012) for the critical threshold in real-valued correlation screening. However, in the complex-valued case the coefficient $c_{m,\delta}$ and the exponent of the term $c_{m,\delta}(p-1)$ are different from the real case. This generally results in smaller values of $\rho_{c,\delta}$ for fixed m and δ.

Figure 3.4 shows the value of $\rho_{c,\delta}$ obtained via (3.33) as a function of m for different values of δ and p. The critical threshold decreases as either the sample size m increases, the number of variables p decreases, or the vertex degree δ increases. Note that even for ten billion (10^{10}) dimensions (upper triplet of curves in the figure) only a relatively small number of samples are necessary for complex-valued correlation screening to be useful. For example, with $m = 200$ one can reliably discover connected vertices ($\delta = 1$ in the figure) having correlation greater than $\rho_{c,\delta} = 0.5$.

3.5 Application to spectral screening of multivariate Gaussian time series

In this section, the complex-valued correlation hub screening method of Section 3.4 is applied to stationary multivariate Gaussian time series. Assume that the time series $X^{(1)}, \cdots, X^{(p)}$ defined in Section 3.3 satisfy the conditions of Corollary III.3. Assume also that a total of $N = n \times m$ time samples of $X^{(1)}, \cdots, X^{(p)}$ are available. We divide the N samples into m parts of n consecutive samples and we take the n-point DFT of each part. Therefore, for each time series, at each frequency $f_i = (i-1)/n$, $1 \leq i \leq n$, m samples are available. This allows us to construct a (partial) correlation graph corresponding to each frequency. We denote the (partial) correlation graph corresponding to frequency f_i and correlation threshold ρ_i as $\mathcal{G}_{f_i,\rho_i}$. $\mathcal{G}_{f_i,\rho_i}$ has p vertices

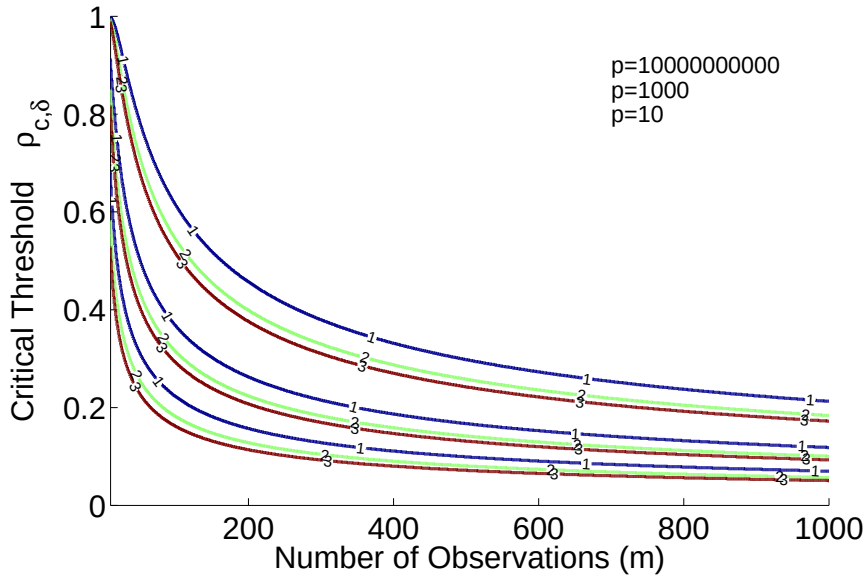

Figure 3.4: The critical threshold $\rho_{c,\delta}$ as a function of the sample size m for $\delta = 1, 2, 3$ (curve labels) and $p = 10, 1000, 10^{10}$ (bottom to top triplets of curves). The figure shows that the critical threshold decreases as either m or δ increases. When the number of samples m is small the critical threshold is close to 1 in which case reliable hub discovery is impossible. However a relatively small increment in m is sufficient to reduce the critical threshold significantly. For example for $p = 10^{10}$, only $m = 200$ samples are enough to bring $\rho_{c,1}$ down to 0.5.

$v_1, v_2, \cdots, v_p$ corresponding to time series $X^{(1)}, X^{(2)}, \cdots, X^{(p)}$, respectively. Vertices v_k and v_l are connected if the magnitude of the sample (partial) correlation between the DFTs of $X^{(k)}$ and $X^{(l)}$ at frequency f_i (i.e. the sample (partial) correlation between $Y^{(k)}(i-1)$ and $Y^{(l)}(i-1)$) is at least ρ_i.

Consider a single frequency f_i and the null hypothesis, $\mathcal{H}_0$, that the correlations among the time series $X^{(1)}, X^{(2)}, \cdots, X^{(p)}$ at frequency f_i are block sparse in the sense of Section 3.4.5. As discussed in Sec. 3.4.5, under $\mathcal{H}_0$ the expected number of δ-hubs and the probability of discovery of at least one δ-hub in graph $\mathcal{G}_{f_i,\rho_i}$ are not functions of the unknown underlying distribution of the data. Therefore the results of Corollary III.6 may be used to quantify the statistical significance of declaring vertices of $\mathcal{G}_{f_i,\rho_i}$ to be δ-hubs. The statistical significance is represented by the p-value,

defined in general as the probability of having a test statistic at least as extreme as the value actually observed assuming that the null hypothesis $\mathcal{H}_0$ is true. In the case of correlation hub screening, the p-value $pv_\delta(j)$ assigned to vertex v_j for being a δ-hub is the maximal probability that v_j maintains degree δ given the observed sample correlations, assuming that the block-sparse hypothesis $\mathcal{H}_0$ is true. The detailed procedure for assigning p-values is similar to the procedure in (*Hero and Rajaratnam*, 2012) for real-valued correlation screening and is illustrated in Algorithm 1. Equation (3.33) helps in choosing the initial threshold ρ^*.

The bottleneck of the computational complexity of Algorithm 1 is finding δth greatest element of the jth row of the sample (partial) correlation matrix $\mathbf{\Psi}$. This can be done by performing approximate k-NN algorithm on the U-scores associated with $\mathbf{\Psi}$ in $O(m^2p)$. Hence the overall computational complexity of performing spectral hub screening in all n frequencies is $O(nm^2p)$. Without using approximate k-NN methods, the overall computational complexity is $O(nmp^2)$.

Algorithm 1: Spectral hub screening of multivariate time series.

- initialization:
  - Select a screening threshold ρ^*;
  - Calculate the degree d_j^x of each vertex in $\mathcal{G}_{\rho^*}(\mathbf{\Psi})$;
  - Select a value of $\delta \in \{1, \cdots, \max_{1\leq j\leq p} d_j^x\}$;
- **for** $j = 1$ **to** p **do**
  - find $\rho_j(\delta)$ as the δth greatest element of the jth row of the sample (partial) correlation matrix;
  - Approximate the p-value corresponding to vertex v_j as $pv_\delta(j) \approx 1 - \exp(-\mathbb{E}[N_{\delta,\rho_j(\delta)}]/\varphi(\delta))$, where $\mathbb{E}[N_{\delta,\rho_j(\delta)}]$ is approximated by the limiting expression (3.31) using $J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}}) = 1$;
  - Approximate the p-value corresponding to the ith independent variable X_i as $pv_\delta(i) \approx 1 - \exp(-\xi_{p,m,\delta,\rho_i^{\mathrm{mod}}(\delta)})$;
- Screen variables by thresholding the p-values $pv_\delta(j)$ at desired significance level;

Given Corollary III.3, for $i \neq j$ the correlation graphs $\mathcal{G}_{f_i,\rho_i}$ and $\mathcal{G}_{f_j,\rho_j}$ and their associated inferences are approximately independent. Thus we can solve multiple inference problems by first performing correlation hub screening on each graph as

discussed above and then aggregating the inferences at each frequency in a straightforward manner. Examples of aggregation procedures are described below.

3.5.1 Disjunctive hubs

One task that can be easily performed is finding the p-value for a given time series to be a hub in at least one of the graphs $\mathcal{G}_{f_1,\rho_1}, \cdots, \mathcal{G}_{f_n,\rho_n}$. More specifically, for each $j = 1, \ldots, p$ denote the p-values for vertex v_j being a δ-hub in $\mathcal{G}_{f_1,\rho_1}, \cdots, \mathcal{G}_{f_n,\rho_n}$ by $pv_{f_1,\rho_1,\delta}(j), \cdots, pv_{f_n,\rho_n,\delta}(j)$ respectively. These p-values are obtained using Algorithm 1. Then $pv_\delta(j)$, the p-value for the vertex v_j being a δ-hub in at least one of the frequency graphs $\mathcal{G}_{f_1,\rho_1}, \cdots, \mathcal{G}_{f_n,\rho_n}$ can be approximated as:

$$\mathbb{P}(\exists i : d_{j,f_i} \geq \delta \mid \mathcal{H}_0) \approx \hat{pv}_\delta(j) = 1 - \prod_{i=1}^{n}(1 - pv_{f_i,\rho_i,\delta}(j)),$$

in which d_{j,f_i} is the degree of v_j in the graph $\mathcal{G}_{f_i,\rho_i}$.

3.5.2 Conjunctive hubs

Another property of interest is the existence of a hub at all frequencies for a particular time series. In this case we have:

$$\mathbb{P}(\forall i : d_{j,f_i} \geq \delta \mid \mathcal{H}_0) \approx \check{pv}_\delta(j) = \prod_{i=1}^{n} pv_{f_i,\rho_i,\delta}(j).$$

3.5.3 General persistent hubs

The general case is the event that at least K frequencies have hubs of degree at least δ at vertex v_j. For this general case we have:

$$\mathbb{P}(\exists i_1,\ldots,i_K : d_{j,f_{i_1}} \geq \delta, \ldots d_{j,f_{i_K}} \geq \delta \mid \mathcal{H}_0) =$$
$$\sum_{k'=K}^{n} \sum_{\substack{i_1<\ldots<i_{k'},i_{k'+1}<\ldots<i_n \\ \{i_1,\ldots,i_n\}=\{1,\ldots,n\}}} \prod_{l=1}^{k'} pv_{f_{i_l},\rho_{i_l},\delta}(j) \prod_{l'=k'+1}^{n} \left(1 - pv_{f_{i_{l'}},\rho_{i_{l'}},\delta}(j)\right).$$

3.6 Experimental results

3.6.1 Phase transition phenomenon and mean number of hubs

We first performed numerical simulations to confirm Theorem III.5 and Corollary III.6 for complex-valued correlation screening. Samples were generated from p uncorrelated complex Gaussian random variables. Figure 3.5 shows the number of discovered 1-hubs for $p = 1000$ and several sample sizes m. The plots from left to right correspond to $m = 2000, 1000, 500, 100, 50, 20, 10, 6$ and 4, respectively. The phase transition phenomenon is clearly observed in the plot. Table 3.1 shows the predicted value obtained from formula (3.33) for the critical threshold. As can be seen in Fig. 3.5, the empirical phase transition thresholds approximately match the predicted values of Table 3.1. Moreover, to confirm the accuracy of equation (3.31) in Corollary III.6, we list the number of hubs for $m = 100$ in Table 3.2. The left column shows the empirical average number of hubs of degree at least $\delta = 1, 2, 3, 4$ in a network of i.i.d. complex Gaussian random variables. The numbers in this column are obtained by averaging 1000 independent experiments. The right column shows the predicted value of $\mathbb{E}[N_{\delta,\rho}]$ obtained via formula (3.31) with $J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}}) = 1$ for the i.i.d. case. As we see the empirical and predicted values are close to each other.

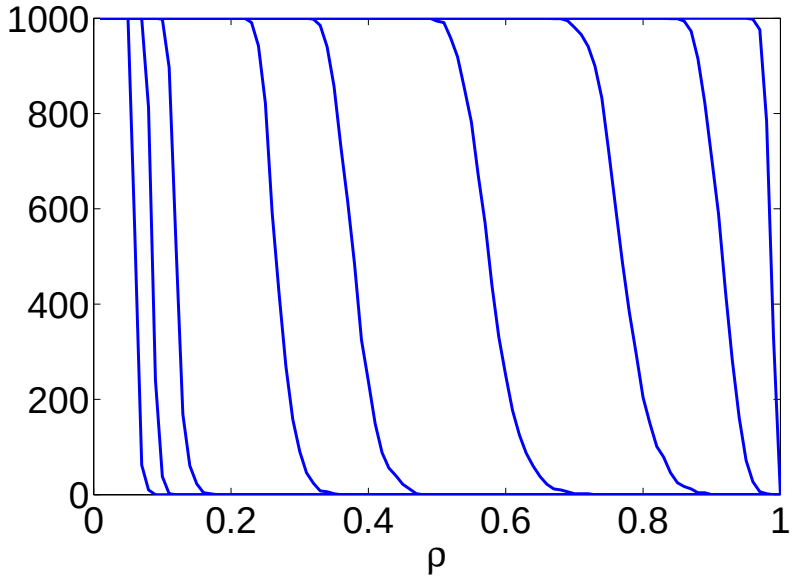

Figure 3.5: Phase transition phenomenon: the number of 1-hubs in the sample correlation graph corresponding to uncorrelated complex Gaussian variables as a function of correlation threshold ρ. Here, $p = 1000$ and the plots from left to right correspond to $m = 2000, 1000, 500, 100, 50, 20, 10, 6$ and 4, respectively.

m	2000	1000	500	100	50	20	10	6	4
$\rho_{c,\delta}$	0.05	0.07	0.10	0.24	0.35	0.56	0.78	0.94	0.99

Table 3.1: The value of critical threshold $\rho_{c,\delta}$ obtained from formula (3.33) for $p = 1000$ complex variables and $\delta = 1$. The predicted $\rho_{c,\delta}$ approximates the phase transition thresholds in Fig. 3.5.

degree threshold	empirical ($\mathbb{E}[N_{\delta,\rho}]$)	predicted ($\mathbb{E}[N_{\delta,\rho}]$)
$d_i \geq \delta = 1$	284	335
$d_i \geq \delta = 2$	45	56
$d_i \geq \delta = 3$	5	6
$d_i \geq \delta = 4$	0	0

Table 3.2: Empirical average number of discovered hubs vs. predicted average number of discovered hubs in an uncorrelated complex Gaussian network. Here $p = 1000$, $m = 100$, $\rho = 0.28$. The empirical values are obtained by performing 1000 independent experiments.

3.6.2 Asymptotic independence of spectral components for AR(1) model

To illustrate the asymptotic independence property and convergence rate of Theorem III.1, we considered the simple case of an AR(1) process,

$$X(k) = \varphi_1 X(k-1) + \varepsilon(k),\ k \geq 1, \tag{3.34}$$

in which $X(0) = 0, \varphi_1 = 0.9$ and $\varepsilon(.)$ is a stationary Gaussian process with no temporal correlation and standard deviation 1. We performed Monte-Carlo simulations to compute the correlation between spectral components at different frequencies for window sizes $n = 10, 20, \ldots, 250$. More specifically, we set $k = 1$ and $l = 2$ and empirically estimated $|\mathrm{cor}\,(Y(k), Y(l))\,|$ using 50000 Monte-Carlo trials for each value of window size n. Figure 3.6 shows the result of this experiment. It is observable that the magnitude of $\mathrm{cor}\,(Y(k), Y(l))$ is bounded above by the function $10/n$. This observation is consistent with Theorem III.1.

3.6.3 Spectral correlation screening of a band-pass multivariate time series

Next we analyzed the performance of the proposed complex-valued correlation screening framework on a synthetic data set for which the expected results are known.

We synthesized a multivariate stationary Gaussian time series using the the following procedure. Here we set $p = 1000, N = 12000$ and $m = n = 100$. The discrepancy between N and the product mn is explained below. Let $X(k), 0 \leq k \leq N-1$ be a sequence of i.i.d. zero-mean Gaussian random variables (i.e. white Gaussian noise) with standard deviation of 1. The p time series $X^{(1)}(k), \ldots, X^{(p)}(k), 0 \leq k \leq N-1$ are obtained from $X(k)$ by band-pass filtering and adding independent white Gaussian

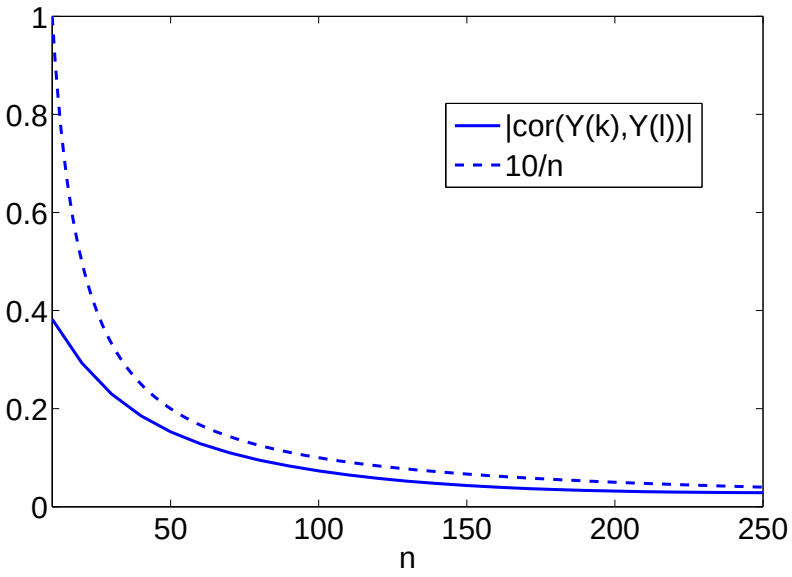

Figure 3.6: Correlation coefficient $|\mathrm{cor}\,(Y(1), Y(2))\,|$ as a function of window size n, empirically estimated using 50000 Monte-Carlo trials. Here $Y(.)$ is the DFT of the AR(1) process (3.34). The magnitude of the correlation for $n = 10, 20, \ldots, 250$ is bounded above by the function $10/n$. This observation is consistent with the convergence rate in Theorem III.1.

noise. Specifically,

$$X^{(i)}(k) = h_i(k) \star X(k) + N_i(k),\ 1 \leq i \leq p, 0 \leq k \leq N-1,$$

in which $\star$ represents the convolution operator, $h_i(.)$ is the impulse response of the ith band-pass filter and $N_i(.)$ is an independent white Gaussian noise series whose standard deviation is 0.1. Since stable filtering of a stationary series results in another stationary series, the obtained series $X^{(1)}(k), \ldots, X^{(p)}(k)$ are stationary and Gaussian. For $i = 10l, 1 \leq l \leq 50$, $h_i(k)$ is the impulse response of a band-pass filter with pass band $f \in [(4l-1)/400, 4l/400]$. We approximate the ideal band-pass filters with finite impulse response (FIR) Chebyshev filters (*Oppenheim et al.*, 1989). Also for $i = 500 + 10l, 1 \leq l \leq 50$ we set $h_i(k) = h_{i-500}(k)$. For all of the other values of i (i.e. $i \neq 10l$) we set $h_i(k) = 0, 0 \leq k \leq N-1$.

Figure 3.7 shows the signal part of the time series (i.e. $h_i(k) \star X(k)$) for $i = 100, 200, 300, 400$. It is seen that the first 2000 samples of the signals reflect the transient response of the filters. These 2000 samples are not included for the purpose of correlation screening. Hence the actual number of time samples considered is $mn = 10000$. Figure 3.8 shows the magnitude of the DFTs of the signals, $Y^{(i)}(k)$, for $i = 50, 100, \ldots, 500$. The band-pass structure of the signals is clearly observable in the figure.

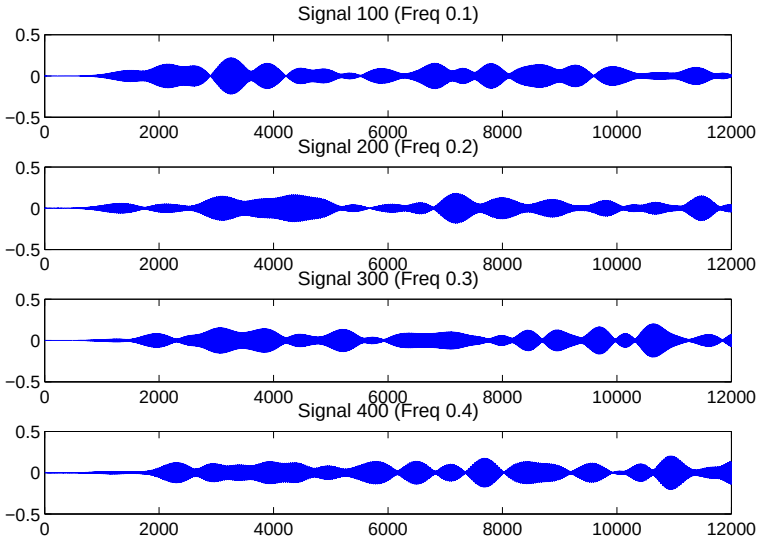

Figure 3.7: Signal part of the band-pass time series $X^{(i)}(k)$ (i.e. $h_i(k) \star X(k)$) for $i = 100, 200, 300, 400$.

We first constructed a correlation matrix for the time series $X^{(1)}(k), \ldots, X^{(p)}(k)$ from their simultaneous time samples. Figure 3.9 illustrates the structure of the thresholded sample correlation matrix and the corresponding correlation graph. Note that this is a real-valued correlation screening problem in the time domain. The correlation threshold used here is $\rho = 0.2$ which is well above the critical threshold $\rho_{c,1} = 0.028$ obtained via formula (10) in (*Hero and Rajaratnam*, 2012) for $p = 1000$ and $N = 10000$.

To examine the spectral structure of the correlations in Fig. 3.9, we then performed

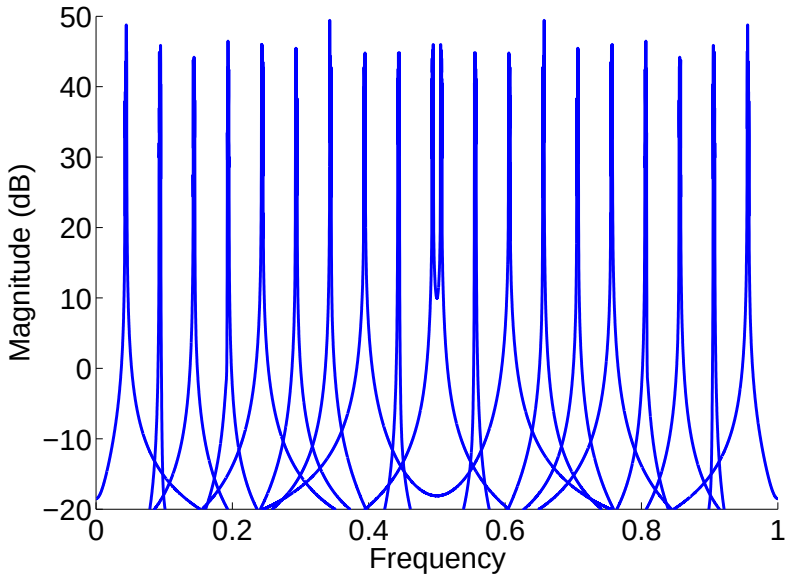

Figure 3.8: DFT magnitude of the band-pass signals $h_i(k) \star X(k)$ (i.e. $20\log_{10}(|Y^{(i)}(.)|)$) as a function of frequency for $i = 50, 100, \ldots, 500$.

complex-valued correlation screening on the spectra of the time series $X^{(1)}(k), \ldots, X^{(p)}(k)$. Figure 3.10 shows the constructed correlation graphs $\mathcal{G}_{f,\rho}$ for $f = [0.1, 0.2, 0.3, 0.4]$ and correlation threshold $\rho = 0.9$, which corresponds to a $\delta = 1$ false positive rate $\mathbb{P}(N_{\delta,\rho} > 0) \approx 10^{-65}$ (using $\delta = 1$ in the asymptotic relation (3.32) with $\Lambda_\infty = e^{\delta}_{m,\delta}/\delta!$ as specified by (3.31)). Note that the value of the correlation threshold is set to be higher than the critical threshold $\rho_c = 0.24$. It can be observed that performing complex-valued spectral correlation screening at each frequency correctly discovers the correlations between the time series which are active around that frequency. As an example, for $f = 0.2$ the discovered hubs (for $\delta = 1$) are the time series $X^{(i)}(k)$ for $i \in \{200, 700\}$. These time series are the ones that are active at frequency $f = 0.2$. Under the null hypothesis of diagonal covariance matrices, the p-values for the discovered hubs are of order 10^{-65} or smaller. These results show that complex-valued spectral correlation screening is able to resolve the sources of correlation between time series in the spectral domain.

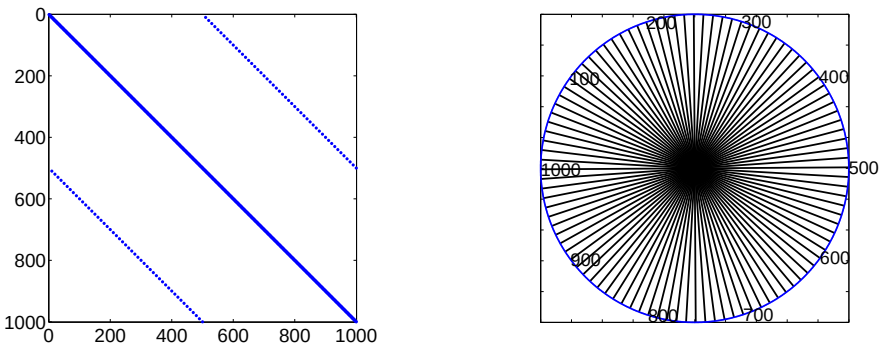

Figure 3.9: (Left) The structure of the thresholded sample correlation matrix in the time domain. (Right) The correlation graph corresponding to the thresholded sample correlation matrix in the time domain.

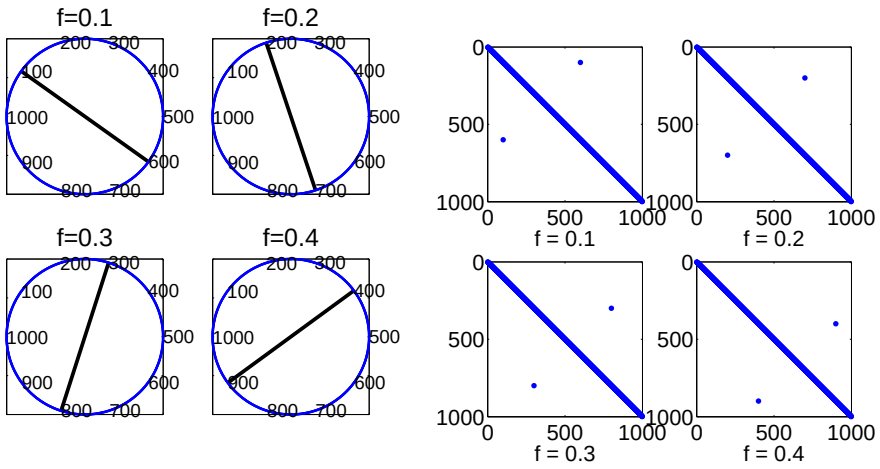

Figure 3.10: Spectral correlation graphs $\mathcal{G}_{f,\rho}$ for $f = [0.1, 0.2, 0.3, 0.4]$ and correlation threshold $\rho = 0.9$, which corresponds to a false positive probability of 10^{-65}. The data used here is a set of synthetic time series obtained by band-pass filtering of a Gaussian white noise series with the band-pass filters shown in Fig. 3.8. As can be seen, complex correlation screening is able to extract the correlations at specific frequencies. This is not directly feasible in the time domain analysis.

3.6.4 Vulnerable asset discovery in financial markets

Asset-wise analysis. We applied the spectral correlation screening method to a financial data set.

Stock prices are commonly modeled by a geometric Brownian motion (*Tsay*, 2005).

Sector	# of stocks
Basic Industries	122
Capital Goods	209
Consumer Durables	78
Consumer Non-Durables	117
Consumer Services	335
Energy	110
Finance	334
Health Care	205
Miscellaneous	36
Public Utilities	103
Technology	256
Transportation	37

Table 3.3: Number of stocks in each sector out of the 1942 selected stocks in Russell 3000 index.

This results in a normal distribution for the log-returns which fits our distributional assumption on the data (see Sec. 3.4.1).

The data set consists of the daily log-returns for those components of Russell 3000 index for which the stock prices from January 2nd 2003 to May 2nd 2013 are available at Yahoo! Finance. There are 1942 such stocks. A total of $N = 2600$ samples of daily log-returns are available for each stock between the mentioned dates. The selected 1942 stocks are from 12 different sectors which cover 96 different industries. The names of the different sectors and the number of stocks corresponding to each sector are shown in Table 3.3.

We divided the 2600 samples of each time series into $m = 51$ half intersecting windows of length $n = 100$ (i.e. each window intersects with the previous window for 50 consecutive samples). The 100-point DFT is then applied to each window to obtain 51 samples of the spectra of the 1942 time series at 100 different frequencies $f_i = (i-1)/100, 1 \leq i \leq 100$. We constructed the correlation and partial correlation graphs $\mathcal{G}_{f_i,\rho}$ at each frequency f_i. The correlation and partial correlation thresholds are set to $\rho = 0.8$ and $\rho = 0.9$, respectively, which correspond to respective false

positive probabilities of approximately 10^{-14} and 10^{-27} under the null hypothesis of diagonal covariance matrices. Note that the correlation and partial correlation thresholds used in this experiment are greater than the critical threshold $\rho_c = 0.36$, obtained via formula (3.33) using $p = 1942, m = 51, \delta = 1$ and $J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}}) = 1$ (for larger values of δ, the critical threshold ρ_c is less than 0.36).

At each frequency $f_1, \cdots, f_{100}$ we identified a set of 100 hubs by picking the stocks with the 100 smallest p-values for $\delta = 1$. Then we computed the top 100 most frequent stocks among the 100 sets corresponding to each frequency. Let $\mathcal{S}_{\text{cor}}$ and $\mathcal{S}_{\text{parcor}}$ denote such sets for complex-valued correlation screening and partial correlation screening, respectively. For the case of correlation screening, $\mathcal{S}_{\text{cor}}$ only covers 4 (out of 12) different sectors and 14 (out of 96) different industries. More than half of the stocks in $\mathcal{S}_{\text{cor}}$ are from the industry "Real Estate Investment Trust" (REIT) (see Table 3.4) due to the rather dense correlation network among the REIT stocks. This suggests that using direct correlations can be rather misleading about the drivers of the market (or vulnerable assets). More specifically due to such clique-type interconnections in many real-world data sets, correlation hubs may not necessarily be the most important variables in the data.

One may think that using pairwise partial correlations would be sufficient since it only considers causal relationships (assuming Gaussian data). The set $\mathcal{S}_{\text{parcor}}$ obtained by partial correlation screening, covers all 12 sectors and 61 industries (see Table 3.5). However, as opposed to expectations, there are relatively few hubs in important sectors like Energy and Finance which cover around 23 percent of the original 1942 stocks. Below we show that considering (partial) correlations between subsets of stocks (i.e. industries in this case) can lead to more intuitive results. Such analysis further reveals the necessity of hyper-graphs in this context.

Industry-wise analysis. Asset-wise analysis of market data is aimed at discovering a subset of assets that are potential drivers of the market. This type of analysis

Sector	# of stocks
Capital Goods	3
Consumer Services	60
Energy	10
Finance	27

Table 3.4: Number of stocks in each sector for the set $\mathcal{S}_{\text{cor}}$.

Sector	# of stocks
Basic Industries	4
Capital Goods	12
Consumer Durables	1
Consumer Non-Durables	6
Consumer Services	19
Energy	4
Finance	9
Health Care	16
Miscellaneous	5
Public Utilities	10
Technology	13
Transportation	1

Table 3.5: Number of stocks in each sector for the set $\mathcal{S}_{\text{parcor}}$.

can be insightful but it mainly relies on the assumption that single stocks can drive the market. However, due to the small capitalization of each stock compared to the whole market, this assumption may not be realistic. As an alternative, industry-wise analysis of the market data could reveal stronger associations.

In order to perform industry-wise analysis, defining an appropriate industry-wise (partial) correlation matrix is necessary. We define industry-wise (partial) correlations by averaging over asset-wise (partial) correlation coefficients. More precisely, let $\mathcal{S}_i, 1 \leq i \leq 96$, be the set of stocks in industry number i. Note that $\{\mathcal{S}_i\}_{i=1}^{96}$ is a partition of the complete set of stocks, $\mathcal{S}$ (i.e. $\cup_{i=1}^{96}\mathcal{S}_i = \mathcal{S}$ and $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset, 1 \leq i \neq j \leq 96$). Moreover, let $\mathbf{R}^{\mathcal{S},f_k}$ (respectively, $\mathbf{P}^{\mathcal{S},f_k}$) denote the 1942×1942 asset-wise correlation (respectively, partial correlation) matrix at frequency f_k. Using $\mathbf{R}^{\mathcal{S},f_k}$ we define the $(i,j)th$ entry of the 96×96 matrix $\tilde{\mathbf{R}}^{\mathcal{I},f_k}$ as:

$$\tilde{\mathbf{R}}_{ij}^{\mathcal{I},f_k} = \frac{1}{|\mathcal{S}_i||\mathcal{S}_j|}\sum_{r\in\mathcal{S}_i}\sum_{l\in\mathcal{S}_j}\mathbf{R}_{rl}^{\mathcal{S},f_k}. \tag{3.35}$$

Similarly define the $(i,j)th$ entry of the 96×96 matrix $\tilde{\mathbf{P}}^{\mathcal{I},f_k}$ as:

$$\tilde{\mathbf{P}}_{ij}^{\mathcal{I},f_k} = \frac{1}{|\mathcal{S}_i||\mathcal{S}_j|}\sum_{r\in\mathcal{S}_i}\sum_{l\in\mathcal{S}_j}\mathbf{P}_{rl}^{\mathcal{S},f_k}. \tag{3.36}$$

The industry-wise correlation and partial correlation matrices at frequency f_k are then defined as:

$$\mathbf{R}^{\mathcal{I},f_k} = \mathbf{D}_{\tilde{\mathbf{R}}^{\mathcal{I},f_k}}^{-\frac{1}{2}}\tilde{\mathbf{R}}^{\mathcal{I},f_k}\mathbf{D}_{\tilde{\mathbf{R}}^{\mathcal{I},f_k}}^{-\frac{1}{2}}, \tag{3.37}$$

and

$$\mathbf{P}^{\mathcal{I},f_k} = \mathbf{D}_{\tilde{\mathbf{P}}^{\mathcal{I},f_k}}^{-\frac{1}{2}}\tilde{\mathbf{P}}^{\mathcal{I},f_k}\mathbf{D}_{\tilde{\mathbf{P}}^{\mathcal{I},f_k}}^{-\frac{1}{2}}, \tag{3.38}$$

respectively. In other words, the (partial) correlation coefficient between industry i and industry j is defined as the normalized average of the (partial) cross-correlation coefficients between all stocks in industry i and all stocks in industry j.

It can be shown that the industry-wise correlation and partial correlation matrices constructed above are symmetric and positive definite. Thus we can perform complex-valued correlation and partial correlation screening at each frequency as before to discover hub industries. Note that since the notion of number of samples m is not well defined in the construction of the industry-wise (partial) correlation matrices, p-values cannot be computed. However, due to the fact that the p-values assigned via Algorithm 1 in Chapter III are decreasing functions of the quantities $\rho_i^{\text{mod}}(\delta)$, the ordering of p-values can be easily obtained via sorting $\rho_i^{\text{mod}}(\delta)$ for $1 \leq i \leq p$.

Similar to the asset-wise analysis of the previous subsection, we discovered a set of 15 hub industries at each frequency by selecting the industries with smallest p-values at that frequency for $\delta = 1$. The sets $\mathcal{I}_{\text{cor}}$ and $\mathcal{I}_{\text{parcor}}$ are defined as the 15 most frequent industries among the hub industries obtained at each frequency via complex-valued correlation and partial correlation screening, respectively. The result of this analysis is shown in Table 3.6. It is evident that most of the discovered industries are in the finance, public utilities and energy sectors.

3.7 Conclusion

This chapter presented a spectral method for correlation analysis of stationary multivariate Gaussian time series with a focus on identifying correlation hubs. The asymptotic independence of spectral components at different frequencies allows the problem to be decomposed into independent problems at each frequency, thus improving computational and statistical efficiency for high-dimensional time series. The method of complex-valued correlation screening is then applied to detect hub variables at each frequency. Using a characterization of the number of hubs discovered

$\mathcal{I}_{\text{cor}}$	Corresponding sector	$\mathcal{I}_{\text{parcor}}$	Corresponding sector
Bank	Finance	Bank	Finance
Bank (Midwest)	Finance	Bank (Midwest)	Finance
Diversified Co.	Consumer Non-Durables	Electric Util. (Central)	Public Utilities
Machinery	Consumer Non-Durables	Electric Utility (East)	Public Utilities
Petroleum (Producing)	Energy	Electric Utility (West)	Public Utilities
Natural Gas (Div.)	Energy	Natural Gas (Div.)	Energy
Electric Util. (Central)	Public Utilities	Petroleum (Producing)	Energy
Electric Utility (West)	Public Utilities	Oilfield Svcs/Equip.	Energy
Chemical (Specialty)	Capital Goods	Petroleum (Integrated)	Energy
Thrift	Finance	Semiconductor	Public Utilities
Financial Svcs. (Div.)	Finance	Semiconductor Equip	Finance
Electric Utility (East)	Public Utilities	Thrift	Finance
Electrical Equipment	Consumer Durables	Metals & Mining (Div.)	Basic Industries
Electronics	Energy	Steel	Basic Industries
Industrial Services	Energy	Natural Gas Utility	Public Utilities

Table 3.6: Industries in $\mathcal{I}_{\text{cor}}$ and $\mathcal{I}_{\text{parcor}}$. These industries which are obtained by complex-valued correlation and partial correlation screening, can be interpreted as the drivers of the market. It is evident that a majority of the discovered industries fall in to the finance, public utilities and energy sectors.

by the method, thresholds for hub screening can be selected to avoid an excessive number of false positives or negatives, and the statistical significance of hub discoveries can be quantified. The theory specifically considers the high-dimensional case where the number of samples at each frequency can be significantly smaller than the number of time series. Experimental results validated the theory and illustrated the applicability of complex-valued correlation screening to the spectral domain.

CHAPTER IV

Variable selection and prediction in high dimensional linear regression using hub screening

4.1 Introduction

In Chapter II we considered a local hub screening method in (partial) correlation graphs. In Chapter III we generalized the hub screening theory to the case of complex-valued random variables. In this Chapter we generalize the hub screening theory to the case of bipartite graphs. Our goal for such a generalization is to proposes a general adaptive procedure for budget-limited predictor design in high dimension called two-stage Sampling, Prediction and Adaptive Regression via Correlation Screening (SPARCS). SPARCS can be applied to high dimensional prediction problems in experimental science, medicine, finance, and engineering, as illustrated by the following. Suppose one wishes to run a sequence of experiments to learn a sparse multivariate predictor of a dependent variable Y (disease prognosis) based on a p dimensional set of independent variables $\mathbf{X} = [X_1, \ldots, X_p]^T$ (assayed biomarkers). Assume that the cost of acquiring the full set of variables $\mathbf{X}$ increases linearly in its dimension. SPARCS breaks the data collection into two stages in order to achieve an optimal tradeoff between sampling cost and predictor performance. In the first stage we collect a few (n) expensive samples $\{y_i, \mathbf{x}_i\}_{i=1}^n$, at the full dimension $p \gg n$ of $\mathbf{X}$, winnowing the

number of variables down to a smaller dimension $l < p$ using some form of variable selection. In the second stage we collect a larger number $(t - n)$ of cheaper samples of the l variables that passed the screening of the first stage. After the second stage, a low dimensional predictor is constructed by solving the regression problem using all t samples of the selected variables. Note that the SPARCS approach is embedded in the sampling process and is therefore closer to adaptive sampling, such as distilled sensing of Haupt, Castro and Nowak (2010), than it is to correlation learning, such as sure independence screening (SIS) of Fan and Lv (2007). SPARCS implements false positive control on the selected variables, is well suited to small sample sizes, and is scalable to high dimensions. We establish asymptotic bounds for the Familywise Error Rate (FWER), specify high dimensional convergence rates for support recovery, and establish optimal sample allocation rules to the first and second stages.

Much effort has been invested in the sparse regression problem where the objective is to learn a sparse linear predictor from training data $\{y_i, x_{i1}, x_{i2}, \ldots, x_{ip}\}_{i=1}^{n}$ where the number p of predictor variables is much larger that the number n of training samples. Applications in science and engineering where such "small n large p" problems arise include: sparse signal reconstruction (*Candés et al.*, 2005; *Donoho*, 2006); channel estimation in multiple antenna wireless communications (*Hassibi and Hochwald*, 2003; *Biguesh and Gershman*, 2006); text processing of internet documents (*Forman*, 2003; *Ding et al.*, 2002); gene expression array analysis (*Golub et al.*, 1999); combinatorial chemistry (*Suh et al.*, 2009); environmental sciences (*Rong*, 2011); and others (*Guyon and Elisseeff*, 2003). In this $n \ll p$ setting training a linear predictor becomes difficult due to rank deficient normal equations, overfitting errors, and high computation complexity.

A large number of methods for solving this sparse regression problem have been proposed. These include methods that simultaneously perform variable selection and predictor design and the methods that perform these two operations separately. The

former class of methods includes, for example, least absolute shrinkage and selection operator (LASSO), elastic LASSO, and group LASSO (*Guyon and Elisseeff*, 2003; *Tibshirani*, 1996; *Efron et al.*, 2004; *Bühlmann*, 2006; *Yuan and Lin*, 2005; *Friedman et al.*, 2001; *Bühlmann and Van De Geer*, 2011). The latter class of methods includes sequential thresholding approaches such as sure independence screening (SIS); and marginal regression (*Fan and Lv*, 2008; *Genovese et al.*, 2009, 2012; *Fan et al.*, 2010). All of these methods are offline in the sense that they learn the predictor from a batch of precollected samples of all the variables. In this chapter we propose an online framework, called two-stage Sampling, Prediction and Adaptive Regression via Correlation Screening (SPARCS), which unequally and adaptively samples the variables in the process of constructing the predictor. One of the principal results of this chapter is that, as compared under common sampling budget constraints, the proposed SPARCS method results in better prediction performance than offline methods.

Specifically, the SPARCS method for online sparse regression operates in two-stages. The first stage, which we refer to as the *SPARCS screening stage*, collects a small number of full dimensional samples and performs variable selection on them. Variable selection can be performed one of two ways, i.e., by screening the sample cross-correlation between Y and $\mathbf{X}$, as in sure independence screening (SIS), or by thresholding the generalized Ordinary Least Squares (OLS) solution, which we call predictive correlation screening (PCS). The second stage of SPARCS, referred to as the *SPARCS regression stage*, collects a larger number of reduced dimensional samples, consisting of the variables selected at the first stage, and regresses the responses on the the selected variables to build the predictor.

We establish the following theoretical results on SPARCS. First, under a sparse correlation assumption, we establish a Poisson-like limit theorem for the number of variables that pass the SPARCS screening stage as $p \to \infty$ for fixed n. This yields

a Poisson approximation to the probability of false discoveries that is accurate for small n and large p. The Poisson-like limit theorem also specifies a phase transition threshold for the false discovery probability. Second, with n the number of samples in the first stage, and t the total number of samples, we establish that n needs only be of order $\log p$ for SPARCS to succeed in recovering the support set of the optimal OLS predictor. Third, given a cost-per-sample that is linear in the number of assayed variables, we show that the optimal value of n is on the order of $\log t$. The above three results, established for our SPARCS framework, can be compared to theory for correlation screening (*Hero and Rajaratnam*, 2011, 2012), support recovery for multivariate LASSO (*Obozinski et al.*, 2011), and optimal exploration vs. exploitation allocation in multi-armed bandits (*Audibert et al.*, 2007).

SPARCS can of course also be applied offline. When implemented in this way, it can be viewed as an alternative to LASSO-type regression methods (*Tibshirani*, 1996; *Paul et al.*, 2008; *Wainwright*, 2009; *Huang and Jojic*, 2011; *Wauthier et al.*, 2013). LASSO based methods try to perform simultaneous variable selection and regression via minimizing an ℓ_1-regularized Mean Squared Error (MSE) objective function. Since the ℓ_1-regularized objective function is not differentiable, such an optimization is computationally costly, specially for large p. Several approaches such as LARS (*Efron et al.*, 2004; *Khan et al.*, 2007; *Hesterberg et al.*, 2008), gradient projection methods (*Figueiredo et al.*, 2007; *Quattoni et al.*, 2009), interior point methods (*Kim et al.*, 2007; *Koh et al.*, 2007) and active-set-type algorithms (*Kim and Park*, 2010; *Wen et al.*, 2010, 2012) have been developed to optimize the LASSO objective function. SPARCS however differs from LASSO as it does not consider a regularized objective function and instead performs variable selection via thresholding the min-norm solution to the non-regularized OLS problem.

Offline implementation of the proposed SPARCS method is a variant of correlation learning, also called marginal regression, simple thresholding, and sure independence

screening (*Genovese et al.*, 2009, 2012; *Fan and Lv*, 2008), wherein the simple sample cross-correlation vector between the response variable and the predictor variables is thresholded. The theory developed in this chapter yields phase transitions for the familywise false discovery rate for these methods.

The SPARCS screening stage has some similarity to recently developed correlation screening and hub screening in graphical models (*Hero and Rajaratnam*, 2011, 2012). However, there are important and fundamental differences. The methods in (*Hero and Rajaratnam*, 2011, 2012) screen for connectivity in the correlatrion graph, i.e., they only screen among the predictor variables $\{X_1, \ldots, X_p\}$. SPARCS screens for the connections in the bi-partite graph between the response variable Y and the predictor variables $X_1, ..., X_p$. Thus SPARCS is a supervised learning method that accounts for Y while the methods of (*Hero and Rajaratnam*, 2011, 2012) are unsupervised methods.

SPARCS can also be compared to sequential sampling methods, originating in the pioneering work of (*Wald et al.*, 1945). This work has continued in various directions such as sequential selection and ranking and adaptive sampling schemes (*Bechhofer et al.*, 1968; *Gupta and Panchapakesan*, 1991). Recent advances include the many multi-stage adaptive support recovery methods that have been collectively called distilled sensing (*Haupt et al.*, 2009, 2011; *Wei and Hero*, 2013a,b) in the compressive sensing literature. While bearing some similarities, our SPARCS approach differs from distilled sensing (DS). Like SPARCS, DS performs initial stage thresholding in order to reduce the number of measured variables in the second stage. However, in distilled sensing the objective is to recover a few variables with high mean amplitudes from a larger set of initially measured predictor variables. In contrast, SPARCS seeks to recover a few variables that are strongly predictive of the response variable from a large number of initially measured predictor variables and the corresponding response variable. Furthermore, unlike in DS, in SPARCS the final predictor uses all

the information on selected variables collected during both stages.

The chapter is organized as follows. Section 4.2 provides a practical motivation for SPARCS from the perspective of an experimental design problem in biology. It introduces the under-determined multivariate regression problem and formally defines the two stages of the SPARCS algorithm. Section 4.3 develops high dimensional convergence results for screening and support recovery performance of SPARCS. Section 4.3 also provides theory that specifies optimal sample allocation between the two stages of SPARCS. Section 4.4 presents simulation comparisons and an application to symptom prediction from gene expression data.

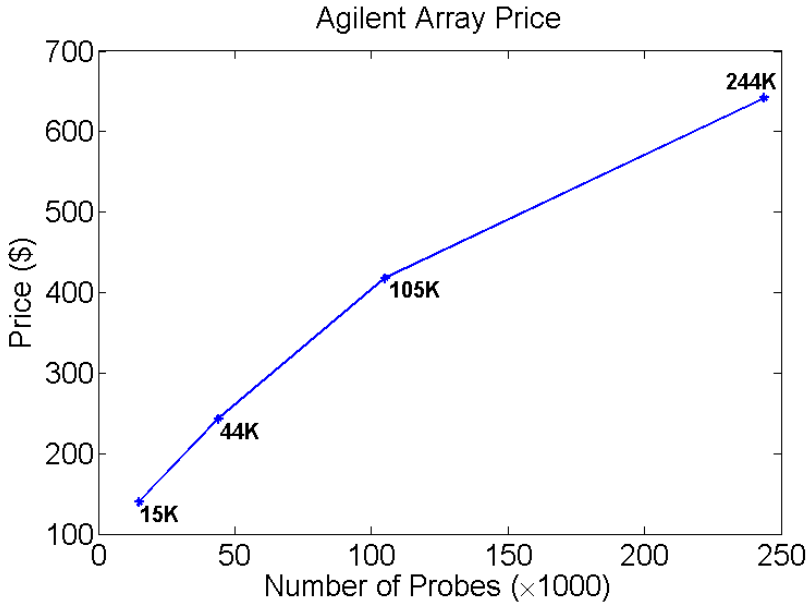

Figure 4.1: Price of arrays as a function of the number of probes. The dots represent pricing per slide for Agilent Custom Microarrays G2509F, G2514F, G4503A, G4502A (May 2014). The cost increases as a function of probeset size. Source: BMC Genomics and RNA Profiling Core.

4.2 Two-stage SPARCS method for online sparse regression

In this section we motivate the two-stage SPARCS method for online sparse regression via an experimental design problem in biology. Moreover, we formally define

each stage of the two-stage SPARCS method.

4.2.1 Motivation and definition for SPARCS

As a practical motivation for SPARCS consider the following sequential design problem that is relevant to applications where the cost of samples increases with the number p of variables. This is often the case for example, in gene microarray experiments: a high throughput "full genome" gene chip with $p = 40,000$ gene probes can be significantly more costly than a smaller assay that tests fewer than $p = 15,000$ gene probes (see Fig. 4.1). In this situation a sensible cost-effective approach would be to use a two-stage procedure: first select a smaller number l of variables on a few expensive high throughput samples and then construct the predictor on additional cheaper low throughput samples.

Motivated by the above practical example, we propose SPARCS as the following two-stage procedure. The first stage of SPARCS, also referred to as the SPARCS screening stage, performs variable selection and the second stage, also referred to as the SPARCS regression stage, constructs a predictor using the variables selected at the first stage. More specifically, assume that there are a total of t samples $\{y_i, \mathbf{x}_i\}_{i=1}^{t}$ available. During the first stage a number $n \leq t$ of these samples are assayed for all p variables and during the second stage the rest of the $t - n$ samples are assayed for a subset of $l < p$ of the variables selected in the first stage. Variable selection at SPARCS screening stage can be performed one of two ways, i.e., by screening the sample cross-correlation between Y and $\mathbf{X}$, as in sure independence screening (SIS), or by thresholding the solution to the generalized Ordinary Least Squares (OLS) problem, which we refer to as predictive correlation screening (PCS). Subsequently, the SPARCS regression stage uses standard OLS to design a l-variable predictor using all t samples collected during both stages.

An asymptotic analysis (as the total number of samples $t \to \infty$) of the above

two-stage predictor is undertaken in Sec. 4.3 to obtain the optimal sample allocation for stage 1 and stage 2. Assuming that a sample of a single variable has unit cost and that the total available budget for all of the samples is μ, the asymptotic analysis yields minimum Mean Squared Error (MSE) when n, t, p, and k satisfy the budget constraint:

$$np + (t - n)k \leq \mu, \tag{4.1}$$

where k is the true number of active variables in the underlying linear model. The condition in (4.1) is relevant in cases where there is a bound on the total sampling cost of the experiment and the cost of a sample increases linearly in its dimension p.

4.2.2 SPARCS screening stage

We start out with some notations. Assume that n i.i.d. paired realizations of $\mathbf{X} = [X_1, \ldots, X_p]$ and Y are available, where $\mathbf{X}$ is a random vector of predictor variables and Y is a scalar response variable to be predicted. We represent the $n \times p$ predictor data matrix as $\mathbb{X}$ and the $n \times 1$ response data vector as $\mathbb{Y}$. The $p \times p$ sample covariance matrix $\mathbf{S}^x$ for the rows of the data matrix $\mathbb{X}$ is defined as:

$$\mathbf{S}^x = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})^T (\mathbf{x}_i - \overline{\mathbf{x}}), \tag{4.2}$$

where $\mathbf{x}_i$ is the i-th row of data matrix $\mathbb{X}$, and $\overline{\mathbf{x}}$ is the vector average of all n rows of $\mathbb{X}$. We also denote the sample variance of the elements of $\mathbb{Y}$ as s^y.

Consider the $n \times (p+1)$ concatenated matrix $\mathbb{W} = [\mathbb{X}, \mathbb{Y}]$. The sample cross-covariance vector $\mathbf{S}^{xy}$ is defined as the upper right $p \times 1$ block of the $(p+1) \times (p+1)$ sample covariance matrix obtained by (4.2) using $\mathbb{W}$ as the data matrix instead of $\mathbb{X}$. The $p \times p$ sample correlation matrix $\mathbf{R}^x$ is defined as

$$\mathbf{R}^x = \mathbf{D}_{\mathbf{S}^x}^{-\frac{1}{2}} \mathbf{S}^x \mathbf{D}_{\mathbf{S}^x}^{-\frac{1}{2}}, \tag{4.3}$$

where $\mathbf{D_A}$ represents a matrix that is obtained by zeroing out all but diagonal entries of $\mathbf{A}$. Moreover, the $p \times 1$ sample cross-correlation vector $\mathbf{R}^{xy}$ is defined as:

$$\mathbf{R}^{xy} = \mathbf{D}_{\mathbf{S}^x}^{-\frac{1}{2}} \mathbf{S}^{xy} (s^y)^{-\frac{1}{2}}. \tag{4.4}$$

The SIS method for the SPARCS screening stage selects the desired number of variables, l, by picking the l variables that have the largest absolute sample correlation with the response variable Y. Therefore, SIS performs support recovery by discovering the entries of $\mathbf{R}^{xy}$ whose absolute value is larger than some threshold.

Next we introduce the under-determined ordinary lest squares (OLS) multivariate regression problem.

Assume that $n < p$. We define the generalized Ordinary Least Squares (OLS) estimator of Y given $\mathbf{X}$ as the min-norm solution of the under-determined least squares regression problem

$$\min_{\mathbf{B}^{xy} \in \mathbb{R}^p} \|\mathbb{Y} - \mathbb{X}\mathbf{B}^{xy}\|_F^2, \tag{4.5}$$

where $\|\mathbf{A}\|_F$ represents the Frobenius norm of matrix $\mathbf{A}$. The min-norm solution to (4.5) is the vector of regression coefficients

$$\mathbf{B}^{xy} = (\mathbf{S}^x)^\dagger \mathbf{S}^{xy}, \tag{4.6}$$

where $\mathbf{A}^\dagger$ denotes the Moore-Penrose pseudo-inverse of the matrix $\mathbf{A}$. If the i-th entry of the regression coefficient vector $\mathbf{B}^{xy}$ is zero then the i-th predictor variable is not included in the OLS estimator. This is the main motivation for the PCS method for variable selection at the SPARCS screening stage. More specifically, the PCS method selects the l entries of $\mathbf{B}^{xy}$ having the largest absolute values. Equivalently, PCS performs support recovery by discovering the entries of the generalized OLS solution

$\mathbf{B}^{xy}$ whose absolute value is larger than some threshold.

In Sec. 4.3.3 we will see that, under certain assumptions, SIS and PCS admit similar asymptotic support recovery guarantees. However, our experimental results in Sec. 4.4 show that for $n \ll p$, if SIS (or LASSO) is used instead of PCS in the SPARCS screening stage, the performance of the two-stage predictor suffers. This empirical observation suggests that pre-multiplication of $\mathbf{S}^{xy}$ by the pseudo-inverse $(\mathbf{S}^{x})^{\dagger}$ instead of by the diagonal matrix $\mathbf{D}_{\mathbf{S}^{x}}^{-1/2}$, can improve the performance of the SPARCS procedure.

4.2.3 SPARCS regression stage

In the second stage of SPARCS, a number $t-n$ of additional samples are collected for the $l < p$ variables found by the SPARCS screening stage. Subsequently, a sparse OLS predictor of Y is constructed using only the l variables designated at the SPARCS screening stage. Specifically, the predictor coefficients are determined from all of the t samples according to

$$(\mathbf{S}_{(l)}^{x})^{-1}\mathbf{S}_{(l)}^{xy}, \tag{4.7}$$

where $\mathbf{S}_{(l)}^{x}$ and $\mathbf{S}_{(l)}^{xy}$ are the $l \times l$ sample covariance matrix and the $l \times 1$ sample cross-covariance vector obtained for the set of l variables selected by the SPARCS screening stage.

In Sec. 4.3 we establish high dimensional convergence rates for the two stage online SPARCS procedure and we obtain asymptotically optimal sample allocation proportions n/t and $(t-n)/t$ for the first and second stage.

4.3 Convergence analysis

4.3.1 Notations and assumptions

In this section we introduce some additional notations and state the required assumptions for our convergence analysis of SPARCS.

The following notations are required for the theorems in this section. The surface area of the $(n-2)$-dimensional unit sphere S_{n-2} in $\mathbb{R}^{n-1}$ is denoted by by a_n. In the sequel we often refer to a vector on S_{n-2} as a *unit norm* vector.

Our convergence analysis for SPARCS uses the U-score representations of the data. More specifically, there exist a $(n-1) \times p$ matrix $\mathbb{U}^x$ with unit norm columns, and a $(n-1) \times 1$ unit norm vector $\mathbb{U}^y$ such that the following representations hold (*Hero and Rajaratnam*, 2011, 2012):

$$\mathbf{R}^x = (\mathbb{U}^x)^T \mathbb{U}^x, \tag{4.8}$$

and

$$\mathbf{R}^{xy} = (\mathbb{U}^x)^T \mathbb{U}^y. \tag{4.9}$$

Assume that $\mathbf{U}, \mathbf{V}$ are two independent and uniformly distributed random vectors on S_{n-2}. For a threshold $\rho \in [0,1]$, let $r = \sqrt{2(1-\rho)}$. $P_0(\rho, n)$ is then defined as the probability that either $\|\mathbf{U} - \mathbf{V}\|_2 \leq r$ or $\|\mathbf{U} + \mathbf{V}\|_2 \leq r$. $P_0(\rho, n)$ can be computed using the formula for the area of spherical caps on S_{n-2} (cf. (*Li*, 2011)):

$$P_0 = I_{1-\rho^2}((n-2)/2, 1/2), \tag{4.10}$$

in which $I_x(a,b)$ is the regularized incomplete beta function.

$S \subseteq \{1, \ldots, p\}$ denotes the set of indices of the variables selected by the SPARCS

screening stage. Moreover, l refers to the number of variables selected at the SPARCS screening stage, i.e., $|S| = l$.

For arbitrary joint densities $f_{\mathbf{U}_i^x,\mathbf{U}^y}(\mathbf{u},\mathbf{v}), 1 \leq i \leq p$ defined on the Cartesian product $S_{n-2} \times S_{n-2}$, define

$$\overline{f_{\mathbf{U}_*^x,\mathbf{U}^y}(\mathbf{u},\mathbf{v})} = \frac{1}{4p}\sum_{i=1}^{p}\sum_{s,t\in\{0,1\}} f_{\mathbf{U}_i^x,\mathbf{U}^y}(s\mathbf{u},t\mathbf{v}). \tag{4.11}$$

The quantity $\overline{f_{\mathbf{U}_*^x,\mathbf{U}^y}(\mathbf{u},\mathbf{v})}$ is key in determining the expected number of discoveries in screening the entries of the vector $\mathbf{\Phi}^{xy}$ in (4.30).

In the theorems of this chapter, q represents an upper bound on the number of entries in any row or column of covariance matrix $\mathbf{\Sigma}_x$ or cross-covariance vector $\mathbf{\Sigma}_{xy}$ that do not converge to zero as $p \to \infty$. We define $\|\Delta_{p,n,q}^{xy}\|_1$, the average dependency coefficient, as:

$$\|\Delta_{p,n,q}^{xy}\|_1 = \frac{1}{p}\sum_{i=1}^{p}\Delta_{p,n,q}^{xy}(i) \tag{4.12}$$

with

$$\Delta_{p,n,q}^{xy}(i) = \left\|(f_{\mathbf{U}_i^x,\mathbf{U}^y|\mathbf{U}_{A_q(i)}} - f_{\mathbf{U}_i^x,\mathbf{U}^y})/f_{\mathbf{U}_i^x,\mathbf{U}^y}\right\|_\infty, \tag{4.13}$$

in which $A_q(i)$ is defined as the set complement of indices of the q-nearest neighbors of $\mathbf{U}_i^x$ (i.e. the complement of indices of the q entries with largest magnitude in the i-th row of $\mathbf{\Sigma}_x$). Finally, the function J of the joint density $f_{\mathbf{U},\mathbf{V}}(\mathbf{u},\mathbf{v})$ is defined as:

$$J(f_{\mathbf{U},\mathbf{V}}) = |S_{n-2}| \int\limits_{S_{n-2}} f_{\mathbf{U},\mathbf{V}}(\mathbf{w},\mathbf{w})d\mathbf{w}. \tag{4.14}$$

The function $J(f_{\mathbf{U},\mathbf{V}})$ plays a key role in the asymptotic expression for the mean number of discoveries. Note that when observations are independent, by symmetry,

the marginal distributions of U-scores are exchangeable, i.e.,

$$f_{\mathbf{U}}(\mathbf{u}) = f_{\mathbf{U}}(\mathbf{\Pi u}) \text{ and } f_{\mathbf{V}}(\mathbf{v}) = f_{\mathbf{V}}(\mathbf{\Pi v}), \tag{4.15}$$

for any $(n-1) \times (n-1)$ permutation matrix $\mathbf{\Pi}$. Therefore, the joint distribution $f_{\mathbf{U},\mathbf{V}}$ must yield exchangeable marginals.

We now present two examples for which $J(f_{\mathbf{U},\mathbf{V}})$ has a closed form expression.

Example 1. If the joint distribution $f_{\mathbf{U},\mathbf{V}}$ is uniform over the product $S_{n-2} \times S_{n-2}$,

$$J(f_{\mathbf{U},\mathbf{V}}) = |S_{n-2}| \int\limits_{S_{n-2}} \frac{1}{|S_{n-2}|^2} d\mathbf{w} = \frac{|S_{n-2}|^2}{|S_{n-2}|^2} = 1. \tag{4.16}$$

Example 2. Consider the case where the joint distribution $f_{\mathbf{U},\mathbf{V}}$ is separable of the form

$$f_{\mathbf{U},\mathbf{V}}(\mathbf{u}, \mathbf{v}) = f_{\mathbf{U}}(\mathbf{u}) f_{\mathbf{V}}(\mathbf{v}), \tag{4.17}$$

i.e., $\mathbf{U}$ and $\mathbf{V}$ are independent. Let the marginals be von Mises-Fisher distributions over the sphere S_{n-2}

$$f_{\mathbf{U}}(\mathbf{u}) = C_{n-1}(\kappa) \exp(\kappa \boldsymbol{\mu}^T \mathbf{u}), \ \mathbf{u} \in S_{n-2}, \tag{4.18}$$

in which $\boldsymbol{\mu}$ and $\kappa \geq 0$ are the location parameter and the concentration parameter, respectively, and $C_{n-1}(\kappa)$ is a normalization constant, calculated as:

$$C_{n-1}(\kappa) = \frac{\kappa^{(n-1)/2-1}}{(2\pi)^{(n-1)/2} \overline{I}_{(n-1)/2-1}(\kappa)}, \tag{4.19}$$

where $\overline{I}_m$ is the modified Bessel function of the first kind of order m. $\overline{I}_m(x)$ can be

computed up to the desired precision using the expansion:

$$\overline{I}_m(x) = \sum_{l=0}^{\infty} \frac{(x/2)^{2l+n}}{l!\Gamma(l+m+1)}, \tag{4.20}$$

in which $\Gamma(.)$ is the gamma function.

Due to exchangeability of $f_{\mathbf{U}}(\mathbf{u})$, the only two feasible choices for $\boldsymbol{\mu}$ are $\boldsymbol{\mu} = \mathbf{1}$ and $\boldsymbol{\mu} = -\mathbf{1}$, where $\mathbf{1} = [1, 1, \ldots, 1]^T$. Hence the joint distribution can be written as:

$$\begin{aligned} f_{\mathbf{U},\mathbf{V}}(\mathbf{u},\mathbf{v}) = f_{\mathbf{U}}(\mathbf{u}) f_{\mathbf{V}}(\mathbf{v}) &= C_{n-1}(\kappa_1)\exp(\kappa_1 \boldsymbol{\mu}_1^T \mathbf{u}) C_{n-1}(\kappa_2)\exp(\kappa_2 \boldsymbol{\mu}_2^T \mathbf{v}) \\ &= C_{n-1}(\kappa_1) C_{n-1}(\kappa_2)\exp(\kappa_1 \boldsymbol{\mu}_1^T \mathbf{u} + \kappa_2 \boldsymbol{\mu}_2^T \mathbf{v}) \end{aligned} \tag{4.21}$$

Assuming $\boldsymbol{\mu}_1 = \alpha_1 \mathbf{1}$ and $\boldsymbol{\mu}_2 = \alpha_2 \mathbf{1}$, where $\alpha_1, \alpha_2 \in \{-1, 1\}$, we obtain:

$$f_{\mathbf{U},\mathbf{V}}(\mathbf{u},\mathbf{v}) = C_{n-1}(\kappa_1) C_{n-1}(\kappa_2)\exp\left(\mathbf{1}^T(\alpha_1\kappa_1\mathbf{u} + \alpha_2\kappa_2\mathbf{v})\right). \tag{4.22}$$

This yields:

$$\begin{aligned} J(f_{\mathbf{U},\mathbf{V}}) &= |S_{n-2}| \int_{S_{n-2}} C_{n-1}(\kappa_1) C_{n-1}(\kappa_2)\exp\left((\alpha_1\kappa_1 + \alpha_2\kappa_2)\mathbf{1}^T\mathbf{w}\right) d\mathbf{w} \\ &= |S_{n-2}| C_{n-1}(\kappa_1) C_{n-1}(\kappa_2) \int_{S_{n-2}} \exp\left((\alpha_1\kappa_1 + \alpha_2\kappa_2)\mathbf{1}^T\mathbf{w}\right) d\mathbf{w} \\ &= \frac{|S_{n-2}| C_{n-1}(\kappa_1) C_{n-1}(\kappa_2)}{C_{n-1}(|\alpha_1\kappa_1 + \alpha_2\kappa_2|)}. \end{aligned} \tag{4.23}$$

Therefore, using (4.19) and (4.20), $J(f_{\mathbf{U},\mathbf{V}})$ can be computed up to the desired precision.

Further properties as well as intuitive interpretations of $J(f_{\mathbf{U},\mathbf{V}})$ have also been considered in (*Hero and Rajaratnam*, 2011).

For the convergence analysis we assume that the response Y is generated from the

following statistical model:

$$Y = a_{i_1} X_{i_1} + a_{i_2} X_{i_2} + \cdots + a_{i_k} X_{i_k} + N, \tag{4.24}$$

where $\pi_0 = \{i_1, \cdots, i_k\}$ is a set of distinct indices in $\{1, \ldots, p\}$, $\mathbf{X} = [X_1, X_2, \cdots, X_p]$ is the vector of predictors, Y is the response variable, and N is a noise variable. $X_{i_1}, \cdots, X_{i_k}$ are called active variables and the remaining $p - k$ variables are called inactive variables. In the sequel, we refer to the set π_0 as the support set, and $|\pi_0| = k$ denotes the number of active variables.

For the purpose of convergence analysis of SPARCS we impose the following three assumptions on the linear model (4.24).

Assumption IV.1. *The rows of the $n \times p$ data matrix $\mathbb{X}$ are i.i.d. realizations of a p-dimensional vector $\mathbf{X}$ which follows a multivariate elliptically contoured distribution with mean $\boldsymbol{\mu}_x$ and $p \times p$ dispersion matrix $\boldsymbol{\Sigma}_x$, i.e. the probability density function (pdf) is of the form $f_{\mathbf{X}}(\mathbf{x}) = g\left((\mathbf{x} - \boldsymbol{\mu}_x)^T {\boldsymbol{\Sigma}_x}^{-1} (\mathbf{x} - \boldsymbol{\mu}_x)\right)$, where g is a non-negative function. Also, N is statistically independent of $\mathbf{X}$ and follows a univariate elliptically contoured distribution $f_N(.)$ with mean 0 and variance σ_N^2. Moreover, the density functions $f_{\mathbf{X}}(.)$ and $f_N(.)$ are bounded and differentiable.*

Assumption IV.2. *Let ρ_{yi} represent the true correlation coefficient between response variable Y and predictor variable X_i. The quantity*

$$\rho_{\min} = \min_{i \in \pi_0, j \in \{1, \cdots, p\} \setminus \pi_0} \{|\rho_{yi}| - |\rho_{yj}|\}, \tag{4.25}$$

is strictly positive and independent of p.

Assumption IV.3. *The $(n-1) \times p$ matrix of U-scores satisfies:*

$$\frac{n-1}{p} \mathbb{U}^x (\mathbb{U}^x)^T = \mathbf{I}_{n-1} + \boldsymbol{o}(1), \ \text{as } p \to \infty, \tag{4.26}$$

in which $\boldsymbol{o}(1)$ *is a* $(n-1)\times(n-1)$ *matrix whose entries are* $o(1)$.

Assumption IV.2 is a common assumption that one finds in performance analysis of support recovery algorithms (cf. (*Obozinski et al.*, 2011; *Fan and Lv*, 2008)). In particular, Assumption IV.2 can be compared to the conditions on the sparsity-overlap function in (*Obozinski et al.*, 2011) which impose assumptions on the population covariance matrix in relation to the true regression coefficients. Assumption IV.2 can also be compared to Condition 3 introduced in (*Fan and Lv*, 2008) that imposes lower bounds on the magnitudes of the true regression coefficients as well as on the true correlation coefficients between predictors and the response. Assumption IV.3 can be related to assumptions (A1)-(A3) in (*Obozinski et al.*, 2011) in the sense that they both lead to regularity conditions on the entries and the eigenspectrum of the correlation matrix. Assumption IV.3 is also similar the concentration property introduced in (*Fan and Lv*, 2008) as they both yield regularity conditions on the inner products of the rows of the data matrix. Moreover, Assumption IV.3 can also be considered as an incoherence-type condition on the U-scores, similar to the incoherence conditions on the design matrix assumed in the compressive sensing literature (*Candes and Romberg*, 2007; *Tropp and Gilbert*, 2007; *Carin et al.*, 2011). It is worth mentioning that a special case in which Assumption IV.3 is satisfied is the orthogonal setting where $\mathbb{X}\mathbb{X}^T/n = \mathbf{I}_n$.

Lemma IV.4 specifies a class of $p\times p$ correlation matrices $\boldsymbol{\Omega}_x$ for which Assumption IV.3 is satisfied.

Lemma IV.4. *Assume that the population correlation matrix* $\boldsymbol{\Omega}_x = \mathbf{D}_{\boldsymbol{\Sigma}_x}^{-1/2}\boldsymbol{\Sigma}_x\mathbf{D}_{\boldsymbol{\Sigma}_x}^{-1/2}$ *is of the following weakly block-sparse form*

$$\boldsymbol{\Omega}_x = \boldsymbol{\Omega}_{bs} + \boldsymbol{\Omega}_e, \tag{4.27}$$

in which $\boldsymbol{\Omega}_{bs}$ *is a* $p\times p$ *block-sparse matrix of degree* d_x *(i.e., by re-arranging rows and*

columns of $\mathbf{\Omega}_{bs}$ all non-zero off-diagonal entries can be collected in a $d_x \times d_x$ block), and $\mathbf{\Omega}_e = [\omega_{ij}]_{1\leq i,j\leq p}$ is a $p \times p$ matrix such that $\omega_{ij} = O\left(f(|i-j|)\right)$ for some function $f(.)$ with $f(t) = O(|t|^{-\gamma})$ where $\gamma > 1$. If $d_x = o(p)$, then Assumption IV.3 holds.

Proof. By block sparsity of $\mathbf{\Omega}_{bs}, \mathbb{U}^x$ can be partitioned as:

$$\mathbb{U}^x = [\underline{\mathbb{U}}^x, \overline{\mathbb{U}}^x], \tag{4.28}$$

where $\underline{\mathbb{U}}^x = [\underline{\mathbf{U}}^x_1, \cdots, \underline{\mathbf{U}}^x_{d_x}]$ are the U-scores corresponding to the dependent block of $\mathbf{\Omega}_{bs}$ and $\overline{\mathbb{U}}^x = [\overline{\mathbf{U}}^x_1, \cdots, \overline{\mathbf{U}}^x_{p-d_x}]$ are the remaining U-scores. Using relations (4.68) and (4.69) we have:

$$\begin{aligned}\frac{n-1}{p}\mathbb{U}^x(\mathbb{U}^x)^T &= \frac{n-1}{p}\left(\underline{\mathbb{U}}^x(\underline{\mathbb{U}}^x)^T + \overline{\mathbb{U}}^x(\overline{\mathbb{U}}^x)^T\right) \\ &= \mathbf{I}_{n-1} + (n-1)\mathbf{O}(d_x/p).\end{aligned} \tag{4.29}$$

Noting that $d_x = o(p)$ the result follows. □

Using Schur's complement formula it can be shown that if a matrix $\mathbf{\Omega}_x$ is weakly block-sparse of the form (4.27) then its inverse is also weakly block-sparse. It is also worth mentioning that in our high dimensional analysis, a weakly block-sparse matrix asymptotically behaves similar to a block-sparse matrix.

4.3.2 High dimensional convergence rates for screening

In this section, we establish a Poisson-like limit theorem for the number of variables that pass the SPARCS screening stage as $p \to \infty$ for fixed n. This yields a Poisson approximation to the probability of false discoveries that is accurate for small n and large p. The Poisson-like limit theorem also specifies a phase transition threshold for the false discovery probability.

Below we show that both SIS and PCS methods for discovering the support set

are equivalent to discovering the largest entries of some $p \times 1$ vector $\mathbf{\Phi}^{xy}$ having the following representation:

$$\mathbf{\Phi}^{xy} = (\mathbb{Z}^x)^T \mathbb{Z}^y, \tag{4.30}$$

in which $\mathbb{Z}^x$ is a $(n-1) \times p$ matrix whose columns are unit norm vectors, and $\mathbb{Z}^y$ is a $(n-1) \times 1$ unit norm vector.

Using the U-score representation of the correlation matrices, there exist a $(n-1)\times p$ matrix $\mathbb{U}^x$ with unit norm columns, and a $(n-1) \times 1$ unit norm vector $\mathbb{U}^y$ such that (*Hero and Rajaratnam*, 2011, 2012):

$$\mathbf{R}^{xy} = (\mathbb{U}^x)^T \mathbb{U}^y. \tag{4.31}$$

Representation (4.31) immediately shows that SIS is equivalent to discovering non-zero entries of a vector with representation (4.30). Moreover, we have

$$\mathbf{S}^{xy} = \mathbf{D}_{\mathbf{S}^x}^{\frac{1}{2}} (\mathbb{U}^x)^T \mathbb{U}^y (s^y)^{\frac{1}{2}}, \tag{4.32}$$

and:

$$(\mathbf{S}^x)^\dagger = \mathbf{D}_{\mathbf{S}^x}^{-\frac{1}{2}} ((\mathbb{U}^x)^T (\mathbb{U}^x (\mathbb{U}^x)^T)^{-2} \mathbb{U}^x) \mathbf{D}_{\mathbf{S}^x}^{-\frac{1}{2}}, \tag{4.33}$$

where $\mathbf{D}_{\mathbf{A}}$ denotes the diagonal matrix obtained by zeroing out the off-diagonals of square matrix $\mathbf{A}$. We refer the interested reader to (*Hero and Rajaratnam*, 2012; *Anderson*, 2003) for more information about the calculations of U-scores. Using representations (4.32) and (4.33), one can write:

$$\hat{Y} = ((\mathbf{S}^x)^\dagger \mathbf{S}^{xy})^T \mathbf{X} = (s^y)^{\frac{1}{2}} (\mathbb{U}^y)^T (\mathbb{U}^x (\mathbb{U}^x)^T)^{-1} \mathbb{U}^x \mathbf{D}_{\mathbf{S}^x}^{-\frac{1}{2}} \mathbf{X}. \tag{4.34}$$

Defining $\tilde{\mathbb{U}}^x = (\mathbb{U}^x(\mathbb{U}^x)^T)^{-1}\mathbb{U}^x\mathbf{D}^{-\frac{1}{2}}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x}$, we have:

$$\hat{Y} = (s^y)^{\frac{1}{2}}(\mathbb{U}^y)^T\tilde{\mathbb{U}}^x\mathbf{D}^{\frac{1}{2}}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x}\mathbf{D}^{-\frac{1}{2}}_{\mathbf{S}^x}\mathbf{X} \tag{4.35}$$

$$= (s^y)^{\frac{1}{2}}(\mathbf{H}^{xy})^T\mathbf{D}^{\frac{1}{2}}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x}\mathbf{D}^{-\frac{1}{2}}_{\mathbf{S}^x}\mathbf{X}, \tag{4.36}$$

where

$$\mathbf{H}^{xy} = (\tilde{\mathbb{U}}^x)^T\mathbb{U}^y. \tag{4.37}$$

Note that the columns of the matrix $\tilde{\mathbb{U}}^x$ lie on S_{n-2} since the diagonal entries of the $p \times p$ matrix $(\tilde{\mathbb{U}}^x)^T\tilde{\mathbb{U}}^x$ are equal to one. Therefore, a U-score representation of the generalized OLS solution $\mathbf{B}^{xy}$ can be obtained as:

$$\mathbf{B}^{xy} = (\mathbf{S}^x)^\dagger\mathbf{S}^{xy} = \mathbf{D}^{-\frac{1}{2}}_{\mathbf{S}^x}\mathbf{D}^{\frac{1}{2}}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x}\mathbf{H}^{xy}(s^y)^{\frac{1}{2}}, \tag{4.38}$$

Under the condition that $\mathbf{D}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x}$ has non-zero diagonal entries, the i-th entry of $\mathbf{B}^{xy}$ is zero if and only if the i-th entry of $\mathbf{H}^{xy}$ is zero, for $1 \leq i \leq p$. This motivates screening for non-zero entries of the vector $\mathbf{H}^{xy}$ instead of the entries of $\mathbf{B}^{xy}$. In particular, for a threshold $\rho \in [0, 1]$, we can undertake variable selection by discovering the entries of the vector $\mathbf{H}^{xy}$ in (4.37) that have absolute values at least ρ. This implies that discovering the support via PCS is equivalent to discovering the non-zero entries of $\mathbf{H}^{xy}$ in (4.37) which admits the representation (4.30).

Now for a threshold $\rho \in [0, 1]$, let N^{xy}_ρ denote the number of entries of a $p \times 1$ vector of the form (4.30) whose magnitude is at least ρ. The following theorem gives an asymptotic expression for the expected number of discoveries $\mathbb{E}[N^{xy}_\rho]$, for fixed n, as $p \to \infty$ and $\rho \to 1$. It also states that under certain assumptions, the probability of having at least one discovery converges to a given limit. This limit is equal to the probability that a certain Poisson random variable N^* with rate equal

to $\lim_{p\to\infty,\rho\to 1}\mathbb{E}[N_\rho^{xy}]$ satisfies: $N^* > 0$.

Theorem IV.5. *Consider the linear model* (4.24). *Let* $\{\rho_p\}_p$ *be a sequence of threshold values in* $[0,1]$ *such that* $\rho_p \to 1$ *as* $p \to \infty$ *and* $p(1-\rho_p^2)^{(n-2)/2} \to e_n$. *Under the Assumptions IV.1 and IV.3, if the number of active variables* k *grows at a rate slower than* p, *i.e.,* $k = o(p)$, *then for the number of discoveries* $N_{\rho_p}^{xy}$ *we have:*

$$\lim_{p\to\infty}\mathbb{E}[N_{\rho_p}^{xy}] = \lim_{p\to\infty}\xi_{p,n,\rho_p} = \zeta_n, \tag{4.39}$$

where $\xi_{p,n,\rho_p} = pP_0(\rho,n)$ *and* $\zeta_n = e_n \mathrm{a}_n/(n-2)$. *Moreover:*

$$\lim_{p\to\infty}\mathbb{P}(N_{\rho_p}^{xy} > 0) = 1 - \exp(\zeta_n). \tag{4.40}$$

Proof. In order to obtain stronger bounds, we prove the Theorem IV.5 under the weakly block-sparse assumption (4.27). However the proof for the general case where Assumption IV.3 is satisfied follow similarly. The proof follows directly from Theorem IV.6 and Lemma IV.7 presented below. □

It is worth mentioning that Theorem IV.5 can be generalized to the case where Assumption IV.3 is not required. However the asymptotic rates for $\mathbb{E}[N_{\rho_p}^{xy}]$ and $\mathbb{P}(N_{\rho_p}^{xy} > 0)$ depend on the underlying distribution of the data in the case that Assumption IV.3 is not satisfied. The following theorem asserts such a generalization.

Theorem IV.6. *Consider the linear model* (4.24) *for which Assumption IV.1 is satisfied. Let* $\mathbb{U}^x = [\mathbf{U}_1^x, \mathbf{U}_2^x, ..., \mathbf{U}_p^x]$ *and* $\mathbb{U}^y = [\mathbf{U}^y]$ *be* $(n-1)\times p$ *and* $(n-1)\times 1$ *random matrices with unit norm columns. Let* $\{\rho_p\}_p$ *be a sequence of threshold values in* $[0,1]$ *such that* $\rho_p \to 1$ *as* $p \to \infty$ *and* $p(1-\rho_p^2)^{(n-2)/2} \to e_n$. *Throughout this theorem* N_ρ^{xy} *denotes the number of entries of the* $p\times 1$ *vector* $\mathbf{G}^{xy} = (\mathbb{U}^x)^T\mathbb{U}^y$ *whose*

magnitude is at least ρ. We have:

$$\lim_{p\to\infty} \mathbb{E}[N_{\rho_p}^{xy}] = \lim_{p\to\infty} \xi_{p,n,\rho_p} J(\overline{f_{\mathbf{U}_*^x,\mathbf{U}^y}}) = \zeta_n \lim_{p\to\infty} J(\overline{f_{\mathbf{U}_*^x,\mathbf{U}^y}}), \tag{4.41}$$

where $\xi_{p,n,\rho_p} = pP_0(\rho, n)$ and $\zeta_n = e_n \mathrm{a}_n/(n-2)$.

Assume also that $q = o(p)$ and that the limit of average dependency coefficient satisfies $\lim_{p\to\infty} \|\Delta_{p,n,q}^{xy}\|_1 = 0$. Then:

$$\mathbb{P}(N_{\rho_p}^{xy} > 0) \to 1 - \exp(-\Lambda^{xy}), \tag{4.42}$$

with

$$\Lambda^{xy} = \lim_{p\to\infty} \mathbb{E}[N_{\rho_p}^{xy}]. \tag{4.43}$$

Proof. Let d_i^x denote the degree of vertex X_i in part x of the graph $\mathcal{G}_\rho(\mathbf{G}^{xy})$. We have:

$$N_\rho^{xy} = \sum_{i=1}^{p} d_i^x. \tag{4.44}$$

The following representation for d_i^x holds:

$$d_i^x = I(\mathbf{U}^y \in A(r, \mathbf{U}_i^x)), \tag{4.45}$$

where $A(r, \mathbf{U}_i^x)$ is the union of two anti-polar caps in S_{n-2} of radius $\sqrt{2(1-\rho)}$ centered at $\mathbf{U}_i^x$ and $-\mathbf{U}_i^x$. The following inequality will be helpful:

$$\mathbb{E}[d_i^x] = \int_{S_{n-2}} d\mathbf{u} \int_{A(r,\mathbf{u})} d\mathbf{v}\ f_{\mathbf{U}_i^x,\mathbf{U}^y}(\mathbf{u}, \mathbf{v}) \tag{4.46}$$

$$\leq P_0 a_n M_{1|1}^{yx}, \tag{4.47}$$

where $M_{1|1}^{yx} = \max_i \|f_{\mathbf{U}^y|\mathbf{U}_i^x}\|_\infty$, and P_0 is a simplified notation for $P_0(\rho, n)$. Also for $i \neq j$ we have:

$$\mathbb{E}[d_i^x d_j^x] \leq P_0^2 a_n^2 M_{2|1}^{xy}, \tag{4.48}$$

where $M_{2|1}^{xy}$ is a bound on the conditional joint densities of the form $f_{\mathbf{U}_i^x, \mathbf{U}_j^x|\mathbf{U}^y}$.
Application of the mean value theorem to the integral representation (4.46) yields:

$$|\mathbb{E}[d_i^x] - P_0 J(f_{\mathbf{U}_i^x, \mathbf{U}^y})| \leq \tilde{\gamma}^{yx} P_0 r, \tag{4.49}$$

where $\tilde{\gamma}^{yx} = 2a_n^2 \dot{M}_{1|1}^{yx}$ and $\dot{M}_{1|1}^{yx}$ is a bound on the norm of the gradient:

$$\dot{M}_{1|1}^{yx} = \max_i \|\nabla_{\mathbf{U}^y} f_{\mathbf{U}^y|\mathbf{U}_i^x}(\mathbf{u}^y|\mathbf{u}_i^x)\|_\infty. \tag{4.50}$$

Using (4.49) and the relation $r = O\left((1-\rho)^{1/2}\right)$ we conclude:

$$|\mathbb{E}[d_i^x] - P_0 J(\overline{f_{\mathbf{U}_i^x, \mathbf{U}^y}})| \leq O\left(P_0(1-\rho)^{1/2}\right). \tag{4.51}$$

Summing up over i we conclude:

$$\begin{aligned} |\mathbb{E}[N_\rho^{xy}] - \xi_{p,n,\rho} J(\overline{f_{\mathbf{U}_*^x, \mathbf{U}^y}})| &\leq O\left(pP_0(1-\rho)^{1/2}\right) \\ &= O\left(\eta_p^{xy}(1-\rho)^{1/2}\right), \end{aligned} \tag{4.52}$$

where $\eta_p^{xy} = pP_0$. This concludes (4.41).

To prove the second part of the theorem, we use Chen-Stein method (*Arratia et al.*, 1990). Define the index set $B^{xy}(i) = \mathcal{N}_q^{xy}(i) - \{i\}, 1 \leq i \leq p$, where $\mathcal{N}_q^{xy}(i)$ is the set of indices of the q-nearest neighbors of $\mathbf{U}_i^x$. Note that $|B^{xy}(i)| \leq q$. Assume N_ρ^{*xy} is a

Poisson random variable with $\mathbb{E}[N_\rho^{*xy}] = \mathbb{E}[N_\rho^{xy}]$. Using theorem 1 of (*Arratia et al.*, 1990), we have:

$$2 \max_A |\mathbb{P}(N_\rho^{xy} \in A) - \mathbb{P}(N_\rho^{*xy} \in A)| \leq b_1 + b_2 + b_3, \tag{4.53}$$

where:

$$b_1 = \sum_{i=1}^{p} \sum_{j \in B^{xy}(i)} \mathbb{E}[d_i^x]\mathbb{E}[d_j^x], \tag{4.54}$$

$$b_2 = \sum_{i=1}^{p} \sum_{j \in B^{xy}(i)} \mathbb{E}[d_i^x d_j^x], \tag{4.55}$$

and

$$b_3 = \sum_{i=1}^{p} E\left[E\left[d_i^x - \mathbb{E}[d_i^x] | d_j^x : j \in A_q(i)\right]\right], \tag{4.56}$$

where $A_q(i) = (B^{xy}(i))^c - \{i\}$. Using the bound (4.47), $\mathbb{E}[d_i^x]$ is of order $O(P_0)$. Therefore:

$$b_1 \leq O(pkP_0^2) = O((\eta_p^{xy})^2 q/p). \tag{4.57}$$

Since $i \notin B^{xy}(i)$, applying (4.48) to each term of the summation (4.55) gives:

$$b_2 \leq O(pqP_0^2) = O((\eta_p^{xy})^2 q/p). \tag{4.58}$$

Finally, to bound b_3 we have:

$$\begin{aligned} b_3 &= \sum_{i=1}^{p} \mathbb{E}\left[\mathbb{E}\left[d_i^x - \mathbb{E}[d_i^x]|\mathbf{U}_{A_q(i)}\right]\right] \\ &= \sum_{i=1}^{p} \int\limits_{S_{n-2}^{|A_q(i)|}} d\mathbf{u}_{A_q(i)} \int\limits_{S_{n-2}} d\mathbf{u}_i^x \int\limits_{A(r,\mathbf{u}_i^x)} d\mathbf{u}^y \\ &\quad \left(\frac{f_{\mathbf{U}_i^x,\mathbf{U}^y|\mathbf{U}_{A_q(i)}}(\mathbf{u}_i^x,\mathbf{u}^y|\mathbf{u}_{A_q(i)}) - f_{\mathbf{U}_i^x,\mathbf{U}^y}(\mathbf{u}_i^x,\mathbf{u}^y)}{f_{\mathbf{U}_i^x,\mathbf{U}^y}(\mathbf{u}_i^x,\mathbf{u}^y)}\right) \\ &\quad f_{\mathbf{U}_i^x,\mathbf{U}^y}(\mathbf{u}_i^x,\mathbf{u}^y) f_{\mathbf{U}_{A_q(i)}}(\mathbf{u}_{A_q(i)}) \qquad (4.59) \\ &\leq O(pP_0\|\Delta_{p,n,q}^{xy}\|_1) = O(\eta_p^{xy}\|\Delta_{p,n,q}^{xy}\|_1). \end{aligned}$$

Therefore using bound (4.52) we obtain:

$$\begin{aligned} |\mathbb{P}(N_\rho^{xy} > 0) - (1 - \exp(-\Lambda^{xy}))| &\leq \\ |\mathbb{P}(N_\rho^{xy} > 0) - (1 - \exp(-\mathbb{E}[N_\rho^{xy}]))| &+ |\exp(-\mathbb{E}[N_\rho^{xy}]) - \exp(-\Lambda^{xy})| \leq \\ b_1 + b_2 + b_3 &+ O(|\mathbb{E}[N_\rho^{xy}] - \Lambda^{xy}|) \leq \\ b_1 + b_2 + b_3 &+ O\left(\eta_p^{xy}(1-\rho)^{1/2}\right). \qquad (4.60) \end{aligned}$$

Combining this with the bounds on b_1, b_2 and b_3, completes the proof of (4.42). □

Lemma IV.7. *Assume the hypotheses of Theorem IV.5. Assume also that the correlation matrix $\mathbf{\Omega}_x$ is of the weakly block-sparse from (4.27) with $d_x = o(p)$. We have:*

$$\tilde{\mathbb{U}}^x = \mathbb{U}^x(1 + O(d_x/p)). \qquad (4.61)$$

Moreover, the 2-fold average function $J(\overline{f_{\mathbf{U}_*^x,\mathbf{U}^y}})$ and the average dependency coefficient $\|\Delta_{p,n,q}^{xy}\|$ satisfy

$$J(\overline{f_{\mathbf{U}_*^x,\mathbf{U}^y}}) = 1 + O((k + d_x)/p), \qquad (4.62)$$

$$\|\Delta^{xy}_{p,n,q}\|_1 = 0. \tag{4.63}$$

Furthermore,

$$J(\overline{f_{\tilde{\mathbf{U}}^x_*, \mathbf{U}^y}}) = 1 + O(\max\{d_x/p, d_{xy}/p\}) \tag{4.64}$$

$$\|\Delta^{\tilde{x}y}_{p,n,q}\|_1 = O(d_x/p). \tag{4.65}$$

Proof. We have:

$$\tilde{\mathbb{U}}^x = (\mathbb{U}^x(\mathbb{U}^x)^T)^{-1}\mathbb{U}^x\mathbf{D}^{-\frac{1}{2}}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x}. \tag{4.66}$$

By block sparsity of $\mathbf{\Omega}_{bs}, \mathbb{U}^x$ can be partitioned as:

$$\mathbb{U}^x = [\underline{\mathbb{U}}^x, \overline{\mathbb{U}}^x], \tag{4.67}$$

where $\underline{\mathbb{U}}^x = [\underline{\mathbf{U}}^x_1, \cdots, \underline{\mathbf{U}}^x_{d_x}]$ are the U-scores corresponding to the dependent block of $\mathbf{\Omega}_{bs}$ and $\overline{\mathbb{U}}^x = [\overline{\mathbf{U}}^x_1, \cdots, \overline{\mathbf{U}}^x_{p-d_x}]$ are the remaining U-scores.

Using the law of large numbers for a sequence of correlated variables (see, e.g., Example 11.18 in (*Severini*, 2005)) since the off-diagonal entries of $\mathbf{\Omega}_x$ that are not in the dependent block converge to 0 as $|i-j|$ grows, we have

$$\frac{1}{p-d_x}\overline{\mathbb{U}}^x(\overline{\mathbb{U}}^x)^T \to \mathbb{E}[\overline{\mathbf{U}}^x_1(\overline{\mathbf{U}}^x_1)^T] = \frac{1}{n-1}\mathbf{I}_{n-1}. \tag{4.68}$$

Since the entries of $1/d_x\underline{\mathbb{U}}^x(\underline{\mathbb{U}}^x)^T$ are bounded by one, we have:

$$\frac{1}{p}\underline{\mathbb{U}}^x(\underline{\mathbb{U}}^x)^T = \mathbf{O}(d_x/p), \tag{4.69}$$

where $\mathbf{O}(u)$ is an $(n-1)\times(n-1)$ matrix whose entries are $O(u)$. Hence:

$$\begin{aligned}(\mathbb{U}^x(\mathbb{U}^x)^T)^{-1}\mathbb{U}^x &= (\underline{\mathbb{U}}^x\left(\underline{\mathbb{U}}^x\right)^T + \overline{\mathbb{U}}^x(\overline{\mathbb{U}}^x)^T\big)^{-1}\,\mathbb{U}^x \\ &= \frac{n-1}{p}(\mathbf{I}_{n-1}+\mathbf{O}(d_x/p))^{-1}\mathbb{U}^x \\ &= \frac{n-1}{p}\mathbb{U}^x(1+O(d_x/p)). \end{aligned} \tag{4.70}$$

Hence, as $p\to\infty$:

$$\begin{aligned}&(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x = \\ &= (\frac{n-1}{p})^2(\mathbb{U}^x)^T\mathbb{U}^x(1+O(d_x/p)).\end{aligned} \tag{4.71}$$

Thus:

$$\mathbf{D}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x} = \left(\frac{p}{n-1}\mathbf{I}_{n-1}(1+O(d_x/p))\right). \tag{4.72}$$

Combining (4.72) and (4.70) concludes (4.61).

Now we prove relations (4.62)-(4.65). Define the partition $\{1,\ldots,p\}=\mathcal{D}\cup\mathcal{D}^c$ of the index set $\{1,\ldots,p\}$, where $\mathcal{D}=\{i:\mathbf{U}_i^x$ is asymptotically uncorrelated of $\mathbb{U}^y\}$. We have:

$$J(\overline{f_{\mathbf{U}_*^x,\mathbf{U}^y}}) = \frac{1}{4p}\sum_{s,t\in\{-1,1\}}(\sum_{i\in\mathcal{D}}+\sum_{i\in\mathcal{D}^c})J(f_{s\mathbf{U}_i^x,t\mathbf{U}^y}), \tag{4.73}$$

and

$$\|\Delta_{p,n,q}^{xy}\|_1 = \frac{1}{p}(\sum_{i\in\mathcal{D}}+\sum_{i\in\mathcal{D}^c})\Delta_{p,n,q}^{xy}(i). \tag{4.74}$$

But, $J(f_{s\mathbf{U}_i^x,t\mathbf{U}^y})=1$ for $i\in\mathcal{D}$ and $\Delta_{p,n,q}^{xy}(i)=0$ for $1\le i\le p$. Moreover, we have

$|\mathcal{D}^c| \leq d_{xy}$, where $d_{xy} = k + d_x$. Therefore,:

$$J(\overline{f_{\mathbf{U}^x_*,\mathbf{U}^y}}) = 1 + O(d_{xy}/p). \tag{4.75}$$

Moreover, since $\tilde{\mathbb{U}}^x = \mathbb{U}^x\,(1 + O(d_x/p))$, $f_{\tilde{\mathbf{U}}^x_i,\mathbf{U}^y} = f_{\mathbf{U}^x_i,\mathbf{U}^y}\,(1 + O(d_x/p))$. This concludes:

$$J(\overline{f_{\tilde{\mathbf{U}}^x_*,\mathbf{U}^y}}) = 1 + O(\max\{d_x/p, d_{xy}/p\}), \tag{4.76}$$

and

$$\|\Delta^{\tilde{x}y}_{p,n,q}\|_1 = O(d_x/p). \tag{4.77}$$

□

Theorem IV.5 plays an important role in identifying phase transitions and in approximating p-values associated with individual predictor variables. More specifically, under the assumptions of Theorem IV.5:

$$\mathbb{P}(N^{xy}_{\rho_p} > 0) \to 1 - \exp(-\xi_{p,n,\rho_p}) \text{ as } p \to \infty. \tag{4.78}$$

The above limit provides an approach for calculating approximate p-values in the setting where the dimension p is very large. For a threshold $\rho \in [0, 1]$ define $\mathcal{G}_\rho(\mathbf{\Phi}^{xy})$ as the undirected bipartite graph (Fig. 5.1) with parts labeled x and y, and vertices $\{X_1, X_2, ..., X_p\}$ in part x and Y in part y. For $1 \leq i \leq p$, vertices X_i and Y are connected if $|\phi^{xy}_i| > \rho$, where ϕ^{xy}_i is the i-th entry of $\mathbf{\Phi}^{xy}$ defined in (4.30). Denote by d^x_i the degree of vertex X_i in $\mathcal{G}_\rho(\mathbf{\Phi}^{xy})$. Note that $d^x_i \in \{0, 1\}$. For each $1 \leq i \leq p$, denote by $\rho(i)$ the maximum value of the threshold ρ for which $d^x_i = 1$ in $\mathcal{G}_\rho(\mathbf{\Phi}^{xy})$. By this definition, we have $\rho(i) = |\phi^{xy}_i|$. Using Theorem IV.5 the p-value associated

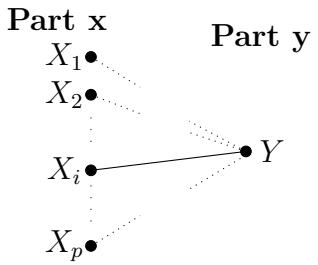

Figure 4.2: The first stage of SPARCS is equivalent to discovering the non-zero entries of the $p \times 1$ vector $\mathbf{\Phi}^{xy}$ in (4.30) to find variables X_i that are most predictive of the response Y. This is equivalent to finding sparsity in a bipartite graph $\mathcal{G}_\rho(\mathbf{\Phi}^{xy})$ with parts x and y which have vertices $\{X_1, \ldots, X_p\}$ and Y, respectively. For $1 \leq i \leq p$, vertex X_i in part x is connected to vertex Y in part y if $|\phi_i^{xy}| > \rho$.

with predictor variable X_i can now be approximated as:

$$pv(i) \approx 1 - \exp(-\xi_{p,n,\rho(i)}). \tag{4.79}$$

Similar to the result in (*Hero and Rajaratnam*, 2011, 2012), there is a phase transition in the p-values as a function of the threshold ρ. More exactly, there is a critical threshold ρ_c such that if $\rho > \rho_c$, the average number $\mathbb{E}[N_\rho^{xy}]$ of discoveries abruptly decreases to 0 and if $\rho < \rho_c$ the average number of discoveries abruptly increases to p. Motivated by this, we define the critical threshold ρ_c as the threshold that satisfies the equation $\partial \mathbb{E}[N_\rho^{xy}]/\partial \rho = -p$. Using (4.39), the value of the critical threshold can be approximated as:

$$\rho_c = \sqrt{1 - (\mathrm{a}_n p)^{-2/(n-4)}}. \tag{4.80}$$

Note that the expression given in (4.80) bears resemblance to the expression (3.14) in (*Hero and Rajaratnam*, 2011). Expression (4.80) is useful in choosing the screening threshold ρ. Selecting ρ slightly greater than ρ_c will prevent the bipartite graph $\mathcal{G}_\rho(\mathbf{\Phi}^{xy})$ from having an overwhelming number of edges.

4.3.3 High dimensional convergence rates for support recovery

In this section we give theoretical upper bounds on the Family-Wise Error Rate (FWER) when performing variable selection in SPARCS screening stage.

Theorems IV.8 and IV.11 give upper bounds on the probability of selection error for the SPARCS screening stage by thresholding the vector $\mathbf{R}^{xy}$ (i.e. using SIS), or the vector $\mathbf{B}^{xy}$ (i.e. using PCS), respectively.

Theorem IV.8. *Under Assumptions IV.1 and IV.2, if $n \geq \Theta(\log p)$ then for any $l \geq k$, SIS recovers the support π_0, with probability at least $1 - 1/p$, i.e.*

$$\mathbb{P}\left(\pi_0 \subseteq S\right) \geq 1 - 1/p. \tag{4.81}$$

Proof. Since $\mathbb{P}\left(\pi_0 \subseteq S\right)$ increases as the size of the recovered set S increases, it suffices to prove the theorem for $l = k$. Define an auxiliary random variable X_{ax} such that $\text{Cor}(Y, X_{\text{ax}}) = \left(\max_{j\in\{1,\cdots,p\}\setminus\pi_0} |\rho_{yj}| + \min_{i\in\pi_0} |\rho_{yi}|\right)/2$. Note that by Assumption IV.2 $\max_{j\in\{1,\cdots,p\}\setminus\pi_0} |\rho_{yj}| < \text{Cor}(Y, X_{\text{ax}}) < \min_{i\in\pi_0} |\rho_{yi}|$. For $l = k$ we have:

$$\begin{aligned}
&\mathbb{P} \quad (\pi_0 \nsubseteq S) = \mathbb{P}\left(\pi_0 \neq S\right) = \\
&\leq \quad \mathbb{P}\left(\bigcup_{i\in\pi_0} \{|r_{yi}| < |\text{SampCor}(Y, X_{\text{ax}})|\} \bigcup_{j\in\{1,\ldots,p\}\setminus\pi_0} \{|r_{yj}| > |\text{SampCor}(Y, X_{\text{ax}})|\}\right) \\
&\leq \quad \sum_{i\in\pi_0} \mathbb{P}\left(|r_{yi}| < |\text{SampCor}(Y, X_{\text{ax}})|\right) + \sum_{j\in\{1,\ldots,p\}\setminus\pi_0} \mathbb{P}\left(|r_{yj}| > |\text{SampCor}(Y, X_{\text{ax}})|\right)
\end{aligned} \tag{4.82}$$

Now since Assumptions IV.1 and IV.2 are satisfied, by Lemma IV.9 there exist constants $C_i > 0, 1 \leq i \leq p$ and a constant N such that

$$\begin{aligned}
&\mathbb{P} \quad (\pi_0 \neq S) \leq \sum_{i\in\pi_0} \exp(-C_i n) + \sum_{j\in\{1,\ldots,p\}\setminus\pi_0} \exp(-C_j n) \\
&\leq \quad p\exp(-C_{\min} n),\ \forall n > N,
\end{aligned} \tag{4.83}$$

in which $C_{\min} = \min_{1 \leq i \leq p} C_i = \rho_{\min}/6$. Hence by letting $C = 2/C_{\min} = 12/\rho_{\min}$ and $n = C \log p$ we have:

$$\mathbb{P}\left(\pi_0 \neq S\right) \leq \frac{1}{p}, \tag{4.84}$$

and

$$\mathbb{P}\left(\pi_0 = S\right) = 1 - \mathbb{P}\left(\pi_0 \neq S\right) \geq 1 - \frac{1}{p}, \tag{4.85}$$

which completes the proof. □

The following lemma was used in the proof of Theorem IV.8.

Lemma IV.9. *Assume Z_1, Z_2 and Z are jointly elliptically contoured distributed random variables from which n joint observations are available. Let $\rho_1 = Cor(Z, Z_1)$ and $\rho_2 = Cor(Z, Z_2)$. Also let $r_1 = SampCor(Z, Z_1)$ and $r_2 = SampCor(Z, Z_2)$, be the corresponding sample correlation coefficients. Assume that $|\rho_1| > |\rho_2|$. Then, there exists $C > 0$ and N such that:*

$$\mathbb{P}\left\{|r_2| > |r_1|\right\} \leq \exp(-Cn), \tag{4.86}$$

for all $n > N$.

Proof. Let $\mathbf{Z} = [Z_2, Z_1, Z]^T$. Assume $\mathbf{Z}$ follows an elliptically contoured density function of the form $f_{\mathbf{Z}}(\mathbf{z}) = |\mathbf{\Sigma}_z|^{-1/2} g\left((\mathbf{z} - \boldsymbol{\mu}_z)^T {\mathbf{\Sigma}_z}^{-1} (\mathbf{z} - \boldsymbol{\mu}_z)\right)$. Without loss of generality assume $\mathrm{Var}(Z_1) = \mathrm{Var}(Z_2) = \mathrm{Var}(Z) = 1$. Using a Cholesky factorization we can represent Z_1, Z_2 and Z as linear combination of uncorrelated random variables W_1, W_2 and W which follow a spherically contoured distribution:

$$\begin{bmatrix} Z_2 \\ Z_1 \\ Z \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ a & b & 0 \\ c & d & e \end{bmatrix} \times \begin{bmatrix} W_2 \\ W_1 \\ W \end{bmatrix} \tag{4.87}$$

where

$$\rho_1 = ac + bd, \tag{4.88}$$

$$\rho_2 = c, \tag{4.89}$$

$$a^2 + b^2 = 1, \tag{4.90}$$

and

$$c^2 + d^2 + e^2 = 1. \tag{4.91}$$

Let $\mathbf{W} = [W_2, W_1, W]^T$. Assume $\mathbf{W}$ follows a spherically contoured density function of the form $f_{\mathbf{W}}(\mathbf{w}) = |\mathbf{\Sigma}_w|^{-1/2} h\left((\mathbf{w} - \boldsymbol{\mu}_w)^T {\mathbf{\Sigma}_w}^{-1} (\mathbf{w} - \boldsymbol{\mu}_w)\right)$. Since $\mathbf{W}$ follows a spherically contoured distribution, it has a stochastic representation of the form $\mathbf{W} = R\mathbf{U}$, where R has a marginal density $f_R(r) = \alpha h(r^2) r^2$, in which α is a normalizing constant. Moreover $\mathbf{U}$ is independent of R and the distribution of $\mathbf{U}$ does not depend on the function h (see, e.g., Chapter 2 in (*Anderson*, 2003) for more details about such stochastic representation). Therefore for the U-score analysis, without loss of generality, we can assume that $\mathbf{W}$ follows a multivariate normal distribution. Now let $\mathbf{U}_1^z, \mathbf{U}_2^z$ and $\mathbf{U}^z$ denote the U-scores corresponding to Z_1, Z_2 and Z, respectively. Similarly, let $\mathbf{U}_1^w, \mathbf{U}_2^w$ and $\mathbf{U}^w$ denote the U-scores corresponding to W_1, W_2 and W,

respectively. Using (4.87) we have the following relations:

$$\begin{aligned}
&\mathbf{U}_2^z = \mathbf{U}_2^w, \\
&\mathbf{U}_1^z = (a\mathbf{U}_2^w + b\mathbf{U}_1^w)/\|a\mathbf{U}_2^w + b\mathbf{U}_1^w\|_2, \\
&\mathbf{U}^z = (c\mathbf{U}_2^w + d\mathbf{U}_1^w + e\mathbf{U}^w)/\|c\mathbf{U}_2^w + d\mathbf{U}_1^w + e\mathbf{U}^w\|_2. \qquad (4.92)
\end{aligned}$$

$$(4.93)$$

Hence

$$\begin{aligned}
r_1 &= (\mathbf{U}^z)^T\mathbf{U}_1^z \\
&= \frac{ac + bd + bc(\mathbf{U}_2^w)^T\mathbf{U}_1^w + ad(\mathbf{U}_1^w)^T\mathbf{U}_2^w + ae(\mathbf{U}^w)^T\mathbf{U}_2^w + be(\mathbf{U}^w)^T\mathbf{U}_1^w}{\|c\mathbf{U}_2^w + d\mathbf{U}_1^w + e\mathbf{U}^w\|_2\|a\mathbf{U}_2^w + b\mathbf{U}_1^w\|_2} \qquad (4.94)
\end{aligned}$$

and

$$r_2 = (\mathbf{U}^z)^T\mathbf{U}_2^z = \frac{c + d(\mathbf{U}_1^w)^T\mathbf{U}_2^w + e(\mathbf{U}^w)^T\mathbf{U}_2^w}{\|c\mathbf{U}_2^w + d\mathbf{U}_1^w + e\mathbf{U}^w\|_2}. \qquad (4.95)$$

Now let $E = \{|r_2| > |r_1|\}$. We have:

$$\begin{aligned}
E &= \left\{|\mathbf{U}^T\mathbf{U}_2| > |\mathbf{U}^T\mathbf{U}_1|\right\} = \\
&\left\{\|a\mathbf{U}_2^w + b\mathbf{U}_1^w\|_2 \left|c + d(\mathbf{U}_1^w)^T\mathbf{U}_2^w + e(\mathbf{U}^w)^T\mathbf{U}_2^w\right|\right. \\
&\left. > \left|ac + bd + bc(\mathbf{U}_2^w)^T\mathbf{U}_1^w + ad(\mathbf{U}_1^w)^T\mathbf{U}_2^w + ae(\mathbf{U}^w)^T\mathbf{U}_2^w + be(\mathbf{U}^w)^T\mathbf{U}_1^w\right|\right\}. \qquad (4.96)
\end{aligned}$$

Since

$$\begin{aligned}
\|a\mathbf{U}_2^w + b\mathbf{U}_1^w\|_2 &= \sqrt{(a\mathbf{U}_2^w + b\mathbf{U}_1^w)^T(a\mathbf{U}_2^w + b\mathbf{U}_1^w)} = \sqrt{a^2 + b^2 + 2ab(\mathbf{U}_2^w)^T\mathbf{U}_1^w} \\
&= \sqrt{1 + 2ab(\mathbf{U}_2^w)^T\mathbf{U}_1^w} \leq 1 + 2|ab|.|(\mathbf{U}_2^w)^T\mathbf{U}_1^w|, \qquad (4.97)
\end{aligned}$$

and, by using triangle inequality, we have

$$
\begin{aligned}
&E \subseteq \big\{2|abc|.|(\mathbf{U}_2^w)^T\mathbf{U}_1^w|^2 + 2|e|.|(\mathbf{U}^w)^T\mathbf{U}_2^w|.|(\mathbf{U}_2^w)^T\mathbf{U}_1^w| + |ad+bc|.|(\mathbf{U}_2^w)^T\mathbf{U}_1^w| + \\
&|ae|.|(\mathbf{U}^w)^T\mathbf{U}_1^w| + |be|.|(\mathbf{U}^w)^T\mathbf{U}_1^w| > |ac+bd| - |c|\ \} \subseteq \\
&\{2|abc|.|(\mathbf{U}_2^w)^T\mathbf{U}_1^w|^2 > |ac+bd| - |c|\} \cup \{2|e|.|(\mathbf{U}^w)^T\mathbf{U}_2^w|.|(\mathbf{U}_2^w)^T\mathbf{U}_1^w| > |ac+bd| - |c|\} \\
&\cup\{|ad+bc|.|(\mathbf{U}_2^w)^T\mathbf{U}_1^w| > |ac+bd| - |c|\} \cup \\
&\{|ae|.|(\mathbf{U}^w)^T\mathbf{U}_1^w| > |ac+bd| - |c|\} \cup \{|be|.|(\mathbf{U}^w)^T\mathbf{U}_1^w| > |ac+bd| - |c|\} \subseteq \\
&\{|(\mathbf{U}_2^w)^T\mathbf{U}_1^w| > (|ac+bd| - |c|)/2|abc|\} \cup \{|(\mathbf{U}_2^w)^T\mathbf{U}_1^w| > (|ac+bd| - |c|)/2|e|\} \cup \\
&\{|(\mathbf{U}_2^w)^T\mathbf{U}_1^w| > (|ac+bd| - |c|)/|ad+bc|\} \cup \\
&\{|(\mathbf{U}^w)^T\mathbf{U}_1^w| > (|ac+bd| - |c|)/|ae|\} \cup \{|(\mathbf{U}^w)^T\mathbf{U}_1^w| > (|ac+bd| - |c|)/|be|\}. \qquad (4.98)
\end{aligned}
$$

Note that by assumption $|ac+bd| = |\rho_1| > |\rho_2| = |c|$. Now by Lemma IV.10 we get

$$\mathbb{P}(E) \leq 5\exp(-\alpha n), \qquad (4.99)$$

with

$$\alpha = \frac{|ac+bd| - |c|}{\max\{2|abc|, 2|e|, |ad+bc|, |ae|, |be|\}} \geq \frac{\rho_1 - \rho_2}{2}, \qquad (4.100)$$

where the last inequality is obtained via equations (4.88)-(4.91). Letting $C = (\rho_1 - \rho_2)/3$ and $N = 12/(\rho_1 - \rho_2)$ we have

$$\mathbb{P}(E) = \mathbb{P}\{|r_2| > |r_1|\} \leq \exp(-Cn), \qquad (4.101)$$

for $n > N$. □

The following lemma was used in the proof of Lemma IV.9.

Lemma IV.10. *Let* $\mathbf{U}$ *and* $\mathbf{V}$ *be two independent uniformly distributed random vec-*

tors on S_{n-2}. For any fixed $\epsilon > 0$, there exists $C > 0$ such that:

$$\mathbb{P}\{|\mathbf{U}^T\mathbf{V}| > \epsilon\} \leq \exp(-Cn). \tag{4.102}$$

Proof. Without loss of generality assume $U = [1, 0, \ldots, 0]^T$. We have

$$\{|\mathbf{U}_2^T\mathbf{U}_1| > \epsilon\} = \{|v_1| > \epsilon\}, \tag{4.103}$$

in which v_1 is the first entry of the vector $\mathbf{V}$. Using the formula for the area of spherical cap (*Li*, 2011) we obtain

$$\mathbb{P}\{|\mathbf{U}_2^T\mathbf{U}_1| > \epsilon\} = I_\lambda(n/2, 1/2), \tag{4.104}$$

where $\lambda = 1 - \epsilon^2$, and

$$I_x(a, b) = \frac{\int_0^x t^{a-1}(1-t)^{b-1}dt}{\int_0^1 t^{a-1}(1-t)^{b-1}dt} \tag{4.105}$$

is the regularized incomplete beta function. Note that:

$$\begin{aligned} 1/I_\lambda(n/2, 1/2) = \frac{\int_0^\lambda t^{(n-2)/2}/\sqrt{1-t}dt + \int_\lambda^1 t^{(n-2)/2}/\sqrt{1-t}dt}{\int_0^\lambda t^{(n-2)/2}/\sqrt{1-t}dt} = \\ 1 + \frac{\int_\lambda^1 t^{(n-2)/2}/\sqrt{1-t}dt}{\int_0^\lambda t^{(n-2)/2}/\sqrt{1-t}dt} \geq 1 + \frac{\int_\lambda^1 t^{(n-2)/2}/\sqrt{1-\lambda}dt}{\int_0^\lambda t^{(n-2)/2}/\sqrt{1-\lambda}dt} = \\ 1 + \frac{1-\lambda^{n/2}}{\lambda^{n/2}} = (\sqrt{\lambda})^n. \end{aligned} \tag{4.106}$$

Therefore by letting $C = -\frac{1}{2}\log(\lambda) = -\frac{1}{2}\log(1-\epsilon^2)$ we obtain

$$\mathbb{P}\{|\mathbf{U}_2^T\mathbf{U}_1| > \epsilon\} \leq \exp(-Cn). \tag{4.107}$$

□

Theorem IV.11. *Under Assumptions IV.1-IV.3, if* $n \geq \Theta(\log p)$ *then for any* $l \geq k$*, PCS recovers the support* π_0*, with probability at least* $1 - 1/p$*, i.e.*

$$\mathbb{P}\left(\pi_0 \subseteq S\right) \geq 1 - 1/p. \tag{4.108}$$

Proof. By Assumption IV.3 we have

$$\mathbb{U}^x(\mathbb{U}^x)^T = \frac{p}{n-1}\left(\mathbf{I}_{n-1} + \mathbf{o}(1)\right). \tag{4.109}$$

Therefore:

$$\left(\mathbb{U}^x(\mathbb{U}^x)^T\right)^{-1} = \frac{n-1}{p}\left(\mathbf{I}_{n-1} + \mathbf{o}(1)\right). \tag{4.110}$$

Since columns of $\mathbb{U}^x$ have unit norm we obtain:

$$(\mathbb{U}^x(\mathbb{U}^x)^T)^{-1}\mathbb{U}^x = \frac{n-1}{p}\mathbb{U}^x(1 + o(1)), \tag{4.111}$$

and

$$(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x = (\frac{n-1}{p})^2(\mathbb{U}^x)^T\mathbb{U}^x(1 + o(1)). \tag{4.112}$$

This yields

$$\mathbf{D}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x} = (\frac{n-1}{p})^2\mathbf{I}_p(1 + o(1)), \tag{4.113}$$

which implies

$$\tilde{\mathbb{U}}^x = (\mathbb{U}^x(\mathbb{U}^x)^T)^{-1}\mathbb{U}^x\mathbf{D}^{-\frac{1}{2}}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x} = \mathbb{U}^x(1 + o(1)). \tag{4.114}$$

Therefore screening the entries of $\mathbf{B}^{xy}$ or $\mathbf{H}^{xy}$ is asymptotically equivalent to select-

ing the support via thresholding the entries of $(\mathbb{U}^x)^T\mathbb{U}^y$, i.e., the sample correlation coefficients. Therefore the proof follows from Theorem IV.8. □

The constant in $\Theta(\log p)$ of Theorem IV.8 and Theorem IV.11 is increasing in $\rho_{\min}$. It is shown in the proof of the theorems that $12/\rho_{\min}$ is an upper bound for the constant in $\Theta(\log p)$. Note that these theorems on support recovery allow all types of non-zero correlations (i.e., correlations between active variables, correlations between inactive variables, and correlations between active and inactive variables) as long as the corresponding assumptions are satisfied.

Theorems IV.8 and IV.11 can be compared to Thm. 2 in (*Obozinski et al.*, 2011) and Thm. 1 in (*Fan and Lv*, 2008) for recovering the support set π_0. More specifically, Thm. 2 in (*Obozinski et al.*, 2011) asserts a similar result as in Theorem IV.8 and Theorem IV.11 for support recovery via minimizing a LASSO-type objective function. Also Thm. 1 in (*Fan and Lv*, 2008) asserts that if $n = \Theta((\log p)^\alpha)$ for some $\alpha > 1$, SIS recovers the true support with probability no less than $1 - 1/p$. Note also that Theorem IV.8 and Theorem IV.11 state stronger results than the similar results proven in (*Fan and Lv*, 2008) and in (*Obozinski et al.*, 2011), respectively, in the sense that the support recovery guarantees presented in (*Fan and Lv*, 2008; *Obozinski et al.*, 2011) are proven for the class of multivariate Gaussian distributions whereas Theorem IV.8 and Theorem IV.11 consider the larger class of multivariate elliptically contoured distributions.

4.3.4 High dimensional convergence rates for prediction

The following theorem states the optimal sample allocation rule for the two-stage SPARCS predictor, in order to minimize the expected MSE as $t \to \infty$.

Theorem IV.12. *The optimal sample allocation rule for the SPARCS online proce-*

dure introduced in Sec. 4.2 under the cost condition (4.1) *is*

$$n = \begin{cases} O(\log t), & c(p-k)\log t + kt \leq \mu \\ 0, & o.w. \end{cases} \tag{4.115}$$

where c is a positive constant that is independent of p.

Proof. First we consider a two-stage predictor similar to the one introduced in Sec. 4.2 with the difference that the n samples which are used in stage 1 are not used in stage 2. Therefore, there are n and $t-n$ samples used in the first and the second stages, respectively. We represent this two-stage predictor by $n|(t-n)$. Similarly, $n|t$ denotes the SPARCS algorithm which uses n samples at the first stage and all of the t samples at the second stage. The asymptotic results for the $n|(t-n)$ two-stage predictor will be shown to hold as well for the $n|t$ two-stage predictor.

Using inequalities of the form (5.24) and the union bound, it is straightforward to see that for any subset $\pi \neq \pi_0$ of k elements of $\{1, \cdots, p\}$, the probability that π is the outcome of variable selection via SPARCS, is bounded above by pc_π^n, in which $0 < c_\pi < 1$ is a constant that is bounded above by $\exp(-C_{\min})$. The expected MSE of the $n|(t-n)$ algorithm can be written as:

$$\mathbb{E}[\text{MSE}] = \sum_{\pi \in S_k^p, \pi \neq \pi_0} \mathbb{P}(\pi)\mathbb{E}[\text{MSE}_\pi] + \mathbb{P}(\pi_0)\mathbb{E}[\text{MSE}_{\pi_0}], \tag{4.116}$$

where S_k^p is the set of all k-subsets of $\{1, \cdots, p\}$, $\mathbb{P}(\pi)$ is the probability that the outcome of variable selection via SPARCS is the subset π, and MSE_π is the MSE of OLS stage when the indices of the selected variables are the elements of π. Therefore the expected MSE is upper bounded as below:

$$\mathbb{E}[\text{MSE}] \leq (1 - pc_0^n)\mathbb{E}[\text{MSE}_\pi] + p \sum_{\pi \in S_k^p, \pi \neq \pi_0} c_\pi^n \mathbb{E}[\text{MSE}_\pi], \tag{4.117}$$

where c_0 is a constant which is upper bounded by $\exp(-C_{\min})$. It can be shown that if there is at least one wrong variable selected ($\pi \neq \pi_0$), the OLS estimator is biased and the expected MSE converges to a positive constant M_π as $(t-n) \to \infty$. When all the variables are selected correctly (subset π_0), MSE goes to zero with rate $O(1/(t-n))$. Hence:

$$\begin{aligned} \mathbb{E}[\text{MSE}] &\leq (1 - pc_0^n)O(1/(t-n)) + p \sum_{\pi \in S_k^p, \pi \neq \pi_0} c_\pi^n M_\pi \\ &\leq (1 - pc_0^n)C_2/(t-n) + p^{k+1} C_1 C^n, \end{aligned} \tag{4.118}$$

where C, C_1 and C_2 are constants that do not depend on n or p but depend on the quantities $\sum_{j \in \pi_0} a_j^2$ and $\min_{j \in \pi_0} |a_j| / \sum_{l \in \pi_0} |a_l|$. Note that $C = \max_{\pi \in S_k^p, \pi \neq \pi_0} c_\pi \leq \exp(-C_{\min})$. This quantity is an increasing function $\rho_{\min}$.

On the other hand since at most t variables could be used in OLS stage, the expected MSE is lower bounded:

$$\mathbb{E}[\text{MSE}] \geq \Theta(1/t). \tag{4.119}$$

It can be seen that the minimum of (4.118) as a function of n, subject to the constraint (4.1), happens for $n = O(\log t)$ if $c \log t \leq \frac{\mu - tk}{p-k}$ with $c = -1/\log C$ (therefore, similar to C, c is increasing in $\rho_{\min}$); otherwise it happens for 0. If $\Theta(\log t) \leq \frac{\mu - tk}{p-k}$, the minimum value attained by the upper bound (4.118) is $\Theta(1/t)$ which is as low as the lower bound (4.119). This shows that for large t, the optimal number of samples that should be assigned to the SPARCS stage of the $n|(t-n)$ predictor is $n = O(\log t)$. As $t \to \infty$, since $n = O(\log t)$, the MSE of the $n|t$ predictor proposed in Sec. 4.2 converges to the MSE of the $n|(t-n)$ predictor. Therefore, as $t \to \infty$, $n = O(\log t)$ becomes optimal for the $n|t$ predictor as well. □

The constant c above is an increasing function of the quantity $\rho_{\min}$ defined in

(4.25). Theorem IV.12 implies that for a generous budget (μ large) the optimal first stage sampling allocation is $O(\log t)$. However, when the budget is tight it is better to skip stage 1 ($n = 0$). Figure 4.3 illustrates the allocation region (for $c = 1$) as a function of the sparsity coefficient $\rho = 1 - k/p$. Note that Theorem IV.12 is generally true for any two-stage predictor which at the first stage, uses a support recovery method that satisfies the performance bound proposed by Theorem IV.8 or Theorem IV.11, and at the second stage uses OLS.

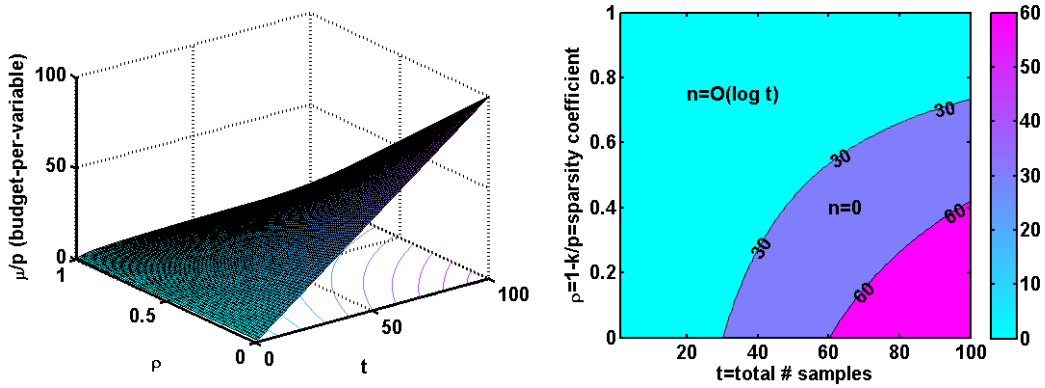

Figure 4.3: (Left) surface $\mu/p = c\rho \log t + (1-\rho)t$, for $c = 1$. (Right) contours indicating optimal allocation regions for $\mu/p = 30$ and $\mu/p = 60$ ($\rho = 1 - k/p$). As the coefficient c increases, the surface $c\rho \log t + (1-\rho)t$ moves upward and the regions corresponding to $n = O(\log t)$ and $n = 0$, become smaller and larger, respectively.

4.4 Numerical comparisons

We now present experimental results which demonstrate the performance of SPARCS when applied to both synthetic and real world data. Throughout this section we refer to the SPARCS predictors which use SIS or PCS at the first stage as SIS-SPARCS or PCS-SPARCS, respectively.

a) Efficiency of SPARCS screening stage. We illustrate the performance of the SPARCS screening stage (i.e., the first stage of the SPARCS predictor) using SIS or PCS and compare to LASSO (*Tibshirani*, 1996; *Genovese et al.*, 2012).

In the first set of simulations we generated an $n \times p$ data matrix $\mathbb{X}$ with independent rows, each of which is drawn from a p-dimensional multivariate normal distribution with mean $\mathbf{0}$ and block-sparse covariance matrix satisfying (4.27). The $p \times 1$ coefficient vector $\mathbf{a}$ is then generated such that exactly 100 entries of $\mathbf{a} \in \mathbb{R}^p$ are active. Each active entry of $\mathbf{a}$ is an independent draw from $\mathcal{N}(0,1)$ distribution, and each inactive entry of $\mathbf{a}$ is zero. Finally, a synthetic response vector $\mathbb{Y}$ is generated by a simple linear model

$$\mathbb{Y} = \mathbb{X}\mathbf{a} + \mathbb{N}, \tag{4.120}$$

where $\mathbb{N}$ is $n \times 1$ noise vector whose entries are i.i.d. $\mathcal{N}(0, 0.05)$. The importance of a variable is measured by the magnitude of the corresponding entry of $\mathbf{a}$.

We implemented LASSO on the above data set using an active set type algorithm - asserted to be one the fastest methods for solving LASSO (*Kim and Park*, 2010). In all of our implementations of LASSO, the regularization parameter is tuned to minimize prediction MSE using 2-fold cross validation. To illustrate SPARCS screening stage for a truly high dimensional example, we set $p = 10000$ and compared SIS and PCS methods with LASSO, for a small number of samples. Figure 4.4 shows the results of this simulation over an average of 400 independent experiments for each value of n. As we see for small number of samples, PCS and SIS methods perform significantly better in selecting the important predictor variables. Moreover, the advantage of the extra pseudo-inverse factor used for variable selection in PCS as compared to SIS is evident in Fig. 4.4.

b) Efficiency of the SPARCS predictor. To test the efficiency of the proposed SPARCS predictor, a total of t samples are generated using the linear model (4.120) from which $n = 25 \log t$ are used for the task of variable selection at the first stage. All t samples are then used to compute the OLS estimator restricted to the selected

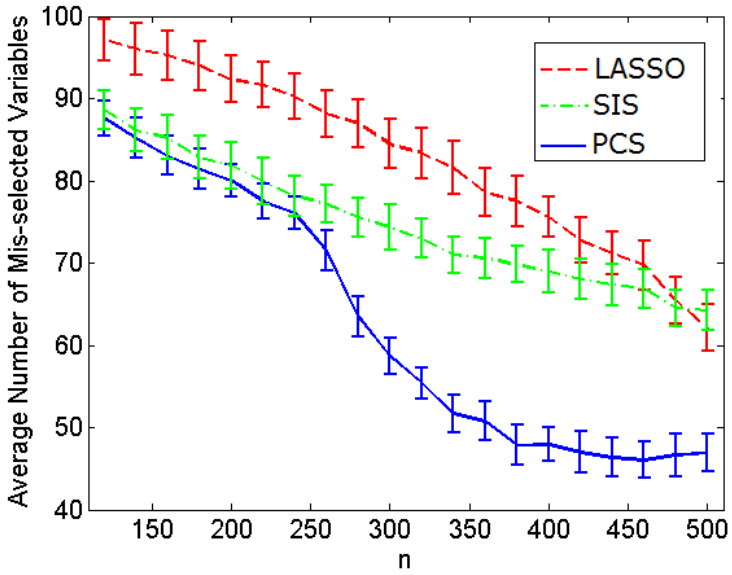

Figure 4.4: Average number of mis-selected variables. Active set implementation of LASSO (red-dashed) vs. SIS (green-dashed) vs. PCS (solid), $p = 10000$. The data is generated via model (4.120). The regularization parameter of LASSO is set using 2-fold cross validation. It is evident that PCS has a lower miss-selection error compared to SIS and LASSO.

variables. We chose t such that $n = (130 : 10 : 200)$. The performance is evaluated by the empirical Root Mean Squared Error

$$\text{RMSE} = \sqrt{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2/m}, \tag{4.121}$$

where m is the number of simulation trials. Similar to the previous experiment, exactly 100 entries of $\mathbf{a}$ are active and the predictor variables follow a multivariate normal distribution with mean $\mathbf{0}$ and block-sparse covariance matrix. Figure 4.5 shows the result of this simulation for $p = 10000$, in terms of performance (left) and running time (right). Each point on these plots is an average of 1000 independent experiments. Observe that in this low sample regime, when LASSO or SIS are used instead of PCS in the first stage, the performance suffers. More specifically we observe that the RMSE of the PCS-SPARCS predictor is uniformly lower than

n	130	140	150	160	170	180	190	200
PCS-SPARCS vs. SIS-SPARCS	7.7×10^{-3}	6.7×10^{-09}	3.2×10^{-11}	2.4×10^{-22}	7.8×10^{-29}	8.1×10^{-36}	9.2×10^{-42}	5.3×10^{-46}
PCS-SPARCS vs. LASSO	3.1×10^{-4}	8.0×10^{-10}	7.2×10^{-14}	3.0×10^{-25}	1.8×10^{-30}	5.6×10^{-39}	1.1×10^{-42}	6.5×10^{-48}

Table 4.1: p-values of the one-sided paired t-test for testing the null hypothesis $\mathcal{H}_0$: PCS-SPARCS and SIS-SPARCS (LASSO) have the same average prediction RMSE in the experiment corresponding to Fig 4.5. Small p-values suggest that PCS-SPARCS significantly outperforms the others.

the SIS-SPARCS predictor or the two-stage predictor that uses LASSO in the first stage. Table 4.1 shows the p-values of the one-sided paired t-tests performed for each value of n, testing the null hypothesis $\mathcal{H}_0$: RMSE for PCS-SPARCS and SIS-SPARCS (LASSO) have the same average in the experiment corresponding to Fig. 4.5. Small p-values confirm that the null hypothesis is rejected.

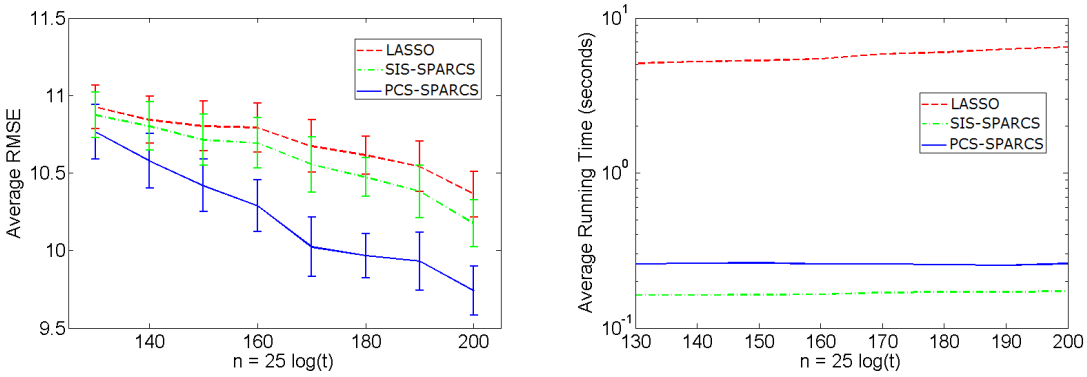

Figure 4.5: (Left) Prediction RMSE for the two-stage predictor when $n = 25 \log t$ samples are used for screening at the first stage and all t samples are used for computing the OLS estimator coefficients at the second stage. The solid plot shows the RMSE for PCS-SPARCS while the green and red dashed plots show the RMSE for SIS-SPARCS and LASSO, respectively. Here, $p = 10000$. The Oracle OLS (not shown), which is the OLS predictor constructed on the true support set, has average RMSE performance that is a factor of 2 lower than the curves shown in the figure. This is due to the relatively small sample size available to these algorithms. (Right) Average running time as a function of n for the experiment of the plot on the left. It is evident that due to lower computational complexity, SIS-SPARCS and PCS-SPARCS run an order of magnitude faster than LASSO.

To further indicate the advantage of the PCS-SPARCS predictor compared to the SIS-SPARCS predictor, we performed simulations in which the number of samples

used at the first stage, $n = 500$, and the number of samples used at the second stage, $t = 2000$, are fixed while the number of variables p increases from $p = 1000$ to $p = 100000$. Moreover, exactly 100 entries of the coefficient vector $\mathbf{a}$ are active. Similar to the previous experiments, samples are generated using the linear model (4.120). However, in order to generate a data set with high multicollinearity, a situation that is likely to happen in high dimensional data sets (see (*Rajaratnam et al.*, 2014) and the references therein), here the inactive variables are consecutive samples of an Auto-Regressive (AR) process of the form:

$$\begin{aligned} W(1) &= \epsilon(1), \\ W(i) &= \phi W(i-1) + \epsilon(i),\ i = 2, \ldots, p - 100, \end{aligned} \tag{4.122}$$

in which $\epsilon(i)$'s are independent draws of $\mathcal{N}(0, 1)$. The result of this experiment for $\phi = 0.99$ is shown in Fig. 4.6 (Left). The average RMSE values are computed using 1000 independent experiments. The advantage of using PCS-SPARCS over SIS-SPARCS is evident in Fig. 4.6 (Left). Note that as the number of variables p becomes significantly larger than the number of samples n, the performance of both of the predictors converge to the performance of a random selection and estimation scheme in which variables are selected at random in the first stage.

Furthermore, to analyze the performance of PCS-SPARCS and SIS-SPARCS for different levels of multicollinearity in the data, we performed similar experiments for $p = [1000, 5000, 10000]$ as the value of ϕ increases from 0.9 to 0.999. Figure 4.6 (Right) shows the result of this simulation. Each point on these plots is the average of 500 independent experiments. It is evident that similar to the previous experiment, the PCS-SPARCS predictor outperforms the SIS-SPARCS predictor. An interesting observation in Fig 4.6 (Right) is that as the multicollinearity coefficient $-\log_{10}(1-\phi)$ increases the performance of the PCS-SPARCS predictor improves.

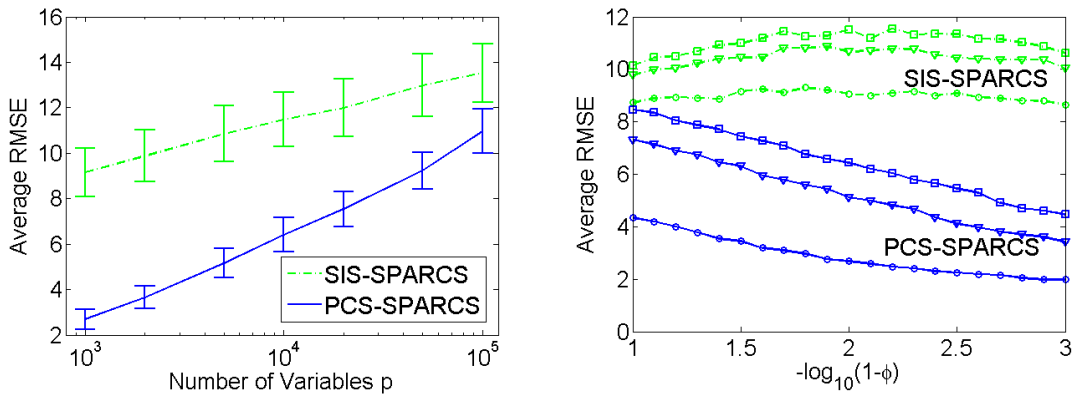

Figure 4.6: (Left) Prediction RMSE for the two-stage predictor when $n = 500$ samples are used at the first stage, and a total of $t = 2000$ samples are used at the second stage. The number of variables varies from $p = 1000$ to $p = 100000$. In this experiment, inactive variables are generated via realizations of an Auto-Regressive process of the form (4.122) with $\phi = 0.99$ ($-\log_{10}(1-\phi) = 2$). The solid and dashed plots show the RMSE for PCS-SPARCS and SIS-SPARCS, respectively. The plots show the advantage of using PCS instead of SIS at the SPARCS screening stage. (Right) Prediction RMSE as function of the multicollinearity coefficient $-\log_{10}(1-\phi)$ for $p = [1000, 5000, 10000]$. For both PCS-SPARCS (solid) and SIS-SPARCS (dashed) predictors, the plots with square, triangle and circle markers correspond to $p = 10000, p = 5000$ and $p = 1000$, respectively. These plots show that the PCS-SPARCS predictor uniformly outperforms the SIS-SPARCS predictor. Observe also that as the multicollinearity coefficient $-\log_{10}(1-\phi)$ increases the performance of the PCS-SPARCS predictor improves.

c) Estimation of FWER using Monte Carlo simulation. We set $p = 1000, k = 10$ and $n = (100 : 100 : 1000)$ and using Monte Carlo simulation, we computed the probability of support recovery error for the PCS method. In order to prevent the coefficients $a_j, j \in \pi_0$ from getting close to zero, the active coefficients were generated via a Bernoulli-Gaussian distribution of the form:

$$a \sim 0.5\mathcal{N}(1, \sigma^2) + 0.5\mathcal{N}(-1, \sigma^2), \tag{4.123}$$

Figure 4.7 shows the estimated probabilities. Each point of the plot is an average of $N = 10^4$ experiments. As the value of σ decreases the quantity $\rho_{\min}$ defined in

(4.25) is bounded away from 0 with high probability and the probability of selection error degrades. As we can see, the FWER decreases at least exponentially with the number of samples. This behavior is consistent with the result of Theorem IV.11.

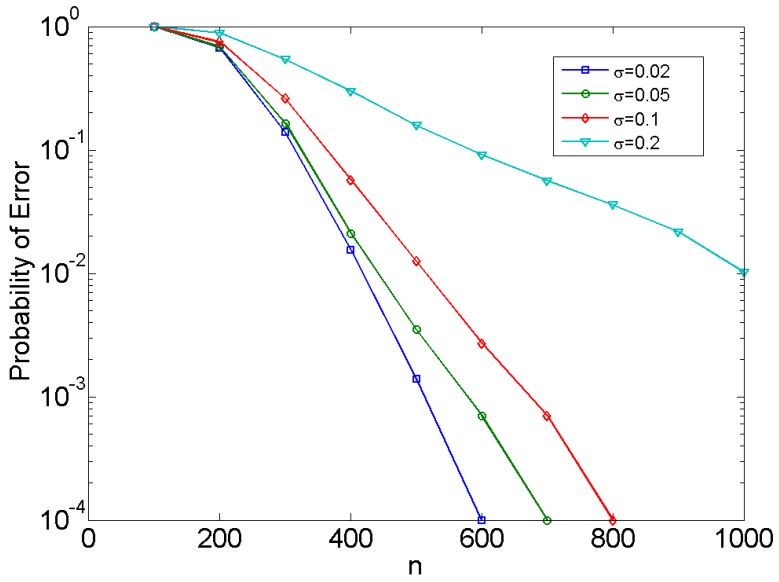

Figure 4.7: Probability of selection error as a function of number of samples for PCS. Probability of selection error is calculated as the ratio of the number of experiments in which the exact support is not recovered over the total number of experiments. The entries of the coefficient matrix are i.i.d. draws from distribution (4.123). Observe that the probability of selection error decreases at least exponentially with the number of samples. This behavior is consistent with Theorem IV.11.

d) Application to experimental data. We illustrate the proposed SPARCS predictor on the Predictive Health and Disease data set, which consists of gene expression levels and symptom scores of 38 different subjects. The data was collected during a challenge study for which some subjects become symptomatically ill with the H3N2 flu virus (*Huang et al.*, 2011). For each subject, the gene expression levels (for $p = 12023$ genes) and the clinical symptoms have been recorded at a large number of time points that include pre-inoculation and post-inoculation sample times. Ten different symptom scores were measured. Each symptom score takes an integer value from 0 to 4, which measures the severity of that symptom at the corresponding time. The goal here is to

learn a predictor that can accurately predict the future symptom scores of a subject based on her last measured gene expression levels.

We considered each symptom as a scalar response variable and applied the SPARCS predictor to each symptom separately. In order to do the prediction task, the data used for the SPARCS predictor consists of the samples of the symptom scores for various subjects at 4 specified time points (t_1, t_2, t_3, t_4) and their corresponding gene expression levels measured at the previous time points $(t_1 - 1, t_2 - 1, t_3 - 1, t_4 - 1)$. The number of predictor variables (genes) selected in the first stage is restricted to 100. Since, the symptom scores take integer values, the second stage uses multinomial logistic regression instead of the OLS predictor. Maximum likelihood estimation is used for computing the multinomial logistic regression coefficients (*Albert and Anderson*, 1984). The performance is evaluated by leave-one-out cross validation. To do this, the data from all except one subject are used as training samples and the data from the remaining subject are used as the test samples. The final RMSE is then computed as the average over the 38 different leave-one-out cross validation trials. In each of the experiments 18 out of the 37 subjects of the training set, are used in first stage and all of the 37 subjects are used in the second stage. It is notable that PCS-SPARCS performs better in predicting the symptom scores for 7 of the 10 symptoms whereas SIS-SPARCS and LASSO perform better in predicting the symptom scores for 2 symptoms and 1 symptom, respectively.

Simulations for the case of multidimensional response. The generalization of PCS to the case where the response $\mathbf{Y}$ is a q-dimensional vector is presented in the Appendix A. Below we briefly present the simulations that was perfomed for the case of $q > 1$.

We set the number of regressor and response variables to $p = 200$ and $q = 20$, respectively, while the number of samples n was varied from 4 to 50. The training data is generated from the multidimensional version of the model (4.120). Figure 4.8

Symptom	RMSE: LASSO	RMSE: SIS-SPARCS	RMSE: PCS-SPARCS
Runny Nose	0.7182	0.6896	**0.6559**
Stuffy Nose	0.9242	**0.7787**	0.8383
Sneezing	0.7453	0.6201	**0.6037**
Sore Throat	0.8235	0.7202	**0.5965**
Earache	**0.2896**	0.3226	0.3226
Malaise	1.0009	**0.7566**	0.9125
Cough	0.5879	0.7505	**0.5564**
Shortness of Breath	0.4361	0.5206	**0.4022**
Headache	0.7896	0.7500	**0.6671**
Myalgia	0.6372	0.5539	**0.4610**
Average for all symptoms	0.6953	0.6463	**0.6016**

Table 4.2: RMSE of the two-stage LASSO predictor, the SIS-SPARCS predictor and the PCS-SPARCS predictor used for symptom score prediction. The data come from a challenge study experiment that collected gene expression and symptom data from human subjects (*Huang et al.*, 2011). Leave-one-out cross validation is used to compute the RMSE values.

shows the average number of mis-selected variables for both methods, as a function of n. The plot is computed by averaging the results of 400 independent experiments for each value of n. Figure 4.9 shows the average run time on a logarithmic scale, as a function of n (MATLAB version 7.14 running on 2.80GHz CPU). As we see, for low number of samples, PCS has better performance than LASSO and is significantly faster.

4.5 Conclusion

We proposed an online procedure for budget-limited predictor design in high dimension called two-stage Sampling, Prediction and Adaptive Regression via Correlation Screening (SPARCS). SPARCS is specifically useful in cases where $n \ll p$ and the high cost of assaying all predictor variables justifies a two-stage design: high throughput variable selection followed by predictor construction using fewer selected variables. We also proposed theories for high dimensional false discovery rates, support recovery guarantees, and optimal stage-wise sample allocation rule associated with the SPARCS online procedure. Simulation and experimental results showed advantages

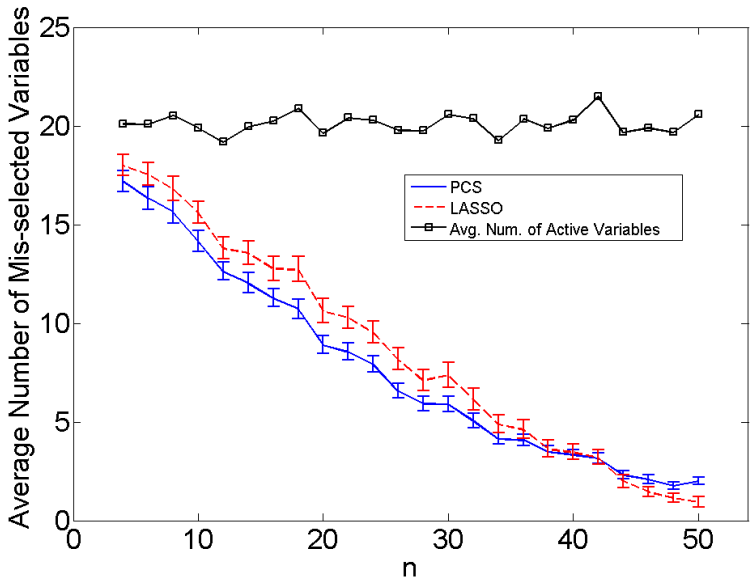

Figure 4.8: Average number of mis-selected variables for active set implementation of LASSO (dashed) vs. Predictive Correlation Screening (solid), $p = 200, q = 20$.

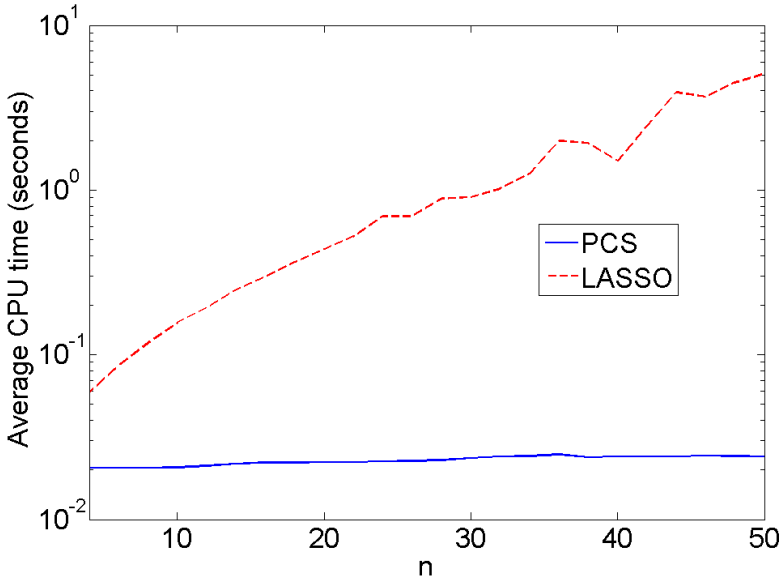

Figure 4.9: Average CPU time for active set implementation of LASSO (dashed) vs. PCS (solid), $p = 200, q = 20$.

of SPARCS as compared to LASSO. Our future work includes using SPARCS in a multi-stage framework. We believe that multi-stage SPARCS can further improve the performance of the algorithm while benefiting from high computational efficiency.

CHAPTER V

Covariance and inverse covariance support recovery via correlation and partial correlation thresholding

5.1 Introduction

In this chapter we consider the problem of (inverse) covariance support recovery using (partial) correlation screening. More specifically, we propose a simple adaptive thresholding method for discovering the structure of covariance matrix or its inverse in a high dimensional regime for which $n \ll p$. As in previous chapters, the proposed method is based on thresholding the magnitudes of the entries of the sample correlation or sample partial correlation matrix. We prove theoretical guarantees, similar to those presented in Chapter IV, for support recovery of the proposed method. The results in this chapter can be viewed as the generalization of the results in chapter IV in the context of covariance support recovery instead of variable selection for online regression.

Discovery of the structure of a high-dimensional covariance matrix or its inverse (also referred to as (inverse) covariance support recovery) is an attractive problem which is useful in various contexts. In the context of covariance estimation, discovering the structure of the covariance matrix or the inverse covariance matrix can be

the first stage of a two-stage esimator of the covriance matrix or its inverse. The second stage of such two-stage procedure is to estimate the non-zero entries of the (inverse) covariance matrix given the support recovered at the first stage. The existing methods on estimation of covariance matrices with pre-specified zeros can be used in the second stage of such two-stage procedure. Example of such methods are constrained maximum likelihood estimation via iterative conditional threshoding (*Chaudhuri et al.*, 2007), maximization of a penalized likelihood with a modified ℓ_1 penalty (*Bien and Tibshirani*, 2011), and partial estimation of a covariance matrix with given structure (*Levina and Vershynin*, 2012).

In the context of graphical models, inverse covariance support recovery can be used to tackle the problem of learning the structure of graphical models. It is well known that the zeros in the inverse covariance matrix of multivariate normal distribution imply the absence of an edge in the corresponding graphical model (*Bishop et al.*, 2006). Discovering such structure is of interest in many applications such as social networks, epidemiology, and finance. In social networks, discovering the structure of the underlying graphical model can identify the friendship links between people in the network (*Sadilek et al.*, 2012). In epidemiology the structure of the graphical model may represent the links between organisms having the potential to spread an infectious disease (*Newman and Watts*, 1999). In finance, a graphical model can be associated with the causal relationships between assets in a market. Discovering the structure of such network can be of use in identification of vulnerable assets for the purpose of risk and portfolio management (*Filiz et al.*, 2012).

In high-dimensional regimes where the number of samples n is relatively small compared to the number of variables p, the sample covariance matrix performs poorly in estimating the population covariance matrix (*Bickel and Levina*, 2008). In this chapter we show that despite the poor estimation, thresholding the sample (partial) correlation coefficients can perform well in discovering the true sparsity structure of

the population (inverse) covariance matrix. To do this we generalize the support recovery results presented in Chapter IV for the problem of covariance structure discovery. More specifically, we show that in a purely high-dimensional regime where n is fixed and p goes to infinity, under certain conditions, the total number of edges in a (partial) correlation network converges to a Poisson random variable. Using the proposed Poisson asymptotic result we introduce an algorithm for discovering the edges of a (partial) correlation network at a specified false discovery rate. We show that, under the assumption of elliptically contoured distribution, such structure recovery method only requires $n = \Theta(\log p)$ samples to recover the true structure with probability converging to one.

Thresholding based methods for (inverse) covariance regularization can be divided into two general categories, i.e., universal thresholding methods and adaptive thresholding methods. Universal thresholding methods (see, e.g., (*Bickel and Levina*, 2008; *Karoui*, 2008; *Rothman et al.*, 2009)) perform thresholding by applying a fixed threshold to all entries of an estimate of the (inverse) covariance matrix (often the sample covariance matrix). In contrast, adaptive thresholding methods (see, e.g., (*Cai and Liu*, 2011)) apply different thresholds to the entries of the sample covariance matrix. The methods provided in this chapter can be considered as adaptive methods for solving the (inverse) covariance support recovery problem since they apply a fixed threshold to the entries of the sample (partial) correlation matrix (i.e., in general they apply different thresholds to the entries of the sample (inverse) covariance matrix).

The rest of this chapter is organized as follows. In Sec. 5.2 we introduce the necessary notation and definitions. Section 5.3 presents the asymptotic theory for the number of edges in a (partial) correlation network along with the algorithms for support recovery. In Sec. 5.4 we present our support recovery guarantees for the proposed algorithm.

5.2 Notations and Definitions

Assume $\mathbf{X} = [X_1, \ldots, X_p]$ and is a random vector, from which n observations are available. We represent the $n \times p$ data matrix as $\mathbb{X}$. Similar to previous chapters, we assume that the vector $\mathbf{X}$ has an elliptically contoured density with mean $\boldsymbol{\mu}_x$ and non-singular $p \times p$ covariance matrix $\boldsymbol{\Sigma}_x$, i.e. the probability density function is of the form $f_{\mathbf{X}}(\mathbf{x}) = g\left((\mathbf{x} - \boldsymbol{\mu}_x)^T {\boldsymbol{\Sigma}_x}^{-1} (\mathbf{x} - \boldsymbol{\mu}_x)\right)$, in which g is a non-negative integrable function. The rest of the notation follows that introduced in Chapter II, unless otherwise specified.

Moreover, we denote the true support of the covariance or the inverse covariance with Ψ, i.e., for the covariance support recovery problem:

$$\Psi = \{(i, j) : \sigma_{ij} \neq 0\}, \tag{5.1}$$

where σ_{ij} denotes the ijth entry of the covariance matrix $\boldsymbol{\Sigma}_x$. Also for the inverse covariance support recovery problem:

$$\Psi = \{(i, j) : \overline{\sigma}_{ij} \neq 0\}, \tag{5.2}$$

where $\overline{\sigma}_{ij}$ denotes the ijth entry of the inverse covariance matrix $\boldsymbol{\Sigma}_x^{-1}$. We denote size of the set Ψ as k.

5.3 Support recovery using (partial) correlation thresholding

In this section we introduce our algorithm for support recovery. We also present an asymptotic Poisson approximation for the number of edges in a (partial) correlation graph. The asymptotic theory is then used to assign p-values to the edges of a (partial) correlation network. As a result, we present a version of our support recovery algorithm that recovers the support at a given statistical significance level.

5.3.1 Main idea

Below we describe the simple idea of (inverse) covariance support recovery using (partial) correlation thresholding. Similar to previous chapters let $\mathbf{\Phi} = [\phi_{ij}]_{1 \leq i,j \leq p}$ be a generic notation for sample correlation matrix $\mathbf{R}$ or its inverse $\mathbf{P}$. For a (partial) correlation threshold $0 \leq \rho \leq 1$ we define the recovered support at the threshold ρ by

$$S = \{(i,j) : |\phi_{ij}| \geq \rho\} \tag{5.3}$$

This algorithm is summarized in Algorithm 2. We show that this simple algorithm has the sure screening property that the true support is a subset of S with probability tending to one.

Algorithm 2: Support recovery via thresholding sample (partial) correlation coefficients

- $S = \{(i,i) : 1 \leq i \leq p\}$;
- Let $L =$ the desired number of upper diagonal entries ;
- **for** $l = 1$ **to** L **do**
  - Let $(i,j) = \arg\max_{(k,m) \in \{(v,w):1 \leq v < w \leq p\} \setminus S} |\phi_{km}|$;
  - $S \leftarrow S \cup \{(i,j),(j,i)\}$;
- Return S;

In the next sub-section we present asymptotic results which allow assigning p-values to the entries of the covariance matrix for being an edge in the (partial) correlation network. This results in a version of the simple algorithm discussed above which recovers the support at a given statistical significance level.

5.3.2 Asymptotic theory

We define $\mathbb{U} = [\mathbf{U}_1, \cdots, \mathbf{U}_p]$ as the generic notation for the U-score representation of matrix $\mathbf{\Phi}$, i.e.:

$$\mathbf{\Phi} = \mathbb{U}^T \mathbb{U}. \tag{5.4}$$

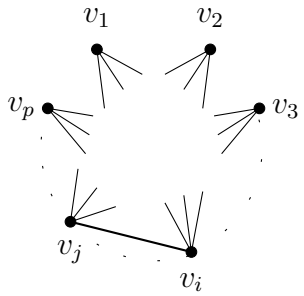

Figure 5.1: (Partial) correlation graph $\mathcal{G}_\rho(\mathbf{\Phi})$ with p vertices $v_1, \cdots, v_p$. For $1 \leq i, j \leq p$, v_i is connected to v_j in $\mathcal{G}_\rho(\mathbf{\Phi})$ if $|\phi_{ij}| \geq \rho$.

Similar to previous chapters, for $\rho \in [0, 1]$, we define the (partial) correlation graph (or network) $\mathcal{G}_\rho(\mathbf{\Phi})$ as follows. The vertices of $\mathcal{G}_\rho(\mathbf{\Phi})$ are $v_1, \cdots, v_p$ which correspond to $\mathbf{U}_1, \cdots, \mathbf{U}_p$ respectively. For $1 \leq i, j \leq p$, v_i and v_j are connected in $\mathcal{G}_\rho(\mathbf{\Phi})$ if the magnitude of the sample (partial) correlation coefficient between X_i and X_j is at least ρ, i.e. $|\phi_{ij}| = |\mathbf{U}_i^T \mathbf{U}_j| \geq \rho$.

The following theorem bounds the total variation distance between the total number of edges in a (partial) correlation graph and a Poisson random variable with rate $\Lambda_{p,n,\rho}$ defined below:

$$\Lambda_{p,n,\rho} = \frac{1}{2} p(p-1) P_0 J(\overline{f_{\mathbf{U}_\bullet, \mathbf{U}_{*-\bullet}}}), \tag{5.5}$$

in which

$$\overline{f_{\mathbf{U}_\bullet, \mathbf{U}_{*-\bullet}}}(\mathbf{u}, \mathbf{v}) = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{p-1} \sum_{j \neq i} \frac{1}{2} \left(f_{\mathbf{U}_i, \mathbf{U}_j}(\mathbf{u}, \mathbf{v}) + f_{\mathbf{U}_i, \mathbf{U}_j}(\mathbf{u}, -\mathbf{v}) \right), \tag{5.6}$$

and the function J is defined in (2.8). Let N_e be the total number of edges in the correlation network corresponding to data matrix $\mathbb{X}$. Also let $Po(\Lambda_{p,n,\rho})$ denote a Poisson random variable with rate $\Lambda_{p,n,\rho}$.

Theorem V.1. *Let the $n \times p$ data matrix $\mathbb{X}$ have associated U-scores $\mathbb{U}$ and assume that $n > 2$. Assume that the joint density of any subset of $\mathbf{U}_i$'s is bounded and*

differentiable. Let the sequence $\{\rho_p\}_p$ *of correlation thresholds be such that* $\rho_p \to 1$ *and* $p(p-1)\left(1-\rho_p^2\right)^{(n-2)/2} \to e_n$ *for some finite constant* e_n*. Then:*

$$d_{TV}\left(N_e, Po(\Lambda_{p,n,\rho})\right) \leq O\left(\max\left\{(d_x/p)^2, \|\Delta_{p,d_x}\|_1, p^{-1}, \sqrt{1-\rho_p}\right\}\right), \tag{5.7}$$

where d_x *is an upper bound on the number of non-zero entries in any row of the covariance matrix* $\mathbf{\Sigma}_x$*.*

Proof. In the proof of Proposition 1 of (*Hero and Rajaratnam*, 2012), the total variation distance between the quantity $\tilde{N}_{\rho,\delta}$ and the corresponding Poisson random variable is bounded. For $\delta = 1$, $\tilde{N}_{\rho,\delta}$ is equal to the number of edges N_e. Therefore, the bound (5.7) follows directly. □

Moreover, using similar arguments as in hub screening (*Hero and Rajaratnam*, 2012) it can be shown that if $\mathbf{\Sigma}_x$ is block sparse of degree d_x, we have:

$$J(\overline{f_{\mathbf{U}_\bullet,\mathbf{U}_{*-\bullet}}}) = 1 + O(d_x/p) \tag{5.8}$$

and

$$\Delta_{i,p,n,d_x} = \begin{cases} 0, & \mathbf{\Phi} = \mathbf{R} \\ O\left(d_x/p\right), & \mathbf{\Phi} = \mathbf{P}. \end{cases} \tag{5.9}$$

Therefore:

$$\Lambda_{p,n,\rho} \approx \frac{1}{2} p(p-1) P_0. \tag{5.10}$$

Theorem V.1 implies that, under appropriate assumptions,

$$\mathbb{P}(N_e > 0) \to 1 - \exp(-\Lambda_{p,n,\rho}), \tag{5.11}$$

as $p \to \infty$, and $\rho \to 1$.

5.3.3 Assigning p-values to edges

Similar to the theory developed in previous chapters, the asymptotic relation (5.11) allows us to assign approximate p-values to pairs $(i, j), 1 \leq i \neq j \leq p$, for being an edge in the (partial) correlation graph, under the null hypothesis of sparse (inverse) covariance matrix

$$pv(i, j) \approx 1 - \exp\left(-\Lambda_{p,n,|\phi_{ij}|}\right), \tag{5.12}$$

in which ϕ_{ij} is the sample (partial) correlation coefficient between X_i and X_j.

5.3.4 Structure discovery using p-value thresholding

Assume that we assign p-values to edges of the correlation graph using (5.12). In order to discover the structure of covariance matrix, we threshold the the p-values at a specified false discovery rate. Since $\Lambda_{p,n,\rho}$ is a decreasing function of ρ, such structure discovery algorithm is equivalent to thresholding the magnitudes of the entries of the sample correlation matrix at a specified threshold. This algorithm is summarized in Algorithm 3.

Algorithm 3: Support recovery via thresholding p-values

- $S = \{(i, i) : 1 \leq i \leq p\}$;
- Let $\alpha =$ the desired significance level;
- **for** $(i, j) \in \{(v, w) : 1 \leq v < w \leq p\}$ **do**
  - Let $pv(i, j) = 1 - \exp\left(-\Lambda_{p,n,|\phi_{ij}|}\right)$;
  - **if** $pv(i, j) \leq \alpha$ **then**
    - $S \leftarrow S \cup \{(i, j), (j, i)\}$;
- Return S;

Algorithm 2 and Algorithm 3 can be considered as adaptive methods for regularizing the (inverse) covariance matrix. In Sec. 5.4 we present our results concerning

the performance guarantees for such thresholding algorithms.

5.4 Performance guarantees

To establish the consistent support recovery property of the (partial) correlation thresholding algorithm we impose similar assumptions to the ones stated in Chapter IV for support recovery of the active variables in the SPARCS online regression procedure.

Assumption V.2. *The rows of the $n \times p$ data matrix $\mathbb{X}$ are i.i.d. realizations of a p-dimensional vector $\mathbf{X}$ which follows a multivariate elliptically contoured distribution with mean $\boldsymbol{\mu}_x$ and $p \times p$ dispersion matrix $\boldsymbol{\Sigma}_x$, i.e. the probability density function (pdf) is of the form $f_{\mathbf{X}}(\mathbf{x}) = g\left((\mathbf{x}-\boldsymbol{\mu}_x)^T {\boldsymbol{\Sigma}_x}^{-1}(\mathbf{x}-\boldsymbol{\mu}_x)\right)$, where g is a non-negative function. Moreover, the density function $f_{\mathbf{X}}(.)$ is bounded and differentiable.*

Assumption V.3. *Let $\boldsymbol{\Gamma} = [\gamma_{ij}]_{1\leq i,j\leq p}$ represent the true $p \times p$ correlation matrix of $\mathbf{X}$. The quantity*

$$\gamma_{\min} = \min_{(i,j)\in\Psi} \{|\gamma_{ij}|\}, \tag{5.13}$$

is strictly positive and independent of p.

Assumption V.4. *The $(n-1) \times p$ matrix of U-scores satisfies:*

$$\frac{n-1}{p}\mathbb{U}^x(\mathbb{U}^x)^T = \mathbf{I}_{n-1} + \boldsymbol{o}(1), \ \text{as } p \to \infty, \tag{5.14}$$

where $\boldsymbol{o}(1)$ is a $(n-1) \times (n-1)$ matrix whose entries are $o(1)$.

Assumption V.3 is similar to the assumption (14) in (*Cai and Liu*, 2011) on the minimum value of the true correlation coefficients. Note that in (*Cai and Liu*, 2011) equation (14) assumes a lower bound for the minimum true covariance. However,

since the variables Y_i introduced there are standardized, assumption (14) is indeed a lower bound for the minimum true correlation coefficient. Assumption V.4 is a common assumption used in the performance analysis of support recovery algorithms that require some estimate of the inverse covariance matrix (cf. (*Obozinski et al.*, 2011; *Fan and Lv*, 2008)). Assumption V.4 can be compared to the conditions on the sparsity-overlap function introduced in (*Obozinski et al.*, 2011) and the concentration property introduced in (*Fan and Lv*, 2008). In the context of inverse covariance estimation, Assumption V.4 can be related to the common regularity assumptions on the eigenspectrum of the true covariance matrix (cf. (*Lam and Fan*, 2009; *Yuan*, 2010)). Note that Assumption V.4 is the same as Assumption IV.3 introduced in Chapter IV. Lemma IV.4 specifies a class of correlation matrices for which Assumption V.4 is satisfied.

The following two theorems give lower bounds on the probability of correct support recovery using Algorithm 2.

Theorem V.5. *Consider the covariance support recovery problem using Algorithm 2. Under Assumptions V.2 and V.3, if* $n \geq \Theta(\log p)$ *then for any* $L \geq (k-p)/2$*, Algorithm 2 recovers the support* Ψ*, with probability at least* $1 - 1/p$*, i.e.,*

$$\mathbb{P}\left(\Psi \subseteq S\right) \geq 1 - 1/p. \tag{5.15}$$

Proof. The proof is similar to the proof of Theorem V.6 and is omitted. □

Theorem V.6. *Consider the inverse covariance support recovery problem using Algorithm 2. Under Assumptions V.2-V.4, if* $n \geq \Theta(\log p)$ *then for any* $L \geq (k-p)/2$*, Algorithm 2 recovers the support* Ψ*, with probability at least* $1 - 1/p$*, i.e.,*

$$\mathbb{P}\left(\Psi \subseteq S\right) \geq 1 - 1/p. \tag{5.16}$$

Proof. By Assumption V.4 we have

$$\mathbb{U}^x(\mathbb{U}^x)^T = \frac{p}{n-1}\left(\mathbf{I}_{n-1} + \mathbf{o}(1)\right). \tag{5.17}$$

Therefore:

$$\left(\mathbb{U}^x(\mathbb{U}^x)^T\right)^{-1} = \frac{n-1}{p}\left(\mathbf{I}_{n-1} + \mathbf{o}(1)\right). \tag{5.18}$$

Since columns of $\mathbb{U}^x$ have unit norm we obtain:

$$(\mathbb{U}^x(\mathbb{U}^x)^T)^{-1}\mathbb{U}^x = \frac{n-1}{p}\mathbb{U}^x(1+o(1)), \tag{5.19}$$

and

$$(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x = (\frac{n-1}{p})^2(\mathbb{U}^x)^T\mathbb{U}^x(1+o(1)). \tag{5.20}$$

This yields

$$\mathbf{D}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x} = (\frac{n-1}{p})^2\mathbf{I}_p(1+o(1)), \tag{5.21}$$

which implies

$$\mathbb{U}^x_{\mathbf{P}} = (\mathbb{U}^x(\mathbb{U}^x)^T)^{-1}\mathbb{U}^x\mathbf{D}^{-\frac{1}{2}}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x} = \mathbb{U}^x(1+o(1)). \tag{5.22}$$

Therefore using representation (2.4) we have

$$\mathbf{P} = (\mathbb{U}^x_{\mathbf{P}})^T\mathbb{U}^x_{\mathbf{P}} = (\mathbb{U}^x)^T\mathbb{U}^x(1+o(1)). \tag{5.23}$$

Hence under Assumption V.4, Algorithm 2 asymptotically selects the support set based on the entries of $(\mathbb{U}^x)^T\mathbb{U}^x$, i.e. sample correlation coefficients.

Since $\mathbb{P}\left(\Psi_0 \subseteq S\right)$ increases as the size of the recovered set S increases, it suffices to prove the proposition for $L = (k-p)/2$ (note that due to the symmetry of inverse covariance matrix, $(k-p)/2$ is necessarily an integer). We have:

$$\begin{aligned}
\mathbb{P} \quad & (\Psi \nsubseteq S) = \mathbb{P}\left(\Psi \neq S\right) \\
\leq \quad & \mathbb{P}\left(\bigcup_{(i,j)\in\Psi,(v,w)\notin\Psi} \{|r_{ij}| < |r_{vw}|\}\right) \\
\leq \quad & \sum_{(i,j)\in\Psi,(v,w)\notin\Psi} \mathbb{P}(|r_{ij}| < |r_{vw}|)
\end{aligned} \tag{5.24}$$

Now by Lemma IV.9 in Chpter IV there exist constants $C_{ij,vw} > 0$ and a constant N such that

$$\begin{aligned}
\mathbb{P} \quad & (\Psi \neq S) \leq \sum_{(i,j)\in\Psi,(v,w)\notin\Psi} \exp(-C_{ij,vw}n) \\
\leq \quad & p^4 \exp(-C_{\min}n),\ \forall n > N,
\end{aligned} \tag{5.25}$$

in which $C_{\min} = \min_{(i,j)\in\Psi,(v,w)\notin\Psi} C_{ij,vw} = \rho_{\min}/6$. Hence by letting $C = 5/C_{\min} = 30/\rho_{\min}$ and $n = C\log p$ we have:

$$\mathbb{P}\left(\Psi \neq S\right) \leq \frac{1}{p}, \tag{5.26}$$

and

$$\mathbb{P}\left(\pi_0 = S\right) = 1 - \mathbb{P}\left(\pi_0 \neq S\right) \geq 1 - \frac{1}{p}, \tag{5.27}$$

which completes the proof. □

The next two theorems propose a lower bound for the probability of exact support recovery for covariance and inverse covariance matrices using Algorithm 3.

Theorem V.7. *Consider the covariance support recovery problem using Algorithm*

3. Under Assumptions V.2 and V.3, if $n \geq \Theta(\log p)$ *then there exists* $0 \leq \alpha_c \leq 1$ *such that using Algorithm 3 at significance level* α_c *recovers the exact support* Ψ *with probability at least* $1 - 1/p$*, i.e.,*

$$\mathbb{P}\left(\Psi = S\right) \geq 1 - 1/p. \tag{5.28}$$

Proof. The proof is similar to the proof of Theorem V.8 and is omitted. □

Theorem V.8. *Consider the inverse covariance support recovery problem using Algorithm 3. Under Assumptions V.2-V.4, if* $n \geq \Theta(\log p)$ *then there exists* $0 \leq \alpha_c \leq 1$ *such that using Algorithm 3 at significance level* α_c *recovers the exact support* Ψ *with probability at least* $1 - 1/p$*, i.e.,*

$$\mathbb{P}\left(\Psi = S\right) \geq 1 - 1/p. \tag{5.29}$$

Proof. Since $\Lambda_{p,n,\rho}$ is a decreasing function of ρ, $pv(i) = 1 - \exp(-\Lambda_{p,n,|\phi_{ij}|})$ is decreasing in ϕ_{ij}. Therefore screening the p-values at a given significance level α is equivalent to screening the quantities ϕ_{ij} at a given threshold ρ. Let ρ_t be the value of threshold for which the size of the recovered support set S is exactly k. Using Algorithm 3 at significance level $\alpha_c = 1 - \exp(-\Lambda_{p,n,\rho_t})$ is equivalent to using Algorithm 2 with $L = (k - p)/2$. Therefore by Theorem V.6, running Algorithm 3 at significance level α_c will recover a support set S for which:

$$\mathbb{P}\left(\Psi_0 = S\right) \geq 1 - \frac{1}{p}. \tag{5.30}$$

□

Theorems V.5 and V.7 can be compared to the support recovery results (also referred to as sparsity or sparsistency in the literature) presented in (*Cai and Liu*, 2011). For distributions with exponential-type tail, Theorem 2 in (*Cai and Liu*, 2011)

requires that $\log p = o(n^{1/3})$. Moreover, for distributions with polynomial-type tail, Theorem 2 in (*Cai and Liu*, 2011) requires that $p \leq c_1 n^{\beta}$ for some $\beta, c_1 > 0$. In contrast, our proposed covariance support recovery theorems V.5 and V.7 only require asymptotic number of samples $n = \Theta(\log p)$ to perform guaranteed support recovery. However, we make the assumption of elliptically contoured distribution with a correlation matrix that satisfies Assumption V.3. It is worth mentioning that the family of elliptically contoured distributions includes distributions with exponential tails (such as multivariate power exponential distribution, cf. (*Gómez et al.*, 1998)) as well as distributions with polynomial tails (such as multivariate t distribution, cf. (*Gupta et al.*, 2013)). Note also that the covariance support recovery theorems V.5 and V.7 do not require sparsity assumptions on the covariance matrix. Also, the inverse covariance support recovery theorems V.6 and V.8 do not directly impose sparsity assumptions on the covariance matrix or its inverse. However, Assumption V.4 reflects sparsity assumption on the population correlation matrix (see Lemma IV.4). Moreover, our asymptotic results for assigning p-values to the edges of the correlation or partial correlation network let $\rho \to 1$ which controls the effective sparsity as $p \to \infty$. Note also that the support recovery results presented in this chapter are different from the conventional convergence analysis performed in the covariance and inverse covariance estimation. Here we consider convergence of the support instead of ℓ_2 norm type of converegnce. Finally, it is worth mentioning that to the best of our knowledge none of the existing methods for covariance and inverse covariance regularization assign p-values to the entries of the discovered support.

CHAPTER VI

Future work

6.1 Introduction

In this chapter, I present five possible directions for future research. The first direction is considering a multi-stage version of the PCS method for SPARCS screening stage presented in Chapter IV. The second direction is proposing a two-stage (inverse) covariance estimator using the support recovery methods introduced in Chapter V. The third direction is screening for general motifs in correlation networks. The fourth direction is about generalization of correlation and hub screening framework to hyper-graphs. We use the simulations on the financial data presented in Chapter III as a motivation for this generalization. Finally, the fifth direction is generalization of the screening results presented in this thesis to the case of complex-valued random variables.

6.2 Multi-stage PCS support recovery for SPARCS screening stage

In Chapter IV we presented theoretical and experimental results which showed the superiority of predictive correlation screening (PCS) algorithm over the well known existing methods of LASSO and SIS, in a truly high-dimensional setting. An idea to

boost such superiority is to use PCS in a multi-stage setting. More specifically, one can consider the following algorithm for variable selection in high-dimensional linear regression. First run PCS on all p variables using the n samples to select a subset S_1 of $[\gamma p]$ variables, where $[x]$ denotes the integer part of x and $0 < \gamma < 1$ is a fixed integer. Next run PCS on variables in S_1 using the n samples to select a subset S_2 of the $[\gamma^2 p]$ variables, and so on. Repeat this procedure k times to obtain subset S_k with $[\gamma^k p]$ predictor variables. Clearly the PCS presented in Chapter IV can be considered as a special case of the above multi-stage procedure which selects all of the predictor variables in one stage. As a result, above multi-stage idea will admit support recovery guarantees as good as those for PCS. More specifically, we believe that above multi-stage procedure enjoys a smaller constant for $\Theta(.)$ in $n = \Theta(\log p)$ relation required for support recovery theorems in Chapter IV.

6.3 Two-stage estimation of the covariance matrix

As we motivated in Chapter V the support recovery methods presented in that chapter can be used as the first stage of a two-stage procedure for (inverse) covariance estimation. Considering such two-stage estimator would be a problem of interest. We showed in Chapter V that the presented support recovery methods are backed with strong guarantees. As a result, we expect that the two-stage estimator of (inverse) covariance to be an estimator which can beat state of the art estimators in very high dimensional situations.

6.4 Screening for general motifs

The hub and edge screening methods discussed in this thesis can be considered as methods which screen for star sub-graphs in a correlation network. A problem that has always been interesting to me is screening for a general sub-graph (motif). More

specifically, can we obtain similar asymptotic results for a general feasible sub-graph instead of the specific case of a star sub-graph? Based on the limited time that I have spent on this problem, I am confident that proving similar Poisson theorems for the number of specific sub-graphs in a (partial) correlation graph is rather straightforward. In fact using similar indicator function expansions (an idea which is used several times across this thesis) along with Chen-Stein method can provide us with asymptotic Poisson limits. What is more challenging in the case of general motifs is obtaining an expression for the expected number of motifs. For the special case of star motifs we were able to come up of with an integral expansion for the expected number of motifs. The integration region for the case of star motifs is a region of the form $A_\rho \times A_\rho \times \cdots \times A_\rho \subset S_{n-2} \times S_{n-2} \times \cdots \times S_{n-2}$ (cf. equation (A.20)). However, for the case of general motifs, it is not generally possible to have a similar separable form for the integration region as a subset of $S_{n-2} \times S_{n-2} \times \cdots \times S_{n-2}$. As a result coming up with an integral expansion for the expectation seems to be a challenging problem. However, an idea that can be useful to overcome this problem is the similarity of correlation graphs with random geometric graphs and Erdös-Rényi random graphs. As an example, if we can bound the total variation distance between the total number of copies of a certain sub-graph in a correlation graph and in an Erdös-Rényi random graph (specifically an Erdös-Rényi random graph with edge connectivity probability P_0), then an approximate expression for the mean number of motifs can be obtained.

6.5 Correlation screening on hyper-graphs

Pairwise relationships are often used to investigate the interconnections among a set of random variables. An important reason for such popularity is that pairwise relationships can be illustrated on graphs. However, in many real-world problems, relations among variables are more complex than simple pairwise relations. Hyper-graph learning studies the higher order relationships among the variables in the data

(*Zhou et al.*, 2006; *Yu et al.*, 2012; *Sun et al.*, 2008).

The vulnerable asset discovery experiment which was introduced as an application of the complex-valued correlation and partial screening procedure proposed in Chapter III, can be an example where the insufficiency of pairwise (partial) correlation relations is evident. We saw in Sec. 3.6.4 that the experimental results became more sensible when we considered the (partial) correlations between groups of assets (i.e., industries) as compared to simply considering (partial) correlations between single assets.

6.6 Generalization of the results to the complex-valued case

Motivated by the spectral correlation and partial correlation screening results of Chapter III, the screening results presented in the rest of the chapters of this thesis can be generalized to the case of complex-valued random variables, in a similar manner. Such generalizations would allow applying the proposed screening methods in the spectral domain.

APPENDICES

APPENDIX A

Generalization of SPARCS to regression with multidimensional response

A.1 Introduction

In this appendix we generalize the asymptotic results of Chapter IV to the case where the response is multidimensional instead of a scalar. We also present another approach for proving some of the results presented in Chapter IV.

A.2 Under-determined multivariate regression with multidimensional response

Assume $\mathbf{X} = [X_1, \ldots, X_p]$ and $\mathbf{Y} = [Y_1, \ldots, Y_q]$ are random vectors of regressor and response variables, from which n observations are available. We represent the $n \times p$ and $n \times q$ data matrices as $\mathbb{X}$ and $\mathbb{Y}$, respectively. We assume that the vector $\mathbf{X}$ has an elliptically contoured density with mean μ_x and non-singular $p \times p$ covariance matrix $\mathbf{\Sigma}_x$, i.e. the probability density function is of the form $f_{\mathbf{X}}(\mathbf{x}) = g\left((\mathbf{x}-\mu_x)^T {\mathbf{\Sigma}_x}^{-1}(\mathbf{x}-\mu_x)\right)$, in which g is a non-negative integrable function. Similarly, the vector $\mathbf{Y}$, is assumed to follow an elliptically contoured density

with mean μ_y and non-singular $q \times q$ covariance matrix $\mathbf{\Sigma}_y$. We assume that the joint density function of $\mathbf{X}$ and $\mathbf{Y}$ is bounded and differentiable. Denote the $p \times q$ population cross covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$ by $\mathbf{\Sigma}_{xy}$.

The $p \times p$ sample covariance matrix $\mathbf{S}$ for data $\mathbb{X}$ is defined as:

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_{(i)} - \overline{\mathbf{X}})^T (\mathbf{X}_{(i)} - \overline{\mathbf{X}}), \tag{A.1}$$

where $\mathbf{X}_{(i)}$ is the ith row of data matrix $\mathbb{X}$, and $\overline{\mathbf{X}}$ is the vector average of all n rows of $\mathbb{X}$.

Consider the $n \times (p+q)$ concatenated matrix $\mathbb{Z} = [\mathbb{X}, \mathbb{Y}]$. The sample cross covariance matrix $\mathbf{S}^{yx}$ is defined as the lower left $q \times p$ block of the $(p+q) \times (p+q)$ sample covariance matrix obtained by (A.1) using $\mathbb{Z}$ as the data matrix instead of $\mathbb{X}$.

Assume that $p \gg n$. We define the ordinary least squares (OLS) estimator of $\mathbf{Y}$ given $\mathbf{X}$ as the min-norm solution of the underdetermined least squares regression problem

$$\min_{\mathbf{B}} \|\mathbb{Y}^T - \mathbf{B}\mathbb{X}^T\|_F^2, \tag{A.2}$$

where $\|\mathbf{A}\|_F$ represents the Frobenius norm of matrix $\mathbf{A}$. The min-norm solution to (A.2) is the $q \times p$ matrix of regression coefficients

$$\mathbf{B} = \mathbf{S}^{yx} (\mathbf{S}^x)^\dagger, \tag{A.3}$$

where $\mathbf{A}^\dagger$ denotes the Moore-Penrose pseudo-inverse of matrix $\mathbf{A}$. If the ith column of $\mathbf{B}$ is zero then the ith variable is not included in the OLS estimator. This is the main motivation for the proposed partial correlation screening procedure.

The PCS procedure for variable selection is based on the U-score representation of the correlation matrices. It is easily shown that there exist matrices $\mathbb{U}^x$ and $\mathbb{U}^y$

of dimensions $(n-1)\times p$ and $(n-1)\times q$ respectively, such that the columns of $\mathbb{U}^x$ and $\mathbb{U}^y$ lie on the $(n-2)$-dimensional unit sphere S_{n-2} in $\mathbb{R}^{n-1}$ and the following representations hold (*Hero and Rajaratnam*, 2012):

$$\mathbf{S}^{yx} = \mathbf{D}_{\mathbf{S}^y}^{\frac{1}{2}}((\mathbb{U}^y)^T\mathbb{U}^x)\mathbf{D}_{\mathbf{S}^x}^{\frac{1}{2}}, \tag{A.4}$$

and:

$$(\mathbf{S}^x)^\dagger = \mathbf{D}_{\mathbf{S}^x}^{-\frac{1}{2}}((\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x)\mathbf{D}_{\mathbf{S}^x}^{-\frac{1}{2}}, \tag{A.5}$$

where $\mathbf{D}_{\mathbf{M}}$ denotes the diagonal matrix obtained by zeroing out the off-diagonals of matrix $\mathbf{M}$. Note that $\mathbb{U}^x$ and $\mathbb{U}^y$ are constructed from data matrices $\mathbb{X}$ and $\mathbb{Y}$, respectively.

Throughout the Appendix, we assume the data matrices $\mathbb{X}$ and $\mathbb{Y}$ have been normalized in such a way that the sample variance of each variable X_i and Y_j is equal to 1 for $1\leq i\leq p$ and $1\leq j\leq q$. This simplifies the representations (A.4) and (A.5) to $\mathbf{S}^{yx} = (\mathbb{U}^y)^T\mathbb{U}^x$ and $(\mathbf{S}^x)^\dagger = (\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x$. Using these representations, one can write:

$$\hat{\mathbf{Y}} = \mathbf{S}^{yx}(\mathbf{S}^x)^\dagger\mathbf{X} = (\mathbb{U}^y)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-1}\mathbb{U}^x\mathbf{X}. \tag{A.6}$$

Defining $\tilde{\mathbb{U}}^x = (\mathbb{U}^x(\mathbb{U}^x)^T)^{-1}\mathbb{U}^x\mathbf{D}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x}^{-\frac{1}{2}}$, we have:

$$\hat{\mathbf{Y}} = (\mathbb{U}^y)^T\tilde{\mathbb{U}}^x\mathbf{D}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x}^{\frac{1}{2}}\mathbf{X} \tag{A.7}$$

$$= (\mathbf{H}^{xy})^T\mathbf{D}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x}^{\frac{1}{2}}\mathbf{X}, \tag{A.8}$$

where

$$\mathbf{H}^{xy} = (\tilde{\mathbb{U}}^x)^T\mathbb{U}^y. \tag{A.9}$$

Note that the columns of matrix $\tilde{\mathbb{U}}^x$ lie on S_{n-2}. This can simply be verified by the fact that diagonal entries of the $p \times p$ matrix $(\tilde{\mathbb{U}}^x)^T\tilde{\mathbb{U}}^x$ are equal to one.

The U-score representations of covariance matrices completely specify the regression coefficient matrix $\mathbf{S}^{yx}(\mathbf{S}^x)^\dagger$.

We define variable selection by discovering columns of the matrix (A.10) that are not close to zero. The expected number of discoveries will play an important role in the theory of false discoveries, discussed below.

From Sec. A.2 we obtain a U-score representation of the regression coefficient matrix:

$$\mathbf{S}^{yx}(\mathbf{S}^x)^\dagger = (\mathbf{H}^{xy})^T\mathbf{D}^{\frac{1}{2}}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x}. \tag{A.10}$$

Under the condition that $\mathbf{D}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x}$ has non-zero diagonal entries, the ith column of $\mathbf{S}^{yx}(\mathbf{S}^x)^\dagger$ is a zero vector if and only if the ith row of $\mathbf{H}^{xy}$ is a zero vector, for $1 \leq i \leq p$. This motivates screening for non-zero rows of the matrix $\mathbf{H}^{xy}$ instead of columns of $\mathbf{S}^{yx}(\mathbf{S}^x)^\dagger$.

Fix an integer $\delta \in \{1, 2, \cdots, p\}$ and a real number $\rho \in [0, 1]$. For each $1 \leq i \leq p$, we call i a discovery at degree threshold δ and correlation threshold ρ if there are at least δ entries in ith row of $\mathbf{H}^{xy}$ of magnitude at least ρ. Note that this definition can be generalized to an arbitrary matrix of the form $(\mathbb{U}^x)^T\mathbb{U}^y$ where $\mathbb{U}^x$ and $\mathbb{U}^y$ are matrices whose columns lie on S_{n-2}. For a general matrix of the form $(\mathbb{U}^x)^T\mathbb{U}^y$ we represent the number of discoveries at degree level δ and threshold level ρ as $N^{xy}_{\delta,\rho}$.

A.3 Asymptotic theory

The following notations are necessary for the theorems in this section. We denote the surface area of the $(n-2)$-dimensional unit sphere S_{n-2} in $\mathbb{R}^{n-1}$ by a_n. Assume that $\mathbf{U}, \mathbf{V}$ are two independent and uniformly distributed random vectors on S_{n-2}.

For a threshold $\rho \in [0,1]$, let $r = \sqrt{2(1-\rho)}$. P_0 is then defined as the probability that either $\|\mathbf{U}-\mathbf{V}\|_2 \leq r$ or $\|\mathbf{U}+\mathbf{V}\|_2 \leq r$. P_0 can be computed using the formula for the area of spherical caps on S_{n-2} (*Hero and Rajaratnam*, 2012).

Define the index set $\mathcal{C}$ as:

$$\mathcal{C} = \{(i_0, i_1, \ldots, i_\delta) : \\ 1 \leq i_0 \leq p, 1 \leq i_1 < \ldots < i_\delta \leq q\}. \tag{A.11}$$

For arbitrary joint density $f_{\mathbf{U}_0,\ldots,\mathbf{U}_\delta}(\mathbf{u}_0,\ldots,\mathbf{u}_\delta)$ defined on the Cartesian product $S_{n-2}^{\delta+1} = S_{n-2} \times \cdots \times S_{n-2}$, define $\overline{f_{\mathbf{U}^x_\bullet,\mathbf{U}^y_{*_1},\ldots,\mathbf{U}^y_{*_\delta}}}(\mathbf{u}_0,\mathbf{u}_1,\ldots,\mathbf{u}_\delta)$ as the average of

$$f_{\mathbf{U}_{\vec{i}}}(s_0\mathbf{u}_0, s_1\mathbf{u}_1, \ldots, s_\delta\mathbf{u}_\delta) = \\ f_{\mathbf{U}^x_{i_0},\mathbf{U}^y_{i_1},\ldots,\mathbf{U}^y_{i_\delta}}(s_0\mathbf{u}_0, s_1\mathbf{u}_1, \ldots, s_\delta\mathbf{u}_\delta), \tag{A.12}$$

for all $\vec{i} = (i_0, i_1, \ldots, i_\delta) \in \mathcal{C}$ and $s_j \in \{-1, 1\}, 0 \leq j \leq \delta$.

In the following theorems, k represents an upper bound on the number of non-zero entries in any row or column of covariance matrix $\mathbf{\Sigma}_x$ or cross covariance matrix $\mathbf{\Sigma}_{xy}$. We define $\|\Delta^{xy}_{p,q,n,k,\delta}\|_1 = |\mathcal{C}|^{-1}\sum_{\vec{i}\in\mathcal{C}}\Delta^{xy}_{p,q,n,k,\delta}(\vec{i})$, the average dependency coefficient, as the average of

$$\Delta^{xy}_{p,q,n,k,\delta}(\vec{i}) = \left\|(f_{\mathbf{U}_{\vec{i}}|\mathbf{U}_{A_k(i_0)}} - f_{\mathbf{U}_{\vec{i}}})/f_{\mathbf{U}_{\vec{i}}}\right\|_\infty, \tag{A.13}$$

in which $A_k(i_0)$ is defined as the set complement of the union of indices of non-zero elements of the i_0-th column of $\mathbf{\Sigma}_{yx}\mathbf{\Sigma}_x^{-1}$. Finally, the function J of the joint density $f_{\mathbf{U}_0,\ldots,\mathbf{U}_\delta}(\mathbf{u}_0,\ldots,\mathbf{u}_\delta)$ is defined as:

$$J(f_{\mathbf{U}_0,\ldots,\mathbf{U}_\delta}) = |S_{n-2}|^\delta \int\limits_{S_{n-2}} f_{\mathbf{U}_0,\ldots,\mathbf{U}_\delta}(\mathbf{u},\ldots,\mathbf{u})d\mathbf{u}. \tag{A.14}$$

The following theorem gives an asymptotic expression for the number of discoveries in a matrix of the form $(\mathbb{U}^x)^T\mathbb{U}^y$, as $p \to \infty$, for fixed n. Also it states that, under certain assumptions, the probability of having at least one discovery converges to a given limit. This limit is equal to the probability that a certain Poisson random variable N^*_{δ,ρ_p} with rate equal to $\lim_{p\to\infty} E[N^{xy}_{\delta,\rho_p}]$ takes a non-zero value, i.e. it satisfies: $N^*_{\delta,\rho_p} > 0$.

Theorem A.1. *Let* $\mathbb{U}^x = [\mathbf{U}^x_1, \mathbf{U}^x_2, ..., \mathbf{U}^x_p]$ *and* $\mathbb{U}^y = [\mathbf{U}^y_1, \mathbf{U}^y_2, ..., \mathbf{U}^y_q]$ *be* $(n-1)\times p$ *and* $(n-1)\times q$ *random matrices respectively, with* $\mathbf{U}^x_i, \mathbf{U}^y_j \in S_{n-2}$ *for* $1 \le i \le p, 1 \le j \le q$. *Fix integers* $\delta \ge 1$ *and* $n > 2$. *Assume that the joint density of any subset of* $\{\mathbf{U}^x_1, ...\mathbf{U}^x_p, \mathbf{U}^y_1, ..., \mathbf{U}^y_q\}$ *is bounded and differentiable. Let* $\{\rho_p\}_p$ *be a sequence in* $[0,1]$ *such that* $\rho_p \to 1$ *as* $p \to \infty$ *and* $p^{\frac{1}{\delta}} q(1-\rho_p^2)^{\frac{(n-2)}{2}} \to e_{n,\delta}$. *Then,*

$$\begin{aligned}\lim_{p\to\infty} E[N^{xy}_{\delta,\rho_p}] &= \lim_{p\to\infty} \xi_{p,q,n,\delta,\rho_p} J(\overline{f_{\mathbf{U}^x_*, \mathbf{U}^y_{\bullet 1}, ..., \mathbf{U}^y_{\bullet\delta}}}) \\ &= \kappa_{n,\delta} \lim_{p\to\infty} J(\overline{f_{\mathbf{U}^x_*, \mathbf{U}^y_{\bullet 1}, ..., \mathbf{U}^y_{\bullet\delta}}}), \end{aligned} \tag{A.15}$$

where $\xi_{p,q,n,\delta,\rho_p} = p\binom{q}{\delta} P_0^\delta$ *and* $\kappa_{n,\delta} = (e_{n,\delta}\mathrm{a}_n/(n-2))^\delta/\delta!$.
Assume also that $k = o((p^{\frac{1}{\delta}}q)^{1/(\delta+1)})$ *and that the average dependency coefficient satisfies*
$\lim_{p\to\infty} \|\Delta^{xy}_{p,q,n,k,\delta}\|_1 = 0$. *Then:*

$$p(N^{xy}_{\delta,\rho_p} > 0) \to 1 - \exp(-\Lambda^{xy}_\delta), \tag{A.16}$$

with

$$\Lambda^{xy}_\delta = \lim_{p\to\infty} E[N^{xy}_{\delta,\rho_p}]. \tag{A.17}$$

Proof. Define $\phi^x_i = I(d^x_i \ge \delta)$, where d^x_i is the degree of vertex i in part x in the thresh-

olded correlation graph. We have: $N^{xy}_{\delta,\rho} = \sum_{i=1}^{p} \phi^x_i$. Define $\phi^{xy}_{ij} = I(\mathbf{U}^y_j \in A(r, \mathbf{U}^x_i))$, where $A(r, \mathbf{U}^x_i)$ is the union of two anti-polar caps in S_{n-2} of radius $\sqrt{2(1-\rho)}$ centered at $\mathbf{U}^x_i$ and $-\mathbf{U}^x_i$. ϕ^x_i can be expressed as:

$$\phi^x_i = \sum_{l=\delta}^{q} \sum_{\vec{k} \in \breve{C}(q,l)} \prod_{j=1}^{l} \phi^{xy}_{ik_j} \prod_{m=l+1}^{q} (1 - \phi^{xy}_{ik_m}), \tag{A.18}$$

where $\vec{k} = (k_1, ..., k_q)$ and $\breve{C}(q,l) = \{\vec{k} : k_1 < k_2 < ... < k_l, k_{l+1} < ... < k_q, k_j \in \{1, 2, ..., q\}, k_i \neq k_j\}$.

By subtracting $\sum_{\vec{k} \in \breve{C}(q,l)} \prod_{j=1}^{\delta} \phi^{xy}_{ik_j}$ from both sides, we get:

$$\begin{aligned} \phi^x_i - \sum_{\vec{k} \in \breve{C}(q,l)} \prod_{j=1}^{\delta} \phi^{xy}_{ik_j} = \\ \sum_{l=\delta+1}^{q} \sum_{\vec{k} \in \breve{C}(q,l)} \prod_{j=1}^{l} \phi^{xy}_{ik_j} \prod_{m=l+1}^{q} (1 - \phi^{xy}_{ik_m}) + \\ \sum_{\vec{k} \in \breve{C}(q,\delta)} \sum_{m=\delta+1}^{q} (-1)^{m-\delta} \prod_{j=1}^{\delta} \phi^{xy}_{ik_j} \\ \sum_{k'_{\delta+1} < ... < k'_m, \{k'_{\delta+1}, ..., k'_m\} \subset \{k_{\delta+1}, ..., k_q\}} \prod_{n=\delta+1}^{m} \phi^{xy}_{ik'_n}. \end{aligned} \tag{A.19}$$

The following inequality will be helpful:

$$\begin{aligned} E[\prod_{i=1}^{k} \phi^{xy}_{ii_j}] = \int_{S_{n-2}} dv \int_{A(r,v)} du_1 ... \int_{A(r,v)} du_k \\ f_{U^y_{i_1}, ..., U^y_{i_k}, U^x_i}(u_1, ..., u_k, v) \end{aligned} \tag{A.20}$$

$$\leq P^k_0 a^k_n M^{yx}_{K|1}, \tag{A.21}$$

where $M^{yx}_{K|1} = \max_{i_1 \neq ... \neq i_k, i} \| f_{U^y_{i_1}, ..., U^y_{i_k} | U^x_i} \|_\infty$.

Also we have:

$$E[\prod_{l=1}^{m}\phi_{i_l j_l}^{xy}] \leq P_0^m a_n^m M_{|Q|}^{yx}, \tag{A.22}$$

where $Q = unique(\{i_l, j_l\})$ is the set of unique indices among the distinct pairs $\{\{i_l, j_l\}\}_{l=1}^{m}$ and $M_{|Q|}^{yx}$ is a bound on the joint density of $\mathbf{U}_Q^{xy}$.

Now define:

$$\theta_i^x = \binom{q}{\delta}^{-1} \sum_{\vec{k}\in\breve{C}(q,\delta)} \prod_{j=1}^{\delta} \phi_{ik_j}^{xy}. \tag{A.23}$$

Now, we show that

$$|E[\phi_i^x] - \binom{q}{\delta} E[\theta_i^x]| \leq \gamma_{q,\delta}(qP_0)^{\delta+1}, \tag{A.24}$$

where $\gamma_{q,\delta} = 2e \max_{\delta+1\leq l\leq q}\{a_n^l M_{l|1}^{yx}\}$. To show this, take expectations from both sides of equation (A.19) and apply the bound (A.21) to obtain:

$$\begin{aligned}
&|E[\phi_i^x - \binom{q}{\delta} E[\theta_i^x]]| \\
&\leq \sum_{l=\delta+1}^{q} \binom{q}{l} P_0^l a_n^l M_{l|1}^{yx} + \\
&\binom{q}{\delta} \sum_{l=1}^{q-\delta} \binom{q-\delta}{l} P_0^{\delta+l} a_n^{\delta+l} M_{\delta+l|1}^{yx} \\
&\leq \max_{\delta+1\leq l\leq q}\{a_n^l M_{l|1}^{yx}\} \\
&(\sum_{l=\delta+1}^{q} \binom{q}{l} P_0^l + \binom{q}{\delta} P_0^\delta \sum_{l=1}^{q-\delta} \binom{q-\delta}{l} P_0^l) \\
&\leq \max_{\delta+1\leq l\leq q}\{a_n^l M_{l|1}^{yx}\} \\
&((e - \sum_{l=1}^{\delta} \frac{1}{l!})(qP_0)^{\delta+1} + \frac{q^\delta}{\delta!} P_0^\delta (e-1)(q-\delta)P_0) \\
&\leq \max_{\delta+1\leq l\leq q}\{a_n^l M_{l|1}^{yx}\} 2e(qP_0)^{\delta+1},
\end{aligned} \tag{A.25}$$

in which, the third inequality follows from the assumption $qP_0 \leq 1$ along with the inequality :

$$\sum_{k=s+1}^{G} \binom{G}{k} (\frac{t}{G})^k \leq \sum_{k=s+1}^{G} \frac{t^k}{k!} \\ \leq (e - \sum_{k=0}^{s} \frac{1}{k!}) t^{s+1}, \ 0 \leq t \leq 1. \tag{A.26}$$

Application of the mean value theorem to the integral representation (A.20) yields:

$$|E[\theta_i^x] - P_0^{\delta} J(\overline{f_{\mathbf{U}_{*1}^y,...,\mathbf{U}_{*\delta}^y,\mathbf{U}_i^x}})| \leq \tilde{\gamma}_{q,\delta}^{yx} (qP_0)^{\delta} r, \tag{A.27}$$

where $\tilde{\gamma}_{q,\delta}^{yx} = 2a_n^{\delta+1} \dot{M}_{\delta+1|1}^{yx}/\delta!$ and $\dot{M}_{\delta+1|1}^{yx}$ is a bound on the norm of the gradient:

$$\nabla_{\mathbf{U}_{i_1}^y,...,\mathbf{U}_{i_\delta}^y} \overline{f_{\mathbf{U}_{*1}^y,...,\mathbf{U}_{*\delta}^y|\mathbf{U}_i^x}}(\mathbf{U}_{i_1}^y, ..., \mathbf{U}_{i_\delta}^y | \mathbf{U}_i^x). \tag{A.28}$$

Combining (A.25) and (A.27) and using the relation $r = O((1-\rho)^{1/2})$ we conclude:

$$|E[\phi_i^x] - \binom{q}{\delta} P_0^{\delta} J(\overline{f_{\mathbf{U}_i^x,\mathbf{U}_{*1}^y,...,\mathbf{U}_{*\delta}^y}})| \leq \\ O(p^{\delta}(qP_0)^{\delta} \max\{pP_0, (1-\rho)^{1/2}\}). \tag{A.29}$$

Summing up over i we conclude:

$$E[N_{\delta,\rho}^{xy}] - \xi_{p,q,n,\delta,\rho}^{xy} J(\overline{f_{\mathbf{U}_*^x,\mathbf{U}_{\bullet 1}^y,...,\mathbf{U}_{\bullet\delta}^y}}) \leq \\ O(p(pP_0)^{\delta} \max\{pP_0, (1-\rho)^{1/2}) \\ = O((\eta_{p,q,\delta}^{xy})^{\delta} \max\{\eta_{p,q,\delta}^{xy} p^{-\frac{1}{\delta}}, (1-\rho)^{1/2}\}), \tag{A.30}$$

where $\eta_{p,q,\delta}^{xy} = p^{1/\delta} qP_0$. This concludes (A.15).

To prove the second part of the theorem, we use Chen-Stein method (*Arratia et al.*,

1990). Define:

$$\tilde{N}^{xy}_{\delta,\rho} = \sum_{0\le i_0\le p, 0\le i_1<\ldots<i_\delta\le q} \prod_{j=1}^{\delta} \phi^{xy}_{i_0 i_j}. \tag{A.31}$$

Assume the vertices i in part x and y of the thresholded graph are shown by i^x and i^y respectively. for $\vec{i} = (i^x_0, i^y_1, ..., i^y_\delta)$, define the index set $B^{xy}_{\vec{i}} = B^{xy}_{(i^x_0, i^y_1, ..., i^y_\delta)} = \{(j^x_0, j^y_1 ..., j^y_\delta) : j^x_1 \in \mathcal{N}^{xy}_k(i^x_1) \cup i^x_1, j^y_l \in \mathcal{N}^{xy}_k(i^y_l) \cup i^y_l, l = 1, ..., \delta\} \cap C^{xy}_<$ where $C^{xy}_< = \{(j_0, ..., j_\delta) : 1 \le j_0 \le p, 1 \le j_1 < ... < j_\delta \le q\}$. Note that $|B^{xy}_{\vec{i}}| \le k^{\delta+1}$. We have:

$$\tilde{N}^{xy}_{\delta,\rho} = \sum_{\vec{i}\in C^{xy}_<} \prod_{j=1}^{\delta} \phi^{xy}_{i_0 i_j}. \tag{A.32}$$

Assume $N^{*xy}_{\delta,\rho}$ is a Poisson random variable with $E[N^{*xy}_{\delta,\rho}] = \tilde{N}^{xy}_{\delta,\rho}$. Using theorem 1 of (*Arratia et al.*, 1990), we have:

$$2 \max_A |p(\tilde{N}^{xy}_{\delta,\rho} \in A) - p(\tilde{N}^{*xy}_{\delta,\rho} \in A)| \le b_1 + b_2 + b_3, \tag{A.33}$$

where:

$$b_1 = \sum_{\vec{i}\in C^{xy}_<} \sum_{\vec{j}\in B^{xy}_{\vec{i}} - \vec{i}} E[\prod_{l=1}^{\delta} \phi^{xy}_{i_0 i_l}] E[\prod_{m=1}^{\delta} \phi^{xy}_{j_0 j_m}], \tag{A.34}$$

$$b_2 = \sum_{\vec{i}\in C^{xy}_<} \sum_{\vec{j}\in B^{xy}_{\vec{i}} - \vec{i}} E[\prod_{l=1}^{\delta} \phi^{xy}_{i_0 i_l} \prod_{m=1}^{\delta} \phi^{xy}_{j_0 j_m}], \tag{A.35}$$

and for $p_{\vec{i}^{xy}} = E[\prod_{l=1}^{\delta} \phi^{xy}_{i_0 i_l}]$:

$$b_3 = \sum_{\vec{i}\in C^{xy}_<} E[E[\prod_{l=1}^{\delta} \phi^{xy}_{i_0 i_l} - p_{\vec{i}^{xy}} | \phi^x_{\vec{j}} : \vec{j} \notin B^{xy}_{\vec{i}}]]. \tag{A.36}$$

Using the bound (A.22), $E[\prod_{l=1}^{\delta}\phi_{i_0 i_l}^{xy}]$ is of order $O(P_0^{\delta})$. Therefore:

$$b_1 \leq O(pq^{\delta}k^{\delta+1}P_0^{2\delta}) =$$
$$= O((\eta_{p,q,\delta}^{xy})^{2\delta}(k/(p^{\frac{1}{\delta+1}}q^{\frac{\delta}{\delta+1}}))^{\delta+1}). \tag{A.37}$$

Note that, since $\vec{i} \neq \vec{j}, \prod_{l=1}^{\delta}\phi_{i_0 i_l}^{xy}\prod_{m=1}^{\delta}\phi_{j_0 j_m}^{xy}$ is a multiplication of at least $\delta+1$ different characteristic functions. Hence by (A.22),

$$E[\prod_{l=1}^{\delta}\phi_{i_0 i_l}^{xy}\prod_{m=1}^{\delta}\phi_{j_0 j_m}^{xy}] = O(P_0^{\delta+1}). \tag{A.38}$$

Hence, $b_2 \leq O(pq^{\delta}k^{\delta+1}P_0^{\delta+1}) = O((\eta_{p,q,\delta}^{xy})^{\delta+1}(k/(p^{\frac{1}{\delta}}q)^{1/(\delta+1)})^{\delta+1})$. Finally, to bound b_3 we have:

$$b_3 = \sum_{\vec{i}\in C_<^{xy}} E[E[\prod_{l=1}^{\delta}\phi_{i_0 i_l}^{xy} - p_{\vec{i}^{xy}}|\mathbf{U}_{A_k^{xy}(\vec{i})}]] = \tag{A.39}$$

$$= \sum_{\vec{i}\in C_<^{xy}} \int_{S_{n-2}^{|A_k^{xy}(\vec{i})|}} dz_{A_k^{xy}(\vec{i})} (\prod_{l=1}^{\delta}\int_{S_{n-2}} dz_{i_0^x} \int_{A(r,\mathbf{u}_{i_0}^x)} d\mathbf{u}_{i_l}^y)$$
$$(\frac{f_{\mathbf{U}_{\vec{i}}^{xy}|\mathbf{U}_{A_k^{xy}(\vec{i})}}(\mathbf{U}_{\vec{i}}^{xy}|\mathbf{U}_{A_k^{xy}(\vec{i})}) - f_{\mathbf{U}_{\vec{i}}^{xy}}(\mathbf{U}_{\vec{i}}^{xy})}{f_{\mathbf{U}_{\vec{i}}^{xy}}(\mathbf{U}_{\vec{i}}^{xy})})$$
$$f_{\mathbf{U}_{\vec{i}}^{xy}}(\mathbf{U}_{\vec{i}}^{xy}) f_{\mathbf{U}_{A_k^{xy}(\vec{i})}}(\mathbf{u}_{A_k^{xy}(\vec{i})}) \tag{A.40}$$
$$\leq O(pq^{\delta}P_0^{\delta+1}\|\Delta_{p,q,n,k,\delta}^{xy}\|_1) = O((\eta_{p,q,\delta}^{xy})^{\delta}\|\Delta_{p,q,n,k,\delta}^{xy}\|_1).$$

Therefore:

$$\begin{aligned} |p(N^{xy}_{\delta,\rho} > 0) - (1 - \exp(-\Lambda^{xy}_{\delta}))| \leq \\ |p(N^{xy}_{\delta,\rho} > 0) - (\tilde{N}^{xy}_{\delta,\rho} > 0)| + \\ |p(\tilde{N}^{xy}_{\delta,\rho} > 0) - (1 - \exp(-E[\tilde{N}^{xy}_{\delta,\rho}]))| + \\ |\exp(-E[\tilde{N}^{xy}_{\delta,\rho}]) - \exp(-\Lambda^{xy}_{\delta})| \\ \leq 0 + b_1 + b_2 + b_3 + O(|E[\tilde{N}^{xy}_{\delta,\rho}] - \Lambda^{xy}_{\delta}|). \end{aligned} \tag{A.41}$$

Hence, it remains to bound $O(|E[\tilde{N}^{xy}_{\delta,\rho}] - \Lambda^{xy}_{\delta}|)$. Application of mean value theorem to the multiple integral (A.20) gives:

$$|E[\prod_{l=1}^{\delta} \phi^{xy}_{ii_l}] - P_0^{\delta} J(f_{\mathbf{U}^y_{i_1},\ldots,\mathbf{U}^y_{i_\delta},\mathbf{U}^x_i})| \leq O(P_0^{\delta} r). \tag{A.42}$$

Using relation (A.32) we conclude:

$$\begin{aligned} |E[\tilde{N}^{xy}_{\delta,\rho}] - p\binom{q}{\delta} P_0^{\delta} J(\overline{f_{\mathbf{U}^y_{*1},\ldots,\mathbf{U}^y_{*\delta},\mathbf{U}^x_{\bullet}}})| \leq \\ O(pq^{\delta} P_0^{\delta} r) = O((\eta^{xy}_{p,q,\delta})^{\delta} r). \end{aligned} \tag{A.43}$$

Combining this with inequality (A.41) along with the bounds on b_1, b_2 and b_3, completes the proof of (A.16). □

The following theorem states that when the rows of data matrices $\mathbb{X}$ and $\mathbb{Y}$ are i.i.d. elliptically distributed with block sparse covariance matrices, the rate (A.15) in Theorem A.1 becomes independent of Σ_x and Σ_{xy}. Specifically, the $(\delta+1)$-fold average $J(\overline{f_{\mathbf{U}^x_{\bullet},\mathbf{U}^y_{*1},\ldots,\mathbf{U}^y_{*\delta}}})$ converges to 1 while the average dependency coefficient $\|\Delta^{xy}_{p,q,n,k,\delta}\|_1$ goes to 0, as $p \to \infty$. This theorem will play an an important role in identifying phase transitions and in approximating p-values.

Theorem A.2. *Assume the hypotheses of Theorem A.1 are satisfied. In addition*

assume that the rows of data matrices $\mathbb{X}$ *and* $\mathbb{Y}$ *are i.i.d. elliptically distributed with block sparse covariance and cross covariance matrices* $\mathbf{\Sigma}_x$ *and* $\mathbf{\Sigma}_{xy}$*. Then* Λ_δ^{xy} *in the limit* (A.17) *in Theorem A.1 is equal to the constant* $\kappa_{n,\delta}$ *given in* (A.15)*. Moreover,* $\tilde{\mathbb{U}}_x \approx \mathbb{U}_x$.

Proof. We prove the more general theorem below. Theorem A.2 is then a direct consequence.

Proposition: Let $\mathbb{X}$ *and* $\mathbb{Y}$ *be* $n \times p$ *and* $n \times q$ *data matrices whose rows are i.i.d. realizations of elliptically distributed* p*-dimensional and* q*-dimensional vectors* $\mathbf{X}$ *and* $\mathbf{Y}$ *with mean parameters* μ_x *and* μ_y *and covariance parameters* $\mathbf{\Sigma}_x$ *and* $\mathbf{\Sigma}_y$*, respectively and cross covariance* $\mathbf{\Sigma}_{xy}$*. Let* $\mathbb{U}^x = [\mathbf{U}_1^x, \ldots, \mathbf{U}_p^x]$ *and* $\mathbb{U}^y = [\mathbf{U}_1^y, \ldots, \mathbf{U}_q^y]$ *be the matrices of correlation U-scores. Assume that the covariance matrices* $\mathbf{\Sigma}_x$ *and* $\mathbf{\Sigma}_y$ *are block-sparse of degrees* d_x *and* d_y*, respectively (i.e. by rearranging their rows and columns, all non-diagonal entries are zero except a* $d_x \times d_x$ *or a* $d_y \times d_y$ *block). Assume also that the cross covariance matrix* $\mathbf{\Sigma}^{xy}$ *is block-sparse of degree* d_1 *for* x *and degree* d_2 *for* y *(i.e. by rearranging its rows and columns, all entries are zero except a* $d_1 \times d_2$ *block), then*

$$\tilde{\mathbb{U}}^x = \mathbb{U}^x(1 + O(d_x/p)). \tag{A.44}$$

Also assume that for $\delta \geq 1$ *the joint density of any distinct set of U-scores* $\mathbf{U}_i^x, \mathbf{U}_{i_1}^y, \ldots, \mathbf{U}_{i_\delta}^y$ *is bounded and differentiable over* $S_{n-2}^{\delta+1}$*. Then the* $(\delta+1)$*-fold average function* $J(\overline{f_{\mathbf{U}_\bullet^x, \mathbf{U}_{*_1}^y, \ldots, \mathbf{U}_{*_\delta}^y}})$ *and the average dependency coefficient* $\|\Delta_{p,n,k,\delta}^{xy}\|$ *satisfy*

$$J(\overline{f_{\mathbf{U}_\bullet^x, \mathbf{U}_{*_1}^y, \ldots, \mathbf{U}_{*_\delta}^y}}) = 1 + O(\max\{\frac{d_1}{p}, \delta\frac{(d_y - 1)}{q}\}), \tag{A.45}$$

$$\|\Delta_{p,q,n,k,\delta}^{xy}\|_1 = 0. \tag{A.46}$$

Furthermore,

$$J(\overline{f_{\tilde{\mathbf{U}}^x_\bullet, \mathbf{U}^y_{*_1}, \ldots, \mathbf{U}^y_{*_\delta}}}) = 1 + O(\max\{\frac{d_x}{p}, \frac{d_1}{p}, \delta\frac{(d_y-1)}{q}\}) \tag{A.47}$$

$$\|\Delta^{\tilde{x}y}_{p,q,n,k,\delta}\|_1 = O\left((d_x/p)\right). \tag{A.48}$$

Proof: We have:

$$\tilde{\mathbb{U}}^x = (\mathbb{U}^x(\mathbb{U}^x)^T)^{-1}\mathbb{U}^x\mathbf{D}^{-\frac{1}{2}}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x}. \tag{A.49}$$

By block sparsity of $\mathbf{\Sigma}_x, \mathbb{U}^x$ can be partitioned as:

$$\mathbb{U}^x = [\underline{\mathbb{U}}^x, \overline{\mathbb{U}}^x], \tag{A.50}$$

where $\underline{\mathbb{U}}^x = [\underline{\mathbf{U}}^x_1, \cdots, \underline{\mathbf{U}}^x_{d_x}]$ and $\overline{\mathbb{U}}^x = [\overline{\mathbf{U}}^x_1, \cdots, \overline{\mathbf{U}}^x_{p-d_x}]$ are dependent and independent columns of $\mathbb{U}^x$, respectively. Similarly, by block sparsity of $\mathbf{\Sigma}_y$,

$$\mathbb{U}^y = [\underline{\mathbb{U}}^y, \overline{\mathbb{U}}^y], \tag{A.51}$$

where $\underline{\mathbb{U}}^y = [\underline{\mathbf{U}}^y_1, \cdots, \underline{\mathbf{U}}^y_{d_y}]$ and $\overline{\mathbb{U}}^y = [\overline{\mathbf{U}}^y_1, \cdots, \overline{\mathbf{U}}^y_{q-d_y}]$ are dependent and independent columns of $\mathbb{U}^y$, respectively. By block sparsity of $\mathbf{\Sigma}_{xy}$, at most d_1 variables among $\overline{\mathbf{U}}^x_1, \cdots, \overline{\mathbf{U}}^x_{p-d_x}$, are correlated with columns of $\mathbb{U}^y$. Assume the correlated variables are among $\overline{\mathbf{U}}^x_1, \cdots, \overline{\mathbf{U}}^x_{d_2}$. Similarly, at most d_2 variables among $\overline{\mathbf{U}}^y_1, \cdots, \overline{\mathbf{U}}^y_{q-d_y}$ are correlated with columns of $\mathbb{U}^x$. Without loss of generality, assume the correlated variables are among $\overline{\mathbf{U}}^y_1, \cdots, \overline{\mathbf{U}}^y_{d_1}$.

The columns of $\overline{\mathbb{U}}^x$, are i.i.d. and uniform over the unit sphere S_{n-2}. Therefore,

as $p \to \infty$:

$$\frac{1}{p-d_x}\overline{\mathbb{U}}^x(\overline{\mathbb{U}}^x)^T \to E[\overline{\mathbf{U}}_1^x(\overline{\mathbf{U}}_1^x)^T] = \frac{1}{n-1}\mathbf{I}_{n-1}. \tag{A.52}$$

Also, since the entries of $1/d_x\underline{\mathbb{U}}^x(\underline{\mathbb{U}}^x)^T$ are bounded by one, we have:

$$\frac{1}{p}\underline{\mathbb{U}}^x(\underline{\mathbb{U}}^x)^T = \mathbf{O}(d_x/p), \tag{A.53}$$

where $\mathbf{O}(u)$ is an $(n-1)\times(n-1)$ matrix whose entries are $O(u)$. Hence:

$$\begin{aligned}(\mathbb{U}^x(\mathbb{U}^x)^T)^{-1}\mathbb{U}^x &= \underline{\mathbb{U}}^x(\underline{\mathbb{U}}^x)^T + \overline{\mathbb{U}}^x(\overline{\mathbb{U}}^x)^T\mathbb{U}^x \\ &= \frac{n-1}{p}(\mathbf{I}_{n-1} + \mathbf{O}(d_x/p))^{-1}\mathbb{U}^x \\ &= \frac{n-1}{p}\mathbb{U}^x(1+O(d_x/p)).\end{aligned} \tag{A.54}$$

Hence, as $p \to \infty$:

$$\begin{aligned}(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x = \\ = (\frac{n-1}{p})^2(\mathbb{U}^x)^T\mathbb{U}^x(1+O(d_x/p)).\end{aligned} \tag{A.55}$$

Thus:

$$\mathbf{D}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x} = \left(\frac{p}{n-1}\mathbf{I}_{n-1}(1+O(d_x/p))\right). \tag{A.56}$$

Combining (A.56) and (A.54) concludes (A.44).

Now we prove relations (A.45) and (A.46). Define the partition $\mathcal{C} = \mathcal{D} \cup \mathcal{D}^c$ of the index set $\mathcal{C}$ defined in (A.11), where $\mathcal{D} = \{\vec{i} = (i_0, i_1, \cdots, i_\delta) : i_0$ is among $p - d_1$ columns of $\mathbb{U}^x$ that are uncorrelated of columns of $\mathbb{U}^y$ and at most one of $i_1, \cdots, i_\delta$ is less than or equal to $d_y\}$ is the set of $(\delta+1)$-tuples restricted to columns of $\mathbb{U}^x$ and

$\mathbb{U}^y$ that are independent. We have:

$$J(\overline{f_{\mathbf{U}^x_\bullet,\mathbf{U}^y_{*_1},\ldots,\mathbf{U}^y_{*_\delta}}}) = |\mathcal{C}|^{-1}2^{-\delta} \sum_{s_1,\ldots,s_\delta\in\{-1,1\}} (\sum_{\vec{i}\in\mathcal{D}} + \sum_{\vec{i}\in\mathcal{D}^c}) J(f_{s_0\mathbf{U}^x_{i_0},s_1\mathbf{U}^y_{i_1},\ldots,s_\delta\mathbf{U}^y_{i_\delta}}), \tag{A.57}$$

and

$$\|\Delta^{xy}_{p,q,n,k,\delta}\|_1 = |\mathcal{C}|^{-1}(\sum_{\vec{i}\in\mathcal{D}} + \sum_{\vec{i}\in\mathcal{D}^c})\Delta^{xy}_{p,q,n,k,\delta}(\vec{i}). \tag{A.58}$$

But, $J(f_{s_0\mathbf{U}^x_{i_0},s_1\mathbf{U}^y_{i_1},\ldots,s_\delta\mathbf{U}^y_{i_\delta}}) = 1$ for $\vec{i}\in\mathcal{D}$ and $\Delta^{xy}_{p,q,n,k,\delta}(\vec{i}) = 0$ for $\vec{i}\in\mathcal{C}$. Moreover, we have:

$$\frac{|\mathcal{D}|}{|\mathcal{C}|} = O(\frac{(p-d_1)(q-d_y+1)^\delta}{pq^\delta}). \tag{A.59}$$

Thus:

$$J(\overline{f_{\mathbf{U}^x_\bullet,\mathbf{U}^y_{*_1},\ldots,\mathbf{U}^y_{*_\delta}}}) = 1 + O(\max\{\frac{d_1}{p}, \delta\frac{(d_y-1)}{q}\}). \tag{A.60}$$

Moreover, since $\tilde{\mathbb{U}}^x = \mathbb{U}^x(1+O(d_x/p))$, $f_{\tilde{\mathbf{U}}^x_{i_0},\mathbf{U}^y_{i_1},\ldots,\mathbf{U}^y_{i_\delta}} = f_{\mathbf{U}^x_{i_0},\mathbf{U}^y_{i_1},\ldots,\mathbf{U}^y_{i_\delta}}(1+O(d_x/p))$. This concludes:

$$J(\overline{f_{\tilde{\mathbf{U}}^x_\bullet,\mathbf{U}^y_{*_1},\ldots,\mathbf{U}^y_{*_\delta}}}) = 1 + O(\max\{\frac{d_x}{p}, \frac{d_1}{p}, \delta\frac{(d_y-1)}{q}\}), \tag{A.61}$$

and

$$\|\Delta^{\tilde{x}y}_{p,q,n,k,\delta}\|_1 = O(d_x/p). \tag{A.62}$$

□

A.4 Predictive Correlation Screening

Under the assumptions of Theorem A.1 and Theorem A.2:

$$p(N^{xy}_{\delta,\rho_p} > 0) \to 1 - \exp(-\xi_{p,q,n,\delta,\rho_p}) \text{ as } p \to \infty \tag{A.63}$$

Using the above limit, approximate p-values can be computed. Fix a degree threshold $\delta \leq q$ and a correlation threshold $\rho^* \in [0, 1]$. Define $\mathcal{G}_{\rho^*}(\mathbf{H}^{xy})$ as the undirected bipartite graph (Fig. A.1) with parts labeled x and y, vertices $\{X_1, X_2, ..., X_p\}$ in part x and $\{Y_1, Y_2, ..., Y_q\}$ in part y. For $1 \leq i \leq p$ and $1 \leq j \leq q$, vertices X_i and Y_j are connected if $|h^{xy}_{ij}| > \rho^*$, where h^{xy}_{ij} is the (i, j)th entry of $\mathbf{H}^{xy}$ defined in (A.9). Denote by d^x_i the degree of vertex X_i in $\mathcal{G}_{\rho^*}(\mathbf{H}^{xy})$. For each value $\delta \in \{1, \cdots, \max_{1\leq i\leq p} d^x_i\}$, and each i, $1 \leq i \leq p$, denote by $\rho_i(\delta)$ the maximum value of the correlation threshold ρ for which $d^x_i \geq \delta$ in $\mathcal{G}_\rho(\mathbf{H}^{xy})$. $\rho_i(\delta)$ is in fact equal to the δth largest value $|h^{xy}_{ij}|, 1 \leq j \leq q$. $\rho_i(\delta)$ can be computed using Approximate Nearest Neighbors (ANN) type algorithms (*Jégou et al.*, 2011; *Arya et al.*, 1998). Now for each i define the modified threshold $\rho^{\text{mod}}_i(\delta)$ as:

$$\rho^{\text{mod}}_i(\delta) = w_i \rho_i(\delta), \qquad 1 \leq i \leq p, \tag{A.64}$$

where $w_i = D(i)/\sum_{j=1}^{p} D(j)$, in which $D(i)$ is the ith diagonal element of the diagonal matrix $\mathbf{D}^{\frac{1}{2}}_{(\mathbb{U}^x)^T(\mathbb{U}^x(\mathbb{U}^x)^T)^{-2}\mathbb{U}^x}$ (recall Sec. A.2).

Using Theorem A.1 and Theorem A.2 the p-value associated with variable X_i at degree level δ can be approximated as:

$$pv_\delta(i) \approx 1 - \exp(-\xi_{p,q,n,\delta,\rho^{\text{mod}}_i(\delta)}). \tag{A.65}$$

The set of p-values (A.65), $i = 1, \ldots, p$, provides a measure of importance of each variable X_i in predicting Y_j's. Under a block-sparsity null hypothesis, the most

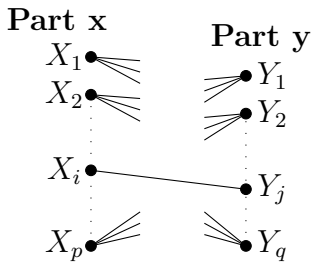

Figure A.1: Predictive correlation screening thresholds the matrix $\mathbf{H}^{xy}$ in (A.10) to find variables X_i that are most predictive of responses Y_j. This is equivalent to finding sparsity in a bipartite graph $\mathcal{G}_{\rho^*}(\mathbf{H}^{xy})$ with parts x and y which have p and q vertices, respectively. For $1 \leq i \leq p$ and $1 \leq j \leq q$, vertex X_i in part x is connected to vertex Y_j in part y if $|h_{ij}^{xy}| > \rho^*$.

important variables would be the ones that have the smallest p-values. Similar to the result in (*Hero and Rajaratnam*, 2011, 2012), there is a phase transition in the p-values as a function of threshold ρ. More exactly, there is a critical threshold $\rho_{c,\delta}$ such that if $\rho > \rho_{c,\delta}$, the average number $E[N_{\delta,\rho}^{xy}]$ of discoveries abruptly decreases to 0 and if $\rho < \rho_{c,\delta}$ the average number of discoveries abruptly increases to p. The value of this critical threshold is:

$$\rho_{c,\delta} = \sqrt{1 - (c_{n,\delta}^{xy} p)^{-2\delta/(\delta(n-2)-2)}}, \tag{A.66}$$

where $c_{n,\delta}^{xy} = \mathrm{a}_n \delta J(\overline{f_{\mathbf{U}_\bullet^x, \mathbf{U}_{*_1}^y, \ldots, \mathbf{U}_{*_\delta}^y}})$. When $\delta = 1$, the expression given in (A.66) is identical, except for the constant $c_{n,\delta}^{xy}$, to the expression (3.14) in (*Hero and Rajaratnam*, 2011).

Expression (A.66) is useful in choosing the PCS correlation threshold ρ^*. Selecting ρ^* slightly greater than $\rho_{c,\delta}$ will prevent the bipartite graph $\mathcal{G}_{\rho^*}(\mathbf{H}^{xy})$ from having an overwhelming number of edges.

Normally $\delta = 1$ would be selected to find all regressor variables predictive of at least 1 response variable Y_j. A value of $\delta = d > 1$ would be used if the experimenter were only interested in variables that were predictive of at least d of the responses.

Pseudo-code for the complete algorithm for variable selection is shown in Fig. 4. The worse case computational complexity of the PCS algorithm is only $O(np \log q)$.

Algorithm 4: Predicive Correlation Screening (PCS) Algorithm

- Initialization:
  1. Choose an initial threshold $\rho^* > \rho_{c,\delta}$;
  2. Calculate the degree of each vertex on side x of the bipartite graph $\mathcal{G}_{\rho^*}(\mathbf{H}^{xy})$;
  3. Select a value of $\delta \in \{1, \cdots, \max_{1 \le i \le p} d_i^x\}$;

for $i = 1$ **to** p **do**

  Find $\rho_i(\delta)$ as the δth greatest element of $\{|h_{ij}|, 1 \le j \le q\}$;
  Compute $\rho_i^{\text{mod}}(\delta)$ using (A.64);
  Approximate the p-value corresponding to the ith independent variable X_i as $pv_\delta(i) \approx 1 - \exp(-\xi_{p,q,n,\delta,\rho_i^{\text{mod}}(\delta)})$;

- Screen variables by thresholding the p-values $pv_\delta(i)$ at desired significance level ;

A.5 Two-stage predictor design

Assume there are a total of t samples $\{\mathbf{Y}_i, \mathbf{X}_i\}_{i=1}^t$ available. During the first stage a number $n \le t$ of these samples are assayed for all p variables and during the second stage the rest of the $t - n$ samples are assayed for a subset of $k \le p$ of the variables. Subsequently, a k-variable predictor is designed using all t samples collected during both stages. The first stage of the PCS predictor is implemented by using the PCS algorithm with $\delta = 1$.

As this two-stage PCS algorithm uses n and t samples in stage 1 and stage 2 respectively, we denote the algorithm above as the $n|t$ algorithm. Experimental results in Sec. A.7 show that for $n \ll p$, if LASSO or correlation learning is used instead of PCS in stage 1 of the two-stage predictor the performance suffers. An asymptotic analysis (as the total number of samples $t \to \infty$) of the above two-stage predictor can be performed to obtain optimal sample allocation rules for stage 1 and stage 2. The

asymptotic analysis discussed in Sec. A.6 provides minimum Mean Squared Error (MSE) under the assumption that n, t, p, and k satisfy the budget constraint:

$$np + (t - n)k \leq \mu, \tag{A.67}$$

where μ is the total budget available. The motivation for this condition is to bound the total sampling cost of the experiment.

A.6 Optimal stage-wise sample allocation

We first give theoretical upper bounds on the Family-Wise Error Rate (FWER) of performing variable selection using p-values obtained via PCS. Then, using the obtained bound, we compute the asymptotic optimal sample size n used in the first stage of the two-stage predictor, introduced in the previous section, to minimize the asymptotic expected MSE.

We assume that the response $\mathbf{Y}$ satisfies the following ground truth model:

$$\mathbf{Y} = \mathbf{a}_{i_1} X_{i_1} + \mathbf{a}_{i_2} X_{i_2} + \cdots + \mathbf{a}_{i_k} X_{i_k} + \mathbf{N}, \tag{A.68}$$

where $\pi_0 = \{i_1, \cdots, i_k\}$ is a set of distinct indices in $\{1, \ldots, p\}$, $\mathbf{X} = [X_1, X_2, \cdots, X_p]$ is the vector of predictors, $\mathbf{Y}$ is the q-dimensional response vector, and $\mathbf{N}$ is a noise vector statistically independent of $\mathbf{X}$. $X_{i_1}, \cdots, X_{i_k}$ are called active variables and the remaining $p - k$ variables are called inactive variables. We assume that the p-dimensional vector $\mathbf{X}$ follows a multivariate normal distribution with mean $\mathbf{0}$ and $p \times p$ covariance matrix $\mathbf{\Sigma} = [\sigma_{ij}]_{1 \leq i,j \leq p}$, where $\mathbf{\Sigma}$ has the following block diagonal structure:

$$\sigma_{ij} = \sigma_{ji} = 0, \ \forall \ i \in \pi_0, j \in \{1, \cdots, p\} \backslash \pi_0. \tag{A.69}$$

In other words active (respectively inactive) variables are only correlated with the other active (respectively inactive) variables. Also, we assume that $\mathbf{N}$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\sigma\mathbf{I}_{q\times q}$.

We use the PCS algorithm of Sec. A.4 with $\delta = 1$ to select the k variables with the smallest p-values. These selected variables will then be used as estimated active variables in the second stage. The following theorem gives an upper bound on the probability of selection error for the PCS algorithm.

Theorem A.3. *If $n \geq \Theta(\log p)$ then with probability at least $1 - q/p$, PCS recovers the exact support π_0.*

Proof. First we prove the theorem for $q = 1$. Without loss of generality assume

$$Y = a_1X_1 + a_2X_2 + \cdots + a_kX_k + \sigma N, \tag{A.70}$$

where N is follows the standard normal distribution. Note that since $q = 1$, $a_1, \cdots, a_k$ are scalars. Defining $\mathbf{b} = \mathbf{\Sigma}^{1/2}\mathbf{a}$, the response Y can be written as:

$$Y = a_1Z_1 + a_2Z_2 + \cdots + a_kZ_k + \sigma N, \tag{A.71}$$

in which $Z_1, \cdots, Z_k$ are i.i.d. standard normal random variables. Assume $\mathbf{U}_1, \cdots, \mathbf{U}_p, \mathbf{U}_N$ represent the U-scores (which are in S_{n-2}) corresponding to $Z_1, \cdots, Z_p, N$, respectively. It is easy to see:

$$\mathbf{U}_y = \frac{b_1\mathbf{U}_1 + b_2\mathbf{U}_2 + \cdots + b_k\mathbf{U}_k + \sigma U_N}{\|b_1\mathbf{U}_1 + b_2\mathbf{U}_2 + \cdots + b_k\mathbf{U}_k + \sigma\mathbf{U}_N\|}. \tag{A.72}$$

If $\mathbf{U}$ and $\mathbf{V}$ are the U-scores corresponding to two random variables, and r is the correlation coefficient between the two random variables, we have:

$$|r| = 1 - \frac{(\min\{\|\mathbf{U} - \mathbf{V}\|, \|\mathbf{U} + \mathbf{V}\|\})^2}{2}. \tag{A.73}$$

Let $r_{y,i}$ represent the sample correlation between Y and X_i. Here, we want to upper bound $\text{prob}\{|r_{y,1}| < |r_{y,k+1}|\}$. We have:

$$\begin{aligned}
\text{prob}\{|r_{y,1}| < |r_{y,k+1}|\} &= \\
\text{prob}\{1 - \frac{(\min\{\|\mathbf{U}_1 - \mathbf{U}_y\|, \|\mathbf{U}_1 + \mathbf{U}_y\|\})^2}{2} &< \\
1 - \frac{(\min\{\|\mathbf{U}_{k+1} - \mathbf{U}_y\|, \|\mathbf{U}_{k+1} + \mathbf{U}_y\|\})^2}{2}\} &= \qquad \text{(A.74)} \\
\text{prob}\{\min\{\|\mathbf{U}_1 - \mathbf{U}_y\|, \|\mathbf{U}_1 + \mathbf{U}_y\|\} &> \\
\min\{\|\mathbf{U}_{k+1} - \mathbf{U}_y\|, \|\mathbf{U}_{k+1} + \mathbf{U}_y\|\}\} &\leq \qquad \text{(A.75)} \\
\text{prob}\{\|\mathbf{U}_1 - \mathbf{U}_y\| &> \\
\min\{\|\mathbf{U}_{k+1} - \mathbf{U}_y\|, \|\mathbf{U}_{k+1} + \mathbf{U}_y\|\}\} &= \\
\text{prob}\{\{\|\mathbf{U}_1 - \mathbf{U}_y\| > \|\mathbf{U}_{k+1} - \mathbf{U}_y\|\} &\cup \\
\{\|\mathbf{U}_1 - \mathbf{U}_y\| > \|\mathbf{U}_{k+1} + \mathbf{U}_y\|\}\} &\leq \\
\text{prob}\{\|\mathbf{U}_1 - \mathbf{U}_y\| > \|\mathbf{U}_{k+1} - \mathbf{U}_y\|\} &+ \\
\text{prob}\{\|\mathbf{U}_1 - \mathbf{U}_y\| > \|\mathbf{U}_{k+1} + \mathbf{U}_y\|\} &= \qquad \text{(A.76)} \\
2\ \text{prob}\{\|\mathbf{U}_1 - \mathbf{U}_y\| > \|\mathbf{U}_{k+1} - \mathbf{U}_y\|\}, & \qquad \text{(A.77)}
\end{aligned}$$

in which, the last inequality holds since $\mathbf{U}_{k+1}$ is uniform over S_{n-2} and is independent of $\mathbf{U}_1$ and $\mathbf{U}_y$. Therefore, it suffices to upper bound $p_1 := \text{prob}\{\|\mathbf{U}_1 - \mathbf{U}_y\| > \|\mathbf{U}_{k+1} - \mathbf{U}_y\|\}$. Define:

$$\mathbf{V} = b_2\mathbf{U}_2 + \cdots + b_k\mathbf{U}_k, \qquad \text{(A.78)}$$

and

$$\mathbf{U}_* = \mathbf{V}/\|\mathbf{V}\|. \qquad \text{(A.79)}$$

By symmetry, $\mathbf{U}_*$ is uniform over S_{n-2}. Hence:

$$\mathbf{U}_y = \frac{b_1\mathbf{U}_1 + \|\mathbf{V}\|\mathbf{U}_*}{\|b_1\mathbf{U}_1 + \|\mathbf{V}\|\mathbf{U}_*\|}. \qquad \text{(A.80)}$$

Since $\|\mathbf{V}\| \leq |b_2| + \cdots + |b_k|$, we have:

$$\frac{|b_1|}{\|\mathbf{V}\|} \geq \frac{|b_1|}{|b_2| + \cdots + |b_k|} = \frac{|b_1|}{c_1}, \tag{A.81}$$

where $c_1 := |b_2| + \cdots + |b_k|$. Define:

$$\theta_1 = \cos^{-1}(\mathbf{U}_y^T \mathbf{U}_1), \tag{A.82}$$

and

$$\theta_1 = \cos^{-1}(\mathbf{U}_y^T \mathbf{U}_*). \tag{A.83}$$

It is easy to see that:

$$\frac{\sin\theta_1}{\sin\theta_2} \leq \frac{c_1}{|b_1|}. \tag{A.84}$$

For each $0 \leq \theta \leq \pi$, define:

$$\beta_1(\theta) = \max_{0 \leq \theta' \leq \pi} \frac{\theta}{\theta + \theta'} \text{ s.t. } \frac{\sin\theta}{\sin\theta'} \leq \frac{c_1}{|b_1|}. \tag{A.85}$$

Now fix the point $\mathbf{U}_1$ on S_{n-2}. Define $f(\theta)$ as the probability distribution of θ_2. Also, define $p(\theta)$ as the probability that the angle between the uniformly distributed (over S_{n-2}) point $\mathbf{U}_{k+1}$ and $\mathbf{U}_y$ is less than θ. Since $\mathbf{U}_1$ is independent of $\mathbf{U}_*$ and $\mathbf{U}_{k+1}$ is independent of $\mathbf{U}_y$, clearly:

$$p(\theta) = \int_0^\theta f(\theta')d\theta'. \tag{A.86}$$

We have:

$$p_1 \leq \int_0^\pi p(\beta_1(\theta)\theta) f(\theta) d\theta$$

$$= \int_0^{\pi/2} \left(p(\beta_1(\theta)\theta) + p(\beta_1(\pi-\theta)(\pi-\theta))\right) f(\theta)d\theta, \tag{A.87}$$

where the last equality holds because $f(\theta) = f(\pi - \theta)$. Noting the fact that:

$$\int_0^{\pi} p(\theta) f(\theta) d\theta = \int_0^{\pi/2} (p(\theta) + p(\pi - \theta)) f(\theta) d\theta = \int_0^{\pi/2} f(\theta) d\theta = \frac{1}{2}. \tag{A.88}$$

we conclude:

$$p_1 \leq \frac{1}{2} - \int_0^{\pi/2} \{(p(\theta) - p(\beta_1(\theta)\theta)) + (p(\pi - \theta) - p(\beta_1(\pi - \theta)(\pi - \theta)))\} f(\theta) d\theta. \tag{A.89}$$

Hence by (A.86), for any $0 < \theta_0 < \pi/2$:

$$p_1 \leq \frac{1}{2} - \int_{\theta_0}^{\pi/2} p_{\gamma_1}(\theta) f(\theta) d\theta, \tag{A.90}$$

in which

$$\begin{aligned} p_{\gamma_1}(\theta) &= p(\theta + \gamma_1\theta) - p(\theta - \gamma_1\theta) \\ &= \text{prob}\{\theta - \gamma_1\theta \leq \theta_2 \leq \theta + \gamma_1\theta\}, \end{aligned} \tag{A.91}$$

with

$$\gamma_1 = \min_{\theta_0 \leq \theta \leq \pi - \theta_0} 1 - \beta_1(\theta) = 1 - \max_{\theta_0 \leq \theta \leq \pi - \theta_0} \beta_1(\theta). \tag{A.92}$$

It is easy to check that $\gamma_1 > 0$. Therefore, since $p_{\gamma_1}(\theta)$ is an increasing functions of θ for $0 \leq \theta \leq \pi/2$, we conclude:

$$p_1 \leq \frac{1}{2} - \int_{\theta_0}^{\pi/2} p_{\gamma_1}(\theta_0) f(\theta) d\theta. \tag{A.93}$$

Choose θ_0 so that $\theta_0 = \frac{\pi}{2+\gamma_1}$. We have:

$$
\begin{aligned}
p_1 &\leq \frac{1}{2} - p_{\gamma_1}(\pi/(2+\gamma_1)) \int_{\pi/(2+\gamma_1)}^{\pi/2} f(\theta)d\theta \\
&= \frac{1}{2} - \int_{\pi(1-\gamma_1)/(2+\gamma_1)}^{\pi(1+\gamma_1)/(2+\gamma_1)} f(\theta)d\theta \int_{\pi/(2+\gamma_1)}^{\pi/2} f(\theta)d\theta \\
&\leq \frac{1}{2} - \int_{\pi/2-\gamma_1\pi/6}^{\pi/2+\gamma_1\pi/6} f(\theta)d\theta \int_{\pi/2-\gamma_1\pi/6}^{\pi/2} f(\theta)d\theta \\
&\leq \frac{1}{2} - 2\left(\int_{\pi/2-\gamma_1\pi/6}^{\pi/2} f(\theta)d\theta\right)^2,
\end{aligned} \tag{A.94}
$$

in which, the last inequality holds, since $0 < \gamma_1 < 1$. Defining $\lambda_1 = \sin(\pi/2 - \gamma_1\pi/6)$ and using the formula for the area of the spherical cap, we will have:

$$
\begin{aligned}
&\int_{\pi/2-\gamma_1\pi/6}^{\pi/2} f(\theta)d\theta = \\
&\frac{I_1((n-2)/2, 1/2) - I_{\lambda_1}((n-2)/2, 1/2)}{2I_1((n-2)/2, 1/2)},
\end{aligned} \tag{A.95}
$$

in which

$$
I_x(a,b) = \frac{\int_0^x t^{a-1}(1-t)^{b-1}dt}{\int_0^1 t^{a-1}(1-t)^{b-1}dt}, \tag{A.96}
$$

is the regularized incomplete beta function. Hence:

$$
\int_{\pi/2-\gamma_1\pi/6}^{\pi/2} f(\theta)d\theta = \frac{\int_{\lambda_1}^1 t^{(n-4)/2}/\sqrt{1-t}dt}{2\int_0^1 t^{(n-4)/2}/\sqrt{1-t}dt}. \tag{A.97}
$$

Note that we have:

$$\begin{aligned}
&\frac{\int_{\lambda_1}^{1} t^{(n-4)/2}/\sqrt{1-t}dt}{2\int_{0}^{\lambda_1} t^{(n-4)/2}/\sqrt{1-t}dt} \\
&\geq \frac{\int_{\lambda_1}^{1} t^{(n-4)/2}/\sqrt{1-\lambda_1}dt}{2\int_{0}^{1} t^{(n-4)/2}/\sqrt{1-\lambda_1}dt} \\
&= \frac{1-\lambda_1^{(n-2)/2}}{\lambda_1^{(n-2)/2}} := \kappa_1. \qquad \text{(A.98)}
\end{aligned}$$

Hence:

$$\begin{aligned}
\int_{\pi/2-\gamma_1\pi/6}^{\pi/2} f(\theta)d\theta &\geq \frac{\int_{\lambda_1}^{1} t^{(n-4)/2}/\sqrt{1-t}dt}{2(1+1/\kappa_1)\int_{\lambda_1}^{1} t^{(n-4)/2}/\sqrt{1-t}dt} \\
&= \frac{\kappa_1}{2(\kappa_1+1)} = \frac{1-\lambda_1^{(n-2)/2}}{2}. \qquad \text{(A.99)}
\end{aligned}$$

Hence by (A.94):

$$p_1 \leq \lambda_1^{(n-2)/2} - \lambda_1^{n-2} \leq \lambda_1^{(n-2)/2}. \qquad \text{(A.100)}$$

Therefore, p_1 decreases at least exponentially by n.

Assume $P(i)$ for $1 \leq i \leq k$, represents the probability that the active variable X_i is not among the selected k variables. By (A.77) and using the union bound we have:

$$P(1) \leq 2(p-k)\lambda_1^{(n-2)/2}. \qquad \text{(A.101)}$$

Similar inequalities can be obtained for $P(2), \cdots, P(k)$ which depend on $\lambda_2, \cdots, \lambda_k$, respectively. Finally, using the union bound, the probability P that all the active variables are correctly selected satisfies:

$$P \geq 1 - 2(p-k)\sum_{i=1}^{k} \lambda_i^{(n-2)/2} \geq 1 - 2k(p-k)\lambda^{(n-2)/2}, \qquad \text{(A.102)}$$

where $\lambda := \max_{1\leq i\leq k} \lambda_i$. This concludes that if $n = \Theta(\log p)$, with probability at least

$1 - 1/p$ the exact support can be recovered using PCS.

For $q > 1$, by union bound, the probability of error becomes at most q times larger and this concludes the statement of theorem A.3. □

Theorem A.3 can be compared to Theorem 1 in (*Obozinski et al.*, 2008) for recovering the support π_0 by minimizing a LASSO-type objective function. The constant in $\Theta(\log p)$ of Theorem A.3 is increasing in the dynamic range coefficient

$$\max_{i=1,\cdots,q} \frac{|\pi_0|^{-1} \sum_{j\in\pi_0} |b_{ij}|}{\min_{j\in\pi_0} |b_{ij}|} \in [1, \infty), \tag{A.103}$$

where $\mathbf{B} = [\mathbf{b}_1, \cdots, \mathbf{b}_p] = \mathbf{\Sigma}^{1/2}\mathbf{A}$. The worst case (largest constant in $\Theta(\log p)$) occurs when there is high dynamic range in some rows of the $q \times p$ matrix $\mathbf{B}$.

The following theorem states the optimal sample allocation rule for the two-stage predictor, as $t \to \infty$.

Theorem A.4. *The optimal sample allocation rule for the two-stage predictor introduced in Sec. A.5 under the cost condition* (A.67) *is*

$$n = \begin{cases} O(\log t), & c(p-k)\log t + kt \le \mu \\ 0, & o.w. \end{cases} \tag{A.104}$$

Proof. Proof is similar to the proof of Theorem IV.12 and is omitted. □

A.7 Simulation results

Simulation results for the case of $q > 1$ can be found in Sec. 4.4.

A.8 Conclusion

In this Appendix, we proposed a generalization of the SPARCS algorithm presented in Chapter IV to the case of multi-dimensional response. Similar to SPARCS,

this generalization is specifically useful in cases where $n \ll p$ and the high cost of assaying all regressor variables justifies a two-stage design: high throughput variable selection followed by predictor construction using fewer selected variables. Asymptotic analysis and experiments showed advantages of PCS compared to LASSO.

BIBLIOGRAPHY

BIBLIOGRAPHY

Albert, A., and J. Anderson (1984), On the existence of maximum likelihood estimates in logistic regression models, *Biometrika*, *71*(1), 1–10.

Anderson, T. W. (2003), An introduction to multivariate statistical analysis.

Arratia, R., L. Goldstein, and L. Gordon (1990), Poisson approximation and the chen-stein method, *Statistical Science*, *5*(4), 403–424.

Arya, S., D. Mount, N. Netanyahu, R. Silverman, and A. Wu (1998), An optimal algorithm for approximate nearest neighbor searching fixed dimensions, *Journal of the ACM (JACM)*, *45*(6), 891–923.

Audibert, J.-Y., R. Munos, and C. Szepesvári (2007), Tuning bandit algorithms in stochastic environments, in *Algorithmic Learning Theory*, pp. 150–165, Springer.

Bechhofer, R. E., J. Kiefer, and M. Sobel (1968), *Sequential identification and ranking procedures: with special reference to Koopman-Darmois populations*, vol. 3, University of Chicago Press Chicago.

Bickel, P. J., and E. Levina (2008), Covariance regularization by thresholding, *The Annals of Statistics*, pp. 2577–2604.

Bien, J., and R. J. Tibshirani (2011), Sparse estimation of a covariance matrix, *Biometrika*, *98*(4), 807–820.

Biguesh, M., and A. B. Gershman (2006), Training-based mimo channel estimation: a study of estimator tradeoffs and optimal training signals, *Signal Processing, IEEE Transactions on*, *54*(3), 884–893.

Bishop, C. M., et al. (2006), *Pattern recognition and machine learning*, vol. 1, springer New York.

Bühlmann, P. (2006), Boosting for high-dimensional linear models, *The Annals of Statistics*, *34*(2), 559–583.

Bühlmann, P., and S. Van De Geer (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer.

Bullmore, E., and O. Sporns (2009), Complex brain networks: graph theoretical analysis of structural and functional systems, *Nature Reviews Neuroscience*, *10*(3), 186–198.

Cai, T., and W. Liu (2011), Adaptive thresholding for sparse covariance matrix estimation, *Journal of the American Statistical Association*, *106*(494), 672–684.

Candes, E., and J. Romberg (2007), Sparsity and incoherence in compressive sampling, *Inverse problems*, *23*(3), 969.

Candés, E., J. Romberg, and T. Tao (2005), Stable signal recovery from incomplete and inaccurate measurements, *Comm. Pure Appl. Math.*, *59*, 1207–1223.

Carin, L., D. Liu, and B. Guo (2011), Coherence, compressive sensing, and random sensor arrays, *Antennas and Propagation Magazine, IEEE*, *53*(4), 28–39.

Chaudhuri, S., M. Drton, and T. S. Richardson (2007), Estimation of a covariance matrix with zeros, *Biometrika*, *94*(1), 199–216.

Chen, X., M. Xu, W. B. Wu, et al. (2013), Covariance and precision matrix estimation for high-dimensional time series, *The Annals of Statistics*, *41*(6), 2994–3021.

Conway, J. B. (1990), *A course in functional analysis*, vol. 96, Springer.

Ding, C., X. He, H. Zha, and H. D. Simon (2002), Adaptive dimension reduction for clustering high dimensional data, in *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pp. 147–154, IEEE.

Donoho, D. L. (2006), Compressed sensing, *Information Theory, IEEE Transactions on*, *52*(4), 1289–1306.

Durrett, R. (2010), *Probability: theory and examples*, vol. 3, Cambridge university press.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004), Least angle regression, *The Annals of statistics*, *32*(2), 407–499.

Fan, J., and J. Lv (2008), Sure independence screening for ultrahigh dimensional feature space, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(5), 849–911.

Fan, J., R. Song, et al. (2010), Sure independence screening in generalized linear models with np-dimensionality, *The Annals of Statistics*, *38*(6), 3567–3604.

Figueiredo, M. A., R. D. Nowak, and S. J. Wright (2007), Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems, *Selected Topics in Signal Processing, IEEE Journal of*, *1*(4), 586–597.

Filiz, I. O., X. Guo, J. Morton, and B. Sturmfels (2012), Graphical models for correlated defaults, *Mathematical Finance*, *22*(4), 621–644.

Firouzi, H., B. Rajaratnam, and A. Hero III (2013), Predictive correlation screening: Application to two-stage predictor design in high dimension, in *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pp. 274–288.

Forman, G. (2003), An extensive empirical study of feature selection metrics for text classification, *The Journal of machine learning research*, *3*, 1289–1305.

Friedman, J., T. Hastie, and R. Tibshirani (2001), *The elements of statistical learning*, vol. 1, Springer Series in Statistics.

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

Friston, K. J., J. T. Ashburner, S. J. Kiebel, T. E. Nichols, and W. D. Penny (2011), *Statistical Parametric Mapping: The Analysis of Functional Brain Images: The Analysis of Functional Brain Images*, Academic Press.

Genovese, C., J. Jin, and L. Wasserman (2009), Revisiting marginal regression.

Genovese, C. R., J. Jin, L. Wasserman, and Z. Yao (2012), A comparison of the lasso and marginal regression, *The Journal of Machine Learning Research*, *98888*, 2107–2143.

Golub, T. R., et al. (1999), Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *science*, *286*(5439), 531–537.

Gómez, E., M. Gomez-Viilegas, and J. Marin (1998), A multivariate generalization of the power exponential family of distributions, *Communications in Statistics-Theory and Methods*, *27*(3), 589–600.

Gray, R. M. (2006), Toeplitz and circulant matrices: A review, *Foundations and Trends in Communications and Information Theory*, *2*(3), 155–239, doi:10.1561/0100000006.

Grenander, U., and G. Szegő (1958), *Toeplitz forms and their applications*, Univ of California Press.

Gupta, A. K., T. Varga, and T. Bodnar (2013), *Elliptically contoured models in statistics and portfolio theory*, Springer.

Gupta, S. S., and S. Panchapakesan (1991), Sequential ranking and selection procedures, *Handbook of sequential analysis*, pp. 363–380.

Guyon, I., and A. Elisseeff (2003), An introduction to variable and feature selection, *The Journal of Machine Learning Research*, *3*, 1157–1182.

Hamilton, J. D. (1994), *Time series analysis*, vol. 2, Princeton university press Princeton.

Hassibi, B., and B. Hochwald (2003), How much training is needed in multiple-antenna wireless links?, *Information Theory, IEEE Transactions on*, *49*(4), 951–963, doi:10.1109/TIT.2003.809594.

Haupt, J., R. M. Castro, and R. Nowak (2011), Distilled sensing: Adaptive sampling for sparse detection and estimation, *Information Theory, IEEE Transactions on*, *57*(9), 6222–6235.

Haupt, J. D., R. G. Baraniuk, R. M. Castro, and R. D. Nowak (2009), Compressive distilled sensing: Sparse recovery using adaptivity in compressive measurements, in *Signals, Systems and Computers, 2009 Conference Record of the Forty-Third Asilomar Conference on*, pp. 1551–1555, IEEE.

He, Y., and A. Evans (2010), Graph theoretical modeling of brain connectivity, *Current opinion in neurology*, *23*(4), 341–350.

Hero, A., and B. Rajaratnam (2011), Large-scale correlation screening, *Journal of the American Statistical Association*, *106*(496), 1540–1552.

Hero, A., and B. Rajaratnam (2012), Hub discovery in partial correlation graphs, *Information Theory, IEEE Transactions on*, *58*(9), 6064–6078.

Hesterberg, T., N. H. Choi, L. Meier, C. Fraley, et al. (2008), Least angle and âĎŞ1 penalized regression: A review, *Statistics Surveys*, *2*, 61–93.

Huang, J. C., and N. Jojic (2011), Variable selection through correlation sifting, in *Research in Computational Molecular Biology*, pp. 106–123, Springer.

Huang, Y., et al. (2011), Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection, *PLoS genetics*, *7*(8), e1002,234.

Jégou, H., M. Douze, and C. Schmid (2011), Product quantization for nearest neighbor search, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *33*(1), 117–128.

Karoui, N. E. (2008), Operator norm consistent estimation of large-dimensional sparse covariance matrices, *The Annals of Statistics*, pp. 2717–2756.

Khan, J. A., S. Van Aelst, and R. H. Zamar (2007), Robust linear model selection based on least angle regression, *Journal of the American Statistical Association*, *102*(480), 1289–1299.

Kim, J., and H. Park (2010), Fast active-set-type algorithms for l1-regularized linear regression, *Proc. AISTAT*, pp. 397–404.

Kim, S.-J., K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky (2007), An interior-point method for large-scale l 1-regularized least squares, *Selected Topics in Signal Processing, IEEE Journal of*, *1*(4), 606–617.

Koh, K., S.-J. Kim, and S. P. Boyd (2007), An interior-point method for large-scale l1-regularized logistic regression., *Journal of Machine learning research*, *8*(8), 1519–1555.

Lam, C., and J. Fan (2009), Sparsistency and rates of convergence in large covariance matrix estimation, *Annals of statistics*, *37*(6B), 4254.

Lauritzen, S. L. (1996), *Graphical models*, Oxford University Press.

Levina, E., and R. Vershynin (2012), Partial estimation of covariance matrices, *Probability Theory and Related Fields*, *153*(3-4), 405–419.

Li, S. (2011), Concise formulas for the area and volume of a hyperspherical cap, *Asian Journal of Mathematics and Statistics*, *4*(1), 66–70.

Li, Y., M. T. Thai, and W. Wu (2008), *Wireless sensor networks and applications*, Springer.

Liu, Y., V. Chandrasekaran, A. Anandkumar, and A. S. Willsky (2012), Feedback message passing for inference in gaussian graphical models, *Signal Processing, IEEE Transactions on*, *60*(8), 4135–4150.

Micheas, A. C., D. K. Dey, and K. V. Mardia (2006), Complex elliptical distributions with application to shape analysis, *Journal of statistical planning and inference*, *136*(9), 2961–2982.

Newman, M. E., and D. J. Watts (1999), Scaling and percolation in the small-world network model, *Physical Review E*, *60*(6), 7332.

Obozinski, G., M. J. Wainwright, M. I. Jordan, et al. (2011), Support union recovery in high-dimensional multivariate regression, *The Annals of Statistics*, *39*(1), 1–47.

Obozinski, G. R., M. J. Wainwright, and M. I. Jordan (2008), High-dimensional support union recovery in multivariate regression, in *Advances in Neural Information Processing Systems*, pp. 1217–1224.

Oppenheim, A. V., R. W. Schafer, J. R. Buck, et al. (1989), *Discrete-time signal processing*, vol. 2, Prentice-hall Englewood Cliffs.

Paffenroth, R., P. du Toit, R. Nong, L. Scharf, A. P. Jayasumana, and V. Bandara (2013), Space-time signal processing for distributed pattern detection in sensor networks, *Selected Topics in Signal Processing, IEEE Journal of*, *7*(1), 38–49.

Paul, D., E. Bair, T. Hastie, and R. Tibshirani (2008), " preconditioning" for feature selection and regression in high-dimensional problems, *The Annals of Statistics*, pp. 1595–1618.

Quattoni, A., X. Carreras, M. Collins, and T. Darrell (2009), An efficient projection for l 1,âĹđ regularization, in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 857–864, ACM.

Rajaratnam, B., S. Roberts, D. Sparks, and O. Dalal (2014), The deterministic bayesian lasso, *arXiv preprint arXiv:1401.2480*.

Rong, Y. (2011), *Practical environmental statistics and data analysis*, ILM Publications.

Rothman, A. J., E. Levina, and J. Zhu (2009), Generalized thresholding of large covariance matrices, *Journal of the American Statistical Association*, *104*(485), 177–186.

Sadilek, A., H. Kautz, and J. P. Bigham (2012), Finding your friends and following them to where you are, in *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 723–732, ACM.

Severini, T. A. (2005), *Elements of distribution theory*, vol. 17, Cambridge University Press.

Simon, M. K. (2007), *Probability distributions involving Gaussian random variables: A handbook for engineers and scientists*, Springer.

Stanley, M., S. Gervais-Ducouret, and J. Adams (2012), Intelligent sensor hub benefits for wireless sensor networks, in *Sensors Applications Symposium (SAS), 2012 IEEE*, pp. 1–6, IEEE.

Suh, C., S. C. Sieg, M. J. Heying, J. H. Oliver, W. F. Maier, and K. Rajan (2009), Visualization of high-dimensional combinatorial catalysis data, *Journal of combinatorial chemistry*, *11*(3), 385–392.

Sun, L., S. Ji, and J. Ye (2008), Hypergraph spectral learning for multi-label classification, in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 668–676, ACM.

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

Tropp, J. A., and A. C. Gilbert (2007), Signal recovery from random measurements via orthogonal matching pursuit, *Information Theory, IEEE Transactions on*, *53*(12), 4655–4666.

Tsay, R. S. (2005), *Analysis of financial time series*, vol. 543, Wiley. com.

Vuran, M. C., Ö. B. Akan, and I. F. Akyildiz (2004), Spatio-temporal correlation: theory and applications for wireless sensor networks, *Computer Networks*, *45*(3), 245–259.

Wainwright, M. J. (2009), Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso), *Information Theory, IEEE Transactions on*, *55*(5), 2183–2202.

Wald, A., et al. (1945), Sequential tests of statistical hypotheses, *Annals of Mathematical Statistics*, *16*(2), 117–186.

Wauthier, F. L., N. Jojic, and M. Jordan (2013), A comparative framework for preconditioned lasso algorithms, in *Advances in Neural Information Processing Systems*, pp. 1061–1069.

Wei, D., and A. O. Hero (2013a), Multistage adaptive estimation of sparse signals, *Selected Topics in Signal Processing, IEEE Journal of*, *7*(5), 783–796.

Wei, D., and A. O. Hero (2013b), Performance guarantees for adaptive estimation of sparse signals, *arXiv preprint arXiv:1311.6360*.

Wen, Z., W. Yin, D. Goldfarb, and Y. Zhang (2010), A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation, *SIAM Journal on Scientific Computing*, *32*(4), 1832–1857.

Wen, Z., W. Yin, H. Zhang, and D. Goldfarb (2012), On the convergence of an active-set method for âĎŞ1 minimization, *Optimization Methods and Software*, *27*(6), 1127–1146.

Wiesel, A., Y. C. Eldar, and A. O. Hero (2010), Covariance estimation in decomposable gaussian graphical models, *Signal Processing, IEEE Transactions on*, *58*(3), 1482–1492.

Yu, J., D. Tao, and M. Wang (2012), Adaptive hypergraph learning and its application in image classification, *Image Processing, IEEE Transactions on*, *21*(7), 3262–3272.

Yuan, M. (2010), High dimensional inverse covariance matrix estimation via linear programming, *The Journal of Machine Learning Research*, *11*, 2261–2286.

Yuan, M., and Y. Lin (2005), Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67.

Zhang, P., Y. Huang, S. Shekhar, and V. Kumar (2003), Correlation analysis of spatial time series datasets: A filter-and-refine approach, in *Advances in Knowledge Discovery and Data Mining*, pp. 532–544, Springer.

Zhou, D., J. Huang, and B. Schölkopf (2006), Learning with hypergraphs: Clustering, classification, and embedding, in *Advances in neural information processing systems*, pp. 1601–1608.