2017

Canonical Correlation Analysis And Network Data Modeling: Statistical And Computational Properties

Zhuang Ma
University of Pennsylvania, kop.mazhuang@gmail.com

Recommended Citation

Ma, Zhuang, "Canonical Correlation Analysis And Network Data Modeling: Statistical And Computational Properties" (2017). *Publicly Accessible Penn Dissertations*. 2460.
https://repository.upenn.edu/edissertations/2460

Canonical Correlation Analysis And Network Data Modeling: Statistical And Computational Properties

Abstract

Classical decision theory evaluates an estimator mostly by its statistical properties, either the closeness to the underlying truth or the predictive ability for new observations. The goal is to find estimators to achieve statistical optimality. Modern "Big Data" applications, however, necessitate efficient processing of large-scale ("big-n-big-p'") datasets, which poses great challenge to classical decision-theoretic framework which seldom takes into account the scalability of estimation procedures. On the one hand, statistically optimal estimators could be computationally intensive and on the other hand, fast estimation procedures might suffer from a loss of statistical efficiency. So the challenge is to kill two birds with one stone. This thesis brings together statistical and computational perspectives to study canonical correlation analysis (CCA) and network data modeling, where we investigate both the optimality and the scalability of the estimators. Interestingly, in both cases, we find iterative estimation procedures based on non-convex optimization can significantly reduce the computational cost and meanwhile achieve desirable statistical properties.

In the first part of the thesis, motivated by the recent success of using CCA to learn low-dimensional feature representations of high-dimensional objects, we propose novel metrics which quantify the estimation loss of CCA by the excess prediction loss defined through a prediction-after-dimension-reduction framework. These new metrics have rich statistical and geometric interpretations, which suggest viewing CCA estimation as estimating the subspaces spanned by the canonical variates.

We characterize, with minimal assumptions, the non-asymptotic minimax rates under the proposed error metrics, especially how the minimax rates depend on the key quantities including the dimensions, the condition number of the covariance matrices and the canonical correlations. Finally, by formulating sample CCA as a non-convex optimization problem, we propose an efficient (stochastic) first order algorithm which scales to large datasets.

In the second part of the thesis, we propose two universal fitting algorithms for networks (possibly with edge covariates) under latent space models: one based on finding the exact maximizer of a convex surrogate of the non-convex likelihood function and the other based on finding an approximate optimizer of the original non-convex objective. Both algorithms are motivated by a special class of inner-product models but are shown to work for a much wider range of latent space models which allow the latent vectors to determine the connection probability of the edges in flexible ways. We derive the statistical rates of convergence of both algorithms and characterize the basin-of-attraction of the non-convex approach. The effectiveness and efficiency of the non-convex procedure is demonstrated by extensive simulations and real-data experiments.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Statistics

First Advisor
Dean P. Foster

Second Advisor
Zongming Ma

Keywords
Canonical Correlation Analysis, computational efficiency, dimension reduction, minimax rates, network data modeling, non-convex optimization

Subject Categories
Statistics and Probability

CANONICAL CORRELATION ANALYSIS AND NETWORK DATA MODELING: STATISTICAL AND COMPUTATIONAL PROPERTIES

Zhuang Ma

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

Dean P. Foster
Marie and Joseph Melone Professor
Professor Emeritus of Statistics

Co-Supervisor of Dissertation

Zongming Ma
Associate Professor of Statistics

Graduate Group Chairperson

Catherine Schrand, Celia Z. Moh Professor, Professor of Accounting

Dissertation Committee

Dean P. Foster, Marie and Joseph Melone Professor, Professor Emeritus of Statistics
Zongming Ma, Associate Professor of Statistics
Lawrence D. Brown, Miers Busch Professor, Professor of Statistics
Robert A. Stine, Professor of Statistics

CANONICAL CORRELATION ANALYSIS AND NETWORK DATA MODELING:
STATISTICAL AND COMPUTATIONAL PROPERTIES

Dedicated to Mom and Dad

ACKNOWLEDGEMENTS

The last four years as a Ph.D. student have been a special and memorable journey with mixed ups and downs, which I cannot possibly finish without tremendous and unconditional love and support from my advisors, friends and families.

First and foremost, I would like to express my sincere gratitude to my advisors Dean Foster and Zongming Ma for their guidance, encouragement and trust, which have made me an independent researcher. I was given maximum freedom to explore research topics I am passionate about. The discussions with them are always inspiring and stimulating. Thank Dean for bringing me to areas beyond classical statistics and raising my interest in natural language processing and deep learning. Thank Zongming for introducing me to network data modeling which interlaces statistics, optimization and real data analysis.

I am also deeply grateful to Lawrence Brown and Robert Stine for sitting in my thesis committee, being great collaborators and sharing their wisdom in both research and life. In particular, I thank Bob for constantly encouraging and reminding me to pursue what I am truly passionate about. I would also like to thank Michael Collins for hosting me at Google Research and being a great mentor. Mike has showed me how theoretical intuitions and domain knowledge could be transformed into state-of-art empirical results. I learned enormously from his expertise in natural language processing and methods of scientific research.

I am very fortunate to have collaborated with many senior colleagues in the department. They generously shared their personal experience and offered me valuable suggestions which have smoothed my research path. Special thanks go to Xiaodong Li, Yichao Lu and Asaf Weinstein. Xiaodong is a role model and always pursues deeper understanding and sharper results. Yichao has showed me how to keep good balance between theory and practice. Asaf's critical thinking is enlightening and thought-provoking.

Thanks go to all my friends at Penn, including but not limited to Sameer Deshpande, Wenjie Dou, Zijian Guo, Colman Humphrey, Kwon Sang Lee, Shaokun Li, Xinyao Ji, Matt Olson, Peichao Peng, Xin Lu Tan and Linjun Zhang. In particular, I thank Peichao for his help all the way through.

Thanks also go to my high school and college friends, especially Kaifeng Chen, Huan He, Yang Kang, Lin Li, Weifeng Sun, Jie Wang, Jun Wang, Jian Yin, Xun Zhang, Ke Zhao, and Yunqing Zheng, for all the old days we spent together. It is always delightful and refreshing when those memories come up every now and then.

Finally, I am indebted to my parents Donglei Ma and Yuanying Zhang for their continuing love and support. They are the ultimate source of my strength and confidence.

Zhuang Ma
Philadelphia, PA
April, 2017

ABSTRACT

CANONICAL CORRELATION ANALYSIS AND NETWORK DATA MODELING: STATISTICAL AND COMPUTATIONAL PROPERTIES

Zhuang Ma

Dean P. Foster

Zongming Ma

Classical decision theory evaluates an estimator mostly by its statistical properties, either the closeness to the underlying truth or the predictive ability for new observations. The goal is to find estimators to achieve statistical optimality. Modern “Big Data” applications, however, necessitate efficient processing of large-scale (“big-n-big-p”) datasets, which poses great challenge to classical decision-theoretic framework which seldom takes into account the scalability of estimation procedures. On the one hand, statistically optimal estimators could be computationally intensive and on the other hand, fast estimation procedures might suffer from a loss of statistical efficiency. So the challenge is to kill two birds with one stone. This thesis brings together statistical and computational perspectives to study canonical correlation analysis (CCA) and network data modeling, where we investigate both the optimality and the scalability of the estimators. Interestingly, in both cases, we find iterative estimation procedures based on non-convex optimization can significantly reduce the computational cost and meanwhile achieve desirable statistical properties.

In the first part of the thesis, motivated by the recent success of using CCA to learn low-dimensional feature representations of high-dimensional objects, we propose novel metrics which quantify the estimation loss of CCA by the excess prediction loss defined through a prediction-after-dimension-reduction framework. These new metrics have rich statistical and geometric interpretations, which suggest viewing CCA estimation as estimating the subspaces spanned by the canonical variates. We characterize, with minimal assumptions, the non-asymptotic minimax rates under the proposed error metrics, especially how the minimax rates depend on the key quantities including the dimensions, the condition number of the covariance matrices and the canonical correlations. Finally, by formulating sample CCA as a non-convex optimization problem, we propose an efficient (stochastic) first order algorithm which scales to large datasets.

In the second part of the thesis, we propose two universal fitting algorithms for networks (possibly with edge covariates) under latent space models: one based on finding the exact maximizer of a convex surrogate of the non-convex likelihood function and the other based on finding an approximate optimizer of the original non-convex objective. Both algorithms are motivated by a special class of inner-product models but are shown to work for a much wider range of latent space models which allow the latent vectors to determine the connection probability of the edges in flexible ways. We derive the statistical rates of convergence of both algorithms and characterize the basin-of-attraction of the non-convex approach. The effectiveness and efficiency of the non-convex procedure is demonstrated by extensive simulations and real-data experiments.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 : Introduction

The age of “Big Data” features cheap and easy availability of large quantities of massive, high-dimensional, complex datasets, the analysis of which interlaces statistics, machine learning and numerical optimization. The challenge/goal is to extract low-dimensional structures from high-dimensional complex objects in a statistically optimal and computationally efficient manner. This thesis brings together statistical and computational perspectives in the study of canonical correlation analysis (CCA) and network data modeling. We aim to answer questions such as:

> What are the proper error metrics to quantify the estimation/prediction loss? Under such metrics, what are the quantities that characterize the fundamental statistical limits (e.g. the minimax rates)? To achieve the optimal error rates on large datasets, what are the efficient algorithms?

1.1. Canonical Correlation Analysis

Canonical correlation analysis (CCA), first introduced by Hotelling (1936), is a fundamental statistical tool to characterize the relationship between two groups of random variables and finds a wide range of applications across many different fields. In recent years, CCA has been successfully applied to learning low dimensional representations of high dimensional objects like images (Rasiwasia et al., 2010), text (Dhillon et al., 2011) and speeches (Arora and Livescu, 2013). Meanwhile, a parallel line of research builds up the theoretical fundation for CCA to achieve sufficient dimension reduction (Kakade and Foster (2007); Foster et al. (2008); Sridharan and Kakade (2008); Fukumizu et al. (2009); Chaudhuri et al. (2009) and many others), especially under a popular multi-view setup.

Motivated by such empirical and theoretical success of CCA, we revisit CCA with a new statistical and computational perspective. Theoretical understanding of the estimation of CCA dates back to the study of the asymptotic distribution of the sample canonical coefficients and sample canonical vectors like Hsu (1941), Izenman (1975), Anderson (1984, 1999) and many others. More recently, Chen et al. (2013) and Gao et al. (2014, 2015b) established the non-asymptotic minimax rates of sparse CCA in a high dimensional setup. Furthermore, there has been a surge of interest in developing scalable algorithms for estimating CCA, to name a few, Avron et al. (2013) Lu and Foster (2014) (before the proposed algorithm was published), Ge et al. (2016a), Wang et al. (2016) and Allen-Zhu and Li (2016) (after the proposed algorithm was published).

Compared with previous work, our major contributions are as follows:

1. We propose novel metrics to quantify the estimation loss of CCA by the excess prediction loss defined through a prediction-after-dimension-reduction framework. These new metrics have rich statistical and geometric interpretations, which suggest viewing CCA estimation as estimating the subspaces spanned by the canonical variates.

2. We characterize, with minimal assumptions, the non-asymptotic minimax rates under the proposed error metrics, especially how the minimax rates depend on key quantities including the dimensions, the condition number of the covariance matrices and the canonical correlations. To the best of our knowledge, this is the first finite sample result that fully captures the effect of the canonical correlations on the minimax rates.

3. We propose an efficient (stochastic) first-order algorithm to compute the leading-k dimensional sample canonical vectors. Compared with the well-known closed form solution, the proposed iterative algorithm avoids multiplying/inverting/factoring large matrices and only requires minimal memory space.

1.2. Network Data Modeling

Network produces a prevalent form of data for quantitative and qualitative analysis in many areas, including but not limited to sociology, engineering and neuroscience. Real-world networks exhibit complex characteristics such as degree heterogeneity, transitivity, homophily and community structure. These pose significant challenges to statistical modeling. To date, researchers have proposed a collection of network models in various fields. These models aim to catch different subsets of the foregoing characteristics, and Goldenberg et al. (2010) provides a comprehensive overview. An important class of network models are *latent space models* (Hoff et al., 2002). Suppose there are n nodes in the observed network. The key idea underlying latent space modeling is that each node i can be represented by a vector z_i in some low dimensional Euclidean space (or some other metric space of choice) that is sometimes called the social space, and nodes that are "close" in the social space are more likely to be connected. Hoff et al. (2002) considered two types of latent space models: distance models and projection models. In both cases, the latent vectors $\{z_i\}_{i=1}^n$ were treated as fixed effects. Later, a series of papers (Hoff, 2003; Handcock et al., 2007; Krivitsky et al., 2009) generalized the original proposal in Hoff et al. (2002) for better modeling of other characteristics of social networks, such as clustering, degree heterogeneity, etc. In these generalizations, the z_i's were treated as random effects generated from certain multivariate Gaussian mixtures. Model fitting and inference in these models has been carried out via Markov Chain Monte Carlo, and it is difficult to scale these methodologies to handle large networks (Goldenberg et al., 2010). Moreover, one needs to use different likelihood function based on choice of model and there is little understanding of the quality of fitting when the model is mis-specified. Albeit these disadvantages, latent space models are attractive due to their friendliness to interpretation and visualization.

We make progress on tackling the foregoing two issues simultaneously in this thesis, which we summarize as the following main contributions:

1. We propose a special class of latent space models, called inner-product models, which

are able to characterize degree heterogeneity, transitivity, homophily and community structure. We design two fitting algorithms: one based on finding the exact maximizer of a convex surrogate of the non-convex likelihood function and the other based on directly finding an approximate optimizer of the non-convex objective. We derive high probability error bounds for both algorithms and characterize the basin-of-attraction of the non-convex optimization approach.

2. We show that these two fitting algorithms are "universal" in the sense that we are able to establish their high probability error bounds for a wide range of latent space models beyond the inner-product model class. For example, they work simultaneously for the distance model and the Gaussian kernel model. Thus, the class of inner-product models as well as the proposed fitting algorithms are indeed flexible and can be used to approximate other latent space models of interest.

3. We demonstrate the effectiveness and efficiency of the model and algorithms on real data examples for several different tasks, including visualization, community detection and network-assisted learning. In particular, we obtain state-of-art performance for the task of community detection on three benchmark datasets.

1.3. Thesis Outline

The reminder of the thesis is organized as follows. In Chapter 2, we propose new error metrics which quantify the estimation loss of CCA by the excess prediction loss defined through a prediction-after-dimension-reduction framework. This framework suggests viewing CCA estimation as estimating the subspaces spanned by the canonical variates. We also characterize, with minimal assumptions, the non-asymptotic minimax rates under the proposed error metrics, especially how the minimax rates depend on the key quantities including the dimensions, the condition number of the covariance matrices and the canonical correlations. In Chapter 3, we propose a novel first-order algorithm and its stochastic variant to compute the sample CCA. This algorithm scales to large datasets. We also show the local linear convergence of the proposed algorithm.

In Chapter 4, we switch to the second part of the thesis: network data modeling. We propose a special class of latent space models, called inner-product models, which could capture typical characteristics of real-world networks. Then we propose two fitting algorithms based on convex and non-convex optimization respectively. We further establish the statistical rates of convergence of both algorithms and characterize the basin-of-attraction of the non-convex approach. Finally, simulations and real-data experiments are provided to support the proposed models and algorithms.

Chapter 2 and Chapter 3 are based on the paper Ma and Li (2016) and Ma et al. (2015) respectively. Chapter 4 is based on the unpublished manuscript Ma and Ma (2017). My research on shrinakge estimation (Ma et al. (2014a), Weinstein et al. (2015)) and reduced rank regression (Ma et al. (2014b)) is not included in this thesis.

CHAPTER 2 : Canonical Correlation Analysis: Subspace Perspective and Minimax Rates

2.1. Introduction

Canonical correlation analysis (CCA), first introduced by Hotelling (1936), is a fundamental statistical tool to characterize the relationship between two groups of random variables and finds a wide range of applications across many different fields. For example, in genome-wide association study (GWAS), CCA is used to discover the genetic associations between the genotype data of single nucleotide polymorphisms (SNPs) and the phenotype data of gene expression levels (Witten et al., 2009; Chen et al., 2012). In information retrieval, CCA is used to embed both the search space (e.g. images) and the query space (e.g. text) into a shared low dimensional latent space such that the similarity between the queries and the candidates can be quantified (Rasiwasia et al., 2010; Gong et al., 2014). In natural language processing, CCA is applied to the word co-occurrence matrix and generates vector representations of the words which capture the semantics (Dhillon et al., 2011; Faruqui and Dyer, 2014). Other applications, to name a few, include fMRI data analysis (Friman et al., 2003), computer vision (Kim et al., 2007) and speech recognition (Arora and Livescu, 2013; Wang et al., 2015).

The enormous empirical success motivates us to revisit the estimation problem of canonical correlation analysis. From a decision-theoretic point of view, two questions are naturally posed: What is the proper error metric to quantify the discrepancy between the population CCA and its sample estimates? And under such a metric, what are the quantities that characterize the fundamental statistical limits?

The justification of loss functions, in the context of CCA, has seldom appeared in the literature. From the first principle that the proper metric to quantify the estimation loss should depend on the specific purpose of using CCA, we find that the applications discussed above mainly fall into two categories: identifying variables of interest and dimension reduction. The first category, mostly in genomic research (Witten et al., 2009; Chen et al., 2012), treats one group of variables as responses and the other group of variables as covariates. The goal is to discover the specific subset of the covariates that are most correlated with the responses. Such applications are characterized by low signal-to-noise ratio and the interpretability of the results. The other category is investigated extensively in statistical machine learning and engineering community where CCA is used to learn low dimensional latent representations of complex objects such as images (Rasiwasia et al., 2010), text (Dhillon et al., 2011) and speeches (Arora and Livescu, 2013). These scenarios are usually accompanied by a relatively high signal-to-noise ratio and prediction accuracy, using the learned low dimensional embeddings as a new set of predictors, is of primary interest. In recent years, a series of publications has established fundamental theoretical guarantees for CCA to achieve sufficient dimension reduction (Kakade and Foster (2007); Foster et al. (2008); Sridharan and Kakade (2008); Fukumizu et al. (2009); Chaudhuri et al. (2009) and many others), especially under a multi-view setup as will be discussed in detail in Section 2.2.4.

In this thesis, we aim to address the problems raised above by treating CCA as a tool for dimension reduction.

2.1.1. Linear Invariance of Canonical Variates

On the population level, CCA extracts the most correlated directions between two sets of random variables: $x \in \mathbb{R}^{p_1}$ and $y \in \mathbb{R}^{p_2}$. To be specific, CCA recursively finds the pairs of vectors $\boldsymbol{\phi}_i \in \mathbb{R}^{p_1}, \boldsymbol{\psi}_i \in \mathbb{R}^{p_2}, 1 \leq i \leq p := \min\{p_1, p_2\}$ such that

$$\begin{gathered}(\boldsymbol{\phi}_i, \boldsymbol{\psi}_i) = \arg \max_{\boldsymbol{\phi}^\top \Sigma_x \boldsymbol{\phi}=1, \boldsymbol{\psi}^\top \Sigma_y \boldsymbol{\psi}=1} \boldsymbol{\phi}^\top \Sigma_{xy} \boldsymbol{\psi} \\ \text{subject to } \ \boldsymbol{\phi}^\top \Sigma_x \boldsymbol{\phi}_j = 0, \ \boldsymbol{\psi}^\top \Sigma_y \boldsymbol{\psi}_j = 0, \ \forall \ 1 \leq j \leq i-1.\end{gathered} \tag{2.1}$$

For $1 \leq i \leq p$, $(\boldsymbol{\phi}_i, \boldsymbol{\psi}_i)$ is the i^{th} pair of canonical coefficients (loading vectors), $(\boldsymbol{\phi}_i^\top x, \boldsymbol{\psi}_i^\top y)$ is the i^{th} pair of canonical variates and $\lambda_i := \boldsymbol{\phi}_i^\top \Sigma_{xy} \boldsymbol{\psi}_i$ is the i^{th} canonical correlation.

Define $\Phi := [\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_p]$, $\Psi := [\boldsymbol{\psi}_1, \cdots, \boldsymbol{\psi}_p]$ and $\Lambda := \mathrm{diag}(\lambda_1, \cdots, \lambda_p)$. Then by definition, $\Sigma_x^{1/2}\Phi, \Sigma_y^{1/2}\Psi$ have orthonormal columns and $\Lambda = \Phi^\top \Sigma_{xy} \Psi$, which further implies that $\Sigma_x^{1/2}\Phi, \Sigma_y^{1/2}\Psi$ are respectively left and right singular vectors of $\Sigma_x^{-1/2}\Sigma_{xy}\Sigma_y^{-1/2}$. With these notations, the first type of applications discussed above can be understood as identifying the support of the top-k canonical vectors: $\Phi_{1:k}$ and $\Psi_{1:k}$, where $\Phi_{1:k} \in \mathbb{R}^{p_1 \times k}$ and $\Psi_{1:k} \in \mathbb{R}^{p_2 \times k}$ consist of the first k columns of Φ and Ψ respectively. Dimension reduction, which motivates this paper, is concerned with the leading k canonical variates: $\Phi_{1:k}^\top x$ and $\Psi_{1:k}^\top y$ (k is assumed to be pre-specified).

What distinguishes CCA from other dimension reduction methods like principal component analysis or partial least squares is its linear invariance. As highlighted in Hotelling (1936) when canonical correlation analysis was first developed:

> The relations between two sets of variates with which we shall be concerned are those that remain invariant under internal linear transformations of each set separately.

Among all the population parameters, Hotelling (1936) noticed that the canonical correlations $\lambda_1, \cdots, \lambda_p$ and the functions of these quantities are the only linear invariants of the system. On the contrary, the canonical coefficients Φ and Ψ will change accordingly either with rotation of axes or scaling of the variables, which diminishes the rationale for using an error metric built directly upon the loadings. If extending Hotelling's notion of invariants to include random vectors, the canonical variates are actually invariant under linear transformations of each set separately. To illustrate, let T_1, T_2 be any pair of nonsingular matrices and define the new random vectors $a = T_1^\top x, b = T_2^\top y$. As will be shown in Section 2.3.1, $T_1^{-1}\Phi_{1:k}, T_2^{-1}\Psi_{1:k}$ are the top-k canonical coefficients of (a, b). Therefore, the top-k canonical variates of (a, b) will be $(T_1^{-1}\Phi_{1:k})^\top a = \Phi_{1:k}^\top x$ and $(T_2^{-1}\Psi_{1:k})^\top b = \Psi_{1:k}^\top y$, which are the same as those of (x, y). This fact substantiates our interest in the canonical variates instead of the loadings. Let $(\widehat{\Phi}_{1:k}, \widehat{\Psi}_{1:k})$ be any generic estimator of the loadings. Then the two questions that we aim to answer can be recast as:

What is the proper error metric to quantify the discrepancy between $(\Phi_{1:k}^\top x, \Psi_{1:k}^\top y)$ and the sample counterparts $(\widehat{\Phi}_{1:k}^\top x, \widehat{\Psi}_{1:k}^\top y)$? And under such a metric, what are the quantities that characterize the fundamental statistical limits?

For the rest of the paper, we will focus on the relationship between $\widehat{\Phi}_{1:k}^\top x$ and $\Phi_{1:k}^\top x$ since similar results can be obtained for the other pair by symmetry.

2.1.2. Subspace Estimation and Subspace Loss

In Section 2.2, we show that when CCA is used for dimension reduction, it is the difference between the predictive power of $\Phi_{1:k}^\top x$ and $\widehat{\Phi}_{1:k}^\top x$ that matters, rather than the Euclidean distance between $\Phi_{1:k}^\top x$ and $\widehat{\Phi}_{1:k}^\top x$. Specifically, we characterize the discrepancy between the predictive power by the excess prediction loss induced by replacing the population canonical variates $\Phi_{1:k}^\top x$ with the sample estimates $\widehat{\Phi}_{1:k}^\top x$. When linear prediction is concerned, such discrepancy is reduced to the difference between the linear span of the population and sample canonical variates, denoted by $\text{span}(x^\top \Phi_{1:k})$ and $\text{span}(x^\top \widehat{\Phi}_{1:k})$, which are subspaces of $\text{span}(x^\top) := \{x^\top w,\ w \in \mathbb{R}^{p_1}\}$. This suggests that CCA estimation can be viewed as subspace estimation, that is, estimating the subspace spanned by the leading-k canonical variates: $\text{span}(x^\top \Phi_{1:k})$. From this perspective, the error metric $\mathcal{L}(\cdot,\cdot)$ we pursue could be rewritten as

$$\mathcal{L}(\Phi_{1:k}^\top x, \widehat{\Phi}_{1:k}^\top x) = \mathcal{L}(\text{span}(x^\top \Phi_{1:k}), \text{span}(x^\top \widehat{\Phi}_{1:k})). \tag{2.2}$$

Interestingly, the error metrics derived through the excess prediction loss is closely related to the principal angles (defined in Section 2.2.3) between $\text{span}(x^\top \widehat{\Phi}_{1:k})$ and $\text{span}(x^\top \Phi_{1:k})$. Suppose $\theta = (\theta_1, \cdots, \theta_k)^\top$ is the vector of such principal angles. As elaborated in Theorem 1,

$$\begin{aligned} \text{Worst case excess prediction loss} &\simeq \left\| P_{\Sigma_x^{1/2}\widehat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2}\Phi_{1:k}} \right\|^2 = \|\sin(\theta)\|_\infty^2 \\ \text{Bayesian excess prediction loss} &\simeq \left\| P_{\Sigma_x^{1/2}\widehat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2}\Phi_{1:k}} \right\|_{\text{F}}^2 /2k = \|\sin(\theta)\|_2^2 /k \end{aligned} \tag{2.3}$$

where $\simeq$ means 'equal up to an absolute constant', $\sin(\theta) = (\sin(\theta_1), \cdots, \sin(\theta_k))$ and $P_{(\cdot)}$ denotes the projection matrix w.r.t. the column space of the matrix in the subscript.

2.1.3. Minimax Rates

In section 2.3, we characterize the non-asymptotic minimax estimation rates for CCA under the error metrics proposed in (2.3), especially how the minimax rates depend on the key quantities, including the dimensions, the condition number of the covariance matrices and the canonical correlations. Informally, with operator norm error as an example, in Theorem 2 and Theorem 3, we show that under certain a sample size condition ($n \geq C_{\lambda_k, \lambda_{k+1}}(p_1 + p_2)$), the minimax rate is characterized by

$$\inf_{\widehat{\Phi}_{1:k}} \sup_{\Sigma \in \mathcal{F}} \mathbb{E}\left[\left\| P_{\Sigma_x^{1/2} \widehat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2} \Phi_{1:k}} \right\|^2 \right] \asymp \frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{(\lambda_k - \lambda_{k+1})^2} \frac{p_1}{n}.$$

To the best of our knowledge, this is the first finite sample result that captures the factor $(1-\lambda_k^2)(1-\lambda_{k+1}^2)$. This term is not negligible because λ_k, λ_{k+1} are parameters depending on the dimensions and should not be treated as constants. In practice, as the number of variables increases, one should expect the canonical correlations to increase as well (e.g. considering the case that the variables are gradually added to the two groups). The other important feature is the independence of the dimension p_2. If only interested in the 'estimation' of the canonical variates of x, then even when $p_2 \gg p_1$, as long as the sample size is large enough, the minimax rate of 'estimating' $\Phi_{1:k}^\top x$ does not depend on p_2. This phenomenon was also revealed in Gao et al. (2014) and Cai and Zhang (2016) with the additional assumption that all the residual canonical correlations are zero: $\lambda_{k+1} = \cdots = \lambda_{p_1} = 0$. Finally, the minimax rates are independent of the condition number of the covariance matrices: $\kappa(\Sigma_x), \kappa(\Sigma_y)$. This is due to the linear invariance of the canonical variates as illustrated in Section 2.3.1. We hope our theoretical findings could provide some guidance for the practical use of CCA because in real applications, these factors matter both computationally and statistically (Ma et al., 2015).

The upper bound of the minimax rates is achieved by sample CCA which is defined in the same manner as (2.1) by replacing the population covariance matrices with the corresponding sample estimates. The sample canonical variates are also linear invariant, which is crucial in reducing the estimation error of sample CCA to the "standard form", as spelled out in Section 2.3.1, in which Σ_x and Σ_y are identity and $\Sigma_{xy} = [\Lambda, 0]$ $(p_1 \leq p_2)$.

Theoretical understanding for the estimation of CCA dates back to the study of the asymptotic distribution of sample CCA, in the low dimensional regime with fixed dimensions and sample size going to infinity, for both sample canonical coefficients and sample canonical correlations (Hotelling, 1936; Hsu, 1941; Izenman, 1975; Anderson, 1984, 1999) (and many others). More recently, Chen et al. (2013) and Gao et al. (2014, 2015b) have studied the non-asymptotic minimax rates of sparse CCA in a high dimensional setup. We defer the detailed comparison between these results and ours to Section 2.3.2.

2.1.4. Notations

Throughout this chapter, we use lower-case and upper-case letters to represent vectors and matrices respectively. For any matrix $U \in \mathbb{R}^{n\times p}$ and vector $u \in \mathbb{R}^p$, $\|U\|, \|U\|_{\mathrm{F}}$ denotes operator (spectral) norm and Frobenius norm respectively, $\|u\|$ denotes the vector l_2 norm, $U_{1:k}$ denotes the submatrix consisting of the first k columns of U, and P_U stands for the projection matrix onto the column space of U. Moreover, we use $\sigma_{\max}(U)$ and $\sigma_{\min}(U)$ to represent the largest and smallest singular value of U respectively, and $\kappa(U) = \sigma_{\max}(U)/\sigma_{\min}(U)$ to denote the condition number of the matrix. We use I_p for the identity matrix of dimension p and $I_{p,k}$ for the submatrix composed of the first k columns of I_p. Further, $\mathcal{O}(m,n)$ (and simply $\mathcal{O}(n)$ when $m = n$) stands for the set of $m \times n$ matrices with orthonormal columns and $\mathbb{S}_+^p$ denotes the set of $p \times p$ strictly positive definite matrices. For a random vector $x \in \mathbb{R}^p$, $\mathrm{span}(x^\top) = \{x^\top w, w \in \mathbb{R}^p\}$ denotes the subspace of all the linear combinations of x. Other notations will be specified within the corresponding context.

2.2. Subspace Perspective: Excess Prediction Loss and Subspace Angles

In this section, we propose a prediction-after-dimension-reduction framework to quantify the loss of any generic dimension reduction algorithm (including CCA). This framework suggests two error metrics, one induced by the worst case excess prediction loss and the other induced by the average excess prediction loss. For CCA, these two error metrics are closely related to the principal angles between the subspaces spanned by the population and sample canonical variates, respectively.

2.2.1. Linear Prediction Revisited

First, we review the basics of linear model theory under the random design setup. Suppose given

$$x_1, \dots, x_p, z \in L^2(\Omega, \mathcal{F}, \mathcal{P}),$$

where $L^2(\Omega, \mathcal{F}, \mathcal{P})$ is the set of random variables with mean zero and finite second moment, and the goal is to predict the response z with the random vector $x := (x_1, \dots, x_p)^\top$. We measure the prediction loss by

$$loss(z|x) := \min_{\beta \in \mathbb{R}^p} \mathbb{E}[(z - x^\top \beta)^2].$$

We further assume that (x, z) has joint covariance matrix:

$$\mathrm{Cov}\left(\begin{bmatrix} x \\ z \end{bmatrix}\right) = \begin{bmatrix} \Sigma_x & \sigma_{xz} \\ \sigma_{xz}^\top & \sigma_z^2 \end{bmatrix}.$$

By classical linear model theory

$$\beta^* := \arg\min_{\beta \in \mathbb{R}^p} \mathbb{E}[(z - x^\top \beta)^2] = \Sigma_x^{-1}\sigma_{xz},$$

$$loss(z|x) = \sigma_z^2 - \sigma_{xz}^\top \Sigma_x^{-1} \sigma_{xz} = \sigma_z^2(1 - \|r_{xz}\|^2)$$

where $r_{xz} := \Sigma_x^{-1/2}\sigma_{xz}(\sigma_z^2)^{-1/2}$ and $\|r_{xz}\|^2$ is the population R^2 which characterizes the proportion of the variability in the response z explained by the predictor x. One notable feature of such prediction loss is its linear invariance. Define the linear subspace spanned by the coordinates of x as

$$\operatorname{span}(x^\top) := \{x^\top w : w \in \mathbb{R}^p\} \subset L^2(\Omega, \mathcal{F}, \mathcal{P}).$$

If for another set of random variables $\{v_1, \ldots, v_p\} \in L^2(\Omega, \mathcal{F}, \mathcal{P})$ with $\operatorname{span}(v^\top) = \operatorname{span}(x^\top)$ and $v := (v_1, \ldots, v_p)^\top$, then by definition, $loss(z|x) = loss(z|v)$. Therefore, we can rewrite

$$loss(z|x) = loss(z|\operatorname{span}(x^\top)).$$

These two notations will be used interchangeably throughout the paper. The linear invariance property can be revealed by noticing that $\|r_{xz}\|^2 = \mathbb{E}[(P_{\operatorname{span}(x)}z)^2]/\mathbb{E}[z^2]$ where $P_{(\cdot)}$ is the projection operator defined in the Hilbert space $L^2(\Omega, \mathcal{F}, \mathcal{P})$ with covariance operator as the inner product.

2.2.2. Competing with Oracles

Consider the scenario where the predictor x is in a high dimensional space and many directions in $\operatorname{span}(x^\top)$ are redundant for predicting the response z. Practitioners usually perform certain kind of dimension reduction on x before applying supervised learning algorithms. Suppose $U \in \mathbb{R}^{p\times k}$ is a reduction matrix obtained by some generic dimension reduction method. The subspace perspective of the prediction loss discussed in the previous section suggests $loss(z|\operatorname{span}(x^\top U)) - loss(z|\operatorname{span}(x^\top))$, or simply $loss(z|\operatorname{span}(x^\top U))$ as the measure of goodness for dimension reduction algorithms.

Given any pair of reduction matrices $U_1, U_2 \in \mathbb{R}^{p\times k}$, the discrepancy between their prediction loss can be quantified by:

$$\begin{aligned} loss(z|\operatorname{span}(x^\top U_1)) - loss(z|\operatorname{span}(x^\top U_2)) &= \mathbb{E}[(P_{\operatorname{span}(x^\top U_2)}z)^2 - (P_{\operatorname{span}(x^\top U_1)}z)^2] \\ &= \sigma_z^2 \left(r_{xz}^\top \left(P_{\Sigma_x^{1/2}U_2} - P_{\Sigma_x^{1/2}U_1} \right) r_{xz} \right). \qquad (2.4) \end{aligned}$$

The first equality is geometrically straightforward, measuring the proportion of the variability in response z explained by the two subspaces: $\operatorname{span}(x^\top U_1)$ and $\operatorname{span}(x^\top U_2)$. The algebraic expression in the second equality (proved in Theorem 1) is less obvious but decouples the loss into an interaction between a supervised learning factor r_{xz} and an unsupervised learning factor $P_{\Sigma_x^{1/2}U_2} - P_{\Sigma_x^{1/2}U_1}$. To shed more light on this excess prediction risk, we parametrize the joint covariance matrix of (x, z) in terms of separate covariance matrices Σ_x, σ_z^2 and the vector r_{xz}, that is

$$\operatorname{Cov}\left(\begin{bmatrix} x \\ z \end{bmatrix} \right) = \begin{bmatrix} \Sigma_x & \sigma_{xz} \\ \sigma_{xz}^\top & \sigma_z^2 \end{bmatrix} = \begin{bmatrix} \Sigma_x^{1/2} & 0 \\ 0 & \sigma_z \end{bmatrix} \begin{bmatrix} I_p & r_{xz} \\ r_{xz}^\top & 1 \end{bmatrix} \begin{bmatrix} \Sigma_x^{1/2} & 0 \\ 0 & \sigma_z \end{bmatrix}. \qquad (2.5)$$

Considering the worst case discrepancy across all possible correlation structures, as proved in Theorem 1,

$$\sup_{r_{xz}:\|r_{xz}\|^2=R^2} \left\{ loss(z|\operatorname{span}(x^\top U_1)) - loss(z|\operatorname{span}(x^\top U_2)) \right\} = \sigma_z^2 R^2 \left\| P_{\Sigma_x^{1/2}U_1} - P_{\Sigma_x^{1/2}U_2} \right\|,$$

which suggests the right hand side of the equation as a sensible metric to quantify the difference between the two reduction matrices. Actually, it is more informative to replace the competitor U_2 with an oracle reduction matrix, denoted by $U_\star$. As suggested by (2.4), we say that a reduction matrix $U_\star$ is an oracle reduction matrix if $P_{\Sigma_x^{1/2}U_\star} r_{xz} = r_{xz}$. Define $\mathcal{A} := \{r : P_{\Sigma_x^{1/2}U_\star} r = r, \|r\|^2 = R^2\}$ as the set of 'correlation' vectors such that $U_\star$ is an oracle reduction matrix with fixed population R^2. If considering the worst case excess prediction loss within $\mathcal{A}$, then according to Theorem 1,

$$\sup_{r_{xz}\in\mathcal{A}} \left\{loss(z|\operatorname{span}(x^\top U)) - loss(z|\operatorname{span}(x^\top U_\star))\right\} = \sigma_z^2 R^2 \left\| P_{\Sigma_x^{1/2}U} - P_{\Sigma_x^{1/2}U_\star}\right\|^2. \quad (2.6)$$

Interestingly, the operator norm is replaced by its square when the competitor is an oracle reduction matrix. On the other hand, from a Bayesian perspective, considering the prior that the vector r_{xz} is sampled with respect to the uniform measure (Haar measure) on $\mathcal{A}$, denoted by π, then the average excess prediction loss will satisfy (also refer to Theorem 1):

$$\mathbb{E}_{r_{xz}\sim\pi} \left\{loss(z|\operatorname{span}(x^\top U)) - loss(z|\operatorname{span}(x^\top U_\star))\right\} = \frac{\sigma_z^2 R^2}{2k} \left\| P_{\Sigma_x^{1/2}U} - P_{\Sigma_x^{1/2}U_\star}\right\|_{\mathrm{F}}^2. \quad (2.7)$$

The analysis above connects the prediction loss for the generic response z with the estimation loss for the oracle reduction matrix $U_\star$ under the metrics derived in (2.6) and (2.7). Therefore, when CCA is used for dimension reduction, it is natural to quantify the discrepancy between $\widehat{\Phi}_{1:k}^\top x$ and $\Phi_{1:k}^\top x$ by the excess prediction loss:

$$\left\| P_{\Sigma_x^{1/2}\widehat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2}\Phi_{1:k}}\right\|^2, \quad \left\| P_{\Sigma_x^{1/2}\widehat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2}\Phi_{1:k}}\right\|_{\mathrm{F}}^2.$$

2.2.3. Measuring Subspace Distance by Principal Angles

In this section, we show that the loss defined in (2.6) and (2.7) are closely related to the principal angles between the two subspaces spanned by the reduced predictors. For any p dimensional random vector x with mean zero and bounded second moments, define the Hilbert space

$$\mathcal{H} = \operatorname{span}(x^\top) = \{X | X = x^\top w, w \in \mathbb{R}^p\}$$

with covariance operator as the inner product, that is, for any $X_1, X_2 \in \mathcal{H}$, $\langle X_1, X_2\rangle := \operatorname{Cov}(X_1, X_2) = \mathbb{E}(X_1 X_2)$. Suppose we have a pair of full column rank matrices $U_1, U_2 \in \mathbb{R}^{p\times k}$ and consider the canonical correlation analysis between the two subspaces of $\mathcal{H}$: $\operatorname{span}(x^\top U_1)$ and $\operatorname{span}(x^\top U_2)$. Let $(W_1, \widehat{W_1}), (W_2, \widehat{W_2}), \ldots, (W_k, \widehat{W_k})$ be the first, second, ..., and k^{th} pair of canonical variates between $\text{span}(x^\top U_1)$ and $\text{span}(x^\top U_2)$. Then $\text{span}(W_1, \ldots, W_k) = \text{span}(x^\top U_1)$, $\text{span}(\widehat{W}_1, \ldots, \widehat{W}_k) = \text{span}(x^\top U_2)$ and $\langle W_i, W_j \rangle = \langle W_i, \widehat{W}_j \rangle = \langle \widehat{W}_i, \widehat{W}_j \rangle = 0$, for any $i \neq j$ and $\text{Var}(W_i) = \text{Var}(\widehat{W}_i) = 1$, for $i = 1, \ldots, k$. The i^{th} principal angle is defined as $\theta_i := \angle(W_i, \widehat{W}_i)$. Without loss of generality we assume $\theta_1 \geq \cdots \geq \theta_k$ and define the distance between the two subspaces as:

$$\mathcal{L}_2(\text{span}(x^\top U_1), \text{span}(x^\top U_2)) := \sum_{i=1}^k \sin^2 \theta_i = \sum_{i=1}^k \left(1 - \left|\left\langle W_i, \widehat{W}_i \right\rangle\right|^2\right).$$

This is a valid metric because the principal angles are uniquely defined though the canonical variates need not be. Since $x^\top \Sigma_x^{-1/2}$ is an orthonormal basis of $\mathcal{H}$ under the covariance inner product, it is convenient to represent the elements in $\mathcal{H}$ by this basis. Let

$$(W_1, \ldots, W_k) = x^\top \Sigma_x^{-1/2} B, \text{ and } (\widehat{W}_1, \ldots, \widehat{W}_k) = x^\top \Sigma_x^{-1/2} \widehat{B},$$

where $B := [b_1, \ldots, b_k]$, $\widehat{B} := [\widehat{b}_1, \ldots, \widehat{b}_k] \in \mathbb{R}^{p \times k}$ are the coordinate representations under $x^\top \Sigma_x^{-1/2}$. Notice that by definition, $\{W_1, \ldots, W_k\}$ and $\{\widehat{W}_1, \ldots, \widehat{W}_k\}$ are orthonormal bases of $\text{span}(x^\top U_1)$ and $\text{span}(x^\top U_2)$, respectively. Then $B, \widehat{B}$ are $p \times k$ basis matrices. Moreover, we have $b_i^\top \widehat{b}_j = \langle W_i, \widehat{W}_j \rangle = 0$, for all $i \neq j$.

Let's now represent $\mathcal{L}_2(\text{span}(x^\top U_1), \text{span}(x^\top U_2))$ in terms of B and $\widehat{B}$. In fact, since

$$1 - \left|\left\langle W_i, \widehat{W}_i \right\rangle\right|^2 = 1 - \left|b_i^\top \widehat{b}_i\right|^2 = \frac{1}{2}\left\|b_i b_i^\top - \widehat{b}_i \widehat{b}_i^\top\right\|_{\text{F}}^2,$$

we have

$$\begin{aligned}
\mathcal{L}_2(\text{span}(x^\top U_1), \text{span}(x^\top U_2)) &= \frac{1}{2} \sum_{i=1}^k \left\|b_i b_i^\top - \widehat{b}_i \widehat{b}_i^\top\right\|_{\text{F}}^2 \\
&= \frac{1}{2}\left\|\sum_{i=1}^k \left(b_i b_i^\top - \widehat{b}_i \widehat{b}_i^\top\right)\right\|_{\text{F}}^2 \\
&= \frac{1}{2}\left\|BB^\top - \widehat{B}\widehat{B}^\top\right\|_{\text{F}}^2,
\end{aligned}$$

where the second equality is due to $b_i^\top b_j = \widehat{b}_i^\top b_j = \widehat{b}_i^\top \widehat{b}_j = 0$, for all $i \neq j$.

Finally, notice that $\operatorname{span}(x^\top U_1) = \operatorname{span}(W_1, \ldots, W_k)$, $x^\top U_1 = (x^\top \Sigma_x^{-1/2})(\Sigma_x^{1/2} U_1)$, and $(W_1, \ldots, W_k) = x^\top \Sigma_x^{-1/2} B$. Then B and $\Sigma_x^{1/2} U_1$ have the same column space. Since $B \in \mathbb{R}^{p \times k}$ is a basis matrix, we have $BB^\top = P_{\Sigma_x^{1/2} U_1}$, which is the orthogonal projector to the column space of $\Sigma_x^{1/2} U_1$. Similarly, we have $\widehat{B}\widehat{B}^\top = P_{\Sigma_x^{1/2} U_2}$, which implies

$$\mathcal{L}_2(\operatorname{span}(x^\top U_1), \operatorname{span}(x^\top U_2)) = \frac{1}{2} \left\| P_{\Sigma_x^{1/2} U_1} - P_{\Sigma_x^{1/2} U_2} \right\|_{\mathrm{F}}^2 .$$

On the other hand, we can also define the distance through the largest principal angle:

$$\mathcal{L}_1(\operatorname{span}(x^\top U_1), \operatorname{span}(x^\top U_2)) := \sin^2 \theta_1 = 1 - \left| \left\langle W_1, \widehat{W_1} \right\rangle \right|^2 .$$

Let $\theta = (\theta_1, \cdots, \theta_k)$, then $BB^\top \widehat{B}\widehat{B}^\top = B \operatorname{diag}(\cos(\theta)) \widehat{B}^\top$, and by Lemma 2.18,

$$\begin{aligned} \left\| P_{\Sigma_x^{1/2} U_1} - P_{\Sigma_x^{1/2} U_2} \right\|^2 &= \left\| BB^\top - \widehat{B}\widehat{B}^\top \right\|^2 = 1 - \sigma_{\min}^2 \left(BB^\top \widehat{B}\widehat{B}^\top \right) \\ &= \sin^2(\theta_1) = \mathcal{L}_1(\operatorname{span}(x^\top U_1), \operatorname{span}(x^\top U_2)) \end{aligned}$$

We summarize the results into the following theorem of which the proof is deferred to Section 2.4.

Theorem 1. *Suppose $(x, z) \sim \mathbb{P}$ for some unknown distribution $\mathbb{P}$ with covariance structure specified in* (2.5) *and subspace angles defined above. For any pair reduction matrices $U, U_\star \in \mathbb{R}^{p \times k}$,*

$$\begin{aligned} \mathit{loss}(z | \operatorname{span}(x^\top U)) - \mathit{loss}(z | \operatorname{span}(x^\top U_\star)) &= \sigma_z^2 \left(r_{xz}^\top \left(P_{\Sigma_x^{1/2} U_\star} - P_{\Sigma_x^{1/2} U} \right) r_{xz} \right) \\ \sup_{r_{xz}\colon \|r_{xz}\|^2 = R^2} \mathit{loss}(z | \operatorname{span}(x^\top U)) &- \mathit{loss}(z | \operatorname{span}(x^\top U_\star)) \\ &= \sigma_z^2 R^2 \left\| P_{\Sigma_x^{1/2} U} - P_{\Sigma_x^{1/2} U_\star} \right\| = \sigma_z^2 R^2 \left\| \sin(\theta) \right\|_\infty . \end{aligned}$$

Let $\mathcal{A} = \{r : P_{\Sigma_x^{1/2} U_\star} r = r, \|r\|^2 = R^2\}$. By treating $U_\star$ as an oracle reduction matrix,

$$\sup_{r_{xz}:\ r_{xz} \in \mathcal{A}} \left\{ loss(z | \operatorname{span}(x^\top U)) - loss(z | \operatorname{span}(x^\top U_\star)) \right\}$$
$$= \sigma_z^2 R^2 \left\| P_{\Sigma_x^{1/2} U} - P_{\Sigma_x^{1/2} U_\star} \right\|^2 = \sigma_z^2 R^2 \left\| \sin(\theta) \right\|_\infty^2 .$$

By treating $U_\star$ as a Bayes oracle, that is $r_{xz} \sim \pi$ where π is the uniform measure (Haar measure) on $\mathcal{A}$, then

$$\mathbb{E}_{r_{xz} \sim \pi} \left\{ loss(z | \operatorname{span}(x^\top U)) - loss(z | \operatorname{span}(x^\top U_\star)) \right\}$$
$$= \frac{\sigma_z^2 R^2}{2k} \left\| P_{\Sigma_x^{1/2} U} - P_{\Sigma_x^{1/2} U_\star} \right\|_{\mathrm{F}}^2 = \frac{\sigma_z^2 R^2}{k} \left\| \sin(\theta) \right\|_2^2$$

2.2.4. CCA for Multi-view Dimension Reduction

In the research and applications of multi-media analytics, data of the same object, is often collected from multiple sources and exhibit heterogeneous properties. Features obtained from different domains are referred to as different ‘views’. Usually each view summarizes a specific aspect of the studied object and different views are complementary to one another. For example, in web-page classification, the hyperlink structure and the words on the page are two different views (Chaudhuri et al., 2009). In video surveillance, images of cameras from different angles constitute different views (Loy et al., 2009). For more recent results, see the survey paper of Xu et al. (2013) and references therein.

Although multiple views provide more potential discriminative information to distinguish the patterns of different classes, the feature vector of each view usually lies in a high dimensional space. It is critical, both statistically and computationally, to perform dimension reduction before applying any supervised learning algorithm. It has been shown by many researchers that canonical correlation analysis can achieve sufficient dimension reduction under certain multi-view assumptions (Kakade and Foster (2007); Foster et al.

(2008); Sridharan and Kakade (2008); Fukumizu et al. (2009); Chaudhuri et al. (2009) and many others)

Suppose the input variable x can be split into two views $x^{(1)}, x^{(2)}$ and the goal is to predict the response z based on the two views. Let $(\Phi_{1:k}, \Psi_{1:k})$ be the top-k population canonical coefficients between $x^{(1)}$ and $x^{(2)}$. Foster et al. (2008) proved the following proposition.

Proposition 2.1. *(Sufficient Dimension Reduction by CCA Foster et al. (2008)) Under certain multi-view assumptions*[1],

$$
\begin{aligned}
loss(z|x^{(1)}) &= loss(z|\operatorname{span}((x^{(1)})^\top \Phi_{1:k})) \\
loss(z|x^{(2)}) &= loss(z|\operatorname{span}((x^{(2)})^\top \Psi_{1:k}))
\end{aligned}
$$

This proposition shows that the predictive power of the original high-dimensional predictors $x^{(1)}$ and $x^{(2)}$ is fully captured by the top k canonical variates. However, the proposition focuses on the population level and does not take into account the estimation error induced by substituting the population canonical coefficients with the sample estimates. Such sample-population discrepancy can be quantified by

$$
\begin{aligned}
&loss(z|\operatorname{span}((x^{(1)})^\top \widehat{\Phi}_{1:k})) - loss(z|\operatorname{span}((x^{(1)})^\top \Phi_{1:k})), \\
&loss(z|\operatorname{span}((x^{(2)})^\top \widehat{\Psi}_{1:k})) - loss(z|\operatorname{span}((x^{(2)})^\top \Psi_{1:k})),
\end{aligned}
$$

or equivalently, the proposed loss functions according to Theorem 1.

2.3. Minimax Upper and Lower Bounds

In this section, we introduce our main results on non-asymptotic upper and lower bounds for estimating CCA under the proposed loss functions. Specifically, the upper bound is achieved by sample CCA.

[1] See Theorem 3 of Foster et al. (2008) for details

2.3.1. Reduction for Sample CCA

In this section, we show that the linear invariance of both population and sample canonical variates enables us to reduce the estimation error of sample CCA to the special case that $\Sigma_x = I_{p_1}, \Sigma_y = I_{p_2}$ and $\Sigma_{xy} = [\Lambda \ 0]$ (we assume $p_1 \leq p_2$ without loss of generality).

Let $X = (x_1, \cdots, x_n)^\top$ be the data matrix where $x_1, \cdots, x_n \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma_x)$ and similarly we define Y. It is well known that the top-k sample canonical coefficients can be defined as a solution to the following optimization problem:

$$\begin{aligned}(\widehat{\Phi}_{1:k}, \widehat{\Psi}_{1:k}) \in \arg\max_{W_x, W_y} \operatorname{tr}(W_x^\top \widehat{\Sigma}_{xy} W_y) \\ \text{subject to} \quad W_x^\top \widehat{\Sigma}_x W_x = I_k, \quad W_y^\top \widehat{\Sigma}_y W_y = I_k.\end{aligned} \tag{2.8}$$

where $\widehat{\Sigma}_x, \widehat{\Sigma}_y, \widehat{\Sigma}_{xy}$ are sample variance and covariance matrices defined as

$$\widehat{\Sigma}_{xy} = \frac{1}{n} X^\top Y, \quad \widehat{\Sigma}_x = \frac{1}{n} X^\top X, \quad \text{and } \widehat{\Sigma}_y = \frac{1}{n} Y^\top Y.$$

Coming back to the definition of population CCA in (2.1), Ψ is a $p_2 \times p_1$ matrix such that $\Sigma_y^{1/2}\Psi \in \mathcal{O}(p_2, p_1)$. In this section, we abuse notation and redefine Ψ as the $p_2 \times p_2$ matrix by arbitrarily padding the rest $p_2 - p_1$ columns such that $\Sigma_y^{1/2}\Psi \in \mathcal{O}(p_2)$. Let $a_i = \Phi^\top x_i$ and $b_i = \Psi^\top y_i$. Then we will have $a_i \overset{i.i.d}{\sim} a = \Phi^\top x$ with distribution $\mathcal{N}(0, I_{p_1})$ and $b_i \overset{i.i.d}{\sim} b = \Psi^\top y$ with distribution $\mathcal{N}(0, I_{p_2})$. Moreover,

$$\Sigma_{ab} := \mathbb{E} a_i b_i^\top = \Phi^\top \Sigma_{xy} \Psi = (\Sigma_x^{1/2}\Phi)^\top \Sigma_x^{-1/2} \Sigma_{xy} \Sigma_y^{-1/2} (\Sigma_y^{\frac{1}{2}} \Psi) = [\Lambda \ 0].$$

where the last equality is due to the fact that $\Sigma_x^{1/2}\Phi, \Sigma_y^{1/2}\Psi$ are respectively left and right singular vectors of $\Sigma_x^{-1/2}\Sigma_{xy}\Sigma_y^{-1/2}$. This implies

$$\begin{bmatrix} a \\ b \end{bmatrix} \overset{i.i.d}{\sim} \mathcal{N}\left(0, \begin{bmatrix} I_{p_1} & \Sigma_{ab} \\ \Sigma_{ba} & I_{p_2} \end{bmatrix}\right),$$

and their top-k population canonical coefficients $(\Phi^a_{1:k}, \Psi^b_{1:k}) = (I_{p_1,k}, I_{p_2,k})$. Let $A = (a_1, \cdots, a_n)^\top = X\Phi$ and $B = (b_1, \cdots, b_n)^\top = Y\Psi$. Then

$$\begin{aligned}\widehat{\Sigma}_a &= \frac{1}{n}A^\top A = \Phi^\top \widehat{\Sigma}_x \Phi,\\ \widehat{\Sigma}_b &= \frac{1}{n}B^\top B = \Psi^\top \widehat{\Sigma}_y \Psi,\\ \widehat{\Sigma}_{ab} &= \frac{1}{n}A^\top B = \Phi^\top \widehat{\Sigma}_{xy} \Psi.\end{aligned} \tag{2.9}$$

Since $(\widehat{\Phi}_{1:k}, \widehat{\Psi}_{1:k})$ is a solution to the sample CCA (2.8), then

$$\begin{aligned}(\Phi^{-1}\widehat{\Phi}_{1:k}, \Psi^{-1}\widehat{\Psi}_{1:k}) &\in \arg\max_{W_x, W_y} \operatorname{tr}(W_x^\top \Phi^\top \widehat{\Sigma}_{xy} \Psi W_y)\\ \text{subject to} \quad & W_x^\top \Phi^\top \widehat{\Sigma}_x \Phi W_x = I_k, \quad W_y^\top \Psi^\top \widehat{\Sigma}_y \Psi W_y = I_k,\end{aligned}$$

or, by (2.9), equivalently,

$$\begin{aligned}(\Phi^{-1}\widehat{\Phi}_{1:k}, \Psi^{-1}\widehat{\Psi}_{1:k}) &\in \arg\max_{W_x, W_y} \operatorname{tr}(W_x^\top \widehat{\Sigma}_{ab} W_y)\\ \text{subject to} \quad & W_x^\top \widehat{\Sigma}_a W_x = I_k, \quad W_y^\top \widehat{\Sigma}_b W_y = I_k.\end{aligned}$$

Therefore, $(\Phi^{-1}\widehat{\Phi}_{1:k}, \Psi^{-1}\widehat{\Psi}_{1:k})$ are the sample canonical coefficients for (a, b), which we denote by $(\widehat{\Phi}^a_{1:k}, \widehat{\Psi}^b_{1:k})$. Then $a^\top \widehat{\Phi}^a_{1:k} = x^\top \widehat{\Phi}_{1:k}$ and $a^\top \Phi^a_{1:k} = x^\top \Phi_{1:k}$ (linear invariance of canonical variates). Hence,

$$\begin{aligned}\left\| P_{\Sigma_x^{1/2}\Phi_{1:k}} - P_{\Sigma_x^{1/2}\widehat{\Phi}_{1:k}} \right\|^2 &= \mathcal{L}_1(\operatorname{span}(x^\top \widehat{\Phi}_{1:k}), \operatorname{span}(x^\top \Phi_{1:k}))\\ &= \mathcal{L}_1(\operatorname{span}(a^\top \widehat{\Phi}^a_{1:k}), \operatorname{span}(a^\top \Phi^a_{1:k}))\\ &= \left\| P_{\Phi^a_{1:k}} - P_{\widehat{\Phi}^a_{1:k}} \right\|^2.\end{aligned}$$

By the same argument, with $\mathcal{L}_1$ replaced by $\mathcal{L}_2$,

$$\left\| P_{\Sigma_x^{1/2}\Phi_{1:k}} - P_{\Sigma_x^{1/2}\widehat{\Phi}_{1:k}} \right\|_{\mathrm{F}}^2 = \left\| P_{\Phi^a_{1:k}} - P_{\widehat{\Phi}^a_{1:k}} \right\|_{\mathrm{F}}^2$$

To sum up, it suffices to consider the special covariance structure $\Sigma_x = I_{p_1}, \Sigma_y = I_{p_2}, \Sigma_{xy} =$

$[\Lambda\ 0]$ to analyze the estimation error of sample CCA.

Remark 2.2. As a byproduct, the reduction argument reveals that the estimation error of sample CCA is independent of the condition numbers of the covariance matrices: $\kappa(\Sigma_x)$ and $\kappa(\Sigma_y)$. This is not obvious because the separate estimation errors, $\|\Sigma_x - \widehat{\Sigma}_x\|, \|\Sigma_y - \widehat{\Sigma}_y\|, \|\Sigma_{xy} - \widehat{\Sigma}_{xy}\|$, are in fact proportional to $\kappa(\Sigma_x)$ and $\kappa(\Sigma_y)$.

2.3.2. Upper and Lower Bounds

In this section, we assume $x \in \mathbb{R}^{p_1}, y \in \mathbb{R}^{p_2}$ are jointly normal with mean zero and joint covariance matrix Σ specified by

$$\Sigma = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^\top & \Sigma_y \end{bmatrix}$$

where Σ_x and Σ_y are nonsingular. Recall that $\lambda_1, \cdots, \lambda_{p_1 \wedge p_2}$ are the canonical correlations and (Φ, Ψ) are the canonical coefficient matrices (loadings) as defined in (2.1). For any $1 \leq k < p_1 \wedge p_2$, define the k_{th} eigen-gap as $\Delta = \lambda_k - \lambda_{k+1}$.

Theorem 2. *(Upper bound) There exists universal positive constants C, C_1, C_2 independent of n, p_1, p_2 and Σ such that if $n \geq C(p_1 + p_2)$, the top-k sample canonical coefficients matrix $\widehat{\Phi}_{1:k}$ satisfies*

$$\mathbb{E}\left[\left\|P_{\Sigma_x^{1/2}\widehat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2}\Phi_{1:k}}\right\|^2\right] \leq C_1 \frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{\Delta^2} \frac{p_1}{n} + C_2 \left(\frac{p_1 + p_2}{n\Delta^2}\right)^2$$

$$\mathbb{E}\left[\left\|P_{\Sigma_x^{1/2}\widehat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2}\Phi_{1:k}}\right\|_{\mathrm{F}}^2 / k\right] \leq C_1 \frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{\Delta^2} \frac{p_1 - k}{n} + C_2 \left(\frac{p_1 + p_2}{n\Delta^2}\right)^2$$

The upper bounds for $\widehat{\Psi}_{1:k}$ can be obtained by switching p_1 and p_2.

This theorem exhibits several notable features:

1. The multiplicative factor $(1-\lambda_k^2)(1-\lambda_{k+1}^2)/\Delta^2$ appears in the principal term. The inverse dependence on the eigen-gap Δ^2 is inherent for spectral estimations. The factor $(1-\lambda_k^2)$ reveals that the estimation error decreases with increasing correlations

between the two sets of random variables. When there is perfect correlation, that is $\lambda_k = 1$, we can recover the CCA directions errorlessly because the observed data along those directions are perfectly co-linear. The factor $(1 - \lambda_{k+1}^2)$ comes as surprise and appears in our lower bound as well. When λ_k is close to 1 and the eigen-gap Δ is close to 0, for example, in the regime that $\lambda_k = 1 - C\Delta$, as $\Delta \to 0$,

$$(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)/\Delta^2 \asymp \text{ constant}$$

which indicates that consistency might still be achieved. In contrast, without the factor $(1 - \lambda_{k+1}^2)$, the principal term will explode since $(1 - \lambda_k^2)/\Delta^2 \asymp 1/\Delta$. We remark that λ_k, λ_{k+1} are parameters depending on the dimensions and should not be treated as constants. As the number of variables increases, one should expect the canonical correlations to increase as well (e.g. considering the case that the variables are gradually added to the two groups).

To the best of our knowledge, this is the first finite sample result to capture the factors: $(1 - \lambda_k^2)$ and $(1 - \lambda_{k+1}^2)$. This is achieved by a careful Taylor expansion of the estimating equations for $\widehat{\Phi}_{1:k}$ and $\widehat{\Psi}_{1:k}$, inspired by the classical multivariate theory of Anderson (1963, 1984, 1999) and Birnbaum et al. (2013), while the analysis of Gao et al. (2014, 2015b) and Cai and Zhang (2016) does not yield this factor.

2. The dimension parameter p_2 only appears in the high order term, which implies that even when $p_2 \gg p_1$, as long as the sample size is large enough (see Corollary 2.4), the 'estimation' error of $\Phi_{1:k}^\top x$ will not depend on p_2. This phenomenon was first revealed by Gao et al. (2014) through multi-stage estimation and sample splitting. The recent work of Cai and Zhang (2016) directly proved such a result for sample CCA without splitting the samples. The results of both Gao et al. (2014) and Cai and Zhang (2016) are based on the artificial assumption that all the residual canonical correlations are zero: $\lambda_{k+1} = \cdots = \lambda_{p_1} = 0$ (or equivalently, the rank of Σ_{xy} is k).

3. The upper bound does not depend on the condition number of the covariance matrices: $\kappa(\Sigma_x)$ and $\kappa(\Sigma_y)$. It is directly implied by the reduction argument but not obvious because the separate estimation errors, $\|\Sigma_x - \widehat{\Sigma}_x\|, \|\Sigma_y - \widehat{\Sigma}_y\|, \|\Sigma_{xy} - \widehat{\Sigma}_{xy}\|$, are in fact proportional to these condition numbers. The success of the reduction argument relies on the linear invariance of both population and sample canonical variates. For loss functions directly based on the loadings (Chen et al., 2013; Gao et al., 2015b), the upper bounds will be proportional to $\kappa(\Sigma_x)$ and $\kappa(\Sigma_y)$.

4. The only assumption made in Theorem 2 is that Σ_x, Σ_y are invertible. Anderson (1999) assumes the canonical correlations are distinct because the argument requires the asymptotic convergence of each individual canonical vector and coefficient. Moreover, the result is asymptotic without finite sample guarantee. Gao et al. (2014) and Cai and Zhang (2016) assume $\lambda_{k+1} = \cdots = \lambda_{p_1} = 0$, and Gao et al. (2014, 2015b) assume the condition number $\kappa(\Sigma_x)$ and $\kappa(\Sigma_y)$ are bounded.

To establish the minimax lower bound, we define the parameter space

$$\mathcal{F}(p_1, p_2, k, \lambda_k, \lambda_{k+1}, \kappa_1, \kappa_2)$$

as the collection of joint covariance matrices Σ satisfying

$$\Sigma_x, \Sigma_y \text{ are nonsingular, } \kappa(\Sigma_x) = \kappa_1, \kappa(\Sigma_y) = \kappa_2$$
$$0 \leq \lambda_{p_1 \wedge p_2} \leq \cdots \leq \lambda_{k+1} < \lambda_k \leq \cdots \leq \lambda_1 \leq 1$$

We deliberately set $\kappa(\Sigma_x) = \kappa_1, \kappa(\Sigma_y) = \kappa_2$ to demonstrate that the lower bound is independent of the condition number. For the rest of the paper, we will use the shorthand $\mathcal{F}$ to represent this parameter space for simplicity.

Theorem 3. *(Lower bound) There exists a universal constant c independent of n, p_1, p_2*

and Σ *such that*

$$\inf_{\widehat{\Phi}_{1:k}} \sup_{\Sigma\in\mathcal{F}} \mathbb{E}\left[\left\|P_{\Sigma_x^{1/2}\widehat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2}\Phi_{1:k}}\right\|^2\right] \geq c^2 \left\{\left(\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{\Delta^2}\frac{p_1-k}{n}\right) \wedge 1 \wedge \frac{p_1-k}{k}\right\}$$

$$\inf_{\widehat{\Phi}_{1:k}} \sup_{\Sigma\in\mathcal{F}} \mathbb{E}\left[\left\|P_{\Sigma_x^{1/2}\widehat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2}\Phi_{1:k}}\right\|_{\mathrm{F}}^2/k\right] \geq c^2 \left\{\left(\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{\Delta^2}\frac{p_1-k}{n}\right) \wedge 1 \wedge \frac{p_1-k}{k}\right\}$$

where $\Delta = \lambda_k - \lambda_{k+1}$. *The lower bounds for* $\widehat{\Psi}_{1:k}$ *can be obtained by switching* p_1 *and* p_2.

The proof of the lower bound is deferred to Section 2.6.

Remark 2.3. The upper and lower bounds together imply that the condition numbers of Σ_x and Σ_y are neither cursing nor blessing when subspace estimation is concerned.

Gao et al. (2015b) obtained minimax lower bounds for sparse CCA in high dimensional regime. Rephrasing their results without sparsity assumptions, they essentially proved

$$\mathbb{E}\left\{\inf_{Q\in\mathcal{O}(p_1)} \mathbb{E}\left[\left\|x^\top\Phi_{1:k} - x^\top\widehat{\Phi}_{1:k}Q\right\|_{\mathrm{F}}^2/k\right]\right\} \geq c^2 \left\{\left(\frac{1-\lambda_k^2}{\kappa_1\lambda_k^2}\frac{p_1-k}{n}\right) \wedge 1 \wedge \frac{p_1-k}{k}\right\},$$

with the parameter space $\lambda_{p_1\wedge p_2} = \cdots = \lambda_{k+1} = 0 < \lambda_k \leq \cdots \leq \lambda_1 \leq 1$. The inner expectation is with respect to an independent sample x and the outer expectation is with respect to the data from which $\widehat{\Phi}_{1:k}$ is constructed. The inverse dependency on the condition number in their results can be removed by noticing that

$$\inf_{Q\in\mathcal{O}(p_1)} \mathbb{E}\left[\left\|x^\top\Phi_{1:k} - x^\top\widehat{\Phi}_{1:k}Q\right\|_{\mathrm{F}}^2\right] \geq \frac{1}{2}\left\|P_{\Sigma_x^{1/2}\widehat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2}\Phi_{1:k}}\right\|_{\mathrm{F}}^2.$$

(see Section 2.7.9 for the proof) and applying Theorem 3 with $\lambda_{k+1} = 0$.

Corollary 2.4. *When* $p_1 \geq 2k$ *and*

$$\frac{p_1+p_2}{n\Delta^2} \leq c\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{(1+p_2/p_1)}, \tag{2.10}$$

for some universal positive constant c, the minimax rates can be characterized by

$$\inf_{\widehat{\Phi}_{1:k}} \sup_{\Sigma\in\mathcal{F}} \mathbb{E}\left[\left\|P_{\Sigma_x^{1/2}\widehat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2}\Phi_{1:k}}\right\|^2\right] \asymp \frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{\Delta^2}\frac{p_1}{n},$$
$$\inf_{\widehat{\Phi}_{1:k}} \sup_{\Sigma\in\mathcal{F}} \mathbb{E}\left[\left\|P_{\Sigma_x^{1/2}\widehat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2}\Phi_{1:k}}\right\|_{\mathrm{F}}^2/k\right] \asymp \frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{\Delta^2}\frac{p_1}{n}.$$

where $\Delta = \lambda_k - \lambda_{k+1}$.

Remark 2.5. For consistency, we only need the left hand side of (2.10) to converge to zero. However in order for the high order term to be dominated by the principal term, the left hand side of (2.10) is required to converge to zero faster than the right hand side.

2.4. Proof of Theorem 1

For any $U_1, U_2 \in \mathbb{R}^{p\times k}$, from classical linear model theory,

$$\boldsymbol{\beta}_i := \arg\min_{\boldsymbol{\beta}} \mathbb{E}\left[\left|z - \boldsymbol{\beta}^\top(U_i^\top x_i)\right|^2\right] = (U_i^\top\Sigma_x U_i)^{-1}U_i^\top\sigma_{xz}, \quad i=1,2,$$

and

$$\begin{aligned} loss(z|\operatorname{span}(x^\top U_i)) &= \mathbb{E}[z^2] - \mathbb{E}[(\boldsymbol{\beta}_i^\top(U_i^\top x_i))^2] \\ &= \sigma_z^2 - \sigma_{xz}^\top U_i(U_i^\top\Sigma_x U_i)^{-1}U_i^\top\Sigma_x U_i(U_i^\top\Sigma_x U_i)^{-1}U_i^\top\sigma_{xz} \\ &= \sigma_z^2\left(1 - r_{xz}^\top(\Sigma_x^{1/2}U_i)(U_i^\top\Sigma_x U_i)^{-1}(\Sigma_x^{1/2}U_i)^\top r_{xz}\right). \end{aligned}$$

Notice that $(\Sigma_x^{1/2}U_i)(U_i^\top\Sigma_x U_i)^{-1/2}$ has orthonormal columns with column space $\operatorname{span}(\Sigma_x^{1/2}U_i)$, then

$$loss(z|\operatorname{span}(x^\top U_i)) = \sigma_z^2\left(1 - r_{xz}^\top P_{\Sigma_x^{1/2}U_i} r_{xz}\right).$$

Therefore,

$$loss(z|\operatorname{span}(x^\top U_1)) - loss(z|\operatorname{span}(x^\top U_2)) = \sigma_z^2 r_{xz}^\top\left(P_{\Sigma_x^{1/2}U_2} - P_{\Sigma_x^{1/2}U_1}\right)r_{xz}.$$

By the variational definition of the leading eigenvalue,

$$\begin{aligned}\sup_{r_{xz}:\ \|r_{xz}\|^2=R^2} loss(z|\operatorname{span}(x^\top U_1)) - loss(z|\operatorname{span}(x^\top U_2)) &= \sigma_z^2 R^2 \lambda_{\max}\left(P_{\Sigma_x^{1/2}U_2} - P_{\Sigma_x^{1/2}U_1}\right) \\ &= \sigma_z^2 R^2 \left\| P_{\Sigma_x^{1/2}U_2} - P_{\Sigma_x^{1/2}U_1} \right\|,\end{aligned}$$

where the second equality is by the characterization of the difference between two projection matrices due to Wedin (1983). This proves the first claim of the theorem. For the second part, since $r_{xz} \in \mathcal{A}$,

$$\begin{aligned}& loss(z|\operatorname{span}(x^\top U)) - loss(z|\operatorname{span}(x^\top U_\star)) \\ &= \sigma_z^2 r_{xz}^\top \left(P_{\Sigma_x^{1/2}U_\star} - P_{\Sigma_x^{1/2}U}\right) r_{xz} \\ &= \sigma_z^2 r_{xz}^\top P_{\Sigma_x^{1/2}U_\star} \left(P_{\Sigma_x^{1/2}U_\star} - P_{\Sigma_x^{1/2}U}\right) P_{\Sigma_x^{1/2}U_2} r_{xz} \\ &= \sigma_z^2 r_{xz}^\top P_{\Sigma_x^{1/2}U_\star} \left(I_p - P_{\Sigma_x^{1/2}U}\right) P_{\Sigma_x^{1/2}U_\star} r_{xz} \\ &\le \sigma_z^2 R^2 \left\| P_{\Sigma_x^{1/2}U_\star} \left(I_p - P_{\Sigma_x^{1/2}U}\right) P_{\Sigma_x^{1/2}U_\star} \right\| \\ &= \sigma_z^2 R^2 \left\| P_{\Sigma_x^{1/2}U_\star} \left(I_p - P_{\Sigma_x^{1/2}U}\right) \right\|^2 \\ &= \sigma_z^2 R^2 \left\| P_{\Sigma_x^{1/2}U_\star} - P_{\Sigma_x^{1/2}U} \right\|^2 .\end{aligned}$$

Notice that $P_{\Sigma_x^{1/2}U_\star} \left(I_p - P_{\Sigma_x^{1/2}U}\right) P_{\Sigma_x^{1/2}U_\star}$ is positive definite. Let u_1 be the leading singular vector of this matrix and define $r_{xz}^* = Ru_1$. Then $r_{xz}^* \in \mathcal{A}$ and with such choice of r_{xz}^*, the inequality above will become equality and this implies that

$$\sup_{r_{xz}^* \in \mathcal{A}} \left\{ loss(z|\operatorname{span}(x^\top U)) - loss(z|\operatorname{span}(x^\top U_\star)) \right\} = \sigma_z^2 R^2 \left\| P_{\Sigma_x^{1/2}U} - P_{\Sigma_x^{1/2}U_\star} \right\|^2 .$$

For the last part of the theorem, let $P_{\Sigma_x^{1/2}U_\star} = QQ^\top$. Then the set $\mathcal{A}$ can be rewritten as

$$\mathcal{A} = \{r : r = RQ\widetilde{r},\ \widetilde{r} \in \mathbb{S}^{k-1}\}.$$

The uniform measure (Haar measure) π on $\mathcal{A}$ can be defined through the uniform measure (Haar measure) $\widetilde{\pi}$ on the sphere $\mathbb{S}^{k-1}$. Because $\widetilde{\pi}$ is uniform on the sphere $\mathbb{S}^{k-1}$, then $\mathrm{Var}(\widetilde{\pi}) = I_k/k$ (by symmetry) and

$$\begin{aligned}
&\mathop{\mathbb{E}}_{r_{xz}\sim\pi}\left[loss(z|\operatorname{span}(x^\top U)) - loss(z|\operatorname{span}(x^\top U_\star))\right]\\
=&\sigma_z^2 R^2\ \mathbb{E}_{\widetilde{r}\sim\widetilde{\pi}}\left[\widetilde{r}^\top Q^\top\left(P_{\Sigma_x^{1/2}U_\star} - P_{\Sigma_x^{1/2}U}\right)Q\widetilde{r}\right]\\
=&\sigma_z^2 R^2 \operatorname{tr}\left(Q^\top\left(P_{\Sigma_x^{1/2}U_\star} - P_{\Sigma_x^{1/2}U}\right)Q\right)/k\\
=&\sigma_z^2 R^2 \operatorname{tr}\left(\left(P_{\Sigma_x^{1/2}U_\star} - P_{\Sigma_x^{1/2}U}\right)QQ^\top\right)/k\\
=&\sigma_z^2 R^2 \operatorname{tr}\left(\left(I_p - P_{\Sigma_x^{1/2}U}\right)P_{\Sigma_x^{1/2}U_\star}\right)/k
\end{aligned}$$

Notice that $P_{\Sigma_x^{1/2}U_\star} = P^2_{\Sigma_x^{1/2}U_\star}$,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{r_{xz}\sim\pi}\left[loss(z|\operatorname{span}(x^\top U)) - loss(z|\operatorname{span}(x^\top U_\star))\right]\\
=&\sigma_z^2 R^2 \operatorname{tr}\left(P_{\Sigma_x^{1/2}U_\star}\left(I_p - P_{\Sigma_x^{1/2}U}\right)P_{\Sigma_x^{1/2}U_\star}\right)/k\\
=&\sigma_z^2 R^2 \left\|P_{\Sigma_x^{1/2}U_\star}\left(I_p - P_{\Sigma_x^{1/2}U}\right)\right\|_{\mathrm{F}}^2/k\\
=&\sigma_z^2 R^2 \left\|P_{\Sigma_x^{1/2}U} - P_{\Sigma_x^{1/2}U_\star}\right\|_{\mathrm{F}}^2/2k,
\end{aligned}$$

where the last equality is due to Lemma 2.18.

2.5. Proof of Theorem 2

Throughout the proof, we assume $p_1 = p_2$ for the ease of presentation and the same argument works for any p_1 and p_2 at the cost of heavier notations (when $p_1 \neq p_2$, in the definition (2.14), $\Lambda_2, \widehat{\Lambda}_2$ will be rectangular instead of square matrices. As a result, the subsequent places where Λ_2 appears in the current proof should be understood as either Λ_2 or $\Lambda_2^\top$ according to the dimensions in the specific context). We will still use p_1, p_2 ($p_1 \leq p_2$) to denote the dimension of x and y separately such that the results will be interpretable when $p_1 \neq p_2$.

By the reduction argument in Section 2.3.1, it suffices to consider

$$\Sigma_x = I_{p_1},\ \Sigma_y = I_{p_2},\ \Sigma_{xy} = \Lambda$$

where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{p_1}) \in \mathbb{R}^{p_1 \times p_1}$ is the diagonal matrix with $1 \geq \lambda_1 \geq \ldots \geq \lambda_{p_1} \geq 0$. Under this setup

$$\Phi_{1:p_1} = I_{p_1},\ \Psi_{1:p_1} = I_{p_2,p_1}.$$

and $\lambda_1, \lambda_2, \cdots, \lambda_{p_1}$ are the canonical correlations. Then the error metric is reduced to

$$\left|\left|\left| P_{\Sigma_x^{1/2}\widehat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2}\Phi_{1:k}} \right|\right|\right| = \left|\left|\left| P_{\widehat{\Phi}_{1:k}} - P_{\Phi_{1:k}} \right|\right|\right|$$

where $|||\cdot|||$ denotes either operator or Frobenius norm. Divide $\widehat{\Phi}_{1:k}$ into two blocks such that $\widehat{\Phi}_{1:k} = \begin{bmatrix} \widehat{\Phi}^u_{1:k} \\ \widehat{\Phi}^l_{1:k} \end{bmatrix}$ where $\widehat{\Phi}^u_{1:k}$ and $\widehat{\Phi}^l_{1:k}$ are the upper $k \times k$ and lower $(p_1 - k) \times k$ sub-matrices of $\widehat{\Phi}_{1:k}$ respectively. Let $U = \begin{bmatrix} \widehat{\Phi}^u_{1:k} \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{p_1 \times k}$. Then $P_{\Phi_{1:k}} = P_U$ and by Wedin's $\sin\theta$ law (Wedin, 1972), there exists universal constant C such that

$$\left|\left|\left| P_{\widehat{\Phi}_{1:k}} - P_{\Phi_{1:k}} \right|\right|\right|^2 = \left|\left|\left| P_{\widehat{\Phi}_{1:k}} - P_U \right|\right|\right|^2 \leq \frac{C\left|\left|\left| \widehat{\Phi}_{1:k} - U \right|\right|\right|^2}{\left(\sigma_k(\widehat{\Phi}_{1:k}) - \sigma_{k+1}(U)\right)^2} = \frac{C\left|\left|\left| \widehat{\Phi}^l_{1:k} \right|\right|\right|^2}{\sigma_k^2(\widehat{\Phi}_{1:k})} \tag{2.11}$$

The denominator in (2.11) is close to 1 with high probability when $n \geq C(p_1 + p_2)$ for some constant C. The remaining proof will focus on obtaining upper bounds for the numerator. The upper bound in terms of operator norm is involved and we will present detailed proof. The Frobenius norm bound can be obtained in a very similar (but simpler) manner of which the proof is only sketched. The proof mainly contains 3 parts:

1. Express explicitly (principal term + high order term) each cell of the matrix $\widehat{\Phi}^l_{1:k}$ by a careful Taylor expansion of the estimating equations.

2. Derive two separate deterministic upper bounds for the principal part of $\|\widehat{\Phi}^l_{1:k}\|$. One of the them is tight when λ_{k+1} is bounded away from 1 and the other one is tight when λ_{k+1} is bounded away from 0.

3. Upper bound the high order term and put pieces together.

Throughout the proof, the constants $c, C, \cdots$ might change from line to line.

2.5.1. Taylor Expansion for $\widehat{\Phi}^l_{1:k}$

Recall that $\widehat{\Phi} \in \mathbb{R}^{p_1 \times p_1}$, $\widehat{\Psi} \in \mathbb{R}^{p_2 \times p_1}$ are the sample canonical coefficients. By definition, the sample canonical coefficients satisfy the following two estimating equations (because $\widehat{\Sigma}_x^{1/2}\widehat{\Phi}$ and $\widehat{\Sigma}_y^{1/2}\widehat{\Psi}$ are left and right singular vectors of $\widehat{\Sigma}_x^{-1/2}\widehat{\Sigma}_{xy}\widehat{\Sigma}_y^{-1/2}$ respectively),

$$\widehat{\Sigma}_{xy}\widehat{\Psi} = \widehat{\Sigma}_x\widehat{\Phi}\widehat{\Lambda} \tag{2.12}$$

$$\widehat{\Sigma}_{yx}\widehat{\Phi} = \widehat{\Sigma}_y\widehat{\Psi}\widehat{\Lambda} \tag{2.13}$$

Divide the matrices into blocks,

$$\widehat{\Sigma}_x = \begin{bmatrix} \widehat{\Sigma}_x^{11} & \widehat{\Sigma}_x^{12} \\ \widehat{\Sigma}_x^{21} & \widehat{\Sigma}_x^{22} \end{bmatrix}, \ \widehat{\Sigma}_y = \begin{bmatrix} \widehat{\Sigma}_y^{11} & \widehat{\Sigma}_y^{12} \\ \widehat{\Sigma}_y^{21} & \widehat{\Sigma}_y^{22} \end{bmatrix}, \ \widehat{\Sigma}_{xy} = \begin{bmatrix} \widehat{\Sigma}_{xy}^{11} & \widehat{\Sigma}_{xy}^{12} \\ \widehat{\Sigma}_{xy}^{21} & \widehat{\Sigma}_{xy}^{22} \end{bmatrix}, \ \widehat{\Sigma}_{yx} = \begin{bmatrix} \widehat{\Sigma}_{yx}^{11} & \widehat{\Sigma}_{yx}^{12} \\ \widehat{\Sigma}_{yx}^{21} & \widehat{\Sigma}_{yx}^{22} \end{bmatrix}$$

where $\widehat{\Sigma}_x^{11}, \widehat{\Sigma}_y^{11}, \widehat{\Sigma}_{xy}^{11}, \widehat{\Sigma}_{yx}^{11}$ are $k \times k$ matrices. Similarly, we define

$$\Lambda = \begin{bmatrix} \Lambda_1 & \\ & \Lambda_2 \end{bmatrix}, \ \ \widehat{\Lambda} = \begin{bmatrix} \widehat{\Lambda}_1 & \\ & \widehat{\Lambda}_2 \end{bmatrix}, \tag{2.14}$$

where $\Lambda_1, \widehat{\Lambda}_1$ are also $k \times k$ matrices. Finally, we define $\widehat{\Psi}^u_{1:k} \in \mathbb{R}^{k \times k}$, $\widehat{\Psi}^l_{1:k} \in \mathbb{R}^{(p_2-k)\times k}$ in the same way as $\widehat{\Phi}^u_{1:k}, \widehat{\Phi}^l_{1:k}$. With these notations, we can write the lower left $(p_1 - k) \times k$

sub-matrix of (2.12) and (2.13) explicitly as

$$\widehat{\Sigma}_{xy}^{21}\widehat{\Psi}_{1:k}^{u} + \widehat{\Sigma}_{xy}^{22}\widehat{\Psi}_{1:k}^{l} = \widehat{\Sigma}_{x}^{21}\widehat{\Phi}_{1:k}^{u}\widehat{\Lambda}_1 + \widehat{\Sigma}_{x}^{22}\widehat{\Phi}_{1:k}^{l}\widehat{\Lambda}_1, \tag{2.15}$$

$$\widehat{\Sigma}_{yx}^{21}\widehat{\Phi}_{1:k}^{u} + \widehat{\Sigma}_{yx}^{22}\widehat{\Phi}_{1:k}^{l} = \widehat{\Sigma}_{y}^{21}\widehat{\Psi}_{1:k}^{u}\widehat{\Lambda}_1 + \widehat{\Sigma}_{y}^{22}\widehat{\Psi}_{1:k}^{l}\widehat{\Lambda}_1. \tag{2.16}$$

Similarly, the upper left $k \times k$ sub-matrix of (2.12) and (2.13) can be written explicitly as

$$\widehat{\Sigma}_{xy}^{11}\widehat{\Psi}_{1:k}^{u} + \widehat{\Sigma}_{xy}^{12}\widehat{\Psi}_{1:k}^{l} = \widehat{\Sigma}_{x}^{11}\widehat{\Phi}_{1:k}^{u}\widehat{\Lambda}_1 + \widehat{\Sigma}_{x}^{12}\widehat{\Phi}_{1:k}^{l}\widehat{\Lambda}_1, \tag{2.17}$$

$$\widehat{\Sigma}_{yx}^{11}\widehat{\Phi}_{1:k}^{u} + \widehat{\Sigma}_{yx}^{12}\widehat{\Phi}_{1:k}^{l} = \widehat{\Sigma}_{y}^{11}\widehat{\Psi}_{1:k}^{u}\widehat{\Lambda}_1 + \widehat{\Sigma}_{y}^{12}\widehat{\Psi}_{1:k}^{l}\widehat{\Lambda}_1. \tag{2.18}$$

Manipulate the terms in (2.17),

$$\Lambda_1\widehat{\Psi}_{1:k}^{u} + (\widehat{\Sigma}_{xy}^{11} - \Lambda_1)\widehat{\Psi}_{1:k}^{u} + \widehat{\Sigma}_{xy}^{12}\widehat{\Psi}_{1:k}^{l} = \widehat{\Phi}_{1:k}^{u}\widehat{\Lambda}_1 + (\widehat{\Sigma}_{x}^{11} - I_k)\widehat{\Phi}_{1:k}^{u}\widehat{\Lambda}_1 + \widehat{\Sigma}_{x}^{12}\widehat{\Phi}_{1:k}^{l}\widehat{\Lambda}_1.$$

Therefore,

$$\begin{aligned} \Lambda_1\widehat{\Psi}_{1:k}^{u} - \widehat{\Phi}_{1:k}^{u}\Lambda_1 &= \widehat{\Phi}_{1:k}^{u}(\widehat{\Lambda}_1 - \Lambda_1) + (\widehat{\Sigma}_{x}^{11} - I_k)\widehat{\Phi}_{1:k}^{u}\widehat{\Lambda}_1 + \widehat{\Sigma}_{x}^{12}\widehat{\Phi}_{1:k}^{l}\widehat{\Lambda}_1 \\ &\quad - (\widehat{\Sigma}_{xy}^{11} - \Lambda_1)\widehat{\Psi}_{1:k}^{u} - \widehat{\Sigma}_{xy}^{12}\widehat{\Psi}_{1:k}^{l} := \delta_1. \end{aligned} \tag{2.19}$$

To give some intuition, the equation essentially implies $\Lambda_1\widehat{\Psi}_{1:k}^{u} \approx \widehat{\Phi}_{1:k}^{u}\Lambda_1$ because δ_1 will be proved to be a higher order term. Apply the same argument to (2.18), and we will obtain

$$\begin{aligned} \Lambda_1\widehat{\Phi}_{1:k}^{u} - \widehat{\Psi}_{1:k}^{u}\Lambda_1 &= \widehat{\Psi}_{1:k}^{u}(\widehat{\Lambda}_1 - \Lambda_1) + (\widehat{\Sigma}_{y}^{11} - I_k)\widehat{\Psi}_{1:k}^{u}\widehat{\Lambda}_1 + \widehat{\Sigma}_{y}^{12}\widehat{\Psi}_{1:k}^{l}\widehat{\Lambda}_1 \\ &\quad - (\widehat{\Sigma}_{yx}^{11} - \Lambda_1)\widehat{\Phi}_{1:k}^{u} - \widehat{\Sigma}_{yx}^{12}\widehat{\Phi}_{1:k}^{l} := \delta_2. \end{aligned} \tag{2.20}$$

Similarly, massage the terms in (2.15),

$$\begin{aligned} \widehat{\Sigma}_{xy}^{21}\widehat{\Psi}_{1:k}^{u} + \Lambda_2\widehat{\Psi}_{1:k}^{l} + (\widehat{\Sigma}_{xy}^{22} - \Lambda_2)\widehat{\Psi}_{1:k}^{l} &= \widehat{\Sigma}_{x}^{21}\widehat{\Phi}_{1:k}^{u}\Lambda_1 + \widehat{\Sigma}_{x}^{21}\widehat{\Phi}_{1:k}^{u}(\widehat{\Lambda}_1 - \Lambda_1) \\ &\quad + \widehat{\Phi}_{1:k}^{l}\Lambda_1 + (\widehat{\Sigma}_{x}^{22}\widehat{\Phi}_{1:k}^{l}\widehat{\Lambda}_1 - \widehat{\Phi}_{1:k}^{l}\Lambda_1), \end{aligned}$$

which can be equivalently written as

$$\begin{aligned}\widehat{\Sigma}_{xy}^{21}\widehat{\Psi}_{1:k}^{u}+\Lambda_2\widehat{\Psi}_{1:k}^{l}-\widehat{\Sigma}_{x}^{21}\widehat{\Phi}_{1:k}^{u}\Lambda_1-\widehat{\Phi}_{1:k}^{l}\Lambda_1&=\widehat{\Sigma}_{x}^{21}\widehat{\Phi}_{1:k}^{u}(\widehat{\Lambda}_1-\Lambda_1)\\&+(\widehat{\Sigma}_{x}^{22}\widehat{\Phi}_{1:k}^{l}\widehat{\Lambda}_1-\widehat{\Phi}_{1:k}^{l}\Lambda_1)-(\widehat{\Sigma}_{xy}^{22}-\Lambda_2)\widehat{\Psi}_{1:k}^{l}:=\delta_3.\end{aligned}\tag{2.21}$$

Apply the same argument to (2.16), we will obtain

$$\begin{aligned}\widehat{\Sigma}_{yx}^{21}\widehat{\Phi}_{1:k}^{u}+\Lambda_2\widehat{\Phi}_{1:k}^{l}-\widehat{\Sigma}_{y}^{21}\widehat{\Psi}_{1:k}^{u}\Lambda_1-\widehat{\Psi}_{1:k}^{l}\Lambda_1&=\widehat{\Sigma}_{y}^{21}\widehat{\Psi}_{1:k}^{u}(\widehat{\Lambda}_1-\Lambda_1)\\&+(\widehat{\Sigma}_{y}^{22}\widehat{\Psi}_{1:k}^{l}\widehat{\Lambda}_1-\widehat{\Psi}_{1:k}^{l}\Lambda_1)-(\widehat{\Sigma}_{yx}^{22}-\Lambda_2)\widehat{\Phi}_{1:k}^{l}:=\delta_4.\end{aligned}\tag{2.22}$$

Consider (2.21) × $(-\Lambda_1)-\Lambda_2$ × (2.22), then

$$\begin{aligned}\widehat{\Phi}_{1:k}^{l}\Lambda_1^2-\Lambda_2^2\widehat{\Phi}_{1:k}^{l}+\widehat{\Sigma}_{x}^{21}\widehat{\Phi}_{1:k}^{u}\Lambda_1^2-\widehat{\Sigma}_{xy}^{21}\widehat{\Psi}_{1:k}^{u}\Lambda_1-\Lambda_2\widehat{\Sigma}_{yx}^{21}\widehat{\Phi}_{1:k}^{u}+\Lambda_2\widehat{\Sigma}_{y}^{21}\widehat{\Psi}_{1:k}^{u}\Lambda_1\\=-(\delta_3\Lambda_1+\Lambda_2\delta_4):=\delta_5,\end{aligned}$$

that is

$$\begin{aligned}\widehat{\Phi}_{1:k}^{l}\Lambda_1^2-\Lambda_2^2\widehat{\Phi}_{1:k}^{l}=&\widehat{\Sigma}_{xy}^{21}\widehat{\Psi}_{1:k}^{u}\Lambda_1+\Lambda_2\widehat{\Sigma}_{yx}^{21}\widehat{\Phi}_{1:k}^{u}\\&-\widehat{\Sigma}_{x}^{21}\widehat{\Phi}_{1:k}^{u}\Lambda_1^2-\Lambda_2\widehat{\Sigma}_{y}^{21}\widehat{\Psi}_{1:k}^{u}\Lambda_1+\delta_5.\end{aligned}\tag{2.23}$$

The equation above indicates that, ignoring $\widehat{\Phi}_{1:k}^{u}$, $\widehat{\Psi}_{1:k}^{u}$ and the high order term δ_5, the target is expressed as a linear function of the sample covariance matrices. Later we will show that the sample covariance matrices and $\widehat{\Phi}_{1:k}^{u}$, $\widehat{\Psi}_{1:k}^{u}$ can be decoupled and bounded separately.

2.5.2. Upper Bounds for $\|\widehat{\Phi}_{1:k}^{l}\|$

In this section, we frequently use the following lemma on the Hadamard operator norm for some structured matrices and the proof is deferred to Section 2.7.1.

Lemma 2.6. *(Hadamard Operator Norm) For $A\in\mathbb{R}^{m\times n}$, define the Hadamard operator* *norm as*

$$|||A||| = \sup\left\{\|A \circ B\| : \|B\| \leq 1, B \in \mathbb{R}^{m\times n}\right\}$$

Let $\alpha_1, \cdots, \alpha_m$ and $\beta_1, \cdots, \beta_n$ be arbitrary positive numbers lower bounded by a positive constant δ. Define $A_1, A_2, A_3 \in \mathbb{R}^{m\times n}$ by

$$[A_1]_{ij} = \frac{1}{\alpha_i + \beta_j},\ [A_2]_{ij} = \frac{\min\{\alpha_i, \beta_j\}}{\alpha_i + \beta_j},\ [A_3]_{ij} = \frac{\max\{\alpha_i, \beta_j\}}{\alpha_i + \beta_j}$$

Then

$$|||A_1||| \leq \frac{1}{2\delta},\ |||A_2||| \leq \frac{1}{2},\ |||A_3||| \leq \frac{3}{2}.$$

Upper Bound I: tight for $\lambda_{k+1} \leq 1/2$. Multiplying both sides of (2.19) by Λ_1 on the right will yield

$$\Lambda_1 \widehat{\Psi}^u_{1:k}\Lambda_1 - \widehat{\Phi}^u_{1:k}\Lambda_1^2 = \delta_1\Lambda_1. \tag{2.24}$$

Substitute (2.19), (2.20) and (2.24) into (2.23),

$$\begin{aligned}
\widehat{\Phi}^l_{1:k}\Lambda_1^2 - \Lambda_2^2\widehat{\Phi}^l_{1:k} &= \widehat{\Sigma}^{21}_{xy}\widehat{\Psi}^u_{1:k}\Lambda_1 + \Lambda_2\widehat{\Sigma}^{21}_{yx}\widehat{\Phi}^u_{1:k} - \widehat{\Sigma}^{21}_{x}\Lambda_1\widehat{\Psi}^u_{1:k}\Lambda_1 + \widehat{\Sigma}^{21}_{x}\delta_1\Lambda_1 \\
&\quad - \Lambda_2\widehat{\Sigma}^{21}_{y}\Lambda_1\widehat{\Phi}^u_{1:k} - \Lambda_2\widehat{\Sigma}^{21}_{y}\delta_1 + \delta_5 \\
&= (\widehat{\Sigma}^{21}_{xy} - \widehat{\Sigma}^{21}_{x}\Lambda_1)\widehat{\Psi}^u_{1:k}\Lambda_1 + \Lambda_2(\widehat{\Sigma}^{21}_{yx} - \widehat{\Sigma}^{21}_{y}\Lambda_1)\widehat{\Phi}^u_{1:k} \\
&\quad + \widehat{\Sigma}^{21}_{x}\delta_1\Lambda_1 - \Lambda_2\widehat{\Sigma}^{21}_{y}\delta_1 + \delta_5 \\
&:= B_1\widehat{\Psi}^u_{1:k}\Lambda_1 + \Lambda_2 B_2\widehat{\Phi}^u_{1:k} + \delta_6.
\end{aligned} \tag{2.25}$$

where $B_1 := \widehat{\Sigma}^{21}_{xy} - \widehat{\Sigma}^{21}_{x}\Lambda_1$, $B_2 := \widehat{\Sigma}^{21}_{yx} - \widehat{\Sigma}^{21}_{y}\Lambda_1$ and

$$\delta_6 := \widehat{\Sigma}^{21}_{x}\delta_1\Lambda_1 - \Lambda_2\widehat{\Sigma}^{21}_{y}\delta_1 + \delta_5.$$

Further, define the $(p_1 - k) \times k$ matrices A_1, A_2 by

$$[A_1]_{ij} = \frac{1}{\lambda_j + \lambda_{k+i}},\ [A_2]_{ij} = \frac{1}{\lambda_j - \lambda_{k+i}},\ 1 \le i \le p_1 - k, 1 \le j \le k$$

Then we can rewrite (2.25) as

$$\begin{aligned}\widehat{\Phi}^l_{1:k} &= A_1 \circ A_2 \circ (B_1 \widehat{\Psi}^u_{1:k} \Lambda_1) + A_1 \circ A_2 \circ (\Lambda_2 B_2 \widehat{\Phi}^u_{1:k}) + A_1 \circ A_2 \circ \delta_6 \\ &= (A_1 \Lambda_1) \circ A_2 \circ (B_1 \widehat{\Psi}^u_{1:k}) + (\Lambda_2 A_1) \circ A_2 \circ (B_2 \widehat{\Phi}^u_{1:k}) + A_1 \circ A_2 \circ \delta_6,\end{aligned}$$

where $\circ$ denotes the matrix Hadamard (element-wise) product. Define $\alpha_j := \lambda_j, 1 \le j \le k$ and $\beta_i := \lambda_{k+i}, 1 \le i \le p_1 - k$, then

$$[A_1 \Lambda_1]_{ij} = \frac{\lambda_j}{\lambda_j + \lambda_{k+i}} = \frac{\max\{\alpha_j, \beta_i\}}{\alpha_j + \beta_i},\ [\Lambda_2 A_1]_{ij} = \frac{\lambda_{k+i}}{\lambda_j + \lambda_{k+i}} = \frac{\min\{\alpha_j, \beta_i\}}{\alpha_j + \beta_i} \tag{2.26}$$

Therefore, by Lemma 2.6,

$$\begin{aligned}\|(A_1\Lambda_1) \circ A_2 \circ (B_1 \widehat{\Psi}^u_{1:k})\| &\le \frac{3}{2} \|A_2 \circ (B_1 \widehat{\Psi}^u_{1:k})\|, \\ \|(\Lambda_2 A_1) \circ A_2 \circ (B_2 \widehat{\Phi}^u_{1:k})\| &\le \frac{1}{2} \|A_2 \circ (B_2 \widehat{\Phi}^u_{1:k})\|.\end{aligned}$$

Observe that

$$[A_2]_{ij} = \frac{1}{\lambda_j - \lambda_{k+i}} = \frac{1}{(\lambda_j - \lambda_k + \Delta/2) + (\lambda_k - \Delta/2 - \lambda_{k+i})} = \frac{1}{a_j + b_i}, \tag{2.27}$$

where $a_j := \lambda_j - (\lambda_k - \Delta/2), 1 \le j \le k$ and $b_i := (\lambda_k - \Delta/2) - \lambda_{i+k}, 1 \le i \le p_1 - k$. Then $a_j, b_i \ge \Delta/2$ and again apply Lemma 2.6,

$$\begin{aligned}\left\|A_2 \circ (B_1 \widehat{\Psi}^u_{1:k})\right\| &\le \frac{1}{\Delta} \left\|B_1 \widehat{\Psi}^u_{1:k}\right\| \le \frac{1}{\Delta} \|B_1\| \left\|\widehat{\Psi}^u_{1:k}\right\|, \\ \left\|A_2 \circ (B_2 \widehat{\Phi}^u_{1:k})\right\| &\le \frac{1}{\Delta} \left\|B_2 \widehat{\Phi}^u_{1:k}\right\| \le \frac{1}{\Delta} \|B_2\| \left\|\widehat{\Phi}^u_{1:k}\right\|,\end{aligned}$$

which further implies

$$\left\|\widehat{\Phi}^l_{1:k}\right\| \leq \frac{1}{2\Delta}\left(3\left\|B_1\right\|\left\|\widehat{\Psi}^u_{1:k}\right\| + \left\|B_2\right\|\left\|\widehat{\Phi}^u_{1:k}\right\|\right) + \left\|A_1 \circ A_2 \circ \delta_6\right\|. \tag{2.28}$$

In Section 2.5.4, we will show that this bound is tight and matches Theorem 2 when λ_{k+1} is away from 1. Now we switch to deriving the other upper bound which will be tight when λ_{k+1} is close to 1.

Upper Bound II: tight for $\lambda_{k+1} \geq 1/2$. Notice that $\Lambda_1 \times$ (2.19) + $\Lambda_1 \times$ (2.20) yields

$$\Lambda_1^2 \widehat{\Psi}^u_{1:k} - \widehat{\Psi}^u_{1:k}\Lambda_1^2 = \Lambda_1\delta_1 + \delta_2\Lambda_1. \tag{2.29}$$

Substitute (2.19), (2.20) and (2.29) into (2.23),

$$\begin{aligned}\widehat{\Phi}^l_{1:k}\Lambda_1^2 - \Lambda_2^2\widehat{\Phi}^l_{1:k} &= \widehat{\Sigma}^{21}_{xy}\Lambda_1\widehat{\Phi}^u_{1:k} + \widehat{\Sigma}^{21}_{xy}\delta_1 + \Lambda_2\widehat{\Sigma}^{21}_{yx}\widehat{\Phi}^u_{1:k} - \widehat{\Sigma}^{21}_x\Lambda_1^2\widehat{\Phi}^u_{1:k} \\ &\quad + \widehat{\Sigma}^{21}_x(\Lambda_1\delta_1 + \delta_2\Lambda_1) - \Lambda_2\widehat{\Sigma}^{21}_y\Lambda_1\widehat{\Phi}^u_{1:k} + \Lambda_2\widehat{\Sigma}^{21}_y\delta_2 + \delta_5 \\ &= B\widehat{\Phi}^u_{1:k} + \delta_7,\end{aligned}$$

where we define

$$\begin{aligned}B &= \widehat{\Sigma}^{21}_{xy}\Lambda_1 + \Lambda_2\widehat{\Sigma}^{21}_{yx} - \widehat{\Sigma}^{21}_x\Lambda_1^2 - \Lambda_2\widehat{\Sigma}^{21}_y\Lambda_1, \\ \delta_7 &= \widehat{\Sigma}^{21}_{xy}\delta_1 + \widehat{\Sigma}^{21}_x(\Lambda_1\delta_1 + \delta_2\Lambda_1) + \Lambda_2\widehat{\Sigma}^{21}_y\delta_2 + \delta_5.\end{aligned}$$

Again with the definition of A_1 and A_2,

$$\widehat{\Phi}^l_{1:k} = A_1 \circ A_2 \circ (B\widehat{\Phi}^u_{1:k}) + A_1 \circ A_2 \circ \delta_7.$$

Notice that we can rewrite A_1 as

$$[A_1]_{ij} = \frac{1}{\lambda_j + \lambda_{k+i}} = \frac{1}{(\lambda_j - \lambda_k/2) + (\lambda_{k+i} + \lambda_k/2)} = \frac{1}{\alpha_j + \beta_i},$$

where $\alpha_j := \lambda_j - \lambda_k/2, 1 \leq j \leq k$ and $\beta_i := \lambda_{k+i} + \lambda_k/2, 1 \leq i \leq p_1 - k$. Hence $\alpha_j, \beta_i \geq \lambda_k/2$, and apply Lemma 2.6,

$$\|\widehat{\Phi}^l_{1:k}\| \leq \frac{1}{\lambda_k}\left(\left\|A_2 \circ (B\widehat{\Phi}^u_{1:k})\right\| + \|A_2 \circ \delta_7\|\right). \tag{2.30}$$

Define k^* as the largest index i such that

$$k+1 \leq i \leq p_1, \quad \lambda_i \geq 2\lambda_k - 1 - \Delta.$$

Divide the indexes $1, \cdots, p_1 - k$ into two sets:

$$\mathcal{I}_1 = \{i - k : k+1 \leq i \leq k^*\}, \quad \mathcal{I}_2 = \{i - k : k^* + 1 \leq i \leq p_1\},$$

and accordingly divide A_2 and $B\widehat{\Phi}^u_{1:k}$ into two blocks:

$$A_2 = \begin{bmatrix} A_2^{(1)} \\ A_2^{(2)} \end{bmatrix}, \quad B\widehat{\Phi}^u_{1:k} = \begin{bmatrix} B^{(1)}\widehat{\Phi}^u_{1:k} \\ B^{(2)}\widehat{\Phi}^u_{1:k} \end{bmatrix},$$

where $A_2^{(1)}, B^{(1)}$ corresponds to the rows indexed by $\mathcal{I}_1$ and $A_2^{(2)}, B^{(2)}$ corresponds to the rows indexed by $\mathcal{I}_2$. Then

$$A_2 \circ (B\widehat{\Phi}^u_{1:k}) = \begin{bmatrix} A_2^{(1)} \circ (B^{(1)}\widehat{\Phi}^u_{1:k}) \\ A_2^{(2)} \circ (B^{(2)}\widehat{\Phi}^u_{1:k}) \end{bmatrix},$$

and by triangle inequality,

$$\left\|A_2 \circ (B\widehat{\Phi}^u_{1:k})\right\| \leq \left\|A_2^{(1)} \circ (B^{(1)}\widehat{\Phi}^u_{1:k})\right\| + \left\|A_2^{(2)} \circ (B^{(2)}\widehat{\Phi}^u_{1:k})\right\|.$$

For the first part, by the same argument as in (2.27)

$$\|A_2^{(1)} \circ (B^{(1)}\widehat{\Phi}^u_{1:k})\| \leq \frac{1}{\Delta}\|B^{(1)}\|\|\widehat{\Phi}^u_{1:k}\|. \tag{2.31}$$

For the second part, let $D = \operatorname{diag}(\lambda_k - \Delta/2 - \lambda_{k^*+1}, \cdots, \lambda_k - \Delta/2 - \lambda_{p_1}) \in \mathbb{R}^{(p_1-k^*)\times(p_1-k^*)}$. Then

$$A_2^{(2)} \circ (B^{(2)} \widehat{\Phi}^u_{1:k}) = (DA_2^{(2)}) \circ (D^{-1} B^{(2)} \widehat{\Phi}^u_{1:k}).$$

Notice that for $1 \le i \le p_1 - k^*$, $1 \le j \le k$

$$\begin{aligned}
[DA_2^{(2)}]_{ij} &= \frac{(\lambda_k - \Delta/2 - \lambda_{k^*+i})}{\lambda_j - \lambda_{i+k}} \\
&= \frac{(\lambda_k - \Delta/2 - \lambda_{k^*+i})}{\lambda_j - (\lambda_k - \Delta/2) + (\lambda_k - \Delta/2 - \lambda_{k^*+i})} \\
&= \frac{b_i}{a_j + b_i},
\end{aligned}$$

where $a_j := \lambda_j - (\lambda_k - \Delta/2), 1 \le j \le k$ and $b_i := (\lambda_k - \Delta/2) - \lambda_{k^*+i}$ for $1 \le i \le p_1 - k^*$. Further observe that by definition of k^*,

$$b_i - a_j = (\lambda_k - \Delta/2) - \lambda_{k^*+i} - \lambda_j + (\lambda_k - \Delta/2) \ge 2\lambda_k - 1 - \Delta - \lambda_{k^*+i} \ge 0.$$

This implies

$$[DA_2^{(2)}]_{ij} = \frac{\max\{a_j, b_i\}}{a_j + b_i}.$$

Again, by Lemma 2.6,

$$\|A_2^{(2)} \circ (B^{(2)} \widehat{\Phi}^u_{1:k})\| \le \frac{3}{2} \|D^{-1} B^{(2)} \widehat{\Phi}^u_{1:k}\| \le \frac{3}{2} \|D^{-1} B^{(2)}\| \|\widehat{\Phi}^u_{1:k}\|. \tag{2.32}$$

Substitute (2.31) and (2.32) into (2.30),

$$\begin{aligned}
\|\widehat{\Phi}^l_{1:k}\| &\le \frac{1}{\lambda_k} \left(\left\| A_2 \circ (B \widehat{\Phi}^u_{1:k}) \right\| + \|A_2 \circ \delta_7\| \right) \\
&\le \frac{1}{2\lambda_k \Delta} \|B^{(1)}\| \|\widehat{\Phi}^u_{1:k}\| + \frac{3}{2\lambda_k} \|D^{-1} B^{(2)}\| \|\widehat{\Phi}^u_{1:k}\| + \frac{1}{\lambda_k} \|A_2 \circ \delta_7\|.
\end{aligned} \tag{2.33}$$

2.5.3. Upper Bounds in Expectation for the Principal Terms

We state upper bounds for key quantities in (2.28) and (2.33) with proofs deferred to the supplement. Both Lemma 2.8 and Lemma 2.9 are proved by a covering argument to upper bound quadratic forms of Gaussian random variables.

Lemma 2.7. *There exists a universal constant C such that*

$$\mathbb{E}[\|B_1\|^2],\ \mathbb{E}[\|B_2\|^2] \leq C\frac{p_1}{n}(1-\lambda_k^2).$$

See Section 2.7.2 for the proof of this lemma.

Lemma 2.8. *There exist universal constants c, C_1 such that the following inequality holds with probability at least $1-2e^{-ct^2}$,*

$$\|D^{-1}B^{(2)}\| \leq \sqrt{\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{\Delta^2}}\max\{\delta,\delta^2\} \qquad \delta = C_1\left(\sqrt{\frac{p_1}{n}}+\frac{t}{\sqrt{n}}\right).$$

As a corollary, there exists constant C_2,

$$\mathbb{E}\|D^{-1}B^{(2)}\|^2 \leq C_2\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{\Delta^2}\frac{p_1}{n}.$$

See Section 2.7.3 for the proof of this lemma.

Lemma 2.9. *There exists universal constants c, C_1 such that the following inequality holds with probability at least $1-2e^{-ct^2}$,*

$$\|B^{(1)}\| \leq \sqrt{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}\max\{\delta,\delta^2\} \qquad \delta = C_1\left(\sqrt{\frac{p_1}{n}}+\frac{t}{\sqrt{n}}\right).$$

As a corollary, there exists constant C_2,

$$\mathbb{E}\|B^{(1)}\|^2 \leq C_2\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)p_1}{n}.$$

See Section 2.7.4 for the proof of this lemma.

2.5.4. Upper Bound for Operator Norm

For the ease of presentation, we introduce $z = (x^\top, y^\top)^\top$ as the concatenation of x and y. Then the population and sample covariances of z can be written as,

$$\Sigma_z = \begin{bmatrix} I_{p_1} & \Sigma_{xy} \\ \Sigma_{yx} & I_{p_2} \end{bmatrix}, \quad \widehat{\Sigma}_z = \begin{bmatrix} \widehat{\Sigma}_x & \widehat{\Sigma}_{xy} \\ \widehat{\Sigma}_{yx} & \widehat{\Sigma}_y \end{bmatrix}.$$

The advantage of introducing z is that the sample-population discrepancy for x and y can be simultaneously bounded by that of z.

Lemma 2.10. *There exists a universal constant C such that the following inequality holds deterministically,*

$$\|A_1 \circ A_2 \circ \delta_6\|, \ \|A_2 \circ \delta_7\| \leq \frac{C\|\Sigma_z - \widehat{\Sigma}_z\|^2}{\Delta^2}(2 + \|\widehat{\Sigma}_z\|)^2(\|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \widehat{\Sigma}_z\|),$$

where $\Delta = \lambda_k - \lambda_{k+1}$ is the eigen-gap.

See Section 2.7.5 for the proof.

Lemma 2.11. *There exists universal constant c, C_1, C_2 such that when $n \geq C_1(p_1 + p_2)$, the following inequality holds*

$$\sigma_k^2(\widehat{\Phi}_{1:k}) \geq 1/2,$$
$$(2 + \|\widehat{\Sigma}_z\|)^2(\|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \widehat{\Sigma}_z\|) \leq C_2,$$

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

See Section 2.7.6 for the proof.

Let G be the event that the inequalities in Lemma 2.11 hold. Notice that for any pair of projection matrices (P_1, P_2), $\|P_1 - P_2\| \le \|P_1\| + \|P_2\| \le 2$. Substitute into equation (2.11),

$$\begin{aligned}\mathbb{E}\left\|P_{\widehat{\Phi}_{1:k}} - P_{\Phi_{1:k}}\right\|^2 &\le 4\mathbb{P}(G^c) + \mathbb{E}\left[\left\|P_{\widehat{\Phi}_{1:k}} - P_{\Phi_{1:k}}\right\|^2 I_G\right] \\ &\le 4\exp(-cn) + 2\mathbb{E}\left[\|\widehat{\Phi}^l_{1:k}\|^2 I_G\right].\end{aligned}$$

Now we plug in the upper bounds of $\|\widehat{\Phi}^l_{1:k}\|$ obtained in (2.28) and (2.33) respectively. On event G, (2.28) can be reduced to

$$\|\widehat{\Phi}^l_{1:k}\|^2 \le \frac{C}{\Delta^2}\left(\|B_1\|^2 + \|B_2\|^2\right) + C\left\|A_1 \circ A_2 \circ \delta_6\right\|^2.$$

Further, by Lemma 2.10 and Lemma 2.11, on event G,

$$\|\widehat{\Phi}^l_{1:k}\|^2 \le \frac{C}{\Delta^2}\left(\|B_1\|^2 + \|B_2\|^2\right) + \frac{C}{\Delta^4}\|\Sigma_z - \widehat{\Sigma}_z\|^4.$$

Therefore,

$$\mathbb{E}\left[\|\widehat{\Phi}^l_{1:k}\|^2 I_G\right] \le \frac{C}{\Delta^2}\mathbb{E}\left(\|B_1\|^2 + \|B_2\|^2\right) + \frac{C}{\Delta^4}\mathbb{E}\|\Sigma_z - \widehat{\Sigma}_z\|^4.$$

By Lemma 2.7 and Lemma 2.20 (notice that $\|\Sigma_z\| \le 2$),

$$\mathbb{E}\left[\|\widehat{\Phi}^l_{1:k}\|^2 I_G\right] \le \frac{C(1-\lambda_k^2)p_1}{\Delta^2 n} + C\left(\frac{p_1 + p_2}{n\Delta^2}\right)^2. \tag{2.34}$$

This upper bound implies the result in Theorem 2 when λ_{k+1} is bounded away from 1, for instance, when $\lambda_{k+1} \le 1/2$. Now, we use (2.33) for the case that $\lambda_{k+1} \ge 1/2$. On event G, with $\lambda_{k+1} \ge 1/2$, (2.33) can be reduced to

$$\|\widehat{\Phi}^l_{1:k}\|^2 \le C\left(\frac{1}{\Delta^2}\|B^{(1)}\|^2 + \|D^{-1}B^{(2)}\|^2 + \|A_2 \circ \delta_7\|^2\right).$$

Apply Lemma 2.10 and Lemma 2.11, on event G,

$$\|\widehat{\Phi}^l_{1:k}\|^2 \le C\left(\frac{1}{\Delta^2}\|B^{(1)}\|^2 + \|D^{-1}B^{(2)}\|^2 + \frac{1}{\Delta^4}\|\Sigma_z - \widehat{\Sigma}_z\|^4\right).$$

By Lemma 2.8, Lemma 2.9 and Lemma 2.10,

$$\mathbb{E}\left[\|\widehat{\Phi}^l_{1:k}\|^2 I_G\right] \le \frac{C(1-\lambda_k^2)(1-\lambda_{k+1}^2)p_1}{\Delta^2 n} + C\left(\frac{p_1+p_2}{n\Delta^2}\right)^2. \tag{2.35}$$

Combine the results of (2.34) and (2.35),

$$\begin{aligned}\mathbb{E}\left\|P_{\widehat{\Phi}_{1:k}} - P_{\Phi_{1:k}}\right\|^2 &\le 4\exp(-cn) + \frac{C(1-\lambda_k^2)(1-\lambda_{k+1}^2)p_1}{\Delta^2 n} + C\left(\frac{p_1+p_2}{n\Delta^2}\right)^2 \\ &\le C\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)p_1}{\Delta^2 n} + C_2\left(\frac{p_1+p_2}{n\Delta^2}\right)^2\end{aligned}$$

2.5.5. Upper Bound for Frobenius Norm

A quick upper bound in terms of Frobenius norm can be obtained by noticing that $P_{\widehat{\Phi}_{1:k}} - P_{\Phi_{1:k}}$ has rank at most $2k$ and therefore,

$$\begin{aligned}\frac{1}{k}\mathbb{E}\left\|P_{\widehat{\Phi}_{1:k}} - P_{\Phi_{1:k}}\right\|_{\mathrm{F}}^2 \le 2\mathbb{E}\left\|P_{\widehat{\Phi}_{1:k}} - P_{\Phi_{1:k}}\right\|^2 \\ \le C\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)p_1}{\Delta^2 n} + C\left(\frac{p_1+p_2}{n\Delta^2}\right)^2.\end{aligned}$$

In fact, the factor p_1 in the main term can be reduced to $p_1 - k$ by similar (but much simpler) arguments as done for the operator norm. We state the corresponding results in this section without proof. Specifically, the following Frobenius norm counterparts for (2.28) and (2.33) can be obtained.

Lemma 2.12. *Let $\widetilde{D} = diag(\lambda_k - \lambda_{k+1}, \cdots, \lambda_k - \lambda_{p_1})$, then*

$$\left\|\widehat{\Phi}^l_{1:k}\right\|_{\mathrm{F}} \leq \frac{1}{2\Delta}\left(3\left\|B_1\right\|_{\mathrm{F}}\left\|\widehat{\Psi}^u_{1:k}\right\| + \left\|B_2\right\|_{\mathrm{F}}\left\|\widehat{\Phi}^u_{1:k}\right\|\right) + \left\|A_1 \circ A_2 \circ \delta_6\right\|_{\mathrm{F}},$$

$$\left\|\widehat{\Phi}^l_{1:k}\right\|_{\mathrm{F}} \leq \frac{1}{\lambda_k}\|\widetilde{D}^{-1}B\|_{\mathrm{F}}\|\widehat{\Phi}^u_{1:k}\| + \frac{1}{\lambda_k}\left\|A_2 \circ \delta_7\right\|_{\mathrm{F}}.$$

For the second inequality, the divide-and-conquer analysis used in Section 2.5.2 is no longer necessary due to the observation that

$$\|A \circ M\|_{\mathrm{F}} \leq \|M\|_{\mathrm{F}}, \ \forall A \text{ satisfying } \max_{i,j}|A_{ij}| \leq 1$$

while the inequality is not true for the operator norm. Similarly, parallel results to the lemmas in Section 2.5.3 can be derived as follows (see Section 2.7.7 for the proof of the second inequality in the lemma as illustration).

Lemma 2.13. *There exists a universal constant C such that*

$$\mathbb{E}[\|B_1\|^2_{\mathrm{F}}], \ \mathbb{E}[\|B_2\|^2_{\mathrm{F}}] \leq C\frac{(1-\lambda_k^2)(p_1-k)k}{n},$$

$$\mathbb{E}\left[\left\|\widetilde{D}^{-1}B\right\|^2_{\mathrm{F}}\right] \leq 2\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)(p_1-k)k}{n\Delta^2},$$

$$\|A_1 \circ A_2 \circ \delta_6\|_{\mathrm{F}}, \ \|A_2 \circ \delta_7\|_{\mathrm{F}} \leq \frac{C\sqrt{k}\|\Sigma_z - \widehat{\Sigma}_z\|^2}{\Delta^2}(2 + \|\widehat{\Sigma}_z\|)^2(\|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \widehat{\Sigma}_z\|).$$

Substituting these lemmas into the procedure of Section 2.5.4 will yield

$$\mathbb{E}\left\|P_{\widehat{\Phi}_{1:k}} - P_{\Phi_{1:k}}\right\|^2_{\mathrm{F}}/k \leq \frac{C(1-\lambda_k^2)(1-\lambda_{k+1}^2)(p_1-k)}{\Delta^2 n} + C\left(\frac{p_1+p_2}{n\Delta^2}\right)^2.$$

2.6. Proof of Theorem 3

2.6.1. On Kullback-Leibler Divergence

The construction in equation (2.36) of the following lemma is crucial to prove the lower bound. The proof of the lemma can be found in Section 2.6.4.

Lemma 2.14. *For $i = 1, 2$ and $p_2 \geq p_1 \geq k$, let $\begin{bmatrix} U_{(i)}, & W_{(i)} \end{bmatrix} \in \mathcal{O}(p_1, p_1)$, $\begin{bmatrix} V_{(i)}, & Z_{(i)} \end{bmatrix} \in \mathcal{O}(p_2, p_1)$ where $U_{(i)} \in \mathbb{R}^{p_1 \times k}, V_{(i)} \in \mathbb{R}^{p_2 \times k}$. For $0 \leq \lambda_2 < \lambda_1 < 1$, let $\Delta = \lambda_1 - \lambda_2$ and define*

$$\Sigma_{(i)} = \begin{bmatrix} \Sigma_x & \Sigma_x^{1/2}(\lambda_1 U_{(i)} V_{(i)}^\top + \lambda_2 W_{(i)} Z_{(i)}^\top)\Sigma_y^{1/2} \\ \Sigma_y^{1/2}(\lambda_1 V_{(i)} U_{(i)}^\top + \lambda_2 Z_{(i)} W_{(i)}^\top)\Sigma_x^{1/2} & \Sigma_y \end{bmatrix} \quad i = 1, 2,$$

Let $\mathbb{P}_{(i)}$ denote the distribution of a random i.i.d. sample of size n from $N(0, \Sigma_{(i)})$. If we further assume

$$[U_{(1)}, W_{(1)}] \begin{bmatrix} V_{(1)}^\top \\ Z_{(1)}^\top \end{bmatrix} = [U_{(2)}, W_{(2)}] \begin{bmatrix} V_{(2)}^\top \\ Z_{(2)}^\top \end{bmatrix}, \tag{2.36}$$

Then one can show that

$$D(\mathbb{P}_{(1)} || \mathbb{P}_{(2)}) = \frac{n\Delta^2(1 + \lambda_1\lambda_2)}{2(1 - \lambda_1^2)(1 - \lambda_2^2)} \|U_{(1)} V_{(1)}^\top - U_{(2)} V_{(2)}^\top\|_{\mathrm{F}}^2.$$

See Section 2.6.4 for the proof.

2.6.2. Packing Number and Fano's Lemma

The following result on the packing number is based on the metric entropy of the Grassmannian manifold $G(k, r)$ due to Szarek (1982). We use the version adapted from Lemma 1 of Cai et al. (2013) which is also used in Gao et al. (2015b).

Lemma 2.15. *For any fixed $U_0 \in \mathcal{O}(p, k)$ and $\mathcal{B}_{\epsilon_0} = \{U \in \mathcal{O}(p, k) : \|UU^\top - U_0 U_0^\top\|_{\mathrm{F}} \leq \epsilon_0\}$ with $\epsilon_0 \in (0, \sqrt{2[k \wedge (p - k)]})$. Define the semi-metric $\rho(\cdot, \cdot)$ on $\mathcal{B}_{\epsilon_0}$ by*

$$\rho(U_1, U_2) = \|U_1 U_1^\top - U_2 U_2^\top\|_{\mathrm{F}}.$$

Then there exists universal constant C such that for any $\alpha \in (0, 1)$, the packing number

$\mathcal{M}(\mathcal{B}_{\epsilon_0}, \rho, \alpha\epsilon_0)$ satisfies

$$\mathcal{M}(\mathcal{B}_{\epsilon_0}, \rho, \alpha\epsilon_0) \geq \left(\frac{1}{C\alpha}\right)^{k(p-k)}.$$

The following corollary is used to prove the lower bound.

Corollary 2.16. *If we change the set in Lemma 2.15 to $\widetilde{\mathcal{B}}_{\epsilon_0} = \{U \in \mathcal{O}(p,k) : \|U - U_0\|_{\mathrm{F}} \leq \epsilon_0\}$, then we still have*

$$\mathcal{M}(\widetilde{\mathcal{B}}_{\epsilon_0}, \rho, \alpha\epsilon_0) \geq \left(\frac{1}{C\alpha}\right)^{k(p-k)}.$$

Proof. Apply Lemma 2.15 to $\mathcal{B}_{\epsilon_0}$, there exists $U_1, \cdots, U_n$ with $n \geq (1/C\alpha)^{k(p-k)}$ such that

$$\|U_iU_i^\top - U_0U_0^\top\|_{\mathrm{F}} \leq \epsilon_0,\ 1 \leq i \leq n,\ \|U_iU_i^\top - U_jU_j^\top\|_{\mathrm{F}} \geq \alpha\epsilon_0, 1 \leq i \leq j \leq n.$$

Define $\widetilde{U}_i = \arg\min_{U \in \{U_iQ,\ Q \in \mathcal{O}(k)\}} \|U - U_0\|_{\mathrm{F}}$, by Lemma 2.25,

$$\|\widetilde{U}_i - U_0\|_{\mathrm{F}} \leq \|\widetilde{U}_i\widetilde{U}_i^\top - U_0U_0^\top\|_{\mathrm{F}} \leq \epsilon_0.$$

Therefore, $\widetilde{U}_1, \cdots, \widetilde{U}_n \in \widetilde{\mathcal{B}}_{\epsilon_0}$ and

$$\|\widetilde{U}_i\widetilde{U}_i^\top - \widetilde{U}_j\widetilde{U}_j^\top\|_{\mathrm{F}} = \|U_iU_i^\top - U_jU_j^\top\|_{\mathrm{F}} \geq \alpha\epsilon_0.$$

which implies,

$$\mathcal{M}(\widetilde{\mathcal{B}}_{\epsilon_0}, \rho, \alpha\epsilon_0) \geq n \geq \left(\frac{1}{C\alpha}\right)^{k(p-k)}.$$

□

Lemma 2.17 (Fano's Lemma Yu (1997)). *Let (Θ, ρ) be a (semi)metric space and $\{\mathbb{P}_\theta : \theta \in \Theta\}$ a collection of probability measures. For any totally bounded $T \subset \Theta$, denote $\mathcal{M}(T, \rho, \epsilon)$ the ϵ-packing number of T with respect to the metric ρ, i.e. , the maximal number of points* *in T whoese pairwise minimum distance in ρ is at least ϵ. Define the Kullback-Leibler diameter of T by*

$$d_{KL}(T) = \sup_{\theta,\theta' \in T} D(\mathbb{P}_\theta || \mathbb{P}_{\theta'}).$$

Then,

$$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \left[\rho^2(\widehat{\theta},\theta)\right] \geq \sup_{T\subset\Theta} \sup_{\epsilon>0} \frac{\epsilon^2}{4}\Big(1 - \frac{d_{KL}(T) + log\ 2}{log\ \mathcal{M}(T,\rho,\epsilon)}\Big)$$

2.6.3. Proof of Lower Bound

For any fixed $\begin{bmatrix} U_{(0)}, & W_{(0)} \end{bmatrix} \in \mathcal{O}(p_1,p_1)$ and $\begin{bmatrix} V_{(0)}, & Z_{(0)} \end{bmatrix} \in \mathcal{O}(p_2,p_1)$ where $U_{(0)} \in \mathbb{R}^{p_1\times k}, V_{(0)} \in \mathbb{R}^{p_2\times k}, W_{(0)} \in \mathbb{R}^{p_1\times(p_1-k)}, V_{(0)} \in \mathbb{R}^{p_2\times(p_2-k)}$, define

$$\begin{aligned}\mathcal{H}_{\epsilon_0} = \Big\{ \big(U,W,V,Z\big) : \ \begin{bmatrix} U, & W \end{bmatrix} \in \mathcal{O}(p_1,p_1) \text{ with } U \in \mathbb{R}^{p_1\times k},\ \begin{bmatrix} V, & Z \end{bmatrix} \in \mathcal{O}(p_2,p_1) \\ \text{with } V \in \mathbb{R}^{p_2\times k}, \|U - U_{(0)}\|_{\mathrm{F}} \leq \epsilon_0,\ [U,W]\begin{bmatrix} V^\top \\ Z^\top \end{bmatrix} = [U_{(0)}, W_{(0)}]\begin{bmatrix} V_{(0)}^\top \\ Z_{(0)}^\top \end{bmatrix} \Big\}.\end{aligned}$$

For any fixed $\Sigma_x \in \mathbb{S}_+^{p_1}, \Sigma_y \in \mathbb{S}_+^{p_2}$ with $\kappa(\Sigma_x) = \kappa_x, \kappa(\Sigma_y) = \kappa_y$, consider the parametrization $\Sigma_{xy} = \Sigma_x \Phi\Lambda\Psi^\top\Sigma_y$, for $0 \leq \lambda_{k+1} < \lambda_k < 1$, define

$$\begin{aligned}\mathcal{T}_{\epsilon_0} = \Big\{ \Sigma = \begin{bmatrix} \Sigma_x & \Sigma_x^{1/2}(\lambda_k UV^\top + \lambda_{k+1}WZ^\top)\Sigma_y^{1/2} \\ \Sigma_y^{1/2}(\lambda_k VU^\top + \lambda_{k+1}ZW^\top)\Sigma_x^{1/2} & \Sigma_y \end{bmatrix}, \\ \Phi = \Sigma_x^{-1/2}[U,W], \Psi = \Sigma_y^{-1/2}[V,Z], \big(U,W,V,Z\big) \in \mathcal{H}_{\epsilon_0} \Big\}.\end{aligned}$$

It is straightforward to verify that $\mathcal{T}_{\epsilon_0} \subset \mathcal{F}(p_1,p_2,k,\lambda_k,\lambda_{k+1},\kappa_x,\kappa_y)$. For any $\Sigma_{(i)} \in \mathcal{T}_{\epsilon_0}$, $i = 1,2$, they yield to the parametrization,

$$\Sigma_{(i)} = \begin{bmatrix} \Sigma_x & \Sigma_x^{1/2}(\lambda_k U_{(i)}V_{(i)}^\top + \lambda_{k+1}W_{(i)}Z_{(i)}^\top)\Sigma_y^{1/2} \\ \Sigma_y^{1/2}(\lambda_k V_{(i)}U_{(i)}^\top + \lambda_{k+1}Z_{(i)}W_{(i)}^\top)\Sigma_x^{1/2} & \Sigma_y \end{bmatrix},$$

where $\left(U_{(i)}, W_{(i)}, V_{(i)}, Z_{(i)}\right) \in \mathcal{H}_{\epsilon_0}$ and the leading-k canonical vectors are $\Phi_{1:k}^{(i)} = \Sigma_x^{-1/2} U_{(i)}, \Psi_{1:k}^{(i)} = \Sigma_y^{-1/2} V_{(i)}$. We define a semi-metric on $\mathcal{T}_{\epsilon_0}$ as

$$\rho(\Sigma_{(1)}, \Sigma_{(2)}) = \left\| P_{\Sigma_x^{1/2}\Phi_{1:k}^{(1)}} - P_{\Sigma_x^{1/2}\Phi_{1:k}^{(2)}} \right\|_{\mathrm{F}} = \left\| P_{U_{(1)}} - P_{U_{(2)}} \right\|_{\mathrm{F}}.$$

By Lemma 2.14,

$$D(\mathbb{P}_{\Sigma_1} || \mathbb{P}_{\Sigma_2}) = \frac{n\Delta^2(1+\lambda_k\lambda_{k+1})}{2(1-\lambda_k^2)(1-\lambda_{k+1}^2)} \|U_{(1)}V_{(1)}^\top - U_{(2)}V_{(2)}^\top\|_{\mathrm{F}}^2.$$

Further by the definition of $d_{KL}(T)$,

$$d_{KL}(T) = \frac{n\Delta^2(1+\lambda_k\lambda_{k+1})}{2(1-\lambda_k^2)(1-\lambda_{k+1}^2)} \sup_{\Sigma_{(1)}, \Sigma_{(2)} \in \mathcal{T}_{\epsilon_0}} \|U_{(1)}V_{(1)}^\top - U_{(2)}V_{(2)}^\top\|_{\mathrm{F}}^2. \tag{2.37}$$

To bound the Kullback-Leibler diameter, for any $\Sigma_{(1)}, \Sigma_{(2)} \in \mathcal{T}_{\epsilon_0}$, by definition,

$$[U_{(1)}, W_{(1)}] \begin{bmatrix} V_{(1)}^\top \\ Z_{(1)}^\top \end{bmatrix} = [U_{(2)}, W_{(2)}] \begin{bmatrix} V_{(2)}^\top \\ Z_{(2)}^\top \end{bmatrix},$$

which implies that they are singular value decompositions of the same matrix. Therefore, there exists $Q \in \mathcal{O}(p_1, p_1)$ such that

$$[U_{(2)}, W_{(2)}] = [U_{(1)}, W_{(1)}]Q \ , \ [V_{(2)}, Z_{(2)}] = [V_{(1)}, Z_{(1)}]Q. \tag{2.38}$$

Decompose Q into four blocks such that

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}.$$

Substitute into (2.38),

$$U_{(2)} = U_{(1)}Q_{11} + W_{(1)}Q_{21}, \ V_{(2)} = V_{(1)}Q_{11} + Z_{(1)}Q_{21}.$$

Then,

$$\begin{aligned}\|U_{(2)} - U_{(1)}\|_{\mathrm{F}}^2 &= \|U_{(1)}(Q_{11} - I_k) + W_{(1)}Q_{21}\|_{\mathrm{F}}^2 \\ &= \|U_{(1)}(Q_{11} - I_k)\|_{\mathrm{F}}^2 + \|W_{(1)}Q_{21}\|_{\mathrm{F}}^2 \\ &= \|Q_{11} - I_k\|_{\mathrm{F}}^2 + \|Q_{21}\|_{\mathrm{F}}^2.\end{aligned}$$

The second equality is due to the fact that $U_{(1)}$ and $W_{(1)}$ have orthogonal column space and the third equality is valid because $U_{(1)}, W_{(1)} \in \mathcal{O}(p_1, k)$. By the same argument, we will have

$$\|V_{(2)} - V_{(1)}\|_{\mathrm{F}}^2 = \|Q_{11} - I_k\|_{\mathrm{F}}^2 + \|Q_{21}\|_{\mathrm{F}}^2.$$

Notice that

$$\begin{aligned}\|U_{(1)}V_{(1)}^\top - U_{(2)}V_{(2)}^\top\|_{\mathrm{F}}^2 &= \|(U_{(1)} - U_{(2)})V_{(1)} + U_{(2)}(V_{(1)} - V_{(2)})\|_{\mathrm{F}}^2 \\ &\leq 2\|U_{(1)} - U_{(2)}\|_{\mathrm{F}}^2 + 2\|V_{(1)} - V_{(2)}\|_{\mathrm{F}}^2 \\ &= 4\|(U_{(1)} - U_{(2)})\|_{\mathrm{F}}^2 \\ &\leq 8\left(\|(U_{(1)} - U_{(0)})\|_{\mathrm{F}}^2 + \|(U_{(0)} - U_{(2)})\|_{\mathrm{F}}^2\right) \\ &\leq 16\epsilon_0^2.\end{aligned}$$

Then, substitute into (2.37)

$$d_{KL}(T) \leq \frac{8n\Delta^2(1 + \lambda_k\lambda_{k+1})}{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)}\epsilon_0^2. \tag{2.39}$$

Let $\mathcal{B}_{\epsilon_0} = \{U \in O(p_1, k) : \|U - U_{(0)}\|_{\mathrm{F}} \leq \epsilon_0\}$. Under the semi-metric $\widetilde{\rho}(U_{(1)}, U_{(2)}) = \|U_{(1)}U_{(1)}^\top - U_{(2)}U_{(2)}^\top\|_{\mathrm{F}}$, we claim that the packing number of $\mathcal{H}_{\epsilon_0}$ is lower bounded by the packing number of $\mathcal{B}_{\epsilon_0}$. To prove this claim, it suffices to show that for any $U \in \mathcal{B}_{\epsilon_0}$, there exists corresponding W, V, Z such that $(U, W, V, Z) \in \mathcal{H}_{\epsilon_0}$. First of all, by definition, $\|U - U_0\|_{\mathrm{F}} \leq \epsilon_0$. Let $W \in \mathcal{O}(p_1, p_1 - k)$ be the orthogonal complement of U. Then

$[U,W] \in \mathcal{O}(p_1,p_1)$ and therefore there exists $Q \in \mathcal{O}(p_1,p_1)$ such that

$$[U,W] = [U_{(0)}, W_0]Q.$$

Set $[V,Z] = [V_{(0)}, Z_0]Q \in \mathcal{O}(p_2,p_1)$, then

$$[U,W]\begin{bmatrix} V^\top \\ \\ Z^\top \end{bmatrix} = [U_{(0)}, W_{(0)}]\begin{bmatrix} V_{(0)}^\top \\ \\ Z_{(0)}^\top \end{bmatrix},$$

which implies $(U,W,V,Z) \in \mathcal{H}_{\epsilon_0}$. Let

$$\epsilon = \alpha\epsilon_0 = c\left(\sqrt{k \wedge (p_1-k)} \wedge \sqrt{\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{n\Delta^2(1+\lambda_k\lambda_{k+1})}k(p_1-k)}\right),$$

where $c \in (0,1)$ depends on α and is chosen small enough such that $\epsilon_0 = \epsilon/\alpha \in (0, \sqrt{2[k \wedge (p_1-k)]}\,)$. By Corollary 2.16,

$$\mathcal{M}(\mathcal{T}_{\epsilon_0}, \rho, \alpha\epsilon_0) = \mathcal{M}(\mathcal{H}_{\epsilon_0}, \widetilde{\rho}, \alpha\epsilon_0) \geq \mathcal{M}(\mathcal{B}_{\epsilon_0}, \widetilde{\rho}, \alpha\epsilon_0) \geq \left(\frac{1}{C\alpha}\right)^{k(p_1-k)}.$$

Apply Lemma 2.17 with $\mathcal{T}_{\epsilon_0}, \rho, \epsilon$,

$$\inf_{\widehat{\Phi}_{1:k}} \sup_{\Sigma\in\mathcal{F}} \mathbb{E}\left[\ \left\| P_{\Sigma_x^{1/2}\widehat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2}\Phi_{1:k}} \right\|_{\mathrm{F}}^2\right] \geq \sup_{T\subset\Theta} \sup_{\epsilon>0} \frac{\epsilon^2}{4}\left(1 - \frac{8c^2k(p_1-k)+\log 2}{k(p_1-k)\log\frac{1}{C\alpha}}\right).$$

Choose α small enough such that

$$1 - \frac{8c^2k(p_1-k)+\log 2}{k(p_1-k)\log\frac{1}{C\alpha}} \geq \frac{1}{2}.$$

Then the lower bound is reduced to

$$\inf_{\widehat{\Phi}_{1:k}} \sup_{\Sigma\in\mathcal{F}} \mathbb{E}\left[\ \left\|P_{\Sigma_x^{1/2}\widehat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2}\Phi_{1:k}}\right\|_{\mathrm{F}}^2\right] \geq \frac{c^2}{8}\left\{\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{n\Delta^2(1+\lambda_k\lambda_{k+1})}k(p_1-k)\wedge k\wedge(p_1-k)\right\}$$

$$\geq C^2k\left\{\left(\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{\Delta^2}\frac{p_1-k}{n}\right)\wedge 1\wedge\frac{p_1-k}{k}\right\}$$

By symmetry,

$$\inf_{\widehat{\Psi}_{1:k}} \sup_{\Sigma\in\mathcal{F}} \mathbb{E}\left[\ \left\|P_{\Sigma_y^{1/2}\widehat{\Psi}_{1:k}} - P_{\Sigma_y^{1/2}\Psi_{1:k}}\right\|_{\mathrm{F}}^2\right] \geq C^2k\left\{\left(\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{\Delta^2}\frac{p_1-k}{n}\right)\wedge 1\wedge\frac{p_1-k}{k}\right\}$$

The lower bound for operator norm error can be immediately obtained by noticing that $P_{\Sigma_y^{1/2}\widehat{\Psi}_{1:k}} - P_{\Sigma_y^{1/2}\Psi_{1:k}}$ has at most rank $2k$ and

$$\left\|P_{\Sigma_x^{1/2}\widehat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2}\Phi_{1:k}}\right\|^2 \geq \frac{1}{2k}\left\|P_{\Sigma_x^{1/2}\widehat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2}\Phi_{1:k}}\right\|_{\mathrm{F}}^2$$

2.6.4. Proof of Lemma 2.14

By simple algebra, the Kullback-Leibler divergence between two multivariate gaussian distributions satisfies

$$D(\mathbb{P}_{\Sigma_{(1)}}||\mathbb{P}_{\Sigma_{(2)}}) = \frac{n}{2}\Big\{\mathrm{Tr}\left(\Sigma_{(2)}^{-1}(\Sigma_{(1)}-\Sigma_{(2)})\right) - \log\det(\Sigma_{(2)}^{-1}\Sigma_{(1)})\Big\}.$$

Notice that

$$\Sigma_{(i)} = \begin{bmatrix}\Sigma_x^{1/2} & 0\\ 0 & \Sigma_y^{1/2}\end{bmatrix}\Omega_{(i)}\begin{bmatrix}\Sigma_x^{1/2} & 0\\ 0 & \Sigma_y^{1/2}\end{bmatrix},$$

where

$$\Omega_{(i)} = \begin{bmatrix} I_{p_1} & \lambda_1U_{(i)}V_{(i)}^\top + \lambda_2W_{(i)}Z_{(i)}^\top\\ \lambda_1V_{(i)}U_{(i)}^\top + \lambda_2Z_{(i)}W_{(i)}^\top & I_{p_2}\end{bmatrix}.$$

Then,

$$D(\mathbb{P}_{\Sigma_{(1)}}||\mathbb{P}_{\Sigma_{(2)}}) = \frac{n}{2}\Big\{\mathrm{Tr}(\Omega_{(2)}^{-1}\Omega_{(1)}) - (p_1+p_2) - \log\det(\Omega_{(2)}^{-1}\Omega_{(1)})\Big\}.$$

Also notice that

$$\Omega_{(i)} = \begin{bmatrix} I_{p_1} & 0 \\ 0 & I_{p_2} \end{bmatrix} + \frac{\lambda_1}{2} \begin{bmatrix} U_{(i)} \\ V_{(i)} \end{bmatrix} \begin{bmatrix} U_{(i)}^\top & V_{(i)}^\top \end{bmatrix} - \frac{\lambda_1}{2} \begin{bmatrix} U_{(i)} \\ -V_{(i)} \end{bmatrix} \begin{bmatrix} U_{(i)}^\top & -V_{(i)}^\top \end{bmatrix}$$
$$+ \frac{\lambda_2}{2} \begin{bmatrix} W_{(i)} \\ Z_{(i)} \end{bmatrix} \begin{bmatrix} W_{(i)}^\top & Z_{(i)}^\top \end{bmatrix} - \frac{\lambda_2}{2} \begin{bmatrix} W_{(i)} \\ -Z_{(i)} \end{bmatrix} \begin{bmatrix} W_{(i)}^\top & -Z_{(i)}^\top \end{bmatrix}.$$

Therefore $\Omega_{(1)}, \Omega_{(2)}$ share the same set of eigenvalues: $1 + \lambda_1$ with multiplicity k, $1 - \lambda_1$ with multiplicity k, $1 + \lambda_2$ with multiplicity $p_1 - k$, $1 - \lambda_2$ with multiplicity $p_1 - k$ and 1 with multiplicity $2(p_2 - p_1)$. This implies $\log \det(\Omega_{(2)}^{-1}\Omega_{(1)})) = 0$. On the other hand, by block inversion formula, we can compute

$$\Omega_{(2)}^{-1} = \begin{bmatrix} I_{p_1} + \frac{\lambda_1^2}{1-\lambda_1^2} U_{(2)} U_{(2)}^\top + \frac{\lambda_2^2}{1-\lambda_2} W_{(2)} W_{(2)}^\top & -\frac{\lambda_1}{1-\lambda_1^2} U_{(2)} V_{(2)}^\top - \frac{\lambda_2}{1-\lambda_2} W_{(2)} Z_{(2)}^\top \\ -\frac{\lambda_1}{1-\lambda_1^2} V_{(2)} U_{(2)}^\top - \frac{\lambda_2}{1-\lambda_2} Z_{(2)} W_{(2)}^\top & I_{p_2} + \frac{\lambda_1^2}{1-\lambda_1^2} V_{(2)} V_{(2)}^\top + \frac{\lambda_2^2}{1-\lambda_2} Z_{(2)} Z_{(2)}^\top \end{bmatrix}.$$

Divide $\Omega_{(2)}^{-1}\Omega_{(1)}$ into blocks such that

$$\Omega_{(2)}^{-1}\Omega_{(1)} = \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix} \quad \textit{where} \quad J_{11} \in \mathbb{R}^{p_1 \times p_1},\ J_{22} \in \mathbb{R}^{p_2 \times p_2},$$

and

$$\begin{aligned} J_{11} = {} & \frac{\lambda_1^2}{1-\lambda_1^2}(U_{(2)}U_{(2)}^\top - U_{(2)}V_{(2)}^\top V_{(1)}U_{(1)}^\top) + \frac{\lambda_2^2}{1-\lambda_2^2}(W_{(2)}W_{(2)} - W_{(2)}Z_{(2)}^\top Z_{(1)}W_{(1)}^\top) \\ & - \frac{\lambda_1\lambda_2}{1-\lambda_1^2}(U_{(2)}V_{(2)}^\top Z_{(1)}W_{(1)}^\top) - \frac{\lambda_1\lambda_2}{1-\lambda_2^2}(W_{(2)}Z_{(2)}^\top V_{(1)}U_{(1)}^\top) \\ J_{22} = {} & \frac{\lambda_1^2}{1-\lambda_1^2}(V_{(2)}V_{(2)}^\top - V_{(2)}U_{(2)}^\top U_{(1)}V_{(1)}^\top) + \frac{\lambda_2^2}{1-\lambda_2^2}(Z_{(2)}Z_{(2)} - Z_{(2)}W_{(2)}^\top W_{(1)}Z_{(1)}^\top) \\ & - \frac{\lambda_1\lambda_2}{1-\lambda_1^2}(V_{(2)}U_{(2)}^\top W_{(1)}Z_{(1)}^\top) - \frac{\lambda_1\lambda_2}{1-\lambda_2^2}(Z_{(2)}W_{(2)}^\top U_{(1)}V_{(1)}^\top). \end{aligned}$$

We spell out the algebra for $tr(J_{11})$, and $tr(J_{22})$ can be computed in exactly the same fashion.

$$\begin{aligned} tr(U_{(2)}U_{(2)}^\top - U_{(2)}V_{(2)}^\top V_{(1)}U_{(1)}^\top) &= \frac{1}{2} tr(U_{(2)}V_{(2)}^\top V_{(2)}U_{(2)}^\top + U_{(1)}V_{(1)}^\top V_{(1)}U_{(1)}^\top - 2U_{(2)}V_{(2)}^\top V_{(1)}U_{(1)}^\top) \\ &= \frac{1}{2}\|U_{(1)}V_{(1)}^\top - U_{(2)}V_{(2)}\|_{\mathrm{F}}^2. \end{aligned}$$

Similarly,

$$tr(W_{(2)}W_{(2)} - W_{(2)}Z_{(2)}^\top Z_{(1)}W_{(1)}^\top) = \frac{1}{2}\|W_{(1)}Z_{(1)}^\top - W_{(2)}Z_{(2)}\|_{\mathrm{F}}^2.$$

By the assumption, $U_{(1)}V_{(1)}^\top + W_{(1)}Z_{(1)}^\top = U_{(2)}V_{(2)}^\top + W_{(2)}Z_{(2)}^\top$, which implies

$$tr(W_{(2)}W_{(2)} - W_{(2)}Z_{(2)}^\top Z_{(1)}W_{(1)}^\top) = \frac{1}{2}\|U_{(1)}V_{(1)}^\top - U_{(2)}V_{(2)}\|_{\mathrm{F}}^2.$$

Further,

$$\begin{aligned} tr(U_{(2)}V_{(2)}^\top Z_{(1)}W_{(1)}^\top) &= tr\left(U_{(2)}V_{(2)}^\top (U_{(2)}V_{(2)}^\top + W_{(2)}Z_{(2)}^\top - U_{(1)}V_{(1)}^\top)^\top\right) \\ &= tr\left(U_{(2)}V_{(2)}^\top (U_{(2)}V_{(2)}^\top - U_{(1)}V_{(1)}^\top)^\top\right) \\ &= \frac{1}{2}\|U_{(1)}V_{(1)}^\top - U_{(2)}V_{(2)}\|_{\mathrm{F}}^2, \end{aligned}$$

and by the same argument,

$$tr(W_{(2)}Z_{(2)}^\top V_{(1)}U_{(1)}^\top) = \frac{1}{2}\|U_{(1)}V_{(1)}^\top - U_{(2)}V_{(2)}\|_{\mathrm{F}}^2.$$

Sum these equations,

$$\begin{aligned} \mathrm{tr}(J_{11}) &= \frac{1}{2}\left\{\frac{\lambda_1^2}{1-\lambda_1^2} + \frac{\lambda_2^2}{1-\lambda_2^2} - \frac{\lambda_1\lambda_2}{1-\lambda_1^2} - \frac{\lambda_1\lambda_2}{1-\lambda_2^2}\right\}\|U_{(1)}V_{(1)}^\top - U_{(2)}V_{(2)}\|_{\mathrm{F}}^2 \\ &= \frac{\Delta^2(1+\lambda_1\lambda_2)}{2(1-\lambda_1^2)(1-\lambda_2^2)}\|U_{(1)}V_{(1)}^\top - U_{(2)}V_{(2)}\|_{\mathrm{F}}^2. \end{aligned}$$

Repeat the argument for J_{22}, one can show that

$$\mathrm{tr}(J_{22}) = \mathrm{tr}(J_{11}) = \frac{\Delta^2(1+\lambda_1\lambda_2)}{2(1-\lambda_1^2)(1-\lambda_2^2)}\|U_{(1)}V_{(1)}^\top - U_{(2)}V_{(2)}\|_{\mathrm{F}}^2.$$

Therefore,

$$\begin{aligned} D(\mathbb{P}_{\Sigma_{(1)}}||\mathbb{P}_{\Sigma_{(2)}}) &= \frac{n}{2}tr(\Omega_{(2)}^{-1}\Omega_{(1)}) = \frac{n}{2}\left(tr(J_{11}) + tr(J_{22})\right) \\ &= \frac{n\Delta^2(1+\lambda_1\lambda_2)}{2(1-\lambda_1^2)(1-\lambda_2^2)}\|U_{(1)}V_{(1)}^\top - U_{(2)}V_{(2)}\|_{\mathrm{F}}^2. \end{aligned}$$

2.7. Proof of Technical Lemmas

The proofs in this section rely on the following supporting lemmas.

Lemma 2.18. *(Subspace Angles Wedin (1983)) For $U_i \in \mathcal{O}(p,k)$ and $P_i = U_iU_i^\top,\ i = 1,2,$*

$$\begin{gathered} \|P_1 - P_2\|_{\mathrm{F}}^2 = 2\|(I_p - P_1)P_2\|_{\mathrm{F}}^2 = 2\|(I_p - P_2)P_1\|_{\mathrm{F}}^2, \\ \|P_1 - P_2\|^2 = \|(I_p - P_1)P_2\|^2 = \|(I_p - P_2)P_1\|^2 = 1 - \sigma_{\min}^2(P_1P_2) \end{gathered}$$

Lemma 2.19. *(Covariance Matrix Estimation, Remark 5.40 of Vershynin (2010)) Assume $A \in \mathbb{R}^{n\times p}$ has independent sub-gaussian random rows with second moment matrix Σ. Then there exists universal constant C such that for every $t \geq 0$, the following inequality holds with probability at least $1 - e^{-ct^2}$,*

$$\|\frac{1}{n}A^\top A - \Sigma\| \leq \max\{\delta, \delta^2\}\|\Sigma\| \qquad \delta = C\sqrt{\frac{p}{n}} + \frac{t}{\sqrt{n}}.$$

Lemma 2.20. *Assume $A \in \mathbb{R}^{n\times p}, n \geq p$ has independent sub-gaussian random rows with second moment matrix Σ. Then there exists universal constant C such that*

$$\mathbb{E}\|\frac{1}{n}A^\top A - \Sigma\|^4 \leq C\frac{p^2}{n^2}\|\Sigma\|^4.$$

Proof. Without loss of generality, we can assume $\|\Sigma\| = 1$ or else we can scale A by $1/\sqrt{\|\Sigma\|}$. Let $J = \|\frac{1}{n}A^\top A - \Sigma\|$ and by Lemma 2.19, there exists positive constants c_1, C_1 such that

$$P(J \geq \max\{\delta, \delta^2\}) \leq e^{-c_1 t^2}, \quad \delta = C_1\sqrt{\frac{p}{n}} + \frac{t}{\sqrt{n}}.$$

Notice that $J \geq 0$, then

$$\begin{aligned}
\mathbb{E}[J^4] &= \int_0^{+\infty} P(J \geq x^{1/4})dx \\
&= \int_0^{C_1^2p^2/n^2} P(J \geq x^{1/4})dx + \int_{C_1^2p^2/n^2}^1 P(J \geq x^{1/4})dx + \int_1^{+\infty} P(J \geq x^{1/4})dx \\
&\leq C_1^2p^2/n^2 + \int_{C_1^2p^2/n^2}^1 e^{-(\sqrt{n}x^{1/4}-C_1\sqrt{p})^2}dx + \int_1^{+\infty} e^{-(\sqrt{n}x^{1/8}-C_1\sqrt{p})^2}dx \\
&= C_1^2p^2/n^2 + \int_{C_1^2p^2/n^2}^1 4e^{-y^2}\left(\frac{y+C_1\sqrt{p}}{\sqrt{n}}\right)^3 \frac{1}{\sqrt{n}}dy + \int_1^{+\infty} 8e^{-y^2}\left(\frac{y+C_1\sqrt{p}}{\sqrt{n}}\right)^7 \frac{1}{\sqrt{n}}dy.
\end{aligned}$$

There exists a large constant C_2 such that

$$\begin{aligned}
\mathbb{E}[J^4] &\leq C_1^2p^2/n^2 + \frac{4}{n^2}\int_{C_1^2p^2/n^2}^1 4e^{-y^2}C_2(y^3 + p^{2/3})dy + \frac{8}{n^4}\int_1^{+\infty} C_2e^{-y^2}(y^7 + p^{7/2})dy \\
&\leq C_1^2p^2/n^2 + \frac{4}{n^2}\int_0^{+\infty} 4e^{-y^2}C_2(y^3 + p^{2/3})dy + \frac{8}{n^4}\int_0^{+\infty} C_2e^{-y^2}(y^7 + p^{7/2})dy.
\end{aligned}$$

Notice that $\int_0^{+\infty} e^{-y^2}y^k dy$ is bounded for any $k \in \mathbb{Z}_+$ and $n \geq p$. There exists a large constant C_3 such that

$$\mathbb{E}[J^4] \leq C_3\frac{p^2}{n^2}.$$

□

Lemma 2.21. *(Bernstein inequality, Proposition 5.16 of Vershynin (2010)) Let* $X_1, \cdots, X_n$ *be independent centered sub-exponential random variables and* $K = \max_i \|X_i\|_{\psi_1}$. *Then for* *every $a = (a_1, \cdots, a_n) \in \mathbb{R}^n$ and every $t \geq 0$, we have*

$$P\left\{|\sum_{i=1}^{n} a_i X_i| \geq t\right\} \leq 2exp\left\{-c\min\left(\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right)\right\}.$$

Lemma 2.22. *(Hanson-Wright inequality, Theorem 1.1 of Rudelson and Vershynin (2013)) Let $x = (x_1, \cdots, x_p)$ be a random vectors with independent components x_i which satisfy $\mathbb{E}x_i = 0$ and $\|x_i\|_{\psi_2} \leq K$, Let $A \in \mathbb{R}^{p\times p}$. Then there exists universal constant c such that for every $t \geq 0$,*

$$P\left\{|x^\top Ax - \mathbb{E}x^\top Ax| \geq t\right\} \leq 2exp\left\{-c\min\left(\frac{t^2}{K^4\|A\|_{\mathrm{F}}^2}, \frac{t}{K^2\|A\|}\right)\right\}.$$

Lemma 2.23. *(Covering Number of the Sphere, Lemma 5.2 of Vershynin (2010)). The unit Euclidean sphere $\mathbb{S}^{n-1}$ equipped with the Euclidean metric satisfies for every $\epsilon > 0$ that*

$$|\mathcal{N}(\mathbb{S}^{n-1}, \epsilon)| \leq (1 + \frac{2}{\epsilon})^n,$$

where $\mathcal{N}(\mathbb{S}^{n-1}, \epsilon)$ is the ϵ-net of $\mathbb{S}^{n-1}$ with minimal cardinality.

The following variant of Wedin's $\sin\theta$ law (Wedin, 1972) is proved in Proposition 1 of Cai et al. (2015).

Lemma 2.24. *For $A, E \in \mathbb{R}^{m\times n}$ and $\widehat{A} = A + E$, define the singular value decomposition of A and $\widehat{A}$ as*

$$A = UDV^\top, \ \widehat{A} = \widehat{U}\widehat{D}\widehat{V}^\top.$$

Then the following perturbation bound holds,

$$\left\|(I - P_{U_{1:k}})\, P_{\widehat{U}_{1:k}}\right\| = \left\|P_{U_{1:k}} - P_{\widehat{U}_{1:k}}\right\| \leq \frac{2\|E\|}{\sigma_k(A) - \sigma_{k+1}(A)},$$

where $\sigma_k(A), \sigma_{k+1}(A)$ are the k_{th} and $(k+1)_{th}$ singular values of A.

Lemma 2.25. *For any matrices* $U_1, U_2 \in \mathcal{O}(p,k)$*,*

$$\inf_{Q \in \mathcal{O}(k,k)} \|U_1 - U_2 Q\|_{\mathrm{F}} \leq \|P_{U_1} - P_{U_2}\|_{\mathrm{F}}$$

Proof. Since U_1, U_2 are orthonormal matrices,

$$\|U_1 - U_2 Q\|_{\mathrm{F}}^2 = 2k - 2tr(U_1^\top U_2 Q).$$

Let $U_1^\top U_2 = UDV^\top$ be the singular value decomposition. Then $VU^\top \in O(k,k)$ and

$$\begin{aligned} \inf_{Q \in O(k,k)} \|U_1 - U_2 Q\|_{\mathrm{F}}^2 &\leq 2k - 2tr(U_1^\top U_2 V U^\top) \\ &= 2k - 2tr(UDU^\top) \\ &= 2k - 2tr(D). \end{aligned}$$

On the other hand,

$$\begin{aligned} \|P_{U_1} - P_{U_2}\|_{\mathrm{F}}^2 &= \|U_1 U_1^\top - U_2 U_2^\top\|_{\mathrm{F}}^2 \\ &= 2k - 2tr(U_1 U_1^\top U_2 U_2^\top) \\ &= 2k - 2tr(U_1^\top U_2 U_2^\top U_1) \\ &= 2k - 2tr(D^2). \end{aligned}$$

Since $U_1, U_2 \in O(p,k)$, $\|U_1^\top U_2\| \leq 1$ and therefore all the diagonal elements of D is less than 1, which implies that $tr(D) \geq tr(D^2)$ and

$$\inf_{Q \in O(k,k)} \|U_1 - U_2 Q\|_{\mathrm{F}}^2 \leq \|P_{U_1} - P_{U_2}\|_{\mathrm{F}}^2.$$

□

Lemma 2.26. *(Theorem 5.5.18 of Hom and Johnson (1991)) If $A, B \in \mathbb{R}^{n\times n}$ and A is positive semidefinite. Then,*

$$\|A \circ B\| \le \left(\max_{1\le i\le n} A_{ii}\right) \|B\|,$$

where $\|\cdot\|$ is the operator norm.

Lemma 2.27. *(Theorem A of Fiedler (2010)) A symmetric Cauthy matrix*

$$\boldsymbol{C} = \Big(\frac{1}{a_i + a_j}\Big)_{1\le i,j\le n}$$

is positive semidefinite if $a_i > 0, 1 \le i \le n$.

2.7.1. Proof of Lemma 2.6

Define $\gamma_i = \beta_i, 1 \le i \le n$ and $\gamma_i = \alpha_{i-n}, n+1 \le i \le m+n$. Consider the matrix $M_1, M_2 \in \mathbb{R}^{(m+n)\times(m+n)}$ define by

$$[M_1]_{ij} = \frac{1}{\gamma_i + \gamma_j},\ [M_2]_{ij} = \frac{\min\{\gamma_i, \gamma_j\}}{\gamma_i + \gamma_j}.$$

By Lemma 2.27, M_1 is positive semidefinite and by Lemma 2.26,

$$|||M_1||| \le \frac{1}{2 \min\limits_{1\le i\le m+n} \{\gamma_i\}} \le \frac{1}{2\delta}.$$

Notice that A_1 is the lower left sub-matrix of M_1, therefore,

$$|||A_1||| \le |||M_1||| \le \frac{1}{2\delta}.$$

By Theorem 3.2 of Mathias (1993), M_2 is also positive semidefinite. Again, apply Lemma 2.26 and notice that A_2 is the lower left sub-matrix of M_2,

$$|||A_2||| \le |||M_2||| \le \frac{1}{2}.$$

Finally, observe that, by definition, $A_3 \circ B = B - A_2 \circ B$, hence

$$\|A_3 \circ B\| \le \|B\| + \|A_2 \circ B\|,$$

which implies,

$$|||A_3||| \le 1 + |||A_2||| \le \frac{3}{2}.$$

2.7.2. Proof of Lemma 2.7

Divide x, y into two parts,

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \; y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix},$$

where $x_1, y_1 \in \mathbb{R}^k$, $x_2 \in \mathbb{R}^{p_1-k}$ and $y_2 \in \mathbb{R}^{p_2-k}$. By definition,

$$B_1 = \widehat{\text{Cov}}(x_2, y_1 - \Lambda_1 x_1),$$

where $\widehat{\text{Cov}}(\cdot,\cdot)$ denotes the sample covariance operator. When $\lambda_k = 1$, by definition of CCA, $y_1 = x_1$ almost surely, which implies that $B_1 = 0$ almost surely. When $\lambda_k < 1$,

$$B_1 = \widehat{\text{Cov}}(x_2, y_1 - \Lambda_1 x_1) = \sqrt{1-\lambda_k^2}\widehat{\text{Cov}}\left(x_2, \frac{y_1 - \Lambda_1 x_1}{\sqrt{1-\lambda_k^2}}\right) = \sqrt{1-\lambda_k^2}\widehat{\text{Cov}}(w_1, w_2),$$

where we define $w_1 = \frac{y_1 - \Lambda_1 x_1}{\sqrt{1-\lambda_k^2}}$ and $w_2 = x_2$. Let $w = (w_1^\top, w_2^\top)^\top$ be the concatenation of w_1 and w_2. Then

$$\Sigma_w = \text{Var}(w) = \text{diag}\left(\frac{1-\lambda_1^2}{1-\lambda_k^2}, \frac{1-\lambda_2^2}{1-\lambda_k^2}, \cdots, \frac{1-\lambda_k^2}{1-\lambda_k^2}, 1, \cdots, 1\right).$$

Notice that $\|\Sigma_w\| \le 1$ and $\mathbb{E}[\widehat{\text{Cov}}(w_1, w_2)] = 0$. Therefore,

$$\|B_1\|^2 \le (1-\lambda_k^2)\|\widehat{\text{Var}}(w) - \Sigma_w\|^2.$$

By Lemma 2.20,

$$\mathbb{E}[\|B_1\|^2] \leq C\frac{(1-\lambda_k^2)p_1}{n}.$$

The result for B_2 can be derived in the same manner and we skip the proof.

2.7.3. Proof of Lemma 2.8

We use a covering argument to prove the lemma.

Step 1. Reduction. For $\epsilon > 0$ and any pair of vectors $u \in \mathbb{R}^{p_1-k^*}, v \in \mathbb{R}^k$, we can choose $u_\epsilon \in \mathcal{N}(\mathbb{S}^{p_1-k^*-1}, \epsilon), v_\epsilon \in \mathcal{N}(\mathbb{S}^{k-1}, \epsilon)$ such that $\|u-u_\epsilon\|, \|v-v_\epsilon\| \leq \epsilon$. Then

$$\begin{aligned} u^\top D^{-1}B^{(2)}v &= u^\top D^{-1}B^{(2)}v - u_\epsilon^\top D^{-1}B^{(2)}v + u_\epsilon^\top D^{-1}B^{(2)}v - u_\epsilon^\top D^{-1}B^{(2)}v_\epsilon + u_\epsilon^\top D^{-1}B^{(2)}v_\epsilon \\ &\leq \|u-u_\epsilon\|\|D^{-1}B^{(2)}v\| + \|u_\epsilon^\top D^{-1}B^{(2)}\|\|v-v_\epsilon\| + u_\epsilon^\top D^{-1}B^{(2)}v_\epsilon \\ &\leq 2\epsilon\|D^{-1}B^{(2)}\| + u_\epsilon^\top D^{-1}B^{(2)}v_\epsilon \\ &\leq 2\epsilon\|D^{-1}B^{(2)}\| + \max_{u_\epsilon, v_\epsilon} u_\epsilon^\top D^{-1}B^{(2)}v_\epsilon. \end{aligned}$$

Maximize over u and v, we obtain

$$\|D^{-1}B^{(2)}\| \leq 2\epsilon\|D^{-1}B^{(2)}\| + \max_{u_\epsilon, v_\epsilon} u_\epsilon^\top D^{-1}B^{(2)}v_\epsilon. \tag{2.40}$$

Therefore, $\|D^{-1}B^{(2)}\| \leq (1-2\epsilon)^{-1}\max_{u_\epsilon, v_\epsilon} u_\epsilon^\top D^{-1}B^{(2)}v_\epsilon$. Let $\epsilon = 1/4$. Then it suffices to prove with required probability,

$$\max_{u_\epsilon, v_\epsilon} u_\epsilon^\top D^{-1}B^{(2)}v_\epsilon \leq \frac{1}{2}\max\{\delta, \delta^2\}. \tag{2.41}$$

Step 2. Concentration. Notice that for $1 \leq j \leq k < k^* + 1 \leq i \leq p_1$,

$$\begin{aligned}
[D^{-1}B^{(2)}]_{i-k^*,j} &= \frac{1}{\lambda_k - \Delta/2 - \lambda_i} \frac{1}{n} \sum_{\alpha=1}^{n} (\lambda_j x_{\alpha i} y_{\alpha j} - \lambda_j^2 x_{\alpha i} x_{\alpha j} + \lambda_i x_{\alpha j} y_{\alpha i} - \lambda_i \lambda_j y_{\alpha i} y_{\alpha j}) \\
&= \frac{1}{\lambda_k - \Delta/2 - \lambda_i} \frac{1}{n} \sum_{\alpha=1}^{n} \Big\{ (1-\lambda_j^2)\lambda_i \lambda_j x_{\alpha i} x_{\alpha j} - \lambda_j^2 (y_{\alpha i} - \lambda_i x_{\alpha i})(y_{\alpha j} - \lambda_j x_{\alpha j}) \\
&\quad + (1-\lambda_j^2)\lambda_j (y_{\alpha i} - \lambda_i x_{\alpha i}) x_{\alpha j} + (1-\lambda_j^2)\lambda_i (y_{\alpha j} - \lambda_j x_{\alpha j}) x_{\alpha i} \Big\}.
\end{aligned}$$

Let $z_l = (y_l - \lambda_i x_l)/\sqrt{1-\lambda_i^2}, 1 \leq l \leq p_1$. Then

$$\begin{aligned}
[D^{-1}B^{(2)}]_{i-k^*,j} = \frac{1}{\lambda_k - \Delta/2 - \lambda_i} \frac{1}{n} \sum_{\alpha=1}^{n} \Big\{ & (1-\lambda_j^2)\lambda_i \lambda_j x_{\alpha i} x_{\alpha j} - \lambda_j^2 \sqrt{1-\lambda_i^2}\sqrt{1-\lambda_j^2} z_{\alpha i} z_{\alpha j} \\
& + (1-\lambda_j^2)\lambda_j \sqrt{1-\lambda_i^2} z_{\alpha i} x_{\alpha j} + (1-\lambda_j^2)\lambda_i \sqrt{1-\lambda_i^2} z_{\alpha j} x_{\alpha i} \Big\}.
\end{aligned}$$

In this way, $\{x_{\alpha i}, z_{\alpha i}, 1 \leq i \leq p_1, 1 \leq \alpha \leq n\}$ are mutually independent standard gaussian random variables. For any given pair of vectors $u \in \mathbb{R}^{p_1 - k^*}, v \in \mathbb{R}^k$,

$$\begin{aligned}
u^\top D^{-1} B^{(2)} v &= \frac{1}{n} \sum_{\alpha=1}^{n} \sum_{i=k^*+1}^{p_1} \sum_{j=1}^{k} \frac{u_{i-k} v_j}{\lambda_k - \Delta/2 - \lambda_i} \Big\{ (1-\lambda_j^2)\lambda_i \lambda_j x_{\alpha i} x_{\alpha j} \\
&\quad - \lambda_j^2 \sqrt{1-\lambda_i^2}\sqrt{1-\lambda_j^2} z_{\alpha i} z_{\alpha j} + (1-\lambda_j^2)\lambda_j \sqrt{1-\lambda_i^2} z_{\alpha i} x_{\alpha j} \\
&\quad + (1-\lambda_j^2)\lambda_i \sqrt{1-\lambda_i^2} z_{\alpha j} x_{\alpha i} \Big\} \\
&\doteq \frac{1}{n} \sum_{\alpha=1}^{n} w_\alpha,
\end{aligned}$$

where $w_1, \cdots, w_n$ are *i.i.d.* quadratic forms of the concatenated vector $(x^\top, z^\top)$ and the quadratic form can be represented by a matrix A. In order to apply Lemma 2.22, we first compute,

$$\begin{aligned}\|A\|_{\mathrm{F}}^2 = & \sum_{i=k^*+1}^{p_1}\sum_{j=1}^{k}\frac{u_{i-k}^2 v_j^2}{(\lambda_k-\Delta/2-\lambda_i)^2}\Big\{(1-\lambda_j^2)^2\lambda_i^2\lambda_j^2+\lambda_j^4(1-\lambda_i^2)(1-\lambda_j^2) \\ & +(1-\lambda_j^2)^2\lambda_j^2(1-\lambda_i^2)+(1-\lambda_j^2)^2\lambda_i^2(1-\lambda_i^2)\Big\} \\ = & \sum_{i=k^*+1}^{p_1}\sum_{j=1}^{k}\frac{u_{i-k}^2 v_j^2}{(\lambda_k-\Delta/2-\lambda_i)^2}\left(1-\lambda_j^2\right)\left(\lambda_i^2+\lambda_j^2-2\lambda_i^2\lambda_j^2\right).\end{aligned}$$

By definition,

$$\sum_{i=k^*+1}^{p_1}\sum_{j=1}^{k}u_{i-k}^2 v_j^2 = 1.$$

Then $\|A\|_2^2, \|A\|_{\mathrm{F}}^2$ can be upper bounded by

$$\begin{aligned}\|A\|_2^2 \le \|A\|_{\mathrm{F}}^2 & \le \max_{1\le j\le k<k^*+1\le i\le p_1}\frac{(1-\lambda_j^2)(\lambda_i^2+\lambda_j^2-2\lambda_i^2\lambda_j^2)}{(\lambda_k-\Delta/2-\lambda_i)^2} \\ & \le \max_{1\le j\le k<k^*+1\le i\le p_1}\frac{(1-\lambda_k^2)(\lambda_i^2(1-\lambda_j^2)+\lambda_j^2(1-\lambda_i^2))}{(\lambda_k-\Delta/2-\lambda_i)^2} \\ & \le (1-\lambda_k^2)\max_{1\le j\le k<k^*+1\le i\le p_1}\frac{2(1-\lambda_i^2)}{(\lambda_k-\Delta/2-\lambda_i)^2} \\ & \le (1-\lambda_k^2)\max_{1\le j\le k<k^*+1\le i\le p_1}\frac{2(1-\lambda_{k+1}^2)}{(\lambda_k-\Delta/2-\lambda_{k+1})^2} \\ & \le 8\frac{(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{\Delta^2} \doteq K^2,\end{aligned}$$

where the second last inequality is due to the fact that for $\lambda > \lambda_{k+1}$, $f(x) = \frac{1-x^2}{(\lambda-x)^2}$ is increasing in the interval $[0, \lambda_{k+1}]$. Therefore, Lemma 2.22 implies that

$$P\left\{|w_\alpha| \ge t\right\} \le 2exp\left\{-c_0\min\left(\frac{t^2}{K^2},\frac{t}{K}\right)\right\}. \tag{2.42}$$

Observe that $\forall t \ge 0, \min\left(1, 2exp\left\{-c_0\min\left(\frac{t^2}{K^2},\frac{t}{K}\right)\right\}\right) \le exp\left\{1-c_0\frac{t}{K}\right\}$, then

$$P\left\{|w_\alpha| \ge t\right\} \le 2exp\left\{-c_0\left(\frac{t}{K}-1\right)\right\}. \tag{2.43}$$

By Definition 5.13 in Vershynin (2010) , w_α is sub-exponential random variable with

$\|w_\alpha\|_{\psi_1} \leq c_1 K$ for some universal constant c_1. Let $\widetilde{\delta} = \max\{\delta, \delta^2\}$. Apply Bernstein inequality (Lemma 2.21) to w_α / K with $a_i = 1/\sqrt{n}$,

$$
\begin{aligned}
P\left\{ |\frac{1}{\sqrt{n}} \sum_{\alpha=1}^{n} w_\alpha / K| \geq \widetilde{\delta}/2 \right\} &\leq 2exp\left\{ -c_2 n \min\left(\frac{\widetilde{\delta}^2}{4c_1^2}, \frac{\widetilde{\delta}}{2c_1} \right) \right\} \\
&\leq 2exp\left\{ -\frac{c_2}{1+4c_1^2} n\delta^2 \right\} \\
&\leq 2exp\left\{ -\frac{c_2}{1+4c_1^2} (C^2 p_1 + t^2) \right\}.
\end{aligned}
$$

Step 3. Union Bound. By Lemma 2.23, we can choose 1/4-net such that

$$
\begin{aligned}
P\left\{ \max_{u_\epsilon, v_\epsilon} u_\epsilon^\top B^{(2)} v_\epsilon \geq K\widetilde{\delta}/2 \right\} &\leq 9^{p_1 - k^*} 9^k \times 2exp\left\{ -\frac{c_2}{1+4c_1^2} (C^2 p_1 + t^2) \right\} \\
&\leq 2exp\left\{ -\frac{c_2}{1+4c_1^2} t^2 \right\},
\end{aligned}
$$

where the second inequality follows if we choose $C \geq \sqrt{\frac{(1+4c_1^2)\log 9}{c_2}}$. We finish the proof by choosing $c = \frac{c_2}{1+4c_1^2}$. The expectation bound can be obtained using the formula

$$
E[X] = \int_0^{+\infty} P(X \geq t) dt
$$

where X is nonnegative random variable. The calculation is essentially the same as the proof of Lemma 2.20 and we leave out the details.

2.7.4. Proof of Lemma 2.9

The proof is essentially the same as the proof for Lemma 2.8 except that the term $\|A\|_{\mathrm{F}}^2$ will be different (but simpler) and is sketched as follows,

$$\begin{aligned}
\|A\|_{\mathrm{F}}^2 &= \sum_{i=k+1}^{k^*}\sum_{j=1}^{k} u_{i-k}^2 v_j^2 \Big\{(1-\lambda_j^2)^2\lambda_i^2\lambda_j^2 + \lambda_j^4(1-\lambda_i^2)(1-\lambda_j^2) \\
&\quad + (1-\lambda_j^2)^2\lambda_j^2(1-\lambda_i^2) + (1-\lambda_j^2)^2\lambda_i^2(1-\lambda_i^2)\Big\} \\
&= \sum_{i=k+1}^{k^*}\sum_{j=1}^{k} u_{i-k}^2 v_j^2 \left(1-\lambda_j^2\right)\left(\lambda_i^2+\lambda_j^2-2\lambda_i^2\lambda_j^2\right) \\
&\le \max_{1\le j\le k<i\le k^*}(1-\lambda_k^2)(\lambda_i^2(1-\lambda_j^2)+\lambda_j^2(1-\lambda_i^2)) \\
&\le 2(1-\lambda_k^2)\max_{1\le j\le k<i\le k^*}(1-\lambda_i^2).
\end{aligned}$$

Notice that by definition, for $k+1\le i\le k^*$, $\lambda_i \ge 2\lambda_k - 1 - \Delta$, then

$$\begin{aligned}
\|A\|_{\mathrm{F}}^2 &\le 2(1-\lambda_k^2)(1+\lambda_i)(2-2\lambda_k+\Delta) \\
&\le 2(1-\lambda_k^2)(1+\lambda_{k+1})2(1-\lambda_{k+1}) \\
&\le 4(1-\lambda_k^2)(1-\lambda_{k+1}^2).
\end{aligned}$$

The other parts of the argument proceed in the same way as in the proof of Lemma 2.8.

2.7.5. Proof of Lemma 2.10

In this section, we show how to control the higher order terms δ_6 and δ_7. The universal constants $C, C_1, c, \cdots$ might change from line to line. To facilitate presentation, we again introduce $z = (x^\top, y^\top)^\top$ as the concatenation of x and y. Then

$$\Sigma_z = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix} = \begin{bmatrix} I_{p_1} & \Sigma_{xy} \\ \Sigma_{yx} & I_{p_2} \end{bmatrix}, \quad \widehat{\Sigma}_z = \begin{bmatrix} \widehat{\Sigma}_x & \widehat{\Sigma}_{xy} \\ \widehat{\Sigma}_{yx} & \widehat{\Sigma}_y \end{bmatrix}.$$

Lemma 2.28. *There exists universal constant C such that the following inequalities hold* *deterministically*

$$\begin{aligned}
\|\widehat{\Sigma}_x\|, \|\widehat{\Sigma}_y\|, \|\widehat{\Sigma}_{xy}\| &\le \|\widehat{\Sigma}_z\|, \\
\|\Sigma_x - \widehat{\Sigma}_x\|, \|\Sigma_y - \widehat{\Sigma}_y\|, \|\Sigma_{xy} - \widehat{\Sigma}_{xy}\| &\le \|\Sigma_z - \widehat{\Sigma}_z\|, \\
\|\widehat{\Lambda} - \Lambda\| \le \|\widehat{\Sigma}_x^{-1/2}\widehat{\Sigma}_{xy}\widehat{\Sigma}_y^{-1/2} - \Sigma_{xy}\| &\le \|\Sigma_z - \widehat{\Sigma}_z\| \left(2 + \|\widehat{\Sigma}_z\|\right), \\
\|\widehat{\Phi}^l_{1:k}\|, \|\widehat{\Psi}^l_{1:k}\| &\le C\|\Sigma_z - \widehat{\Sigma}_z\| \left(\frac{2 + \|\widehat{\Sigma}_z\|}{\Delta} + \|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|\right),
\end{aligned}$$

where $\Delta = \lambda_k - \lambda_{k+1}$ *is the eigen-gap.*

Lemma 2.28 (proved in Section 2.7.8) and triangle inequality are frequently applied in this proof. Notice that $\Sigma_x^{ij}, \Sigma_y^{ij}, \Sigma_{xy}^{ij}, \Sigma_{yx}^{ij}, 1 \le i, j \le 2$ are sub-matrices of $\Sigma_x, \Sigma_y, \Sigma_{xy}, \Sigma_{yx}$. We will also repeatedly use the fact that the operator norm of a matrix is no less than that of its sub-matrices.

$$\begin{aligned}
\|\delta_1\| &\le \|\widehat{\Phi}^u_{1:k}(\widehat{\Lambda}_1 - \Lambda_1) + (\widehat{\Sigma}^{11}_x - I_k)\widehat{\Phi}^u_{1:k}\widehat{\Lambda}_1 - (\widehat{\Sigma}^{11}_{xy} - \Lambda_1)\widehat{\Psi}^u_{1:k} + \widehat{\Sigma}^{12}_x\widehat{\Phi}^l_{1:k}\widehat{\Lambda}_1 - \widehat{\Sigma}^{12}_{xy}\widehat{\Psi}^l_{1:k}\| \\
&\le \|\widehat{\Phi}_{1:k}\|\|\Sigma_z - \widehat{\Sigma}_z\|(2 + \|\widehat{\Sigma}_z\|) + \|\Sigma_z - \widehat{\Sigma}_z\|\|\widehat{\Phi}_{1:k}\| + \|\Sigma_z - \widehat{\Sigma}_z\|\|\widehat{\Psi}_{1:k}\| \\
&\qquad + 2C\|\Sigma_z - \widehat{\Sigma}_z\|^2 \left(\frac{2 + \|\widehat{\Sigma}_z\|}{\Delta} + \|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|\right) \\
&\le C_1\|\Sigma_z - \widehat{\Sigma}_z\|(2 + \|\widehat{\Sigma}_z\|)(\|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|) \\
&\quad + C_2\|\Sigma_z - \widehat{\Sigma}_z\|^2 \left(\frac{2 + \|\widehat{\Sigma}_z\|}{\Delta} + \|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|\right) \\
&\le C_3\|\Sigma_z - \widehat{\Sigma}_z\|(2 + \|\widehat{\Sigma}_z\|)(\|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \widehat{\Sigma}_z\|/\Delta).
\end{aligned}$$

where in the last inequality we use $\|\Sigma_z\| \le 2$ and

$$\|\Sigma_z - \widehat{\Sigma}_z\|^2 \left(\|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|\right) \le \|\Sigma_z - \widehat{\Sigma}_z\|(\|\Sigma_z\| + \|\widehat{\Sigma}_z\|) \left(\|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|\right)$$

By the same argument,

$$\|\delta_2\| \leq C_3 \|\Sigma_z - \widehat{\Sigma}_z\|(2 + \|\widehat{\Sigma}_z\|)(\|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \widehat{\Sigma}_z\|/\Delta).$$

We can also bound δ_3, δ_4 in the same manner and will obtain,

$$\|\delta_3\|, \|\delta_4\| \leq C\|\Sigma_z - \widehat{\Sigma}_z\|^2(2 + \|\widehat{\Sigma}_z\|)^2(\|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|)/\Delta$$

Recall that $\delta_6 = \widehat{\Sigma}_x^{21}\delta_1\Lambda_1 - \Lambda_2\widehat{\Sigma}_y^{21}\delta_1 + \delta_5$ and $\delta_5 = -\delta_3\Lambda_1 + \Lambda_2\delta_4$, then

$$\begin{aligned}
\|A_1 \circ \delta_6\| &\leq \|A_1 \circ (\widehat{\Sigma}_x^{21}\delta_1\Lambda_1 - \Lambda_2\widehat{\Sigma}_y^{21}\delta_1 + \delta_5)\| \\
&\leq \|A_1 \circ (\widehat{\Sigma}_x^{21}\delta_1\Lambda_1)\| + \|A_1 \circ (\Lambda_2\widehat{\Sigma}_y^{21}\delta_1)\| + \|A_1 \circ (\delta_3\Lambda_1)\| + \|A_1 \circ (\Lambda_2\delta_4)\| \\
&= \|(A_1\Lambda_1) \circ (\widehat{\Sigma}_x^{21}\delta_1)\| + \|(\Lambda_2 A_1) \circ (\widehat{\Sigma}_y^{21}\delta_1)\| + \|(A_1\Lambda_1) \circ \delta_3\| + \|(\Lambda_2 A_1) \circ \delta_4\|.
\end{aligned}$$

By the same argument as in (2.26),

$$\begin{aligned}
\|A_1 \circ \delta_6\| &\leq \frac{3}{2}\|\widehat{\Sigma}_x^{21}\delta_1\| + \frac{1}{2}\|\widehat{\Sigma}_y^{21}\delta_1\| + \frac{3}{2}\|\delta_3\| + \frac{1}{2}\|\delta_4\| \\
&\leq 2\|\Sigma_z - \widehat{\Sigma}_z\|\|\delta_1\| + \frac{3}{2}\|\delta_3\| + \frac{1}{2}\|\delta_4\| \\
&\leq C\|\Sigma_z - \widehat{\Sigma}_z\|^2(2 + \|\widehat{\Sigma}_z\|)(\|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \widehat{\Sigma}_z\|/\Delta) \\
&\qquad + C\|\Sigma_z - \widehat{\Sigma}_z\|^2(2 + \|\widehat{\Sigma}_z\|)^2(\|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|)/\Delta \\
&\leq C\|\Sigma_z - \widehat{\Sigma}_z\|^2(2 + \|\widehat{\Sigma}_z\|)^2(\|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \widehat{\Sigma}_z\|)/\Delta.
\end{aligned}$$

By the same argument as in (2.27),

$$\begin{aligned}
\|A_1 \circ A_2 \circ \delta_6\| &\leq \frac{1}{\Delta}\|A_1 \circ \delta_6\| \\
&\leq C\|\Sigma_z - \widehat{\Sigma}_z\|^2(2 + \|\widehat{\Sigma}_z\|)^2(\|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \widehat{\Sigma}_z\|)/\Delta^2.
\end{aligned}$$

Recall that $\delta_7 = \widehat{\Sigma}_{xy}^{21}\delta_1 + \widehat{\Sigma}_x^{21}(\Lambda_1\delta_1 + \delta_2\Lambda_1) + \Lambda_2\widehat{\Sigma}_y^{21}\delta_2 + \delta_5$, then

$$\begin{aligned}
\|\delta_7\| &\le \|\Sigma_z - \widehat{\Sigma}_z\|(2\|\delta_1\| + 2\|\delta_2\|) + \|\delta_3\| + \|\delta_4\| \\
&\le C\|\Sigma_z - \widehat{\Sigma}_z\|^2(2 + \|\widehat{\Sigma}_z\|)(\|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \widehat{\Sigma}_z\|/\Delta) \\
&\quad + C\|\Sigma_z - \widehat{\Sigma}_z\|^2(2 + \|\widehat{\Sigma}_z\|)^2(\|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|)/\Delta \\
&\le C\|\Sigma_z - \widehat{\Sigma}_z\|^2(2 + \|\widehat{\Sigma}_z\|)^2(\|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \widehat{\Sigma}_z\|)/\Delta.
\end{aligned}$$

Again, by the same argument in (2.27),

$$\|A_2 \circ \delta_7\| \le C\|\Sigma_z - \widehat{\Sigma}_z\|^2(2 + \|\widehat{\Sigma}_z\|)^2(\|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \widehat{\Sigma}_z\|)/\Delta^2.$$

2.7.6. Proof of Lemma 2.11

By definition, $\widehat{\Phi}^\top\widehat{\Sigma}_x\widehat{\Phi} = I_{p_1}$, then

$$\widehat{\Phi}^\top\widehat{\Phi} - I_{p_1} = -\widehat{\Phi}^\top(\widehat{\Sigma}_x - I_{p_1})\widehat{\Phi}.$$

Notice that $\widehat{\Sigma}_x^{1/2}\widehat{\Phi} \in \mathcal{O}(p_1)$,

$$\begin{aligned}
\|\widehat{\Phi}^\top\widehat{\Phi} - I_{p_1}\| \le \|\widehat{\Phi}^\top(\widehat{\Sigma}_x - I_{p_1})\widehat{\Phi}\| &\le \|\widehat{\Phi}^\top\widehat{\Sigma}_x^{1/2}\|\|\widehat{\Sigma}_x^{-1/2}(\widehat{\Sigma}_x - I_{p_1})\widehat{\Sigma}_x^{-1/2}\|\|\widehat{\Sigma}_x^{1/2}\widehat{\Phi}\| \\
&= \|\widehat{\Sigma}_x^{-1/2}(\widehat{\Sigma}_x - I_{p_1})\widehat{\Sigma}_x^{-1/2}\|.
\end{aligned}$$

As a submatrix,

$$\begin{aligned}
\|\widehat{\Phi}_{1:k}^\top\widehat{\Phi}_{1:k} - I_k\| &\le \|\widehat{\Sigma}_x^{-1/2}(\widehat{\Sigma}_x - I_{p_1})\widehat{\Sigma}_x^{-1/2}\| \\
&\le \|\widehat{\Sigma}_x^{-1}\|\|\widehat{\Sigma}_x - I_{p_1}\| \\
&\le \frac{1}{1 - \|\widehat{\Sigma}_x - I_{p_1}\|}\|\widehat{\Sigma}_x - I_{p_1}\| \\
&\le \frac{\|\widehat{\Sigma}_z - \Sigma_z\|}{1 - \|\widehat{\Sigma}_z - \Sigma_z\|},
\end{aligned}$$

which implies that

$$\sigma_k^2(\widehat{\Phi}_{1:k}) \geq 1 - \frac{\|\widehat{\Sigma}_z - \Sigma_z\|}{1 - \|\widehat{\Sigma}_z - \Sigma_z\|}, \ \|\widehat{\Phi}_{1:k}\|^2 \leq 1 + \frac{\|\widehat{\Sigma}_z - \Sigma_z\|}{1 - \|\widehat{\Sigma}_z - \Sigma_z\|}.$$

Notice that $\|\Sigma_z\| \leq 2$. By Lemma 2.19, for any given positive constant τ, there exists constants c, C depending on τ such that when $n \geq C(p_1 + p_2)$,

$$\|\widehat{\Sigma}_z - \Sigma_z\| \leq \tau$$

holds with probability at least $1 - e^{-cn}$. Choose τ small enough such that $\sigma_k^2(\widehat{\Phi}_{1:k}) \geq 1/2$ and $\|\widehat{\Phi}_{1:k}\|^2 \leq 3/2$. By the same argument,

$$\sigma_k^2(\widehat{\Psi}_{1:k}) \geq 1/2, \|\widehat{\Psi}_{1:k}\|^2 \leq 3/2$$

will hold as well and therefore

$$(2 + \|\widehat{\Sigma}_z\|)^2(\|\widehat{\Phi}_{1:k}\| + \|\widehat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \widehat{\Sigma}_z\|) \leq (2 + \|\Sigma_z\| + \tau)^2 \times 3 \times (1 + \tau) \leq 150$$

2.7.7. Proof of Lemma 2.13

We can write down explicitly, for $1 \leq j \leq k < i \leq p_1$

$$[B]_{i-k,j} = \frac{1}{n} \sum_{\alpha=1}^{n} \left(\lambda_j x_{\alpha i} y_{\alpha j} - \lambda_j^2 x_{\alpha i} x_{\alpha j} + \lambda_i x_{\alpha j} y_{\alpha i} - \lambda_i \lambda_j y_{\alpha i} y_{\alpha j} \right).$$

Notice that $(x_{\alpha i}, y_{\alpha i})$ are mutually uncorrelated pairs for any $1 \leq \alpha \leq n, 1 \leq i \leq p_1$. It is easy to compute

$$\mathbb{E}[B]_{i-k,j}^2 = (1 - \lambda_j^2)(\lambda_i^2 + \lambda_j^2 - 2\lambda_i^2 \lambda_j^2)/n,$$

and thus

$$\begin{aligned}
\mathbb{E}\|\widetilde{D}^{-1}B\|_{\mathrm{F}}^2 &= \frac{1}{n}\sum_{1\le j\le k<i\le p_1}(1-\lambda_j^2)\frac{\lambda_i^2+\lambda_j^2-2\lambda_i^2\lambda_j^2}{(\lambda_k-\lambda_i)^2}\\
&\le \frac{1-\lambda_k^2}{n}\sum_{1\le j\le k<i\le p_1}\frac{\lambda_i^2(1-\lambda_j^2)+\lambda_j^2(1-\lambda_i^2)}{(\lambda_k-\lambda_i)^2}\\
&\le \frac{2(1-\lambda_k^2)}{n}\sum_{1\le j\le k<i\le p_1}\frac{1-\lambda_i^2}{(\lambda_k-\lambda_i)^2}\\
&\le \frac{2(1-\lambda_k^2)}{n}\sum_{1\le j\le k<i\le p_1}\frac{1-\lambda_{k+1}^2}{(\lambda_k-\lambda_{k+1})^2}\\
&\le \frac{2(1-\lambda_k^2)(1-\lambda_{k+1}^2)}{n\Delta^2}(p_1-k)k,
\end{aligned}$$

where the second last inequality is due to the fact that for $\lambda > \lambda_{k+1}$, $f(x)=\frac{1-x^2}{(\lambda-x)^2}$ is increasing in the interval $[0,\lambda_{k+1}]$.

2.7.8. Proof of Lemma 2.28

The first two inequalities are trivial because the operator norm of a matrix is not less than that of its sub-matrices. Notice that $\widehat{\Lambda}$ and Λ are singular values of $\widehat{\Sigma}_x^{-1/2}\widehat{\Sigma}_{xy}\widehat{\Sigma}_y^{-1/2}$ and Σ_{xy} respectively. Hence by Weyl's inequality,

$$\|\widehat{\Lambda}-\Lambda\| \le \|\widehat{\Sigma}_x^{-1/2}\widehat{\Sigma}_{xy}\widehat{\Sigma}_y^{-1/2}-\Sigma_{xy}\|.$$

Further observe that

$$\begin{aligned}
\widehat{\Sigma}_x^{-1/2}\widehat{\Sigma}_{xy}\widehat{\Sigma}_y^{-1/2}-\Sigma_{xy} &= (I_{p_1}-\widehat{\Sigma}_x^{1/2})\widehat{\Sigma}_x^{-1/2}\widehat{\Sigma}_{xy}\widehat{\Sigma}_y^{-1/2}\\
&\quad+\widehat{\Sigma}_x^{1/2}\widehat{\Sigma}_x^{-1/2}\widehat{\Sigma}_{xy}\widehat{\Sigma}_y^{-1/2}(I_{p_2}-\widehat{\Sigma}_y^{1/2})+(\widehat{\Sigma}_{xy}-\Sigma_{xy}).
\end{aligned}$$

and $\|\widehat{\Sigma}_x^{-1/2}\widehat{\Sigma}_{xy}\widehat{\Sigma}_y^{-1/2}\| = \widehat{\lambda}_1 \le 1$. Then

$$\|\widehat{\Sigma}_x^{-1/2}\widehat{\Sigma}_{xy}\widehat{\Sigma}_y^{-1/2}-\Sigma_{xy}\| \le \|I_{p_1}-\widehat{\Sigma}_x^{1/2}\|+\|\widehat{\Sigma}_x\|\|I_{p_2}-\widehat{\Sigma}_y^{1/2}\|+\|\widehat{\Sigma}_{xy}-\Sigma_{xy}\|.$$

Also notice that

$$\|I - \widehat{\Sigma}_y\| = \|(I - \Sigma_y^{1/2})(I + \Sigma_y^{1/2})\| \geq \sigma_{\min}(I + \Sigma_y^{1/2})\|I - \Sigma_y^{1/2}\| \geq \|I - \Sigma_y^{1/2}\|.$$

Therefore,

$$\begin{aligned}\|\widehat{\Sigma}_x^{-1/2}\widehat{\Sigma}_{xy}\widehat{\Sigma}_y^{-1/2} - \Sigma_{xy}\| &\leq \|I_{p_1} - \widehat{\Sigma}_x\| + \|\widehat{\Sigma}_x\|\|I_{p_2} - \widehat{\Sigma}_y\| + \|\widehat{\Sigma}_{xy} - \Sigma_{xy}\| \\ &\leq \|\Sigma_z - \widehat{\Sigma}_z\| \left(2 + \|\widehat{\Sigma}_z\|\right).\end{aligned}$$

The last inequality in the lemma relies on the fact that $\widehat{\Sigma}_x^{1/2}\widehat{\Phi}_{1:k}$ and $I_{p_1,k}$ are leading k singular vectors of $\widehat{\Sigma}_x^{-1/2}\widehat{\Sigma}_{xy}\widehat{\Sigma}_y^{-1/2}$ and Σ_{xy} respectively. By a variant of Wedin's $\sin\theta$ law as stated in Lemma 2.24,

$$\left\|P_{\widehat{\Sigma}_x^{1/2}\widehat{\Phi}_{1:k}}(I_{p_1} - P_{I_{p_1,k}})\right\| \leq \frac{C\|\widehat{\Sigma}_x^{-1/2}\widehat{\Sigma}_{xy}\widehat{\Sigma}_y^{-1/2} - \Sigma_{xy}\|}{\Delta}.$$

On the other hand,

$$\begin{aligned}\left\|P_{\widehat{\Sigma}_x^{1/2}\widehat{\Phi}_{1:k}}(I_{p_1} - P_{I_{p_1,k}})\right\| &= \left\|\widehat{\Sigma}_x^{1/2}\widehat{\Phi}_{1:k}(\widehat{\Sigma}_x^{1/2}\widehat{\Phi}_{1:k})^\top(I_{p_1} - P_{I_{p_1,k}})\right\| \\ &= \left\|(\widehat{\Sigma}_x^{1/2}\widehat{\Phi}_{1:k})^\top(I_{p_1} - P_{I_{p_1,k}})\right\| \\ &= \left\|(\widehat{\Sigma}_x^{1/2}\widehat{\Phi}_{1:k})^l\right\|,\end{aligned}$$

where the second equality is due to the fact that $\widehat{\Sigma}_x^{1/2}\widehat{\Phi}_{1:k}$ has orthonormal columns and $(\widehat{\Sigma}_x^{1/2}\widehat{\Phi}_{1:k})^l$ denotes the lower $(p_1 - k) \times k$ sub-matrix of $\widehat{\Sigma}_x^{1/2}\widehat{\Phi}_{1:k}$. Again, by triangle inequality,

$$\begin{aligned}\left\|\widehat{\Phi}_{1:k}^l\right\| &= \left\|(\widehat{\Sigma}_x^{1/2}\widehat{\Phi}_{1:k})^l - \left((\widehat{\Sigma}_x^{1/2} - I_{p_1})\widehat{\Phi}_{1:k}\right)^l\right\| \\ &\leq \left\|(\widehat{\Sigma}_x^{1/2}\widehat{\Phi}_{1:k})^l\right\| + \left\|(\widehat{\Sigma}_x^{1/2} - I_{p_1})\widehat{\Phi}_{1:k}\right\| \\ &\leq \frac{C\|\widehat{\Sigma}_x^{-1/2}\widehat{\Sigma}_{xy}\widehat{\Sigma}_y^{-1/2} - \Sigma_{xy}\|}{\Delta} + \left\|\widehat{\Sigma}_z - \widehat{\Sigma}\right\|\left\|\widehat{\Phi}_{1:k}\right\|.\end{aligned}$$

The last inequality is obtained by substituting the upper bound for $\|\widehat{\Sigma}_x^{-1/2}\widehat{\Sigma}_{xy}\widehat{\Sigma}_y^{-1/2} - \Sigma_{xy}\|$ obtained above.

2.7.9. Connection with the Loss Function in Gao et al. (2014)

The proposed Frobenius norm loss function is upper bounded by the loss function of Gao et al. (2015b). Specifically, we are going to show

$$\mathcal{L}_2(\operatorname{span}(x^\top \Phi_{1:k}), \operatorname{span}(x^\top \widehat{\Phi}_{1:k})) = 2 \inf_{Q \in \mathbb{R}^{k\times k}} \mathbb{E} \left\| x^\top \Phi_{1:k} - x^\top \widehat{\Phi}_{1:k} Q \right\|^2, \tag{2.44}$$

and therefore,

$$\begin{aligned} \inf_{Q \in \mathcal{O}(k,k)} \mathbb{E} \left\| x^\top \Phi_{1:k} - x^\top \widehat{\Phi}_{1:k} Q \right\|^2 &\geq \inf_{Q \in \mathbb{R}^{k\times k}} \mathbb{E} \left\| x^\top \Phi_{1:k} - x^\top \widehat{\Phi}_{1:k} Q \right\|^2 \\ &\geq \frac{1}{2} \mathcal{L}_2(\operatorname{span}(x^\top \Phi_{1:k}), \operatorname{span}(x^\top \widehat{\Phi}_{1:k})). \end{aligned}$$

To prove (2.44), simple algebra yields

$$\begin{aligned} \mathbb{E} \left\| x^\top \Phi_{1:k} - x^\top \widehat{\Phi}_{1:k} Q \right\|_2^2 &= \operatorname{trace}\left((\Phi_{1:k} - \widehat{\Phi}_{1:k} Q)^\top \Sigma_x (\Phi_{1:k} - \widehat{\Phi}_{1:k} Q) \right) \\ &= \left\| {\Sigma_x}^{1/2} \Phi_{1:k} - {\Sigma_x}^{1/2} \widehat{\Phi}_{1:k} Q \right\|_{\mathrm{F}}^2 . \end{aligned}$$

Then the least squares solution $Q^* = (\widehat{\Phi}_{1:k}^\top \Sigma_x \widehat{\Phi}_{1:k})^{-1} \widehat{\Phi}_{1:k}^\top \Sigma_x \Phi_{1:k}$ achieves the minimum. Substitute Q^* into the objective will show that

$$\inf_{Q \in \mathbb{R}^{k\times k}} \left\| \Sigma_x^{1/2} \Phi_{1:k} - \Sigma_x^{1/2} \widehat{\Phi}_{1:k} Q \right\|_{\mathrm{F}}^2 = \left\| \left(I_{p_1} - P_{\Sigma_x^{1/2} \widehat{\Phi}_{1:k}} \right) \Sigma_x^{1/2} \Phi_{1:k} \right\|_{\mathrm{F}}^2$$

Notice that $\Sigma_x^{1/2} \Phi_{1:k} \in \mathcal{O}(p_1, k)$, then

$$\begin{aligned} \inf_{Q \in \mathbb{R}^{k\times k}} \left\| \Sigma_x^{1/2} \Phi_{1:k} - \Sigma_x^{1/2} \widehat{\Phi}_{1:k} Q \right\|_{\mathrm{F}}^2 &= \left\| \left(I_{p_1} - P_{\Sigma_x^{1/2} \widehat{\Phi}_{1:k}} \right) \Sigma_x^{1/2} \Phi_{1:k} \left(\Sigma_x^{1/2} \Phi_{1:k} \right)^\top \right\|_{\mathrm{F}}^2 \\ &= \left\| \left(I_{p_1} - P_{\Sigma_x^{1/2} \widehat{\Phi}_{1:k}} \right) P_{\Sigma_x^{1/2} \Phi_{1:k}} \right\|_{\mathrm{F}}^2 \\ &= \frac{1}{2} \left\| P_{\Sigma_x^{1/2} \Phi_{1:k}} - P_{\Sigma_x^{1/2} \widehat{\Phi}_{1:k}} \right\|_{\mathrm{F}}^2 . \end{aligned}$$

where the last equality is due to Lemma 2.18.

CHAPTER 3 : Canonical Correlation Analysis: Iterative Algorithm

3.1. Introduction

In modern machine learning applications, CCA has been successfully applied to massive multi-view datasets to extract low-dimensional feature representations of high-dimensional complex objects, like images (Rasiwasia et al., 2010), text (Dhillon et al., 2011, 2012) and speeches (Arora and Livescu, 2013). The scale of the datasets necessitates efficient estimation of leading-k canonical vectors. According to results obtained in previous chapter, the corresponding sample canonical vectors are minimax optimal under a certain sample size condition, which is usually satisfied in the interesting large-n-large-p regime ($n \gg p_1, p_2 \gg 1$) we consider in this chapter. It is well-known that sample canonical vectors have the following closed form solution.

Proposition 3.1. $\widehat{\Phi} = \widehat{\Sigma}_x^{-1/2}\widehat{U}$, $\widehat{\Psi} = \widehat{\Sigma}_y^{-1/2}\widehat{V}$, $\widehat{\Lambda} = \widehat{D}$ *where* $\widehat{\Sigma}_x^{-1/2}\widehat{\Sigma}_{xy}\widehat{\Sigma}_y^{-1/2} = \widehat{U}\widehat{D}\widehat{V}^\top$ *is the singular value decomposition. This also implies* $\widehat{\Sigma}_{xy} = \widehat{\Sigma}_x\widehat{\Phi}\widehat{\Lambda}\widehat{\Psi}^\top\widehat{\Sigma}_y$.

Here, we continue using the notation in Chapter 2. With $p = p_1 \wedge p_2$, recall that $\widehat{\Sigma}_x \in \mathbb{R}^{p_1\times p_1}, \widehat{\Sigma}_y \in \mathbb{R}^{p_2\times p_2}, \widehat{\Sigma}_{xy} \in \mathbb{R}^{p_1\times p_2}$ are the sample covariance matrices, $\widehat{\Phi} = (\widehat{\phi}_1, \cdots, \widehat{\phi}_p) \in \mathbb{R}^{p_1\times p}, \widehat{\Psi} = (\widehat{\psi}_1, \cdots .\widehat{\psi}_p) \in \mathbb{R}^{p_2\times p}$ are the sample canonical vectors, $\widehat{\Lambda} \in \mathbb{R}^{p\times p}$ is the sample canonical correlation matrix. This proposition reveals that the leading-k sample canonical vectors can be obtained by the following three step algorithm.

1. Whitening: $\widetilde{X} = X\widehat{\Sigma}_x^{-1/2}, \widetilde{Y} = Y\widehat{\Sigma}_y^{-1/2}, \widetilde{\Sigma}_{xy} = \frac{1}{n-1}\widetilde{X}^\top\widetilde{Y}$

2. k-truncated SVD: $\widetilde{\Sigma}_{xy} \approx \widehat{U}_k\widehat{D}_k\widehat{V}_k^\top$

3. $\widehat{\Phi}_{1:k} = \widehat{\Sigma}_x^{-1/2}\widehat{U}_k, \widehat{\Psi}_{1:k} = \widehat{\Sigma}_y^{-1/2}\widehat{V}_k$

This algorithm works well when the sample size and feature dimension is of moderate size but it will be very slow and numerically unstable for large-scale datasets which are ubiquitous in the age of 'Big Data'. The bottleneck of this algorithm is the whitening step, which involves:

- large matrix multiplication $X^\top X, Y^\top Y, \widetilde{X}^\top \widetilde{Y}$ to obtain $\widehat{\Sigma}_x, \widehat{\Sigma}_y, \widetilde{\Sigma}_{xy}$ with computational complexity $O(np_1^2 + np_2^2)$;

- large matrix decomposition to compute $\widehat{\Sigma}_x^{-1/2}$ and $\widehat{\Sigma}_y^{-1/2}$ with computational complexity $O(p_1^3 + p_2^3)$.

Remark 3.2. The whitening step dominates the k-truncated SVD step because the top k dimensional singular vectors can be efficiently computed by randomized SVD algorithms (see Halko et al. (2011) for a nice review).

Remark 3.3. Another classical algorithm (built-in function in Matlab) introduced by Bjorck and Golub (1973) whitens the data matrices in a different but equivalent way. It starts with QR decomposition, $X = Q_x R_x$ and $Y = Q_y R_y$ and then performs a SVD on $Q_x^\top Q_y$, which has the same computational complexity $O(np_1^2 + np_2^2)$.

Besides the heavy computational cost, extra $O(p_1^2 + p_2^2)$ space is necessary to store the matrices $\widehat{\Sigma}_x^{-1/2}$ and $\widehat{\Sigma}_y^{-1/2}$ (typically dense). In high-dimensional applications where the number of features is huge, this is another bottleneck since data retrieval could further slow down the algorithm. In distributed storage systems, operations involving $\widehat{\Sigma}_x^{-1/2}$ and $\widehat{\Sigma}_y^{-1/2}$ will incur heavy communication cost. Therefore, it is natural to ask: is there a scalable algorithm that avoids huge matrix decomposition and huge matrix multiplication? Is it memory efficient? Or even more ambitiously, is there an online algorithm that generates decent approximation given a fixed computational power (e.g. CPU time, number of operations)?

3.1.1. Related Work

The scalability of estimation/learning algorithms has become increasingly important in modern data processing tasks. Since matrix manipulation is the building block of many machine learning and statistical algorithms, lots of efforts have been devoted to developing fast randomized algorithms for large-scale matrix mulitplications and factorizations such as Sarlos (2006), Liberty et al. (2007), Woolfe et al. (2008), Halko et al. (2011), etc. For example, these techniques are successfully applied to approximately finding the leading-k singular value decomposition and other partial matrix decompositions. However, existing results do not directly solve CCA due to the whitening step where full matrix decomposition is unavoidable. Several authors have tried to devise a scalable CCA algorithm. Avron et al. (2013) proposed an efficient approach for CCA between two tall and thin matrices ($p_1, p_2 \ll n$) by exploiting recently developed tools, notably *Subsampled Randomized Hadamard Transform*, which only subsampled a small proportion of the n data points to approximate the matrix product. However, when the size of the features, p_1 and p_2, are large, the sampling scheme does not work. Later, Lu and Foster (2014) consider sparse design matrices and formulate CCA as iterative least squares, where in each iteration a fast regression algorithm that exploits sparsity is applied.

Another related line of research considers stochastic optimization algorithms for PCA such as Oja and Karhunen (1985), Arora et al. (2012), Mitliagkas et al. (2013), Balsubramani et al. (2013). Compared with batch algorithms, these stochastic versions empirically converge much faster with similar accuracy. Moreover, these stochastic algorithms can be applied to streaming setting where data comes sequentially (one pass or several pass) without being stored. As mentioned in Arora et al. (2012), stochastic optimization algorithm for CCA is more challenging and remains an open problem because of the whitening step.

3.1.2. Main Contribution

The main contribution of this paper is to directly tackle CCA as a non-convex optimization problem and propose a novel Augmented Approximate Gradient (*AppGrad*) scheme and its stochastic variant for finding the top k dimensional canonical subspace. Its advantages over state-of-art CCA algorithms are three fold. First, *AppGrad* only involves a large matrix multiplying a thin matrix of width k and small matrix decomposition of dimension $k \times k$. Therefore to some extent it is free from the two bottlenecks. It also benefits if X and Y are sparse while the classical algorithm still needs to invert the dense matrices $X^\top X$ and $Y^\top Y$. Second, *AppGrad* achieves optimal storage complexity $O(k(p_1 + p_2))$, the space necessary to store the output, compared with classical algorithms which usually require $O(p_1^2 + p_2^2)$ for storing $\widehat{\Sigma}_x^{-1/2}$ and $\widehat{\Sigma}_y^{-1/2}$. Third, the stochastic (online) variant of *AppGrad* is especially efficient for large scale datasets if moderate accuracy is desired. It is well-suited to the case when computational resources are limited or data comes as a stream. To the best of our knowledge, it is the first stochastic algorithm for CCA, which partly gives an affirmative answer to a question left open in Arora et al. (2012).

For simplicity, we first focus on the leading canonical pair $(\widehat{\phi}_1, \widehat{\psi}_1)$ to motivate the proposed algorithms. Results for general scenario can be obtained in the same manner and will be briefly discussed in the later part of this section.

3.2. Algorithm: Augmented Approximate Gradient Descent

Throughout the paper, we assume X and Y are of full rank. We use $\|\cdot\|$ for L_2 norm. $\forall\, u \in \mathbb{R}^{p_1}$, $v \in \mathbb{R}^{p_2}$, we define $\|u\|_x = (u^\top \widehat{\Sigma}_x u)^{\frac{1}{2}}$ and $\|v\|_y = (v^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{y}} v)^{\frac{1}{2}}$, which are norms induced by X and Y.

3.2.1. Background

To begin, we recast sample CCA as the solution to a non-convex optimization problem (Golub and Zha, 1995).

Lemma 3.4. $(\widehat{\phi}_1, \widehat{\psi}_1)$ *is the solution to:*

$$
\begin{aligned}
&\min \frac{1}{2n}\|X\phi - Y\psi\|^2 \\
&\textit{subject to} \quad \phi^\top \widehat{\Sigma}_x \phi = 1,\ \psi^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{y}} \psi = 1
\end{aligned}
\tag{3.1}
$$

Although (3.1) is non-convex (objective function is convex but the constraint set is non-convex), Golub and Zha (1995) showed that an alternating minimization strategy (Algorithm 1), or rather iterative least squares, converges to the leading canonical pair. However, each update $\phi^{t+1} = \widehat{\Sigma}_x^{-1}\widehat{\Sigma}_{xy}\psi^t$ is computationally intensive. Essentially, the

Algorithm 2 CCA via Naive Gradient Descent

```
  **Input:** Data matrix $X \in \mathbb{R}^{n\times p_1}, Y \in \mathbb{R}^{n\times p_2}$, initialization $(\phi^0, \psi^0)$, step size $\eta_1, \eta_2$
  **Output :** NAN (incorrect algorithm)
  **repeat**
    $\phi^{t+1} = \phi^t - \eta_1 X^\top(X\phi^t - Y\psi^t)/n$
    $\phi^{t+1} = \phi^{t+1}/\|\phi^{t+1}\|_x$
    $\psi^{t+1} = \psi^t - \eta_2 Y^\top(Y\psi^t - X\phi^t)/n$
    $\psi^{t+1} = \psi^{t+1}/\|\psi^{t+1}\|_y$
  **until** convergence
```

alternating least squares algorithm acts like a second order method, which is usually recognized to be inefficient for large-scale datasets, especially when the current estimate is not close enough to the optimum. Therefore, it is natural to ask: is there a valid first order method that solves (3.1)?

Algorithm 1 CCA via Alternating Least Squares

```
  **Input:** Data matrix $X \in \mathbb{R}^{n\times p_1}, Y \in \mathbb{R}^{n\times p_2}$ and initialization $(\phi^0, \psi^0)$
  **Output:** $(\phi_{\text{ALS}}, \psi_{\text{ALS}})$
  **repeat**
    $\phi^{t+1} = \arg\min_{\phi} \frac{1}{2n}\|X\phi - Y\psi^t\|^2 = \widehat{\Sigma}_x^{-1}\widehat{\Sigma}_{xy}\psi^t$
    $\phi^{t+1} = \phi^{t+1}/\|\phi^{t+1}\|_x$
    $\psi^{t+1} = \arg\min_{\psi} \frac{1}{2n}\|Y\psi - X\phi^t\|^2 = \widehat{\Sigma}_y^{-1}\widehat{\Sigma}_{yx}\phi^t$
    $\psi^{t+1} = \psi^{t+1}/\|\psi^{t+1}\|_y$
  **until** convergence
```

Heuristics borrowed from the convex optimization literature give rise to the projected gradient scheme summarized in Algorithm 2. Instead of completely solving a least squares problem in each iteration, a single gradient step of (3.1) is performed and then the estimates are projected back to the constrained domain, which avoids inverting a huge matrix. Unfortunately, the following proposition shows that Algorithm 2 fails to converge to the leading canonical pair.

Proposition 3.5. *If leading canonical correlation $\widehat{\lambda}_1 \neq 1$ and either $\widehat{\phi}_1$ is not an eigenvector*

Algorithm 3 CCA via AppGrad

```
  **Input:** Data matrix $X \in \mathbb{R}^{n\times p_1}, Y \in \mathbb{R}^{n\times p_2}$, initialization $(\phi^0, \psi^0, \widetilde{\phi}^0, \widetilde{\psi}^0)$, step size $\eta_1, \eta_2$
  **Output:** $(\phi_{\mathrm{AG}}, \psi_{\mathrm{AG}}, \widetilde{\phi}_{\mathrm{AG}}, \widetilde{\psi}_{\mathrm{AG}})$
  **repeat**
    $\widetilde{\phi}^{t+1} = \widetilde{\phi}^t - \eta_1 X^\top (X\widetilde{\phi}^t - Y\psi^t)/n$
    $\phi^{t+1} = \widetilde{\phi}^{t+1} / \|\widetilde{\phi}^{t+1}\|_x$
    $\widetilde{\psi}^{t+1} = \widetilde{\psi}^t - \eta_2 Y^\top (Y\widetilde{\psi}^t - X\phi^t)/n$
    $\phi^{t+1} = \widetilde{\psi}^{t+1} / \|\widetilde{\psi}^{t+1}\|_y$
  **until** convergence
```

of $\widehat{\Sigma}_x$ or $\widehat{\psi}_1$ is not an eigenvector of $\widehat{\Sigma}_y$, then $\forall \eta_1, \eta_2 > 0$, the leading canonical pair $(\widehat{\phi}_1, \widehat{\psi}_1)$ is not a fixed point of the naive gradient scheme in Algorithm 2. Therefore, the algorithm does not converge to $(\widehat{\phi}_1, \widehat{\psi}_1)$.

Proof of Proposition 3.5. The proof is similar to the proof of Proposition 3.6 and we leave out the details here. □

The failure of Algorithm 2 is due to the non-convex nature of (3.1). Although every gradient step might decrease the objective function, this property no longer persists after projected to the non-convex domain $\{(\phi, \psi) \mid \phi^\top \widehat{\Sigma}_x \phi = 1,\ \psi^\top \widehat{\Sigma}_y \psi = 1\}$ (the normalization step). On the contrary, decreases triggered by gradient descent are maintained if the estimates are projected to a convex region.

3.2.2. AppGrad Scheme for Leading Canonical Pair

As a remedy, we propose the novel Augmented Approximate Gradient (*AppGrad*) descent scheme summarized in Algorithm 3. It inherits the convergence guarantee of alternating least squares as well as the scalability and memory efficiency of first order methods, which only involves matrix-vector multiplication and only requires $O(p_1 + p_2)$ extra space.

AppGrad seems unnatural at first sight but has nice intuitions behind as we will discuss later. The differences and similarities between these algorithms are subtle but crucial. Compared with the naive gradient descent, we introduce two auxiliary variables $(\widetilde{\phi}^t, \widetilde{\psi}^t)$, an unnormalized version of (ϕ^t, ψ^t). During each iterate, we update $\widetilde{\phi}^t$ and $\widetilde{\psi}^t$ without scaling them to have unit norm, which in turn produces the 'correct' normalized counterpart, (ϕ^t, ψ^t). It turns out that $(\widehat{\phi}_1, \widehat{\psi}_1, \widehat{\lambda}_1\widehat{\phi}_1, \widehat{\lambda}_1\widehat{\psi}_1)$ is a fixed point of the dynamic system $\{(\phi^t, \psi^t, \widetilde{\phi}^t, \widetilde{\psi}^t)\}_{t=0}^{\infty}$.

Proposition 3.6. *$\forall\, i \leq p_1$, let $\widetilde{\phi}_i = \widehat{\lambda}_i\widehat{\phi}_i, \widetilde{\psi}_i = \widehat{\lambda}_i\widehat{\psi}_i$, then $(\widehat{\phi}_i, \widehat{\psi}_i, \widetilde{\phi}_i, \widetilde{\psi}_i)$ are the fixed points of AppGrad scheme.*

To prove the proposition, we need the following lemma that characterizes the relations among some key quantities.

Proof of Proposition 3.6. Substitute $(\phi^t, \psi^t, \widetilde{\phi}^t, \widetilde{\psi}^t) = (\widehat{\phi}_i, \widehat{\psi}_i, \widetilde{\phi}_i, \widetilde{\psi}_i)$ into the iterative formula in Algorithm 3.

$$\begin{aligned}
\widetilde{\phi}^{t+1} &= \widetilde{\phi}_i - \eta_1(\widehat{\Sigma}_x\widetilde{\phi}_i - \widehat{\Sigma}_{xy}\widehat{\psi}_i) \\
&= \widetilde{\phi}_i - \eta_1(\widehat{\Sigma}_x\widetilde{\phi}_i - \widehat{\Sigma}_x\widehat{\Phi}\widehat{\Lambda}\widehat{\Psi}^\top\widehat{\Sigma}_y\widehat{\psi}_i) \\
&= \widetilde{\phi}_i - \eta_1(\widehat{\Sigma}_x\widetilde{\phi}_i - \widehat{\lambda}_i\widehat{\Sigma}_x\widehat{\phi}_i) \\
&= \widetilde{\phi}_i
\end{aligned}$$

The second equality is direct application of Proposition 3.1. The third equality is due to the fact that $\widehat{\Psi}^\top\widehat{\Sigma}_y\widehat{\Psi} = I_{p_2}$. Then,

$$\phi^{t+1} = \widetilde{\phi}_i / \|\widetilde{\phi}_i\|_x = \widetilde{\phi}_i / \widehat{\lambda}_i = \widehat{\phi}_i$$

Therefore $(\widetilde{\phi}^{t+1}, \phi^{t+1}) = (\widetilde{\phi}^t, \phi^t) = (\widetilde{\phi}_i, \widehat{\phi}_i)$. A symmetric argument will show that $(\widetilde{\psi}^{t+1}, \psi^{t+1}) = (\widetilde{\psi}^t, \psi^t) = (\widetilde{\psi}_i, \widehat{\psi}_i)$, which completes the proof. □

The connection between *AppGrad* and the alternating minimization strategy is subtle. Intuitively, when (ϕ^t, ψ^t) is not close to $(\widehat{\phi}_1, \widehat{\psi}_1)$, solving the least squares completely as carried out in Algorithm 1 is a waste of computational power (informally by regarding it as a second order method, the Newton Step has fast convergence only when the current estimate is close to the optimum). Instead of solving a sequence of possibly irrelevant least squares problems, the following lemma shows that *AppGrad* directly targets the least squares problem that involves the leading canonical pair.

Lemma 3.7. *Let $(\widehat{\phi}_1, \widehat{\psi}_1)$ be the leading canonical pair and $(\widetilde{\phi}_1, \widetilde{\psi}_1) = \widehat{\lambda}_1(\widehat{\phi}_1, \widehat{\psi}_1)$. Then,*

$$\begin{aligned} \widetilde{\phi}_1 &= \arg\min_{\phi} \frac{1}{2n} \|X\phi - Y\widehat{\psi}_1\|^2 \\ \widetilde{\psi}_1 &= \arg\min_{\psi} \frac{1}{2n} \|Y\psi - X\widehat{\phi}_1\|^2 \end{aligned} \tag{3.2}$$

Proof of Lemma 3.7. Let $\phi^* = \arg\min_{\phi} \frac{1}{2n}\|X\phi - Y\widehat{\psi}_1\|^2$. By the optimality condition, $\widehat{\Sigma}_x \phi^* = \widehat{\Sigma}_{xy}\widehat{\psi}_1$. Apply Proposition 3.1,

$$\phi^* = \widehat{\Sigma}_x^{-1}\widehat{\Sigma}_x\widehat{\Phi}\widehat{\Lambda}\widehat{\Psi}^\top\widehat{\Sigma}_y\widehat{\psi}_1 = \widehat{\lambda}_1\widehat{\phi}_1 = \widetilde{\phi}_1$$

A similar argument gives $\psi^* = \widetilde{\psi}_1$ □

Lemma 3.7 characterizes the relationship between the leading canonical pair $(\widehat{\phi}_1, \widehat{\psi}_1)$ and its unnormalized counterpart $(\widetilde{\phi}_1, \widetilde{\psi}_1)$, which sheds some insight on how *AppGrad* works. The intuition is that (ϕ^t, ψ^t) and $(\widetilde{\phi}^t, \widetilde{\psi}^t)$ are current estimates of $(\widehat{\phi}_1, \widehat{\psi}_1)$ and $(\widetilde{\phi}_1, \widetilde{\psi}_1)$, and the updates of $(\widetilde{\phi}^{t+1}, \widetilde{\psi}^{t+1})$ in Algorithm 3 are actually gradient steps of the least squares in (3.2), with the unknown truth $(\widehat{\phi}_1, \widehat{\psi}_1)$ approximated by (ϕ^t, ψ^t). In terms of mathematics,

$$\begin{aligned} \widetilde{\phi}^{t+1} &= \widetilde{\phi}^t - \eta_1 X^\top(X\widetilde{\phi}^t - Y\psi^t)/n \\ &\approx \widetilde{\phi}^t - \eta_1 X^\top(X\widetilde{\phi}^t - Y\widehat{\psi}_1)/n \\ &= \widetilde{\phi}^t - \eta_1 \nabla_\phi \frac{1}{2n}\|X\phi - Y\widehat{\psi}_1\|^2|_{\phi=\widetilde{\phi}^t} \end{aligned} \tag{3.3}$$

The normalization step in Algorithm 3 corresponds to generating new approximations of $(\widehat{\phi}_1, \widehat{\psi}_1)$, namely (ϕ^{t+1}, ψ^{t+1}), using the updated $(\widetilde{\phi}^{t+1}, \widetilde{\psi}^{t+1})$ through the relationship $(\widehat{\phi}_1, \widehat{\psi}_1) = (\widetilde{\phi}_1/\|\widetilde{\phi}_1\|_x, \widetilde{\psi}_1/\|\widetilde{\psi}_1\|_y)$. Therefore, one can interpret *AppGrad* as an approximate gradient scheme for solving (3.2). When $(\widetilde{\phi}^t, \widetilde{\psi}^t)$ converge to $(\widetilde{\phi}_1, \widetilde{\psi}_1)$, its scaled version

Algorithm 4 CCA via *AppGrad* (Rank-k)

```
Input: Data matrix $X \in \mathbb{R}^{n\times p_1}, Y \in \mathbb{R}^{n\times p_2}$, initialization $(\widehat{\Phi}^0, \widehat{\Psi}^0, \widetilde{\Phi}^0, \widetilde{\Psi}^0)$, step size $\eta_1, \eta_2$
Output : $(\widehat{\Phi}_{\mathrm{AG}}, \widehat{\Psi}_{\mathrm{AG}}, \widetilde{\Phi}_{\mathrm{AG}}, \widetilde{\Psi}_{\mathrm{AG}})$
repeat
  $\widetilde{\Phi}^{t+1} = \widetilde{\Phi}^t - \eta_1 X^\top (X\widetilde{\Phi}^t - Y\Psi^t)/n$
  SVD: $(\widetilde{\Phi}^{t+1})^\top \widehat{\Sigma}_x \widetilde{\Phi}^{t+1} = U_x D_x U_x^\top$
  $\Phi^{t+1} = \widetilde{\Phi}^{t+1} U_x D_x^{-\frac{1}{2}} U_x^\top$
  $\widetilde{\Psi}^{t+1} = \widetilde{\Psi}^t - \eta_2 Y^\top (Y\widetilde{\Psi}^t - X\Phi^t)/n$
  SVD: $(\widetilde{\Psi}^{t+1})^\top \widehat{\Sigma}_y \widetilde{\Psi}^{t+1} = U_y D_y U_y^\top$
  $\Psi^{t+1} = \widetilde{\Psi}^{t+1} U_y D_y^{-\frac{1}{2}} U_y^\top$
until convergence
```

(ϕ^t, ψ^t) converges to the leading canonical pair $(\widehat{\phi}_1, \widehat{\psi}_1)$.

The following theorem shows that when the estimates enter a neighborhood of the true canonical pair, *AppGrad* is contractive. Define the error metric $e_t = \|\Delta\widetilde{\phi}^t\|^2 + \|\Delta\widetilde{\psi}^t\|^2$ where $\Delta\widetilde{\phi}^t = \widetilde{\phi}^t - \widetilde{\phi}_1, \Delta\widetilde{\psi}^t = \widetilde{\psi}^t - \widetilde{\psi}_1$.

Theorem 4. *Assume $\widehat{\lambda}_1 > \widehat{\lambda}_2$, and $\lambda_{max}(\widehat{\Sigma}_x), \lambda_{max}(\widehat{\Sigma}_y) \leq L_1, \lambda_{min}(\widehat{\Sigma}_x), \lambda_{min}(\widehat{\Sigma}_y) \geq L_2^{-1}$ for positive constants L_1, L_2, where $\lambda_{min}(\cdot), \lambda_{max}(\cdot)$ denote smallest and largest eigenvalues. If $e_0 < 2(\widehat{\lambda}_1^2 - \widehat{\lambda}_2^2)/L_1$ and set $\eta_1 = \eta_2 = \eta = \delta/6L_1$, AppGrad achieves linear convergence in the sense that $\forall\, t \in \mathbb{N}_+$*

$$e_t \leq \left(1 - \frac{\delta^2}{6L_1L_2}\right)^t e_0$$

where $\delta = 1 - \left(1 - \frac{2(\widehat{\lambda}_1^2 - \lambda_2^2) - L_1 e_0}{2\widehat{\lambda}_1^2}\right)^{\frac{1}{2}} > 0$

Remark 3.8. The theorem reveals that the larger the eigen-gap $\widehat{\lambda}_1 - \widehat{\lambda}_2$, the broader is the basin of attraction. We didn't try to optimize the conditions above and empirically as shown in the experiments, a randomized initialization always suffices to capture most of the correlation.

3.2.3. AppGrad for General Rank-k Case

Following the spirit of rank-one case, *AppGrad* can be easily generalized to compute the top k dimesional canonical subspace as summarized in Algorithm 4. The only difference is that the original scalar normalization is replaced by its matrix counterpart, that is to multiply the inverse of the square root matrix $\Phi^{t+1} = \widetilde{\Phi}^{t+1} U_x D_x^{-\frac{1}{2}} U_x^\top$, ensuring that $(\Phi^{t+1})^\top X^\top X \Phi^{t+1} = I_k$.

Notice that the gradient step only involves a large matrix multiplying a thin matrix of width k and the SVD is performed on a small $k \times k$ matrix. Therefore, the computational complexity per iteration is dominated by the gradient step, of order $O(n(p_1 + p_2)k)$. The cost will be further reduced when the data matrices X, Y are sparse.

Compared with classical spectral algorithm which first whitens the data matrices and then performs a SVD on the whitened covariance matrix, *AppGrad* merges these two steps together. This is the key of its efficiency. At a high level, whitening the whole data matrix is not necessary and we only want to whiten the directions that contain the leading CCA subspace. However, these directions are unknown and therefore for two-step procedures, whitening the whole data matrix is unavoidable. Instead, *AppGrad* tries to identify (gradient step) and whiten (normalization step) these directions simultaneously. In this way, every normalization step is only performed on the potential k dimensional target CCA subspace and therefore only deals with a small $k \times k$ matrix.

Parallel results of Lemma 3.4, Proposition 3.5, Proposition 3.6, Lemma 3.7 for this general scenario can be established in a similar manner. Here, to make Algorithm 4 more clear, we state the fixed point result of which the proof is similar to Proposition 3.6.

Proposition 3.9. *Let $\widehat{\Lambda}_k = diag(\widehat{\lambda}_1, \cdots, \widehat{\lambda}_k)$ be the diagonal matrix of top k canonical correlations and let $\widehat{\Phi}_k = (\widehat{\phi}_1, \cdots, \widehat{\phi}_k), \widehat{\Psi}_k = (\widehat{\phi}_1, \cdots, \widehat{\phi}_k)$ be the top k CCA vectors. Also denote $\widetilde{\Phi}_k = \widehat{\Phi}_k \widehat{\Lambda}_k$ and $\widetilde{\Psi}_k = \widehat{\Psi}_k \widehat{\Lambda}_k$. Then for any $k \times k$ orthogonal matrix Q, $(\widehat{\Phi}_k, \widehat{\Psi}_k, \widetilde{\Phi}_k, \widetilde{\Psi}_k)Q$ is a fixed point of AppGrad scheme.*

The top k dimensional canonical subspace is identifiable up to a rotation matrix and

Algorithm 5 CCA via Stochastic *AppGrad* (Rank-k)

```
Input: Data matrix $X \in \mathbb{R}^{n\times p_1}, Y \in \mathbb{R}^{n\times p_2}$, initialization $(\widehat{\Phi}^0, \widehat{\Psi}^0, \widetilde{\Phi}^0, \widetilde{\Psi}^0)$, step size $\eta_{1t}, \eta_{2t}$, minibatch size $m$
Output : $(\widehat{\Phi}_{\mathrm{SAG}}, \widehat{\Psi}_{\mathrm{SAG}}, \widetilde{\Phi}_{\mathrm{SAG}}, \widetilde{\Psi}_{\mathrm{SAG}})$
repeat
    Randomly pick a subset $\mathcal{I} \subset \{1, 2, \cdots, n\}$ of size $m$
    $\widetilde{\Phi}^{t+1} = \widetilde{\Phi}^t - \eta_{1t} X_{\mathcal{I}}^\top (X_{\mathcal{I}}\widetilde{\Phi}^t - Y_{\mathcal{I}}\Psi^t)/m$
    SVD: $(\widetilde{\Phi}^{t+1})^\top (\frac{1}{m} X_{\mathcal{I}}^\top X_{\mathcal{I}})\widetilde{\Phi}^{t+1} = U_x^\top D_x U_x$
    $\Phi^{t+1} = \widetilde{\Phi}^{t+1} U_x^\top D_x^{-\frac{1}{2}} U_x$
    $\widetilde{\Psi}^{t+1} = \widetilde{\Psi}^t - \eta_{2t} Y_{\mathcal{I}}^\top (Y_{\mathcal{I}}\widetilde{\Psi}^t - X_{\mathcal{I}}\Phi^t)/m$
    SVD: $(\widetilde{\Psi}^{t+1})^\top (\frac{1}{m} Y_{\mathcal{I}}^\top Y_{\mathcal{I}})\widetilde{\Psi}^{t+1} = U_y^\top D_y U_y$
    $\Psi^{t+1} = \widetilde{\Psi}^{t+1} U_y^\top D_y^{-\frac{1}{2}} U_y$
until convergence
```

Proposition 3.9 shows that every optimum is a fixed point of *AppGrad* scheme.

3.2.4. Stochastic AppGrad

Recently, there is a growing interest in stochastic optimization which is shown to have better performance for large-scale learning problems Bousquet and Bottou (2008); Bottou (2010). Especially in the so-called 'data laden regime', where data is abundant and the bottleneck is runtime, stochastic optimization dominates batch algorithms both empirically and theoretically. Given these advantages, lots of efforts have been spent developing stochastic algorithms for principal component analysis Oja and Karhunen (1985); Arora et al. (2012); Mitliagkas et al. (2013); Balsubramani et al. (2013). Despite promising progress in PCA, as mentioned in Arora et al. (2012), stochastic CCA is more challenging and remains an open problem due to the whitening step.

As a gradient scheme, *AppGrad* naturally generalizes to the stochastic regime and we summarize this as Algorithm 5. Compared with the batch version, only a small subset of samples are used to compute the gradient, which reduces the computational cost per iteration from $O(n(p_1 + p_2)k)$ to $O(m(p_1 + p_2)k)$ ($m = |\mathcal{I}|$ is the size of the minibatch).

Table 1: Brief Summary of Datasets

Datasets	Description	p_1	p_2	n
Mediamill	image and its labels	100	120	$3 * 10^4$
Mnist	left and right halves of images	392	392	$6 * 10^4$
Penn Treebank	word co-ocurrance	10^4	10^4	$5 * 10^5$
URL	host and lexical features	10^5	10^5	10^6

Empirically, this makes stochastic *AppGrad* much faster than the batch version as we will see in the experiments. Also, for large scale applications when fully calculating the CCA subspace is prohibitive, stochastic *AppGrad* can generate a decent approximation given a fixed computational effort, while other algorithms only give a one-shot estimate after the whole procedure is carried out completely. Moreover, when there is a generative model, as shown in Bousquet and Bottou (2008), due to the tradeoff between statistical and numerical accuracy, fully solving an empirical risk minimization is unnecessary since the statistical error will finally dominate. On the contrary, stochastic optimization directly tackles the problem in the population level and therefore is more statistically efficient.

It is worth mentioning that the normalization step is accomplished using a sampled Gram matrix $\frac{1}{m}X_{\mathcal{I}}^\top X_{\mathcal{I}}$ and $\frac{1}{m}Y_{\mathcal{I}}^\top Y_{\mathcal{I}}$. A key observation is that when $m \in O(k)$, $(\widetilde{\Phi}^{t+1})^\top(\frac{1}{m}X_{\mathcal{I}}^\top X_{\mathcal{I}})\widetilde{\Phi}^{t+1} \approx (\widetilde{\Phi}^{t+1})^\top(\frac{1}{m}X^\top X)\widetilde{\Phi}^{t+1}$ using a standard concentration inequality, because the matrix we want to approximate $(\widetilde{\Phi}^{t+1})^\top(\frac{1}{m}X^\top X)\widetilde{\Phi}^{t+1}$ is a $k \times k$ matrix, while generally $O(p)$ sample is needed to have $\frac{1}{m}X_{\mathcal{I}}^\top X_{\mathcal{I}} \approx \frac{1}{n}X^\top X$. As we have argued in the previous section, this bonus is a byproduct of the fact that *AppGrad* tries to identify and whiten the directions that contains the CCA subspace simultaneously, or else $O(p)$ samples are necessary for whitening the whole data matrices.

3.3. Experiments

In this section, we present experiments on four real datasets to evaluate the effectiveness of the proposed algorithms for computing the top 20 (k=20) dimensional canonical subspace. A short summary of the datasets is in Table 1.

Mediamill is an annotated video dataset from the Mediamill Challenge Snoek et al. (2006). Each image is a representative frame of a video shot annotated with 101 labels and consists of 120 features. CCA is performed to explore the correlation structure between the images and its labels.

MNIST is a database of handwritten digits. CCA is used to learn correlated representations between the left and right halves of the images.

Penn Tree Bank dataset is extracted from Wall Street Journal, which consists of 1.17 million tokens and a vocabulary size of $43,000$ Lamar et al. (2010). CCA has been successfully used on this dataset to build low dimensional word embeddings Dhillon et al. (2011, 2012). The task here is a CCA between words and their context. We only consider the 10, 000 most frequent words to avoid sample sparsity.

URL Reputation dataset Ma et al. (2009) is extracted from UCI machine learning repository. The dataset contains 2.4 million URLs each represented by 3.2 million features. For simplicity we only use the first 2 million samples. 38% of the features are host based features like WHOIS info, IP prefix and 62% are lexical based features like host name and primary domain. We run a CCA between a subset of host based features and a subset of lexical based features.

3.3.1. Implementations

Evaluation Criterion: The evaluation criterion we use for the first three datasets (Mediamill, MNIST, Penn Tree Bank) is *Proportion of Correlations Captured* (PCC). To introduce this term, we first introduce the concept of *Total Correlations Captured* (TCC) between two data matrices. Suppose $A, B \in \mathbb{R}^{n\times k}$. Consider the sample canonical correlation analysis between A and B by treating the rows of A, B as pairwise observations, and denote $\lambda_1(A, B), \cdots, \lambda_k(A, B)$ as the sample canonical correlations. Then we define

Total Correlations Captured by

$$\text{TCC} := \lambda_1(A, B) + \cdots + \lambda_k(A, B)$$

Finally, we can define the *Proportion of Correlations Captured* by the estimated top-k dimensional canonical subspace against the true top-k dimensional sample canonical subspace by

$$\text{PCC} = \frac{\text{TCC}(X\overline{\Phi}_k, Y\overline{\Psi}_k)}{\text{TCC}(X\widehat{\Phi}_k, Y\widehat{\Psi}_k)}$$

Intuitively PCC characterizes the proportion of correlations captured by certain algorithm compared with the true sample CCA subspace. Therefore, the higher is PCC the better is the estimated CCA subspace.

However, for URL Reputation dataset, the number of samples and features are too large for the algorithm to compute the true sample CCA subspace in a reasonable amount of time and instead we only compare the numerator $\text{TCC}(X\widetilde{\Phi}_k, Y\widetilde{\Psi}_k)$ (monotone w.r.t. PCC) for different algorithms.

Initialization We initialize (Φ^0, Ψ^0) by first drawing *i.i.d.* samples from the standard Gaussian distribution and then normalize such that $(\Phi^0)^\top \widehat{\Sigma}_x \Phi^0 = I_k$ and $(\Psi^0)^\top \widehat{\Sigma}_y \Psi^0 = I_k$

Step size For both *AppGrad* and stochastic *AppGrad*, a small part of the training set is held out and cross-validation is used to choose the step size adaptively.

Regularization For all the algorithms, a little regularization is added for numerical stability which means we replace Gram matrix $X^\top X$ with $X^\top X + \lambda I$ for some small positive λ.

Oversampling Oversampling means when aiming for the top k dimensional subspace, people usually computes the top $k+l$ dimensional subspace from which a best k dimensional subspace is extracted. In practice, $l = 5 \sim 10$ suffices to improve the performance. We only do a oversampling of 5 in the URL dataset.

3.3.2. Summary of Results

For the first three datasets (Mediamill, MNIST, Penn Tree Bank), both in-sample and out-of-sample PCC are computed for *AppGrad* and Stochastic *AppGrad* as summarized in Figure 1. As you can see, both algorithms capture most of the correlations compared with the true sample CCA subspace and stochastic *AppGrad* consistently achieves the same PCC with much less computational cost than its batch version. Moreover, the larger the size of the data, the bigger advantage will stochastic *AppGrad* obtain. One thing to notice is that, as revealed in Mediamill dataset, out-of-sample PCC is not necessarily less than in-sample PCC because both denominator and numerator change on the hold out set.

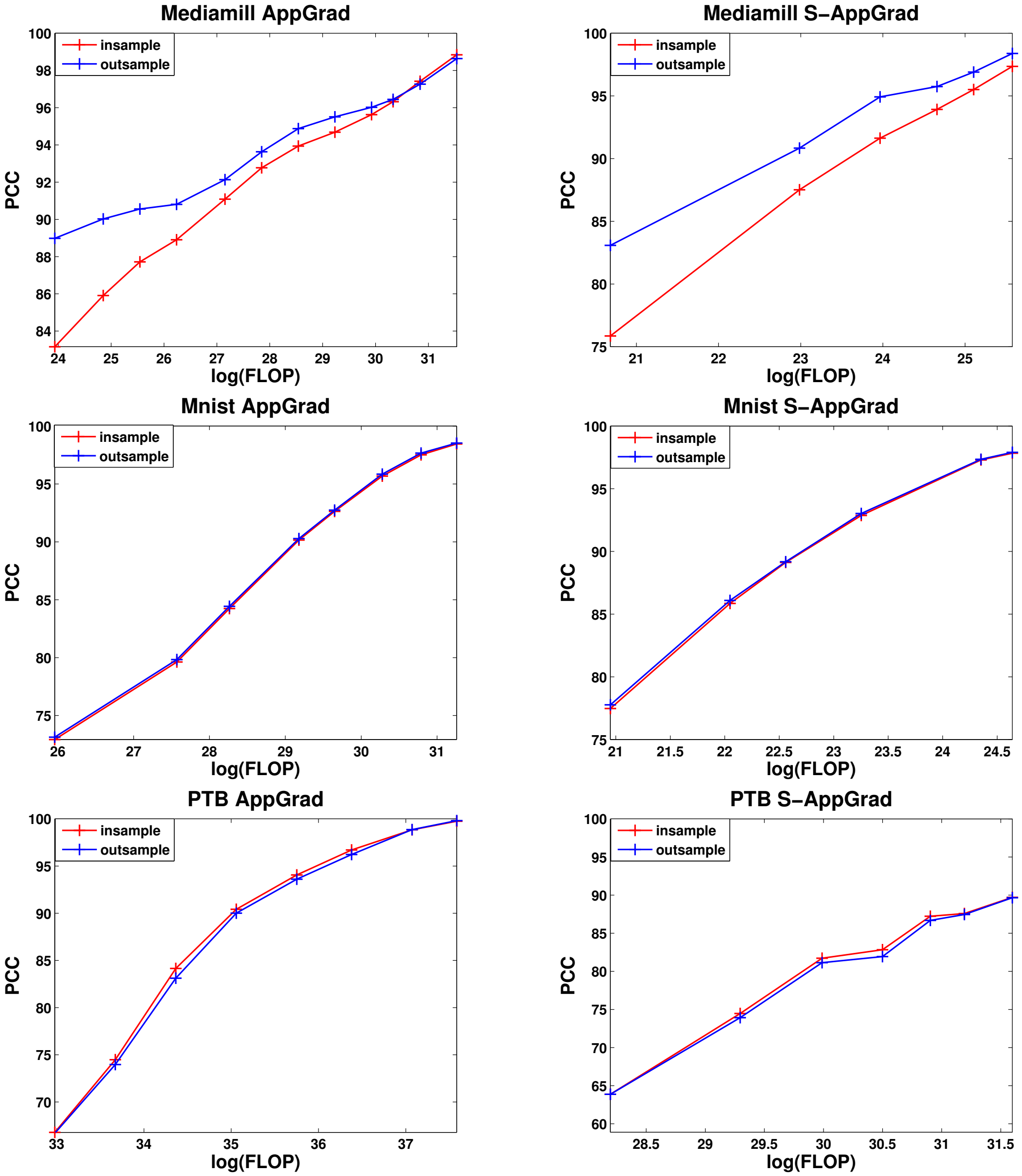

Figure 1: Proportion of Correlations Captured (PCC) by *AppGrad* and stochastic *AppGrad* on different datasets

For URL Reputation dataset, as we mentioned earlier, classical algorithms fail on a personal desktop. The reason is that these algorithms only produce a one-shot estimate after the whole procedure is completed, which is usually prohibitive for huge datasets. In this scenario, the advantage of online algorithms like stochastic *AppGrad* becomes crucial. Further, the stochastic nature makes the algorithm cost-effective and generate decent approximations given fixed computational resources (e.g. FLOP). As revealed by Figure 2, as the number of iterations increases, stochastic *AppGrad* captures more and more correlations.

Since the true sample CCA subspaces for URL dataset are too slow to compute, we compare our algorithm with some naive heuristics which can be carried out efficiently in large scale and catch a reasonable amount of correlation. Below is a brief description of them.

- Non-Whitening (NW-CCA): directly perform SVD on the unwhitened covariance matrix $X^\top Y$. This strategy is also used in Witten et al. (2009)

- Diagonally Whitening (DW-CCA) (Lu and Foster, 2014): avoid inverting matrices by approximating $\widehat{\Sigma}_x^{-1/2}, \widehat{\Sigma}_y^{-1/2}$ with $(\text{diag}(\widehat{\Sigma}_x))^{-1/2}$ and $(\text{diag}(\widehat{\Sigma}_y))^{-1/2}$.

- Whitening the leading m Principal Component Directions (PCA-CCA): First compute the leading m dimensional principal component subspace and project the data matrices X and Y to the subspace, denote them U_x and U_y. Then compute the top k dimensional CCA subspace of the pair (U_x, U_y). At last, transform the CCA subspace of (U_x, U_y) back to the CCA subspace of original matrix pair (X, Y). Specifically for this example, we choose $m = 1200$ (log(FLOP)=35, dominating the computational cost of Stochastic *AppGrad*) .

Remark 3.10. For all the heuristics mentioned above, SVD and PCA steps are carried out using the randomized algorithms in Halko et al. (2011). For PCA-CCA, as the number of Principal Components (m) increases, more correlation will be captured but the computational cost will also increase. When $m = p_1$, PCA-CCA is reduced to the original CCA.

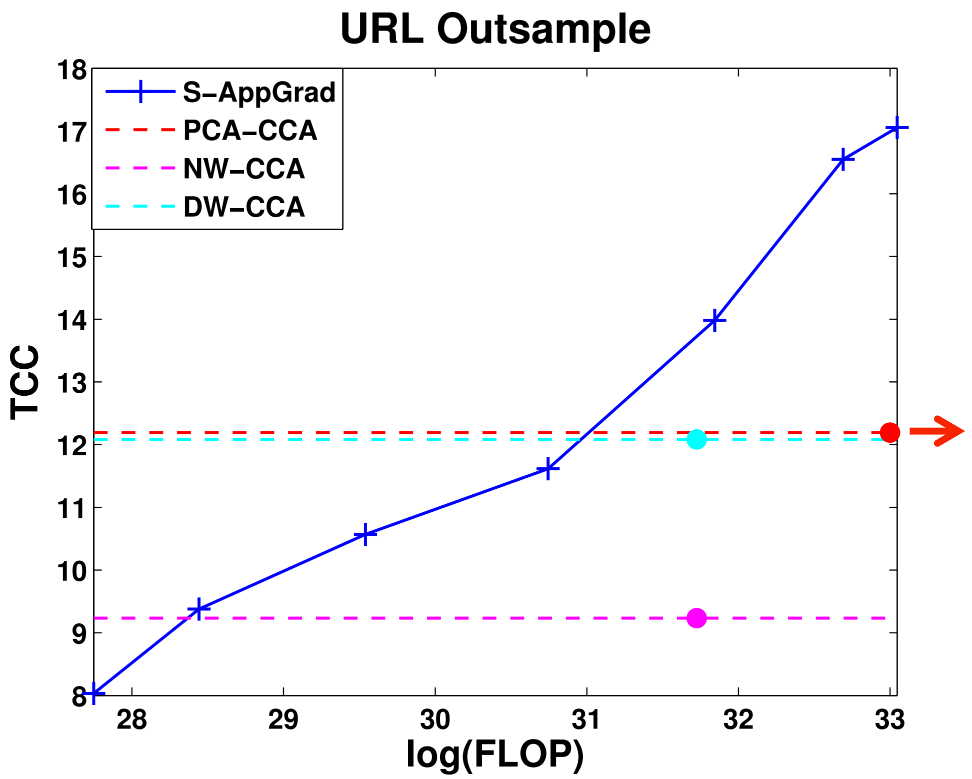

Figure 2: Total Correlations Captured (TCC) by NW-CCA, DW-CCA, PCA-CCA and stochastic *AppGrad* on URL dataset. The dashed lines indicate TCC for those heuristics and the colored dots denote corresponding computational cost. Red arrow means log(FLOP) of PCA-CCA is more than 33.

Essentially, all the heuristics are incorrect algorithms and try to approximately whiten the data matrices. As suggested by Figure 2, stochastic *AppGrad* significantly captures much more correlations.

3.4. Proof of Theorem 4

A brief review of the notations in the main paper:

$$\|u\|_x = (u^\top \widehat{\Sigma}_x u)^{\frac{1}{2}},\ \|v\|_y = (v^\top \widehat{\Sigma}_y v)^{\frac{1}{2}},\ \widetilde{\psi}_1 = \widehat{\lambda}_1 \widehat{\phi}_1,\ \widetilde{\psi}_1 = \widehat{\lambda}_1 \widehat{\psi}_1$$

$$\Delta\widetilde{\phi}^t = \widetilde{\phi}^t - \widetilde{\phi}_1,\ \Delta\widetilde{\psi}^t = \widetilde{\psi}^t - \widetilde{\psi}_1,\ \Delta\phi^t = \phi^t - \widehat{\phi}_1,\ \Delta\psi^t = \psi^t - \widehat{\psi}_1$$

Further, we define $\cos_x(u,v) = \frac{u^\top \widehat{\Sigma}_x v}{\|u\|_x \|v\|_x}$, the cosine of the angle between two vectors induced by the inner product $\langle u, v\rangle = u^\top \widehat{\Sigma}_x v$. Similarly, we define $\cos_y(u,v) = \frac{u^\top \widehat{\Sigma}_y v}{\|u\|_y \|v\|_y}$. To prove the theorem, we will repeatedly use the following lemma.

Lemma 3.11. $\|\Delta\phi^t\|_x \le \frac{1}{\widehat{\lambda}_1}\sqrt{\frac{2}{1+\cos_x(\phi^t,\widehat{\phi}_1)}}\|\Delta\widetilde{\phi}^t\|_x$ *and* $\|\Delta\psi^t\|_y \le \frac{1}{\widehat{\lambda}_1}\sqrt{\frac{2}{1+\cos_y(\psi^t,\widehat{\psi}_1)}}\|\Delta\widetilde{\psi}^t\|_y$

Proof of Lemma 3.11. Notice that $\cos_x(\widetilde{\phi}^t, \widetilde{\phi}_1) = \cos_x(\phi^t, \widehat{\phi}_1)$, then

$$\|\Delta\widetilde{\phi}^t\|_x^2 = \|\widetilde{\phi}^t - \widetilde{\phi}_1\|_x^2 \ge \|\widetilde{\phi}_1\|^2 \sin_x^2(\widetilde{\phi}^t, \widetilde{\phi}_1) = \widehat{\lambda}_1^2 \sin_x^2(\phi^t, \widehat{\phi}_1)$$

Also notice that $\|\phi^t\|_x = \|\widehat{\phi}_1\|_x = 1$, which implies $\cos_x(\phi^t, \widehat{\phi}_1) = 1 - \|\phi^t - \widehat{\phi}_1\|_x^2/2 = 1 - \|\Delta\phi^t\|_x^2/2$. Further

$$\|\Delta\widetilde{\phi}^t\|_x^2 \ge \widehat{\lambda}_1^2 \sin_x^2(\phi^t, \widehat{\phi}_1) = \widehat{\lambda}_1^2(1 - \cos_x^2(\phi^t, \widehat{\phi}_1)) = \frac{\widehat{\lambda}_1^2}{2}\|\Delta\phi^t\|_x^2(1 + \cos_x(\phi^t, \widehat{\phi}_1))$$

Square root both sides,

$$\|\Delta\phi^t\|_x \le \frac{1}{\widehat{\lambda}_1}\sqrt{\frac{2}{1+\cos_x(\phi^t,\widehat{\phi}_1)}}\|\Delta\widetilde{\phi}^t\|_x$$

Similar argument will show that

$$\|\Delta\psi^t\|_y \le \frac{1}{\widehat{\lambda}_1}\sqrt{\frac{2}{1+\cos_y(\psi^t,\widehat{\psi}_1)}}\|\Delta\widetilde{\psi}^t\|_y$$

□

Without loss of generality, we can always assume $\cos_x(\widetilde{\phi}^t, \widetilde{\phi}_1), \cos_y(\widetilde{\psi}^t, \widetilde{\psi}_1) \ge 0$ because the canonical vectors are only identifiable up to a flip in sign and we can always choose $\widetilde{\phi}_1, \widetilde{\psi}_1$ such that the cosines are nonnegative. Apply simple algebra to the gradient step

$\widetilde{\phi}^{t+1} = \widetilde{\phi}^t - \eta(\widehat{\Sigma}_x\widetilde{\phi}^t - \widehat{\Sigma}_{xy}\psi^t)$, we have

$$\begin{aligned} \widetilde{\phi}^{t+1} - \widetilde{\phi}_1 &= \widetilde{\phi}^t - \widetilde{\phi}_1 - \eta(\widehat{\Sigma}_x(\widetilde{\phi}^t - \widetilde{\phi}_1) + \widehat{\Sigma}_x\widetilde{\phi}_1 - \widehat{\Sigma}_{xy}(\psi^t - \widehat{\psi}_1) - \widehat{\Sigma}_{xy}\widehat{\psi}_1) \\ \Delta\widetilde{\phi}^{t+1} &= \Delta\widetilde{\phi}^t - \eta(\widehat{\Sigma}_x\Delta\widetilde{\phi}^t - \widehat{\Sigma}_{xy}\Delta\phi^t) - \eta(\widehat{\Sigma}_x\widetilde{\phi}_1 - \widehat{\Sigma}_{xy}\widehat{\psi}_1) \end{aligned}$$

By Proposition 3.1, $\eta(\widehat{\Sigma}_x\widetilde{\phi}_1 - \widehat{\Sigma}_{xy}\widehat{\psi}_1) = \eta(\widehat{\Sigma}_x\widehat{\phi}_1 - \widehat{\lambda}_1\widehat{\Sigma}_x\widehat{\phi}_1) = 0$, which implies

$$\Delta\widetilde{\phi}^{t+1} = \Delta\widetilde{\phi}^t - \eta(\widehat{\Sigma}_x\Delta\widetilde{\phi}^t - \widehat{\Sigma}_{xy}\Delta\psi^t)$$

Square both sizes,

$$\|\Delta\widetilde{\phi}^{t+1}\|^2 = \|\Delta\widetilde{\phi}^t\|^2 + \eta^2\|\widehat{\Sigma}_x\Delta\widetilde{\phi}^t - \widehat{\Sigma}_{xy}\Delta\psi^t\|^2 - 2\eta(\Delta\widetilde{\phi}^t)^\top(\widehat{\Sigma}_x\Delta\widetilde{\phi}^t - \widehat{\Sigma}_{xy}\Delta\psi^t) \tag{3.4}$$

Again apply Proposition 3.1,

$$\begin{aligned} \|\widehat{\Sigma}_{xy}\Delta\psi^t\| &= \|\widehat{\Sigma}_x\widehat{\Phi}\widehat{\Lambda}\widehat{\Psi}^T\widehat{\Sigma}_y\Delta\psi^t\| \\ &\le \|\widehat{\Sigma}_x^{1/2}\|\|\widehat{\Sigma}_x^{1/2}\widehat{\Phi}\|\|\widehat{\Lambda}\|\|\widehat{\Psi}^\top\widehat{\Sigma}_y^{1/2}\|\|\widehat{\Sigma}_y^{1/2}\Delta\psi^t\| \\ &\le \widehat{\lambda}_1 L_1^{\frac{1}{2}}\|\Delta\psi^t\|_y \end{aligned}$$

The last inequality uses the assumption that $\lambda_{max}(\widehat{\Sigma}_x), \lambda_{max}(\widehat{\Sigma}_y) \le L_1$. By Lemma3.11, $\|\Delta\psi^t\|_y \le \frac{\sqrt{2}}{\widehat{\lambda}_1}\|\Delta\widetilde{\psi}^t\|_y$. Hence, $\|\widehat{\Sigma}_{xy}\Delta\psi^t\| \le \sqrt{2L_1}\|\Delta\widetilde{\psi}^t\|_y$. Also notice that $\|\widehat{\Sigma}_x\Delta\widetilde{\phi}^t\| \le \|\widehat{\Sigma}_x^{1/2}\|\|\widehat{\Sigma}_x^{1/2}\Delta\widetilde{\phi}^t\| \le L_1^{\frac{1}{2}}\|\Delta\widetilde{\phi}^t\|_x$, then

$$\|\widehat{\Sigma}_x\Delta\widetilde{\phi}^t - \widehat{\Sigma}_{xy}\Delta\psi^t\|^2 \le (L_1^{\frac{1}{2}}\|\Delta\widetilde{\phi}^t\|_x + \sqrt{2}L_1^{\frac{1}{2}}\|\Delta\widetilde{\psi}^t\|_y)^2 \le 2L_1(\|\Delta\widetilde{\phi}^t\|_x^2 + 2\|\Delta\widetilde{\psi}^t\|_y^2)$$

Substitute into (3.4),

$$\|\Delta\widetilde{\phi}^{t+1}\|^2 \le \|\Delta\widetilde{\phi}^t\|^2 - 2\eta\|\Delta\widetilde{\phi}^t\|_x^2 + 2L_1\eta^2(\|\Delta\widetilde{\phi}^t\|_x^2 + 2\|\Delta\widetilde{\psi}^t\|_y^2) + 2\eta(\Delta\widetilde{\phi}^t)^\top\widehat{\Sigma}_{xy}\Delta\psi^t \tag{3.5}$$

Now, we are going to bound the term $(\Delta\widetilde{\phi}^t)^T\widehat{\Sigma}_{xy}\Delta\psi^t$. Because $\widehat{\Sigma}_y^{1/2}\widehat{\Psi}$ is an orthonormal matrix and $\widehat{\Sigma}_y^{1/2}\psi_t$ is a unit vector, there exist coefficients $\alpha_1, \cdots, \alpha_p, \alpha_\perp$ and unit vector

$\psi_\perp \in \text{colspan}(\widehat{\Sigma}_y^{1/2}\widehat{\Psi})^\perp$ such that $\widehat{\Sigma}_y^{1/2}\psi_t = \sum_{i=1}^p \alpha_i \widehat{\Sigma}_y^{1/2}\widehat{\psi}_i + \alpha_\perp \widehat{\Sigma}_y^{1/2}\psi_\perp, \sum_{i=1}^p \alpha_i^2 + \alpha_\perp^2 = 1$. Therefore,

$$\begin{aligned}(\Delta\widetilde{\phi}^t)^\top \widehat{\Sigma}_x \widehat{\Phi}\widehat{\Lambda}\widehat{\Psi}^\top \widehat{\Sigma}_y \Delta\psi^t &= \Delta\widetilde{\phi}^t \widehat{\Sigma}_x \widehat{\Phi}\widehat{\Lambda}(\widehat{\Sigma}_y^{1/2}\widehat{\Psi})^\top \Big\{(\alpha_1 - 1)\widehat{\Sigma}_y^{1/2}\widehat{\psi}_1 \\ &\quad + \sum_{i=2}^p \alpha_i \widehat{\Sigma}_y^{1/2}\widehat{\psi}_i + \alpha_\perp \widehat{\Sigma}_y^{1/2}\psi_\perp\Big\} \\ &= \widehat{\lambda}_1(\alpha_1 - 1)(\Delta\widetilde{\phi}^t)^\top \widehat{\Sigma}_x \widehat{\phi}_1 + \sum_{i=2}^p \alpha_i \widehat{\lambda}_i (\Delta\widetilde{\phi}^t)^\top \widehat{\Sigma}_x \widehat{\phi}_i\end{aligned}$$

By Cauchy-Schwartz inequality,

$$\begin{aligned}(\Delta\widetilde{\phi}^t)^\top \widehat{\Sigma}_x \widehat{\Phi}\widehat{\Lambda}\widehat{\Psi}^\top \widehat{\Sigma}_y \Delta\psi^t &\le \left(\widehat{\lambda}_1^2(1-\alpha_1)^2 + \sum_{i=2}^p \alpha_i^2 \widehat{\lambda}_i^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^p \left((\Delta\widetilde{\phi}^t)^\top \widehat{\Sigma}_x \widehat{\phi}_1\right)^2\right)^{\frac{1}{2}} \\ &\le \left(\widehat{\lambda}_1^2(1-\alpha_1)^2 + \widehat{\lambda}_2^2(1-\alpha_1^2)\right)^{\frac{1}{2}} \|\Delta\widetilde{\phi}^t\|_x \\ &= \left(\widehat{\lambda}_1^2 \frac{1-\alpha_1}{1+\alpha_1} + \widehat{\lambda}_2^2\right)^{\frac{1}{2}} (1-\alpha_1^2)^{\frac{1}{2}} \|\Delta\widetilde{\phi}^t\|_x\end{aligned}$$

By definition, $1 - \alpha_1 = 1 - \cos_y(\psi^t, \widehat{\psi}_1) = \frac{\|\Delta\psi^t\|_y^2}{2}$. Further by Lemma 3.11,

$$1 - \alpha_1 \le \frac{1}{\widehat{\lambda}_1^2(1+\alpha_1)} \|\Delta\widetilde{\psi}^t\|_y^2$$

Therefore,

$$\begin{aligned}(\Delta\widetilde{\phi}^t)^\top \widehat{\Sigma}_x \Phi\widehat{\Lambda}\Psi^\top \widehat{\Sigma}_y \Delta\psi^t &\le \left(\frac{1-\alpha_1}{1+\alpha_1} + \frac{\widehat{\lambda}_2^2}{\widehat{\lambda}_1^2}\right)^{\frac{1}{2}} \|\Delta\widetilde{\phi}^t\|_x \|\Delta\widetilde{\psi}^t\|_y \\ &\le \left(\frac{\|\Delta\widetilde{\psi}^t\|_y^2}{\widehat{\lambda}_1^2(1+\alpha_1)^2} + \frac{\widehat{\lambda}_2^2}{\widehat{\lambda}_1^2}\right)^{\frac{1}{2}} \|\Delta\widetilde{\phi}^t\|_x \|\Delta\widetilde{\psi}^t\|_y \\ &\le \frac{1}{2}\left(\frac{\|\Delta\widetilde{\psi}^t\|_y^2}{\widehat{\lambda}_1^2} + \frac{\widehat{\lambda}_2^2}{\widehat{\lambda}_1^2}\right)^{\frac{1}{2}} \left(\|\Delta\widetilde{\phi}^t\|_x^2 + \|\Delta\widetilde{\psi}^t\|_y^2\right)\end{aligned}$$

Substitute into (3.5),

$$\begin{aligned}\|\Delta\widetilde{\phi}^{t+1}\|^2 \le \|\Delta\widetilde{\phi}^t\|^2 - 2\eta\|\Delta\widetilde{\phi}^t\|_x^2 + 2L_1\eta^2\Big(\|\Delta\widetilde{\phi}^t\|_x^2 + 2\|\Delta\widetilde{\psi}^t\|_y^2\Big)\\ + \eta\left(\frac{\|\Delta\widetilde{\psi}^t\|_y^2}{\widehat{\lambda}_1^2} + \frac{\widehat{\lambda}_2^2}{\widehat{\lambda}_1^2}\right)^{\frac{1}{2}}\Big(\|\Delta\widetilde{\phi}^t\|_x^2 + \|\Delta\widetilde{\psi}^t\|_y^2\Big)\end{aligned}$$

Similar analysis implies that,

$$\begin{aligned}\|\Delta\widetilde{\psi}^{t+1}\|^2 \le \|\Delta\widetilde{\psi}^t\|^2 - 2\eta\|\Delta\widetilde{\psi}^t\|_y^2 + 2L_1\eta^2\Big(\|\Delta\widetilde{\psi}^t\|_y^2 + 2\|\Delta\widetilde{\phi}^t\|_x^2\Big)\\ + \eta\left(\frac{\|\Delta\widetilde{\phi}^t\|_x^2}{\widehat{\lambda}_1^2} + \frac{\widehat{\lambda}_2^2}{\widehat{\lambda}_1^2}\right)^{\frac{1}{2}}\Big(\|\Delta\widetilde{\phi}^t\|_x^2 + \|\Delta\widetilde{\psi}^t\|_y^2\Big)\end{aligned}$$

Add these two inequalities,

$$\begin{aligned}\|\Delta\widetilde{\phi}^{t+1}\|^2 + \|\Delta\widetilde{\psi}^{t+1}\|^2 \le \Big(\|\Delta\widetilde{\phi}^t\|^2 + \|\Delta\widetilde{\psi}^t\|^2\Big) - 2\eta\left\{1 - \frac{1}{2}\left(\frac{\|\Delta\widetilde{\psi}^t\|_y^2}{\widehat{\lambda}_1^2} + \frac{\widehat{\lambda}_2^2}{\widehat{\lambda}_1^2}\right)^{\frac{1}{2}}\right.\\ \left.-\frac{1}{2}\left(\frac{\|\Delta\widetilde{\phi}^t\|_x^2}{\widehat{\lambda}_1^2} + \frac{\widehat{\lambda}_2^2}{\widehat{\lambda}_1^2}\right)^{\frac{1}{2}} - 3L_1\eta\right\}\Big(\|\Delta\widetilde{\phi}^t\|_x^2 + \|\Delta\widetilde{\psi}^t\|_y^2\Big)\end{aligned}$$

Notice that $\sqrt{a} + \sqrt{b} \le \sqrt{2(a+b)}$, we have

$$\begin{aligned}\left(\frac{\|\Delta\widetilde{\psi}^t\|_y^2}{\widehat{\lambda}_1^2} + \frac{\widehat{\lambda}_2^2}{\widehat{\lambda}_1^2}\right)^{\frac{1}{2}} + \left(\frac{\|\Delta\widetilde{\phi}^t\|_x^2}{\widehat{\lambda}_1^2} + \frac{\widehat{\lambda}_2^2}{\widehat{\lambda}_1^2}\right)^{\frac{1}{2}} &\le \left(\frac{2\|\Delta\widetilde{\psi}^t\|_y^2}{\widehat{\lambda}_1^2} + \frac{2\|\Delta\widetilde{\phi}^t\|_x^2}{\widehat{\lambda}_1^2} + \frac{4\widehat{\lambda}_2^2}{\widehat{\lambda}_1^2}\right)^{\frac{1}{2}}\\ &\le \left(\frac{2L_1\|\Delta\widetilde{\psi}^t\|^2}{\widehat{\lambda}_1^2} + \frac{2L_1\|\Delta\widetilde{\phi}^t\|^2}{\widehat{\lambda}_1^2} + \frac{4\widehat{\lambda}_2^2}{\widehat{\lambda}_1^2}\right)^{\frac{1}{2}}\\ &= \frac{1}{2\widehat{\lambda}_1}\left(\frac{L_1}{2}\|\Delta\widetilde{\psi}^t\|^2 + \frac{L_1}{2}\|\Delta\widetilde{\phi}^t\|^2 + \widehat{\lambda}_2^2\right)^{\frac{1}{2}}\end{aligned}$$

Then,

$$\begin{aligned}\|\Delta\widetilde{\phi}^{t+1}\|^2 + \|\Delta\widetilde{\psi}^{t+1}\|^2 \le \Big(\|\Delta\widetilde{\phi}^t\|^2 + \|\Delta\widetilde{\psi}^t\|^2\Big) - 2\eta\Big(\|\Delta\widetilde{\phi}^t\|_x^2 + \|\Delta\widetilde{\psi}^t\|_y^2\Big)\\ \times\left\{1 - \frac{1}{\widehat{\lambda}_1}\left(\frac{L_1}{2}\|\Delta\widetilde{\psi}^t\|^2 + \frac{L_1}{2}\|\Delta\widetilde{\phi}^t\|^2 + \widehat{\lambda}_2^2\right)^{\frac{1}{2}} - 3L_1\eta\right\}\end{aligned} \tag{3.6}$$

By definition, $\delta = 1 - \widehat{\lambda}_1^{-1} \left(\frac{L_1}{2} \|\Delta\widetilde{\psi}^0\|^2 + \frac{L_1}{2} \|\Delta\widetilde{\phi}^0\|^2 + \widehat{\lambda}_2^2 \right)^{\frac{1}{2}}$ and $\eta = \delta/6L_1$. Substitute in (3.6) with $t = 0$,

$$\begin{aligned}
\|\Delta\widetilde{\phi}^1\|^2 + \|\Delta\widetilde{\psi}^1\|^2 &= \left(\|\Delta\widetilde{\phi}^0\|^2 + \|\Delta\widetilde{\psi}^0\|^2 \right) - \frac{\delta^2}{6L_1} \left(\|\Delta\widetilde{\phi}^0\|_x^2 + \|\Delta\widetilde{\psi}^0\|_y^2 \right) \\
&\leq \left(\|\Delta\widetilde{\phi}^0\|^2 + \|\Delta\widetilde{\psi}^t\|^2 \right) - \frac{\delta^2}{6L_1L_2} \left(\|\Delta\widetilde{\phi}^0\|^2 + \|\Delta\widetilde{\psi}^0\|^2 \right) \\
&\leq \left(1 - \frac{\delta^2}{6L_1L_2} \right) \left(\|\Delta\widetilde{\phi}^0\|^2 + \|\Delta\widetilde{\psi}^0\|^2 \right)
\end{aligned}$$

It follows by induction that $\forall\, t \in \mathbb{N}_+$

$$\|\Delta\widetilde{\phi}^{t+1}\|^2 + \|\Delta\widetilde{\psi}^{t+1}\|^2 \leq \left(1 - \frac{\delta^2}{6L_1L_2} \right) \left(\|\Delta\widetilde{\phi}^t\|^2 + \|\Delta\widetilde{\psi}^t\|^2 \right)$$

CHAPTER 4 : Network Data Modeling

4.1. Introduction

Network is a prevalent form of data for quantitative and qualitative analysis in a number of fields, including but not limited to sociology, computer science, neuroscience, etc. Moreover, due to advances in science and technology, the sizes of the networks we encounter are ever increasing. Therefore, to explore, to visualize and to utilize the information in large networks poses significant challenges to Statistics. Unlike traditional datasets in which a number of features are recorded for each subject, network datasets provide information on the relation among all subjects under study, sometimes together with additional features. In this paper, we focus on the modeling, visualization and exploration of networks in which additional features might be observed for each node pair.

On real world networks, people oftentimes observe the following characteristics. First, the degree distributions of nodes are often right-skewed and so networks exhibit degree heterogeneity. In addition, connections in networks often demonstrate transitivity, that is nodes with common neighbors are more likely to be connected. Moreover, nodes that are similar in certain ways (students in the same grade, brain regions that are close physically, etc.) are more likely to form bonds. Such a phenomenon is usually called homophily in network studies. Furthermore, nodes in some networks exhibit clustering effect and in such cases it is desirable to partition the nodes into different communities.

An efficient way to explore network data and to extract key information is to fit appropriate statistical models on them. To date, there have been a collection of network models proposed by researchers in various fields. These models aim to catch different subsets of the foregoing characteristics, and Goldenberg et al. (2010) provides a comprehensive overview. An important class of network models are *latent space models* (Hoff et al., 2002). Suppose there are n nodes in the observed network. The key idea underlying latent space modeling is that each node i can be represented by a vector z_i in some low dimensional Euclidean space (or some other metric space of choice) that is sometimes called the social space, and nodes that are "close" in the social space are more likely to be connected. Hoff et al. (2002) considered two types of latent space models: distance models and projection models. In both cases, the latent vectors $\{z_i\}_{i=1}^n$ were treated as fixed effects. Later, a series of papers (Hoff, 2003; Handcock et al., 2007; Krivitsky et al., 2009) generalized the original proposal in Hoff et al. (2002) for better modeling of other characteristics of social networks, such as clustering, degree heterogeneity, etc. In these generalizations, the z_i's were treated as random effects generated from certain multivariate Gaussian mixtures. Moreover, model fitting and inference in these models has been carried out via Markov Chain Monte Carlo, and it is difficult to scale these methodologies to handle large networks (Goldenberg et al., 2010). Moreover, one needs to use different likelihood function based on choice of model and there is little understanding of the quality of fitting when the model is mis-specified. Albeit these disadvantages, latent space models are attractive due to their friendliness to interpretation and visualization.

In this paper, we aim to tackle the following two key issues in latent space modeling of network data. First, we seek a class of latent space models that is special enough so that we can design fast fitting algorithms for them and hence be able to handle networks of very large sizes. In addition, we would like to be able to fit a class of models that are flexible enough to well approximate a wide range of latent space models of interest so that fitting methods for this flexible class continue to work even when the model is mis-specified. From a practical viewpoint, if one is able to find such a class of models and design fast algorithms for fitting them, then one would be able to use this class as working models and to use the associated fast algorithms to effectively explore large networks.

Main contributions We make progress on tackling the foregoing two issues simultaneously in the present paper, which we summarize as the following main contributions:

1. We propose a special class of latent space models, called inner-product models, and

two fast fitting algorithms for this class. Let the observed n-by-n adjacency matrix and covariate matrix be A and X, respectively. The inner-product model assumes that for any $i < j$,

$$\begin{aligned} A_{ij} = A_{ji} &\overset{ind.}{\sim} \text{Bernoulli}(P_{ij}), \quad \text{with} \\ \text{logit}(P_{ij}) = \Theta_{ij} &= \alpha_i + \alpha_j + \beta X_{ij} + z_i^\top z_j, \end{aligned} \tag{4.1}$$

where for any $x \in (0,1)$, $\text{logit}(x) = \log[x/(1-x)]$. Here, α_i, $1 \le i \le n$, are parameters modeling degree heterogeneity. The parameter β is the coefficient for the observed covariate, and $z_i^\top z_j$ is the inner-product between the latent vectors. As we will show later in Section 4.2, this class of models can incorporate degree heterogeneity, transitivity and homophily explicitly. From a matrix estimation viewpoint, the matrix $G = (G_{ij}) = (z_i^\top z_j)$ is of rank at most k that is much smaller than n. Motivated by recent advances in low rank matrix estimation, we design two fast algorithms for fitting (4.1). One algorithm is based on lifting and nuclear norm penalization of the negative log-likelihood function. The other is based on directly optimizing the negative log-likelihood function via projected gradient descent. For both algorithms, we establish high probability error bounds for inner-product models.

2. We further show that these two fitting algorithms are “universal” in the sense they can work simultaneously for a wide range of latent space models beyond the inner-product model class. For example, they work for the distance model and the Gaussian kernel model in which the inner-product term $z_i^\top z_j$ in (4.1) is replaced with $-\|z_i - z_j\|$ and $c\exp(-\|z_i - z_j\|^2/\sigma^2)$, respectively. Thus, the class of inner-product models is flexible and can be used to approximate many other latent space models of interest. In addition, the associated algorithms can be applied to networks generated from a wide range of mis-specified models and still yield reasonable results. The key mathematical insight that enables such flexibility is introduced in Section 4.2 as the Schoenberg Condition (4.7).

3. We demonstrate the effectiveness of the model and algorithms on real data examples. In particular, we fit inner-product models by the proposed algorithms on five different real network datasets for several different tasks, including visualization, clustering and network-assisted classification. On three popular benchmark datasets for testing community detection on networks, a simple k-means clustering on the estimated latent vectors obtained by our algorithm yields as good result on one dataset and better results on the other two when compared with four state-of-the-art methods. The same "model fitting followed by k-means clustering" approach also yields nice clustering of nodes on a network with edge covariates. Due to the nature of latent space models, for all datasets on which we fit the model, we obtain natural visualizations of the networks by plotting latent vectors. Furthermore, we illustrate how network information can be incorporated in traditional learning problems using a document classification example.

Related works When fitting a network model, we are essentially modeling and estimating the edge probability matrix. From this viewpoint, the present paper is related to the literature on graphon estimation and edge probability matrix estimation for block models. See, for instance, Bickel and Chen (2009); Airoldi et al. (2013); Wolfe and Olhede (2013); Gao et al. (2015a); Klopp et al. (2015); Gao et al. (2016) and the references therein. However, the block models have stronger structural assumptions than the latent space models we are going to investigate. In addition, it is relatively difficult to introduce edge covariates in block models while Binkiewicz et al. (2015) has made an interesting attempt.

The algorithmic and theoretical aspects of the paper is also closely connected to the line of research on low rank matrix estimation, which plays an important role in many applications such as phase retrieval Candes et al. (2015); Candès et al. (2015) and matrix completion Candès and Tao (2010); Keshavan et al. (2010a,b); Candès and Recht (2012); Koltchinskii et al. (2011). Indeed, the idea of nuclear norm penalization has originated from matrix completion Candès and Tao (2010). The idea of directly optimizing a non-convex objective function involving a low rank matrix has been studied recently in a series of important papers. See, for instance, Burer and Monteiro (2005); Sun and Luo (2016); Tu et al. (2015); Chen and Wainwright (2015); Zheng and Lafferty (2016); Ge et al. (2016b) and the references therein. Among these papers, the one that is the most related to the projected gradient descent algorithm we are to propose and analyze is Chen and Wainwright (2015) which focused on estimating a positive semi-definite matrix of exact low rank in a collection of interesting problems. However, we will obtain tighter error bounds for latent space models and we will go beyond the exact low rank scenario. We would like to reiterate that the aforementioned works are mostly related to algorithmic and theoretical components of the present paper, but they have little to do with the modeling aspect.

Organization After a brief introduction of standard notation used throughout the paper, the rest of the paper is organized as follows. Section 4.2 introduces both inner-product models and a broader class of latent space models on which our fitting methods work. The two fitting methods are described in detail in Section 4.3, followed by their theoretical guarantees under both inner-product models and the broader class. The theoretical results are further corroborated by simulated examples in Section 4.4. Section 4.5 demonstrates the competitive performance of the modeling approach and fitting methods on five different real network datasets. We discuss interesting related problems in Section 4.6 and present proofs of the main results in Section 4.7. Technical details justifying the initialization methods for the project gradient descent approach are deferred to the appendix.

Notation For $X, Y \in \mathbb{R}^{m\times n}$, $X \circ Y$ denotes the Hadamard (element-wise) product between X and Y and $\langle X, Y \rangle = tr(X^\top Y)$ defines an inner product between them. For any function f, $f(X)$ is the shorthand for applying $f(\cdot)$ element-wisely to X, that is $f(X) \in \mathbb{R}^{m\times n}$ and $[f(X)]_{ij} = f(X_{ij})$. $vec(X)$ is the vector constructed by stacking the matrix X column by column. X_{i*} and X_{*j} respectively denote the i_{th} row and j_{th} column of X. $\mathbb{S}_+^n$ is the set of all $n \times n$ positive semidefinite matrices and $O(m, n)$ is the set of all $m \times n$ orthonormal matrices. P_X is the projection matrix onto the column space of X.

4.2. Latent Space Models

In this section, we first give a detailed introduction of the inner-product model (4.1) and conditions for its identifiability. In addition, we introduce a more general class of latent space models that includes the inner-product model as a special case. The methods we propose later will be motivated by the inner-product model and can also be applied to the more general class.

4.2.1. Inner-product models

Recall the inner-product model defined in (4.1), i.e., for any observed A and X and any $i < j$,

$$A_{ij} = A_{ji} \overset{ind.}{\sim} \text{Bernoulli}(P_{ij}), \quad \text{with} \quad \text{logit}(P_{ij}) = \Theta_{ij} = \alpha_i + \alpha_j + \beta X_{ij} + z_i^\top z_j.$$

Fixing all other parameters, if we increase α_i, then node i has higher chances of connecting with other nodes. Therefore, the α_i's model degree heterogeneity of nodes and we call them degree heterogeneity parameters. Next, the regression coefficient β moderates the contribution of covariate to edge formation. For instance, if X_{ij} indicates whether nodes i and j share some common attribute such as gender, then a positive β value implies that nodes that share common feature are more likely to connect. Such a phenomenon is called *homophily* in the social network literaute. Last but not least, the latent variables $\{z_i\}_{i=1}^n$ enter the model through their inner-product $z_i^\top z_j$, and hence is the name of the model. We impose no additional structural/distributional assumptions on the latent variables for the sake of modeling flexibility.

We note that model (4.1) also allows the latent variables to enter the second equation in the form of $g(z_i, z_j) = -\frac{1}{2}\|z_i - z_j\|^2$. To see this, note that $g(z_i, z_j) = -\frac{1}{2}\|z_i\|^2 - \frac{1}{2}\|z_j\|^2 + z_i^\top z_j$, and we may re-parameterize by setting $\tilde{\alpha}_i = \alpha_i - \frac{1}{2}\|z_i\|^2$ for all i. Then we have

$$\Theta_{ij} = \alpha_i + \alpha_j + \beta X_{ij} - \frac{1}{2}\|z_i - z_j\|^2 = \tilde{\alpha}_i + \tilde{\alpha}_j + \beta X_{ij} + z_i^\top z_j.$$

An important implication of this observation is that the function $g(z_i, z_j) = -\frac{1}{2}\|z_i - z_j\|^2$ directly models *transitivity*, i.e., nodes with common neighbors are more likely to connect since their latent variables are more likely to be close to each other in the latent space. In view of the foregoing discussion, the inner-product model (4.1) also enjoys this nice modeling capacity.

In matrix form, we have

$$\Theta = \alpha 1_n{}^\top + 1_n \alpha^\top + \beta X + G \tag{4.2}$$

where 1_n is the all one vector in $\mathbb{R}^n$ and $G = ZZ^\top$ with $Z = (z_1, \cdots, z_n)^\top \in \mathbb{R}^{n \times k}$. Since there is no self-edge and Θ is symmetric, only the upper diagonal elements of Θ are well defined, which we denote by Θ^u. Nonetheless we define the diagonal element of Θ as in (4.2) since it is inconsequential. To ensure identifiability of model parameters in (4.1), we assume the latent variables are centered, that is

$$JZ = Z \quad \text{where} \quad J = I_n - \frac{1}{n} 1_n 1_n{}^\top. \tag{4.3}$$

Note that this condition uniquely identifies Z up to an orthogonal transformation of the rows while $G = ZZ^\top$ is now directly identifiable.

4.2.2. A more general class of latent space models

Model (4.1) is a special case of a more general class of latent space models, which can be defined by

$$\begin{gathered} A_{ij} = A_{ji} \overset{ind.}{\sim} \text{Bernoulli}(P_{ij}), \quad \text{with} \\ \text{logit}(P_{ij}) = \Theta_{ij} = \tilde{\alpha}_i + \tilde{\alpha}_j + \beta X_{ij} + h(z_i, z_j) \end{gathered} \tag{4.4}$$

where $h(\cdot, \cdot)$ is a smooth symmetric function on $\mathbb{R}^k \times \mathbb{R}^k$. We shall impose an additional constraint on h following the discussion below. In matrix form, for $\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n)'$ and $H = (h(z_i, z_j))$, we can write

$$\Theta = \tilde{\alpha} 1_n{}^\top + 1_n \tilde{\alpha}^\top + \beta X + H.$$

To better connect with (4.2), let

$$G = JHJ, \quad \text{and} \quad \alpha 1_n{}^\top + 1_n \alpha^\top = \tilde{\alpha} 1_n{}^\top + 1_n \tilde{\alpha}^\top + H - JHJ. \tag{4.5}$$

Note that the second equation in the last display holds since the expression on its right hand side is symmetric and of rank at most two. Then we can rewrite the second last display as

$$\Theta = \alpha 1_n{}^\top + 1_n \alpha^\top + \beta X + G \tag{4.6}$$

which reduces to (4.2) and G satisfies $JG = G$. Our additional constraint on h is the following Schoenberg Condition:

$$\begin{aligned}&\text{For any positive integer } n \geq 2 \text{ and any } z_1, \ldots, z_n \in \mathbb{R}^k,\\ &G = JHJ \text{ is positive semi-definite for } H = (h(z_i, z_j)) \text{ and } J = I_n - \tfrac{1}{n} 1_n 1_n{}^\top.\end{aligned} \tag{4.7}$$

Condition (4.7) may seem abstract, while the following lemma provides two important sets of symmetric functions for which it is satisfied.

Lemma 4.1. *Condition* (4.7) *is satisfied in the following cases:*

1. *h is a positive semi-definite kernel function on $\mathbb{R}^k \times \mathbb{R}^k$;*

2. *$h(x, y) = -\|x - y\|_p^q$ for some $p \in (0, 2]$ and $q \in (0, p]$ where $\| \cdot \|_p$ is the p-norm (or p-seminorm when $p < 1$) on $\mathbb{R}^k$.*

The first claim of Lemma 4.1 is a direct consequence of the definition of positive semi-definite kernel function which ensures that the matrix H itself is positive semi-definite and so is $G = JHJ$ since J is also positive semi-definite. The second claim is a direct consequence of the famous Hilbert space embedding result by Schoenberg (Schoenberg, 1937, 1938). See, for instance, Theorems 1 and 2 of Schoenberg (1937).

4.3. Two Model Fitting Methods

In this section, we propose two methods for fitting models (4.1) and (4.4)–(4.7) on network datasets. Both methods are motivated by minimizing the negative log-likelihood function of the inner-product model, and can be regarded as pseudo-likelihood approaches for more general models. In what follows, we first motivate and describe both methods for the inner-product model and establish their theoretical guarantees. Then we extend these guarantees to the general class. From a methodological viewpoint, a key advantage of these methods, in particular the projected gradient descent method, is scalability to networks of large sizes.

4.3.1. A convex approach via penalized MLE

We first focus on the inner-product model (4.1) in which the parameter Θ belongs to the following set

$$\begin{aligned}\mathcal{F}(n,k,M_1,M_2,X) = \big\{\Theta | \Theta = \alpha 1_n{}^\top + 1_n \alpha^\top + \beta X + ZZ^\top,\ JZ = Z \in \mathbb{R}^{n\times k}, \\ -M_1 \le \Theta_{ij} \le -M_2 \text{ for } 1 \le i \ne j \le n, \max_{1\le i\le n} |\Theta_{ii}| \le M_1\big\}\end{aligned} \tag{4.8}$$

where k is the latent space dimension and both M_1 and M_2 are positive. In what follows, we allow all these quantities to scale on n. Notice that for any $\Theta \in \mathcal{F}(n,k,M_1,M_2,X)$, the corresponding edge probabilities satisfy

$$\frac{1}{2}e^{-M_1} \le \frac{1}{1+e^{M_1}} \le P_{ij} \le \frac{1}{1+e^{M_2}} \le e^{-M_2},\ 1 \le i \ne j \le n. \tag{4.9}$$

Thus M_1 controls the conditioning of the problem and M_2 controls the sparsity of the network.

Let $\sigma(x) = 1/(1+e^{-x})$ be the sigmoid function, then for any $i \ne j$, $P_{ij} = \sigma(\Theta_{ij})$ and the log-likelihood function of model (4.1) can be written as

$$\begin{aligned}\ell(\Theta^u|A) &= \sum_{i<j}\Big\{A_{ij}\log\Big(\sigma(\Theta_{ij})\Big) + (1-A_{ij})\log\Big(1-\sigma(\Theta_{ij})\Big)\Big\}\\ &= \sum_{i<j}\Big\{A_{ij}\Theta_{ij} + \log\Big(1-\sigma(\Theta_{ij})\Big)\Big\}.\end{aligned}$$

Recall that $G = ZZ^\top$. The maximum likelihood estimate of Θ^u is the solution of the following rank constrained minimization problem:

$$\begin{aligned}\min_{\Theta^u,\alpha,\beta,G} \quad & -\sum_{i<j}\Big\{A_{ij}\Theta_{ij} + \log\Big(1-\sigma(\Theta_{ij})\Big)\Big\},\\ \text{subject to} \quad & \Theta = \alpha 1_n{}^\top + 1_n\alpha^\top + \beta X + G, \quad -M_1 \le \Theta_{ij} \le -M_2,\\ & GJ = G, \quad G \in \mathbb{S}_+^n, \quad \operatorname{rank}(G) \le k.\end{aligned} \tag{4.10}$$

This optimization problem is non-convex and generally intractable. To overcome this difficulty, we consider a convex relaxation that replaces the rank constraint on G in (4.10) with a penalty term on its nuclear norm. Since G is positive semi-definite, its nuclear norm equals its trace. Thus, our first model fitting scheme solves the following convex program:

$$\begin{aligned}\min_{\alpha,\beta,G} \quad & -\sum_{i,j}\Big\{A_{ij}\Theta_{ij} + \log\Big(1-\sigma(\Theta_{ij})\Big)\Big\} + \lambda_n \operatorname{tr}(G)\\ \text{subject to} \quad & \Theta = \alpha 1_n{}^\top + 1_n\alpha^\top + \beta X + G, GJ = G, \quad G \in \mathbb{S}_+^n\\ & -M_1 \le \Theta_{ij} \le -M_2.\end{aligned} \tag{4.11}$$

Remark 4.2. We remark that in addition to the introduction of the trace penalty, the first term in the objective function in (4.11) now sums over all (i,j) pairs. Due to symmetry, after scaling, the difference from the sum in (4.10) lies in the inclusion of all diagonal terms in Θ. This slight modification leads to no noticeable difference in practice. However, it allows easier implementation and simplifies the theoretical investigation. We would also like to comment that the constraint $-M_1 \le \Theta_{ij} \le -M_2$ is included partially for obtaining theoretical guarantees. In simulated examples, we have found that the convex program worked equally well without this constraint.

Theoretical guarantees. We now turn to establishing theoretical guarantees for the optimizer of (4.11). When X in nonzero, we make the following assumption for the identifiability of β.

Assumption 4.3.1. The *stable rank* of the covariate matrix X satisfies $\mathrm{r}_{\text{stable}}(X) = \|X\|_{\mathrm{F}}^2 / \|X\|_{\mathrm{op}}^2 \geq M_0 k$ for some large enough constant M_0.

The linear dependence on k of $\mathrm{r}_{\text{stable}}(X)$ is in some sense necessary in order for β to be identifiable as otherwise the effect of the covariates could be absorbed into the latent component $ZZ^\top$.

Let $(\widehat{\alpha}, \widehat{\beta}, \widehat{G})$ be the solution to the optimization problem (4.11) and $(\alpha_\star, \beta_\star, G_\star)$ be the true parameter that governs the data generation process. Let $\widehat{\Theta}$ and $\Theta_\star$ be defined as in (4.2) but with the estimates and the true parameter values respectively. Define the error terms $\Delta_{\widehat{\Theta}} = \widehat{\Theta} - \Theta_\star$, $\Delta_{\widehat{\alpha}} = \widehat{\alpha} - \alpha_\star$, $\Delta_{\widehat{\beta}} = \hat{\beta} - \beta_\star$ and $\Delta_{\widehat{G}} = \widehat{G} - G_\star$. The following theorem gives both deterministic and high probability error bounds for estimating both the latent vectors and logit-transformed probability matrix.

Theorem 5. *Under Assumption 4.3.1, for any λ_n satisfying $\lambda_n \geq \max\{2 \|A-P\|_{\mathrm{op}}, \ |\langle A - P, X/\|X\|_{\mathrm{F}}\rangle|/\sqrt{k}, \ 1\}$, there exists a constant C such that*

$$\|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2, \|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \leq Ce^{2M_1}\lambda_n^2 k.$$

Specifically, setting $\lambda_n = C_0\sqrt{\max\{ne^{-M_2}, \log n\}}$ for a large enough constant C_0, there exist positive constants c, C such that with probability at least $1 - n^{-c}$,

$$\|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2, \|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \leq Ce^{2M_1 - M_2} nk \times \max\left\{1, \frac{e^{M_2}\log n}{n}\right\}.$$

If we turn the error metrics in Theorem 5 to mean squared errors, namely $\|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2/n^2$ and

$\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2/n^2$, then we obtain the familiar k/n rate in low rank matrix estimation problems.

Remark 4.3. Note that the choice of the penalty parameter λ_n depends on e^{-M_2} which by (4.9) controls the sparsity of the observed network. In practice, we do not know this quantity and we propose to estimate M_2 with $\widehat{M_2} = \operatorname{logit}(\sum_{ij} A_{ij}/n^2)$.

4.3.2. A non-convex approach via projected gradient descent

Although the foregoing convex relaxation method is conceptually neat, state-of-the-art algorithms to solve the nuclear (trace) norm minimization problem (4.11) such as iterative singular value thresholding usually require computing a full singular value decomposition at every iteration, which can still be time consuming on large networks.

To further improve scalability of model fitting, we propose an efficient first order algorithm that directly tackles the following non-convex objective function:

$$\begin{aligned} \min_{Z,\alpha,\beta}\ h(Z,\alpha,\beta) &= -\sum_{i,j}\Big\{A_{ij}\Theta_{ij} + \log\Big(1-\sigma(\Theta_{ij})\Big)\Big\} \\ \text{where } \Theta &= \alpha 1_n{}^\top + 1_n\alpha^\top + \beta X + ZZ^\top. \end{aligned} \tag{4.12}$$

The detailed description of the method is presented in Algorithm 6.

Algorithm 6 A projected gradient descent model fitting method.

```
Input: Adjacency matrix: $A$; covariate matrix: $X$; latent space dimension: $k \geq 1$; initial estimates: $Z^0, \alpha^0, \beta^0$; step sizes: $\eta_Z, \eta_\alpha, \eta_\beta$; constraint sets: $\mathcal{C}_Z, \mathcal{C}_\alpha, \mathcal{C}_\beta$.
Output: $\widehat{Z} = Z^T$, $\widehat{\alpha} = \alpha^T$, $\widehat{\beta} = \beta^T$.
for $t = 0, 1, \cdots, T-1$ do
    $\widetilde{Z}^{t+1} = Z^t - \eta_Z \nabla_Z h(Z,\alpha,\beta) = Z^t + 2\eta_Z\left(A - \sigma(\Theta^t)\right)Z^t$;
    $\widetilde{\alpha}^{t+1} = \alpha^t - \eta_\alpha \nabla_\alpha h(Z,\alpha,\beta) = \alpha^t + 2\eta_\alpha(A - \sigma(\Theta^t))1_n$;
    $\widetilde{\beta}^{t+1} = \beta^t - \eta_\beta \nabla_\beta h(Z,\alpha,\beta) = \beta^t + \eta_\beta\langle A - \sigma(\Theta^t), X\rangle$;
    $Z^{t+1} = \mathcal{P}_{\mathcal{C}_Z}(\widetilde{Z}^{t+1}),\ \alpha^{t+1} = \mathcal{P}_{\mathcal{C}_\alpha}(\widetilde{\alpha}^{t+1}),\ \beta^{t+1} = \mathcal{P}_{\mathcal{C}_\beta}(\widetilde{\beta}^{t+1})$;
end for
```

After initialization, Algorithm 6 iteratively updates the estimates for the three parameters, namely Z, α and β. In each iteration, for each parameter, the algorithm first descend along the gradient direction by a pre-specified step size. The descent step is then followed by an additional projection step which projects the updated estimate to a pre-specified constraint set. The details on the step sizes and the constraint sets will be given in the statement of Theorem 6.

For each iteration, the update on the latent part is performed in the space of Z (that is $\mathbb{R}^{n\times k}$) rather than the space of all $n \times n$ Gram matrices as was required in the convex approach. In this way, it reduces the computational cost per iteration from $O(n^3)$ to $O(n^2k)$. Since we are most interested in cases where $k \ll n$, such a reduction leads to improved scalability of the non-convex approach to large networks. To implement this non-convex algorithm, we need to assume the knowledge of the latent space dimension k, which was not needed for the convex program (4.11). We defer the discussion on the data-driven choice of k to Section 4.6.

We note that Algorithm 6 is not guaranteed to find any global minimizer, or even any local minimizer, of the objective function (4.12). However, as we shall show next, under appropriate conditions, the estimates generated by the algorithm will quickly enter a neighborhood of the true parameters $(Z_\star, \alpha_\star, \beta_\star)$ and any element in this neighborhood is statistically at least as good as the estimator obtained from the convex method (4.11). This approach has close connection to the investigation of various non-convex methods for other statistical and signal processing applications. See for instance Candès et al. (2015), Chen and Wainwright (2015) and the references therein. In what follows, we first characterize statistical accuracy of the outputs under certain conditions on the initializers of the algorithm. Then we discuss how to construct initializers by which such conditions are satisfied.

Theoretical guarantees. We now investigate the statistical properties of the outputs of Algorithm 6. A key step is to characterize the evolution of the iterates.

As a first step, we introduce an error metric that is equivalent to $\|\Delta_{\Theta^t}\|_{\mathrm{F}}^2 = \|\Theta^t - \Theta_\star\|_{\mathrm{F}}^2$ while at the same time is more convenient for establishing an inequality satisfied by all iterates. Note that the latent vectors are only identifiable up to an orthogonal transformation of $\mathbb{R}^k$, for any $Z_1, Z_2 \in \mathbb{R}^{n\times k}$, we define the distance measure

$$\mathrm{dist}(Z_1, Z_2) = \min_{R\in O(k)} \|Z_1 - Z_2 R\|_{\mathrm{F}}$$

where $O(k)$ collects all $k \times k$ orthogonal matrices. Let $R^t = \arg\min_{R\in O(k)} \left\|Z^t - Z_\star R\right\|_{\mathrm{F}}$ and $\Delta_{Z^t} = Z^t - Z_\star R^t$, and further let $\Delta_{\alpha^t} = \alpha^t - \alpha_\star$, $\Delta_{G^t} = Z^t (Z^t)^\top - Z_\star Z_\star^\top$ and $\Delta_{\beta^t} = \beta^t - \beta_\star$. Then the error metric we use is

$$e_t = \|Z_\star\|_{\mathrm{op}}^2 \|\Delta_{Z^t}\|_{\mathrm{F}}^2 + 2\left\|\Delta_{\alpha^t} 1_n{}^\top\right\|_{\mathrm{F}}^2 + \left\|\Delta_{\beta^t} X\right\|_{\mathrm{F}}^2. \tag{4.13}$$

Let $\kappa_{Z_\star}$ be the condition number of $Z_\star$ (i.e., the ratio of the largest to the smallest singular values). The following lemma shows that the two error metrics e_t and Δ_{Θ^t} are equivalent up to a constant multiple of $\kappa_{Z_\star}^2$.

Lemma 4.4. *Under Assumption 4.3.1, there exists constant* $0 \le c_0 < 1$ *such that*

$$e_t \le \frac{\kappa_{Z_\star}^2}{2(\sqrt{2}-1)} \|\Delta_{G^t}\|_{\mathrm{F}}^2 + 2\left\|\Delta_{\alpha^t} 1_n{}^\top\right\|_{\mathrm{F}}^2 + \left\|\Delta_{\beta^t} X\right\|_{\mathrm{F}}^2 \le \frac{\kappa_{Z_\star}^2}{2(\sqrt{2}-1)(1-c_0)} \|\Delta_{\Theta^t}\|_{\mathrm{F}}^2.$$

Moreover, if dist$(Z^t, Z_\star) \le c\|Z_\star\|_{\mathrm{op}}$,

$$e_t \ge \frac{1}{(c+2)^2} \|\Delta_{G^t}\|_{\mathrm{F}}^2 + 2\left\|\Delta_{\alpha^t} 1_n{}^\top\right\|_{\mathrm{F}}^2 + \left\|\Delta_{\beta^t} X\right\|_{\mathrm{F}}^2 \ge \frac{1}{(c+2)^2(1+c_0)} \|\Delta_{\Theta^t}\|_{\mathrm{F}}^2.$$

Next, we focus on the following class of models on which our theoretical result holds:

$$\begin{aligned}\mathcal{F}_0(n,k,M_1,M_2,X) = \Big\{\Theta | \Theta = \alpha 1_n{}^\top + 1_n \alpha^\top + \beta X + ZZ^\top,\ JZ = Z,& \\ \max_{1\le i\le n} \|Z_{i*}\|^2,\ \|\alpha\|_\infty,\ |\beta| \max_{1\le i<j\le n} |X_{ij}| \le M_1/3,& \\ \max_{1\le i\ne j\le n} \Theta_{ij} \le -M_2\Big\}.&\end{aligned} \tag{4.14}$$

By the triangle inequality, we have $\mathcal{F}_0(n,k,M_1,M_2,X) \subset \mathcal{F}(n,k,M_1,M_2,X)$ where the latter was defined in (4.8). In other words, this is a slightly more restrictive class than that we have considered for the convex method.

Furthermore, our theorem depends on the following condition on the initializers.

Assumption 4.3.2. The initializers Z^0, α^0, β^0 in Algorithm 6 satisfy $e_0 \leq ce^{-2M_1} \|Z_\star\|_{\text{op}}^4 / \kappa_{Z_\star}^4$ for a sufficiently small positive constant c.

The following theorem states that the error sequence converges linearly till it reaches the desired statistical precision.

Theorem 6. *Let Assumptions 4.3.1 and 4.3.2 be satisfied. Set the constraint sets as*

$$\mathcal{C}_Z = \{Z \in \mathbb{R}^{n \times k}, JZ = Z, \max_{1 \leq i \leq n} \|Z_{i*}\| \leq M_1/3\},$$

$$\mathcal{C}_\alpha = \{\alpha \in \mathbb{R}^n, \|\alpha\|_\infty \leq M_1/3\},\ \mathcal{C}_\beta = \{\beta \in \mathbb{R}, \beta\|X\|_\infty \leq M_1/3\},$$

and the step sizes as $\eta_Z = \eta / \left\|Z^0\right\|_{\text{op}}^2, \eta_\alpha = \eta/(2n), \eta_\beta = \eta/(2\|X\|_{\text{F}}^2)$ *for any* $\eta \leq c$ *where* c *is a universal positive constant. Let* $\zeta_n = \max\{2\|A-P\|_{\text{op}},\ |\langle A-P, X/\|X\|_{\text{F}}\rangle|/\sqrt{k},\ 1\}$. *Then we have*

- Deterministic errors of iterates: *if* $\|Z_\star\|_{\text{op}}^2 \geq C_1 \kappa_{Z_\star}^2 e^{M_1} \zeta_n^2 \times \max\left\{\sqrt{\eta k e^{M_1}}, 1\right\}$ *for a sufficiently large constant* C_1*, there exist positive constants* ρ *and* C *such that*

$$e_t \leq 2\left(1 - \frac{\eta}{e^{M_1}\kappa_{Z_\star}^2}\rho\right)^t e_0 + \frac{C\kappa_{Z_\star}^2}{\rho} e^{2M_1}\zeta_n^2 k.$$

- Probabilistic errors of iterates: *if* $\|Z_\star\|_{\text{op}}^2 \geq C_1 \kappa_{Z_\star}^2 \sqrt{n} e^{M_1 - M_2/2} \max\left\{\sqrt{\eta k e^{M_1}}, 1\right\}$ *for a sufficiently large constant* C_1*, there exist positive constants* ρ, c_0 *and* C *such that with probability at least* $1 - n^{-c_0}$,

$$e_t \leq 2\left(1 - \frac{\eta}{e^{M_1}\kappa_{Z_\star}^2}\rho\right)^t e_0 + \frac{C\kappa_{Z_\star}^2}{\rho} e^{2M_1 - M_2} nk \times \max\left\{1, \frac{e^{M_2}\log n}{n}\right\}.$$

For any $T > T_0 = \log\left(\frac{M_1^2}{\kappa_{Z_\star}^2 e^{4M_1 - M_2}}\frac{n}{k^2}\right) / \log\left(1 - \frac{\eta}{e^{M_1}\kappa_{Z_\star}^2}\rho\right)$,

$$\|\Delta_{G^T}\|_{\mathrm{F}}^2,\ \|\Delta_{\Theta^T}\|_{\mathrm{F}}^2 \le C'\kappa_{Z_\star}^2 e^{2M_1 - M_2} nk \times \max\left\{1, \frac{e^{M_2}\log n}{n}\right\}$$

for some constant C'.

Remark 4.5. When both M_1 and M_2 are constants and the covariate matrix X is absent, the result in Section 4.5 of Chen and Wainwright (2015), in particular Corollary 5, implies the error rate of $O(nk)$ in Theorem 6. However, when $M_1 \to \infty$ and M_2 remains bounded as $n \to \infty$, the error rate in Chen and Wainwright (2015) becomes[1] $O(e^{8M_1}M_1^2 nk)$, which can be much larger than the rate $O(e^{2M_1}nk)$ given by Theorem 6 even when X is still absent. In addition, Algorithm 6 enjoys nice theoretical guarantees on its performance even when the model is mis-specified and the Θ matrix is only approximately low rank. See Theorem 9 below. These important cases are not covered by the general theory in Chen and Wainwright (2015).

Remark 4.6. In view of Lemma 4.4, the rate obtained by the non-convex approach in terms of $\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2$ matches the upper bound achieved by the convex method, up to a multiple of $\kappa_{Z_\star}^2$. As suggested by Lemma 4.4, the extra factor comes partly from the fact that e_t is a stronger loss function than $\|\Delta_{\Theta^t}\|_{\mathrm{F}}^2$ and in the worst case can be $c\kappa_{Z_\star}^2$ times larger than $\|\Delta_{\Theta^t}\|_{\mathrm{F}}^2$.

Remark 4.7. Under the setup in Theorem 6, the projection steps for α, β in Algorithm 6 are straightforward and have the following closed form expressions:

$$\alpha_i^{t+1} = \widetilde{\alpha}_i^{t+1}\min\left(1, \frac{M_1}{3|\widetilde{\alpha}_i^{t+1}|}\right),\ \beta^{t+1} = \widetilde{\beta}^{t+1}\min\left(1, \frac{M_1}{3|\widetilde{\beta}^{t+1}|\max_{i,j}|X_{ij}|}\right).$$

The projection step for Z is slightly more involved. Notice that $\mathcal{C}_Z = \mathcal{C}_Z^1 \bigcap \mathcal{C}_Z^2$ where

$$\mathcal{C}_Z^1 = \{Z \in \mathbb{R}^{n\times k}, JZ = Z\},\ \mathcal{C}_Z^2 = \{Z \in \mathbb{R}^{n\times k}, \max_{1\le i\le n}\|Z_{i*}\|^2 \le M_1/3\}.$$

[1] One can verify that in this case we can identify the quantities in Corollary 5 of Chen and Wainwright (2015) as $\sigma = 1$, $p = 1$, $d = n$, $r = k$, $\nu \asymp M_1$, $L_{4\nu} \asymp 1$ and $\ell_{4\nu} \asymp e^{4M_1}$.

Projecting to either of them has closed form solution, that is

$$\mathcal{P}_{\mathcal{C}_Z^1}(Z) = JZ, \ \ \left[\mathcal{P}_{\mathcal{C}_Z^2}(Z)\right]_{i*} = Z_{i*} \min\left(1, \sqrt{\frac{M_1}{3\|Z_{i*}\|^2}}\right).$$

Then Dykstra's projection algorithm (Dykstra, 1983) (or alternating projection algorithm) can be applied to obtain $\mathcal{P}_{\mathcal{C}_Z}(\widetilde{Z}^{t+1})$. We note that projections induced by the boundedness constraints for Z, α, β are needed for establishing the error bounds theoretically. However, when implementing the algorithm, users are at liberty to drop these projections and to only center the columns of the Z iterates. We did not see any noticeable difference on simulated examples caused by dropping them.

4.3.3. Initialization

Assumption 4.3.2 plays a key role in obtaining the desired error rates in Theorem 6. We now present two ways to initialize Algorithm 6 so that Assumption 4.3.2 can be satisfied under different circumstances.

Initialization by projected gradient descent in the lifted space The first initialization method is summarized in Algorithm 7, which is essentially running the projected gradient descent algorithm on the following regularized objective function for a small number of steps:

$$f(G, \alpha, \beta) = -\sum_{i,j} \{A_{ij}\Theta_{ij} + \log(1 - \sigma(\Theta_{ij}))\} + \lambda_n \operatorname{tr}(G) + \frac{\gamma_n}{2}\left(\|G\|_{\mathrm{F}}^2 + 2\left\|\alpha 1_n{}^\top\right\|_{\mathrm{F}}^2 + \|X\beta\|_{\mathrm{F}}^2 \right).$$

Except for the third term, this is the same as the objective function in (4.11). However, the inclusion of the additional proximal term ensures that Assumption 4.3.2 can be satisfied after a small number of projected gradient descent steps.

We further assume the strength of the latent effect $\|G_\star\|_{\mathrm{F}}$ is comparable to the strength of the degree heterogeneity effect $\left\|\alpha_\star 1_n{}^\top\right\|_{\mathrm{F}}$ and the homophily effect $\|X\beta_\star\|_{\mathrm{F}}$.

Theorem 7. *Suppose that Assumption 4.3.1 holds and that* $\|\alpha_\star 1_n{}^\top\|_{\mathrm{F}}, \|\beta_\star X\|_{\mathrm{F}} \le C\|G_\star\|_{\mathrm{F}}$

Algorithm 7 Initialization of Algorithm 6 by Projected Gradient Descent

```
Input: Adjacency matrix: $A$; covariate matrix $X$; initial values: $G^0 = 0, \alpha^0 = 0, \beta^0 = 0$; step size: $\eta$; constraint set: $\mathcal{C}_G, \mathcal{C}_\alpha, \mathcal{C}_\beta$; regularization parameter: $\lambda_n, \gamma_n$; latent dimension: $k$; number of steps: T.
for $t = 1, 2, \cdots, \mathrm{T}$ do
    $\widetilde{G}^{t+1} = G^t - \eta \nabla_Z f(Z, \alpha, \beta) = G^t + \eta \left( A - \sigma(\Theta^t) - \lambda_n I_n - \gamma_n G^t \right)$
    $\widetilde{\alpha}^{t+1} = \alpha^t - \eta \nabla_\alpha f(Z, \alpha, \beta)/n = \alpha^t + \eta \left( (A - \sigma(\Theta^t)) 1_n / 2n - \gamma_n \alpha^t \right)$
    $\widetilde{\beta}^{t+1} = \beta^t - \eta \nabla_\beta f(Z, \alpha, \beta) / \|X\|_{\mathrm{F}}^2 = \beta^t + \eta \left( \langle A - \sigma(\Theta^t), X \rangle / \|X\|_{\mathrm{F}}^2 - \gamma_n \beta^t \right)$
    $G^{t+1} = \mathcal{P}_{\mathcal{C}_G}(\widetilde{G}^{t+1}),\ \alpha^{t+1} = \mathcal{P}_{\mathcal{C}_\alpha}(\widetilde{\alpha}^{t+1}),\ \beta^{t+1} = \mathcal{P}_{\mathcal{C}_\beta}(\widetilde{\beta}^{t+1})$
end for
Top-$k$ eigen-decomposition: $G^{\mathrm{T}} \approx U_k D_k U_k^\top$. Set $Z^{\mathrm{T}} = U_k D_k^{1/2}$
Ouput: $Z^{\mathrm{T}}, \alpha^{\mathrm{T}}, \beta^{\mathrm{T}}$
```

for a numeric constant $C > 0$. *Let* λ_n *satisfy* $C_0 \sqrt{\max\{ne^{-M_2}, \log n\}} \leq \lambda_n \leq c_0 \|G_\star\|_{\mathrm{op}} / (e^{2M_1} \sqrt{k} \kappa_{Z_\star}^3)$ *for sufficiently large constant* C_0 *and sufficiently small constant* c_0*, let* γ_n *satisfy* $\gamma_n \leq \delta \lambda_n / \|G_\star\|_{\mathrm{op}}$ *for sufficiently small constant* δ*. Choose step size* $\eta \leq 2/9$ *and set the constraint sets as*

$$\mathcal{C}_G = \{G \in \mathbb{S}_+^{n \times n}, JG = G, \max_{1 \leq i,j \leq n} |G_{ij}| \leq M_1/3\},$$

$$\mathcal{C}_\alpha = \{\alpha \in \mathbb{R}^n, \|\alpha\|_\infty \leq M_1/3\},\ \mathcal{C}_\beta = \{\beta \in \mathbb{R}, \beta \|X\|_\infty \leq M_1/3\}.$$

If the latent vectors contain strong enough signal in the sense that

$$\|G_\star\|_{\mathrm{op}}^2 \geq C \kappa_{Z_\star}^6 e^{4M_1 - M_2} nk \times \max\left\{1, \frac{e^{M_2} \log n}{n}\right\}, \tag{4.15}$$

for some sufficiently large constant C*, there exist positive constants* c, C_1 *such that with probability at least* $1 - n^{-c}$*, for any given constant* $c_1 > 0$*,* $e_T \leq c_1^2 e^{-2M_1} \|Z_\star\|_{\mathrm{op}}^4 / \kappa_{Z_\star}^4$ *as long as* $T \geq T_0$*, where*

$$T_0 = \log\left(\frac{C_1 e^{2M_1} k \kappa_{Z_\star}^6}{c_1^2}\right) \left(\log\left(\frac{1}{1 - \gamma_n \eta}\right)\right)^{-1}. \tag{4.16}$$

Theorem 7 gives the range of λ_n and γ_n such that implementing Algorithm 7 with $T \geq T_0$ and using the output as the initializers for Algorithm 6, the condition on e_0 in Assumption 4.3.2 will be satisfied. To go one step further, the following corollary characterizes the ideal choices of γ_n and λ_n in Algorithm 7. It is worth noting that the choice of λ_n here does not coincide with that in Theorem 5. Interestingly, the corollary shows that when M_1, k and $\kappa_{Z_\star}$ are all upper bounded by universal constants, for appropriate choices of γ_n and λ_n in Algorithm 7, the number of iterations needed does not depend on the graph size n.

Corollary 4.8. *Specifically in Theorem 7, if we choose $\gamma_n = \gamma = c_0/(e^{2M_1}\sqrt{k}\kappa_{Z_\star}^3)$ for some sufficiently small constant c_0, and $\lambda_n = C_0\gamma_n \left\|G_\star\right\|_{\mathrm{op}}$ for some sufficiently large constant C_0, there exist positive constants c, C_1 such that with probability at least $1 - n^{-c}$, for any given constant $c_1 > 0$, $e_T \leq c_1^2 e^{-2M_1} \left\|Z_\star\right\|_{\mathrm{op}}^4 / \kappa_{Z_\star}^4$ as long as $T \geq T_0$, where*

$$T_0 = \log\left(\frac{C_1 e^{2M_1} k \kappa_{Z_\star}^6}{c_1^2}\right)\left(\log\left(\frac{1}{1-\gamma\eta}\right)\right)^{-1}. \tag{4.17}$$

Remark 4.9. Similar to computing $\mathcal{P}_{\mathcal{C}_Z}(\cdot)$ in Algorithm 6, $\mathcal{P}_{\mathcal{C}_G}(\cdot)$ could also be implemented by Dykstra's projection algorithm since $\mathcal{C}_G$ is the intersection of two convex sets. The boundedness constraint $\max_{i,j}|G_{ij}| \leq M/3$ is only for the purpose of proof. In practice, if ignoring this constraint, G_{t+1} will have closed form solution $G_{t+1} = \mathcal{P}_{\mathbb{S}_+^n}(J\widetilde{G}_{t+1}J)$ where $\mathcal{P}_{\mathbb{S}_+^n}(\cdot)$ can be computed by singular value thresholding.

Initialization by universal singular value thresholding Another way to construct the initialization is to first estimate the probability matrix P by universal singular value thresholding (USVT) proposed by Chatterjee (2015) and then recover the initial estimates of α, Z, β heuristically by inverting the logit transform. The procedure is summarized in Algorithm 8.

The estimate of P by USVT is consistent when $\|P\|_*$ is "small". Following the arguments

Algorithm 8 Initialization of Algorithm 6 by Singular Value Thresholding

```
**Input:** Adjacency matrix: $A$; covariate matrix $X$; latent dimension $k$; threshold $\tau$.
  1. Singular value thresholding: $\widetilde{P} = \sum_{\sigma_i \geq \tau} \sigma_i u_i v_i^\top$ where $A = \sum_{i=1}^n \sigma_i u_i v_i^\top$ is the singular value decomposition. Elementwisely project $\widetilde{P}$ to the interval $\left[\frac{1}{2}e^{-M_1}, \frac{1}{2}\right]$ to obtain $\widehat{P}$. Estimate the logit matrix by $\widehat{\Theta} = \operatorname{logit}((\widehat{P} + \widehat{P}^\top)/2)$.
  2. $\alpha^0, \beta^0 = \arg\min_{\alpha,\beta} \left\| \widehat{\Theta} - \left(\alpha 1_n^\top + 1_n \alpha^\top + \beta X\right) \right\|_{\mathrm{F}}^2$
  3. $\widehat{G} = \mathcal{P}_{\mathbb{S}_+^n}(R)$ where $R = J\left(\widehat{\Theta} - \left(\alpha^0 1_n^\top + 1_n(\alpha^0)^\top + \beta^0 X\right)\right) J$
  4. $Z^0 = U_k D_k^{1/2}$ where $\widehat{G} \approx U_k D_k U_k^\top$ is the top-$k$ singular value decomposition.
**Output:** $\alpha^0, Z^0, \beta^0$
```

in Theorems 2.6 and 2.7 of Chatterjee (2015), such condition is satisfied when the covariate matrix $X = 0$ or when X has "simple" structure. Such "simple" structure could be $X_{ij} = f(x_i, x_j)$ where $x_1, \cdots, x_n \in \mathbb{R}^d$ are feature vectors associated with the n nodes and $f(\cdot,\cdot)$ characterizes the distance/similarity between node i and node j. For instance, one could have $X_{ij} = \mathbf{1}_{\{x_i = x_j\}}$ where $x_1, \cdots, x_n \in \{1, \cdots, K\}$ is a categorical variable such as gender, race, nationality, etc; or $X_{ij} = g(|x_i - x_j|)$ where $g(\cdot)$ is a continuous monotone link function and $x_1, \cdots, x_n \in \mathbb{R}$ is a continuous variable such as age, income, years of education, etc.

Remark 4.10. The least squares problem in step 2 of Algorithm 8 has closed form solution and can be computed in $O(n^2)$ operations. The computational cost of Algorithm 8 is dominated by matrix decompositions in step 1 and step 3.

In particular, the following proposition shows that for large enough network with no edge covariates included in the latent space model, the α^0 and Z^0 generated by Algorithm 8 satisfy the initialization condition specified in Assumption 4.3.2 with high probability.

Proposition 4.11. *If no covariates are included in the latent space model and $\|G_\star\|_{\mathrm{F}} \geq c_0 n$ for some numeric constant $c_0 > 0$, then there exists constant c_1 such that with probability at least $1 - n^{c_1}$, for any $n \geq C(k, M_1, \kappa_{Z_\star})$ where $C(k, M_1, \kappa_{Z_\star})$ is a constant depending on k, M_1 and $\kappa_{Z_\star}$, the outputs of Algorithm 8 with $\tau \geq 1.1\sqrt{n}$ satisfies the initialization condition in Assumption 4.3.2.*

4.3.4. Results for general models

Following the introduction in Section 4.2.2, we consider the following parameter space for the more general class of latent space models

$$\begin{aligned}\mathcal{F}_g(n, M_1, M_2, X) = \Big\{\Theta | \Theta = \alpha 1_n{}^\top + 1_n \alpha^\top + \beta X + G, G \in \mathbb{S}_+^n, JG = G,& \\ \max_{1\le i\le n} G_{ii},\ \|\alpha\|_\infty,\ |\beta| \max_{1\le i<j\le n} |X_{ij}| \le M_1/3,& \\ \max_{1\le i\neq j\le n} \Theta_{ij} \le -M_2\Big\}.&\end{aligned} \tag{4.18}$$

Note that the latent space dimension k is no longer a parameter in (4.18). Then for any positive integer k, let $U_k D_k U_k^\top$ be the best rank-k approximation to $G_\star$. In this case, with slight abuse of notation, we let

$$Z_\star = U_k D_k^{1/2} \quad \text{and} \quad \overline{G}_k = G_\star - U_k D_k U_k^\top.$$

Performance of the penalized MLE method The following theorem is a generalization of Theorem 5 to the general class.

Theorem 8. *For any $k \in \mathbb{N}_+$ such that Assumption 4.3.1 holds and any λ_n satisfying $\lambda_n \ge \max\{2\,\|A-P\|_{\mathrm{op}},\ |\langle A-P, X/\|X\|_{\mathrm{F}}\rangle|/\sqrt{k},\ 1\}$, there exists a constant C such that*

$$\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \le C\left(e^{2M_1}\lambda_n^2 k + e^{M_1}\lambda_n \|\overline{G}_k\|_*\right).$$

Specifically, setting $\lambda_n = C_0\sqrt{\max\{ne^{-M_2}, \log n\}}$ for a large enough constant C_0, there exists positive constants c, C such that with probability at least $1-n^{-c}$,

$$\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \le C\left(e^{2M_1-M_2} nk \times \max\left\{1, \frac{e^{M_2}\log n}{n}\right\} + e^{M_1-M_2/2}\sqrt{n}\|\overline{G}_k\|_*\right). \tag{4.19}$$

We continue using the error metric e_t defined in equation (4.37).

Theorem 9. *Under Assumption 4.3.1, 4.3.2, set the constraint sets as*

$$\mathcal{C}_Z = \{Z \in \mathbb{R}^{n\times k}, JZ = Z, \max_{1\le i\le n} \|Z_{i*}\| \le M_1/3\},$$

$$\mathcal{C}_\alpha = \{\alpha \in \mathbb{R}^n, \|\alpha\|_\infty \le M_1/3\},\ \mathcal{C}_\beta = \{\beta \in \mathbb{R}, \beta\|X\|_\infty \le M_1/3\}.$$

and choose step sizes by $\eta_Z = \eta/\left\|Z^0\right\|_{\mathrm{op}}^2, \eta_\alpha = \eta/(2n), \eta_\beta = \eta/(2\left\|X\right\|_{\mathrm{F}}^2)$ *for any* $\eta \le c$ *where* c *is some positive constant. Let* $\zeta_n = \max\{2\left\|A-P\right\|_{\mathrm{op}},\ |\langle A-P, X/\left\|X\right\|_{\mathrm{F}}\rangle|/\sqrt{k},\ 1\}$.

- *If* $\|G_\star\|_{\mathrm{op}} \ge C_1\kappa_{Z_\star}^2 e^{M_1}\zeta_n^2 \times \max\left\{\sqrt{\eta k e^{M_1}}, \sqrt{\eta\left\|\overline{G}_k\right\|_{\mathrm{F}}^2/\zeta_n^2}, 1\right\}$, *there exist positive constants* ρ *and* C *such that*

  $$e_t \le 2\left(1 - \frac{\eta}{e^{M_1}\kappa_{Z_\star}^2}\rho\right)^t e_0 + \frac{C\kappa_{Z_\star}^2}{\rho}\left(e^{2M_1}\zeta_n^2 k + e^{M_1}\left\|\overline{G}_k\right\|_{\mathrm{F}}^2\right).$$

- *If* $\|G_\star\|_{\mathrm{op}} \ge C_1\kappa_{Z_\star}^2\sqrt{n}e^{M_1-M_2/2}\max\left\{\sqrt{\eta k e^{M_1}}, \sqrt{\eta\left\|\overline{G}_k\right\|_{\mathrm{F}}^2/\zeta_n^2}, 1\right\}$ *for a sufficiently large constant* C_1, *there exist positive constants* ρ, c_0 *and* C *such that with probability at least* $1 - n^{-c_0}$, *the iterates generated by Algorithm 6 satisfying*

  $$e_t \le 2\left(1 - \frac{\eta}{e^{M_1}\kappa_{Z_\star}^2}\rho\right)^t e_0 + \frac{C\kappa_{Z_\star}^2}{\rho}\left(e^{2M_1-M_2}nk \times \max\left\{1, \frac{e^{M_2}\log n}{n}\right\} + e^{M_1}\left\|\overline{G}_k\right\|_{\mathrm{F}}^2\right).$$

  For any $T > T_0 = \log\left(\frac{M_1^2}{\kappa_{Z_\star}^2 e^{4M_1-M_2}}\frac{n}{k^2}\right)/\log\left(1 - \frac{\eta}{e^{M_1}\kappa_{Z_\star}^2}\rho\right)$,

  $$\|\Delta_{G^T}\|_{\mathrm{F}}^2,\ \|\Delta_{\Theta^T}\|_{\mathrm{F}}^2 \le C'\kappa_{Z_\star}^2\left(e^{2M_1-M_2}nk \times \max\left\{1, \frac{e^{M_2}\log n}{n}\right\} + e^{M_1}\left\|\overline{G}_k\right\|_{\mathrm{F}}^2\right)$$

  for some constant C'.

Theorem 10. *Suppose that Assumption 4.3.1 holds and that* $\|\alpha_\star 1_n{}^\top\|_{\mathrm{F}}, \|\beta_\star X\|_{\mathrm{F}} \le C\|G_\star\|_{\mathrm{F}}$ *for a numeric constant* $C > 0$. *Let* λ_n *satisfy* $C_0\sqrt{\max\{ne^{-M_2}, \log n\}} \le \lambda_n \le c_0\left\|G_\star\right\|_{\mathrm{op}}/(e^{2M_1}\sqrt{k}\kappa_{Z_\star}^3)$ *for sufficiently large constant* C_0 *and sufficiently small constant* c_0, *let* γ_n *satisfy* $\gamma_n \le \delta\lambda_n/\left\|G_\star\right\|_{\mathrm{op}}$ *for sufficiently small constant* δ. *Choose step size*

$\eta \leq 2/9$ and set the constraint sets as

$$\mathcal{C}_G = \{G \in \mathbb{S}_+^{n\times n}, JG = G, \max_{1\leq i,j\leq n} |G_{ij}| \leq M_1/3\},$$

$$\mathcal{C}_\alpha = \{\alpha \in \mathbb{R}^n, \|\alpha\|_\infty \leq M_1/3\},\ \mathcal{C}_\beta = \{\beta \in \mathbb{R}, \beta\|X\|_\infty \leq M_1/3\}.$$

If the latent vectors contain strong enough signal in the sense that

$$\|G_\star\|_{\mathrm{op}}^2 \geq C\kappa_{Z_\star}^6 e^{2M_1} \max\left\{e^{2M_1-M_2} nk \times \max\left\{1, \frac{e^{M_2}\log n}{n}\right\},\ \|\overline{G}_k\|_*^2/k,\ \left\|\overline{G}_k\right\|_{\mathrm{F}}^2\right\},$$

for some sufficiently large constant C, there exist positive constants c, C_1 such that with probability at least $1 - n^{-c}$, for any given constant $c_1 > 0$, $e_T \leq c_1^2 e^{-2M_1} \|Z_\star\|_{\mathrm{op}}^4 / \kappa_{Z_\star}^4$ as long as $T \geq T_0$, where

$$T_0 = \log\left(\frac{C_1 e^{2M_1} k\kappa_{Z_\star}^6}{c_1^2}\right)\left(\log\left(\frac{1}{1-\gamma_n\eta}\right)\right)^{-1}.$$

4.4. Simulation Studies

In this section, we present simulation studies of three different aspects of the proposed methods: (1) scaling of estimation errors and computational costs with network sizes, (2) impact of initialization on Algorithm 6, and (3) performance of the methods on general models.

Estimation errors and computational costs We first investigate how estimation errors scale with network size. To this end, we fix $\beta_\star = -\sqrt{2}$ and for any $(n, k) \in \{500, 1000, 2000, 4000, 8000\} \times \{2, 4, 8\}$, we set the other model parameters randomly following these steps:

1. Generate the degree heterogeneity parameters: $(\alpha_\star)_i = -\alpha_i / \sum_{j=1}^n \alpha_j$ for $1 \leq i \leq n$, where $\alpha_1, \cdots, \alpha_n \overset{iid}{\sim} U[1, 3]$.

2. Generate $\mu_1,\ \mu_2 \in \mathbb{R}^k$ with coordinates iid following $U[-1,1]$ as two latent vector centers;

3. Generate latent vectors: for $i = 1, \ldots, k$, let $(z_1)_i, \cdots, (z_{\lfloor n/2 \rfloor})_i \overset{iid}{\sim} (\mu_1)_i + N_{[-2,2]}(0,1)$ and $(z_{\lfloor n/2 \rfloor + 1})_i, \cdots, (z_n)_i \overset{iid}{\sim} (\mu_2)_i + N_{[-2,2]}(0,1)$ where $N_{[-2,2]}(0,1)$ is the standard normal distribution restricted onto the interval $[-2,2]$, then set $Z_\star = JZ$ where $Z = [z_1, \cdots, z_n]^\top$ and J is as defined in (4.3). Finally, we normalize $Z_\star$ such that $\|G_\star\|_{\mathrm{F}} = n$;

4. Generate the covariate matrix: $X = n \times \widetilde{X}/\|\widetilde{X}\|_{\mathrm{F}}$ where $\widetilde{X}_{ij} \overset{iid}{\sim} \min\{|N(1,1)|, 2\}$.

For each generated model, we further generated 30 independent copies of the adjacency matrix for each model configuration. Unless otherwise specified, for all experiments in this section, with given (n,k), the model parameters are set randomly following the above four steps and algorithms are run on 30 independent copies of the adjacency matrix.

The results of the estimation error for varying (n,k) are summarized in the log-log boxplots in Figure 3, where "Relative Error - Z" is defined as $\|\widehat{Z}\widehat{Z}^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}/\|Z_\star Z_\star^\top\|_{\mathrm{F}}$ and "Relative Error - Θ" is defined as $\|\widehat{\Theta} - \Theta_\star\|_{\mathrm{F}}/\|\Theta_\star^\top\|_{\mathrm{F}}$. From the boxplots, for each fixed latent space dimension k, the estimation errors for both $Z_\star$ and $\Theta_\star$ scale at the order of $1/\sqrt{n}$. This agrees well with the theoretical results in Section 4.3. For different latent space dimension k, the error curve with respect to network size n only differs in the intercept.

Next we consider the running time of Algorithm 6 with $T = 100$. To this end, we fix $k = 4$ and vary n from 500 to 16000. The results are presented in the log-log scatterplot of Figure 4 which indicates the computational cost scales quadratically.

Impact of initialization on Algorithm 6 We now turn to the comparison of three different initialization methods for Algorithm 6: the convex method (Algorithm 7), singular value thresholding (Algorithm 8), and random initialization. To this end, we fixed $n = 4000, k = 4$. The relative estimation errors are summarized as boxplots in Figure 5. Clearly,

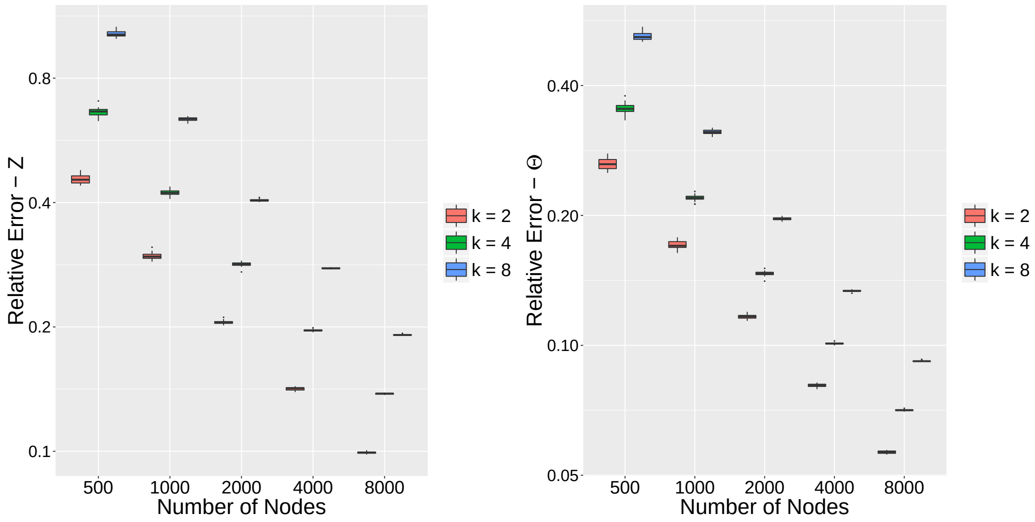

Figure 3: log-log boxplot for relative estimation errors with varying network size and latent space dimension.

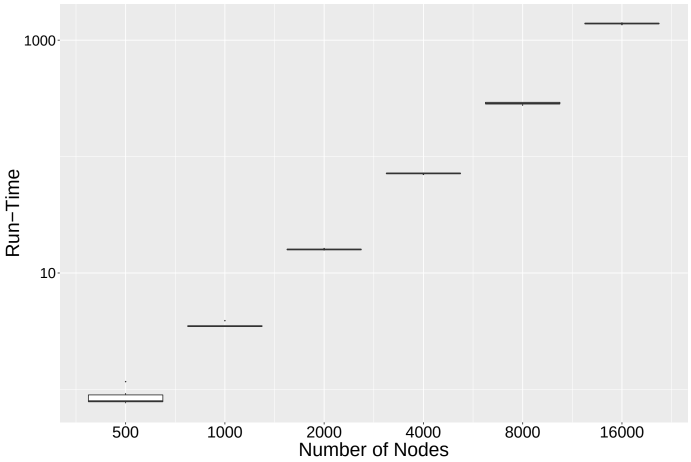

Figure 4: log-log plot for average run time with varying network size.

the non-convex algorithm is very robust to the initial estimates. Similar phenomenon is observed in real data analysis where different initializations yield nearly the same clustering accuracy.

Performance on the general model class Finally, to investigate the performance of the proposed method under the general model (4.4), we try two frequently used kernel functions, distance kernel $h_d(z_i, z_j) = -\|z_i - z_j\|$ and Gaussian kernel $h_g(z_i, z_j) = 4\exp(-\|z_i - z_j\|^2/9)$. In this part, we use d to represent the dimension of the latent vectors (that is, $z_1, \cdots, z_n \in$

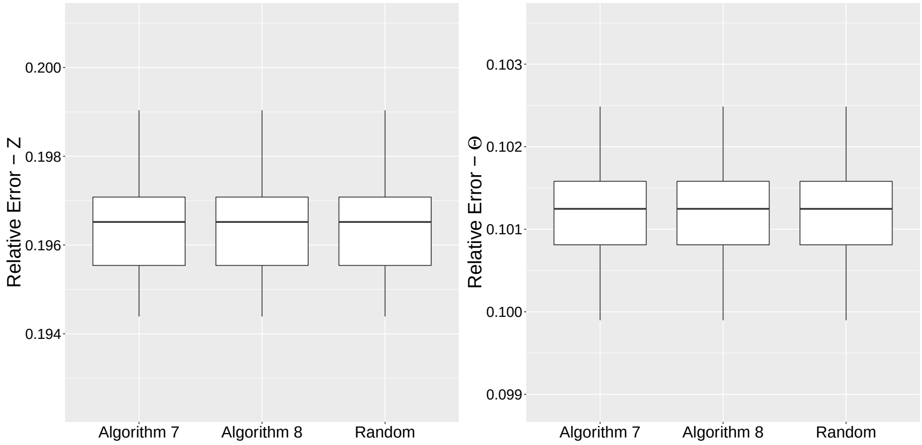

Figure 5: Boxplot for relative estimation error with different initialization methods.

$\mathbb{R}^d$) and k to represent the fitting dimension in Algorithm 6. We fix $d = 4$ and network size $n = 4000$. Model parameters are set randomly in the same manner as the four step procedure except that the third step is changed to:

> Generate latent vectors: for $i = 1, \ldots, d$, let $(z_1)_i, \cdots, (z_{\lfloor n/2 \rfloor})_i \overset{iid}{\sim} (\mu_1)_i + N_{[-2,2]}(0,1)$ and $(z_{\lfloor n/2 \rfloor+1})_i, \cdots, (z_n)_i \overset{iid}{\sim} (\mu_2)_i + N_{[-2,2]}(0,1)$ where $N_{[-2,2]}(0,1)$ is the standard normal distribution restricted onto the interval $[-2, 2]$. Finally for given kernel function $h(\cdot, \cdot)$, set $G_\star = JHJ$ where $H_{ij} = h(z_i, z_j)$.

We run both the convex approach and Algorithm 6 with different fitting dimensions. The boxplot for the relative estimation errors and the singular value decay of the kernel matrix under distance kernel and Gaussian kernel are summarized in Figure 6 and Figure 7 respectively.

As we can see, under the generalized model, the non-convex algorithm exhibits bias-variance tradeoff with respect to the fitting dimension, which dependens on the singular value decay of the kernel matrix. The advantage of the convex method is the adaptivity to the unknown kernel function.

As indicated by Theorem 9, the optimal choice of fitting dimension k should depend on the size of the network. To illustrate such dependency, we vary both network size and fitting

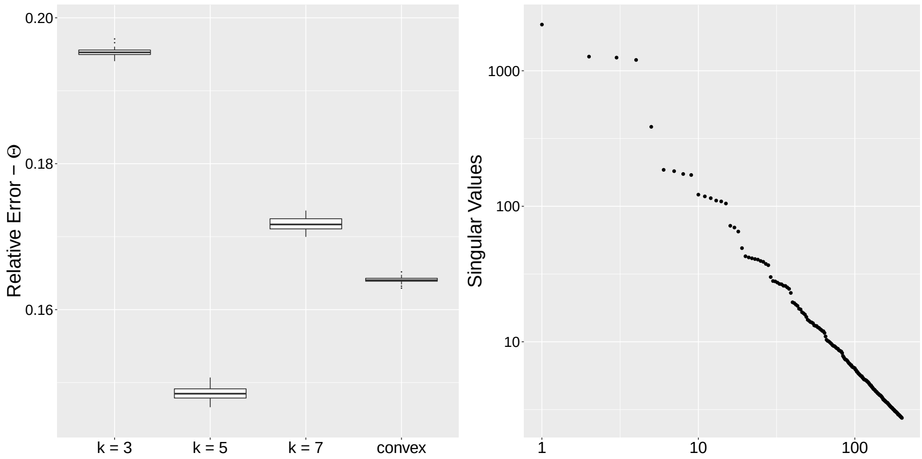

Figure 6: log-log plot for the relative estimation errors of both convex and non-convex approach under the distance kernel $h_d(z_i, z_j) = -\|z_i - z_j\|$.

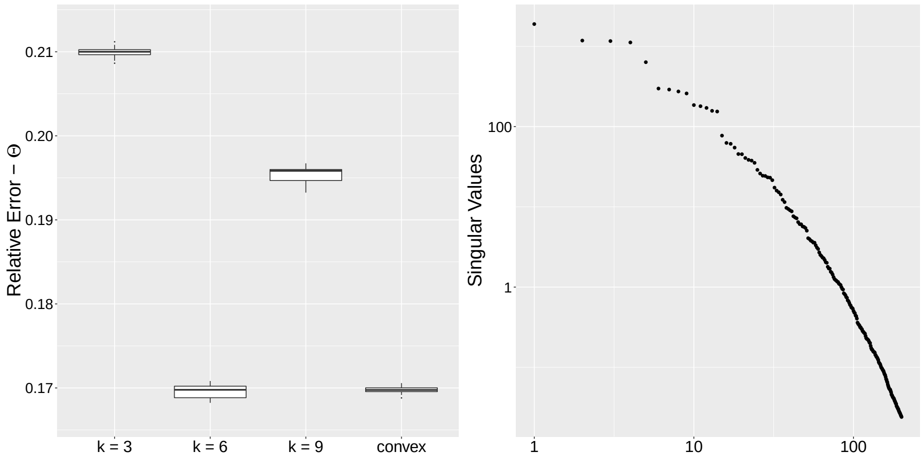

Figure 7: log-log plot for the relative estimation errors of both convex and non-convex approach under the Gaussian kernel $h_g(z_i, z_j) = 4\exp(-\|z_i - z_j\|^2/9)$

dimension, of which the results are summarized in Figure 8. As the size of the network increases, the optimal choice of fitting dimension increases as well.

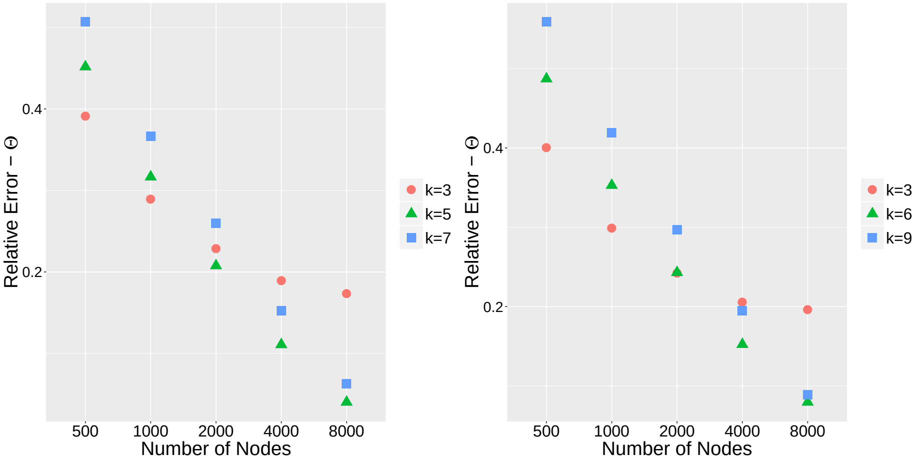

Figure 8: log-log plot for the relative estimation error with varying network size under distance kernel $h_d(z_i, z_j) = -\|z_i - z_j\|$ (left panel) and under the Gaussian kernel $h_g(z_i, z_j) = 4\exp(-\|z_i - z_j\|^2/9)$ (right panel).

4.5. Real Data Examples

In this section, we demonstrate how the model and fitting methods can be used to explore real world datasets that involve large networks. In view of the discussion in Section 4.3.4 and Section 4.4, we can always use the inner-product model (4.1) – (4.3) as our working model. In particular, we illustrate three different aspects. First, we consider community detection on networks without covariate. To this end, we compare the performance of simple k-means clustering on fitted latent variables with several state-of-the-art methods. Next, we investigate community detection on networks with covariates. In this case, we could still apply k-means clustering on fitted latent variables. Whether there is covariate or not, we can always visualize the network by plotting fitted latent variables in some appropriate way. Furthermore, we study how fitting the model can generate new feature variables to aid content-based classification of documents. The ability of feature generation also makes the model and the fitting methods potentially useful in other learning scenarios when additional network information is present.

4.5.1. Community detection without covariate

Community detection on networks without covariate has been intensively studied from both theoretical and methodological viewpoints. Thus, it naturally serves as a test example of the usefulness of the model and fitting methods we have proposed in previous sections. To adapt our method to community detection, we propose to partition the network nodes by the following two step procedure:

1. Fit the inner-product model to data with Algorithm 6;

2. Apply a simple k-means clustering on the fitted latent variables.

In what follows, we call this two step procedure LSCD (Latent Space based Community Detection). We shall compare it with three state-of-the-art methods:

- SCORE (Jin, 2015): a normalized spectral clustering method developed under DCBM;

- OCCAM (Zhang et al., 2014): a normalized and regularized spectral clustering method for potentially overlapping community detection;

- CMM (Chen et al., 2015): a convexified modularity maximization method developed under DCBM.

- Latentnet Krivitsky and Handcock: a hierachical bayesian method based on the latent space clustering model (Handcock et al., 2007).

To avoid biasing toward our own method, we compare these methods on three datasets that have been previously used in the original papers to justify the three methods at comparison: a political blog dataset (Adamic and Glance, 2005) that was studied in Jin (2015) and two Facebook datasets (friendship networks of Simmons College and Caltech) (Traud et al., 2012) that were studied in Chen et al. (2015). To make fair comparison, for all the methods, we supplied the true number of communities in each dataset. When fitting our model, we set the latent space dimension to be the same as the number of communities.

In the latentnet package Krivitsky and Handcock, there are three different ways to predict the community membership. In the notations of the package Krivitsky and Handcock, they are mkl$Z.K, mkl$mbc$Z.K and mle$Z.K respectively. We found that mkl$mbc$Z.K consistently outperforms the other two on these data examples. Due to the stochastic nature of the Bayesian approach, we repeated it 20 times and report the average performance as well as the standard deviation (numbers in parentheses).

Table 2 summarizes the performance of all four methods on the three datasets. Among all the methods at comparison, all methods performed well on the political blog dataset with Latentnet being the best, and LSCD outperformed all other methods on the two Facebook datasets. On the Caltech friendship dataset, it improved the best result out of the other three methods by almost 15% in terms of number of mis-clustered nodes.

Dataset	# Clusters	LSCD	SCORE	OCCAM	CMM	Latentnet
Political Blog	2	4.75%	4.75%	5.32%	5.07%	**4.51**% (0.12%)
Simmons	4	**11.79%**	23.57%	23.43%	12.04%	29.09% (1.23%)
Caltech	8	**17.97%**	31.02%	32.03%	21.02%	38.47% (1.19%)

Table 2: Proportions of mis-clustered nodes by different methods on three datasets.

In what follows, we provide more details on each dataset and on the performance of these community detection methods on them.

Political Blog This well-known dataset was recorded by Adamic and Glance (2005) during the 2004 U.S. Presidential Election. The original form is a directed network of hyperlinks between 1490 political blogs. The blogs were manually labeled as either liberal or conservative according to their political leanings. The labels were treated as true community memberships. Following the literature, we removed the direction information and focused on the largest connected component which contains 1222 nodes and 16714 edges. All five methods performed comparably on this dataset with Latentnet achieving the smallest mis-clustered proportion.

Simmons College The Simmons College Facebook network is an undirected graph that contains 1518 nodes and 32988 undirected edges. For comparison purpose, we followed the same pre-processing steps as in Chen et al. (2015) by considering the largest connected component of the students with graduation year between 2006 and 2009, which led to a subgraph of 1137 nodes and 24257 edges. It was observed in Traud et al. (2012) that the class year has the highest assortativity values among all available demographic characteristics, and so we treated the class year as the true community label. On this dataset, LSCD achieved the lowest mis-clustered proportion among these methods, with CMM a close second lowest.

An important advantage of model (4.1) is that it can provide a natural visualization of the network. To illustrate, the left panel of Figure 9 is a 3D visualization of the network with the first three coordinates of the estimated latent variables. From the plot, one can immediately see three big clusters: class year 2006 and 2007 combined (red), class year 2008 (green) and class year 2009 (blue). The right panel zooms into the cluster that includes class year 2006 and 2007 by projecting the the estimated four dimensional latent vectors onto a two dimensional discriminant subspace that was estimated from the fitted latent variables and the clustering results of LSCD. It turned out that class year 2006 and 2007 could also be reasonably distinguished by the latent vectors.

Caltech Data In contrast to the Simmons College network in which communities are formed according to class years, communities in the Caltech friendship network are formed according to dorms Traud et al. (2011, 2012). In particular, students spread across eight different dorms which we treated as true community labels. Following the same pre-processing steps as in Chen et al. (2015), we excluded the students whose residence information was missing and considered the largest connected component of the remaining graph, which contained 590 nodes and 12822 undirected edges. This dataset is more challenging than the Simmons College network. Not only the size of the network halves but the number of communities doubles. In some sense, it serves the purpose of testing

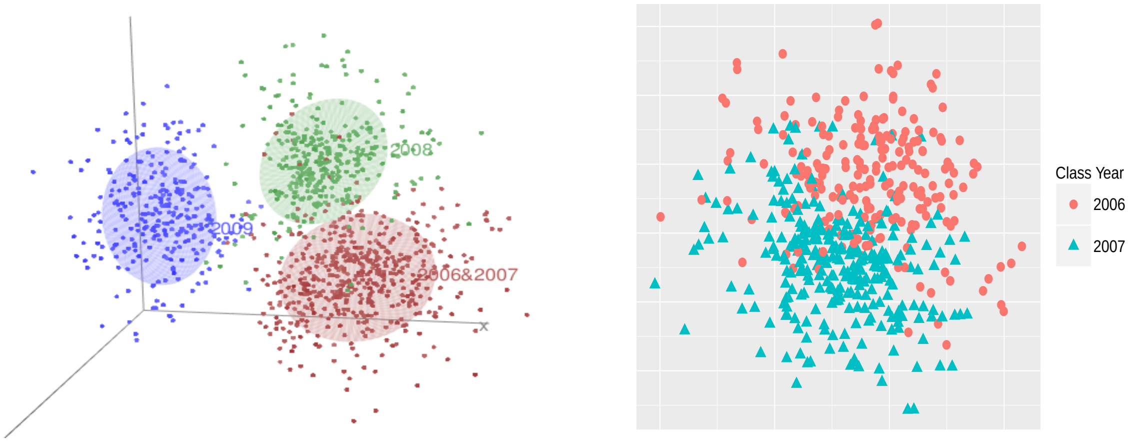

Figure 9: The left panel is a visualization of the network with the first three coordinates of the estimated latent vectors. The right panel is a visualization of students in class year 2006 and 2007 by projecting the four dimensional latent vectors to an estimated two dimensional discriminant subspace.

these methods when the signal is weak. LSCD also achieved the highest overall accuracy on this dataset, reducing the second best error rate (achieved by CMM) by nearly 15%. See the last row of Table 2. Moreover, LSCD achieved the lowest maximum community-wise misclustering error among the four methods. See Figure 10 on page 126 for a detailed comparison of community-wise misclustering rates of the five methods.

It is worth noting that the two spectral methods, SCORE and OCCAM, fell far behind on the two Facebook datasets. One possible explanation is that the structures of these Facebook networks are more complex than the political blog network and so DCBM suffers more under-fitting on them. In contrast, the latent space model (4.1) is more expressive and goes well beyond simple block structure. The Latentnet approach did not perform well on the Facebook datasets, either. One possible reason is the increased numbers of communities compared to the political blog dataset.

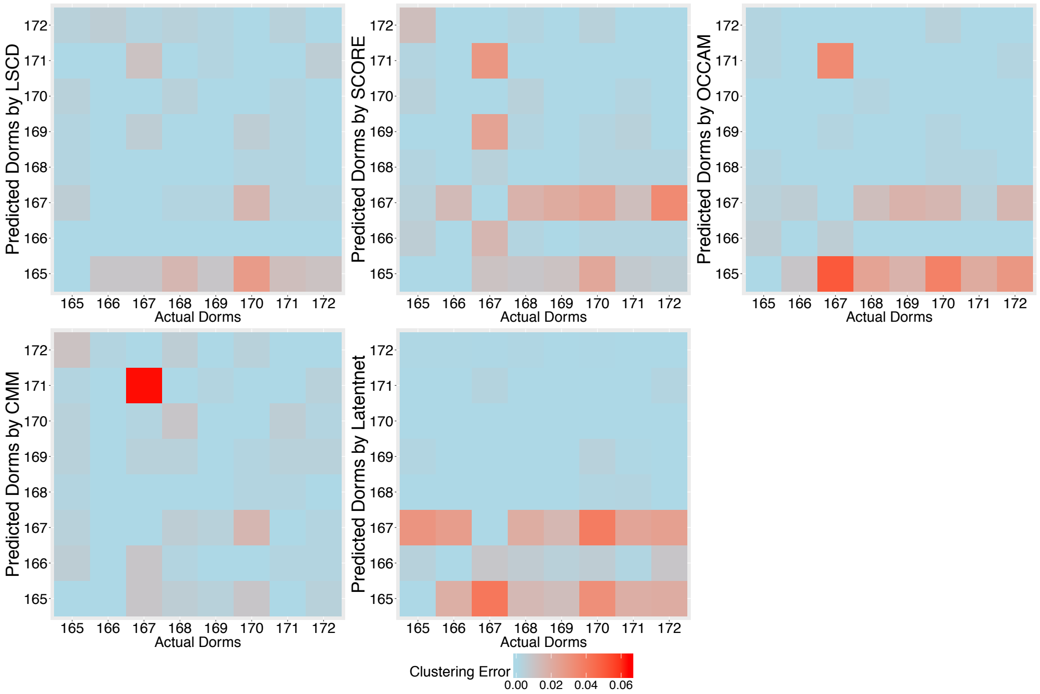

Figure 10: Comparison of community-wise misclustering errors in Caltech friendship network. Top row, left to right: LSCD, SCORE and OCCAM; bottom row, left to right: CMM and Latentnet.

4.5.2. Community detection with covariate

We now further demonstrate the power of the model and our proposed fitting methods by considering community detection on networks with covariates. Again, we used the LSCD procedure laid out in the previous subsection for community detection.

To this end, we consider a lawyer network dataset which was introduced in Lazega (2001) that studied the relations among 71 lawyers in a New England law firm. The lawyers were asked to check the names of those who they socialized with outside work, who they knew their family and vice versa. There are also several node attributes contained in the dataset: status (partner or associate), gender, office, years in the firm, age, practice (litigation or corporate), and law school attended, among which status is most assortative. Following Zhang et al. (2015), we took status as the true community label. Furthermore, we symmetrized the adjacency matrix, excluded two isolated nodes and finally ended up with 69 lawyers connected by 399 undirected edges.

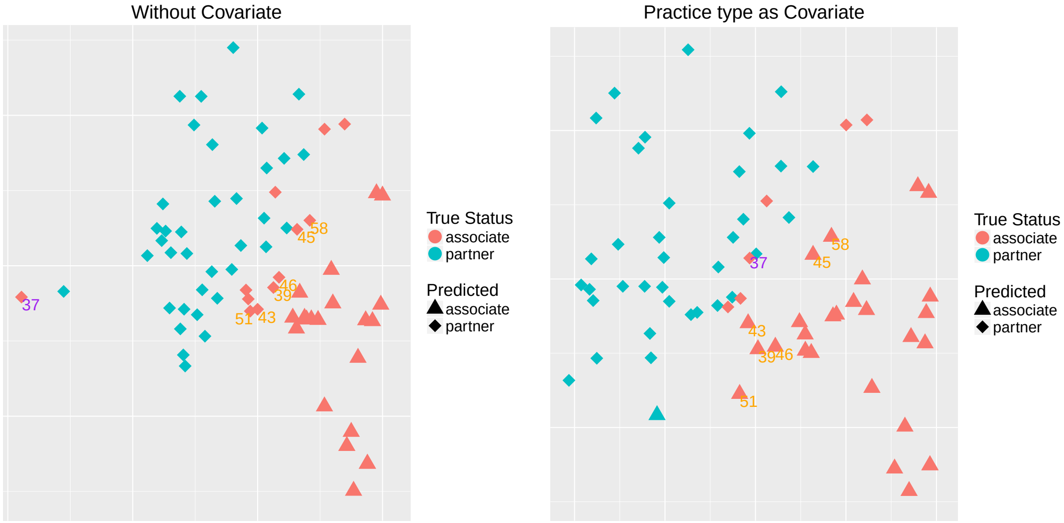

Figure 11: Visualization of the lawyer network using the estimated two dimensional latent vectors. The left panel shows results without including any covariate while the right panel shows results that used practice type information.

Visualization and clustering results with and without covariate are shown in Figure 11. On the left panel, as we can see, the latent vectors without adjustment by any covariate worked reasonably well in separating the lawyers of different status and most of the 12 errors (red diamonds) were committed on the boundary. On the right panel, we included a covariate 'practice' into the latent space model: $X_{ij} = X_{ji} = 1$ if $i \neq j$ and the ith and the jth lawyers shared the same practice, and $X_{ij} = X_{ji} = 0$ otherwise. Ideally, the influence on the network of being the same type of lawyer should be 'ruled out' this way and the remaining influence on connecting probabilities should mainly be the effect of having different status. In other words, the estimated latent vectors should mainly contain the information of lawyers' status and the effect of lawyers' practice type should be absorbed into the factor βX. The predicted community memberships of lawyers indexed by orange numbers (39, 43, 45, 46, 51, 58) were successfully corrected after introducing this covariate. So the number of misclustered nodes was reduced by 50%. We also observed that lawyer 37, though still misclassified, was significantly pushed towards the right spot.

4.5.3. Network assisted learning

In this section, we demonstrate how fitting model (4.1) can generate new features to be used in machine learning applications when additional network information is available. Consider a network with n nodes and observed adjacency matrix A. Suppose the profile of the nodes is represented by d dimensional features, denoted by $x_1, \cdots, x_n \in \mathbb{R}^d$. Assume each node is associated with a label (or say, variable of interest), denoted by y, either continuous or categorical. Suppose the labels are only observed for part of the nodes in the network. Without loss of generality, we assume $y_1, \cdots, y_m$ are observed where $m < n$. The goal here is to predict the rest of the labels $y_{m+1}, \cdots, y_n$ based on available information. Without considering the network information, this is typical setup of supervised learning with labeled traning set $(x_1, y_1), \cdots, (x_m, y_m)$ and unlabeled test set $x_{m+1}, \cdots, x_n$. As one way to utilize the network information, we propose to supplement the existing features in the prediction task with the latent vectors estimated by Algorithm 6 (without including edge covariates).

To give a specific example, we considered the Cora dataset (McCallum et al., 2000). It contains 2708 machine learning papers which were manually classified into 7 categories: Neural Networks, Rule Learning, Reinforcement Learning, Probabilistic Methods, Theory, Genetic Algorithms and Case Based. The dataset also includes the contents of the papers and a citation network, which are represented by a document-word matrix (the vocabulary contains 1433 frequent words) and an adjacency matrix respectively. The task is to predict the category of the papers based on the available information. For demonstration purpose, we only consider distinguishing neural network papers from the other categories.

As usually done in latent semantic analysis, to represent the text information as vectors, we extract leading-d principal components from the document-word matrix as the features. We chose $d = 100$ by maximizing the prediction accuracy using cross-validation.

However, how to utilize the information contained in the citation network for the desired learning problem is less straightforward. We propose to augment the latent semantic features with the latent vectors estimated from the citation network. Based on the simple intuition that papers in the same category are more likely to cite each other, we expect that the latent vectors, as low dimensional summary of the network, should contain information about the paper category. The key message we want to deliver with this data example is that with vector representation of the nodes obtained from fitting the latent space model, network information can be incorporated in many supervised and unsupervised learning problems and other exploratory data analysis tasks.

Back to the Cora dataset, for illustration purpose, we fitted standard logistic regressions with the following three sets of features:

1. the leading 100 principal components;

2. estimated degree parameters $\hat{\alpha}_i$ and latent vectors $\hat{z}_i$ obtained from Algorithm 6;

3. the combination of features in 1 and 2.

We considered three different latent space dimensions: $k = 2, 5, 10$. As we can see from Figure 12, the latent vectors contained a considerable amount of predictive power for the category. Adding the latent vectors to the principal components of the word-document matrix substantially reduced misclassification rate.

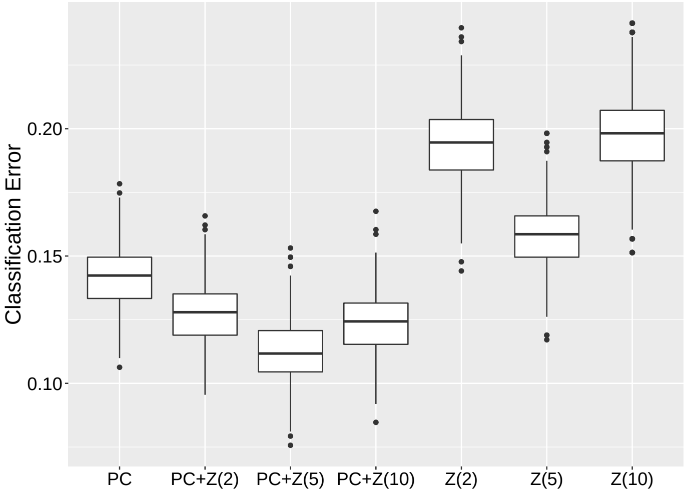

Figure 12: Boxplots of misclassification errors using logistic regression with different feature sets. We randomly split the dataset into training and test sets with size ratio 3:1 for 500 times and computed misclassification errors for each configuration. PC represents the leading 100 principal components of the document-word matrix. $Z(k)$ represents the feature matrix where the i^{th} row is the concatenation of the estimated degree parameter $\widehat{\alpha}_i$ and the estimated latent vector $\widehat{z}_i$ with latent dimension k.

4.6. Discussion

In this section, we discuss a number of related issues and potential future research problems.

Data-driven choice of latent space dimension For the projected gradient descent method, i.e., Algorithm 6, one needs to specify the latent space dimension k as an input. Although Theorem 9 suggests that the algorithm could still work reasonably well if the specified latent space dimension is slightly off the target, it is desirable to have a systematic approach to selecting k based on data. One possibility is to inspect the eigenvalues of G^T in Algorithm 7 and set k to be the number of eigenvalues larger than the parameter λ_n used in the algorithm. Alternatively, one may consider adapting the bi-cross-validation technique Owen and Perry (2009) to the current context. We leave systematic investigation of such a choice for future research.

Undirected networks with multiple covariates and weighted edges The model (4.1) and the fitting methods can easily be extended to handle multiple covariates. When the number of covariates is fixed, error bounds analogous to those in Section 4.3 can also be established when network sizes are sufficiently large. We omit the details since they do not seem to offer additional insights and the proof arguments are essentially the same.

Moreover, as pointed out in Goldenberg et al. (2010), latent space models for binary networks such as (4.1) can readily be generalized to weighted networks, i.e., networks with non-binary edges. We refer interested readers to the general recipe prescribed in Section 3.9 of Goldenberg et al. (2010). If the latent variables enter a model for weighted networks in the same way as in model (4.1), we expect the key ideas behind our proposed fitting methods to continue to work.

Directed networks In many real world networks, edges are directed. Thus, it is a natural next step to generalize model (4.1) to handle such data. Suppose for any $i \neq j$, $A_{ij} = 1$ if there is an edge pointing from node i to node j, and $A_{ij} = 0$ otherwise. We can consider the following model: for any $i \neq j$,

$$A_{ij} \overset{ind.}{\sim} \text{Bernoulli}(P_{ij}), \quad \text{with} \quad \text{logit}(P_{ij}) = \Theta_{ij} = \alpha_i + \gamma_j + \beta X_{ij} + z_i^\top w_j. \tag{4.20}$$

Here, the α_i's $\in \mathbb{R}$ model degree heterogeneity of outgoing edges while the γ_j's $\in \mathbb{R}$ model heterogeneity of incoming edges. The meaning of β is the same as in model (4.1). To further accommodate asymmetry, we associate with each node two sets of latent variables $z_i, w_i \in \mathbb{R}^k$, where the z_i's are latent variables influencing outgoing edges and the w_i's incoming edges. Such a model has been proposed and used in the study of recommender system Agarwal and Chen (2009). Under this model, the idea behind the convex programming fitting method in Section 4.3.1 can be extended. However, it is not clear whether one could devise a non-convex fitting method with similar theoretical guarantees to what we have in the undirected case. On the other hand, it should be relatively straightforward to further extend the ideas to directed networks with multiple covariates and weighted edges.

4.7. Proofs

Throughout the proof, let $P = (\sigma(\Theta_{\star,ij}))$ and $P^0 = (P_{ij}\mathbf{1}_{i\neq j})$. Thus, $\mathbb{E}(A) = P^0$.

4.7.1. Proofs of Theorems 8

Let $Z_\star \in \mathbb{R}^{n\times k}$ such that $Z_\star Z_\star^\top$ is the best rank k approximation to $G_\star$. For any matrix M, let $\mathrm{col}(M)$ be the subspace spanned by the column vectors of M and $\mathrm{row}(M) = \mathrm{col}(M^\top)$. For any subspace $\mathcal{S}$ of $\mathbb{R}^n$ (or $\mathbb{R}^{n\times n}$), let $\mathcal{S}^\perp$ be its orthogonal complement, and $\mathcal{P}_{\mathcal{S}}$ the projection operator onto the subspace. The proof relies on the following two lemmas.

Lemma 4.12. *Let* $\mathcal{M}_k^\perp = \{M \in \mathbb{R}^{n\times n} : \mathrm{row}(M) \subset \mathrm{col}(Z_\star)^\perp \text{ and } \mathrm{col}(M) \subset \mathrm{col}(Z_\star)^\perp\}$ *and* $\mathcal{M}_k$ *be its orthogonal complement in* $\mathbb{R}^{n\times n}$ *under trace inner product. If* $\lambda_n \geq 2\left\|A - P\right\|_{\mathrm{op}}$, *then for* $\overline{G}_k = \mathcal{P}_{\mathcal{M}_k^\perp} G_\star$, *we have*

$$\|\Delta_{\widehat{G}}\|_* \leq 4\sqrt{2k}\left\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\right\|_{\mathrm{F}} + 2\left\|\Delta_{\widehat{\alpha}}1_n{}^\top\right\|_{\mathrm{F}} + \frac{2}{\lambda_n}|\langle A - P, \Delta_{\widehat{\beta}}X\rangle| + 4\|\overline{G}_k\|_*. \tag{4.21}$$

Lemma 4.13. *For any* $k \geq 1$ *such that Assumption 4.3.1 holds. Choose* $\lambda_n \geq \max\{2\|A - P\|_{\mathrm{op}}, 1\}$ *and* $|\langle A - P, X\rangle| \leq \lambda_n\sqrt{k}\|X\|_{\mathrm{F}}$. *There exist constants* $C > 0$ *and* $0 \leq c < 1$ *such that*

$$\begin{aligned}
&\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \geq (1-c)\big(\|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widehat{\beta}}X\|_{\mathrm{F}}^2\big) - C\|\overline{G}_k\|_*^2/k, \quad \text{and} \\
&\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \leq (1+c)\big(\|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widehat{\beta}}X\|_{\mathrm{F}}^2\big) + C\|\overline{G}_k\|_*^2/k.
\end{aligned} \tag{4.22}$$

Proof of part one of Theorem 8 Recall the definition of h in (4.35). Observe that $\widehat{\Theta} = \widehat{\alpha}1_n{}^\top + 1_n\widehat{\alpha}^\top + \widehat{\beta}X + \widehat{G}$ is the optimal solution to (4.11), and that the true parameter $\Theta_\star = \alpha_\star 1_n{}^\top + 1_n\alpha_\star^\top + \beta_\star X + G_\star$ is feasible. Thus, we have the basic inequality

$$h(\widehat{\Theta}) - h(\Theta_\star) + \lambda_n(\|\widehat{G}\|_* - \|G_\star\|_*) \leq 0. \tag{4.23}$$

For any Θ in the parameter space $\mathcal{F}(n,k,M_1,M_2,X)$, $|\Theta_{ij}| \le M_1$ for all $i,j \in [n]$ and so for $\tau = e^{M_1}/(1+e^{M_1})^2$, the Hessian

$$\nabla^2 h(\Theta) = \mathrm{diag}\big(\mathrm{vec}\big(\sigma(\Theta) \circ (1-\sigma(\Theta))\big)\big) \succeq \tau I_{n^2 \times n^2}.$$

For any vector b, $\mathrm{diag}(b)$ is the diagonal matrix with elements of a on its diagonals. For any matrix $B = [b_1, \ldots, b_n] \in \mathbb{R}^{n\times n}$, $\mathrm{vec}(B) \in \mathbb{R}^{n^2}$ is obtained by stacking $b_1, \ldots, b_n$ in order. With the last display, Taylor expansion gives

$$h(\widehat{\Theta}) - h(\Theta_\star) \ge \langle \nabla_\Theta h(\Theta_\star), \Delta_{\widehat{\Theta}} \rangle + \frac{\tau}{2}\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2.$$

On the other hand, triangular inequality implies

$$\lambda_n(\|\widehat{G}\|_* - \|G_\star\|_*) \ge -\lambda_n \|\Delta_G\|_*.$$

Together with (4.23), the last two displays imply

$$\langle \nabla_\Theta h(\Theta_\star), \Delta_{\widehat{\Theta}} \rangle + \frac{\tau}{2}\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}} - \lambda_n \|\Delta_{\widehat{G}}\|_* \le 0.$$

Triangle inequality further implies

$$\begin{aligned}
\frac{\tau}{2}\|\Delta_\Theta\|_F^2 &\le \lambda_n \|\Delta_G\|_* + |\langle \nabla_\Theta h(\Theta_\star), \Delta_{\widehat{G}} + \Delta_{\widehat{\alpha}} 1_n{}^\top + 1_n \Delta_{\widehat{\alpha}}^\top \rangle| + |\Delta_{\widehat{\beta}} \langle \nabla_\Theta h(\Theta_\star), X\rangle| \\
&= \lambda_n \|\Delta_G\|_* + |\langle A-P, \Delta_{\widehat{G}} + 2\Delta_{\widehat{\alpha}} 1_n{}^\top \rangle| + |\Delta_{\widehat{\beta}} \langle A - P, X\rangle| \qquad (4.24)\\
&\le \lambda_n \|\Delta_G\|_* + |\langle A-P, \Delta_{\widehat{G}} + 2\Delta_{\widehat{\alpha}} 1_n{}^\top \rangle| + \lambda_n \sqrt{k} \|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}}.
\end{aligned}$$

Here the equality is due to the symmetry of $A - P$ and the last inequality is due to the condition imposed on λ_n. We now further upper bound the first two terms on the rightmost side. First, by Lemma 4.12 and the assumption that $|\langle A-P, X\rangle| \le \lambda_n \sqrt{k}\|X\|_{\mathrm{F}}$, we have

$$\|\Delta_G\|_* \le 4\sqrt{2k}\, \|\mathcal{P}_{\mathcal{M}_k} \Delta_{\widehat{G}}\|_{\mathrm{F}} + 2\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + 2\sqrt{k}\, \|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}} + 4\|\overline{G}_k\|_*. \qquad (4.25)$$

Moreover, Hölder's inequality implies

$$
\begin{aligned}
|\langle A - P, \Delta_{\widehat{G}} + 2\Delta_{\widehat{\alpha}} 1_n{}^\top \rangle| &\le \|A - P\|_{\mathrm{op}} (\|\Delta_{\widehat{G}}\|_* + 2\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_*) \\
&= \|A - P\|_{\mathrm{op}} (\|\Delta_{\widehat{G}}\|_* + 2\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}}) \\
&\le \frac{\lambda_n}{2} (\|\Delta_{\widehat{G}}\|_* + 2\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}}).
\end{aligned}
\tag{4.26}
$$

Here the equality holds since $\Delta_{\widehat{\alpha}} 1_n{}^\top$ is a rank one matrix. Substituting (4.25) and (4.26) into (4.24), we obtain that

$$
\begin{aligned}
\frac{\tau}{2}\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 &\le \frac{3\lambda_n}{2}\|\Delta_{\widehat{G}}\|_* + \lambda_n \|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \lambda_n \sqrt{k} \|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}} \\
&\le \frac{3\lambda_n}{2} (4\sqrt{2k} \|\mathcal{P}_{\mathcal{M}_k} \Delta_{\widehat{G}}\|_{\mathrm{F}} + 2\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + 2\sqrt{k}\|\Delta_{\widehat{\beta}} X\|_F + 4\|\overline{G}_k\|_*) \\
&\qquad\qquad + \lambda_n \|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \lambda_n \sqrt{k} \|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}} \\
&\le C_1 \lambda_n \big(\sqrt{k}\, (\|\mathcal{P}_{\mathcal{M}_k} \Delta_{\widehat{G}}\|_{\mathrm{F}} + \|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}}) + \|\overline{G}_k\|_* \big).
\end{aligned}
\tag{4.27}
$$

By Lemma 4.13, we can further bound the righthand side as

$$
\begin{aligned}
\frac{\tau}{2}\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 &\le C_2 \lambda_n \sqrt{k}\, (\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}} + \|\overline{G}_k\|_* / \sqrt{k}) + C_1 \lambda_n \|\overline{G}_k\|_* \\
&\le C_2 \lambda_n \sqrt{k}\, \|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}} + (C_1 + C_2)\lambda_n \|\overline{G}_k\|_*.
\end{aligned}
\tag{4.28}
$$

Solving the quadratic inequality, we obtain

$$
\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \le C' \left(\frac{\lambda_n^2 k}{\tau^2} + \frac{\lambda_n \|\overline{G}_k\|_*}{\tau} \right). \tag{4.29}
$$

Note that $\tau \ge c e^{-M_1}$ for some positive constant c. Therefore,

$$
\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \le C \left(e^{2M_1} \lambda_n^2 k + e^{M_1} \lambda_n \|\overline{G}_k\|_* \right).
$$

This completes the proof. □

Proof of part two of Theorem 8 The proof relies on the following lemma.

Lemma 4.14. *For any $\Theta \in \mathcal{F}(n, k, M_1, M_2, X)$, there exists absolute constants c, C such* *that with probability at least* $1 - n^{-c}$*, the following inequality holds*

$$\|A - P\|_{\mathrm{op}}, \ \frac{\langle A - P, X\rangle}{\sqrt{k}\, \|X\|_{\mathrm{F}}} \leq C\sqrt{\max\left\{ne^{-M_2}, \log n\right\}}. \tag{4.30}$$

Proof of Lemma 4.14. For any Θ in the parameter space, the off diagonal elements of Θ are uniformly bounded from above by $-M_2$, and so $\max_{i,j} P^0_{ij} \leq e^{-M_2}$. Moreover, $\max_i P_{ii} \leq 1$ under our assumption. Thus, $\|A - P\|_{\mathrm{op}} \leq \|A - P^0\|_{\mathrm{op}} + \|P^0 - P\|_{\mathrm{op}} \leq \|A - P^0\|_{\mathrm{op}} + 1$. Together with Lemma 4.23, this implies that there exist absolute constants $c_1, C > 0$ such that uniformly over the parameter space

$$\mathbb{P}\left(\|A - P\|_{\mathrm{op}} \leq C\sqrt{\max\left\{ne^{-M_2}, \log n\right\}}\right) \geq 1 - n^{-c_1}. \tag{4.31}$$

Since the diagonal entries of X are all zeros, we have $\langle A - P, X\rangle = \langle A - P^0, X\rangle$. Hence, Lemma 4.24 implies that uniformly over the parameter space,

$$\begin{aligned}\mathbb{P}\left(\frac{\langle A - P, X\rangle}{\sqrt{k}\, \|X\|_{\mathrm{F}}} \leq C\sqrt{\max\left\{ne^{-M_2}, \log n\right\}}\right) &\geq 1 - 3\exp\left(-C^2 \max\left\{ne^{-M_2}, \log n\right\} k/8\right) \\ &\geq 1 - 3n^{-C^2 k/8}\end{aligned} \tag{4.32}$$

Combining (4.31) and (4.31) finishes the proof. □

By Lemma 4.14, there exist constants c_1, C_1 such that for any $\lambda_n \geq 2C_1\sqrt{\max\left\{ne^{-M_2}, \log n\right\}}$, we have uniformly over the parameter space that

$$\mathbb{P}\left(\lambda_n \geq 2\max\left\{\|A - P\|_{\mathrm{op}}, \frac{\langle A - P, X\rangle}{\sqrt{k}\, \|X\|_{\mathrm{F}}}\right\}\right) \geq 1 - n^{-c_1}. \tag{4.33}$$

Denote such event as E. Since the conditions on λ_n in the first part of Theorem 8 are satisfied on E, it follows that there exists an absolute constant $C > 0$ such that uniformly over the parameter space, with probability at least $1-n^{-c_1}$,

$$\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \leq C(e^{2M_1-M_2}nk + e^{M_1-M_2/2}\sqrt{n}\|\overline{G}_k\|_*). \tag{4.34}$$

This completes the proof. □

4.7.2. Proof of Lemma 4.12

Let

$$h(\Theta) = -\sum_{1\leq i,j\leq n}\{A_{ij}\Theta_{ij} + \log(1-\sigma(\Theta_{ij}))\}. \tag{4.35}$$

By the convexity of $h(\Theta)$,

$$\begin{aligned}
h(\widehat{\Theta}) - h(\Theta_\star) &\geq \langle \nabla_\Theta h(\Theta_\star), \Delta_{\widehat{\Theta}}\rangle \\
&= -\langle A-P,\, \Delta_{\widehat{G}} + 2\Delta_{\widehat{\alpha}}1_n{}^\top + \Delta_{\widehat{\beta}}X\rangle \\
&\geq -\|A-P\|_{\mathrm{op}}\left(\|\Delta_{\widehat{G}}\|_* + 2\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_*\right) - |\langle A-P, \Delta_{\widehat{\beta}}X\rangle| \\
&\geq -\frac{\lambda_n}{2}\left(\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_* + \|\mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widehat{G}}\|_* + 2\left\|\Delta_{\widehat{\alpha}}1_n{}^\top\right\|_{\mathrm{F}}\right) - |\langle A-P, \Delta_{\widehat{\beta}}X\rangle|\,.
\end{aligned}$$

The last inequality holds since $\lambda_n \geq 2\|A-P\|_{\mathrm{op}}$ and $\mathcal{P}_{\mathcal{M}_k} + \mathcal{P}_{\mathcal{M}_k^\perp}$ equals identity. On the other hand, by the definition of $\overline{G}_k$,

$$\begin{aligned}
\|\widehat{G}\|_* - \|G_\star\|_* &= \|\mathcal{P}_{\mathcal{M}_k}G_\star + \overline{G}_k + \mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}} + \mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widehat{G}}\|_* - \|\mathcal{P}_{\mathcal{M}_k}G_\star + \overline{G}_k\|_* \\
&\geq \|\mathcal{P}_{\mathcal{M}_k}G_\star + \mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widehat{G}}\|_* - \|\overline{G}_k\|_* - \|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_* - \|\mathcal{P}_{\mathcal{M}_k}G_\star\|_* - \|\overline{G}_k\|_* \\
&= \|\mathcal{P}_{\mathcal{M}_k}G_\star\|_* + \|\mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widehat{G}}\|_* - 2\|\overline{G}_k\|_* - \|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_* - \|\mathcal{P}_{\mathcal{M}_k}G_\star\|_* \\
&= \|\mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widehat{G}}\|_* - \|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_* - 2\|\overline{G}_k\|_*\,.
\end{aligned}$$

Here, the second last equality holds since $\mathcal{P}_{\mathcal{M}_k}G_\star$ and $\mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widehat{G}}$ have orthogonal column and row spaces. Furthermore, since $\widehat{\Theta}$ is the optimal solution to (4.11), and $\Theta_\star$ is feasible, the basic inequality and the last two displays imply

$$
\begin{aligned}
0 &\geq h(\widehat{\Theta}) - h(\Theta_\star) + \lambda_n \big(\|\widehat{G}\|_* - \|G_\star\|_* \big) \\
&\geq -\frac{\lambda_n}{2} \big(\|\mathcal{P}_{\mathcal{M}_k} \Delta_{\widehat{G}}\|_* + \|\mathcal{P}_{\mathcal{M}_k^\perp} \Delta_{\widehat{G}}\|_* + 2 \left\| \Delta_{\widehat{\alpha}} 1_n{}^\top \right\|_{\mathrm{F}} \big) \\
&\qquad\qquad - |\langle A - P, \Delta_{\widehat{\beta}} X \rangle| + \lambda_n \big(\|\mathcal{P}_{\mathcal{M}_k^\perp} \Delta_{\widehat{G}}\|_* - \|\mathcal{P}_{\mathcal{M}_k} \Delta_{\widehat{G}}\|_* - 2\|\overline{G}_k\|_* \big) \\
&= \frac{\lambda_n}{2} \big(\|\mathcal{P}_{\mathcal{M}_k^\perp} \Delta_{\widehat{G}}\|_* - 3\|\mathcal{P}_{\mathcal{M}_k} \Delta_{\widehat{G}}\|_* - 4\|\overline{G}_k\|_* - 2 \left\| \Delta_{\widehat{\alpha}} 1_n{}^\top \right\|_{\mathrm{F}} \big) - |\langle A - P, \Delta_{\widehat{\beta}} X \rangle| \,.
\end{aligned}
$$

Rearranging the terms leads to

$$
\|\mathcal{P}_{\mathcal{M}_k^\perp} \Delta_{\widehat{G}}\|_* \leq 3\|\mathcal{P}_{\mathcal{M}_k} \Delta_{\widehat{G}}\|_* + 2 \left\| \Delta_{\widehat{\alpha}} 1_n{}^\top \right\|_{\mathrm{F}} + \frac{2}{\lambda_n} |\langle A - P, \Delta_{\widehat{\beta}} X \rangle| + 4\|\overline{G}_k\|_* \,,
$$

and triangle inequality further implies

$$
\|\Delta_{\widehat{G}}\|_* \leq 4\|\mathcal{P}_{\mathcal{M}_k} \Delta_{\widehat{G}}\|_* + 2 \left\| \Delta_{\widehat{\alpha}} 1_n{}^\top \right\|_{\mathrm{F}} + \frac{2}{\lambda_n} |\big\langle A - P, \Delta_{\widehat{\beta}} X \big\rangle| + 4\|\overline{G}_k\|_* \,.
$$

Last but not least, note that the rank of $\mathcal{P}_{\mathcal{M}_k} \Delta_{\widehat{G}}$ is at most $2k$, and so we complete the proof by further bounding the first term on the righthand side of the last display by $4\sqrt{2k} \left\| \mathcal{P}_{\mathcal{M}_k} \Delta_{\widehat{G}} \right\|_{\mathrm{F}}$.

4.7.3. Proof of Lemma 4.13

By definition, we have the decomposition

$$
\begin{aligned}
\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 &= \|\Delta_{\widehat{G}} + \Delta_{\widehat{\alpha}} 1_n{}^\top + 1_n \Delta_{\widehat{\alpha}}^\top + \Delta_{\widehat{\beta}} X\|_{\mathrm{F}}^2 \\
&= \|\Delta_{\widehat{G}} + \Delta_{\widehat{\alpha}} 1_n{}^\top + 1_n \Delta_{\widehat{\alpha}}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}}^2 + 2 \,\langle \Delta_{\widehat{G}} + \Delta_{\widehat{\alpha}} 1_n{}^\top + 1_n \Delta_{\widehat{\alpha}}^\top, \Delta_{\widehat{\beta}} X \rangle \\
&= \|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}}^2 + 2 \,\mathrm{tr}(\Delta_{\widehat{\alpha}} 1_n{}^\top \Delta_{\widehat{\alpha}} 1_n{}^\top) \\
&\quad + \|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}}^2 + 2 \,\langle \Delta_{\widehat{G}} + 2\Delta_{\widehat{\alpha}} 1_n{}^\top, \Delta_{\widehat{\beta}} X \rangle \,.
\end{aligned}
$$

Here the last equality is due to the symmetry of X and the fact that $\Delta_{\widehat{G}} 1_n = 0$. Since $\operatorname{tr}(\Delta_{\widehat{\alpha}} 1_n{}^\top \Delta_{\widehat{\alpha}} 1_n{}^\top) = \operatorname{tr}(1_n{}^\top \Delta_{\widehat{\alpha}} 1_n{}^\top \Delta_{\widehat{\alpha}}) = |1_n{}^\top \Delta_{\widehat{\alpha}}|^2 \geq 0$, the last display implies

$$\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \geq \|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}}^2 + 2\,\langle \Delta_{\widehat{G}} + 2\Delta_{\widehat{\alpha}} 1_n{}^\top, \Delta_{\widehat{\beta}} X\rangle. \tag{4.36}$$

Furthermore, we have

$$\begin{aligned}
&|\langle \Delta_{\widehat{G}} + 2\Delta_{\widehat{\alpha}} 1_n{}^\top, \Delta_{\widehat{\beta}} X\rangle| \\
&\leq \|\Delta_{\widehat{G}}\|_* \|\Delta_{\widehat{\beta}} X\|_{\mathrm{op}} + 2\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_* \|\Delta_{\widehat{\beta}} X\|_{\mathrm{op}} \\
&\leq \big(4\sqrt{2k}\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_{\mathrm{F}} + 4\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \frac{2}{\lambda_n}|\langle A - P, \Delta_{\widehat{\beta}} X\rangle| + 4\|\overline{G}_k\|_*\big)\|\Delta_{\widehat{\beta}} X\|_{\mathrm{op}} \\
&\leq \big(4\sqrt{2k}\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_{\mathrm{F}} + 4\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + 2\sqrt{k}\|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}} + 4\|\overline{G}_k\|_*\big)\frac{\|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}}}{\sqrt{\mathrm{r}_{\mathrm{stable}}(X)}} \\
&\leq \frac{C_0\sqrt{k}}{\sqrt{\mathrm{r}_{\mathrm{stable}}(X)}}\big(\|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}}^2\big) + \frac{4\|\overline{G}_k\|_*}{\sqrt{\mathrm{r}_{\mathrm{stable}}(X)}}\|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}} \\
&\leq \frac{C_0\sqrt{k}}{\sqrt{\mathrm{r}_{\mathrm{stable}}(X)}}\big(\|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}}^2\big) + \frac{2\|\overline{G}_k\|_*^2}{c_0\,\mathrm{r}_{\mathrm{stable}}(X)} + 2c_0\|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}}^2
\end{aligned}$$

for any constant $c_0 \geq 0$. Here, the first inequality holds since the operator norm and the nuclear norm are dual norms under trace inner product. The second inequality is due to Lemma 4.12 and the fact that $\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_* = \|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}}$ since $\Delta_{\widehat{\alpha}} 1_n{}^\top$ is of rank one. The third inequality is due to the definition of $\mathrm{r}_{\mathrm{stable}}(X)$ and that $|\langle A - P, X\rangle| \leq \lambda_n\sqrt{k}\|X\|_{\mathrm{F}}$ by assumption and $\Delta_{\widehat{\beta}}$ is a scalar. The fourth inequality is due to Assumption 4.3.1 and the last due to $2ab \leq a^2 + b^2$ for any $a, b \in \mathbb{R}$. Substituting these inequalities into (4.36) leads to

$$\begin{aligned}
\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \geq{}& \left(1 - \frac{2C_0\sqrt{k}}{\sqrt{\mathrm{r}_{\mathrm{stable}}(X)}}\right)\|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2 + \left(2 - \frac{2C_0\sqrt{k}}{\sqrt{\mathrm{r}_{\mathrm{stable}}(X)}}\right)\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}}^2 \\
&+ \left(1 - \frac{2C_0\sqrt{k}}{\sqrt{\mathrm{r}_{\mathrm{stable}}(X)}} - 4c_0\right)\|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}}^2 - \frac{4\|\overline{G}_k\|_*^2}{c_0\,\mathrm{r}_{\mathrm{stable}}(X)}\,.
\end{aligned}$$

On the other hand, notice that $\operatorname{tr}(\Delta_{\widehat{\alpha}} 1_n{}^\top \Delta_{\widehat{\alpha}} 1_n{}^\top) \le \|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_F^2$, we have

$$\begin{aligned}\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \le& \left(1 + \frac{2C_0\sqrt{k}}{\sqrt{\mathrm{r_{stable}}(X)}}\right)\|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2 + \left(4 + \frac{2C_0\sqrt{k}}{\sqrt{\mathrm{r_{stable}}(X)}}\right)\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}}^2 \\ &+ \left(1 + \frac{2C_0\sqrt{k}}{\sqrt{\mathrm{r_{stable}}(X)}} + 4c_0\right)\|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}}^2 + \frac{4\|\overline{G}_k\|_*^2}{c_0\,\mathrm{r_{stable}}(X)}\,.\end{aligned}$$

Together with Assumption 4.3.1, the last two displays complete the proof.

4.7.4. Proof of Lemma 4.4 and Theorem 9

Again, we directly prove the results under the general model. Recall that $G_\star \approx U_k D_k U_k^\top$ is the top-k eigen-decomposition of $G_\star$, $Z_\star = U_k D_k^{1/2}$, $\overline{G}_k = G_\star - U_k D_k U_k^\top$ and $\Delta_{G^t} = Z^t(Z^t)^\top - Z_\star Z_\star^\top$. For the convenience of analysis, we will instead analyze the following quantity,

$$\widetilde{e}_t = \left\|Z^0\right\|_{\mathrm{op}}^2 \|\Delta_{Z^t}\|_{\mathrm{F}}^2 + 2\left\|\Delta_{\alpha^t} 1_n{}^\top\right\|_{\mathrm{F}}^2 + \left\|\Delta_{\beta^t} X\right\|_{\mathrm{F}}^2. \tag{4.37}$$

Under Assumption 4.3.2,

$$\|\Delta_{Z^0}\|_{\mathrm{op}} \le \delta\, \|Z_\star\|_{\mathrm{op}}\,, \quad (1-\delta)e_t \le \widetilde{e}_t \le (1+\delta)e_t. \tag{4.38}$$

for some sufficiently small constant $\delta \in (0,1)$. The rest of the proof relies on the following lemmas.

Lemma 4.15. *For any $\Theta_\star \in \mathcal{F}_g(n, M_1, M_2, X)$, $\max_{1\le i\le n} \|(Z_\star)_i\|_2^2 \le M_1/3$.*

Proof. By definition, $G_\star - Z_\star Z_\star^\top \in \mathcal{S}_+^n$, which implies, $e_i^\top\left(G_\star - Z_\star Z_\star^\top\right) e_i = G_{ii} - \|(Z_\star)_i\|_2^2 \ge 0$, that is $\|(Z_\star)_i\|_2^2 \le G_{ii} \le M_1/3$ for any $1 \le i \le n$. □

Lemma 4.16. *If Assumption 4.3.1 holds, there exist constants $0 \le c_0 < 1$ and C_0 such that*

$$\begin{aligned}\|\Delta_{\Theta^t}\|_{\mathrm{F}}^2 &\ge (1-c_0)\left(\left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\mathrm{F}}^2 + 2\left\|\Delta_{\alpha^t} 1_n{}^\top\right\|_{\mathrm{F}}^2 + \left\|\Delta_{\beta^t} X\right\|_{\mathrm{F}}^2\right) - C_0\left\|\overline{G}_k\right\|_{\mathrm{F}}^2, \\ \|\Delta_{\Theta^t}\|_{\mathrm{F}}^2 &\le (1+c_0)\left(\left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\mathrm{F}}^2 + 2\left\|\Delta_{\alpha^t} 1_n{}^\top\right\|_{\mathrm{F}}^2 + \left\|\Delta_{\beta^t} X\right\|_{\mathrm{F}}^2\right) + C_0\left\|\overline{G}_k\right\|_{\mathrm{F}}^2.\end{aligned}$$

Lemma 4.17. *Under Assumption 4.3.1, let* $\zeta_n = \max\{2\left\|A-P\right\|_{\rm op},\ |\langle A-P, X/\left\|X\right\|_{\rm F}\rangle|/\sqrt{k},\ 1\}$, *if* $\left\|\Delta_{Z^t}\right\|_{\rm F} \leq c_0 e^{-M_1}\left\|Z_\star\right\|_{\rm op}/\kappa_{Z_\star}^2$ *and* $\left\|Z_\star\right\|_{\rm op}^2 \geq C_0 e^{M_1}\kappa_{Z_\star}^2\zeta_n^2$ *for sufficiently small constant* c_0 *and sufficiently large constant* C_0*, there exist a constant* c *such that, for any* $\eta \leq c$*, there exist positive constants* ρ *and* C*,*

$$\widetilde{e}_{t+1} \leq \left(1 - \frac{\eta}{e^{M_1}\kappa^2}\rho\right)\widetilde{e}_t + \eta C\left(\left\|\overline{G}_k\right\|_{\rm F}^2 + e^{M_1}\zeta_n^2 k\right).$$

Lemma 4.18. *Under Assumption 4.3.1, let* $\zeta_n = \max\{2\left\|A-P\right\|_{\rm op},\ |\langle A-P, X/\left\|X\right\|_{\rm F}\rangle|/\sqrt{k},\ 1\}$, *if* $\left\|Z_\star\right\|_{\rm op}^2 \geq C_1\kappa_{Z_\star}^2\zeta_n^2 e^{M_1}\max\left\{\sqrt{\eta\left\|\overline{G}_k\right\|_{\rm F}^2/\zeta_n^2}, \sqrt{\eta k e^{M_1}}, 1\right\}$ *for a sufficiently large constant* C_1 *and* $\widetilde{e}_0 \leq c_0^2 e^{-2M_1}\left\|Z_\star\right\|_{\rm op}^4/4\kappa_{Z_\star}^4$*, then for all* $t \geq 0$*,*

$$\left\|\Delta_{Z^t}\right\|_{\rm F} \leq \frac{c_0}{e^{M_1}\kappa_{Z_\star}^2}\left\|Z_\star\right\|_{\rm op}. \tag{4.39}$$

Proof of Lemma 4.4 By Lemma 4.16, notice that $\overline{G}_k = 0$ under the inner product model,

$$\begin{aligned}\left\|\Delta_{\Theta^t}\right\|_{\rm F}^2 &\geq (1-c_0)\left(\left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\rm F}^2 + 2\left\|\Delta_{\alpha^t}1_n{}^\top\right\|_{\rm F}^2 + \left\|\Delta_{\beta^t}X\right\|_{\rm F}^2\right)\\ \left\|\Delta_{\Theta^t}\right\|_{\rm F}^2 &\leq (1+c_0)\left(\left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\rm F}^2 + 2\left\|\Delta_{\alpha^t}1_n{}^\top\right\|_{\rm F}^2 + \left\|\Delta_{\beta^t}X\right\|_{\rm F}^2\right)\end{aligned} \tag{4.40}$$

By Lemma 4.20,

$$\left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\rm F}^2 \geq 2(\sqrt{2}-1)\kappa_{Z_\star}^{-2}\left\|Z_\star\right\|_{\rm op}^2\left\|\Delta_{Z^t}\right\|_{\rm F}^2 \tag{4.41}$$

which implies,

$$
\begin{aligned}
e_t &\leq \frac{\kappa_{Z_\star}^2}{2(\sqrt{2}-1)} \left\| Z^t (Z^t)^\top - Z_\star Z_\star^\top \right\|_{\mathrm{F}}^2 + 2 \left\| \Delta_{\alpha^t} 1_n{}^\top \right\|_{\mathrm{F}}^2 + \left\| \Delta_{\beta^t} X \right\|_{\mathrm{F}}^2 \\
&\leq \frac{\kappa_{Z_\star}^2}{2(\sqrt{2}-1)(1-c_0)} \left\| \Delta_{\Theta^t} \right\|_{\mathrm{F}}^2 .
\end{aligned}
\tag{4.42}
$$

Similarly, by Lemma 4.21, when $\operatorname{dist}(Z^t, Z_\star) \leq c \left\| Z_\star \right\|_{\mathrm{op}}$,

$$
\left\| Z^t (Z^t)^\top - Z_\star Z_\star^\top \right\|_{\mathrm{F}}^2 \leq (2+c)^2 \left\| Z_\star \right\|_{\mathrm{op}}^2 \left\| \Delta_{Z^t} \right\|_{\mathrm{F}}^2 ,
\tag{4.43}
$$

and this implies,

$$
\begin{aligned}
e_t &\geq \frac{1}{(2+c)^2} \left\| Z^t (Z^t)^\top - Z_\star Z_\star^\top \right\|_{\mathrm{F}}^2 + 2 \left\| \Delta_{\alpha^t} 1_n{}^\top \right\|_{\mathrm{F}}^2 + \left\| \Delta_{\beta^t} X \right\|_{\mathrm{F}}^2 \\
&\geq \frac{1}{(2+c)^2(1+c_0)} \left\| Z_\star \right\|_{\mathrm{op}}^2 \left\| \Delta_{Z^t} \right\|_{\mathrm{F}}^2 .
\end{aligned}
\tag{4.44}
$$

Proof of Part one of Theorem 9 By Lemma 4.18, for all $t \geq 0$,

$$
\left\| \Delta_{Z^t} \right\|_{\mathrm{F}} \leq \frac{c_0}{e^{M_1} \kappa_{Z_\star}^2} \left\| Z_\star \right\|_{\mathrm{op}}
\tag{4.45}
$$

Then apply Lemma 4.17, there exists positive constants ρ and M such that for all $t \geq 0$,

$$
\widetilde{e}_{t+1} \leq \left(1 - \frac{\eta}{e^{M_1} \kappa_{Z_\star}^2} \rho \right) \widetilde{e}_t + \eta C \left(\left\| \overline{G}_k \right\|_{\mathrm{F}}^2 + e^{M_1} \lambda_n^2 k \right)
$$

Therefore,

$$
\begin{aligned}
\widetilde{e}_t &\leq \left(1 - \frac{\eta}{e^{M_1} \kappa_{Z_\star}^2} \rho \right)^t \widetilde{e}_0 + \sum_{i=0}^{t} \eta C \left(\left\| \overline{G}_k \right\|_{\mathrm{F}}^2 + e^{M_1} \lambda_n^2 k \right) \left(1 - \frac{\eta}{e^{M_1} \kappa_{Z_\star}^2} \rho \right)^i \\
&\leq \left(1 - \frac{\eta}{e^{M_1} \kappa_{Z_\star}^2} \rho \right)^t \widetilde{e}_0 + \frac{C \kappa^2}{\rho} \left(e^{2M_1} \lambda_n^2 k + e^{M_1} \left\| \overline{G}_k \right\|_{\mathrm{F}}^2 \right) .
\end{aligned}
\tag{4.46}
$$

Notice that $0.9e_t \leq \widetilde{e}_t \leq 1.1e_t$,

$$e_t \leq 2\left(1 - \frac{\eta}{e^{M_1}\kappa_{Z_\star}^2}\rho\right)^t e_0 + \frac{2C\kappa^2}{\rho}\left(e^{2M_1}\lambda_n^2 k + e^{M_1}\left\|\overline{G}_k\right\|_{\mathrm{F}}^2\right). \tag{4.47}$$

Proof of Part two of Theorem 9 The proof is nearly the same as that of part two of Theorem 8 and we leave out the details.

4.7.5. Proof of Lemma 4.16

By definition,

$$\begin{aligned}
\|\Delta_{G^t}\|_{\mathrm{F}}^2 &= \left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top - \overline{G}_k\right\|_{\mathrm{F}}^2 \\
&\geq \left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\mathrm{F}}^2 + \left\|\overline{G}_k\right\|_{\mathrm{F}}^2 - 2|\langle Z^t(Z^t)^\top - Z_\star Z_\star^\top, \overline{G}_k\rangle| \\
&\geq \left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\mathrm{F}}^2 + \left\|\overline{G}_k\right\|_{\mathrm{F}}^2 - 2\left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\mathrm{F}}\left\|\overline{G}_k\right\|_{\mathrm{F}} \\
&\geq \left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\mathrm{F}}^2 + \left\|\overline{G}_k\right\|_{\mathrm{F}}^2 - c_1\left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\mathrm{F}}^2 - c_1^{-1}\left\|\overline{G}_k\right\|_{\mathrm{F}}^2 \\
&\geq (1-c_1)\left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\mathrm{F}}^2 - (c_1^{-1}-1)\left\|\overline{G}_k\right\|_{\mathrm{F}}^2
\end{aligned} \tag{4.48}$$

where the second last inequality comes from $a^2 + b^2 \geq 2ab$ and holds for any $c_1 \geq 0$. Similarly, it could be shown that

$$\|\Delta_{G^t}\|_{\mathrm{F}}^2 \leq (1+c_1)\left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\mathrm{F}}^2 + (1+c_1^{-1})\left\|\overline{G}_k\right\|_{\mathrm{F}}^2 \tag{4.49}$$

Expand the term $\|\Delta_{\Theta^t}\|_{\mathrm{F}}^2$,

$$\begin{aligned}
\|\Delta_{\Theta^t}\|_{\mathrm{F}}^2 &= \left\|\Delta_{G^t} + \Delta_{\alpha^t}1_n{}^\top + 1_n\Delta_{\alpha^t}^\top + \Delta_{\beta^t}X\right\|_{\mathrm{F}}^2 \\
&= \left\|\Delta_{G^t} + \Delta_{\alpha^t}1_n{}^\top + 1_n\Delta_{\alpha^t}^\top\right\|_{\mathrm{F}}^2 + \left\|\Delta_{\beta^t}X\right\|_{\mathrm{F}}^2 + 2\left\langle\Delta_{G^t} + \Delta_{\alpha^t}1_n{}^\top + 1_n\Delta_{\alpha^t}^\top, \Delta_{\beta^t}X\right\rangle \\
&= \|\Delta_{G^t}\|_{\mathrm{F}}^2 + 2\left\|\Delta_{\alpha^t}1_n{}^\top\right\|_{\mathrm{F}}^2 + 2\operatorname{tr}(\Delta_{\alpha^t}1_n{}^\top\Delta_{\alpha^t}1_n{}^\top) + \left\|\Delta_{\beta^t}X\right\|_{\mathrm{F}}^2 \\
&\qquad + 2\left\langle\Delta_{G^t} + 2\Delta_{\alpha^t}1_n{}^\top, \Delta_{\beta^t}X\right\rangle
\end{aligned}$$

where the last equality is due to the symmetry of X. Notice that $\operatorname{tr}(\Delta_{\widehat{\alpha}} 1_n{}^\top \Delta_{\widehat{\alpha}} 1_n{}^\top) = \operatorname{tr}(1_n{}^\top \Delta_{\widehat{\alpha}} 1_n{}^\top \Delta_{\widehat{\alpha}}) = |1_n{}^\top \Delta_{\widehat{\alpha}}|^2 \geq 0$,

$$\begin{aligned}
\|\Delta_{\Theta^t}\|_{\mathrm{F}}^2 \geq{}& (1-c_1)\left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\mathrm{F}}^2 - (c_1^{-1}-1)\left\|\overline{G}_k\right\|_{\mathrm{F}}^2 + 2\left\|\Delta_{\alpha^t} 1_n{}^\top\right\|_{\mathrm{F}}^2 + \left\|\Delta_{\beta^t} X\right\|_{\mathrm{F}}^2 \\
&+ 2\left\langle Z^t(Z^t)^\top - Z_\star Z_\star^\top + 2\Delta_{\alpha^t} 1_n{}^\top, \Delta_{\beta^t} X\right\rangle - 2\left\langle \overline{G}_k, \Delta_{\beta^t} X\right\rangle
\end{aligned} \tag{4.50}$$

By Holder's inequality,

$$\begin{aligned}
|\langle Z^t(Z^t)^\top - Z_\star Z_\star^\top + 2\Delta_{\alpha^t} 1_n{}^\top, \Delta_{\beta^t} X\rangle| &\leq \left(\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_* + 2\|\Delta_{\alpha^t} 1_n{}^\top\|_*\right)\left\|\Delta_{\beta^t} X\right\|_{\mathrm{op}} \\
&\leq \left(\sqrt{2k}\left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\mathrm{F}} + 2\left\|\Delta_{\alpha^t} 1_n{}^\top\right\|_{\mathrm{F}}\right)\left\|\Delta_{\beta^t} X\right\|_{\mathrm{op}} \\
&\leq \left(\sqrt{2k}\left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\mathrm{F}} + 2\left\|\Delta_{\alpha^t} 1_n{}^\top\right\|_{\mathrm{F}}\right)\left\|\Delta_{\beta^t} X\right\|_{\mathrm{F}} / \sqrt{\mathrm{r}_{\mathrm{stable}}(X)} \\
&\leq C_1\sqrt{\frac{k}{\mathrm{r}_{\mathrm{stable}}(X)}}\left(\left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\mathrm{F}}^2 + \left\|\Delta_{\alpha^t} 1_n{}^\top\right\|_{\mathrm{F}}^2 + \left\|\Delta_{\beta^t} X\right\|_{\mathrm{F}}^2\right)
\end{aligned}$$

and for any $c > 0$,

$$|\langle \overline{G}_k, \Delta_{\beta^t} X\rangle| \leq \left\|\overline{G}_k\right\|_{\mathrm{F}}\left\|\Delta_{\beta^t} X\right\|_{\mathrm{F}} \leq c\left\|\Delta_{\beta^t} X\right\|_{\mathrm{F}}^2 + \frac{1}{4c}\left\|\overline{G}_k\right\|_{\mathrm{F}}^2. \tag{4.51}$$

Substitute these inequalities into (4.50),

$$\begin{aligned}
\left\|\Delta_{\widehat{\Theta}}\right\|_{\mathrm{F}}^2 \geq{}& \left(1 - 2C_1\sqrt{\frac{k}{\mathrm{r}_{\mathrm{stable}}(X)}} - c_1\right)\left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\mathrm{F}}^2 \\
&+ \left(2 - 2C_1\sqrt{\frac{k}{\mathrm{r}_{\mathrm{stable}}(X)}}\right)\left\|\Delta_{\alpha^t} 1_n{}^\top\right\|_{\mathrm{F}}^2 \\
&+ \left(1 - 2C_1\sqrt{\frac{k}{\mathrm{r}_{\mathrm{stable}}(X)}} - 2c\right)\left\|\Delta_{\beta^t} X\right\|_{\mathrm{F}}^2 - (c_1^{-1} + 1/2c)\left\|\overline{G}_k\right\|_{\mathrm{F}}^2
\end{aligned}$$

On the other hand, notice that $\mathrm{tr}(\Delta_{\alpha^t}1_n{}^\top \Delta_{\alpha^t}1_n{}^\top) \leq \|\Delta_{\alpha^t}1_n\|_F^2$, we have

$$\begin{aligned}\left\|\Delta_{\widehat{\Theta}}\right\|_{\mathrm{F}}^2 \leq& \left(1 + 2C_1\sqrt{\frac{k}{\mathrm{r}_{\mathrm{stable}}(X)}} + c_1\right) \left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\mathrm{F}}^2 \\ &+ \left(2 + 2C_1\sqrt{\frac{k}{\mathrm{r}_{\mathrm{stable}}(X)}}\right) \left\|\Delta_{\widehat{\alpha}} 1_n{}^\top\right\|_{\mathrm{F}}^2 \\ &+ \left(1 + 2C_1\sqrt{\frac{k}{\mathrm{r}_{\mathrm{stable}}(X)}} + 2c\right) \left\|\Delta_{\beta^t} X\right\|_{\mathrm{F}}^2 + (c_1^{-1} + 1/2c) \left\|\overline{G}_k\right\|_{\mathrm{F}}^2\end{aligned}$$

4.7.6. Proof of Lemma 4.17

Let $\Theta^t = \alpha^t 1_n{}^\top + 1_n(\alpha^t)^\top + \beta^t X + Z^t(Z^t)^\top \in \mathcal{F}$, $R^t = \arg \min_{R\in\mathbb{R}^{r\times r}, RR^\top = I_r} \left\|Z^t - Z_\star R\right\|_{\mathrm{F}}$, $\widetilde{R}^t = \arg \min_{R\in\mathbb{R}^{r\times r}, RR^\top = I_r} \left\|\widetilde{Z}_t - Z_\star R\right\|_{\mathrm{F}}$ and $\Delta_{Z^t} = Z^t - Z_\star R^t$, then

$$\begin{aligned}\left\|Z^{t+1} - Z_\star R^{t+1}\right\|_{\mathrm{F}}^2 &\leq \left\|Z^{t+1} - Z_\star \widetilde{R}^{t+1}\right\|_{\mathrm{F}}^2 \\ &\leq \left\|\widetilde{Z}_{t+1} - Z_\star \widetilde{R}^{t+1}\right\|_{\mathrm{F}}^2 \\ &\leq \left\|\widetilde{Z}_{t+1} - Z_\star R^t\right\|_{\mathrm{F}}^2\end{aligned}$$

The first and last inequality are due to the definition of R^{t+1} and $\widetilde{R}^{t+1}$, and the second inequality is due to the projection step. Plug in the definition of $\widetilde{Z}^{t+1}$,

$$\begin{aligned}\left\|Z^{t+1} - Z_\star R^{t+1}\right\|_{\mathrm{F}}^2 &\leq \left\|Z^t - Z_\star R^t\right\|_{\mathrm{F}}^2 + \eta_Z^2 \left\|\nabla h(\Theta^t) Z^t\right\|_{\mathrm{F}}^2 - 2\eta_Z \left\langle \nabla h(\Theta^t) Z^t, Z^t - Z_\star R^t\right\rangle \\ &= \left\|Z^t - Z_\star R^t\right\|_{\mathrm{F}}^2 + \eta_Z^2 \left\|\nabla h(\Theta^t) Z^t\right\|_{\mathrm{F}}^2 \\ &\quad - 2\eta_Z \left\langle \nabla h(\Theta^t), (Z^t - Z_\star R^t)(Z^t)^\top\right\rangle\end{aligned}$$

Notice that,

$$Z^t(Z^t)^\top - Z_\star R^t (Z^t)^\top = \frac{1}{2}(Z^t(Z^t)^\top - Z_\star Z_\star^\top) + \frac{1}{2}(Z^t(Z^t)^\top + Z_\star Z_\star^\top) - Z_\star R(Z^t)^\top$$

Also due to the symmetry of $\nabla h(\Theta^t)$,

$$\big\langle \nabla h(\Theta^t), \frac{1}{2}(Z^t(Z^t)^\top + Z_\star Z_\star^\top) - Z_\star R(Z^t)^\top \big\rangle = \frac{1}{2}\big\langle \nabla h(\Theta^t), \Delta_{Z^t}\Delta_{Z^t}^\top \big\rangle$$

Therefore, combine the above three equations,

$$\begin{aligned} \left\|Z^{t+1} - Z_\star R^{t+1}\right\|_{\mathrm{F}}^2 \leq \left\|Z^t - Z_\star R^t\right\|_{\mathrm{F}}^2 + \eta_Z^2 \left\|\nabla h(\Theta^t) Z^t\right\|_{\mathrm{F}}^2 - \eta_Z \big\langle \nabla h(\Theta^t), \Delta_{Z^t}\Delta_{Z^t}^\top \big\rangle \\ - \eta_Z \big\langle \nabla h(\Theta^t), (Z^t(Z^t)^\top - Z_\star Z_\star^\top) \big\rangle \end{aligned} \tag{4.52}$$

By similar while much simpler argument, one will obtain

$$\begin{aligned} \|\alpha^{t+1} - \alpha_\star\|^2 &\leq \|\widetilde{\alpha}_{t+1} - \alpha_\star\|^2 \\ &= \|\alpha^t - \alpha_\star\|^2 + \eta_\alpha^2 \left\|\nabla h(\Theta^t) 1_n\right\|_{\mathrm{F}}^2 - 2\eta_\alpha \big\langle \nabla h(\Theta^t) 1_n, \alpha^t - \alpha_\star \big\rangle \end{aligned} \tag{4.53}$$

$$\begin{aligned} \|\beta^{t+1} - \beta_\star\|^2 &\leq \|\widetilde{\beta}_{t+1} - \beta_\star\|^2 \\ &= \|\beta^t - \beta_\star\|^2 + \eta_\beta^2 \big\langle \nabla h(\Theta^t), X \big\rangle^2 - 2\eta_\beta \big\langle \nabla h(\Theta^t), (\beta^t - \beta_\star) X \big\rangle \end{aligned} \tag{4.54}$$

Let $H(\Theta) = \mathbb{E}[\, h(\Theta)]$. With $\eta_Z = \eta / \left\|Z^0\right\|_{\mathrm{op}}^2, \eta_\alpha = \eta/2n, \eta_\beta = \eta/2\left\|X\right\|_{\mathrm{F}}^2$, the weighted sum $\left\|Z^0\right\|_{\mathrm{op}}^2 \times$(4.52)+$2n\times$(4.53)+$\|X\|_{\mathrm{F}}^2 \times$(4.54) is equivalent to

$$\begin{aligned} \widetilde{e}_{t+1} &\leq \widetilde{e}_t - \eta \big\langle \nabla h(\Theta^t), Z^t(Z^t)^\top - Z_\star Z_\star^\top + 2(\alpha^t - \alpha_\star)1_n{}^\top + (\beta^t - \beta_\star)X + \Delta_{Z^t}\Delta_{Z^t}^\top \big\rangle \\ &\quad + \Big(\left\|Z^0\right\|_{\mathrm{op}}^2 \eta_Z^2 \left\|\nabla h(\Theta^t) Z^t\right\|_{\mathrm{F}}^2 + 2n\eta_\alpha^2 \left\|\nabla h(\Theta^t) 1_n\right\|_{\mathrm{F}}^2 + \|X\|_{\mathrm{F}}^2 \eta_\beta^2 \big\langle \nabla h(\Theta^t), X \big\rangle^2 \Big) \\ &\leq \widetilde{e}_t - \eta \big\langle \nabla h(\Theta^t), \Delta_{\overline{\Theta}^t} \big\rangle - \eta \big\langle \nabla h(\Theta^t), \Delta_{Z^t}\Delta_{Z^t}^\top \big\rangle \\ &\quad + \Big(\frac{\eta^2}{\left\|Z^0\right\|_{\mathrm{op}}^2} \left\|\nabla h(\Theta^t) Z^t\right\|_{\mathrm{F}}^2 + \frac{\eta^2}{2n} \left\|\nabla h(\Theta^t) 1_n\right\|_{\mathrm{F}}^2 + \frac{\eta^2}{4\left\|X\right\|_{\mathrm{F}}^2} \big\langle \nabla h(\Theta^t), X \big\rangle^2 \Big), \end{aligned}$$

where $\Delta_{\overline{\Theta}^t} = Z^t(Z^t)^\top - Z_\star Z_\star^\top + \Delta_{\alpha^t} 1_n{}^\top + 1_n(\Delta_{\alpha^t})^\top + \Delta_{\beta^t} X = \Delta_{\Theta^t} - \overline{G}_k$. Then,

$$\begin{aligned}
\widetilde{e}_{t+1} &\leq \widetilde{e}_t - \eta\big\langle \nabla h(\Theta^t) - \nabla H(\Theta^t), \Delta_{\overline{\Theta}^t}\big\rangle - \eta\big\langle \nabla H(\Theta^t), \Delta_{\Theta^t}\big\rangle - \eta\big\langle \nabla H(\Theta^t), \overline{G}_k\big\rangle \\
&\quad - \eta\big\langle \nabla h(\Theta^t), \Delta_{Z^t}\Delta_{Z^t}^\top\big\rangle + \Big(\frac{\eta^2}{\|Z^0\|_{\text{op}}^2}\left\|\nabla h(\Theta^t) Z^t\right\|_{\text{F}}^2 + \frac{\eta^2}{2n}\left\|\nabla h(\Theta^t) 1_n\right\|_{\text{F}}^2 \\
&\quad + \frac{\eta^2}{4\|X\|_{\text{F}}^2}\big\langle \nabla h(\Theta^t), X\big\rangle^2\Big) \\
&\leq \widetilde{e}_t - \eta\big\langle \nabla H(\Theta^t), \Delta_{\Theta^t}\big\rangle + \eta|\big\langle \nabla h(\Theta^t) - \nabla H(\Theta^t), \Delta_{\overline{\Theta}^t}\big\rangle| + \eta|\big\langle \nabla h(\Theta), \Delta_{Z^t}\Delta_{Z^t}^\top\big\rangle| \\
&\quad + \eta|\big\langle \nabla H(\Theta^t), \overline{G}_k\big\rangle| + \eta^2\Big(\frac{1}{\|Z^0\|_{\text{op}}^2}\left\|\nabla h(\Theta^t) Z^t\right\|_{\text{F}}^2 + \frac{1}{2n}\left\|\nabla h(\Theta^t) 1_n\right\|_{\text{F}}^2 \\
&\quad + \frac{1}{4\|X\|_{\text{F}}^2}\big\langle \nabla h(\Theta^t), X\big\rangle^2\Big) \\
&= \widetilde{e}_t - \eta I_1 + \eta I_2 + \eta I_3 + \eta I_4 + \eta^2 I_5
\end{aligned} \tag{4.55}$$

Notice that for any $\Theta \in \mathcal{F}(n, k, M_1, M_2, X)$,

$$\frac{1}{4} I_{n^2\times n^2} \succeq \nabla^2 H(\Theta) = diag\Big(vec\big(\sigma(\Theta)\circ(1-\sigma(\Theta))\big)\Big) \succeq \tau I_{n^2\times n^2}$$

where $\tau = e^{M_1}/(1+e^{M_1})^2 \asymp e^{-M_1}$. Hence $H(\cdot)$ is τ-strongly convex and $\frac{1}{4}$-smooth. Further notice that $\nabla H(\Theta_\star) = 0$, then by Lemma 4.22,

$$I_1 = \big\langle \nabla H(\Theta^t), \Delta_{\Theta^t}\big\rangle \geq \frac{\tau/4}{\tau + 1/4}\|\Delta_{\Theta^t}\|_{\text{F}}^2 + \frac{1}{\tau + 1/4}\left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\text{F}}^2$$

By triangle inequality,

$$I_2 \leq |\big\langle \sigma(\Theta_\star) - A, Z^t(Z^t)^\top - Z_\star Z_\star^\top\big\rangle| + 2|\big\langle \sigma(\Theta_\star) - A, \Delta_{\alpha^t} 1_n{}^\top\big\rangle| + |\big\langle \sigma(\Theta_\star) - A, \Delta_{\beta^t} X\big\rangle|.$$

Recall that $\zeta_n = \max\{2\|A - P\|_{\text{op}},\ |\langle A - P, X/\|X\|_{\text{F}}\rangle|/\sqrt{k},\ 1\}$,

$$I_2 \leq \frac{\zeta_n}{2}\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_* + \zeta_n\|\Delta_{\alpha^t} 1_n{}^\top\|_* + \zeta_n\sqrt{k}\left\|\Delta_{\beta^t} X\right\|_{\text{F}}.$$

Notice that $Z^t(Z^t)^\top - Z_\star Z_\star^\top$ has rank at most $2k$,

$$I_2 \leq \frac{\zeta_n\sqrt{2k}}{2}\left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\mathrm{F}} + \zeta_n\left\|\Delta_{\alpha^t}1_n{}^\top\right\|_{\mathrm{F}} + \zeta_n\sqrt{k}\|\Delta_{\beta^t}X\|_{\mathrm{F}}.$$

Further by Cauthy-Schwartz inequality, there exists constant C_2 such that for any positive constant c_2 which we will specify later,

$$I_2 \leq c_2\left(\left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\mathrm{F}}^2 + 2\left\|\Delta_{\alpha^t}1_n{}^\top\right\|_{\mathrm{F}}^2 + \|\Delta_{\beta^t}X\|_{\mathrm{F}}^2\right) + \frac{C_2}{4c_2}\zeta_n^2 k$$

By Lemma 4.16, there exist constants c_1, C_1 such that

$$\begin{aligned} I_1 - I_2 \geq &\left(\frac{(1-c_1)\tau}{4\tau+1} - c_2\right)\left(\left\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\right\|_{\mathrm{F}}^2 + 2\|\Delta_{\alpha^t}1_n{}^\top\|_F^2 + \|\Delta_{\beta^t}X\|_{\mathrm{F}}^2\right) \\ &+ \frac{1}{\tau+1/4}\left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}}^2 - C_1\left\|\overline{G}_k\right\|_{\mathrm{F}}^2 - \frac{C_2}{4c_2}\zeta_n^2 k \end{aligned} \tag{4.56}$$

By Lemma 4.20,

$$\begin{aligned} I_1 - I_2 \geq &\frac{2(\sqrt{2}-1)}{\kappa^2}\left(\frac{(1-c_1)\tau}{4\tau+1} - c_2\right)e_t + \frac{1}{\tau+1/4}\left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}}^2 \\ &- C_1\left\|\overline{G}_k\right\|_{\mathrm{F}}^2 - \frac{C_2}{4c_2}\zeta_n^2 k \end{aligned}$$

To bound I_3, notice that $\Delta_{Z^t}\Delta_{Z^t}^\top$ is a positive definite matrix,

$$\begin{aligned} I_3 &\leq |\langle\nabla h(\Theta^t), \Delta_{Z^t}\Delta_{Z^t}^\top\rangle| \leq \left\|\nabla h(\Theta^t)\right\|_{\mathrm{op}}\left\|\Delta_{Z^t}\Delta_{Z^t}^\top\right\|_* \\ &= \left\|\nabla h(\Theta^t)\right\|_{\mathrm{op}}\operatorname{tr}(\Delta_{Z^t}\Delta_{Z^t}^\top) \leq \left\|\nabla h(\Theta^t)\right\|_{\mathrm{op}}\|\Delta_{Z^t}\|_{\mathrm{F}}^2 \\ &= \left\|\nabla h(\Theta^t) - \nabla H(\Theta^t) + \nabla H(\Theta^t)\right\|_{\mathrm{op}}\|\Delta_{Z^t}\|_{\mathrm{F}}^2 \\ &\leq \left\|\nabla h(\Theta^t) - \nabla H(\Theta^t)\right\|_{\mathrm{op}}\|\Delta_{Z^t}\|_{\mathrm{F}}^2 + \left\|\nabla H(\Theta^t)\right\|_{\mathrm{op}}\|\Delta_{Z^t}\|_{\mathrm{F}}^2 \\ &= \|\sigma(\Theta_\star) - A\|_{\mathrm{op}}\|\Delta_{Z^t}\|_{\mathrm{F}}^2 + \left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{op}}\|\Delta_{Z^t}\|_{\mathrm{F}}^2 \\ &\leq \frac{\zeta_n}{2}\|\Delta_{Z^t}\|_{\mathrm{F}}^2 + \left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}}\|\Delta_{Z^t}\|_{\mathrm{F}}^2 \end{aligned} \tag{4.57}$$

By the assumption that $\|\Delta_{Z^t}\|_{\mathrm{F}} \leq \frac{c_0}{e^{M_1}\kappa^2} \|Z_\star\|_{\mathrm{op}}$,

$$\begin{aligned}\left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}} \|\Delta_{Z^t}\|_{\mathrm{F}}^2 &\leq \frac{c_0}{e^{M_1}\kappa^2} \left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}} \|\Delta_{Z^t}\|_{\mathrm{F}} \|Z_\star\|_{\mathrm{op}} \\ &\leq c_3 \left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}}^2 + \frac{c_0}{4c_3 e^{M_1}\kappa^2} \|\Delta_{Z^t}\|_{\mathrm{F}}^2 \|Z_\star\|_{\mathrm{op}}^2\end{aligned} \tag{4.58}$$

for any constant c_3 to be specified later. Then

$$I_3 \leq \left(\frac{\zeta_n}{2\|Z_\star\|_{\mathrm{op}}^2} + \frac{c_0}{4c_3 e^{M_1}\kappa^2}\right) e_t + c_3 \left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}}^2 \tag{4.59}$$

By the assumption that $\|Z_\star\|_{\mathrm{op}}^2 \geq C_0 \kappa^2 \zeta_n e^{M_1}$ for sufficiently large constant C_0,

$$I_3 \leq \left(\frac{1}{2C_0 e^{M_1}\kappa^2} + \frac{c_0}{4c_3 e^{M_1}\kappa^2}\right) e_t + c_3 \left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}}^2 \tag{4.60}$$

By basically inequalities,

$$\begin{aligned}I_4 &= |\langle \nabla H(\Theta^t), \overline{G}_k\rangle| = |\langle \sigma(\Theta^t) - \sigma(\Theta_\star), \overline{G}_k\rangle| \leq \left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}} \left\|\overline{G}_k\right\|_{\mathrm{F}} \\ &\leq c_4 \left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}}^2 + \frac{1}{4c_4} \left\|\overline{G}_k\right\|_{\mathrm{F}}^2\end{aligned} \tag{4.61}$$

for any constant c_4 to be specified later. The final step is to control I_5 by upper bounding its three terms separately,

$$\begin{aligned}\left\|\nabla h(\Theta^t) Z^t\right\|_{\mathrm{F}}^2 &= \left\|\left(\nabla h(\Theta^t) - \nabla H(\Theta^t)\right) Z^t + \nabla H(\Theta^t) Z^t\right\|_{\mathrm{F}}^2 \\ &\leq 2\Big(\left\|(\nabla h(\Theta^t) - \nabla H(\Theta^t)) Z^t\right\|_{\mathrm{F}}^2 + \left\|\nabla H(\Theta^t) Z^t\right\|_{\mathrm{F}}^2 \Big) \\ &\leq 2\Big(\left\|(\sigma(\Theta_\star) - A) Z^t\right\|_{\mathrm{F}}^2 + \left\|(\sigma(\Theta^t) - \sigma(\Theta_\star)) Z^t\right\|_{\mathrm{F}}^2 \Big) \\ &\leq 2\Big(\|\sigma(\Theta_\star) - A\|_{\mathrm{op}}^2 \left\|Z^t\right\|_{\mathrm{F}}^2 + \left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}}^2 \left\|Z^t\right\|_{\mathrm{op}}^2 \Big) \\ &\leq 2\Big(\frac{\zeta_n^2}{4} \left\|Z^t\right\|_{\mathrm{F}}^2 + \left\|Z^t\right\|_{\mathrm{op}}^2 \left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}}^2 \Big)\end{aligned} \tag{4.62}$$

$$
\begin{aligned}
\left\|\nabla h(\Theta^t) 1_n\right\|^2 &= \left\|\left(\nabla h(\Theta^t) - \nabla H(\Theta^t)\right) 1_n + \nabla H(\Theta^t) 1_n\right\|^2 \\
&\leq 2\Big(\left\|\left(\nabla h(\Theta^t) - \nabla H(\Theta^t)\right) 1_n\right\|^2 + \left\|\nabla H(\Theta^t) 1_n\right\|^2\Big) \\
&\leq 2\Big(\left\|(\sigma(\Theta_\star) - A) 1_n\right\|^2 + \left\|(\sigma(\Theta^t) - \sigma(\Theta_\star)) 1_n\right\|^2 \Big) \\
&\leq 2n\Big(\frac{\zeta_n^2}{4} + \left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}}^2 \Big)
\end{aligned} \tag{4.63}
$$

$$
\begin{aligned}
\left\langle \nabla H(\Theta^t), X\right\rangle^2 &= \Big(\left\langle \nabla h(\Theta^t) - \nabla H(\Theta^t), X\right\rangle + \left\langle \nabla H(\Theta^t), X\right\rangle \Big)^2 \\
&\leq 2\Big(\left\langle \sigma(\Theta_\star) - A, X\right\rangle^2 + \left\langle \sigma(\Theta^t) - \sigma(\Theta_\star), X\right\rangle^2 \Big) \\
&\leq 2\Big(\zeta_n^2 k \left\|X\right\|_{\mathrm{F}}^2 + \left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}}^2 \left\|X\right\|_{\mathrm{F}}^2 \Big)
\end{aligned} \tag{4.64}
$$

When $\operatorname{dist}(Z^t, Z_\star) \leq c \left\|Z_\star\right\|_{\mathrm{op}}$, adding these inequalities will yield

$$
I_5 \leq \left(\frac{\left\|Z^t\right\|_{\mathrm{op}}^2}{\left\|Z_\star\right\|_{\mathrm{op}}^2} \frac{\zeta_n^2 k}{2} + \frac{\zeta_n^2}{4} + \frac{\zeta_n^2 k}{2} \right) + \left(\frac{2\left\|Z^t\right\|_{\mathrm{op}}^2}{\left\|Z_\star\right\|_{\mathrm{op}}^2} + \frac{3}{2} \right) \left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}}^2 \tag{4.65}
$$

By the assumption that $\left\|\Delta_{Z^t}\right\|_{\mathrm{F}} \leq \frac{c_0}{e^{M_1}\kappa^2} \left\|Z_\star\right\|_{\mathrm{op}}$ for some sufficiently small c_0,

$$
I_5 \leq C_5 \left(\zeta_n^2 k + \left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}}^2 \right). \tag{4.66}
$$

Combine equations (4.56), (4.60), (4.61), (4.66),

$$
\begin{aligned}
\widetilde{e}_{t+1} \leq \widetilde{e}_t &- \eta \left(\frac{2(\sqrt{2}-1)}{\kappa^2} \left(\frac{(1-c_1)\tau}{4\tau+1} - c_2 \right) - \frac{1}{2C_0 e^{M_1}\kappa^2} + \frac{c_0}{4c_3 e^{M_1}\kappa^2} \right) e_t \\
&+ \eta \left(C_1 + \frac{1}{4c_4} \right) \left\|\overline{G}_k\right\|_{\mathrm{F}}^2 - \left(\frac{1}{\tau + 1/4} - c_3 - c_4 - C_5\eta \right) \left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}}^2 \\
&+ \eta \frac{C_2}{4c_2} \zeta_n^2 k + \eta^2 C_5 \zeta_n^2 k
\end{aligned} \tag{4.67}
$$

where c_2, c_3, c_4 are arbitrary constants, c_0 is a sufficiently small constant, and C_0 is a sufficiently large constant. Notice that $\tau \asymp e^{-M_1}$. Choose $c_2 = c\tau$ and c, c_3, c_4, η small enough such that

$$\begin{aligned} 2(\sqrt{2}-1)\left(\frac{(1-c_1)\tau}{4\tau+1}-c_2\right)-\frac{1}{2e^{M_1}C_0}-\frac{c_0}{4c_3e^{M_1}} > \widetilde{\rho}e^{-M_1} \\ \frac{1}{\tau+1/4}-c_3-c_4-C_5\eta \geq 0, \end{aligned} \tag{4.68}$$

for some positive constant $\widetilde{\rho}$. Recall that $\widetilde{e}_t \geq (1-\delta)e_t$. Then there exists a universal constant $C>0$ such that

$$\widetilde{e}_{t+1} \leq \left(1-\frac{\eta}{e^{M_1}\kappa^2}\widetilde{\rho}(1-\delta)\right)\widetilde{e}_t+\eta C\left(\left\|\overline{G}_k\right\|_{\mathrm{F}}^2+e^{M_1}\zeta_n^2k\right). \tag{4.69}$$

The proof is finished by setting $\rho=(1-\delta)\widetilde{\rho}$.

4.7.7. Proof of Lemma 4.18

We prove this by induction. For the base case,

$$\|\Delta_{Z^0}\|_{\mathrm{F}} \leq \left(\frac{\widetilde{e}_0}{\left\|Z^0\right\|_{\mathrm{op}}^2}\right)^{\frac{1}{2}} \leq \left(\frac{c_0^2}{4e^{2M_1}\kappa^4}\frac{\|Z_\star\|_{\mathrm{op}}^4}{\left\|Z^0\right\|_{\mathrm{op}}^2}\right)^{\frac{1}{2}} = \frac{c_0}{2e^{M_1}\kappa^2}\|Z_\star\|_{\mathrm{op}}\frac{\|Z_\star\|_{\mathrm{op}}}{\left\|Z^0\right\|_{\mathrm{op}}} \leq \frac{c_0}{e^{M_1}\kappa^2}\|Z_\star\|_{\mathrm{op}},$$

where the last inequality is by the fact that,

$$\left\|Z^0\right\|_{\mathrm{op}} \geq \|Z_\star\|_{\mathrm{op}}-\|\Delta_{Z^0}\|_{\mathrm{op}} \geq \left(1-\frac{c_0}{2e^{M_1}\kappa^2}\right)\|Z_\star\|_{\mathrm{op}} \geq \frac{3}{4}\|Z_\star\|_{\mathrm{op}}.$$

Suppose the claim is true for all $t \leq t_0$, by Lemma 4.17,

$$\begin{aligned} \widetilde{e}_{t_0+1} &\leq \left(1-\frac{\eta}{e^{M_1}\kappa^2}\rho\right)^{t_0}\widetilde{e}_0+\eta C\left(\left\|\overline{G}_k\right\|_{\mathrm{F}}^2+e^{M_1}\zeta_n^2k\right) \\ &\leq \widetilde{e}_0+\eta C\left(\left\|\overline{G}_k\right\|_{\mathrm{F}}^2+e^{M_1}\zeta_n^2k\right) \\ &\leq \frac{c_0^2}{4e^{2M_1}\kappa^4}\|Z_\star\|_{\mathrm{op}}^4+\eta C\left(\left\|\overline{G}_k\right\|_{\mathrm{F}}^2+e^{M_1}\zeta_n^2k\right) \\ &= \frac{c_0^2}{e^{2M_1}\kappa^4}\|Z_\star\|_{\mathrm{op}}^4\left(\frac{1}{4}+\eta\frac{Ce^{2M_1}\zeta_n^2\kappa^4}{c_0^2\|Z_\star\|_{\mathrm{op}}^4}\left(\frac{\left\|\overline{G}_k\right\|_{\mathrm{F}}^2}{\zeta_n^2}+e^{M_1}k\right)\right) \\ &\leq \frac{c_0^2}{e^{2M_1}\kappa^4}\|Z_\star\|_{\mathrm{op}}^4\left(\frac{1}{4}+\frac{C}{c_0^2C_1^2}\right) \end{aligned} \tag{4.70}$$

Choosing C_1 large enough such that $C_1^2 \geq \frac{4C}{c_0^2}$, then

$$\widetilde{e}_{t_0+1} \leq \frac{c_0^2}{2e^{2M_1}\kappa^4} \|Z_\star\|_{\text{op}}^4 \tag{4.71}$$

and therefore,

$$\|\Delta_{Z^{t_0+1}}\|_{\text{F}} \leq \left(\frac{\widetilde{e}_{t_0+1}}{\|Z_\star\|_{\text{op}}^2}\right)^{\frac{1}{2}} \leq \frac{c_0}{\sqrt{2}e^{M_1}\kappa^2} \|Z_\star\|_{\text{op}} \frac{\|Z_\star\|_{\text{op}}}{\|Z^0\|_{\text{op}}} \leq \frac{c_0}{e^{M_1}\kappa^2} \|Z_\star\|_{\text{op}}. \tag{4.72}$$

4.7.8. Technique Lemmas

Lemma 4.19 (Chung and Lu (2002))**.** *Let $X_1, \cdots, X_n$ be independent Bernoulli random variables with $P(X_i = 1) = p_i$. For $S_n = \sum_{i=1}^n a_i X_i$ and $\nu = \sum_{i=1}^n a_i^2 p_i$. Then we have*

$$P(S_n - \mathbb{E}S_n < -\lambda) \leq exp(-\lambda^2/2\nu)$$
$$P(S_n - \mathbb{E}S_n > \lambda) \leq exp\left(-\frac{\lambda^2}{2(\nu + a\lambda/3)}\right)$$

where $a = \max\{a_1, \cdots, a_n\}$.

Lemma 4.20 (Tu et al. (2015))**.** *For any $Z_1, Z_2 \in \mathbb{R}^{n\times k}$, we have*

$$dist(Z_1, Z_2)^2 \leq \frac{1}{2(\sqrt{2}-1)\sigma_k^2(Z_1)} \left\|Z_1 Z_1^\top - Z_2 Z_2^\top\right\|_{\text{F}}^2 \tag{4.73}$$

Lemma 4.21 (Tu et al. (2015))**.** *For any $Z_1, Z_2 \in \mathbb{R}^{n\times k}$ such that $dist(Z_1, Z_2) \leq c\|Z_1\|_{\text{op}}$, we have*

$$\left\|Z_1 Z_1^\top - Z_2 Z_2^\top\right\|_{\text{F}} \leq (2 + c)\|Z_1\|_{\text{op}} dist(Z_1, Z_2) \tag{4.74}$$

Lemma 4.22 (Nesterov (2004))**.** *For a continuously differentiable function f, if it is μ-*

strongly convex and L-smooth on a convex domain $\mathcal{D}$, say for any $x, y \in \mathcal{D}$,

$$\frac{\mu}{2}\|x-y\|^2 \leq f(y) - f(x) - \big\langle f'(x), y-x \big\rangle \leq \frac{L}{2}\|x-y\|^2 \tag{4.75}$$

then

$$\big\langle f'(x) - f'(y), x-y \big\rangle \geq \frac{\mu L}{\mu + L}\|x-y\|^2 + \frac{1}{\mu+L}\|f'(x) - f'(y)\|^2 \tag{4.76}$$

and also

$$\big\langle f'(x) - f'(y), x-y \big\rangle \geq \mu\|x-y\|^2 \tag{4.77}$$

Lemma 4.23 (Lei and Rinaldo (2015))**.** *Let A be the adjacency matrix of a random graph on n nodes in which edges occur independently. Let $\mathbb{E}[A] = P$ and assume that $n \max_{i,j} P_{ij} \leq d$ and for $d \geq c_0 \log(n)$ for some $c_0 \geq 0$. Then for any C_0, there is a constant $C = C(C_0, c_0)$ such that*

$$\|A - P\|_{\mathrm{op}} \leq C\sqrt{d} \tag{4.78}$$

with probability at least $1 - n^{-C_0}$

Lemma 4.24. *Let A be the adjacency matrix of a random graph of n nodes in which edges occur independently and $\mathbb{E}[A] = P$. Then,*

$$|\langle A - P, X\rangle| \leq C\, \|X\|_{\mathrm{F}} \tag{4.79}$$

with probability at least $1 - 2exp(-C^2/8p_{\max}) - exp(-C^2\, \|X\|_{\mathrm{F}}\, /8\|X\|_\infty)$

Proof of Lemma 4.24 Observe that $\langle A - P, X\rangle = 2\sum_{i<j}(A_{ij} - P_{ij})X_{ij}$ and A_{ij} are independent Bernoulli random variables with $\mathbb{E}[A_{ij}] = P_{ij}$. Apply Lemma 4.19 to

$\sum_{i<j}(A_{ij} - P_{ij})X_{ij}$ with $\lambda = C\left\|X\right\|_{\mathrm{F}}/2$, we have $\nu = \sum_{i<j} X_{ij}^2 P_{ij} \leq p_{\max}\left\|X\right\|_{\mathrm{F}}^2$ and

$$\begin{aligned}
P\big(|\langle A-P, X\rangle| \leq C\left\|X\right\|_{\mathrm{F}}\big) &\leq exp(-C^2\left\|X\right\|_{\mathrm{F}}^2/8\nu) + exp\left(-\frac{C^2\left\|X\right\|_{\mathrm{F}}^2}{8\max\{\nu, C\|X\|_\infty\left\|X\right\|_{\mathrm{F}}\}}\right)\\
&\leq 2exp(-C^2\left\|X\right\|_{\mathrm{F}}^2/8\nu) + exp(-C^2\left\|X\right\|_{\mathrm{F}}/8\|X\|_\infty)\\
&\leq 2exp(-C^2/8p_{\max}) + exp(-C^2\left\|X\right\|_{\mathrm{F}}/8\|X\|_\infty)
\end{aligned}$$

□

BIBLIOGRAPHY

L. A. Adamic and N. Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM, 2005.

D. Agarwal and B.-C. Chen. Regression-based latent factor models. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 19–28. ACM, 2009.

E. M. Airoldi, T. B. Costa, and S. H. Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.

Z. Allen-Zhu and Y. Li. Doubly accelerated methods for faster cca and generalized eigendecomposition. *arXiv preprint arXiv:1607.06017*, 2016.

T. W. Anderson. Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1):122–148, 1963.

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, NY, second edition, 1984.

T. W. Anderson. Asymptotic theory for canonical correlation analysis. *Journal of Multivariate Analysis*, 70(1):1–29, 1999.

R. Arora and K. Livescu. Multi-view cca-based acoustic features for phonetic recognition across speakers and domains. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7135–7139. IEEE, 2013.

R. Arora, A. Cotter, K. Livescu, and N. Srebro. Stochastic optimization for pca and pls. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 861–868. IEEE, 2012.

H. Avron, C. Boutsidis, S. Toledo, and A. Zouzias. Efficient dimensionality reduction for canonical correlation analysis. In *ICML (1)*, pages 347–355, 2013.

A. Balsubramani, S. Dasgupta, and Y. Freund. The fast convergence of incremental pca. In *Advances in Neural Information Processing Systems*, pages 3174–3182, 2013.

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

N. Binkiewicz, J. T. Vogelstein, and K. Rohe. Covariate assisted spectral clustering. *stat*, 1050:4, 2015.

A. Birnbaum, I. M. Johnstone, B. Nadler, and D. Paul. Minimax bounds for sparse pca with noisy high-dimensional data. *Annals of statistics*, 41(3):1055, 2013.

A. Bjorck and G. H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of computation*, 27(123):579–594, 1973.

L. Bottou. Large-Scale Machine Learning with Stochastic Gradient Descent. In Y. Lechevallier and G. Saporta, editors, *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pages 177–187, Paris, France, Aug. 2010. Springer.

O. Bousquet and L. Bottou. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 161–168, 2008.

S. Burer and R. D. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.

T. Cai, Z. Ma, and Y. Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability theory and related fields*, 161(3-4):781–815, 2015.

T. T. Cai and A. Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *arXiv preprint arXiv:1605.00353*, 2016.

T. T. Cai, Z. Ma, Y. Wu, et al. Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.

E. Candès and B. Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251, 2015.

E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

S. Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual international conference on machine learning*, pages 129–136. ACM, 2009.

M. Chen, C. Gao, Z. Ren, and H. H. Zhou. Sparse cca via precision adjusted iterative thresholding. *arXiv preprint arXiv:1311.6186*, 2013.

X. Chen, H. Liu, and J. G. Carbonell. Structured sparse canonical correlation analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 199–207, 2012.

Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

Y. Chen, X. Li, and J. Xu. Convexified modularity maximization for degree-corrected stochastic block models. *arXiv preprint arXiv:1512.08425*, 2015.

F. Chung and L. Lu. Connected components in random graphs with given expected degree sequences. *Annals of combinatorics*, 6(2):125–145, 2002.

P. S. Dhillon, D. Foster, and L. Ungar. Multi-view learning of word embeddings via cca. In *Advances in Neural Information Processing Systems (NIPS)*, volume 24, 2011.

P. S. Dhillon, J. Rodu, D. P. Foster, and L. H. Ungar. Two step cca: A new spectral method for estimating vector models of words. In *Proceedings of the 29th International Conference on Machine learning*, ICML'12, 2012.

R. L. Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983.

M. Faruqui and C. Dyer. Improving vector space word representations using multilingual correlation. Association for Computational Linguistics, 2014.

M. Fiedler. Notes on hilbert and cauchy matrices. *Linear Algebra and its Applications*, 432 (1):351–356, 2010.

D. P. Foster, R. Johnson, S. M. Kakade, and T. Zhang. Multi-view dimensionality reduction via canonical correlation analysis. Technical report, 2008.

O. Friman, M. Borga, P. Lundberg, and H. Knutsson. Adaptive analysis of fmri data. *NeuroImage*, 19(3):837–845, 2003.

K. Fukumizu, F. R. Bach, and M. I. Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*, pages 1871–1905, 2009.

C. Gao, Z. Ma, and H. H. Zhou. Sparse cca: Adaptive estimation and computational barriers. *arXiv preprint arXiv:1409.8565*, 2014.

C. Gao, Y. Lu, H. H. Zhou, et al. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015a.

C. Gao, Z. Ma, Z. Ren, H. H. Zhou, et al. Minimax estimation in sparse canonical correlation analysis. *The Annals of Statistics*, 43(5):2168–2197, 2015b.

C. Gao, Y. Lu, Z. Ma, and H. H. Zhou. Optimal estimation and completion of matrices with biclustering structures. *Journal of Machine Learning Research*, 17(161):1–29, 2016.

R. Ge, C. Jin, P. Netrapalli, A. Sidford, et al. Efficient algorithms for large-scale generalized eigenvector computation and canonical correlation analysis. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2741–2750, 2016a.

R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016b.

A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2010.

G. H. Golub and H. Zha. *The canonical correlations of matrix pairs and their numerical computation.* Springer, 1995.

Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 106 (2):210–233, 2014.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, 2011. ISSN 0036-1445.

M. S. Handcock, A. E. Raftery, and J. M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170 (2):301–354, 2007.

P. D. Hoff. Random effects models for network data. In *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers.* Citeseer, 2003.

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

R. A. Hom and C. R. Johnson. Topics in matrix analysis. *Cambridge UP, New York*, 1991.

H. Hotelling. Relations between two sets of variables. *Biometrika*, 28:312–377, 1936.

P. Hsu. On the limiting distribution of the canonical correlations. *Biometrika*, 32(1):38–45, 1941.

A. J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

J. Jin. Fast community detection by score. *The Annals of Statistics*, 43(1):57–89, 2015.

S. M. Kakade and D. P. Foster. Multi-view regression via canonical correlation analysis. In *In Proc. of Conference on Learning Theory*, 2007.

R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010a.

R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(Jul):2057–2078, 2010b.

T.-K. Kim, S.-F. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

O. Klopp, A. B. Tsybakov, and N. Verzelen. Oracle inequalities for network models and sparse graphon estimation. *arXiv preprint arXiv:1507.04118*, 2015.

V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, pages 2302–2329, 2011.

P. N. Krivitsky and M. S. Handcock. Fitting position latent cluster models for social networks with latentnet. In *Journal of Statistical Software*. Citeseer.

P. N. Krivitsky, M. S. Handcock, A. E. Raftery, and P. D. Hoff. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, 31(3):204–213, 2009.

M. Lamar, Y. Maron, M. Johnson, and E. Bienenstock. SVD and Clustering for Unsupervised POS Tagging. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 215–219, Uppsala, Sweden, 2010. Association for Computational Linguistics.

E. Lazega. *The collegial phenomenon: The social mechanisms of cooperation among peers in a corporate law partnership.* Oxford University Press on Demand, 2001.

J. Lei and A. Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

E. Liberty, F. Woolfe, P.-G. Martinsson, V. Rokhlin, and M. Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of the National Academy of Sciences*, 104(51):20167–20172, 2007.

C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1988–1995. IEEE, 2009.

Y. Lu and D. P. Foster. Large scale canonical correlation analysis with iterative least squares. In *Advances in Neural Information Processing Systems*, pages 91–99, 2014.

J. Ma, L. K. Saul, S. Savage, and G. M. Voelker. Identifying suspicious urls: An application of large-scale online learning. In *In Proc. of the International Conference on Machine Learning (ICML*, 2009.

Z. Ma and X. Li. Subspace perspective on canonical correlation analysis: Dimension reduction and minimax rates. *arXiv preprint arXiv:1605.03662*, 2016.

Z. Ma and Z. Ma. Exploration of large networks via fast and universal latent space model fitting. *manuscript*, 2017.

Z. Ma, D. Foster, and R. Stine. Adaptive monotone shrinkage for regression. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 533–542. AUAI Press, 2014a.

Z. Ma, Z. Ma, and T. Sun. Adaptive estimation in two-way sparse reduced-rank regression. *arXiv preprint arXiv:1403.1922*, 2014b.

Z. Ma, Y. Lu, and D. Foster. Finding linear structure in large datasets with scalable canonical correlation analysis. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 169–178, 2015.

R. Mathias. The hadamard operator norm of a circulant and applications. *SIAM journal on matrix analysis and applications*, 14(4):1152–1167, 1993.

A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.

I. Mitliagkas, C. Caramanis, and P. Jain. Memory limited, streaming pca. In *Advances in Neural Information Processing Systems*, pages 2886–2894, 2013.

Y. Nesterov. *Introductory lectures on convex optimization*, volume 87. Springer Science & Business Media, 2004.

E. Oja and J. Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of mathematical analysis and applications*, 106(1):69–84, 1985.

A. B. Owen and P. O. Perry. Bi-cross-validation of the svd and the nonnegative matrix factorization. *The annals of applied statistics*, pages 564–594, 2009.

N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260. ACM, 2010.

M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.*, 18:no. 82, 1–9, 2013. ISSN 1083-589X. doi: 10.1214/ECP.v18-2865. URL `http://ecp.ejpecp.org/article/view/2865`.

T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 143–152. IEEE, 2006.

I. J. Schoenberg. On certain metric spaces arising from Euclidean spaces by a change of metric and their imbedding in Hilbert space. *Annals of Mathematics*, pages 787–793, 1937.

I. J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

C. G. Snoek, M. Worring, J. C. Van Gemert, J.-M. Geusebroek, and A. W. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 421–430. ACM, 2006.

K. Sridharan and S. M. Kakade. An information theoretic framework for multi-view learning. In R. A. Servedio and T. Zhang, editors, *COLT*, pages 403–414. Omnipress, 2008. URL `http://dblp.uni-trier.de/db/conf/colt/colt2008.html#SridharanK08`.

R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

S. J. Szarek. Nets of grassmann manifold and orthogonal group. In *Proceedings of research workshop on Banach space theory (Iowa City, Iowa, 1981)*, volume 169, page 185, 1982.

A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM review*, 53(3):526–543, 2011.

A. L. Traud, P. J. Mucha, and M. A. Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012.

S. Tu, R. Boczar, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

W. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1083–1092, 2015.

W. Wang, J. Wang, D. Garber, and N. Srebro. Efficient globally convergent stochastic optimization for canonical correlation analysis. In *Advances in Neural Information Processing Systems*, pages 766–774, 2016.

P. Å. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

P. Å. Wedin. On angles between subspaces of a finite dimensional inner product space. In *Matrix Pencils*, pages 263–285. Springer, 1983.

A. Weinstein, Z. Ma, L. D. Brown, and C.-H. Zhang. Group-linear empirical bayes estimates for a heteroscedastic normal mean. *arXiv preprint arXiv:1503.08503*, 2015.

D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008, 2009.

P. J. Wolfe and S. C. Olhede. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, 2013.

F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert. A fast randomized algorithm for the approximation of matrices. *Applied and Computational Harmonic Analysis*, 25(3):335–366, 2008.

C. Xu, D. Tao, and C. Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

B. Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

Y. Zhang, E. Levina, and J. Zhu. Detecting overlapping communities in networks using spectral methods. *arXiv preprint arXiv:1412.3432*, 2014.

Y. Zhang, E. Levina, and J. Zhu. Community detection in networks with node features. *arXiv preprint arXiv:1509.01173*, 2015.

Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *stat*, 1050:23, 2016.