STATISTICAL INFERENCE FOR MULTI-VIEW DATA

A Dissertation

by

YUNFENG ZHANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Irina Gaynanova
Committee Members,	Jianhua Huang
	Xianyang Zhang
	Xiaoning Qian
Head of Department,	Daren B.H. Cline

August 2020

Major Subject: Statistics

ABSTRACT

Multi-view data, that is matched sets of measurements on the same subjects, have become increasingly common with technological advances in genomics, neuroscience and wearable technologies, etc. Despite its prevalence, traditional techniques for classification or association analysis cannot be applied to multi-view data since they do not take into account the heterogeneity between the views. In this dissertation, we focus on generalizing the existing high-dimensional methods to multi-view data. First, we propose a framework for the Joint Association and Classification Analysis of multi-view data (JACA). We support the methodology with theoretical guarantees for estimation consistency in high-dimensional settings, and numerical comparisons with existing methods. In addition, our approach is capable of using partial information where class labels or subsets of views are missing. Second, we investigate the Pan-Cancer data with a goal to assess the strength of association between different cellular composition estimations by exploring the Generalized Association Study framework. We extract the shared and individual signals from each view, and evaluate the relationship they have with the survival to find out the bio-markers that are predictive for cancer prognosis. Lastly, we propose a low-rank canonical correlation analysis framework to model heterogeneous data (both Gaussian and non-Gaussian) using exponential family distributions. We exploit a decomposition-based strategy to extract shared and individual structures from underlying natural parameter matrices. In contrast to existing methods, our approach guarantees that there is no shared information embedded in the individual structures. An alternating split orthogonal constraints algorithm is developed to estimate the model parameters, and simulation studies show the advantages of the proposed approach over other classical methods.

DEDICATION

To my family for their endless love.

ACKNOWLEDGMENTS

I would like to express my greatest gratitude to my advisor Dr. Irina Gaynanova, for her enthusiasm and constant support. She guided me through my Ph.D study with her extensive experience and provided lots of valuable suggestions on both research and career decisions. Over the past five years, she always inspired me to explore new fields and encouraged me to implement new ideas. Her professional coaching and immense patience have made this dissertation possible. It is an honor to be able to work with her.

My sincere thanks also go to my committee members, Dr. Jianhua Huang, Dr. Xianyang Zhang and Dr. Xiaoning Qian, for the helpful discussions and comments on my research work. Dr. Huang has been a great and caring mentor since my first year, and his door was always open when I needed his help. From course selection to career choices, his advises have always been prudent and illuminating.

I would also like to thank everyone in the Department of Statistics, including the alumni. They have been very supportive and friendly, which kept me passionate about research and daily life. The conversations and collaborations we had together would be my priceless memories.

Lastly, I want to express my special thanks to my parents and my girlfriend. I have been away from home for nine years and could only visit my parents during the winter/summer breaks. Allowing their only child to go abroad was definitely a hard decision for them to make. I know I would not have made it without their unconditional supports.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Committee Members Dr. Irina Gaynanova, Dr. Jianhua Huang and Dr. Xianyang Zhang of the Department of Statistics and Dr. Xiaoning Qian of the Department of Electrical and Computer Engineering.

The data analyzed for Chapter 2 and Chapter 3 was obtained from TCGA2STAT R package and The Cancer Genome Atlas project, which is available on https://tcga-data.nci.nih.gov/docs/publications. The data analyzed for Chapter 4 was provided by Dr. Wenyi Wang from the MD Anderson Cancer Center.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a graduate assistantship from the Department of Statistics at Texas A&M University and a grant from National Science Foundation DMS-1712943.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 Background

Multi-view data, that is matched sets of measurements on the same subjects, have become increasingly common with technological advances in genomics and other fields. For example, The Cancer Genome Atlas Project (Weinstein et al., 2013) contains multiple views for the same set of subjects, such as gene expression, genotype, metabolic measurements, etc. Each of them can be considered as a view. In contrast to traditional high-dimensional data, multi-view usually have some unique characteristics. On the one hand, since all the views have the same underlying subjects, they share common information and are intrinsically correlated with each other. Therefore, it is of interest to analyze associations across the views and use this information to further perform supervised/unsupervised analysis. On the other hand, the views also contain heterogeneity, such as scales, biological meanings and the types of data. This unique feature has prevented us from applying the traditional methods to multi-view data.

In this dissertation, we aim to address two major challenges encountered in the analysis of multi-view data. First, how to conduct classification analysis and association analysis when there is class information available. In practice, multi-view data share the same underlying subjects, and the subjects are often separated into known classes. Hence how to exploit this extra information to assist association analysis is a popular topic, especially in biomedical studies. Meanwhile, predicting class assignments using multi-view data is also a prime interest in many fields, like cancer study and web page classification (Zhao et al., 2017). In the literature, the association and classification problems are usually addressed separately, and the joint analysis of them should increase the estimation efficiency and provide more insights of the data. One of our goals is to develop a framework that can simultaneously learn the classification rule and interrelationships between the views based on multi-view data, and also apply it to modern datasets to provide scientific discoveries.

The second challenge is how to conduct association analysis for multi-view data with hetero-

geneous types. While the abundance of data sources makes collecting data much easier, the downstream datasets from different platforms also tend to have different types, such as non-negative, proportions or binary. For example, the transcriptomics deconvolution in bulk tumor samples is a popular topic in genome research, and several tools (Newman et al., 2015; Li et al., 2017; Wang et al., 2018) have been developed to estimate the cell composition of tumor samples from different perspectives. Therefore, it is beneficial to find out what information is shared between them and identify which cellular types are predictive of cancer prognosis. Nevertheless, the type of datasets is proportion instead of real-valued. Hence, the classical Canonical correlation analysis (CCA) method is not an appropriate choice in this scenario since correlations are usually not well-defined for proportion data. Another goal of our work is to assess the strength of association between different cell composition estimations by exploring the Generalized association study (GAS) framework (Li and Gaynanova, 2018). Furthermore, we will incorporate this information to perform survival analysis, and therefore identify key signals that are predictive of cancer prognosis.

1.2 Review of Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a commonly used methodology, and our works are based on investigating CCA in different settings. Specifically, CCA aims to find linear associations between the two datasets. It seeks linear combinations of two matrices of real-valued random variables that have the largest correlation.

Consider two mean zero random vectors $\boldsymbol{x}_1 \in \mathbb{R}^{p_1}$ and $\boldsymbol{x}_2 \in \mathbb{R}^{p_2}$ with $\boldsymbol{\Sigma}_1 = \mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1^\top)$, $\boldsymbol{\Sigma}_2 = \mathbb{E}(\boldsymbol{x}_2\boldsymbol{x}_2^\top)$, $\boldsymbol{\Sigma}_{12} = \mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_2^\top)$ and $r = \operatorname{rank}(\boldsymbol{\Sigma}_{12}) > 0$. The population CCA seeks linear combinations $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ that maximize $\operatorname{Cor}(\boldsymbol{\theta}_1^\top\boldsymbol{x}_1, \boldsymbol{\theta}_2^\top\boldsymbol{x}_2)$, that is it seeks at most r pairs $(\boldsymbol{\theta}_1^{(k)}, \boldsymbol{\theta}_2^{(k)})$ that satisfy

$$
\begin{aligned}
(\boldsymbol{\theta}_1^{(k)}, \boldsymbol{\theta}_2^{(k)}) = \underset{\boldsymbol{w}_1^{(k)}, \boldsymbol{w}_2^{(k)}}{\operatorname{argmax}} & \left\{ \boldsymbol{w}_1^{(k)\top}\boldsymbol{\Sigma}_{12}\boldsymbol{w}_2^{(k)} \right\} \\
\text{subject to} \quad & \boldsymbol{w}_1^{(k)\top}\boldsymbol{\Sigma}_1\boldsymbol{w}_1^{(k)} = 1, \boldsymbol{w}_2^{(k)\top}\boldsymbol{\Sigma}_2\boldsymbol{w}_2^{(k)} = 1, \\
& \boldsymbol{w}_1^{(k)\top}\boldsymbol{\Sigma}_1\boldsymbol{w}_1^{(j)} = 0,\ \boldsymbol{w}_2^{(k)\top}\boldsymbol{\Sigma}_2\boldsymbol{w}_2^{(j)} = 0 \quad \text{for} \quad j < k.
\end{aligned}
$$

The pairs $(\boldsymbol{\theta}_1^{(k)}, \boldsymbol{\theta}_2^{(k)})$ are called canonical vectors, and the values $\rho_k = \boldsymbol{\theta}_1^{(k)\top} \boldsymbol{\Sigma}_{12} \boldsymbol{\theta}_2^{(k)}$ are canonical correlations. By definition, $1 \geq \rho_1 \geq \rho_2 \geq \cdots \geq \rho_r > 0$ and $\left\{ \boldsymbol{\theta}_d^{(k)} \right\}_{k=1}^{r}$ are orthonormal with respect to $\boldsymbol{\Sigma}_d$, $\boldsymbol{\theta}_d^{(i)\top} \boldsymbol{\Sigma}_d \boldsymbol{\theta}_d^{(j)} = \mathbb{1}_{\{i=j\}}$. Moreover, the population CCA problem can be equivalently formulated as the matrix decomposition problem of $\boldsymbol{\Sigma}_{12}$, that is the r pairs $(\boldsymbol{\theta}_1^{(k)}, \boldsymbol{\theta}_2^{(k)})$ solve the population CCA problem if and only if (Chen et al., 2013)

$$\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_1 \left(\sum_{k=1}^{r} \rho_k \boldsymbol{\theta}_1^{(k)} \boldsymbol{\theta}_2^{(k)\top} \right) \boldsymbol{\Sigma}_2. \tag{1.1}$$

1.3 Dissertation overview

The rest of the dissertation is organized as follows. In Chapter 2, we focus on the first challenge introduced in Section 1.1. We propose a joint framework for simultaneous classification and association analysis (JACA) of multimodal data by connecting linear discriminant analysis and canonical correlation analysis. A naive approach to conduct classification analysis for multi-view data is to apply the classical single-view methods to the concatenated matrix of views. Unfortunately, this method does not perform well in the high dimensional cases due to the over-fitting problem, especially when one view contains much stronger subtype-specific information. We demonstrate this idea by numerical studies on simulated data. We develop an efficient block-coordinate descent algorithm and use group-lasso type penalty to perform variable selection. We show the advantages of the proposed approach over existing methods by applying them to colorectal and breast cancer data from The Cancer Genome Atlas project.

Chapter 3 studies theoretical properties of the proposed methodology, JACA, in high dimensional settings. To our knowledge, this is the first consistency result for joint learning frameworks. By using an augmented approach and sub-exponential concentration bounds, we obtain the estimation error bound that is of the same order as the known bounds for group-lasso linear regression (Lounici et al., 2011; Nardi and Rinaldo, 2008). We also provide a semi-supervised extension of JACA to handle block-missing structure. The semi-supervised JACA can be applied to the settings with missing class labels, and the settings with missing subsets of views. We contrast two ver-

sions of JACA and show that both classification and estimation accuracy can be improved when the subjects with incomplete information are added to the analysis.

Chapter 4 presents an association analysis of cellular subtype proportions for various cancer types. In the literature, several deconvolution methods have been proposed (Newman et al., 2015; Li et al., 2017; Wang et al., 2018) to estimate the cellular subtype proportions, and they process the tumor samples based on different sources, such as RNAseq and microarray. Driven by the desire to find the connections between these methods and provide scientific discoveries for cancer studies, we focus on assessing the strength of association between cellular purity proportions for prostate, bladder and colorectal cancers. The results are then being incorporated to perform survival analysis, and therefore identify key signals that are predictive of cancer prognosis.

Chapter 5 describes a low-rank model to disentangle the common and individual signals of two views, and extend it to handle non-Gaussian data by utilizing the exponential family. Most methods in the literature assume there is no shared information between joint and individual signals, but fail to provide such guarantee for the individual signals. On the contrary, the proposed low-rank model enforces the orthogonality between the individual scores, thus guarantees no more shared information is embedded in the individual structures. The proposed optimization problem for our method is not convex, and we derive an alternating algorithm to estimate the model parameters. Although the overall global convergence is not guaranteed, the estimators by the proposed algorithm show consistent better performance than existing methods based on our experiments.

1.4 Notation

We consider n independent observations $(\boldsymbol{x}_{1i}, \ldots, \boldsymbol{x}_{Di}, y_i) \in \mathbb{R}^{p_1} \times \cdots \times \mathbb{R}^{p_D} \times \{1, \ldots, K\}$, where $\boldsymbol{x}_{di}$ is the ith sample's measurements from view d, and y_i is the corresponding class assignment. For two scalars $a, b \in \mathbb{R}$, we let $a \vee b = \max(a, b)$. For a vector $\boldsymbol{v} \in \mathbb{R}^p$, we let $\|\boldsymbol{v}\|_2 = (\sum_{j=1}^p v_j^2)^{1/2}$, $\|\boldsymbol{v}\|_1 = \sum_{j=1}^p |v_j|$ and $\|\boldsymbol{v}\|_\infty = \max_j |v_j|$. For matrices $\boldsymbol{M}, \boldsymbol{N} \in \mathbb{R}^{n \times p}$, we let $\|\boldsymbol{M}\|_F = (\sum_{i=1}^n \sum_{j=1}^p m_{ij}^2)^{1/2}$, $\|\boldsymbol{M}\|_{\infty,2} = \max_{1 \le i \le n} (\sum_{j=1}^p m_{ij}^2)^{1/2}$, $\|\boldsymbol{M}\|_{1,2} = \sum_{i=1}^n (\sum_{j=1}^p m_{ij}^2)^{1/2}$ and $\langle \boldsymbol{M}, \boldsymbol{N} \rangle = \mathrm{Tr}(\boldsymbol{M}^\top \boldsymbol{N})$. Define the nuclear norm of matrix $\boldsymbol{M}$ as $\|\boldsymbol{M}\|_* = \sum_{i=1}^{min(p,n)} \sigma_i(\boldsymbol{M})$, where $\sigma_i(\boldsymbol{M})$ is the ith singular value of $\boldsymbol{M}$. We use $\boldsymbol{I} = \boldsymbol{I}_p$

to denote $p \times p$ identity matrix, and $\mathbf{0}$ to denote zero matrix. For two sequences of scalars $a_1, \ldots, a_n, \ldots$ and $b_1, \ldots, b_n, \ldots$, we use $b_n = o(a_n)$ if $\lim_{n\to\infty}(b_n/a_n) = 0$ and $b_n = O(a_n)$ if $\lim_{n\to\infty}(b_n/a_n) < C$ for some finite constant C. For two sequences of random variables $x_1, \ldots, x_n, \ldots$ and $y_1, \ldots, y_n, \ldots$, we use $y_n = o_p(x_n)$ if for any $\varepsilon > 0$ $P(|y_n/x_n| < \varepsilon) \to 0$ as $n \to \infty$, and $y_n = O_p(x_n)$ if for any $\varepsilon > 0$ there exists M_ε such that $P(|y_n/x_n| > M_\varepsilon) < \varepsilon$ for all n. For subspaces A and B, denote the orthogonal complement of B in A as A/B.

2. JOINT ASSOCIATION AND CLASSIFICATION ANALYSIS

2.1 Introduction

This work is motivated by The Cancer Genome Atlas Project (Weinstein et al., 2013). TCGA project contains multiple views for the same set of subjects, such as gene expression, genotype, metabolic measurements, etc. At the same time, the subjects are separated into known classes. For example, the breast cancer patients are typically separated into Basal, HER2, Luminal A and Luminal B subtypes (The Cancer Genome Atlas Network, 2012). Since each view presents complementary information regarding the subject's biological system, it is of interest to answer two questions: (1) how to predict the cancer subtype given information from multiple views? (2) what are the associations between the views that are relevant for subtype prediction?

In the case of one view, the subtype prediction can be done using one of the many classification methods such as multinomial regression, multi-class support vector machines, discriminant analysis, etc. In case of multiple views, however, one has to either apply the chosen method separately to each view, or apply the method to the concatenated matrix of views. The separate approach may lead to inconsistent classification results across views. The concatenation approach ignores heterogeneity between the views in terms of scale and the number of measurements. Moreover, when one view has a much stronger subtype-specific signal, the concatenation may mask the less-dominant signals in other views. This is supported by our numerical results in Section 2.5.1.

To answer the second question, a line of research has focused on finding associations between the views based on canonical correlation analysis (Chen et al., 2013; Gao et al., 2017; Witten et al., 2009). These methods, however, do not use subtype information. Witten and Tibshirani (2009) propose supervised canonical correlation analysis, however the method is designed for the continuous response rather than the discrete class assignment, and only uses the response to filter relevant measurements. Another strategy based on factor models is proposed by Li and Jung (2017), who decompose each views into shared and individual structures that are informed by

covariates. Both Witten and Tibshirani (2009) and Li and Jung (2017) use subtype information indirectly, and are not tailored towards classification.

Recently, several methods have combined the task of finding associations between the views with the task of learning the regression coefficients. Gross and Tibshirani (2015) propose to combine canonical correlation analysis with linear regression. The method, however, is restricted to univariate continuous response and can only be applied to two views. Luo et al. (2016) propose to combine canonical correlation analysis objective with a general class of loss functions. Unlike Gross and Tibshirani (2015), the method could be applied to more than two views, and binary response. Nevertheless, the method is not suited for multi-group classification, has nonconvex optimization objective and requires rank pre-specification for model fitting. Finally, neither Gross and Tibshirani (2015) nor Luo et al. (2016) discuss the underlying population model, and the methods come with no theoretical guarantees.

In this work, we develop a framework for Joint Association and Classification Analysis (JACA) of multi-view data by connecting discriminant analysis with canonical correlation analysis. Since the method of Luo et al. (2016) also allows to perform joint association and classification in the two-class case, we further contrast two approaches. First, we use discriminant analysis rather than the regression framework, which allows us to fix the rank for model fitting to be $K - 1$, where K is the number of classes. In Luo et al. (2016), the rank of the model has to be chosen by the user. Secondly, we are able to formulate our method as a convex optimization problem by using the optimal scoring formulation of multi-class discriminant analysis (Hastie et al., 1994) and fixing the scores to be orthogonally invariant (Gaynanova, 2019). We add group-lasso type penalty to the optimization objective to allow for variable selection, and use block-coordinate descent algorithm to solve the corresponding convex problem. In contrast, the method of Luo et al. (2016) is nonconvex, and requires the use of variable splitting and augmented Lagrangian.

The rest of the chapter is organized as follows. Section 2.2 establishes the connection between canonical correlation analysis and linear discriminant analysis, and describes the proposed JACA method. Section 2.3 describes the method's implementation. Section 2.4 provides numerical com-

parisons with other methods on simulated data. Section 2.5 provides the analysis of colorectal cancer data from The Cancer Genome Atlas project. The technical proofs of the main results are provided in Section 2.7.

2.2 Proposed methodology

2.2.1 Connection between canonical correlation and linear discriminant analysis

In this section, We demonstrate that discriminant vectors in LDA coincide with the subset of canonical vectors in CCA, and use this connection to motivate the proposed method.

Consider LDA under Assumptions 1 and 2.

Assumption 1. $P(y = k) = \pi_k$ *for* $k = 1, \ldots, K$.

Assumption 2. $\boldsymbol{x}_d \in \mathbb{R}^{p_d}$, $d \in \{1, \ldots, D\}$ *are mean zero random vectors and have class-conditional means and covariance matrices as*

$$\mathbb{E}\left[\begin{pmatrix}\boldsymbol{x}_1\\ \vdots \\ \boldsymbol{x}_D\end{pmatrix}\Big| y = k\right] = \begin{pmatrix}\boldsymbol{\mu}_{1k}\\ \vdots \\ \boldsymbol{\mu}_{Dk}\end{pmatrix}, \quad \mathrm{Cov}\left[\begin{pmatrix}\boldsymbol{x}_1\\ \vdots \\ \boldsymbol{x}_D\end{pmatrix}\Big| y = k\right] = \boldsymbol{\Sigma}_y = \begin{pmatrix}\boldsymbol{\Sigma}_{1y} & \ldots & \boldsymbol{\Sigma}_{1Dy}\\ & \vdots & \\ \boldsymbol{\Sigma}_{1Dy}^\top & \ldots & \boldsymbol{\Sigma}_{Dy}\end{pmatrix}. \tag{2.1}$$

In Assumption 1, the random variable y indicates the class assignment. In Assumption 2, we do not specify the distribution of $\boldsymbol{x}_d$, but only assume the existence of the first two moments. As in LDA, we assume that the covariance matrices are equal between the groups (we keep subscript y to differentiate the class-conditional covariance matrix $\boldsymbol{\Sigma}_y$ from the marginal covariance matrix $\boldsymbol{\Sigma}$). We next show that under additional assumptions on $\boldsymbol{\Sigma}_y$, the class-conditional specification (2.1) is equivalent to the factor model.

Proposition 1. *Let* y *be a random variable satisfying Assumption 1, and let* $\boldsymbol{x}_d \in \mathbb{R}^{p_d}$ *be random vectors satisfying Assumption 2. Further, let* $\boldsymbol{\Sigma}_{ldy} = \boldsymbol{0}$ *for all* $l \neq d \in \{1, \ldots, D\}$. *Then each* $\boldsymbol{x}_d$ *can be equivalently specified via the factor model:*

$$\boldsymbol{x}_d = \boldsymbol{\mu}_d + \boldsymbol{\Delta}_d \boldsymbol{u}_y + \boldsymbol{\Sigma}_{dy}^{1/2} \boldsymbol{e}_d, \tag{2.2}$$

where $\boldsymbol{\mu}_d = \mathbf{0}$ *is the overall mean;* $\boldsymbol{u}_y = f(y) \in \mathbb{R}^{K-1}$ *is a random vector indicating class assignment with* $\mathbb{E}(\boldsymbol{u}_y) = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{u}_y) = \boldsymbol{I}$; $\boldsymbol{\Delta}_d \in \mathbb{R}^{p_d \times (K-1)}$ *is such that* $\boldsymbol{\Delta}_d^\top \boldsymbol{\Sigma}_{dy}^{-1} \boldsymbol{\Delta}_d = \boldsymbol{\Lambda}_d$ *is diagonal with* $\mathbb{E}(\boldsymbol{\mu}_d + \boldsymbol{\Delta}_d \boldsymbol{u}_y | y = k) = \boldsymbol{\mu}_{dk}$; $\boldsymbol{\Sigma}_{dy}$ *is class-conditional covariance matrix for view* d, *and* $\boldsymbol{e}_d \in \mathbb{R}^{p_d}$ *are isotropic noise vectors independent from* y.

Remark 1. $\boldsymbol{u}_y$ *represents a transformed class indicator vector. Combined with* $\boldsymbol{\Delta}_d$, *it reflects the difference between the conditional mean and the overall mean. When* $K = 2$, $\boldsymbol{u}_y = f(y) = \sqrt{\pi_2/\pi_1}\,\mathbb{1}\{y = 1\} - \sqrt{\pi_1/\pi_2}\,\mathbb{1}\{y = 2\}$, *case* $K > 2$ *is in Section 2.7 of the Supplementary Materials.*

Remark 2. *If* $\mathrm{rank}(\boldsymbol{\Delta}_d^\top \boldsymbol{\Sigma}_{dy}^{-1} \boldsymbol{\Delta}_d) = r < K-1$, *then* (2.2) *is not identifiable as the effective number of class-specific factors* r *is less than* $K - 1$. *For clarity, we assume throughout that* $r = K - 1$, *but the results can be generalized at the expense of a more technical proof. When* $K = 2$, *the restriction is equivalent to requiring the class-conditional means to be distinct.*

The factor model (2.2) is directly connected with discriminant vectors in LDA. When $K > 2$, Gaynanova et al. (2016) show that the matrix of discriminant vectors can be expressed as $\boldsymbol{W}_d \propto \boldsymbol{\Sigma}_d^{-1} \boldsymbol{\Delta}_d$, where $\propto$ is applied columnwise. Hence, by fixing the magnitude of discriminant vectors in accordance with (2.2), we can rewrite the factor model as

$$\boldsymbol{x}_d = \boldsymbol{\mu}_d + \boldsymbol{\Sigma}_{dy} \boldsymbol{W}_d \boldsymbol{u}_y + \boldsymbol{\Sigma}_{dy}^{1/2} \boldsymbol{e}_d,$$

where $\boldsymbol{W}_d^\top \boldsymbol{\Sigma}_{dy} \boldsymbol{W}_d$ is a diagonal matrix. This representation allows to treat the matrix of discriminant vectors $\boldsymbol{W}_d$ as a covariance-adjusted matrix of loadings in the above factor model. The main limitation of Proposition 1, however, is the requirement $\boldsymbol{\Sigma}_{ldy} = \mathbf{0}$, that is the assumption that $\boldsymbol{u}_y$ are the only common factors between the views.

To consider a more general case with $\boldsymbol{\Sigma}_{ldy} \neq \mathbf{0}$, we adjust the factor model (2.2) as

$$\boldsymbol{x}_d = \boldsymbol{\mu}_d + \boldsymbol{\Delta}_d \boldsymbol{u}_y + \boldsymbol{A}_d \boldsymbol{u} + \widetilde{\boldsymbol{\Sigma}}_d^{1/2} \boldsymbol{e}_d, \tag{2.3}$$

where $\boldsymbol{\mu}_d$, $\boldsymbol{\Delta}_d$, $\boldsymbol{u}_y$ are as in Proposition 1, $\boldsymbol{u} \in \mathbb{R}^q$ represents q extra common factors between the D views, and $\boldsymbol{e}_d \in \mathbb{R}^{p_d}$ is an independent noise vector with $\mathbb{E}(\boldsymbol{e}_d) = \boldsymbol{0}$, $\mathrm{Cov}(\boldsymbol{e}_d) = \boldsymbol{I}$. Here $\widetilde{\boldsymbol{\Sigma}}_d$ is no longer class-conditional covariance matrix, but rather covariance matrix after accounting for both class membership ($\boldsymbol{u}_y$) and other shared factors ($\boldsymbol{u}$). When $\boldsymbol{A}_d = \boldsymbol{0}$, the model reduces to (2.2). We assume $\boldsymbol{A}_d$ is full rank given q (with $\boldsymbol{A}_d = \boldsymbol{0}$ for $q = 0$). As with model (2.2), we can rewrite (2.3) in terms of view-specific discriminant vectors as

$$\boldsymbol{x}_d = \boldsymbol{\mu}_d + \widetilde{\boldsymbol{\Sigma}}_d \boldsymbol{W}_d \boldsymbol{u}_y + \boldsymbol{A}_d \boldsymbol{u} + \widetilde{\boldsymbol{\Sigma}}_d^{1/2} \boldsymbol{e}_d.$$

We next connect the LDA-based factor model (2.3) with the CCA decomposition (1.1).

Theorem 1. *Consider the factor model* (2.3)*, where $\boldsymbol{\mu}_d$, $\boldsymbol{u}_y$ are as in Proposition 1; $\boldsymbol{u} \in \mathbb{R}^q$ is a random vector independent of y with $\mathbb{E}(\boldsymbol{u}) = \boldsymbol{0}$, $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{I}$; and the loadings matrix $\boldsymbol{V}_d = [\widetilde{\boldsymbol{\Sigma}}_d^{-1/2} \boldsymbol{\Delta}_d \ \widetilde{\boldsymbol{\Sigma}}_d^{-1/2} \boldsymbol{A}_d] \in \mathbb{R}^{p_d \times (K-1+q)}$ is orthogonal following standard identifiability conditions for factor models (Mardia et al., 1979, Chapter 9.2). Let $\boldsymbol{\Sigma}_{ld} = \mathbb{E}(\boldsymbol{x}_l \boldsymbol{x}_d^\top)$ be the corresponding marginal cross-covariance matrix between mean zero $\boldsymbol{x}_l$ and $\boldsymbol{x}_d$.*

1. If $q = 0$, (2.3) *reduces to* (2.2) *and*

$$\boldsymbol{\Sigma}_{ld} = \boldsymbol{\Sigma}_l \left(\sum_{k=1}^{K-1} \rho_k \boldsymbol{\theta}_l^{(k)} \boldsymbol{\theta}_d^{(k)\top} \right) \boldsymbol{\Sigma}_d,$$

where $\boldsymbol{\Theta}_d = [\boldsymbol{\theta}_d^{(1)} \dots \boldsymbol{\theta}_d^{(K-1)}] \propto \boldsymbol{\Sigma}_d^{-1} \boldsymbol{\Delta}_d$ is orthonormal with respect to $\boldsymbol{\Sigma}_d$, and ρ_k are diagonal elements of matrix $(\boldsymbol{I} + \boldsymbol{\Lambda}_l)^{-1/2} \boldsymbol{\Lambda}_l^{1/2} \boldsymbol{\Lambda}_d^{1/2} (\boldsymbol{I} + \boldsymbol{\Lambda}_d)^{-1/2}$.

2. If $q > 0$,

$$\boldsymbol{\Sigma}_{ld} = \boldsymbol{\Sigma}_l \left(\sum_{k=1}^{q+K-1} \rho_k \boldsymbol{\theta}_l^{(k)} \boldsymbol{\theta}_d^{(k)\top} \right) \boldsymbol{\Sigma}_d,$$

where $\left\{ \boldsymbol{\theta}_d^{(k)} \right\}_{k=1}^{q+K-1}$ are orthonormal with respect to $\boldsymbol{\Sigma}_d$, $\boldsymbol{\Sigma}_l \left(\sum_{k=1}^{q} \rho_k \boldsymbol{\theta}_l^{(k)} \boldsymbol{\theta}_d^{(k)\top} \right) \boldsymbol{\Sigma}_d = \boldsymbol{A}_l \boldsymbol{A}_d^\top$ and $\boldsymbol{\Sigma}_l \left(\sum_{k=q+1}^{q+K-1} \rho_k \boldsymbol{\theta}_l^{(k)} \boldsymbol{\theta}_d^{(k)\top} \right) \boldsymbol{\Sigma}_d = \boldsymbol{\Delta}_l \boldsymbol{\Delta}_d^\top$ are as in part 1.

If $q = 0$, then the only relationship between the views is due to shared class membership ($\boldsymbol{u}_y$).

In this case, the canonical vectors $\boldsymbol{\Theta}_d$ in CCA coincide with discriminant vectors $\boldsymbol{W}_d$ in LDA. If $q > 0$, then there exists extra q factors that are shared between the views, leading to q extra pairs of canonical vectors in CCA. If the LDA directions correspond to the maximal ρ_k, then the first $K-1$ canonical pairs coincide with discriminant vectors. If the LDA directions do not correspond to the maximal ρ_k, then the first $K-1$ canonical pairs include other shared factors that are independent of class membership.

2.2.2 Joint association and classification analysis

Our goal is to estimate view-specific matrices of canonical vectors that correspond to discriminant directions in LDA, that is to estimate $\boldsymbol{W}_d \propto \boldsymbol{\Sigma}_d^{-1}\boldsymbol{\Delta}_d$. On the one hand, we want to perform well in classification. On the other hand, we want to maximize the correlation between the views. In light of correspondence between CCA and LDA explored in Theorem 1, our proposal is based on combining the strengths of both approaches. When the leading canonical correlations are due to shared class memberships, the leading canonical vectors $\boldsymbol{\Theta}_d$ in CCA coincide with discriminant vectors $\boldsymbol{W}_d$ in LDA. We want to improve the estimation accuracy and efficiency of LDA in this case by analyzing multiple views jointly. In other cases, we want the proposed model to not be fooled by leading canonical correlations that are independent of shared class memberships.

For the classification, we reformulate sparse multi-group discriminant analysis (Gaynanova et al., 2016) as penalized optimal scoring problem (Hastie et al., 1994).

Proposition 2. *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be the column-centered data matrix, $\boldsymbol{Z} \in \mathbb{R}^{n\times K}$ be the corresponding class-indicator matrix, n_k be the number of samples in class k and $s_k = \sum_{i=1}^{k} n_i$. Let $\boldsymbol{H} \in \mathbb{R}^{K\times K-1}$ have columns $\boldsymbol{H}_l \in \mathbb{R}^K$ defined as*

$$\boldsymbol{H}_l = \left(\left\{(nn_{l+1})^{1/2}(s_l s_{l+1})^{-1/2}\right\}_l, \quad -(ns_l)^{1/2}(n_{l+1}s_{l+1})^{-1/2}, \quad \mathbf{0}_{K-1-l}\right)^\top,$$

and let $\widetilde{\boldsymbol{Y}} = \boldsymbol{Z}\boldsymbol{H}$. Then the discriminant vectors in multi-group sparse discriminant analysis

(Gaynanova et al., 2016) correspond to the solution of

$$\underset{\boldsymbol{V}\in\mathbb{R}^{p\times(K-1)}}{\text{minimize}}\Big\{\frac{1}{2n}\|\widetilde{\boldsymbol{Y}}-\boldsymbol{X}\boldsymbol{V}\|_F^2+\lambda\sum_{i=1}^{p}\|\boldsymbol{v}_i\|_2\Big\}. \tag{2.4}$$

Hence, the problem of finding sparse discriminant directions in the multi-group case can be recast as the multi-response penalized least-squares linear regression problem.

For the correlation between the views, we rewrite the sample CCA criterion for column-centered views $\boldsymbol{X}_d$ and $\boldsymbol{X}_l$ as minimization of the least squares objective subject to orthogonality constraints

$$\underset{\boldsymbol{W}_d,\boldsymbol{W}_l}{\text{minimize}}\,\|\boldsymbol{X}_d\boldsymbol{W}_d-\boldsymbol{X}_l\boldsymbol{W}_l\|_F^2 \quad\text{subject to}\quad \frac{1}{n}\boldsymbol{W}_d^\top\boldsymbol{X}_d^\top\boldsymbol{X}_d\boldsymbol{W}_d=\boldsymbol{I},\quad \frac{1}{n}\boldsymbol{W}_l^\top\boldsymbol{X}_l^\top\boldsymbol{X}_l\boldsymbol{W}_l=\boldsymbol{I}. \tag{2.5}$$

We propose to find the matrices of discriminant vectors $\boldsymbol{W}_d\in\mathbb{R}^{p_d\times(K-1)}$ by combining classification objective (2.4) with canonical correlation objective (2.5):

$$\begin{aligned}\underset{\boldsymbol{W}_1,\ldots,\boldsymbol{W}_D}{\text{minimize}}\Big\{&\frac{\alpha}{2nD}\sum_{d=1}^{D}\|\widetilde{\boldsymbol{Y}}-\boldsymbol{X}_d\boldsymbol{W}_d\|_F^2+\frac{1-\alpha}{2nD(D-1)}\sum_{d=1}^{D-1}\sum_{l=d+1}^{D}\|\boldsymbol{X}_d\boldsymbol{W}_d-\boldsymbol{X}_l\boldsymbol{W}_l\|_F^2\\&+\sum_{d=1}^{D}\lambda_d\,\mathrm{Pen}(\boldsymbol{W}_d)\Big\}\quad\text{subject to}\quad\frac{1}{n}\boldsymbol{W}_d^\top\boldsymbol{X}_d^\top\boldsymbol{X}_d\boldsymbol{W}_d=\boldsymbol{I},\quad\text{for } 1\le d\le D.\end{aligned} \tag{2.6}$$

Here $\mathrm{Pen}(\boldsymbol{W}_d)$ can be used to put structural assumptions on $\boldsymbol{W}_d$ such as sparsity, and $\alpha\in[0,1]$ controls the relative weights between LDA and CCA criteria. When $\alpha=0$, (2.6) reduces to sparse CCA. When $\alpha=1$, (2.6) reduces to sparse LDA with additional orthogonality constraints. While the orthogonality constraints are required for CCA criterion (2.5) to avoid trivial zero solution, they are not necessary in (2.6) as long as $\alpha>0$ due to the addition of the optimal scoring loss function. Moreover, the classification rule in discriminant analysis is invariant to both scaling and orthogonal transformation of the matrix of discriminant vectors (Gaynanova et al., 2016). To make the problem convex and simplify computations, we only consider $\alpha>0$, and drop the orthogonality constraints

in (2.6) leading to

$$\underset{\boldsymbol{W}_1,\ldots,\boldsymbol{W}_D}{\text{minimize}} \Big\{ \frac{\alpha}{2nD} \sum_{d=1}^{D} \|\widetilde{\boldsymbol{Y}} - \boldsymbol{X}_d \boldsymbol{W}_d\|_F^2 + \frac{1-\alpha}{2nD(D-1)} \sum_{d=1}^{D-1} \sum_{l=d+1}^{D} \|\boldsymbol{X}_d \boldsymbol{W}_d - \boldsymbol{X}_l \boldsymbol{W}_l\|_F^2 + \sum_{d=1}^{D} \lambda_d \operatorname{Pen}(\boldsymbol{W}_d) \Big\}. \tag{2.7}$$

We call (2.7) JACA for Joint Association and Classification Analysis, and choose convex $\operatorname{Pen}(\boldsymbol{W}_d) = \sum_{i=1}^{p_d} \|\boldsymbol{w}_{di}\|_2$ to encourage variable selection via row-wise sparsity in $\boldsymbol{W}_d$. We do not consider ℓ_1 penalty since it induces element-wise rather than row-wise sparsity in $\boldsymbol{W}_d$, hence it does not completely eliminate the variables from the model and the sparsity pattern is not preserved under orthogonal transformations. We chose a convex penalty for computational reasons, we refer to Huang et al. (2012) for other row-wise sparse penalties that are nonconvex.

Further, problem (2.7) can be rewritten as a multi-response linear regression problem using the augmented data approach. We first illustrate the case $D = 2$. Let $\boldsymbol{W} = (\boldsymbol{W}_1^\top, \boldsymbol{W}_2^\top)^\top$,

$$\boldsymbol{Y}' = \frac{\sqrt{\alpha}}{\sqrt{nD}} \begin{pmatrix} \widetilde{\boldsymbol{Y}} \\ \widetilde{\boldsymbol{Y}} \\ \boldsymbol{0} \end{pmatrix}, \quad \boldsymbol{X}' = \frac{1}{\sqrt{nD}} \begin{pmatrix} \sqrt{\alpha}\boldsymbol{X}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \sqrt{\alpha}\boldsymbol{X}_2 \\ \sqrt{(1-\alpha)/(D-1)}\boldsymbol{X}_1 & -\sqrt{(1-\alpha)/(D-1)}\boldsymbol{X}_2 \end{pmatrix}.$$

Then (2.7) is equivalent to

$$\underset{\boldsymbol{W}}{\text{minimize}} \Big\{ 2^{-1} \|\boldsymbol{Y}' - \boldsymbol{X}'\boldsymbol{W}\|_F^2 + \sum_{d=1}^{D} \lambda_d \operatorname{Pen}(\boldsymbol{W}_d) \Big\}. \tag{2.8}$$

When $D > 2$, let $\boldsymbol{W} = (\boldsymbol{W}_1^\top, \ldots, \boldsymbol{W}_D^\top)^\top$,

$$\boldsymbol{Y}' = \sqrt{\frac{\alpha}{nD}} \Big(\underbrace{\widetilde{\boldsymbol{Y}}^\top \quad \ldots \quad \widetilde{\boldsymbol{Y}}^\top}_{D} \quad \boldsymbol{0} \quad \ldots \quad \boldsymbol{0} \Big)^\top,$$

$$\boldsymbol{X}' = \begin{pmatrix} \sqrt{\alpha}\boldsymbol{X}_1 & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \sqrt{\alpha}\boldsymbol{X}_2 & \mathbf{0} & \dots & \mathbf{0} \\ & & \vdots & & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \sqrt{\alpha}\boldsymbol{X}_D \\ \sqrt{\frac{1-\alpha}{D-1}}\boldsymbol{X}_1 & -\sqrt{\frac{1-\alpha}{D-1}}\boldsymbol{X}_2 & \mathbf{0} & \dots & \mathbf{0} \\ \sqrt{\frac{1-\alpha}{D-1}}\boldsymbol{X}_1 & \mathbf{0} & -\sqrt{\frac{1-\alpha}{D-1}}\boldsymbol{X}_3 & \dots & \mathbf{0} \\ \sqrt{\frac{1-\alpha}{D-1}}\boldsymbol{X}_1 & \mathbf{0} & \mathbf{0} & \dots & -\sqrt{\frac{1-\alpha}{D-1}}\boldsymbol{X}_D \\ & & \vdots & & \\ \mathbf{0} & \mathbf{0} & \dots & \sqrt{\frac{1-\alpha}{D-1}}\boldsymbol{X}_{D-1} & -\sqrt{\frac{1-\alpha}{D-1}}\boldsymbol{X}_D \end{pmatrix} / \sqrt{nD}.$$

Then (2.7) is equivalent to

$$\underset{W}{\text{minimize}} \left\{ 2^{-1} \|\boldsymbol{Y}' - \boldsymbol{X}'\boldsymbol{W}\|_F^2 + \sum_{d=1}^{D} \lambda_d \operatorname{Pen}(\boldsymbol{W}_d) \right\}.$$

2.3 Implementation

2.3.1 Additional regularization via elastic net

It is well known that the lasso-type penalties can lead to erratic solution paths in the presence of highly-correlated variables (Hastie et al., 2015, Chapter 4.2). To overcome this drawback, Zou and Hastie (2005) propose an elastic net penalty which combines ridge and lasso penalties, thus making highly correlated variables either being jointly selected or not selected in the model. Zou and Hastie (2005) also advocate an extra scaling step which in regression context is equivalent to replacing the sample covariance matrix $\boldsymbol{X}^\top\boldsymbol{X}/n$ with the regularized version $(1-\rho)\boldsymbol{X}^\top\boldsymbol{X}/n+\rho\boldsymbol{I}$ for $\rho \in [0, 1]$. We adapt this idea to JACA, and replace $\boldsymbol{X}'^\top\boldsymbol{X}'$ in (2.8) with $(1-\rho)\boldsymbol{X}'^\top\boldsymbol{X}' + \rho\boldsymbol{I}$

for $\rho \in [0, 1]$ leading to

$$\underset{\boldsymbol{W}}{\text{minimize}} \left\{ \frac{1}{2}\|\boldsymbol{Y}' - \boldsymbol{X}'\boldsymbol{W}\|_F^2 - \frac{\rho}{2}\|\boldsymbol{X}'\boldsymbol{W}\|_F^2 + \frac{\rho}{2}\|\boldsymbol{W}\|_F^2 + \sum_{d=1}^{D} \lambda_d \operatorname{Pen}(\boldsymbol{W}_d) \right\}. \tag{2.9}$$

Problem (2.9) is still convex. When $\rho = 0$, problems (2.9) and (2.8) coincide.

2.3.2 Optimization algorithm

We use a block-coordinate descent algorithm to solve (2.9) for fixed values of $\rho \in [0, 1]$ and $\lambda_d \geq 0$. Compared with the original problem, our augmented formulation (2.9) leads to easier implementation, less iterations till convergence, and overall faster algorithm. Let $\boldsymbol{w}_{dj}$ be the jth row of $\boldsymbol{W}_d$, and let $\operatorname{Pen}(\boldsymbol{W}_d) = \sum_{j=1}^{p_d} \|\boldsymbol{w}_{dj}\|_2$. Since (2.9) is convex, and the penalty is separable with respect to each $\boldsymbol{w}_{dj}$, the algorithm is guaranteed to converge to the global optimum from any starting point (Tseng, 2001).

We assume that each $\boldsymbol{X}_d$ is standardized so that the diagonal entries of $n^{-1}\boldsymbol{X}_d^\top \boldsymbol{X}_d$ are equal to one. This standardization is common in the literature (Zou and Hastie, 2005; Witten and Tibshirani, 2009), and effectively results in penalizing each variable proportionally to its standard deviation. Moreover, using $\rho > 0$ with this standardization in (2.9) ensures the uniqueness of solution for any λ_d due to strict convexity of the objective function.

Consider solving (2.9) with respect to a row-vector $\boldsymbol{w}_{dj}$, and let $\boldsymbol{X}'_{dj}$ be the corresponding column of $\boldsymbol{X}'$. The KKT conditions (Boyd and Vandenberghe, 2004) can be written as a set of $\sum_{d=1}^{D} p_d$ equations of the form

$$(1-\rho)\boldsymbol{X}_{dj}'^\top \boldsymbol{X}'\boldsymbol{W} + \rho\boldsymbol{w}_{dj} - \boldsymbol{X}_{dj}'^\top \boldsymbol{Y}' + \lambda_d \boldsymbol{u}_{dj} = 0, \tag{2.10}$$

where $\boldsymbol{u}_{dj}$ is the subgradient of $\|\boldsymbol{w}_{dj}\|_2$, that is $\boldsymbol{u}_{dj} = \boldsymbol{w}_{dj}/\|\boldsymbol{w}_{dj}\|_2$ when $\|\boldsymbol{w}_{dj}\|_2 \neq 0$ and $\boldsymbol{u}_{dj} \in \{\boldsymbol{u} : \|\boldsymbol{u}\|_2 \leq 1\}$ otherwise. Solving (2.10) with respect to $\boldsymbol{w}_{dj}$ leads to

$$\boldsymbol{w}_{dj} = \left\{ \boldsymbol{X}_{dj}'^\top (\boldsymbol{Y}' - (1-\rho)\boldsymbol{X}'\boldsymbol{W} + (1-\rho)\boldsymbol{X}'_{dj}\boldsymbol{w}_{dj}) - \lambda_d \boldsymbol{u}_{dj} \right\} \Big/ \left\{ (1-\rho)\,\|\boldsymbol{X}'_{dj}\|_2^2 + \rho \right\}.$$

Algorithm 1 Block-coordinate descent algorithm for (2.9)

Given: $k = 0$, $\boldsymbol{W}^{(0)}$, $\varepsilon > 0$;
$\boldsymbol{R} \leftarrow \boldsymbol{Y}' - (1-\rho)\boldsymbol{X}'\boldsymbol{W}^{(0)}$;
while $k \neq k_{max}$ *and* $\left|\textit{objective}\left(\boldsymbol{W}^{(k)}\right) - \textit{objective}\left(\boldsymbol{W}^{(k-1)}\right)\right| \geq \varepsilon$ **do**
  $k \leftarrow k + 1$;
  for $d = 1$ **to** D **do**
    for $j = 1$ **to** p_d **do**
      $\boldsymbol{w}_{dj}^{(k)} \leftarrow S_{\lambda_d}(\boldsymbol{X}_{dj}'^{\top}\boldsymbol{R} + (1-\rho)\|\boldsymbol{X}'_{dj}\|_2^2\boldsymbol{w}_{dj}^{(k-1)})/\{(1-\rho)\|\boldsymbol{X}'_{dj}\|_2^2 + \rho\}$;
      $\boldsymbol{R} \leftarrow \boldsymbol{R} + (1-\rho)\boldsymbol{X}'_{dj}(\boldsymbol{w}_{dj}^{(k-1)} - \boldsymbol{w}_{dj}^{(k)})$
    end
  end
end

For a vector $\boldsymbol{v} \in \mathbb{R}^m$ and $\lambda > 0$, let $S_\lambda(\boldsymbol{v}) = \max(0, 1 - \lambda/\|\boldsymbol{v}\|_2)\boldsymbol{v}$ be the vector soft-thresholding operator. Then iterating block updates leads to Algorithm 1.

2.3.3 Selection of tuning parameters

JACA requires the specification of several parameters: $\alpha \in (0, 1]$ that controls the relative weights of LDA and CCA criteria, $\rho \in [0, 1]$ that controls the shrinkage induced by elastic net, and $\lambda_d \geq 0$ that control the sparsity level of each $\boldsymbol{W}_d$ respectively. While it is possible to perform cross-validation over all of the parameters, due to computational considerations we restrict the space as follows. First, we set $\alpha = 0.5$ in all of our simulations studies and data analyses. The results were similar for $\alpha = 0.8$, and slightly worse for $\alpha = 0.2$. Further work is required to determine whether there is an optimal choice. Secondly, we set $\lambda_d = \epsilon\lambda_{\max,d}$ with $\epsilon \in (0, 1)$, where $\lambda_{\max,d}$ is such that $\widehat{\boldsymbol{W}}_d = \boldsymbol{0}$ for any $\lambda \geq \lambda_{\max,d}$, similar strategy is used in Luo et al. (2016) as it allows to control the sparsity of each $\boldsymbol{W}_d$ at similar levels. The value of $\lambda_{\max,d}$ is given below.

Proposition 3. *Let* $\lambda_{\max,d} = \frac{\alpha}{nD}\|\boldsymbol{X}_d^{\top}\widetilde{\boldsymbol{Y}}\|_{\infty,2}$. *Then* $\boldsymbol{W}_d = 0$ *for all* $\lambda \geq \lambda_{\max,d}$.

We use cross-validation with F folds to select $\rho \in [0, 1]$ and $\epsilon \in [10^{-4}, 1]$, with a course grid for ρ and a fine grid for ϵ.

It is typical to minimize the prediction error in cross-validation, for example the least squares error in the linear regression. In our context, however, both classification rules and correlation

measures are invariant to the scale of $\boldsymbol{W}_d$, hence we need a scale-invariant metric. We propose to consider

$$CV(\rho,\varepsilon)=\frac{1}{F}\sum_{f=1}^{F}\Bigg\{\alpha\sum_{d=1}^{D}|\mathrm{Cor}(\widetilde{\boldsymbol{Y}}^{(f)},\boldsymbol{X}_d^{(f)}\widehat{\boldsymbol{W}}_d^{(-f)})| + \frac{(1-\alpha)}{D-1}\sum_{d=1}^{D-1}\sum_{l=d+1}^{D}|\mathrm{Cor}(\boldsymbol{X}_d^{(f)}\boldsymbol{W}_d^{(-f)},\boldsymbol{X}_l^{(f)}\boldsymbol{W}_l^{(-f)})|\Bigg\}, \tag{2.11}$$

where $\widetilde{\boldsymbol{Y}}^{(f)}$, $\boldsymbol{X}_d^{(f)}$ correspond to the samples in the fth fold; and $\widehat{\boldsymbol{W}}_d^{(-f)}$ are solutions to (2.9) with given ρ and ε based on samples in all folds except the fth. We define the correlation between two centered matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ as the square root of the RV-coefficient (Robert and Escoufier, 1976), where

$$\mathrm{RV}(\boldsymbol{X},\boldsymbol{Y}):=\frac{\mathrm{Tr}(\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{Y}^\top)}{\sqrt{\mathrm{Tr}(\boldsymbol{X}\boldsymbol{X}^\top)^2}\sqrt{\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Y}^\top)^2}}.$$

By definition, $\sqrt{\mathrm{RV}(\boldsymbol{X},\boldsymbol{Y})}\in[0,1]$, and is invariant to scale and orthogonal transformation. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are vectors, then $\sqrt{\mathrm{RV}(\boldsymbol{X},\boldsymbol{Y})}=|\mathrm{Cor}(\boldsymbol{X},\boldsymbol{Y})|$.

2.4 Simulation studies

We compare the performance of the following methods: (i) JACA: Joint Association and Classification Analysis, the proposed approach; (ii) Sparse Linear Discriminant Analysis of Gaynanova et al. (2016) as implemented in the R package MGSDA (Gaynanova, 2016), either applied separately to each dataset (SLDA_sep), or jointly on concatenated dataset (SLDA_joint); (iii) Sparse CCA: Sparse Canonical Correlation Analysis of Witten and Tibshirani (2009) as implemented in the R package PMA (Witten et al., 2013). We use cross-validation to choose the tuning parameters instead of the permutation method introduced in Witten and Tibshirani (2009), since the former one gives better results. (iv) Sparse sCCA: Sparse supervised CCA proposed in Witten and Tibshirani (2009). We first choose a set of variables with largest values of F-statistic from a one-way ANOVA, and then apply Sparse CCA with selected variables; (v) CVR: Canonical Variate Regression by Luo et al. (2016) as implemented in the corresponding R package (Luo and

Chen, 2017).

2.4.1 Data generation

We generate the data using factor model (2.3). Specifically, given $\widetilde{\boldsymbol{\Sigma}}_d$, $d = 1, \ldots, D$, we generate the factor loadings in (2.3) as follows

1. Generate row-sparse matrix $\boldsymbol{B}_d \in \mathbb{R}^{p_d \times K-1}$ with $s = 10$ non-zero rows. Draw nonzero elements from uniform distribution on $[-2, -1] \cup [1, 2]$. Given $c_d > 0$, rotate and scale $\boldsymbol{B}_d$ so that $\boldsymbol{B}_d^\top \widetilde{\boldsymbol{\Sigma}}_d \boldsymbol{B}_d = \text{diag}(c_d^2)$, and set $\boldsymbol{\Delta}_d = \widetilde{\boldsymbol{\Sigma}}_d \boldsymbol{B}_d$. According to Theorem 1, this sets $K-1$ canonical correlations ρ_k between datasets d and l to be equal to

$$\rho_k = (c_d c_l)/\sqrt{(1 + c_d^2)(1 + c_l^2)}.$$

2. If $q \neq 0$, generate $\boldsymbol{M}_d \in \mathbb{R}^{p_d \times q}$ with elements from $N(0, 1)$, orthogonalize $\boldsymbol{M}_d$ with respect to $\boldsymbol{\Delta}_d$ as $\boldsymbol{M}_d = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{\Delta}_d})\boldsymbol{M}_d$, where $\boldsymbol{P}_{\boldsymbol{\Delta}_d}$ is the projection matrix onto column space of $\boldsymbol{\Delta}_d$. For canonical correlation $\rho_k \in (0, 1)$, set $c_k = \sqrt{\rho_k/(1 - \rho_k)}$, and rotate and scale $\boldsymbol{M}_d$ so that $\boldsymbol{M}_d^\top \widetilde{\boldsymbol{\Sigma}}_d \boldsymbol{M}_d = \text{diag}(c_k^2)$. Set $\boldsymbol{A}_d = \widetilde{\boldsymbol{\Sigma}}_d \boldsymbol{M}_d$.

We further draw n independent y with $P(y = k) = \pi_k$, n independent $\boldsymbol{u}_q$ from $N(0, \boldsymbol{I}_q)$, and n independent $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d$, each from $N(0, \boldsymbol{I}_{p_d})$. We get n replicas $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d$ according to (2.3) with given $\boldsymbol{\Delta}_d$, $\boldsymbol{A}_d$ and $\boldsymbol{\mu}_d = 0$, $d = 1, \ldots, D$. By construction, the population discriminant vectors are proportional to $\boldsymbol{B}_d$ with corresponding row-sparsity pattern.

2.4.2 Evaluation criteria

We compare the methods in terms of misclassification rate and strength of association between the views. Additional comparisons in terms of the estimation consistency and variable selection results are provided in the Supplementary Material. To compare the classification accuracy, we consider two prediction approaches for each method: prediction based on one view alone out of $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d)$ using the corresponding subset of canonical vectors, and prediction based on the full concatenated dataset. All predictions are made by linear discriminant analysis model. The

misclassification rate of each classifier is calculated as

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left\{\text{label}(\boldsymbol{x}_i)\neq\text{pred}(\boldsymbol{x}_i)\right\},$$

where $\boldsymbol{x}_i s$ are m new samples, label$(\boldsymbol{x}_i)$ denotes the corresponding class membership and pred$(\boldsymbol{x}_i)$ denotes the predicted class membership.

To evaluate the strength of found association between the views, we consider

$$\text{Sum correlation}(\boldsymbol{W}_1,\ldots,\boldsymbol{W}_d)=\sum_{d=1}^{D-1}\sum_{l=d+1}^{D}\text{Cor}_{\boldsymbol{\Sigma}}(\boldsymbol{W}_d,\boldsymbol{W}_l),$$

where

$$\text{Cor}_{\boldsymbol{\Sigma}}(\boldsymbol{W}_d,\boldsymbol{W}_l)=\left(\frac{\text{Tr}(\boldsymbol{W}_d^\top\boldsymbol{\Sigma}_{dl}\boldsymbol{W}_l\boldsymbol{W}_l^\top\boldsymbol{\Sigma}_{dl}\boldsymbol{W}_d)}{\sqrt{\text{Tr}(\boldsymbol{W}_d^\top\boldsymbol{\Sigma}_d\boldsymbol{W}_d)^2}\sqrt{\text{Tr}(\boldsymbol{W}_l^\top\boldsymbol{\Sigma}_l\boldsymbol{W}_l)^2}}\right)^{\frac{1}{2}},$$

$\boldsymbol{\Sigma}_d$ is the marginal covariance matrix of view d, and $\boldsymbol{\Sigma}_{dl}$ is the marginal cross-covariance matrix of view d and l as in Section 2.2.1. This criterion is similar to sum correlation in Gross and Tibshirani (2015), however our definition uses population covariance matrices rather than the sample counterparts.

2.4.3 Two datasets, two groups

We set $n=160$, $K=2$, and generate n independent $y\in\{1,2\}$ with $\pi_1=0.4$, and pairs $(\boldsymbol{x}_1,\boldsymbol{x}_2)\in\mathbb{R}^{p_1}\times\mathbb{R}^{p_2}$ with $(p_1,p_2)\in\{(100,100),(100,500),(500,500)\}$ following Section 2.4.1. We consider autocorrelation structures $\widetilde{\boldsymbol{\Sigma}}_1=(0.8^{|i-j|})_{ij}$, $\widetilde{\boldsymbol{\Sigma}}_2=(0.5^{|i-j|})_{ij}$, and set the value of canonical correlation due to shared class as $\rho=0.8$ by letting $c_1=c_2=\sqrt{\rho/(1-\rho)}$ in generating B_d in Section 2.4.1. We consider the following cases for other shared factors:

Case 1: $q=0$, no shared factors except class y;

Case 2: $q=2$ with corresponding values for canonical correlations being 0.6 and 0.5;

Case 3: $q=2$ with corresponding values for canonical correlations being 0.9 and 0.5.

Table 2.1: Comparison of misclassification rates of Case 1 over 100 replications when $D = 2$, $K = 2$. Standard errors are given in the brackets and the lowest values are highlighted in bold.

(p_1, p_2)	Error rate (%)	JACA	SLDA sep	SLDA joint	Sparse CCA	Sparse sCCA	CVR
(100,100)	$(\boldsymbol{X}_1)$	**4.496 (0.037)**	4.809 (0.070)	4.582 (0.044)	6.376 (0.048)	6.675 (0.043)	6.434 (0.189)
	$(\boldsymbol{X}_2)$	**3.168 (0.040)**	3.533 (0.090)	4.552 (0.127)	4.069 (0.051)	4.415 (0.052)	7.860 (0.415)
	$(\boldsymbol{X}_1, \boldsymbol{X}_2)$	**0.594 (0.011)**	0.729 (0.024)	0.934 (0.030)	1.708 (0.016)	1.862 (0.019)	2.197 (0.118)
(100,500)	$(\boldsymbol{X}_1)$	**4.299 (0.036)**	4.593 (0.075)	4.418 (0.042)	6.286 (0.045)	6.612 (0.043)	6.485 (0.220)
	$(\boldsymbol{X}_2)$	**3.103 (0.050)**	3.279 (0.041)	4.519 (0.107)	3.955 (0.061)	6.445 (0.080)	8.883 (0.418)
	$(\boldsymbol{X}_1, \boldsymbol{X}_2)$	**0.548 (0.013)**	0.695 (0.028)	0.879 (0.027)	1.644 (0.019)	2.283 (0.025)	2.417 (0.118)
(500,500)	$(\boldsymbol{X}_1)$	**4.513 (0.035)**	**4.498 (0.033)**	4.675 (0.041)	6.044 (0.040)	7.250 (0.060)	6.634 (0.167)
	$(\boldsymbol{X}_2)$	**3.537 (0.042)**	3.764 (0.049)	4.938 (0.121)	4.546 (0.047)	6.713 (0.076)	8.732 (0.326)
	$(\boldsymbol{X}_1, \boldsymbol{X}_2)$	**0.629 (0.010)**	0.670 (0.012)	0.953 (0.024)	1.447 (0.015)	2.353 (0.029)	2.408 (0.104)

Table 2.2: Comparison of misclassification rates of Case 2 over 100 replications when $D = 2$, $K = 2$. Standard errors are given in the brackets and the lowest values are highlighted in bold.

(p_1, p_2)	Error rate (%)	JACA	SLDA sep	SLDA joint	Sparse CCA	Sparse sCCA	CVR
(100,100)	$(\boldsymbol{X}_1)$	**4.479 (0.038)**	4.785 (0.064)	4.571 (0.039)	9.992 (0.798)	6.920 (0.077)	8.004 (0.418)
	$(\boldsymbol{X}_2)$	**3.224 (0.041)**	3.687 (0.127)	4.915 (0.146)	8.062 (0.873)	4.675 (0.071)	8.468 (0.364)
	$(\boldsymbol{X}_1, \boldsymbol{X}_2)$	**0.601 (0.011)**	0.779 (0.041)	0.972 (0.027)	5.489 (0.895)	2.010 (0.037)	2.558 (0.114)
(100,500)	$(\boldsymbol{X}_1)$	**4.289 (0.034)**	4.616 (0.084)	4.425 (0.044)	9.096 (0.831)	6.990 (0.079)	8.126 (0.440)
	$(\boldsymbol{X}_2)$	**3.088 (0.048)**	3.264 (0.040)	4.473 (0.094)	6.552 (0.890)	6.542 (0.081)	9.760 (0.392)
	$(\boldsymbol{X}_1, \boldsymbol{X}_2)$	**0.543 (0.011)**	0.687 (0.025)	0.876 (0.024)	4.476 (0.945)	2.495 (0.045)	3.029 (0.147)
(500,500)	$(\boldsymbol{X}_1)$	4.541 (0.040)	**4.493 (0.040)**	4.675 (0.039)	13.489 (1.364)	7.307 (0.059)	7.575 (0.252)
	$(\boldsymbol{X}_2)$	**3.572 (0.042)**	3.785 (0.043)	4.975 (0.126)	12.166 (1.433)	6.804 (0.079)	10.151 (0.416)
	$(\boldsymbol{X}_1, \boldsymbol{X}_2)$	**0.632 (0.011)**	0.671 (0.015)	0.956 (0.026)	9.658 (1.557)	2.391 (0.030)	2.952 (0.135)

In Case 2, the leading canonical correlation between the views is due to shared class membership despite the presence of other shared factors, whereas in Case 3 the leading canonical correlation is due to factors independent from class membership. In order to evaluate the misclassification rates, we further generate $10,000$ new samples as test data, and consider 100 replications for each case. The results are summarized in Tables 2.1–2.4.

JACA gives the best classification results in most scenarios, and has low variance for misclassification rates. JACA also performs the best in terms of sum correlation except for Case 3, where sum correlation for Sparse CCA is stronger. This is not surprising, since in Case 3 the largest canonical correlation is due to the factor independent from class membership. This explanation is

Table 2.3: Comparison of misclassification rates of Case 3 over 100 replications when $D = 2$, $K = 2$. Standard errors are given in the brackets and the lowest values are highlighted in bold.

(p_1, p_2)	Error rate (%)	JACA	SLDA sep	SLDA joint	Sparse CCA	Sparse sCCA	CVR
(100,100)	$(\boldsymbol{X}_1)$	**4.428 (0.034)**	4.647 (0.060)	4.544 (0.040)	40.189 (0.197)	12.862 (1.020)	10.062 (0.511)
	$(\boldsymbol{X}_2)$	**3.295 (0.041)**	3.606 (0.099)	5.278 (0.154)	40.446 (0.257)	11.296 (1.036)	9.638 (0.424)
	$(\boldsymbol{X}_1, \boldsymbol{X}_2)$	**0.609 (0.011)**	0.746 (0.034)	1.030 (0.030)	40.306 (0.217)	8.862 (1.148)	2.538 (0.105)
(100,500)	$(\boldsymbol{X}_1)$	**4.298 (0.034)**	4.686 (0.085)	4.441 (0.044)	40.268 (0.188)	10.256 (0.612)	9.232 (0.448)
	$(\boldsymbol{X}_2)$	**3.091 (0.049)**	3.274 (0.041)	4.456 (0.102)	40.453 (0.236)	8.911 (0.426)	11.364 (0.454)
	$(\boldsymbol{X}_1, \boldsymbol{X}_2)$	**0.544 (0.013)**	0.723 (0.024)	0.872 (0.024)	40.362 (0.201)	5.908 (0.619)	3.029 (0.119)
(500,500)	$(\boldsymbol{X}_1)$	4.537 (0.039)	**4.471 (0.030)**	4.664 (0.039)	40.583 (0.262)	8.947 (0.340)	9.216 (0.483)
	$(\boldsymbol{X}_2)$	**3.577 (0.042)**	3.799 (0.057)	5.017 (0.125)	40.566 (0.255)	8.404 (0.303)	10.944 (0.372)
	$(\boldsymbol{X}_1, \boldsymbol{X}_2)$	**0.626 (0.011)**	0.657 (0.011)	0.960 (0.024)	40.575 (0.261)	4.118 (0.346)	3.067 (0.118)

Table 2.4: Comparison of sum correlation over 100 replications when $D = 2$, $K = 2$. Standard errors are given in the brackets and the highest values are highlighted in bold.

Case	(p_1, p_2)	JACA	SLDA sep	SLDA joint	Sparse CCA	Sparse sCCA	CVR
Case 1	(100,100)	**0.752 (0.001)**	0.744 (0.001)	0.732 (0.002)	0.715 (0.001)	0.708 (0.001)	0.670 (0.006)
	(100,500)	**0.750 (0.001)**	0.743 (0.001)	0.730 (0.001)	0.717 (0.001)	0.685 (0.001)	0.656 (0.006)
	(500,500)	**0.750 (0.001)**	0.747 (0.001)	0.729 (0.002)	0.716 (0.001)	0.677 (0.001)	0.661 (0.005)
Case 2	(100,100)	**0.752 (0.001)**	0.742 (0.002)	0.728 (0.002)	0.686 (0.006)	0.704 (0.001)	0.641 (0.009)
	(100,500)	**0.751 (0.001)**	0.743 (0.001)	0.731 (0.001)	0.681 (0.011)	0.682 (0.001)	0.623 (0.008)
	(500,500)	**0.750 (0.001)**	0.748 (0.001)	0.729 (0.002)	0.604 (0.021)	0.676 (0.001)	0.632 (0.006)
Case 3	(100,100)	0.751 (0.001)	0.744 (0.002)	0.724 (0.002)	**0.874 (0.000)**	0.715 (0.003)	0.549 (0.017)
	(100,500)	0.751 (0.001)	0.742 (0.001)	0.731 (0.001)	**0.861 (0.000)**	0.684 (0.002)	0.553 (0.014)
	(500,500)	0.750 (0.001)	0.748 (0.001)	0.729 (0.002)	**0.854 (0.000)**	0.675 (0.001)	0.573 (0.012)

also supported by the poor classification results for Sparse CCA in Case 3. In Table 2.3, Sparse CCA achieves around 40% misclassification rate, which is no better than random guessing. Finally, CVR is slightly better than Sparse CCA in Case 2 and worse than Sparse CCA in Case 3 in terms of miscalssification rates. However, it performs worse than JACA and SLDA methods. We conjecture this is likely due to CVR using logistic model for estimation rather than factor model (2.3).

2.4.4 Multiple datasets, multiple groups

We set $n = 240$, $K = 3$, and generate n independent $y \in \{1, 2, 3\}$ with $\pi_1 = 0.4$, $\pi_2 = \pi_3 = 0.3$. We also generate n tuples $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \times \mathbb{R}^{p_3}$ with $p_1 = p_2 = p_3 \in \{100, 500\}$

following Section 2.4.1, and set $\widetilde{\mathbf{\Sigma}}_1 = (0.8^{|i-j|})_{ij}$, $\widetilde{\mathbf{\Sigma}}_2 = (0.5^{|i-j|})_{ij}$ and $\widetilde{\mathbf{\Sigma}}_3 = \boldsymbol{I}$. We let canonical correlations due to class membership be $\rho_1 = \rho_2 = 0.8$, and consider the following cases for other shared factors:

Case 1: $q = 0$, no shared factors except class y;

Case 2: $q = 3$ with $\rho_3 = \rho_4 = \rho_5 = 0.6$;

Case 3: $q = 3$ with $\rho_3 = 0.9$, $\rho_4 = 0.9$, $\rho_5 = 0.5$.

Similar to Section 2.4.3, the misclassification rates are evaluated on $10,000$ independently generated test samples. We do not consider Sparse CCA methods because they are not directly applicable to the case of more than two views and more than two classes. While the issue of more than two views can be addressed by Multi CCA generalization (Witten and Tibshirani, 2009), both Sparse CCA and Multi CCA find $K - 1$ pairs of canonical vectors sequentially. As a result, one also needs to tune sparsity parameters sequentially leading to computationally expensive procedure with different sparsity patterns across canonical vector pairs. We also do not consider CVR as it is only implemented for binary classification problem.

The results for JACA and SLDA methods are reported in Tables 2.5 and 2.6. JACA performs the best in terms of misclassification rates in most scenarios, and always performs the best in terms of sum correlation. When predicted based on $\boldsymbol{X}_1$ alone, JACA has similar performance with SLDA_sep, but SLDA_sep's performance decreases significantly as p increases. On the other hand, SLDA_joint performs poorly in most cases.

2.5 Data analysis

2.5.1 TCGA-COAD dataset

We consider the colorectal cancer (COAD) data from The Cancer Genome Atlas project with two views: RNAseq data of normalized counts and miRNA expression. We extracted samples corresponding to primary tumor tissue using TCGA2STAT R package (Wan et al., 2015). To account for data skewness and zero counts, we further log10-transformed both datasets with offset 1, and

Table 2.5: Comparison of misclassification rates over 100 replication when $D = 3$, $K = 3$. Standard errors are given in the brackets and the lowest values are highlighted in bold.

		$p_1 = p_2 = p_3 = 100$			$p_1 = p_2 = p_3 = 500$		
	Error rate (%)	JACA	SLDA sep	SLDA joint	JACA	SLDA sep	SLDA joint
Case 1	$(\boldsymbol{X}_1)$	2.632 (0.051)	**2.511 (0.056)**	7.182 (0.155)	**4.555 (0.076)**	5.398 (0.105)	7.014 (0.150)
	$(\boldsymbol{X}_2)$	**2.112 (0.017)**	2.350 (0.046)	4.746 (0.241)	**1.988 (0.013)**	2.343 (0.062)	4.328 (0.160)
	$(\boldsymbol{X}_3)$	**1.750 (0.016)**	1.802 (0.035)	22.344 (0.957)	**1.450 (0.016)**	1.581 (0.041)	22.041 (1.044)
	$(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3)$	**0.010 (0.001)**	0.015 (0.001)	0.389 (0.034)	**0.025 (0.001)**	0.051 (0.003)	0.430 (0.036)
Case 2	$(\boldsymbol{X}_1)$	2.545 (0.049)	**2.370 (0.040)**	7.389 (0.166)	**4.524 (0.077)**	5.361 (0.100)	7.245 (0.188)
	$(\boldsymbol{X}_2)$	**2.127 (0.017)**	2.363 (0.043)	4.596 (0.151)	**1.994 (0.013)**	2.225 (0.042)	4.441 (0.165)
	$(\boldsymbol{X}_3)$	**1.770 (0.016)**	1.790 (0.032)	22.771 (0.935)	**1.458 (0.017)**	1.550 (0.037)	22.573 (0.992)
	$(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3)$	**0.009 (0.001)**	0.015 (0.001)	0.363 (0.027)	**0.025 (0.001)**	0.048 (0.003)	0.449 (0.036)
Case 3	$(\boldsymbol{X}_1)$	2.384 (0.039)	**2.289 (0.039)**	7.355 (0.140)	**4.391 (0.080)**	5.351 (0.105)	7.259 (0.182)
	$(\boldsymbol{X}_2)$	**2.139 (0.018)**	2.393 (0.049)	4.534 (0.147)	**2.006 (0.015)**	2.337 (0.059)	4.536 (0.190)
	$(\boldsymbol{X}_3)$	**1.818 (0.017)**	**1.809 (0.033)**	23.773 (0.980)	**1.485 (0.018)**	1.589 (0.036)	22.821 (1.004)
	$(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3)$	**0.010 (0.001)**	0.016 (0.001)	0.364 (0.029)	**0.025 (0.001)**	0.052 (0.004)	0.472 (0.038)

Table 2.6: Comparison of sum correlation over 100 replication when $D = 3$, $K = 3$. Standard errors are given in the brackets and the highest values are highlighted in bold.

	$p_1 = p_2 = p_3 = 100$			$p_1 = p_2 = p_3 = 500$		
	JACA	SLDA sep	SLDA joint	JACA	SLDA sep	SLDA joint
Case 1	**2.321 (0.001)**	2.309 (0.004)	1.196 (0.021)	**2.282 (0.001)**	2.185 (0.011)	1.231 (0.023)
Case 2	**2.322 (0.001)**	2.314 (0.002)	1.190 (0.020)	**2.282 (0.001)**	2.186 (0.010)	1.213 (0.023)
Case 3	**2.326 (0.001)**	2.316 (0.003)	1.196 (0.020)	**2.284 (0.002)**	2.183 (0.011)	1.206 (0.022)

filtered the data to select 1572 variables for RNA-Seq and 375 for miRNA with highest standard deviation across samples. Recently, the Colorectal Cancer Consortium determined 4 consensus molecular subtypes (CMS) of colorectal cancer based on gene expression (Guinney et al., 2015), and we have extracted the assigned subtypes for COAD data from the Synapse platform (Synapse ID syn2623706). The resulting data has 282 subjects in total, with Table 2.7 displaying the pattern of available information for each subject. Our primary goal is to identify covarying patterns between RNA-Seq and miRNA data that are relevant for subtype discrimination.

First, we compare different methods from Section 2.4 using the subset of subjects with complete views and subtype information ($n = 167$). We do not consider CVR since it is only imple-

Table 2.7: Number of available samples in COAD data with different missing patterns of CMS class/RNAseq/miRNA.

CMS class	RNAseq	miRNA	Sample size
yes	yes	yes	167
yes	yes	no	27
no	yes	yes	51
no	yes	no	37
			Total: 282

mented for the binary classification problem. We randomly select 137 subjects for training and 35 for testing for the total of 100 random splits. The average misclassification rates and the number of selected variables for each method are presented in Table 2.8. We consider two prediction approaches for each method: prediction based on one view alone (either RNA-seq or miRNA) using the corresponding subset of canonical vectors, and prediction using the concatenated dataset. In general, the performance using miRNA data is worse, which is not surprising since the subtypes were determined using gene expression data alone (Guinney et al., 2015). Although JACA selects more variables than SLDA_sep, it performs the best in terms of misclassification rates, with SLDA_sep being the second best. SLDA_joint achieves a competitive misclassification rate using RNAseq data but not miRNA. We conjecture this is because RNAseq view has a much stronger class-specific signal that masks miRNA's signal when datasets are concatenated. This explanation is supported by the mean number of variables selected by SLDA_joint from each view. Both supervised and unsupervised CCA methods perform poorly in classification. Based on results from Section 2.4, this suggests that the subtype-specific association between the views is weak compared to association due to other common factors.

We also compare the out-of-sample correlation values, that is $\mathrm{Cor}(\boldsymbol{X}_1\widehat{\boldsymbol{W}}_1, \boldsymbol{X}_2\widehat{\boldsymbol{W}}_2)$, where $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ are RNAseq and miRNA data from test samples, and $\widehat{\boldsymbol{W}}_1$, $\widehat{\boldsymbol{W}}_2$ are estimated on the training data. We do not consider CCA methods due to their poor classification performance. The results are presented in Table 2.9, with JACA achieving the strongest correlation value.

Table 2.8: Mean misclassification rates in percentages and mean number of selected features over 100 random splits of 167 samples from COAD data with complete information, standard errors are given in brackets and the lowest values are highlighted in bold.

	Misclassification Rate (%)			Cardinality		
Method	RNAseq	miRNA	Both	RNAseq	miRNA	Both
JACA	**2.06** (0.30)	**6.03** (0.49)	**3.49** (0.35)	385.1 (8.8)	202.3 (3.4)	587.4 (12.2)
SLDA_sep	3.91 (0.42)	7.97 (0.61)	4.03 (0.41)	65.4 (3.1)	57.8 (1.6)	123.2 (3.7)
SLDA_joint	4.26 (0.45)	53.26 (1.82)	4.11 (0.46)	**59.5** (2.8)	**3.3** (0.4)	**62.8** (3.2)
Sparse sCCA	41.89 (0.34)	47.71 (0.47)	42.6 (0.30)	932.5 (4.0)	251.4 (1.0)	1183.9 (4.6)
Sparse CCA	42.11 (0.35)	47.97 (0.46)	42.37 (0.34)	1287.5 (6.9)	369.5 (0.6)	1657 (6.8)

Table 2.9: Analysis based on 167 samples from COAD data with complete view and subtype information based on 100 random splits. Mean correlation between $\boldsymbol{X}_1\widehat{\boldsymbol{W}}_1$ and $\boldsymbol{X}_2\widehat{\boldsymbol{W}}_2$ where $\boldsymbol{X}_1, \boldsymbol{X}_2$ are samples from test data, and $\widehat{\boldsymbol{W}}_1$, $\widehat{\boldsymbol{W}}_2$ are estimated from the training data, standard errors are given in brackets and the highest value is highlighted in bold.

	JACA	SLDA_sep	SLDA_joint
Correlation	**0.95** (0.002)	0.88 (0.003)	0.36 (0.027)

2.5.2 TCGA-BRCA dataset

We consider breast cancer data from The Cancer Genome Atlas project with 4 views: gene expression (GE), DNA methylation (ME), miRNA expression (miRNA), and reverse phase protein array (RPPA). The samples are separated into 4 breast cancer subtypes: Basal, LumA, LumB and Her2 (The Cancer Genome Atlas Network, 2012). Five samples are labelled as Normal-like, and we exclude them from the analyses. Li et al. (2016) incorporate subtypes into supervised singular value decomposition, however only GE view is considered. Lock and Dunson (2013) and Gaynanova and Li (2019) jointly analyze all views, however do not take advantage of the subtypes. In this section, we apply JACA to understand the subtype-driven relationships between the views. We use data from https://tcga-data.nci.nih.gov/docs/publications/brca_2012 and the same data-processing as in Lock and Dunson (2013). While the combined number of subjects is 792, only 377 have complete view/subtype information (see Table 2.10).

Table 2.10: Number of samples in BRCA data with different missing patterns of views and cancer subtype. There are only 377 samples with complete information.

GE	ME	miRNA	RPPA	Cancer type	Count
yes	yes	yes	yes	yes	377
yes	yes	yes	no	yes	114
yes	yes	no	yes	yes	19
yes	yes	no	no	yes	3
yes	no	yes	yes	yes	1
no	yes	yes	yes	no	1
no	yes	yes	no	no	193
no	yes	no	no	no	84
					Total = 792

First, we compare JACA with SLDA_sep and SLDA_joint on the 377 subjects with complete view/subtype information following the same strategy as in Section 2.5.1. We do not consider CVR due to $K > 2$ and $D > 2$, and we do not consider Sparse CCA or Sparse sCCA due to their poor performance on COAD data. Tables 2.11 and 2.12 display the mean misclassification error rates and the number of selected variables for each view, where the predictions are made either separately on each view, or jointly using all views. The results are similar to Section 2.5.1. The error rates are higher when using ME, miRNA or RPPA compared to GE, which is not surprising since BRCA subtypes are originally determined based on gene expression. SLDA_joint achieves a similar error rate using GE alone and higher error rates when using other views. The reason is that it selects very few variables from other views since the subtype-specific signal is the strongest in GE view. JACA has similar performance with SLDA_sep using GE, but outperforms SLDA_sep on other views, which suggests the advantage of taking into account the associations between the views. JACA also has higher cardinality, which is consistent with simulation results in Section 2.4. Table 2.13 displays the sum correlation, with JACA performing best compared to SLDA methods.

Table 2.11: Mean misclassification error rates over 100 splits of 377 samples from BRCA data, standard errors are given in brackets and the lowest values are highlighted in bold.

	Misclassification Rate (%)				
Method	GE	ME	miRNA	RPPA	All
JACA	**10.17** (0.34)	**16.65** (0.51)	**16.4** (0.41)	**21.76** (0.43)	13.54 (0.42)
SLDA_sep	**10.15** (0.42)	20.32 (0.73)	17.32 (0.49)	23.4 (0.46)	12.4 (0.44)
SLDA_joint	10.77 (0.36)	51.33 (1.11)	54.95 (1.52)	42.4 (1.04)	**10.79** (0.36)

Table 2.12: Mean numbers of selected features over 100 splits of 377 samples from BRCA data, standard errors are given in brackets and the lowest values are highlighted in bold.

	Cardinality				
Method	GE	ME	miRNA	RPPA	All
JACA	183.2 (1.8)	191.6 (2)	129.7 (1.5)	82.2 (0.7)	374.8 (3.6)
SLDA_sep	62.8 (2.8)	85.8 (4.5)	48.6 (2.4)	27.2 (1.8)	148.6 (5.6)
SLDA_joint	**48.1** (2.3)	**2.6** (0.3)	**1.8** (0.2)	**3.4** (0.2)	**50.7** (2.5)

Table 2.13: Analysis based on 377 samples from BRCA data with complete view and subtype information based on 100 random splits. Mean correlation between $\boldsymbol{X}_1\widehat{\boldsymbol{W}}_1$ and $\boldsymbol{X}_2\widehat{\boldsymbol{W}}_2$ where $\boldsymbol{X}_1, \boldsymbol{X}_2$ are samples from test data, and $\widehat{\boldsymbol{W}}_1$, $\widehat{\boldsymbol{W}}_2$ are estimated from the training data, standard errors are given in brackets and the highest value is highlighted in bold.

	JACA	SLDA_sep	SLDA_joint
Correlation	**5.54** (0.01)	5.06 (0.014)	1.26 (0.057)

2.6 Additional simulation studies

2.6.1 Alternative evaluation criteria

In this section, we compare different methods in terms of estimation consistency and variable selection. Let $\boldsymbol{\Theta}_d \propto \widetilde{\boldsymbol{\Sigma}}_d^{-1}\boldsymbol{\Delta}_d \in \mathbb{R}^{p_d\times(K-1)}$ be the population matrix of class-specific canonical vectors for view d with $\widetilde{\boldsymbol{\Sigma}}_d$ as in (2.3), and $\boldsymbol{W}_d$ be the estimated matrix. To evaluate estimation

performance, we consider

$$\text{Cor}_{\boldsymbol{\Sigma}}(\boldsymbol{W}_d, \boldsymbol{\Theta}_d) = \left(\frac{\text{Tr}(\boldsymbol{W}_d^\top \widetilde{\boldsymbol{\Sigma}}_d \boldsymbol{\Theta}_d \boldsymbol{\Theta}_d^\top \widetilde{\boldsymbol{\Sigma}}_d \boldsymbol{W}_d)}{\sqrt{\text{Tr}(\boldsymbol{W}_d^\top \widetilde{\boldsymbol{\Sigma}}_d \boldsymbol{W}_d)^2} \sqrt{\text{Tr}(\boldsymbol{\Theta}_d^\top \widetilde{\boldsymbol{\Sigma}}_d \boldsymbol{\Theta}_d)^2}} \right)^{\frac{1}{2}}$$

as a measure of similarity between $\boldsymbol{W}_d$ and $\boldsymbol{\Theta}_d$ with $\text{Cor}_{\boldsymbol{\Sigma}}(\boldsymbol{W}_d, \boldsymbol{\Theta}_d) = 1$ if and only if $\boldsymbol{W}_d$ is equal to $\boldsymbol{\Theta}_d$ up to scaling and orthogonal transformation, and $\text{Cor}_{\boldsymbol{\Sigma}}(\boldsymbol{W}_d, \boldsymbol{\Theta}_d) = \mathbf{0}$ if $\boldsymbol{W}_d^\top \widetilde{\boldsymbol{\Sigma}}_d \boldsymbol{\Theta}_d = 0$. We do not use the Frobenius norm considered since it is not invariant to column scaling and orthogonal transformation, and hence will make the evaluation positively biased towards our proposed method.

We use precision and recall to compare the methods in terms of variable selection. Let $\boldsymbol{A}_d$ be the set of nonzero rows of $\boldsymbol{\Theta}_d$, and let $\widehat{\boldsymbol{A}_d}$ be the set of nonzero rows in $\widehat{\boldsymbol{W}}_d$. Let $\#\{\boldsymbol{A}_d\}$ denote the cardinality of $\boldsymbol{A}_d$. We define the precision and recall as

$$\text{Precision}(\boldsymbol{W}_d) = \frac{\#\{\boldsymbol{A}_d \cap \widehat{\boldsymbol{A}_d}\}}{\#\{\widehat{\boldsymbol{A}_d}\}} \quad \text{and} \quad \text{Recall}(\boldsymbol{W}_d) = \frac{\#\{\boldsymbol{A}_d \cap \widehat{\boldsymbol{A}_d}\}}{\#\{\boldsymbol{A}_d\}}.$$

2.6.2 Two datasets, two groups

We consider the simulation setting from Section 2.4.3. The estimation consistency results are summarized in Tables 2.14–2.16, and the values of precision and recall for different methods are reported in Figure 2.1. Overall, JACA gives the best estimation results in most scenarios, and has low estimation variance. Since in Case 3 the largest canonical correlation is due to the factor independent from class membership, the loadings estimated from sparse CCA are almost orthogonal to the true discriminant vectors $\boldsymbol{\Theta}_d$ as demonstrated by low values of $\text{Cor}_{\boldsymbol{\Sigma}}(\boldsymbol{W}_d, \boldsymbol{\Theta}_d)$. JACA also achieves the best trade off between precision and recall. JACA's precision is second best to SLDA_joint, but SLDA_joint has the lowest values of recall. JACA's recall is comparable to SLDA_sep and worse than the recall of sparse CCA methods, but the latter has low values of precision.

Table 2.14: Comparison of estimation correlation of Case 1 over 100 replications when $D = 2$, $K = 2$. Standard errors are given in the brackets and the highest values are highlighted in bold.

(p_1, p_2)	$\text{Cor}_{\boldsymbol{\Sigma}}$	JACA	SLDA sep	SLDA joint	Sparse CCA	Sparse sCCA	CVR
(100,100)	$(\boldsymbol{W}_1, \boldsymbol{\Theta}_1)$	**0.839** (0.002)	0.823 (0.003)	0.835 (0.002)	0.752 (0.002)	0.740 (0.002)	0.756 (0.007)
	$(\boldsymbol{W}_2, \boldsymbol{\Theta}_2)$	**0.907** (0.003)	0.889 (0.005)	0.825 (0.006)	0.841 (0.003)	0.825 (0.003)	0.704 (0.013)
(100,500)	$(\boldsymbol{W}_1, \boldsymbol{\Theta}_1)$	**0.842** (0.002)	0.824 (0.003)	0.833 (0.002)	0.755 (0.002)	0.742 (0.001)	0.751 (0.007)
	$(\boldsymbol{W}_2, \boldsymbol{\Theta}_2)$	**0.893** (0.003)	0.882 (0.003)	0.816 (0.005)	0.844 (0.003)	0.734 (0.003)	0.666 (0.011)
(500,500)	$(\boldsymbol{W}_1, \boldsymbol{\Theta}_1)$	**0.839** (0.002)	**0.839** (0.002)	0.830 (0.002)	0.758 (0.001)	0.711 (0.002)	0.745 (0.006)
	$(\boldsymbol{W}_2, \boldsymbol{\Theta}_2)$	**0.897** (0.003)	0.883 (0.003)	0.817 (0.006)	0.836 (0.003)	0.738 (0.003)	0.674 (0.009)

Table 2.15: Comparison of estimation correlation of Case 2 over 100 replications when $D = 2$, $K = 2$. Standard errors are given in the brackets and the highest values are highlighted in bold.

(p_1, p_2)	$\text{Cor}_{\boldsymbol{\Sigma}}$	JACA	SLDA sep	SLDA joint	Sparse CCA	Sparse sCCA	CVR
(100,100)	$(\boldsymbol{W}_1, \boldsymbol{\Theta}_1)$	**0.840** (0.002)	0.825 (0.003)	0.836 (0.002)	0.687 (0.016)	0.744 (0.002)	0.726 (0.010)
	$(\boldsymbol{W}_2, \boldsymbol{\Theta}_2)$	**0.907** (0.003)	0.883 (0.006)	0.816 (0.006)	0.755 (0.020)	0.823 (0.003)	0.697 (0.011)
(100,500)	$(\boldsymbol{W}_1, \boldsymbol{\Theta}_1)$	**0.844** (0.001)	0.825 (0.003)	0.834 (0.002)	0.704 (0.017)	0.745 (0.002)	0.718 (0.010)
	$(\boldsymbol{W}_2, \boldsymbol{\Theta}_2)$	**0.895** (0.003)	0.883 (0.002)	0.818 (0.004)	0.780 (0.021)	0.732 (0.003)	0.640 (0.010)
(500,500)	$(\boldsymbol{W}_1, \boldsymbol{\Theta}_1)$	**0.838** (0.002)	**0.840** (0.002)	0.831 (0.002)	0.592 (0.029)	0.711 (0.002)	0.718 (0.007)
	$(\boldsymbol{W}_2, \boldsymbol{\Theta}_2)$	**0.898** (0.003)	0.884 (0.003)	0.817 (0.006)	0.657 (0.033)	0.738 (0.003)	0.637 (0.011)

Table 2.16: Comparison of estimation correlation of Case 3 over 100 replications when $D = 2$, $K = 2$. Standard errors are given in the brackets and the highest values are highlighted in bold.

(p_1, p_2)	$\text{Cor}_{\boldsymbol{\Sigma}}$	JACA	SLDA sep	SLDA joint	Sparse CCA	Sparse sCCA	CVR
(100,100)	$(\boldsymbol{W}_1, \boldsymbol{\Theta}_1)$	**0.843** (0.001)	0.830 (0.003)	0.837 (0.002)	0.098 (0.004)	0.682 (0.015)	0.727 (0.007)
	$(\boldsymbol{W}_2, \boldsymbol{\Theta}_2)$	**0.904** (0.003)	0.888 (0.005)	0.805 (0.006)	0.040 (0.003)	0.720 (0.017)	0.714 (0.009)
(100,500)	$(\boldsymbol{W}_1, \boldsymbol{\Theta}_1)$	**0.844** (0.002)	0.822 (0.004)	0.833 (0.002)	0.117 (0.004)	0.728 (0.007)	0.731 (0.006)
	$(\boldsymbol{W}_2, \boldsymbol{\Theta}_2)$	**0.896** (0.003)	0.884 (0.002)	0.820 (0.005)	0.043 (0.003)	0.700 (0.007)	0.625 (0.010)
(500,500)	$(\boldsymbol{W}_1, \boldsymbol{\Theta}_1)$	0.839 (0.002)	**0.842** (0.001)	0.831 (0.002)	0.033 (0.002)	0.694 (0.004)	0.714 (0.008)
	$(\boldsymbol{W}_2, \boldsymbol{\Theta}_2)$	**0.898** (0.003)	0.885 (0.003)	0.817 (0.006)	0.033 (0.003)	0.716 (0.005)	0.636 (0.009)

2.6.3 Multiple datasets, multiple groups

We consider the simulation setting from Section 2.4.4. The estimation consistency results for JACA and SLDA methods are reported in Table 2.17, and the values of precision and recall for

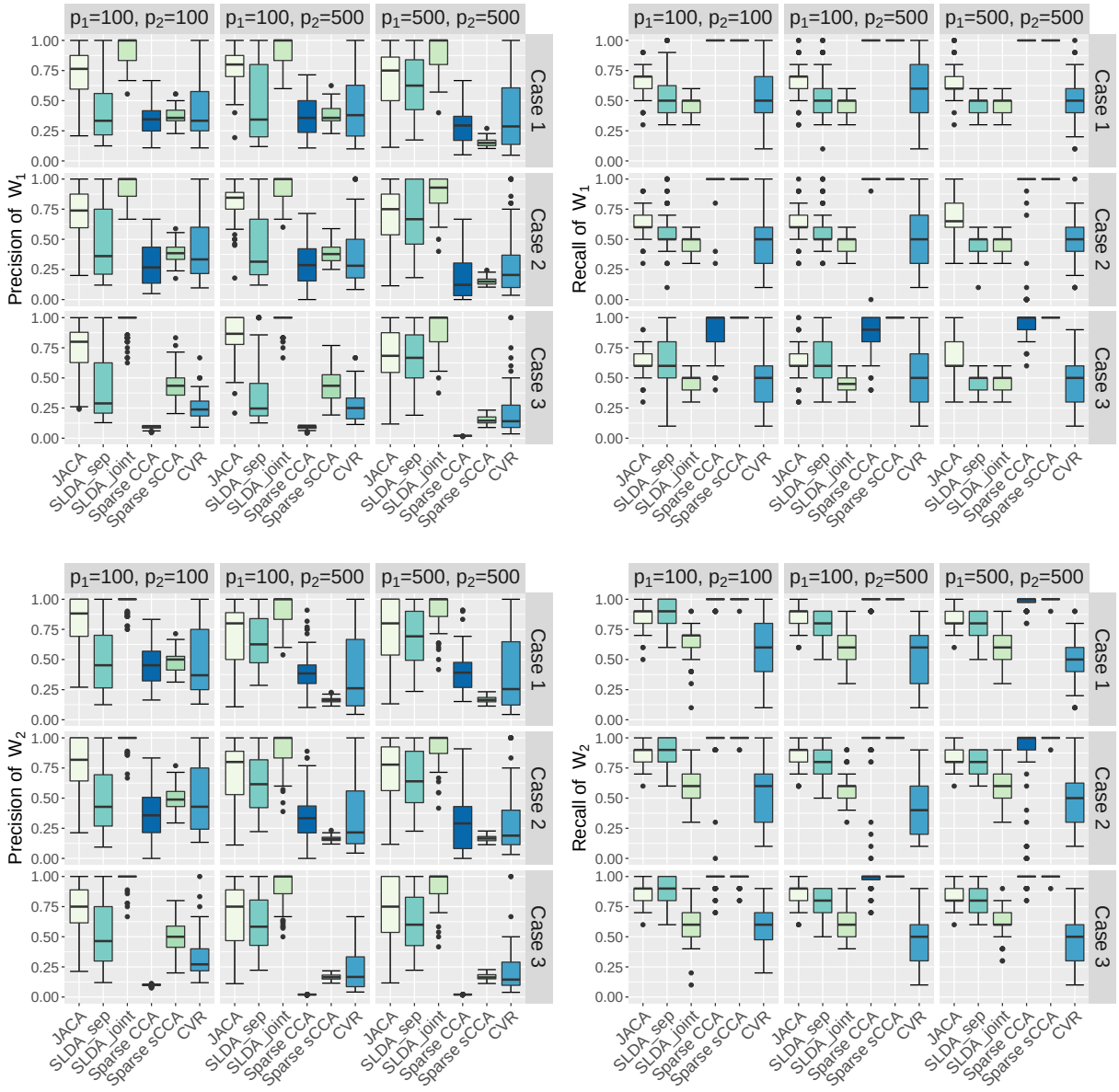

Figure 2.1: Precision and Recall over 100 replications when $D = 2$ and $K = 2$.

different methods are reported in Figure 2.2, and the conclusions are similar to the case of two-groups and two-views. Overall, JACA performs the best in terms of estimation consistency and also achieves the best trade off between precision and recall.

Table 2.17: Comparison of estimation correlation over 100 replication when $D = 3$, $K = 3$. Standard errors are given in the brackets and the highest values are highlighted in bold.

		$p_1 = p_2 = p_3 = 100$			$p_1 = p_2 = p_3 = 500$		
	$\text{Cor}_{\boldsymbol{\Sigma}}$	JACA	SLDA sep	SLDA joint	JACA	SLDA sep	SLDA joint
Case 1	$(\boldsymbol{W}_1, \boldsymbol{\Theta}_1)$	0.903 (0.002)	**0.906** (0.002)	0.795 (0.002)	**0.848** (0.002)	0.825 (0.003)	0.800 (0.002)
	$(\boldsymbol{W}_2, \boldsymbol{\Theta}_2)$	**0.945** (0.001)	0.929 (0.003)	0.794 (0.008)	**0.937** (0.001)	0.913 (0.004)	0.801 (0.008)
	$(\boldsymbol{W}_3, \boldsymbol{\Theta}_3)$	**0.959** (0.001)	**0.960** (0.002)	0.710 (0.010)	**0.969** (0.001)	0.961 (0.003)	0.726 (0.011)
Case 2	$(\boldsymbol{W}_1, \boldsymbol{\Theta}_1)$	0.908 (0.002)	**0.914** (0.002)	0.795 (0.002)	**0.850** (0.002)	0.827 (0.003)	0.798 (0.002)
	$(\boldsymbol{W}_2, \boldsymbol{\Theta}_2)$	**0.946** (0.001)	0.931 (0.003)	0.799 (0.007)	**0.937** (0.001)	0.921 (0.003)	0.797 (0.007)
	$(\boldsymbol{W}_3, \boldsymbol{\Theta}_3)$	0.959 (0.001)	**0.962** (0.002)	0.709 (0.010)	**0.969** (0.001)	0.963 (0.002)	0.726 (0.010)
Case 3	$(\boldsymbol{W}_1, \boldsymbol{\Theta}_1)$	0.917 (0.002)	**0.921** (0.002)	0.802 (0.002)	**0.855** (0.002)	0.828 (0.003)	0.799 (0.002)
	$(\boldsymbol{W}_2, \boldsymbol{\Theta}_2)$	**0.948** (0.001)	0.930 (0.003)	0.805 (0.007)	**0.937** (0.001)	0.914 (0.004)	0.794 (0.008)
	$(\boldsymbol{W}_3, \boldsymbol{\Theta}_3)$	0.957 (0.001)	**0.963** (0.002)	0.702 (0.010)	**0.967** (0.001)	0.960 (0.003)	0.719 (0.010)

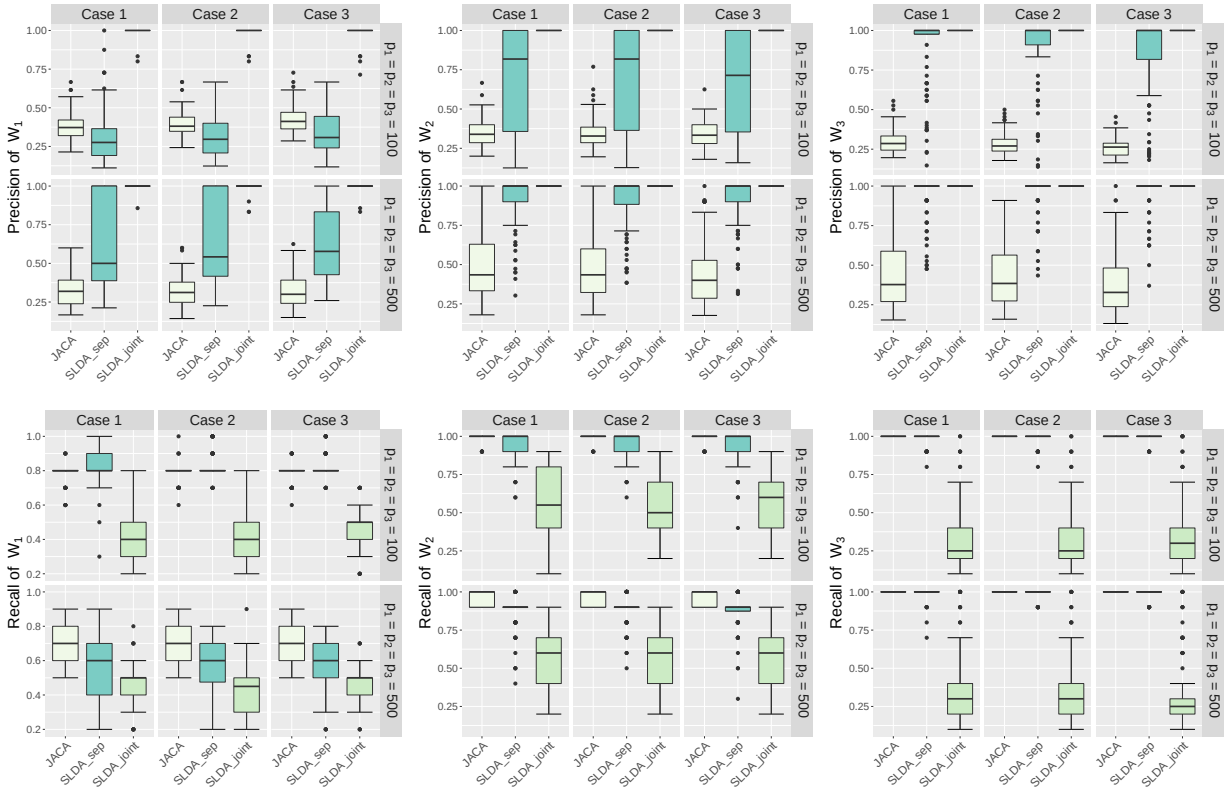

Figure 2.2: Precision and Recall over 100 replications when $D = 3$ and $K = 3$.

2.7 Technical proofs

2.7.1 Proof of Proposition 1

Proof. Under the stated conditions, $\boldsymbol{x}_d$ in (2.2) satisfies (2.1) by construction, therefore it remains to show the reverse. Consider (2.1) with $\boldsymbol{\Sigma}_{ldy} = \mathbf{0}$. Then

$$\boldsymbol{x}_d = \boldsymbol{\mu}_d + \sum_{k=1}^{K} (\boldsymbol{\mu}_{dk} - \boldsymbol{\mu}_d)\, \mathbb{1}\{y = k\} + \boldsymbol{\Sigma}_{dy}^{1/2} \boldsymbol{e}_d,$$

where $\boldsymbol{e}_d$ are independent from y. We next show that there exists function $f : \{1, 2 \ldots, K\} \to \mathbb{R}^{K-1}$ such that $\boldsymbol{\mu}_d + \sum_{k=1}^{K} (\boldsymbol{\mu}_{dk} - \boldsymbol{\mu}_d)\, \mathbb{1}\{y = k\} = \boldsymbol{\mu}_d + \boldsymbol{\Delta}_d f(y) = \boldsymbol{\mu}_d + \boldsymbol{\Delta}_d \boldsymbol{u}_y$ with $\boldsymbol{u}_y$ and $\boldsymbol{\Delta}_d$ satisfying the stated conditions.

Consider $K = 2$. Let $u_y = f(y) = \sqrt{\pi_2/\pi_1}\, \mathbb{1}\{y = 1\} - \sqrt{\pi_1/\pi_2}\, \mathbb{1}\{y = 2\}$, then $\mathbb{E}(u_y) = 0$, $\mathrm{Cov}(u_y) = 1$. Setting $\boldsymbol{\Delta}_d = \sqrt{\pi_1 \pi_2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ gives the desired factor model since

$$\mathbb{E}(\boldsymbol{\mu}_d + \boldsymbol{\Delta}_D u_y | y = 1) = \pi_1 \boldsymbol{\mu}_1 + \pi_2 \boldsymbol{\mu}_2 + \sqrt{\pi_1 \pi_2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\sqrt{\pi_2/\pi_1} = \boldsymbol{\mu}_1,$$

and similarly $\mathbb{E}(\boldsymbol{\mu}_d + \boldsymbol{\Delta}_D u_y | y = 2) = \boldsymbol{\mu}_2$.

Consider $K \geq 2$. Let $\boldsymbol{\Theta} \in \mathbb{R}^{K \times (K-1)}$ have columns $\boldsymbol{\Theta}_l$ with

$$\boldsymbol{\Theta}_l = \left(\left\{ \sqrt{\frac{\pi_{l+1}}{\sum_{i=1}^{l} \pi_i \sum_{i=1}^{l+1} \pi_i}} \right\}_l, -\sqrt{\frac{\sum_{i=1}^{l} \pi_i}{\pi_{l+1} \sum_{i=1}^{l+1} \pi_i}}, 0_{K-1-l} \right)^\top,$$

and let $\boldsymbol{Z} = g(y) \in \mathbb{R}^K$ be a unit norm class-indicator random vector with $z_k = 1$ if observation belongs to class k. Consider $\widetilde{\boldsymbol{u}}_y = \widetilde{f}(y) = \boldsymbol{\Theta}^\top g(y) = \boldsymbol{\Theta}^\top \boldsymbol{Z}$ and let $\boldsymbol{\pi} = (\pi_1 \ldots \pi_K)^\top$. Then

$$\mathbb{E}(\widetilde{\boldsymbol{u}}_y) = \boldsymbol{\Theta}^\top \mathbb{E}(\boldsymbol{Z}) = \boldsymbol{\Theta}^\top \boldsymbol{\pi} = (\boldsymbol{\Theta}_1^\top \boldsymbol{\pi} \ldots \boldsymbol{\Theta}_{K-1}^\top \boldsymbol{\pi}) = \mathbf{0}_{K-1},$$

$$\mathrm{Cov}(\widetilde{\boldsymbol{u}}_y) = \boldsymbol{\Theta}^\top \mathrm{Cov}(\boldsymbol{Z}) \boldsymbol{\Theta} = \boldsymbol{\Theta}^\top (\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top) \boldsymbol{\Theta} = \boldsymbol{\Theta}^\top \mathrm{diag}(\boldsymbol{\pi}) \boldsymbol{\Theta} = \boldsymbol{I}_K.$$

Next define $\widetilde{\boldsymbol{\Delta}}_d \in \mathbb{R}^{p\times(K-1)}$ to have columns $\widetilde{\boldsymbol{\Delta}}_{dr}$ with

$$\widetilde{\boldsymbol{\Delta}}_{dr} = \frac{\sqrt{\pi_{r+1}}\left\{\sum_{i=1}^{r}\pi_i(\boldsymbol{\mu}_{di}-\boldsymbol{\mu}_{d(r+1)})\right\}}{\sqrt{\sum_{i=1}^{r}\pi_i\sum_{i=1}^{r+1}\pi_i}}.$$

Then

$$\begin{aligned}\mathbb{E}(\boldsymbol{\mu}_d+\widetilde{\boldsymbol{\Delta}}_d\widetilde{f}(y)|y=k) =& \sum_{m=1}^{K}\pi_m\boldsymbol{\mu}_{dm}+\widetilde{\boldsymbol{\Delta}}_d\boldsymbol{\Theta}^\top g(k)\\ =& \sum_{m=1}^{K}\pi_m\boldsymbol{\mu}_{dm}-\sqrt{\frac{\sum_{i=1}^{k-1}\pi_i}{\pi_k\sum_{i=1}^{k}\pi_i}}\widetilde{\boldsymbol{\Delta}}_{d(k-1)}+\sum_{l=k}^{K-1}\sqrt{\frac{\pi_{l+1}}{\sum_{i=1}^{l}\pi_i\sum_{i=1}^{l+1}\pi_i}}\widetilde{\boldsymbol{\Delta}}_{dl}\\ =& \boldsymbol{\mu}_{dk},\end{aligned}$$

where in the last step we used the properties of orthogonal group-mean contrasts for unbalanced data, see Searle (2006) and also Proposition 2 in Gaynanova et al. (2016). Consider the eigendecomposition $\widetilde{\boldsymbol{\Delta}}_d^\top\boldsymbol{\Sigma}_{dy}^{-1}\widetilde{\boldsymbol{\Delta}}_d = \boldsymbol{R}_d\boldsymbol{\Lambda}_d\boldsymbol{R}_d^\top$. Setting $\boldsymbol{\Delta}_d = \widetilde{\boldsymbol{\Delta}}_d\boldsymbol{R}_d$ and $\boldsymbol{u}_y = \boldsymbol{R}_d^\top\widetilde{\boldsymbol{u}}_y$ leads to desired factor model. □

2.7.2 Proof of Theorem 1

Proof. 1. When $q=0$, $\boldsymbol{\Sigma}_{ld} = \boldsymbol{\Delta}_l\boldsymbol{\Delta}_d^\top = \boldsymbol{\Sigma}_l\left[\boldsymbol{\Sigma}_l^{-1}\boldsymbol{\Delta}_l\boldsymbol{\Delta}_d^\top\boldsymbol{\Sigma}_d^{-1}\right]\boldsymbol{\Sigma}_d$. Let $\boldsymbol{\Lambda}_d = \boldsymbol{\Delta}_d^\top\widetilde{\boldsymbol{\Sigma}}_d^{-1}\boldsymbol{\Delta}_d$, where $\boldsymbol{\Lambda}_d$ is diagonal by definition of factor model (2.3). Using Woodbury matrix identity,

$$\boldsymbol{\Delta}_d^\top\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\Delta}_d = \boldsymbol{\Delta}_d^\top(\widetilde{\boldsymbol{\Sigma}}_d+\boldsymbol{\Delta}_d\boldsymbol{\Delta}_d^\top)^{-1}\boldsymbol{\Delta}_d = \boldsymbol{\Lambda}_d^{1/2}(\boldsymbol{I}+\boldsymbol{\Lambda}_d)^{-1}\boldsymbol{\Lambda}_d^{1/2}.$$

Let $\boldsymbol{\Theta}_d = \boldsymbol{\Sigma}_d^{-1}\boldsymbol{\Delta}_d\boldsymbol{\Lambda}_d^{-1/2}(\boldsymbol{I}+\boldsymbol{\Lambda}_d)^{1/2}$, then $\boldsymbol{\Theta}_d^\top\boldsymbol{\Sigma}_d\boldsymbol{\Theta}_d = \boldsymbol{I}$, and

$$\boldsymbol{\Sigma}_{ld} = \boldsymbol{\Sigma}_l\left[\boldsymbol{\Theta}_l(I+\boldsymbol{\Lambda}_l)^{-1/2}\boldsymbol{\Lambda}_l^{1/2}\boldsymbol{\Lambda}_d^{1/2}(I+\boldsymbol{\Lambda}_d)^{-1/2}\boldsymbol{\Theta}_d^\top\right]\boldsymbol{\Sigma}_d = \boldsymbol{\Sigma}_l\left(\sum_{k=1}^{K-1}\rho_k\boldsymbol{\theta}_l^{(k)}\boldsymbol{\theta}_d^{(k)\top}\right)\boldsymbol{\Sigma}_d,$$

where ρ_k are the diagonal elements of matrix $(\boldsymbol{I}+\boldsymbol{\Lambda}_l)^{-1/2}\boldsymbol{\Lambda}_l^{1/2}\boldsymbol{\Lambda}_d^{1/2}(\boldsymbol{I}+\boldsymbol{\Lambda}_d)^{-1/2}$, and $\boldsymbol{\theta}_l^{(k)}$, $\boldsymbol{\theta}_d^{(k)}$ are corresponding columns of $\boldsymbol{\Theta}_l$, $\boldsymbol{\Theta}_d$.

2. Consider

$$\begin{aligned}\boldsymbol{\Sigma}_{ld} &= \boldsymbol{A}_l\boldsymbol{A}_d^\top + \boldsymbol{\Delta}_l\boldsymbol{\Delta}_d^\top \\ &= \boldsymbol{\Sigma}_l\left\{\boldsymbol{\Sigma}_l^{-1/2}\Big(\boldsymbol{\Sigma}_l^{-1/2}\boldsymbol{A}_l\boldsymbol{A}_d^\top\boldsymbol{\Sigma}_d^{-1/2} + \boldsymbol{\Sigma}_l^{-1/2}\boldsymbol{\Delta}_l\boldsymbol{\Delta}_d^\top\boldsymbol{\Sigma}_d^{-1/2}\Big)\boldsymbol{\Sigma}_d^{-1/2}\right\}\boldsymbol{\Sigma}_d \\ &= \boldsymbol{\Sigma}_l\left\{\boldsymbol{\Sigma}_l^{-1/2}\Big(\boldsymbol{R}_q\boldsymbol{D}_q\boldsymbol{P}_q^\top + \boldsymbol{R}_{K-1}\boldsymbol{D}_{K-1}\boldsymbol{P}_{K-1}^\top\Big)\boldsymbol{\Sigma}_d^{-1/2}\right\}\boldsymbol{\Sigma}_d,\end{aligned}$$

where we used singular value decomposition

$$\boldsymbol{\Sigma}_l^{-1/2}\boldsymbol{A}_l\boldsymbol{A}_d^\top\boldsymbol{\Sigma}_d^{-1/2} = \boldsymbol{R}_q\boldsymbol{D}_q\boldsymbol{P}_q^\top$$

and

$$\boldsymbol{\Sigma}_l^{-1/2}\boldsymbol{\Delta}_l\boldsymbol{\Delta}_d^\top\boldsymbol{\Sigma}_d^{-1/2} = \boldsymbol{R}_{K-1}\boldsymbol{D}_{K-1}\boldsymbol{P}_{K-1}^\top.$$

Since $\boldsymbol{A}_d^\top\widetilde{\boldsymbol{\Sigma}}_d^{-1}\boldsymbol{\Delta}_d = \mathbf{0}$, by Woodbury matrix identity $\boldsymbol{A}_d^\top\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\Delta}_d = \mathbf{0}$, and therefore $\boldsymbol{R}_q^\top\boldsymbol{R}_{K-1} = \mathbf{0}$ and $\boldsymbol{P}_q^\top\boldsymbol{P}_{k-1} = \mathbf{0}$. From the above display,

$$\boldsymbol{\Sigma}_{ld} = \boldsymbol{\Sigma}_l\left\{\boldsymbol{\Sigma}_l^{-1/2}\boldsymbol{R}_{q+K-1}\boldsymbol{D}_{q+K-1}\boldsymbol{P}_{q+K-1}^\top\boldsymbol{\Sigma}_d^{-1/2}\right\}\boldsymbol{\Sigma}_d.$$

The result follows by setting $\boldsymbol{\Theta}_d = \boldsymbol{\Sigma}_d^{-1/2}\boldsymbol{P}_{q+K-1}$, and using the results from part 1. □

2.7.3 Proof of Proposition 2

Proof. Gaynanova et al. (2016) consider optimization problem

$$\underset{\boldsymbol{V}\in\mathbb{R}^{p\times(K-1)}}{\text{minimize}}\left\{\frac{1}{2}\operatorname{Tr}(\boldsymbol{V}^\top\boldsymbol{W}\boldsymbol{V}) + \frac{1}{2}\|\boldsymbol{D}^\top\boldsymbol{V} - \boldsymbol{I}\|_F^2 + \lambda\sum_{i=1}^{p}\|\boldsymbol{v}_i\|_2\right\}, \tag{2.12}$$

where $\boldsymbol{W} = \frac{1}{n}\sum_{i=1}^{K}(n_i - 1)\boldsymbol{S}_i$ is the within-class sample covariance matrix, $\boldsymbol{S}_i$ is the sample covariance matrix for class i and $\boldsymbol{D}\in\mathbb{R}^{p\times(K-1)}$ has columns $\boldsymbol{D}_l$ defined as

$$\boldsymbol{D}_l = \frac{\sqrt{n_{l+1}}(\sum_{i=1}^{l} n_i(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}_{l+1}))}{\sqrt{n}\sqrt{\sum_{i=1}^{l} n_i\sum_{i=1}^{l+1} n_i}}.$$

Here, $\bar{\boldsymbol{x}}_i$ is the sample mean for class i. Since $\|\boldsymbol{D}^\top \boldsymbol{V} - \boldsymbol{I}\|_F^2 = \mathrm{Tr}\{(\boldsymbol{D}^\top \boldsymbol{V} - \boldsymbol{I})^\top (\boldsymbol{D}^\top \boldsymbol{V} - \boldsymbol{I})\} = \mathrm{Tr}(\boldsymbol{V}^\top \boldsymbol{D}\boldsymbol{D}^\top \boldsymbol{V} - 2\boldsymbol{D}^\top \boldsymbol{V} + \boldsymbol{I})$, function (2.12) can be written as

$$\underset{\boldsymbol{V} \in \mathbb{R}^{p \times (K-1)}}{\text{minimize}} \left\{ \frac{1}{2} \mathrm{Tr}(\boldsymbol{V}^\top (\boldsymbol{W} + \boldsymbol{D}\boldsymbol{D}^\top) \boldsymbol{V}) - \mathrm{Tr}(\boldsymbol{D}^\top \boldsymbol{V}) + \lambda \sum_{i=1}^{p} \|\boldsymbol{v}_i\|_2 \right\}. \tag{2.13}$$

Since $\boldsymbol{X}$ is centered, $\boldsymbol{W} + \boldsymbol{D}\boldsymbol{D}^\top = \boldsymbol{X}^\top \boldsymbol{X}/n$. By the definition of $\boldsymbol{Z}$ and $\boldsymbol{H}$,

$$\boldsymbol{X}^\top \boldsymbol{Z} = \begin{pmatrix} n_1 \bar{\boldsymbol{x}}_1 & \dots & n_k \bar{\boldsymbol{x}}_k \end{pmatrix}, \quad \boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{H} = n\boldsymbol{D}.$$

Plugging the above equality into (2.13) leads to

$$\underset{\boldsymbol{V} \in \mathbb{R}^{p \times (K-1)}}{\text{minimize}} \left\{ \frac{1}{2n} \mathrm{Tr}(\boldsymbol{V}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{V}) - \frac{1}{n} \mathrm{Tr}(\boldsymbol{V}^\top \boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{H}) + \lambda \sum_{i=1}^{p} \|\boldsymbol{v}_i\|_2 \right\}.$$

Denote $\boldsymbol{Z}\boldsymbol{H}$ by $\widetilde{\boldsymbol{Y}}$, then the above display can be expressed as

$$\underset{\boldsymbol{V} \in \mathbb{R}^{p \times (K-1)}}{\text{minimize}} \left\{ \frac{1}{2n} \|\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{V}\|_F^2 + \lambda \sum_{i=1}^{p} \|\boldsymbol{v}_i\|_2 \right\}.$$

□

3. PROPERTIES OF JACA AND ITS SEMI-SUPERVISED EXTENSION

3.1 Introduction

In Chapter 2, we formulate the JACA method for Joint Classification and Association Analysis via a convex optimization problem. In this chapter, we provide theoretical guarantees on the estimation consistency of JACA which are absent for previously proposed joint learning methods (Luo et al., 2016; Gross and Tibshirani, 2015).

While estimation consistency has been established separately for discriminant analysis (Li and Jia, 2017; Gaynanova, 2019) and canonical correlation analysis (Gao et al., 2017), providing similar guarantees for JACA is not straightforward due to the unique structure of our framework. We use the augmented data approach to rewrite our method as a penalized linear regression problem, and use sub-exponential concentration bounds to control the inner-product between the augmented random design matrix and the random matrix of residuals. Despite the dependency between corresponding design matrix and the matrix of residuals, we obtain the estimation error bound that is of the same order as the known bounds for group-lasso linear regression (Lounici et al., 2011; Nardi and Rinaldo, 2008).

Another advantage of the proposed method is that it can be extended to the multi-view data with block-missing structure, that is to cases where a subset of views or class labels is missing for some subjects. In Section 3.5.1 we consider colorectal cancer data, where out of 282 subjects with RNAseq data, only 167 subjects have corresponding miRNA and cancer subtype information. While most methods can only use data from 167 subjects with complete information, our approach can utilize data from 78 extra subjects for which at least two types of information are available (two views with no class labels, or class labels with only one view). Section 3.4 shows an improved classification accuracy of JACA when the subjects with incomplete information are added to the analysis.

In summary, this work has two main contributions. First, we provide finite sample bounds on

estimation consistency of our method in high-dimensional settings. Secondly, we generalize our approach to the settings with block-missing data without the use of imputation.

3.2 Estimation consistency

In this section, we derive the finite sample bound on the estimation error of the minimizer of (2.8) with $\mathrm{Pen}(\boldsymbol{W}_d) = \sum_{i=1}^{p_d} \|\boldsymbol{w}_{di}\|_2$. From Theorem 1, our goal is to estimate the view-specific matrices of discriminant vectors $\boldsymbol{\Theta}_d$, which are equal to $\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\Delta}_d$ up to column scaling. To connect (2.8) with $\boldsymbol{\Theta}_d$, consider the population objective function of (2.8) with $\lambda_d = 0$

$$\mathbb{E}(2^{-1}\|\boldsymbol{Y}' - \boldsymbol{X}'\boldsymbol{W}\|_F^2) = 2^{-1}\,\mathrm{Tr}\{\boldsymbol{W}^\top \mathbb{E}(\boldsymbol{X}'^\top \boldsymbol{X}')\boldsymbol{W}\} - \mathrm{Tr}\{\boldsymbol{W}^\top \mathbb{E}(\boldsymbol{X}'^\top \boldsymbol{Y}')\} + C, \tag{3.1}$$

where C is a constant independent of $\boldsymbol{W}$. Using the definition of augmented $\boldsymbol{X}'$ and $\boldsymbol{Y}'$,

$$\mathbb{E}(\boldsymbol{X}'^\top \boldsymbol{X}') = \begin{pmatrix} \boldsymbol{\Sigma}_1 & -\frac{1-\alpha}{D-1}\boldsymbol{\Sigma}_{12} & \cdots & -\frac{1-\alpha}{D-1}\boldsymbol{\Sigma}_{1D} \\ -\frac{1-\alpha}{D-1}\boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_2 & \cdots & -\frac{1-\alpha}{D-1}\boldsymbol{\Sigma}_{2D} \\ & \vdots & & \\ -\frac{1-\alpha}{D-1}\boldsymbol{\Sigma}_{D1} & -\frac{1-\alpha}{D-1}\boldsymbol{\Sigma}_{D2} & \cdots & \boldsymbol{\Sigma}_D \end{pmatrix} /D := \boldsymbol{G},$$

and using Lemma 8 in Gaynanova and Kolar (2015) for the rth column of $\boldsymbol{X}'^\top \boldsymbol{Y}'$

$$\mathbb{E}(\boldsymbol{X}'^\top \boldsymbol{Y}'_r) = \frac{\alpha}{D}\mathbb{E}\{\frac{1}{n}(\boldsymbol{X}_1^\top \widetilde{\boldsymbol{Y}}_r \ldots \boldsymbol{X}_D^\top \widetilde{\boldsymbol{Y}}_r)^\top\} = \widetilde{\boldsymbol{\Delta}}_r + o(1).$$

Here $o(1)$ term captures the differences between empirical class proportions n_k/n and prior class probabilities π_k, and $\widetilde{\boldsymbol{\Delta}} \in \mathbb{R}^{\left(\sum_{i=1}^D p_i\right)\times(K-1)}$ has rth column defined as

$$\widetilde{\boldsymbol{\Delta}}_r = \frac{\alpha}{D}\frac{\sqrt{\pi_{r+1}}\sum_{k=1}^r \pi_k(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{r+1})}{\sqrt{\sum_{k=1}^r \pi_k \sum_{k=1}^{r+1}\pi_k}}.$$

Therefore, the objective in (3.1) can be written as

$$\mathbb{E}(2^{-1}\|\boldsymbol{Y}' - \boldsymbol{X}'\boldsymbol{W}\|_F^2) = 2^{-1}\operatorname{Tr}\{\boldsymbol{W}^\top \boldsymbol{G}\boldsymbol{W}\} - \operatorname{Tr}\{\boldsymbol{W}^\top \widetilde{\boldsymbol{\Delta}}\} + o(1) + C. \tag{3.2}$$

Let $\boldsymbol{W}^* = \boldsymbol{G}^{-1}\boldsymbol{\Delta}$, then $\boldsymbol{W}^*$ is the minimizer of population loss function in (3.2) up to the $o(1)$ term. We further show that $\boldsymbol{W}^*$ also corresponds to the matrix of discriminant vectors $\boldsymbol{\Theta}_d$ up to orthogonal transformation and column-scaling.

Lemma 1. *Under factor model* (2.3) *and for* $\alpha \in (0, 1]$*, there exists orthogonal matrices* $\boldsymbol{R}_d$ *such that* $\boldsymbol{W}_d^* \boldsymbol{R}_d^\top$ *is equal to* $\boldsymbol{\Theta}_d$ *up to column scaling.*

The proof of Lemma 1 indicates that the choice of α only affects the magnitude of the columns of $\boldsymbol{W}^*$. Thus, $\boldsymbol{W}^*$ corresponds to $\boldsymbol{\Theta}_d$ up to orthogonal transformation and column-scaling for any $\alpha \in (0, 1]$. Thus, the population loss (3.2) can be viewed as the quadratic loss with respect to discriminant vectors $\boldsymbol{\Theta}_d$ with a particular choice of orthogonal transformation and scaling, which affect neither the classification rule nor the row-sparsity pattern. In what follows, we show that minimizer $\widehat{\boldsymbol{W}}$ of (2.8) is consistent at estimating $\boldsymbol{W}^*$ under the following assumptions.

Assumption 3. $\boldsymbol{\Theta}_d$ *is row-sparse with the support* $S_d = \{j : \|e_j^\top \boldsymbol{\Theta}_d\|_2 \neq 0\}$ *and* $s_d = \operatorname{card}(S_d)$ *. Hence* $\boldsymbol{W}_d^*$ *is also row-sparse with the same support, and* $\boldsymbol{W}^*$ *is row-sparse with the support* $S = (S_1, \ldots, S_D)$ *and* $s = \operatorname{card}(S) = \sum_{d=1}^D s_d$*.*

Assumption 4. $P(y = k) = \pi_k$ *for* $k = 1, \ldots, K$ *with* $0 < \pi_{\min} \leq \pi_k \leq \pi_{\max} < 1$*.*

Assumption 5. $\boldsymbol{x}_d | y = k \sim \mathcal{N}(\boldsymbol{\mu}_{dk}, \boldsymbol{\Sigma}_{dy})$ *for all* $k = 1, \ldots, K$*.*

Assumption 6. *Let* $p_{max} = \max_d p_d$ *and* $p_{min} = \min_d p_d$*. Then for some constant* $C > 0$

$$\frac{\log(p_{max})}{\log(p_{min})} \leq C \text{ and } \log p_d = o(n), \text{ for all } d = 1, \ldots, D.$$

These assumptions are typical for multivariate analysis methods and high-dimensional settings. Assumption 3 states that population matrices of discriminant vectors are row-sparse. Assump-

tion 4 states that the class proportions are not degenerate. Assumption 5 states that the measurements are normally distributed conditionally on the class membership, and it can be relaxed to sub-gaussianity without affecting the rates. Assumption 6 allows to have a larger number of measurements than the number of samples, and states that the views have comparable numbers of measurements on the log scale. Because of the log scale, this assumption is mild. For example, taking $p_{max} = 1,000,000$ and $p_{min} = 100$ leads to $C = \log(p_{max})/\log(p_{min}) = 3$.

Similar to the assumptions required for estimation consistency in linear regression with group-lasso penalty (Nardi and Rinaldo, 2008; Lounici et al., 2011), we also require restricted eigenvalue condition satisfied on the weighted cone.

Definition 1 (Weighted cone)**.** *Let* $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$ *and* $S = (S_1, \ldots, S_D)$*. Then*

$$C(S, \boldsymbol{\lambda}) = \Big\{ \boldsymbol{M} \in \mathbb{R}^{\sum_{d=1}^{D} p_d \times (K-1)} : \sum_{d=1}^{D} \lambda_d \|\boldsymbol{M}_{d,S_d^c}\|_{1,2} \leq 3 \sum_{d=1}^{D} \boldsymbol{\lambda}_d \|\boldsymbol{M}_{d,S_d}\|_{1,2} \Big\}.$$

Definition 2. *A matrix* $\boldsymbol{Q} \in \mathbb{R}^{q \times p}$ *satisfies restricted eigenvalue condition RE*$(S, \boldsymbol{\lambda})$ *with parameter* $\gamma_Q = \gamma(S, \boldsymbol{\lambda}, \boldsymbol{Q})$ *if for some set* S*, and for all* $\boldsymbol{A} \in \mathcal{C}(S, \boldsymbol{\lambda})$ *it holds that*

$$\|\boldsymbol{Q}\boldsymbol{A}\|_F^2 \geq \gamma_{\boldsymbol{Q}} \|\boldsymbol{A}\|_F^2.$$

We are now ready to state the main result. Let $\delta = \|\widetilde{\boldsymbol{\Delta}}\|_{\infty,2}$, let $g = \max_j \{\boldsymbol{G}^{-1}\}_{jj}$ be the largest diagonal entry of $\boldsymbol{G}^{-1}$, and let $\tau = \max_j \sqrt{\sigma_j^2 + \max_k \mu_{k,j}^2}$, where σ_j are diagonal elements of $\boldsymbol{\Sigma}_y$ and $\mu_{k,j}$ are elements of $\boldsymbol{\mu}_k$.

Theorem 2. *Under Assumptions 3–6, if* $\lambda_d = C\,(\tau \vee \tau^2 \delta g)\, D^{-1} \sqrt{(K-1)\log[(K-1)p_d]/n}$ *for some constant* $C > 0$*,* $s_d^2 \log[(K-1)p_d] = o(n)$ *and* $\boldsymbol{G}^{-1/2}$ *satisfies condition RE*$(S, \boldsymbol{\lambda})$ *with parameter* $\gamma = \gamma(S, \boldsymbol{\lambda}, \boldsymbol{G}^{-1/2})$*, then*

$$\|\widehat{\boldsymbol{W}} - \boldsymbol{W}^*\|_F = O_p\left(\left(\tau \vee \tau^2 \delta g\right) \frac{1}{D\gamma} \sqrt{\frac{K-1}{n} \sum_{d=1}^{D} s_d \log[(K-1)p_d]} \right).$$

Remark 3. *If $p_d \geq K$ for all d, then $\log[(K-1)p_d] = \log(K-1) + \log p_d < 2\log p_d$, and the rate could be simplified to*

$$\|\widehat{\boldsymbol{W}} - \boldsymbol{W}^*\|_F = O_p\left(\left(\tau \vee \tau^2\delta g\right)\frac{1}{D\gamma}\sqrt{\frac{K-1}{n}\sum_{d=1}^{D} s_d \log(p_d)}\right).$$

Our results allow both the number of variable p_d and the number of classes K to grow with n. The scaling requirement $s_d^2 \log[(K-1)p_d] = o(n)$ is needed to ensure that restricted eigenvalue condition on $\boldsymbol{G}$ implies restricted eigenvalue condition on random $\boldsymbol{X}'^{\top}\boldsymbol{X}'$ via the infinity norm bound. When $K = 2$, $\widehat{\boldsymbol{W}}$ and $\boldsymbol{W}^*$ are vectors, and this condition can be dropped using the results of Rudelson and Zhou (2013). Nevertheless, the estimation error itself has the same rate as estimation error in linear regression with group-lasso (Lounici et al., 2011; Nardi and Rinaldo, 2008). While our method can be viewed as multi-response linear regression due to formulation (2.8), the group lasso results cannot be directly applied for several reasons. First, both $\boldsymbol{X}'$ and $\boldsymbol{Y}'$ have dependencies across rows and contain fixed blocks of 0 values. Second, the linear model assumption between $\boldsymbol{Y}'$ and $\boldsymbol{X}'$ does not hold. Third, the residuals $\boldsymbol{\Psi} = \boldsymbol{Y}' - \boldsymbol{X}'\boldsymbol{W}^*$ do not have normal distribution and are dependent with $\boldsymbol{X}'$. These challenges required the use of different proof techniques, and the full proof of Theorem 2 can be found in the Supplementary Material.

3.3 Missing data case - semi-supervised learning

In the joint analysis of multi-view data, it is typical to perform complete case analysis, that is only consider the subjects for which all the views are available. This is often not the case in practice. One example is described in Section 3.5.1, where out of 282 subjects with RNAseq data, only 218 have also available miRNA measurements. Moreover, 51 subjects out of these 218 have no class labels, and therefore can not be used to train supervised classification algorithms. Most of the available methods require either imputation of missing views/group labels, or perform complete case analysis (only use samples for which all the views and the group labels are available). A particular advantage of our framework is that we can also use the samples for which we have either a class-label or at least two views available without the need to impute the missing values. In

other words, our proposal allows to perform semi-supervised learning, that is to use information from both labeled and unlabeled subjects to construct classification rules. In what follows, we assume that for each view and each subject, the measurements are rather completely missing, or not missing at all, that is we do not consider the case where a subset of measurements from one view is missing.

Consider an equivalent representation of (2.7) as

$$\underset{\boldsymbol{W}_1,\ldots,\boldsymbol{W}_D}{\text{minimize}} \Big\{ \frac{\alpha}{2nD} \sum_{d=1}^{D} \sum_{i=1}^{n} \|\widetilde{\boldsymbol{y}}_i - \boldsymbol{x}_{id}^\top \boldsymbol{W}_d\|_2^2 + \frac{\alpha}{2nD(D-1)} \sum_{d=1}^{D} \sum_{l=d+1}^{D} \sum_{i=1}^{n} \|\boldsymbol{x}_{id}^\top \boldsymbol{W}_d - \boldsymbol{x}_{il}^\top \boldsymbol{W}_l\|_2^2 + \sum_{d=1}^{D} \lambda_d \operatorname{Pen}(\boldsymbol{W}_d) \Big\}, \tag{3.3}$$

where $\boldsymbol{x}_{id}$ is the ith row of matrix $\boldsymbol{X}_d$. Next, assume some samples have missing views or class labels. Let A_{dy} be the subset of samples (out of n) for which both class label and view d are available, and let B_{dl} be the subset of samples for which both views d and l available. In case there are no missing labels/views, $A_{dy} = B_{dl} = \{1, \ldots, n\}$ for all $d = 1, \ldots, D$, $l = 1, \ldots, D$. Then (3.3) can be rewritten as

$$\underset{\boldsymbol{W}_1,\ldots,\boldsymbol{W}_D}{\text{minimize}} \Big\{ \frac{\alpha}{2nD} \sum_{d=1}^{D} \sum_{i \in A_{dy}} \|\widetilde{\boldsymbol{y}}_i - \boldsymbol{x}_{id}^\top \boldsymbol{W}_d\|_2^2 + \frac{\alpha}{2nD(D-1)} \sum_{d=1}^{D-1} \sum_{l=d+1}^{D} \sum_{i \in B_{dl}} \|\boldsymbol{x}_{id}^\top \boldsymbol{W}_d - \boldsymbol{x}_{il}^\top \boldsymbol{W}_l\|_2^2 + \sum_{d=1}^{D} \lambda_d \operatorname{Pen}(\boldsymbol{W}_d) \Big\}, \tag{3.4}$$

that is we can use all samples with class labels and at least one view for the first part (discriminant analysis), and all samples with at least two views for the second part (canonical correlation analysis). The only samples that we can not use are the ones for which there is no class label and only one view. Like (3.3), problem (3.4) is convex, and can be rewritten as multi-response linear regression problem using augmented data approach similar to Section 2.2.2. This means

that the implementation of Section 2.3 can also be used for problem (3.4). We refer to (3.4) as semi-supervised JACA (ssJACA).

To adopt the proposed cross-validation scheme in Section 2.3.3 for ssJACA, we stratify the samples based on the patterns of "missingness", and split each stratum into F folds. For clarity, we illustrate the case when $D = 3$. Let H be the subset of samples (out of n) with no missing labels/views. Let M_y be the subset of samples for which only class labels are missing, and M_d be the subset of samples for which only view d is missing. Similarly, let M_{dy} be the subset of samples for which only class label and view d are missing, and M_{dl} be the subset of samples for which only views d and l are missing. We first randomly divide each of these strata into F folds: $H^{(f)}, M_y^{(f)}, M_d^{(f)}, M_{dy}^{(f)}$ and $M_{dl}^{(f)}$ where $f = 1, 2, \ldots, F$. For each f, we then hold out the union of $H^{(f)}, M_y^{(f)}, M_d^{(f)}, M_{dy}^{(f)}$ and $M_{dl}^{(f)}$ for testing, and use the remaining samples for training so that the criterion (3.5) can still be applied.

$$CV(\rho, \varepsilon) = \frac{1}{F}\sum_{f=1}^{F}\Bigg\{\alpha\sum_{d=1}^{D}|\text{Cor}(\widetilde{\boldsymbol{Y}}^{(f)}, \boldsymbol{X}_d^{(f)}\widehat{\boldsymbol{W}}_d^{(-f)})| + \frac{(1-\alpha)}{D-1}\sum_{d=1}^{D-1}\sum_{l=d+1}^{D}|\text{Cor}(\boldsymbol{X}_d^{(f)}\boldsymbol{W}_d^{(-f)}, \boldsymbol{X}_l^{(f)}\boldsymbol{W}_l^{(-f)})|\Bigg\}, \tag{3.5}$$

3.4 Simulation studies

We compare the performance of the following methods: (i) JACA: Joint Association and Classification Analysis, the proposed approach; (ii) ssJACA: semi-supervised Joint Association and Classification Analysis. We generate the data using factor model (2.3) as in Section 2.4.1. We compare the methods in terms of misclassification rate, strength of association between the views, estimation consistency and variable selection results.

3.4.1 Two datasets, two groups

We set $n = 200$, $K = 2$, and generate n independent $y \in \{1, 2\}$ with $\pi_1 = 0.4$, and pairs $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ with $(p_1, p_2) \in \{(100, 100), (100, 500), (500, 500)\}$ following Section 2.4.1.

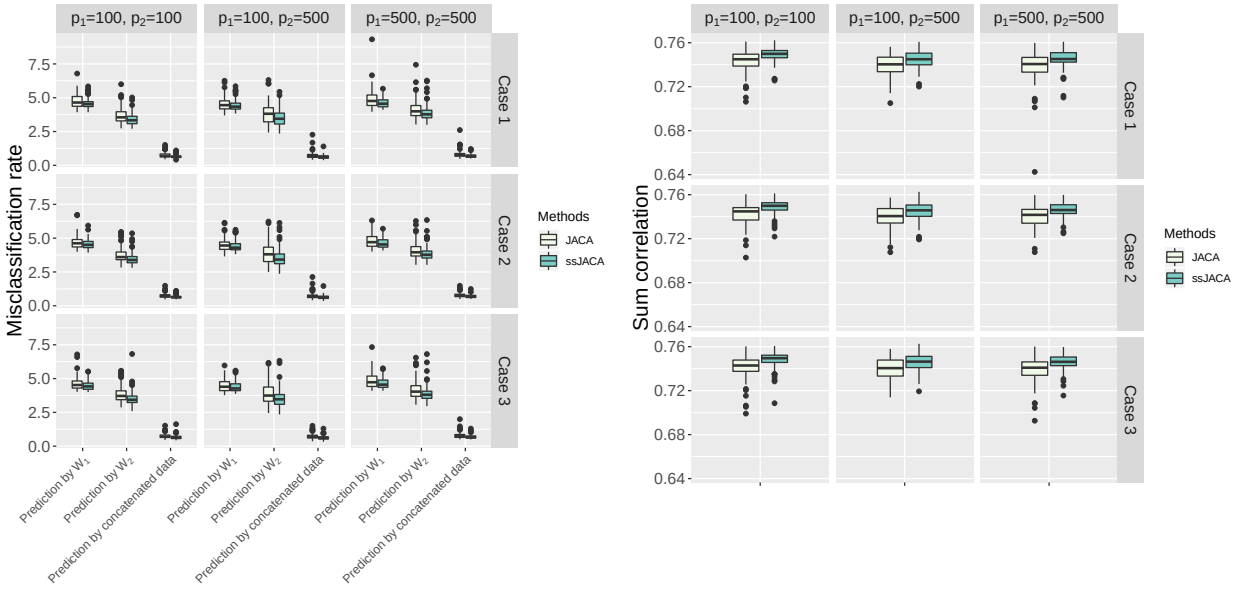

Figure 3.1: Comparison of misclassification rates between JACA and semi-supervised JACA (ssJACA) over 100 replications when $D = 2$, $K = 2$. JACA uses 100 samples with complete view/class information, whereas ssJACA additionally uses 100 samples with missing class information.

Further, we randomly set class information for 100 samples as missing, so that $n_1 = 100$ samples have complete view and class information, whereas the remaining $n_2 = 100$ samples have information on both views but no class assignment. We compare JACA based on $n_1 = 100$ complete samples with ssJACA based on all $n_1 + n_2 = 200$ samples. As before, we generate $10,000$ new samples as test data to evaluate the misclassification rates. The results over 100 replications are displayed in Figure 3.1. In every scenario, ssJACA improves JACA in both miscalssification rates and sum correlation, confirming the advantage of incorporating samples with missing class information in the analysis.

In Figure 3.2 we compare the variable selection performance of JACA with ssJACA. The average performance of both approaches is similar, with ssJACA having lower variability across the replications.

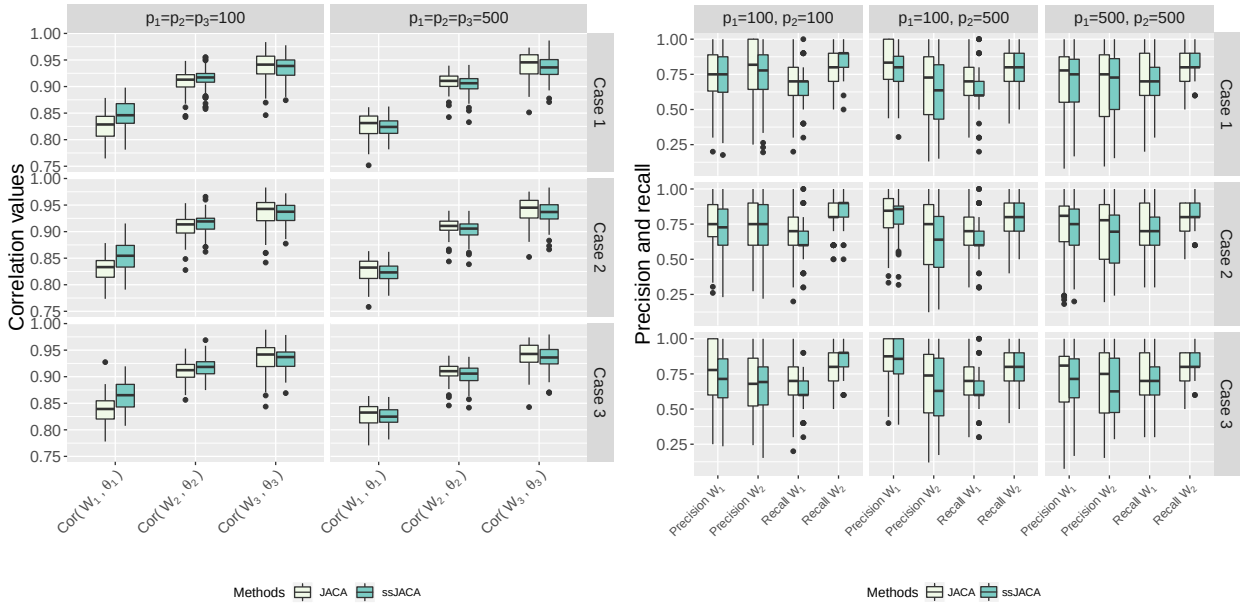

Figure 3.2: Comparison between JACA and semi-supervised JACA (ssJACA) over 100 replications when $D = 2$, $K = 2$. JACA uses $n = 100$ samples with complete view/class information, whereas ssJACA uses extra 100 samples with missing class information. Left: estimation consistency results. Right: variable selection results.

3.4.2 Multiple datasets, multiple groups

We set $n = 200$, $K = 3$, and generate n independent $y \in \{1, 2, 3\}$ with $\pi_1 = 0.4$, $\pi_2 = \pi_3 = 0.3$. We also generate n tuples $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \times \mathbb{R}^{p_3}$ with $p_1 = p_2 = p_3 \in \{100, 500\}$ following Section 2.4.1. We further set class information for 100 samples as missing. We compare JACA based on $n_1 = 100$ complete samples with ssJACA based on all $n_1 + n_2 = 200$ samples, and the misclassification rates are evaluated on $10,000$ test samples as before. The results over 100 replications are displayed in Figure 3.3. When $p_1 = p_2 = p_3 = 100$, ssJACA improves JACA in both misclassification rates and in sum correlation. When $p_1 = p_2 = p_3 = 500$, ssJACA performs slightly better than JACA in misclassification rates, but it obtains higher sum correlation confirming the advantage of incorporating samples with missing class information.

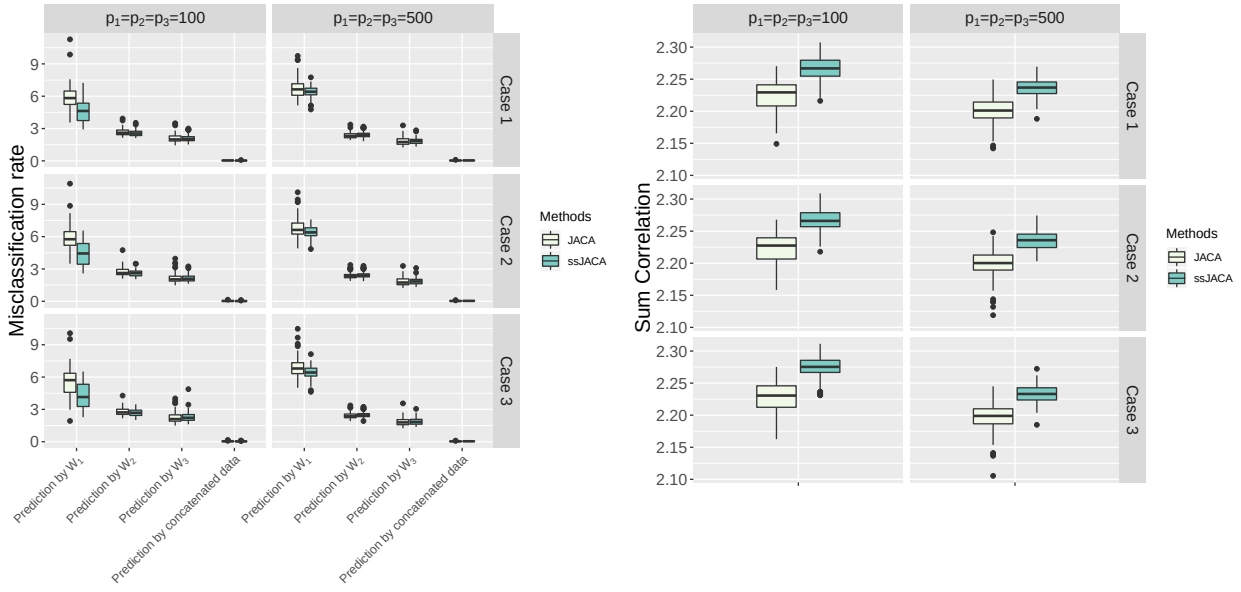

Figure 3.3: Comparison between JACA and semi-supervised JACA (ssJACA) over 100 replications when $D = 3$, $K = 3$. JACA uses 100 samples with complete view/class information, whereas ssJACA additionally uses 100 samples with missing class information.

Table 3.1: Number of available samples in COAD data with different missing patterns of CMS class/RNAseq/miRNA. Complete cases analysis will only be able to use 167 samples, whereas our semi-supervised approach allows to use 245 (all except the last row).

CMS class	RNAseq	miRNA	Sample size
yes	yes	yes	167
yes	yes	no	27
no	yes	yes	51
no	yes	no	37
			Total: 282

3.5 Data analysis

3.5.1 TCGA-COAD dataset

We revisit the colorectal cancer (COAD) data from The Cancer Genome Atlas project. Recall that the data has 282 subjects in total, with Table 3.1 displaying the pattern of available information for each subject.

Table 3.2: Numbers of features selected by JACA and ssJACA on COAD data. JACA is trained using 167 subjects and ssJACA is trained using 245 subjects. The last column corresponds to the number of features shared by both approaches.

	JACA	ssJACA	Intersection
RNA-seq	277	345	227
miRNA	164	188	161

We compare JACA fitted on 167 subjects (all views and subtypes available) with ssJACA fitted on 245 subjects (at least two views available). Both methods achieve the same misclassification rates on 167 subjects. For 27 subjects with missing miRNA data, the subtypes can only be predicted based on RNAseq. JACA has 11.11% misclassification rate on these 27 subjects, whereas ssJACA has 0% misclassification rate. This is perhaps not surprising since these 27 subjects are used by ssJACA for training, however it does show that including additional subjects changes the resulting classification rule. Similarly, for 51 subjects with missing subtype information, the correlation between $\boldsymbol{X}_1^*\widehat{\boldsymbol{W}}_1$ and $\boldsymbol{X}_2^*\widehat{\boldsymbol{W}}_2$ for JACA and ssJACA methods are 0.84 and 0.92 respectively, demonstrating that ssJACA leads to higher associations between the views. Table 3.2 shows the numbers of features selected by both methods. We observe that there is a significant overlap in the selected features, with ssJACA selecting a larger number. Due to the limitations of the data, we can only compare two methods on 245 subjects with at least two views available. However, the simulations in Section 3.4 suggest that incorporating samples with missing information should lead to improved estimation.

The heatmaps of RNAseq and miRNA data with features selected by ssJACA are shown in Figure 3.4. Both views demonstrate different patterns across CMS classes, with the separation on RNASeq being visually much clearer. This is not surprising as CMS classes have been determined based on gene expression data only. Our analysis, however, also allows to determine co-varying patterns in miRNA, with subtype CMS4 being visually the most distinct in that view.

We also consider the visual separation of subtypes based on the projection of RNAseq and miRNA data using discriminant directions found by JACA and ssJACA (Figure 3.5). The triangular

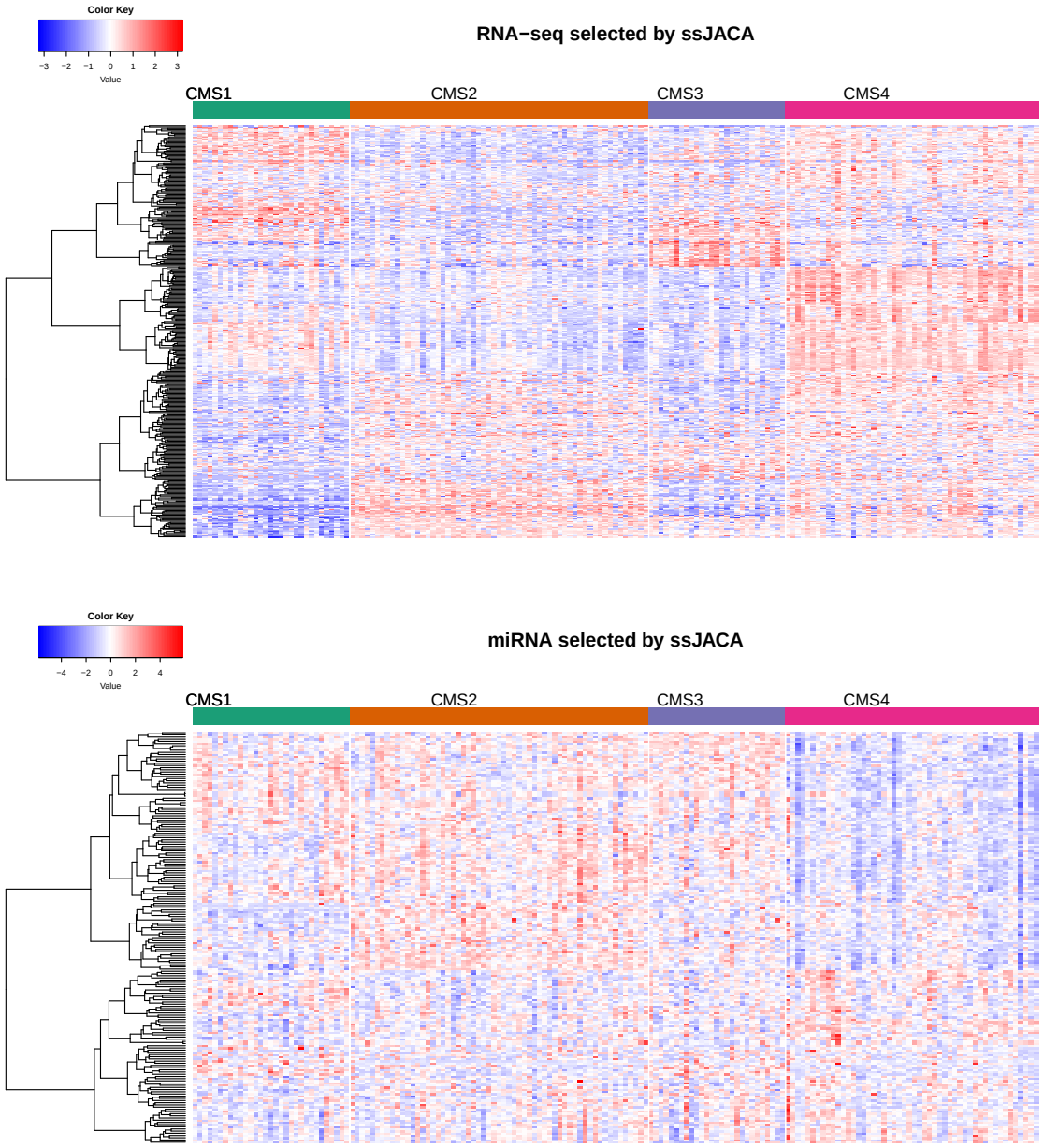

Figure 3.4: Heatmaps of RNAseq and miRNA views from COAD data based on features selected by ssJACA. We use Ward's linkage with euclidean distances for feature ordering.

points in transparent colors indicate 167 subjects with complete view and subtype information. The round points in solid colors are subjects who have missing subtypes, but for whom the subtypes have been previously predicted using random forest classifier (Guinney et al., 2015). We treat these predictions as the gold standard. The square points in solid colors are subjects with no assigned

Table 3.3: Number of samples in BRCA data with different missing patterns of views and cancer subtype. There are only 377 samples with complete information, whereas semi-supervised JACA approach allows to use 708 (all except the last row).

GE	ME	miRNA	RPPA	Cancer type	Count
yes	yes	yes	yes	yes	377
yes	yes	yes	no	yes	114
yes	yes	no	yes	yes	19
yes	yes	no	no	yes	3
yes	no	yes	yes	yes	1
no	yes	yes	yes	no	1
no	yes	yes	no	no	193
no	yes	no	no	no	84
					Total = 792

subtype, which are deemed to have mixed subtype membership (Guinney et al., 2015). The subtype separation is clear based on the projected values, with square points being often in the middle of other subtypes, thus confirming the possibility of mixed subtype membership for those subjects.

3.5.2 TCGA-BRCA dataset

We revisit breast cancer data from Section 2.5.2. Recall that the datasets have 4 views: gene expression (GE), DNA methylation (ME), miRNA expression (miRNA), and reverse phase protein array (RPPA). The samples are separated into 4 breast cancer subtypes: Basal, LumA, LumB and Her2 (The Cancer Genome Atlas Network, 2012). For completeness, we list the number of samples in BRCA data with different missing patterns of views and cancersub type in Table 3.3).

We compare JACA fitted on 377 subjects (all views available) with ssJACA fitted on 708 (at least two views available). In Table 3.4 we compare in-sample misclassification errors based on (i) 377 samples with complete information; and (ii) additional 137 samples for which GE and class information is available, but at least one other view is missing. On 377 samples, ssJACA misclassification rates are better based on GE and ME data, but are somewhat worse for miRNA and RPPA. On 137 samples, ssJACA has a much better performance, likely because ssJACA can incorporate the information from those sample within the estimation procedure. Table 3.5 compares the num-

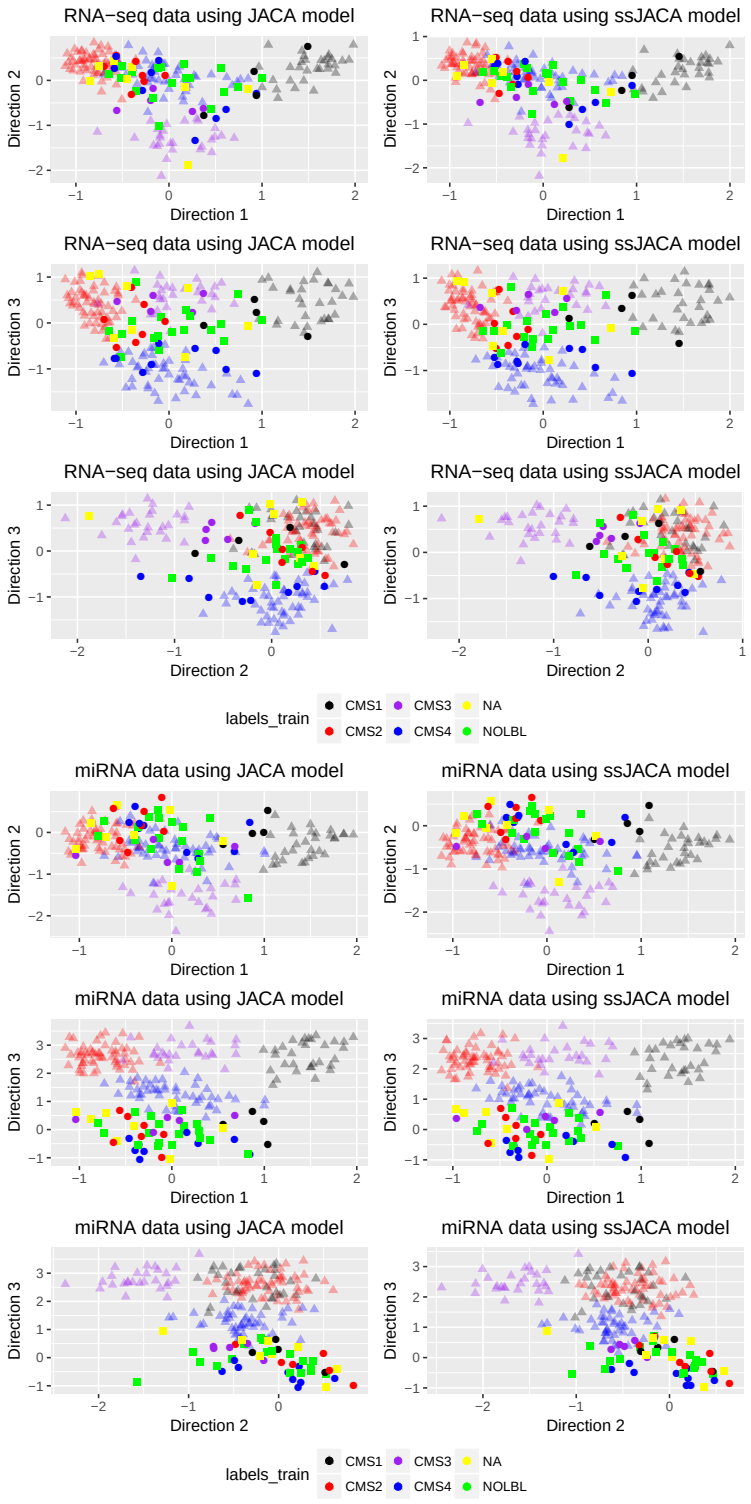

Figure 3.5: Projection of RNAseq and miRNA views from COAD data onto discriminant directions found by JACA and ssJACA.

Table 3.4: Number of misclassified samples on BRCA data. JACA uses 377 subjects and ssJACA is uses 708 subjects. Second to sixth columns correspond to 377 subjects with complete information, whereas the last column corresponds to 137 subject with GE and subtype information, but at least one other view missing.

	out of 377 subjects					out of 137 subjects
Method	GE	ME	miRNA	RPPA	All	GE
JACA	23	40	**38**	**65**	34	13
ssJACA	**19**	**38**	39	72	**33**	**6**

Table 3.5: Cardinality comparison of JACA and ssJACA on BRCA data. JACA is trained using 377 subjects and ssJACA is trained using 708 subjects. The third row is the numbers of features shared by both methods.

	GE	ME	miRNA	RPPA
JACA	465	393	299	129
ssJACA	579	457	318	135
Intersection	394	371	277	128

ber of features selected by both methods. ssJACA tends to select more variables than JACA, with a significant overlap between the two.

3.6 Technical Proofs

3.6.1 Proof of Lemma 1

Proof. By multiplying the covariance matrix $\boldsymbol{G}$ on both sides of $\boldsymbol{W}_d^* \boldsymbol{R}_d^\top \propto \boldsymbol{\Sigma}_d^{-1} \boldsymbol{\Delta}_d$, it remains to show that for some orthogonal matrices $\boldsymbol{R}_d$, $\widetilde{\boldsymbol{\Delta}} \text{diag}(\boldsymbol{R}_1 \cdots, \boldsymbol{R}_D)^\top \propto \boldsymbol{G} \text{diag}(\boldsymbol{\Sigma})^{-1} \boldsymbol{\Delta}$, where

$\text{diag}(\boldsymbol{\Sigma})^{-1} = \text{diag}(\boldsymbol{\Sigma}_1^{-1}, \cdots, \boldsymbol{\Sigma}_D^{-1})$. Expending the right hand side leads to

$$\begin{aligned}\boldsymbol{G}\text{diag}(\boldsymbol{\Sigma})^{-1}\boldsymbol{\Delta} &= \begin{pmatrix} \boldsymbol{I} & -\frac{1-\alpha}{D}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_2^{-1} & \cdots & -\frac{1-\alpha}{D}\boldsymbol{\Sigma}_{1D}\boldsymbol{\Sigma}_D^{-1} \\ & \vdots & & \\ -\frac{1-\alpha}{D}\boldsymbol{\Sigma}_{D1}\boldsymbol{\Sigma}_1^{-1} & -\frac{1-\alpha}{D}\boldsymbol{\Sigma}_{D2}\boldsymbol{\Sigma}_2^{-1} & \cdots & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Delta}_1 \\ \vdots \\ \boldsymbol{\Delta}_D \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\Delta}_1 - \frac{1-\alpha}{D}\sum_{d\neq 1}\boldsymbol{\Sigma}_{1d}\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\Delta}_d \\ \vdots \\ \boldsymbol{\Delta}_D - \frac{1-\alpha}{D}\sum_{d\neq D}\boldsymbol{\Sigma}_{Dd}\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\Delta}_d \end{pmatrix}.\end{aligned}$$

From the factor model decomposition (2.3), $\boldsymbol{A}_d^\top \boldsymbol{\Sigma}_d^{-1}\boldsymbol{\Delta}_d = 0$ holds, and hence

$$\boldsymbol{\Sigma}_{ld}\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\Delta}_d = \boldsymbol{\Delta}_l\boldsymbol{\Delta}_d^\top\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\Delta}_d = \boldsymbol{\Delta}_l\boldsymbol{\Lambda}_d(\boldsymbol{\Lambda}_d + \boldsymbol{I})^{-1},$$

where $\boldsymbol{\Delta}_d^\top\boldsymbol{\Sigma}_{dy}^{-1}\boldsymbol{\Delta}_d = \boldsymbol{\Lambda}_d$. It follows that

$$\boldsymbol{\Delta}_l - \frac{1-\alpha}{D}\sum_{d\neq l}\boldsymbol{\Sigma}_{ld}\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\Delta}_d = \boldsymbol{\Delta}_l - \frac{1-\alpha}{D}\sum_{d\neq l}\boldsymbol{\Delta}_l\boldsymbol{\Lambda}_d(\boldsymbol{\Lambda}_d + \boldsymbol{I})^{-1} \propto \boldsymbol{\Delta}_l.$$

Choosing $\boldsymbol{R}_d$ as an orthogonal matrix such that $\widetilde{\boldsymbol{\Delta}}_d\boldsymbol{R}_d^\top = \boldsymbol{\Delta}_d$ completes the proof. □

3.6.2 Proof of Theorem 2

Proof. Consider the concatenated $\widetilde{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 & \cdots & \boldsymbol{X}_D \end{pmatrix}$. From Lemmas 3 and 7 in Gaynanova (2019), with probability at least $1-\eta$ and some constant C

$$\|\frac{1}{n}\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{X}} - \boldsymbol{\Sigma}_T\|_\infty \leq C\tau^2\sqrt{\frac{\log(\sum_{i=1}^D p_i\eta^{-1})}{n}},$$

where $\tau = \max_j \sqrt{\sigma_j^2 + \max_k \mu_{kj}^2}$, σ_j are diagonal elements of $\boldsymbol{\Sigma}_y$ and $\mu_{k,j}$ are elements of $\boldsymbol{\mu}_k$. Therefore, with probability at least $1 - \eta$

$$\|\boldsymbol{G} - \boldsymbol{X'}^\top \boldsymbol{X'}\|_\infty \leq \frac{1}{D} \|\frac{1}{n} \widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}} - \boldsymbol{\Sigma}_T\|_\infty \leq \frac{C\tau^2}{D} \sqrt{\frac{\log(\sum_{i=1}^D p_i \eta^{-1})}{n}}.$$

From Lemma 5, if $s_d \leq \gamma \lambda_{min}^2 (32 D \lambda_d^2 \|\boldsymbol{G} - \boldsymbol{X'}^\top \boldsymbol{X'}\|_\infty)^{-1}$, then $\boldsymbol{X'}$ satisfies $RE(S, 3, \boldsymbol{\lambda})$ and $\gamma \leq 2\gamma$. Hence, using $\lambda_d = C\left(\tau \vee \tau^2 \delta g\right) D^{-1} \sqrt{(K-1)\log[(K-1)p_d]/n}$, Assumption 6 and the condition $s_d^2 \log[(K-1)p_d] = o(n)$ leads to $s_d \leq \gamma \lambda_{min}^2 (32 D \lambda_d^2 \|\boldsymbol{G} - \boldsymbol{X'}^\top \boldsymbol{X'}\|_\infty)^{-1}$. Therefore, by Theorems 3 and 4

$$\|\widehat{\boldsymbol{W}} - \boldsymbol{W}^*\|_F = O_p \left(\left(\tau \vee \tau^2 \delta g\right) \frac{1}{D\gamma} \sqrt{\frac{K-1}{n} \sum_{d=1}^D s_d \log[(K-1)p_d]} \right).$$

□

Proof of Proposition 3. By the KKT conditions (2.10), $\boldsymbol{W}_d = 0$ leads to $\boldsymbol{X'}_{dj}^\top \boldsymbol{Y'} = \lambda \boldsymbol{u}_{dj}$, hence by the definition of subgradient $\left\|\boldsymbol{X'}_{dj}^\top \boldsymbol{Y'}\right\|_2 = \left\| \left(\frac{\alpha \boldsymbol{X}_d^\top \widetilde{\boldsymbol{Y}}}{nD} \right)_j \right\|_2 = \lambda \|\boldsymbol{u}_{dj}\|_2 \leq \lambda$. This implies that $\boldsymbol{W}_d = 0$ satisfies KKT conditions whenever $\lambda \geq \alpha (nD)^{-1} \|\boldsymbol{X}_d^\top \widetilde{\boldsymbol{Y}}\|_{\infty,2}$. □

3.6.3 Supporting Theorems and Lemmas

Lemma 2. *Let* $\boldsymbol{\phi}_d = \frac{\alpha}{nD} \boldsymbol{X}_d^\top \left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}_d \boldsymbol{W}_d^* \right) + \frac{1-\alpha}{nD(D-1)} \sum_{j \neq d} \left(\boldsymbol{X}_j \boldsymbol{W}_j^* - \boldsymbol{X}_d \boldsymbol{W}_d^* \right)$*, and let* $\widehat{\boldsymbol{W}}$ *be the solution to* (2.8) *with* $\lambda_d \geq 2 \|\boldsymbol{X}_d^\top \boldsymbol{\phi}_d\|_{\infty,2}$*. Let* $\boldsymbol{H} = \widehat{\boldsymbol{W}} - \boldsymbol{W}^*$*, and* S *as defined in Assumption 3, then* $\boldsymbol{H} \in C(S, \lambda)$*.*

Proof. Consider the KKT conditions for (2.8)

$$0 = -\boldsymbol{X'}^\top (\boldsymbol{Y'} - \boldsymbol{X'} \widehat{\boldsymbol{W}}) + \widehat{\boldsymbol{s}},$$

where $\widehat{\boldsymbol{s}}_{dj} \in \partial(\lambda_d \|\boldsymbol{w}_{dj}\|_2)$ evaluated at $\widehat{\boldsymbol{W}}$. Multiplying $(\boldsymbol{W}^* - \widehat{\boldsymbol{W}})^\top$ on both sides gives

$$(\boldsymbol{W}^* - \widehat{\boldsymbol{W}})^\top \left(\boldsymbol{X'}^\top (\boldsymbol{Y'} - \boldsymbol{X'} \widehat{\boldsymbol{W}}) - \widehat{\boldsymbol{s}} \right) = 0.$$

Let $\boldsymbol{\Psi} = \boldsymbol{Y}' - \boldsymbol{X}'\boldsymbol{W}^*$. Replacing $\boldsymbol{Y}'$ with $\boldsymbol{Y}' + \boldsymbol{X}'\boldsymbol{W}^* - \boldsymbol{X}'\boldsymbol{W}^*$ and using properties of subgradient of convex functions leads to

$$\|\boldsymbol{X}'(\boldsymbol{W}^* - \widehat{\boldsymbol{W}})\|_F^2 \leq \langle \boldsymbol{X}'^{\top}\boldsymbol{\Psi}, \widehat{\boldsymbol{W}} - \boldsymbol{W}^* \rangle + \sum_{d=1}^{D} \lambda_d \|\boldsymbol{W}_d^*\|_{1,2} - \sum_{d=1}^{D} \lambda_d \|\widehat{\boldsymbol{W}}_d\|_{1,2}.$$

Since $\langle \boldsymbol{X}'^{\top}\boldsymbol{\Psi}, \widehat{\boldsymbol{W}} - \boldsymbol{W}^* \rangle = \sum_{d=1}^{D} \langle \boldsymbol{X}_d^{\top}\boldsymbol{\phi}_d, \widehat{\boldsymbol{W}}_d - \widehat{\boldsymbol{W}}_d^* \rangle$, applying Hölder inequality twice and using conditions on λ_d leads to

$$\begin{aligned}
\|\boldsymbol{X}'(\boldsymbol{W}^* - \widehat{\boldsymbol{W}})\|_F^2 &\leq \sum_{d=1}^{D} \langle \boldsymbol{X}_d^{\top}\boldsymbol{\phi}_d, \widehat{\boldsymbol{W}}_d - \boldsymbol{W}_d^* \rangle + \sum_{d=1}^{D} \lambda_d \|\boldsymbol{W}_d^*\|_{1,2} - \sum_{d=1}^{D} \lambda_d \|\widehat{\boldsymbol{W}}_d\|_{1,2} \\
&\leq \sum_{d=1}^{D} \|\boldsymbol{X}_d^{\top}\boldsymbol{\phi}_d\|_{\infty,2} \|\widehat{\boldsymbol{W}}_d - \boldsymbol{W}_d^*\|_{1,2} + \sum_{d=1}^{D} \lambda_d \|\boldsymbol{W}_d^*\|_{1,2} - \sum_{d=1}^{D} \lambda_d \|\widehat{\boldsymbol{W}}_d\|_{1,2} \\
&\leq \sum_{d=1}^{D} \frac{\lambda_d}{2} \left(\|\boldsymbol{H}_{d,S_d}\|_{1,2} + \|\boldsymbol{H}_{d,S_d^c}\|_{1,2} \right) + \sum_{d=1}^{D} \lambda_d \|\boldsymbol{W}_d^*\|_{1,2} - \sum_{d=1}^{D} \lambda_d \|\widehat{\boldsymbol{W}}_d\|_{1,2}.
\end{aligned}$$

Since

$$\begin{aligned}
\|\widehat{\boldsymbol{W}}_d\|_{1,2} &= \|\boldsymbol{W}_d^* + \widehat{\boldsymbol{W}}_d - \boldsymbol{W}_d^*\|_{1,2} = \|\boldsymbol{W}_{d,S_d}^* + \boldsymbol{H}_{d,S_d}\|_{1,2} + \|\boldsymbol{H}_{d,S_d^c}\|_{1,2} \\
&\geq \|\boldsymbol{W}_{d,S_d}^*\|_{1,2} - \|\boldsymbol{H}_{d,S_d}\|_{1,2} + \|\boldsymbol{H}_{d,S_d^c}\|_{1,2},
\end{aligned}$$

combining the above two displays gives

$$\|\boldsymbol{X}'(\boldsymbol{W}^* - \widehat{\boldsymbol{W}})\|_F^2 \leq \sum_{d=1}^{D} \frac{3}{2} \lambda_d \|\boldsymbol{H}_{d,S_d}\|_{1,2} - \sum_{d=1}^{D} \frac{1}{2} \lambda_d \|\boldsymbol{H}_{d,S_d^c}\|_{1,2}. \tag{3.6}$$

Since $\|\boldsymbol{X}'(\boldsymbol{W}^* - \widehat{\boldsymbol{W}})\|_F^2 \geq 0$, the statement follows. □

Theorem 3. *Let $\widehat{\boldsymbol{W}}$ be the solution to* (2.8) *with $\lambda_d \geq 2\|\boldsymbol{X}_d^{\top}\boldsymbol{\phi}_d\|_{\infty,2}$, where $\boldsymbol{\phi}_d$ are defined in*

Lemma 2. Under Assumption 3, if $\boldsymbol{X}'$ satisfies RE$(S, \boldsymbol{\lambda})$ with $\gamma = \gamma(S, \boldsymbol{\lambda}, \boldsymbol{X}')$, then

$$\|\widehat{\boldsymbol{W}} - \boldsymbol{W}^*\|_F \leq \frac{3}{2\gamma}\sqrt{\sum_{d=1}^{D} \lambda_d^2 s_d}.$$

Proof. From equation (3.6), using $\boldsymbol{H} = \widehat{\boldsymbol{W}} - \boldsymbol{W}^*$,

$$\|\boldsymbol{X}'(\boldsymbol{W}^* - \widehat{\boldsymbol{W}})\|_F^2 \leq \sum_{d=1}^{D} \frac{3\lambda_d}{2}\|\boldsymbol{W}^*_{d,S_d} - \widehat{\boldsymbol{W}}_{d,S_d}\|_{1,2} \leq \frac{3}{2}\sum_{d=1}^{D} \lambda_d\sqrt{s_d}\|\boldsymbol{W}^*_{d,S_d} - \widehat{\boldsymbol{W}}_{d,S_d}\|_F.$$

Applying Cauchy-Schwartz inequality gives

$$\|\boldsymbol{X}'(\boldsymbol{W}^* - \widehat{\boldsymbol{W}})\|_F^2 \leq \frac{3}{2}\sqrt{\sum_{d=1}^{D} \lambda_d^2 s_d}\|\boldsymbol{W}^*_S - \widehat{\boldsymbol{W}}_S\|_F.$$

Since $\boldsymbol{X}'$ satisfies RE$(S, \boldsymbol{\lambda})$ and $\boldsymbol{H} \in C(S, \boldsymbol{\lambda})$, by Lemma 2

$$\begin{aligned}\|\boldsymbol{W}^* - \widehat{\boldsymbol{W}}\|_F^2 &\leq \frac{1}{\gamma}\|\boldsymbol{X}'(\boldsymbol{W}^* - \widehat{\boldsymbol{W}})\|_2^2 \leq \frac{1}{\gamma}\frac{3}{2}\sqrt{\sum_{d=1}^{D} \lambda_d^2 s_d}\|\boldsymbol{W}^*_S - \widehat{\boldsymbol{W}}_S\|_F \\ &\leq \frac{1}{\gamma}\frac{3}{2}\sqrt{\sum_{d=1}^{D} \lambda_d^2 s_d}\|\boldsymbol{W}^* - \widehat{\boldsymbol{W}}\|_F.\end{aligned}$$

If $\|\boldsymbol{W}^* - \widehat{\boldsymbol{W}}\|_F^2 = 0$, the bound holds trivially. Otherwise, dividing by $\|\boldsymbol{W}^* - \widehat{\boldsymbol{W}}\|_F$ on both sides leads to the desired bound. □

Theorem 4. *Under Assumptions 4–6, there exists $C > 0$ such that*

$$\|\boldsymbol{X}_d^\top \boldsymbol{\phi}_d\|_{\infty,2} \leq C\left(\tau \vee \tau^2\delta g\right)\frac{1}{D}\sqrt{\frac{(K-1)\log((K-1)p_d\eta^{-1})}{n}}, \quad d = 1, \ldots, D,$$

with probability at least $1 - \eta$, where $\boldsymbol{\phi}_d$ are from Lemma 2.

Proof. Without loss of generality, consider $d = 1$ and let $\widetilde{\boldsymbol{\Delta}} = \begin{pmatrix}\widetilde{\boldsymbol{\Delta}}_1^\top & \widetilde{\boldsymbol{\Delta}}_2^\top & \cdots & \widetilde{\boldsymbol{\Delta}}_D^\top\end{pmatrix}^\top$, where

$\widetilde{\boldsymbol{\Delta}}_d \in \mathbb{R}^{p_d \times K-1}$. Applying the triangle inequality gives

$$
\begin{aligned}
&\|\boldsymbol{X}_1^\top \boldsymbol{\phi}_1\|_{\infty,2} \\
&= \|\frac{\alpha}{nD}\boldsymbol{X}_1^\top(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}_1\boldsymbol{W}_1^*) + \frac{1-\alpha}{nD(D-1)}\sum_{l\neq 1}\boldsymbol{X}_1^\top(\boldsymbol{X}_l\boldsymbol{W}_l^* - \boldsymbol{X}_1\boldsymbol{W}_1^*)\|_{\infty,2} \\
&= \|\frac{\alpha}{nD}\boldsymbol{X}_1^\top(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}_1\boldsymbol{W}_1^*) - \widetilde{\boldsymbol{\Delta}}_1 + \widetilde{\boldsymbol{\Delta}}_1 + \frac{1-\alpha}{nD(D-1)}\sum_{l\neq 1}\boldsymbol{X}_1^\top(\boldsymbol{X}_l\boldsymbol{W}_l^* - \boldsymbol{X}_1\boldsymbol{W}_1^*)\|_{\infty,2} \\
&\leq \underbrace{\|\frac{\alpha}{nD}\boldsymbol{X}_1^\top\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{\Delta}}_1\|_{\infty,2}}_{:=I_1} \\
&\quad + \underbrace{\|\widetilde{\boldsymbol{\Delta}}_1 - \frac{\alpha}{nD}\boldsymbol{X}_1^\top\boldsymbol{X}_1\boldsymbol{W}_1^* + \frac{1-\alpha}{nD(D-1)}\sum_{l\neq 1}\boldsymbol{X}_1^\top(\boldsymbol{X}_l\boldsymbol{W}_l^* - \boldsymbol{X}_1\boldsymbol{W}_1^*)\|_{\infty,2}}_{:=I_2}.
\end{aligned}
$$

Consider I_1. From Lemma 4 in Gaynanova (2019), there exists $C_1 > 0$ such that

$$
\|\frac{\alpha}{nD}\boldsymbol{X}_1^\top\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{\Delta}}_1\|_{\infty,2} \leq \frac{C_1}{D}\max_j \sigma_{1,j}\sqrt{\frac{(K-1)\log(p_1\eta^{-1})}{n}} \leq \frac{C_1}{D}\tau\sqrt{\frac{(K-1)\log(p_1\eta^{-1})}{n}}
$$

with probability at least $1-\eta$.

Consider I_2.

$$
\begin{aligned}
I_2 &= \Big\|\widetilde{\boldsymbol{\Delta}}_1 - \frac{1}{n}\boldsymbol{X}_1^\top\Big\{\alpha\frac{1}{D}\boldsymbol{X}_1\boldsymbol{W}_1^* + \frac{1-\alpha}{D}\boldsymbol{X}_1\boldsymbol{W}_1^* - \frac{1-\alpha}{D(D-1)}\sum_{l\neq 1}\boldsymbol{X}_l\boldsymbol{W}_l^*\Big\}\Big\|_{\infty,2} \\
&= \Big\|\widetilde{\boldsymbol{\Delta}}_1 - \frac{1}{Dn}\boldsymbol{X}_1^\top\Big\{\boldsymbol{X}_1\boldsymbol{W}_1^* - \frac{1-\alpha}{D-1}\sum_{l\neq 1}\boldsymbol{X}_l\boldsymbol{W}_l^*\Big\}\Big\|_{\infty,2} \\
&= \|\widetilde{\boldsymbol{\Delta}}_1 - \frac{1}{Dn}\boldsymbol{X}_1^\top\boldsymbol{U}\|_{\infty,2},
\end{aligned}
$$

where $\boldsymbol{U} = \boldsymbol{X}_1\boldsymbol{W}_1^* - \frac{1-\alpha}{D-1}\sum_{l\neq 1}\boldsymbol{X}_l\boldsymbol{W}_l^* \in \mathbb{R}^{n\times(K-1)}$. Since the first p_1 rows of $\boldsymbol{G}$ are

$$
\begin{pmatrix}\boldsymbol{\Sigma}_1 & -\frac{1-\alpha}{D-1}\boldsymbol{\Sigma}_{12} & \cdots & -\frac{1-\alpha}{D-1}\boldsymbol{\Sigma}_{1D}\end{pmatrix}/D,
$$

we have

$$\begin{aligned}\mathbb{E}\left(\frac{1}{Dn}\boldsymbol{X}_1^\top \boldsymbol{U}\right) &= \frac{1}{D}\mathbb{E}\left(\frac{1}{n}\boldsymbol{X}_1^\top \begin{pmatrix}\boldsymbol{X}_1 & -\frac{1-\alpha}{D-1}\boldsymbol{X}_2 & \cdots & -\frac{1-\alpha}{D-1}\boldsymbol{X}_D\end{pmatrix}\boldsymbol{W}^*\right) \\ &= \frac{1}{D}\begin{pmatrix}\boldsymbol{\Sigma}_1 & -\frac{1-\alpha}{D-1}\boldsymbol{\Sigma}_{12} & \cdots & -\frac{1-\alpha}{D-1}\boldsymbol{\Sigma}_{1D}\end{pmatrix}\boldsymbol{G}^{-1}\widetilde{\boldsymbol{\Delta}} \\ &= \begin{pmatrix}I_{p_1} & \boldsymbol{0}\end{pmatrix}\widetilde{\boldsymbol{\Delta}} = \widetilde{\boldsymbol{\Delta}}_1.\end{aligned}$$

Combining the above gives

$$\begin{aligned}I_2 &= \left\|\widetilde{\boldsymbol{\Delta}}_1 - \frac{1}{Dn}\boldsymbol{X}_1^\top \boldsymbol{U}\right\|_{\infty,2} \leq \sqrt{K-1}\left\|\widetilde{\boldsymbol{\Delta}}_1 - \frac{1}{Dn}\boldsymbol{X}_1^\top \boldsymbol{U}\right\|_{\infty} \\ &= \sqrt{K-1}\left\|\mathbb{E}\left(\frac{1}{Dn}\boldsymbol{X}_1^\top \boldsymbol{U}\right) - \frac{1}{Dn}\boldsymbol{X}_1^\top \boldsymbol{U}\right\|_{\infty}.\end{aligned}$$

From Lemma 3 in Gaynanova (2019), all elements of $\boldsymbol{X}_1$ are subgaussian with parameter at most τ. From Lemma 3, all elements of $\boldsymbol{U}$ are subgaussian with parameter at most $2\tau\delta g$. Therefore, by Lemma 4, there exist $C_2 > 0$ such that with probability at least $1-\eta$

$$I_2 \leq C_2 \frac{\tau^2 \delta g}{D}\sqrt{\frac{(K-1)\log((K-1)p_1\eta^{-1})}{n}}.$$

Combining the results for I_1 and I_2 leads to the desired bound. □

Lemma 3. *Under Assumptions 4–5, all elements of* $\boldsymbol{U}_d = \boldsymbol{X}_d\boldsymbol{W}_d^* - \frac{1-\alpha}{D-1}\sum_{l\neq d}\boldsymbol{X}_l\boldsymbol{W}_l^*$, $d = 1,\ldots,D$, *are subgaussian with parameter* $2\tau\delta g$.

Proof. Without loss of generality, let $d = 1$ and $\boldsymbol{V} = \begin{pmatrix}\boldsymbol{X}_1 & -\frac{1-\alpha}{D-1}\boldsymbol{X}_2 & \cdots & -\frac{1-\alpha}{D-1}\boldsymbol{X}_D\end{pmatrix} \in \mathbb{R}^{n\times\sum_{i=1}^D p_i}$ so that $\boldsymbol{U}_1 = \boldsymbol{U} = \boldsymbol{V}\boldsymbol{W}^*$. Let $\boldsymbol{v}_i$ be the i^{th} row of $\boldsymbol{V}$. Under Assumptions 4–5, $\boldsymbol{v}_i|\boldsymbol{y}_i = k$ follows normal distribution with

$$\mathbb{E}\left[\boldsymbol{v}_i\middle|\boldsymbol{y}_i = k\right] = \boldsymbol{P}\boldsymbol{\mu}_k, \mathrm{Cov}\left[\boldsymbol{v}_i\middle|\boldsymbol{y}_i = k\right] = \boldsymbol{P}\boldsymbol{\Sigma}_y\boldsymbol{P} = \bar{\boldsymbol{\Sigma}}_y,$$

where $\boldsymbol{P} = \text{diag}(\boldsymbol{I}_{p_1}, -\frac{1-\alpha}{D-1}\boldsymbol{I}_{p_2}, \ldots, -\frac{1-\alpha}{D-1}\boldsymbol{I}_{p_D})$. Therefore,

$$\begin{aligned}
\boldsymbol{W}^{*\top}\boldsymbol{v}_i &= \widetilde{\boldsymbol{\Delta}}^\top \boldsymbol{G}^{-1}\boldsymbol{v}_i \\
&= \widetilde{\boldsymbol{\Delta}}^\top \boldsymbol{G}^{-1}(\boldsymbol{P}\sum_{k=1}^{K}\boldsymbol{\mu}_k \mathbb{1}\{y_i = k\} + \bar{\boldsymbol{\Sigma}}_y^{1/2}\boldsymbol{e}_i) \\
&= \widetilde{\boldsymbol{\Delta}}^\top \boldsymbol{G}^{-1}\boldsymbol{P}\sum_{k=1}^{K}\boldsymbol{\mu}_k \mathbb{1}\{y_i = k\} + \widetilde{\boldsymbol{\Delta}}^\top \boldsymbol{G}^{-1}\bar{\boldsymbol{\Sigma}}_y^{1/2}\boldsymbol{e}_i \\
&:= \boldsymbol{v}_{1i} + \boldsymbol{v}_{2i},
\end{aligned}$$

where $\boldsymbol{e}_i \sim \mathcal{N}(\boldsymbol{I})$ and $\boldsymbol{v}_{1i}$, $\boldsymbol{v}_{2i}$ are independent random vectors.

Let $\boldsymbol{M} = (\boldsymbol{\mu}_1\ \boldsymbol{\mu}_2\ \cdots\ \boldsymbol{\mu}_K) \in \mathbb{R}^{\sum_{i=1}^{D} p_i \times K}$. Since $\|\boldsymbol{G}^{-1}\|_\infty \leq g$,

$$\begin{aligned}
\|\boldsymbol{v}_{1i}\|_\infty = \|\widetilde{\boldsymbol{\Delta}}^\top \boldsymbol{G}^{-1}\boldsymbol{P}\sum_{k=1}^{K}\boldsymbol{\mu}_k \mathbb{1}\{y_i = k\}\|_\infty &\leq \|\widetilde{\boldsymbol{\Delta}}^\top \boldsymbol{G}^{-1}\boldsymbol{P}\boldsymbol{M}\|_{\infty,2} \\
\leq \|\widetilde{\boldsymbol{\Delta}}\|_{\infty,2}\|\boldsymbol{G}^{-1}\|_\infty\|\boldsymbol{P}\boldsymbol{M}\|_\infty &\leq \delta\tau g,
\end{aligned}$$

where the second inequality holds because of $\|\boldsymbol{A}\boldsymbol{B}\|_{\infty,2} \leq \|\boldsymbol{A}\|_\infty\|\boldsymbol{B}\|_{\infty,2}$ (Obozinski et al., 2011, Lemma 8). Hence all elements of $\boldsymbol{v}_{1i}$ are subgaussian with parameter at most $\delta\tau g$.

On the other hand, $\boldsymbol{v}_{2i}$ is a normally distributed vector with mean $\mathbf{0}$ and covariance $\text{Cov}(\boldsymbol{v}_{2i}) = \widetilde{\boldsymbol{\Delta}}^\top \boldsymbol{G}^{-1}\bar{\boldsymbol{\Sigma}}_y\boldsymbol{G}^{-1}\widetilde{\boldsymbol{\Delta}}$. Since

$$\begin{aligned}
\|\,\text{Cov}(\boldsymbol{v}_{2i})\|_\infty &= \|\widetilde{\boldsymbol{\Delta}}^\top \boldsymbol{G}^{-1}\bar{\boldsymbol{\Sigma}}_y\boldsymbol{G}^{-1}\widetilde{\boldsymbol{\Delta}}\|_\infty \\
&\leq \|\widetilde{\boldsymbol{\Delta}}\|_\infty^2\|\boldsymbol{G}^{-1}\|_\infty^2\|\boldsymbol{\Sigma}_y\|_\infty\|\boldsymbol{P}\|_\infty^2 \\
&\leq \|\widetilde{\boldsymbol{\Delta}}\|_{\infty,2}^2\|\boldsymbol{G}^{-1}\|_\infty^2\|\boldsymbol{\Sigma}_y\|_\infty \leq \delta^2\tau^2 g^2,
\end{aligned}$$

all elements of $\boldsymbol{v}_{2i}$ are also subgaussian with parameter $\delta\tau g$.

Combining the results for $\boldsymbol{v}_{1i}$ and $\boldsymbol{v}_{2i}$,

$$\mathbb{E}(e^{\lambda u_{ij}}) = \mathbb{E}\{e^{\lambda(v_{1ij}+v_{2ij})}\} = \mathbb{E}(e^{\lambda v_{1ij}})\mathbb{E}(e^{\lambda v_{2ij}}) \leq e^{\lambda^2\{2\tau\delta g\}/2}.$$

This implies that all elements of $\boldsymbol{U}_1$ are subgaussian with parameter $2\tau\delta g$. □

Lemma 4. *Let $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathbb{R}^p \times \mathbb{R}^q$ be independent identically distributed pairs of mean zero random vectors with $\mathbb{E}(\boldsymbol{x}_i \boldsymbol{y}_i^\top) = \boldsymbol{\Sigma}_{xy}$, and let all elements of $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ be sub-gaussian with parameters τ_1 and τ_2, respectively. Let $\boldsymbol{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_n]^\top$, $\boldsymbol{Y} = [\boldsymbol{y}_1 \dots \boldsymbol{y}_n]^\top$ If $\log(pq) = o(n)$, then with probability at least $1 - \eta$ for some constant $C > 0$*

$$\left\| \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{Y} - \boldsymbol{\Sigma}_{xy} \right\|_\infty \leq C \tau_1 \tau_2 \sqrt{\frac{\log(pq/\eta)}{n}}.$$

Proof. Let $u_{ikj} = x_{ji} y_{jk}$, then u_{ikj} is sub-exponential with parameter $2\tau_1\tau_2$ (Vershynin, 2012, Lemma 5.14). Let σ_{ik} be elements of $\boldsymbol{\Sigma}_{xy}$, then $u_{ikj} - \sigma_{ik}$ are sub-exponential with parameter $4\tau_1\tau_2$, and using Bernstein's bound (Vershynin, 2012, Proposition 5.16)

$$\operatorname{pr}\left(\left\| \frac{1}{n} \sum_{j=1}^{n} u_{ikj} - \sigma_{ik} \right\|_\infty \geq \varepsilon \right) \leq 2 \exp\left\{ - C \min\left(\frac{\varepsilon^2}{16 \tau_1^2 \tau_2^2}, \frac{\varepsilon}{4\tau_1\tau_2} \right) n \right\}$$

for some $C > 0$. By union bound

$$\operatorname{pr}(\|\boldsymbol{X}^\top \boldsymbol{Y}/n - \boldsymbol{\Sigma}\|_\infty \geq \varepsilon) \leq pq \operatorname{pr}\left(\left\| \frac{1}{n} \sum_{j=1}^{n} u_{ikj} - \sigma_{ik} \right\|_\infty \geq \varepsilon \right).$$

Setting $\varepsilon = C_1 \tau_1 \tau_2 \sqrt{\frac{\log(pq/\eta)}{n}}$ and using $\log(pq) = o(n)$ completes the proof. □

Lemma 5. *Let $\boldsymbol{G}^{1/2}$ satisfy RE$(S, \boldsymbol{\lambda})$ with $\gamma = \gamma(S, \boldsymbol{\lambda}, \boldsymbol{G}^{1/2})$, and let $\lambda_{min} := \min_{d=1,\dots,D} \lambda_d$. If $s_d \leq \gamma \lambda_{min}^2 (32 D \lambda_d^2 \|\boldsymbol{G} - \boldsymbol{X'}^\top \boldsymbol{X'}\|_\infty)^{-1}$, then $\boldsymbol{X'}$ satisfies $RE(S, \boldsymbol{\lambda})$ and*

$$0 < \gamma(S, \boldsymbol{\lambda}, \boldsymbol{X'}) \leq 2\gamma(S, \boldsymbol{\lambda}, \boldsymbol{G}^{1/2}).$$

Proof. Since $\boldsymbol{G}^{1/2}$ satisfies RE$(S, \boldsymbol{\lambda})$, for all $\boldsymbol{A} \in \mathcal{C}(S, \boldsymbol{\lambda})$

$$\begin{aligned}\mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{X}'^\top \boldsymbol{X}' \boldsymbol{A}) &= \mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{G} \boldsymbol{A}) + \mathrm{Tr}\{\boldsymbol{A}^\top (\boldsymbol{G} - \boldsymbol{X}'^\top \boldsymbol{X}') \boldsymbol{A}\} \\ &\geq \gamma \|\boldsymbol{A}\|_F^2 - \|\boldsymbol{A}\|_{1,2}^2 \|\boldsymbol{G} - \boldsymbol{X}'^\top \boldsymbol{X}'\|_\infty.\end{aligned}$$

Since $\boldsymbol{A} \in \mathcal{C}(S, \boldsymbol{\lambda})$, we have

$$\begin{aligned}\|\boldsymbol{A}\|_{1,2} &\leq \sum_{d=1}^{D} \frac{\lambda_d}{\lambda_{min}} (\|\boldsymbol{A}_{d,S_d}\|_{1,2} + \|\boldsymbol{A}_{d,S_d^c}\|_{1,2}) \\ &\leq 4 \sum_{d=1}^{D} \frac{\lambda_d}{\lambda_{min}} \|\boldsymbol{A}_{d,S_d}\|_{1,2} \leq 4 \sum_{d=1}^{D} \sqrt{s_d} \frac{\lambda_d}{\lambda_{min}} \|\boldsymbol{A}_{d,S_d}\|_F \\ &\leq 4 \sqrt{\frac{\sum_{d=1}^{D} \lambda_d^2 s_d}{\lambda_{min}^2}} \|\boldsymbol{A}_S\|_F \leq 4 \sqrt{\frac{\sum_{d=1}^{D} \lambda_d^2 s_d}{\lambda_{min}^2}} \|\boldsymbol{A}\|_F,\end{aligned}$$

Therefore

$$\begin{aligned}\mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{X}'^\top \boldsymbol{X}' \boldsymbol{A}) &\geq \gamma \|\boldsymbol{A}\|_F^2 - 16 \frac{\sum_{d=1}^{D} \lambda_d^2 s_d}{\lambda_{min}^2} \|\boldsymbol{A}\|_F^2 \|\boldsymbol{G} - \boldsymbol{X}'^\top \boldsymbol{X}'\|_\infty \\ &\geq \gamma \|\boldsymbol{A}\|_F^2 - \frac{\gamma}{2} \|\boldsymbol{A}\|_F^2 = \frac{\gamma}{2} \|\boldsymbol{A}\|_F^2,\end{aligned}$$

where the last inequality holds because of the condition on s_d. □

4. PAN-CANCER ASSOCIATION ANALYSIS

4.1 Introduction

The transcriptomics deconvolution in bulk tumor samples is a popular topic in genome research. And in-depth knowledge of the role of genome has been proved to be essential to understand the nature of the cancer and to improve its prognosis. However, analyzing bulk gene expression data is difficult due to the changes in identifying cell composition (Newman et al., 2015), especially when cell subsets are contaminated by unknown mixtures. Therefore, such cell type heterogeneity makes it more complicated to identify genes signatures and bio-markers that are crucial to the interpretation of cancer.

Recently, various work has been developed to investigate the cellular heterogeneity problem. Newman et al. (2015) proposed CIBERSORT method to characterize cell heterogeneity using micro-array data. However, it produces only relative proportions of cellular subtypes and requires the reliability of the reference profiles. Li et al. (2017) developed Tumor Immune Estimation Resource (TIMER) to analyze the abundance of immune infiltrates. Wang et al. (2018) proposed DeMixT to estimate heterogeneous tumor sample compositions using RNA-seq data from a frequentist perspective. These methods utilize different genome information to estimate the cellular purity, thus also contain heterogeneity.

In this work, we apply CIBERSORT, TIMER or DeMixT methods to various cancer types from The Cancer Genome Atlas project (Weinstein et al., 2013). Since these methods produces tumor purity estimations through different channels, our primary goal is to find out what information is shared between them and identify which cellular types are predictive of prognosis. However, one of the major challenges of this task is that the estimations are proportion instead of Gaussian, thus prevents us from applying the commonly used CCA methods.

Several methods have been proposed to extend CCA to handle non-Gaussian data such as non-negative, binary or integer-valued data. Klami et al. (2010) proposed a Bayesian framework by

generalizing probabilistic CCA (Bach and Jordan, 2005) with exponential family. Instead of imposing the normality assumptions, they assume the prior of the sources is a specific combination of Gaussians. However, this method lacks identifiability guarantees and is computationally expensive for large data. Podosinnikova et al. (2016) presented a generalization of the CCA model by factoring the data directly. Nevertheless, this method can only be applied to count data. Li and Gaynanova (2018) proposed the Generalized association study framework (GAS) to extend CCA through a frequentist perspective. It assumes the entries of data matrices follow the exponential family distributions and decomposes the corresponding natural parameter matrices into joint and individual parts. Nevertheless, the model was initially developed for binary and Poisson data instead of proportion data.

In this chapter, we assess the associations between cellular purity estimations by different tools in Pan-Cancer data by exploring the GAS framework. We extended the GAS to handle proportion data. Next, the patients are clustered based on obtained scores from corresponding common or individual signals. The relationship that the estimated common and individual signals have with the survival is then being investigated, using the overall survival probability or the progression-free probability for different types of cancer. The difference in survival probability of different clusters will be assessed to find out which signal is informative for survival and predictive of prognosis.

The rest of the Chapter is organized as follows. Section 4.2 discuss the data cleaning process and describes the methods used in the analysis. Section 4.3 provides association and survival analysis results for prostate, bladder and colorectal cancer. We conclude with a summary of the scientific discoveries in Section 4.4.

4.2 Data and methodology

4.2.1 Data discription

In this work, we consider Prostate Cancer (PROD), Bladder Cancer (BLCA) and Colorectal Cancer (COLON) from the the Cancer Genome Atlas Project (Weinstein et al., 2013). We apply CIBERSORT and DeMixT to BLCA and COLON to obtain proportion estimates, and apply

TIMER and DeMixT to PROD as prostate cancer is known to be immune cold, thus CIBERSORT produce much less meaningful estimates. For each patient, CIBERSORT produces 22 estimates; TIMER produces 6 estimates and DeMixT produces 3 estimates. We further gather clinical results for patients to aid future survival analysis.

4.2.2 Data preprocessing

To facilitate analysis, we first pair the CIBERSORT/TIMER estimates and DeMixT estimated proportions by patients, and only the patients with both types of information available are selected for association analysis. For CIBERSORT estimates, we remove attributes that contains more than 85% of zero entries from further analysis, which lead to removal of T.cells.gamma.delta and Eosinophils for BLCA, T.cells.CD4.naive and T.cells.gamma.delta for COLON. The remaining zero-valued entries of TIMER/CIBERSORT estimates were replaced by 2.42×10^4 in PROD, 3.010×10^{-5} in BLCA and 6.218×10^{-6} in COLON, which correspond to minimal values of non-zero entries in each respective dataset. DeMixT estimates contain three cellular proportions: immune, stroma and tumor. We only keep two of them, as these three proportions sum to one.

4.2.3 Review of the generalized association study framework

In this subsection, we review the Generalized association study framework (GAS) (Li and Gaynanova, 2018) that finds associations between estimations by different tools. The GAS model assumes the entries of data matrices follow the exponential family distributions, and it decomposes the corresponding natural parameter matrices into an intercept, common and individual structure. The common structure contains information shared by both data matrices; whereas the individual structure contains the remaining information. Specifically, let $\boldsymbol{X}_k$ be the data matrix of size $n \times p_k$, $k = 1, 2$. And let $\boldsymbol{\Theta}_k$ be the corresponding natural parameter matrix. GAS decomposed the natural parameter matrices as

$$\boldsymbol{\Theta}_1 = \mathbf{1}_n \boldsymbol{\mu}_1^\top + \boldsymbol{U}_0 \boldsymbol{V}_1^\top + \boldsymbol{U}_1 \boldsymbol{A}_1^\top,$$

$$\boldsymbol{\Theta}_2 = \mathbf{1}_n \boldsymbol{\mu}_2^\top + \boldsymbol{U}_0 \boldsymbol{V}_2^\top + \boldsymbol{U}_2 \boldsymbol{A}_2^\top.$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^{p_k}$ is the intercept vector, $\boldsymbol{U}_0 \in \mathbb{R}^{n \times r_0}$ is the shared score matrix between $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$, and $\boldsymbol{U}_k \in \mathbb{R}^{n \times r_k}$ is the individual scores. $\boldsymbol{V}_k$ and $\boldsymbol{A}_k$ are corresponding loading matrices. r_0, r_1 and r_2 correspond to the ranks of joint and two individual structures, respectively. Essentially, the joint structure implicitly assumes the existence of shared factors between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

The GAS model was originally proposed for binary and Poisson data. Since proportions can be treated as the outcome of multiple binomial trials, We use binomial family to model all cellular estimates. The ranks of common and individual signals were determined based on a two-step data-driven method as described in Li and Gaynanova (2018). Specifically, we first estimate the ranks of the centered natural parameter matrices of $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and concatenated matrix $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ by cross-validation scores. Secondly, we determine the ranks of joint or individual structures by the estimated ranks in the first step. To test the significance of the association, we use the permutation test with 1000 permutations (Li and Gaynanova, 2018).

4.2.4 Association coefficient

To assess the strength of association between the two proportion matrices, we use the association coefficient introduced in Li and Gaynanova (2018). Since the correlation is not well-defined for non-Gaussian data, Li and Gaynanova (2018) propose to use the relative weights of the joint structure of natural parameter matrices as a criterion to evaluate the association. For completeness, we reproduce the definition here.

Definition 3. *Let $\boldsymbol{X}_1 \in \mathbb{R}^{n \times p_1}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{n \times p_2}$ be two matrices, and let $\boldsymbol{\Theta}_1 \in \mathbb{R}^{n \times p_1}$ and $\boldsymbol{\Theta}_2 \in \mathbb{R}^{n \times p_2}$ be the corresponding natural parameter matrices. Further let $\bar{\boldsymbol{\Theta}}_1 \in \mathbb{R}^{n \times p_1}$ and $\bar{\boldsymbol{\Theta}}_2 \in \mathbb{R}^{n \times p_2}$ be the column centered natural parameter matrices. We define the association coefficient as*

$$\textit{association coefficient} = \frac{\|\bar{\boldsymbol{\Theta}}_1^\top \bar{\boldsymbol{\Theta}}_2\|_*}{\|\bar{\boldsymbol{\Theta}}_1\|_F \|\bar{\boldsymbol{\Theta}}_2\|_F}. \tag{4.1}$$

Note that the range of the association coefficient is from 0 to 1, with larger value indicates stronger association. For more details, we refer to Section 3.1 in Li and Gaynanova (2018).

4.3 Analysis results

4.3.1 Prostate cancer

In this section, we consider prostate cancer data in TCGA with TIMER and DeMixT estimates, where we replaced two readings of Dendritic in TIMER estimates $(1.028, 1.052)$ by 1. For DeMixT proportions, we noticed that immune and tumor proportions are highly correlated. Figure 4.1 shows the scatter plots of DeMixT estimates and its corresponding natural parameters. Due to the high correlation between immune and tumor proportions, we keep immune and stroma proportions for further analysis. The resulting dataset contains 293 subjects in total, 6 and 2 proportions for TIMER and DeMixT estimates, respectively.

The corresponding association coefficients is found to be 0.389. The association coefficient is highly significant based on permutation tests (p-value ≤ 0.001), thus indicating presence of moderate associations between TIMER and DeMixT proportions for prostate cancer. Based on the cross-validation scores in Figure 4.2, the rank of the joint structure is equal to 1, and the ranks of individual structures are equal to 2 and 1 for TIMER and DeMixT proportions correspondingly. This implies that DeMixT proportions in PROD contain individual signal that is not shared by TIMER estimates, thus explaining the moderate value of association coefficient.

To assess biological relevance of found associations, we investigated the relationship that the estimated common and individual signals have with the survival, using the progression-free interval (PFI). The subjects are clustered into 2 groups based on the joint/individual scores or TIMER/DeMixT estimates, and the difference in survival probability was assessed. The 2 groups clustered by TIMER estimates and individual signals do not have significant difference in survival probability (log-rank test p-value $= 0.520, 0.479, 0.126$ if clustered by TIMER estimates, individual signals of TIMER and DeMixT estimates, respectively). However, we observe a significant difference in survival based on 2 groups clustered by joint signals (Figure 4.3 left, log-rank test p-value $= 0.002$). The corresponding TIMER cells contributed the most to the common signal were (+)Macrophage, (+)Dendritic, (+)CD4_Tcell. We also observe a significant differ-

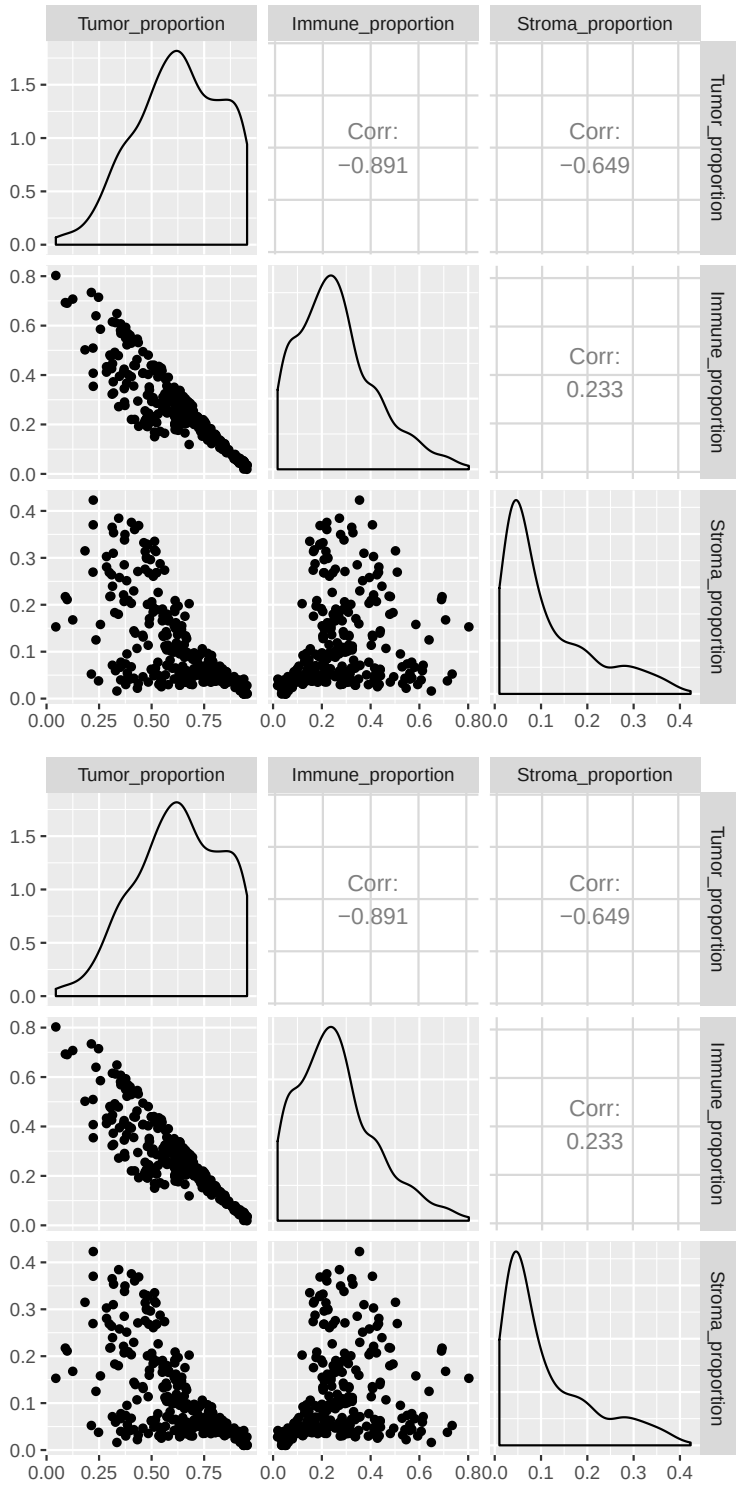

Figure 4.1: Top: scatter plots of DeMixT proportions from Prostate cancer. Bottom: scatter plots of saturated natural parameters of DeMixT proportions from Prostate cancer.

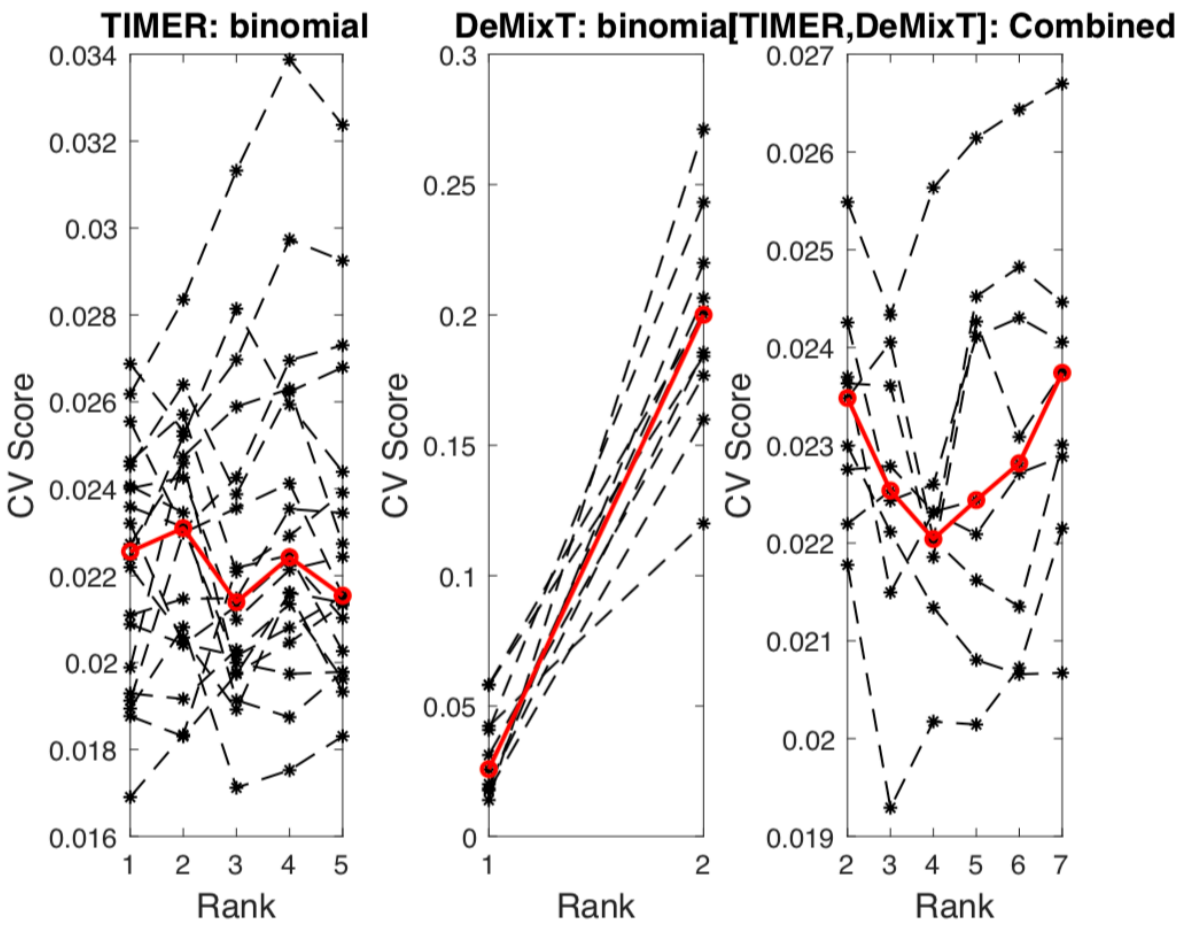

Figure 4.2: Cross-validation scores for TIMER, DeMixT and (TIMER, DeMixT) of PROD, respectively. The red solid line indicates the median of CV scores.

ence in survival based on 2 groups clustered by DeMixT proportions (Figure 4.3 right, log-rank test p-value $= 0.001$). Figure 4.4 shows the distributions of immune and stroma proportions of DeMixT estimates in different groups. We observed that both immune and normal stroma are predictive to survival outcomes. In general, patients with higher immune and higher normal proportions tended to have longer PFI.

4.3.2 Bladder cancer

We consider bladder cancer data in TCGA with CIBERSORT and DeMixT estimates. For DeMixT proportions, we keep immune and stroma proportions as in Section 4.3.1. However,

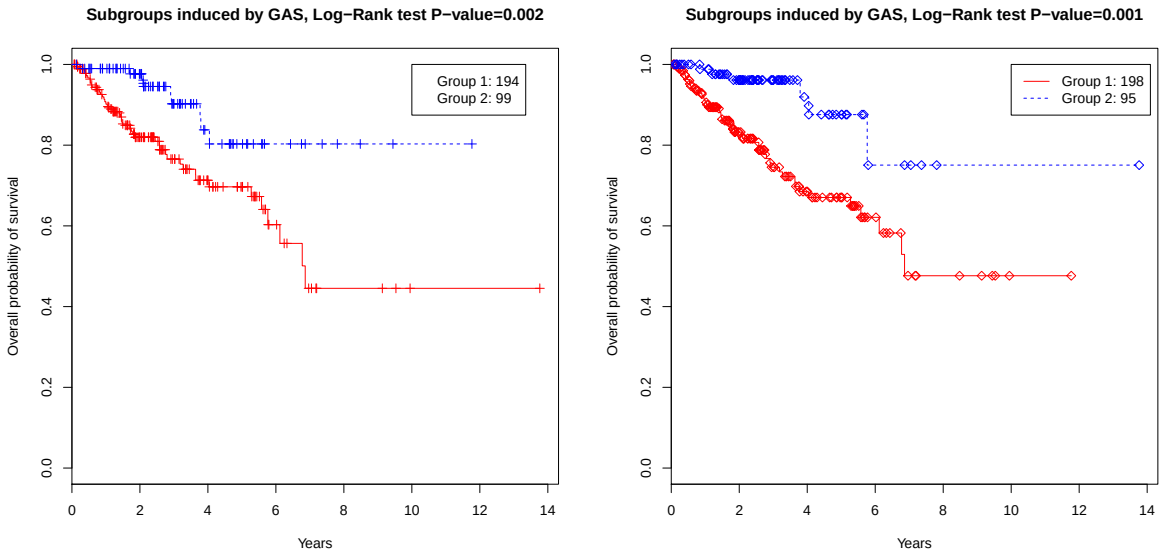

Figure 4.3: Kaplan-Meier plot for progression-free interval for prostate cancer. The log-rank test is used to compare the survival curves of two clusters and calculate p-values. Left: The patients are clustered by common signals. Right: The patients are clustered by DeMixT proportions.

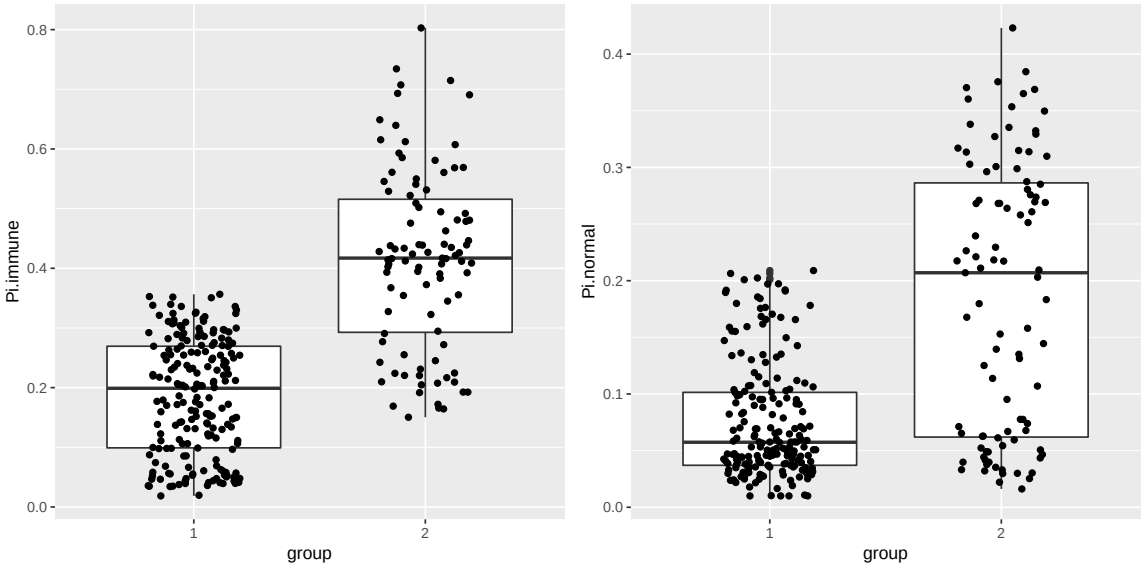

Figure 4.4: The boxplots of immune-normal proportions. Subjects are clustered by DeMixT proportions.

unlike prostate cancer, we do not observe high correlation between immune and tumor proportions for bladder cancer. The resulting data contains 385 samples, 20 and 2 proportions for CIBERSORT

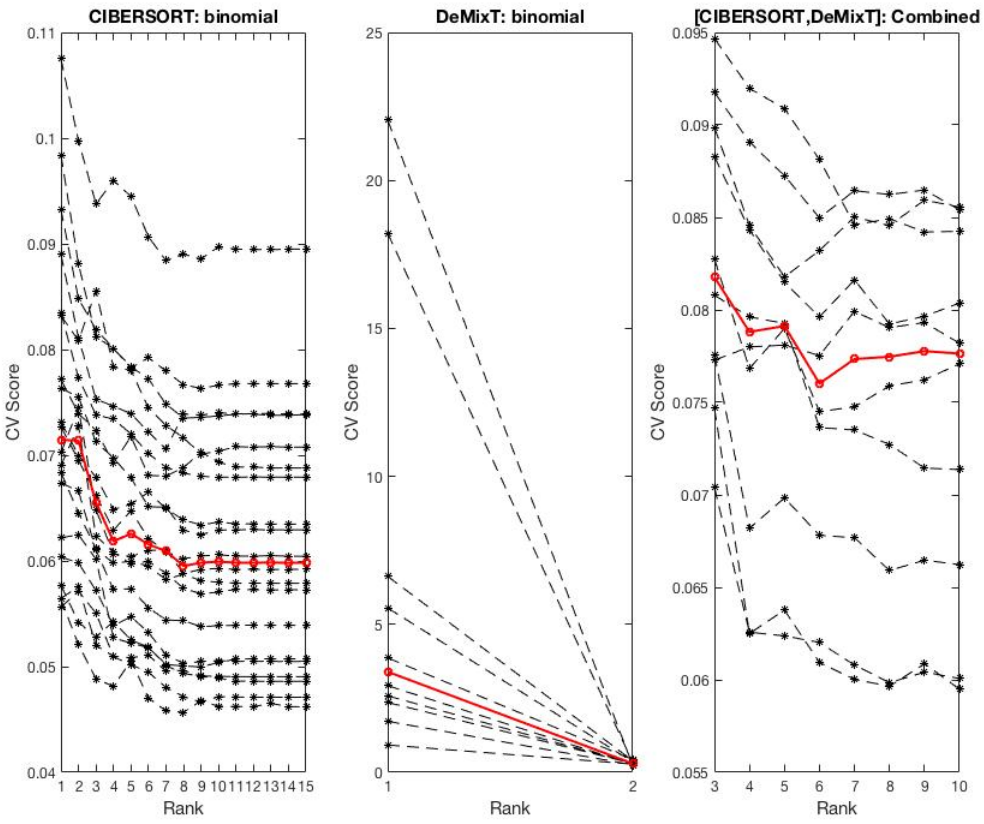

Figure 4.5: Cross-validation scores for CIBERSORT, DeMixT and (CIBERSORT, DeMixT) of BLCA, respectively. The red solid line indicates the median of CV scores.

and DeMixT proportions, respectively.

Figure 4.5 shows the rank selection results. The rank of the joint structure is equal to 2, and the ranks of individual structures are equal to 4 and 0 for CIBERSORT and DeMixT proportions, respectively. This implies that all the signal in DeMixT proportions is shared by CIBERSORT proportions, however there is additional information in CIBERSORT proportions. The association coefficient is found to be 0.553. The association coefficient is highly significant based on permutation tests (p-value ≤ 0.001), thus indicating rather strong association between CIBERSORT and DeMixT proportions.

We investigated the relationship that the estimated common and individual signals had with the survival, using the progression-free interval (PFI). The subjects are clustered into 2 groups based on

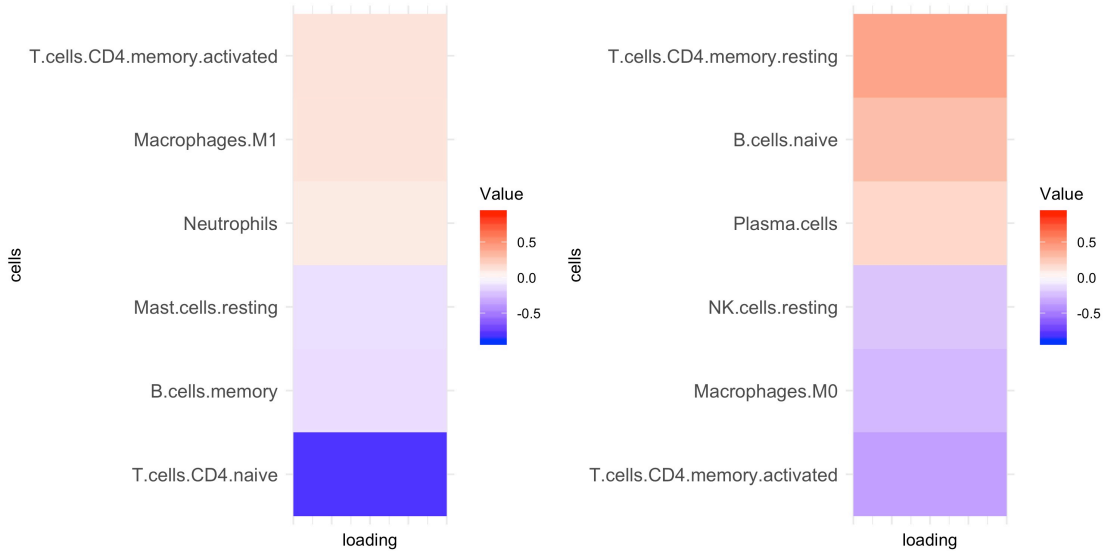

Figure 4.6: Loadings of CIBERSORT proportions correspond to two shared scores.

the joint/individual scores or CIBERSORT/DeMixT estimates, and Figure 4.7 shows the survival results. The 2 groups clustered by CIBERSORT proportions and its individual signals do not have significant difference in survival probability (log-rank test p-value $= 0.483, 0.416$, respectively). On the other hand, the 2 groups clusters based on common signal have significantly different survival probability (log-rank test p-value $= 0.029$). Figure 4.6 shows the six immune subtype cells with the largest contributions to the common scores of CIBERSORT estimates. T.cells.CD4.naive dominates the first loading vector and T.cells.CD4.memory.resting dominates the second loading vector. We also observe a significant difference in survival based on 2 groups clustered by DeMixT proportions (log-rank test p-value $= 0.004$). Note that clustering results by common signals and immune-stroma proportions are similar, and Table 4.1 confirms this observation.

Figure 4.8 shows the distributions of immune and stroma proportions in different groups. Unlike prostate cancer, only immune proportions were predictive to survival outcomes, and patients with higher immune proportions tended to have shorter PFI.

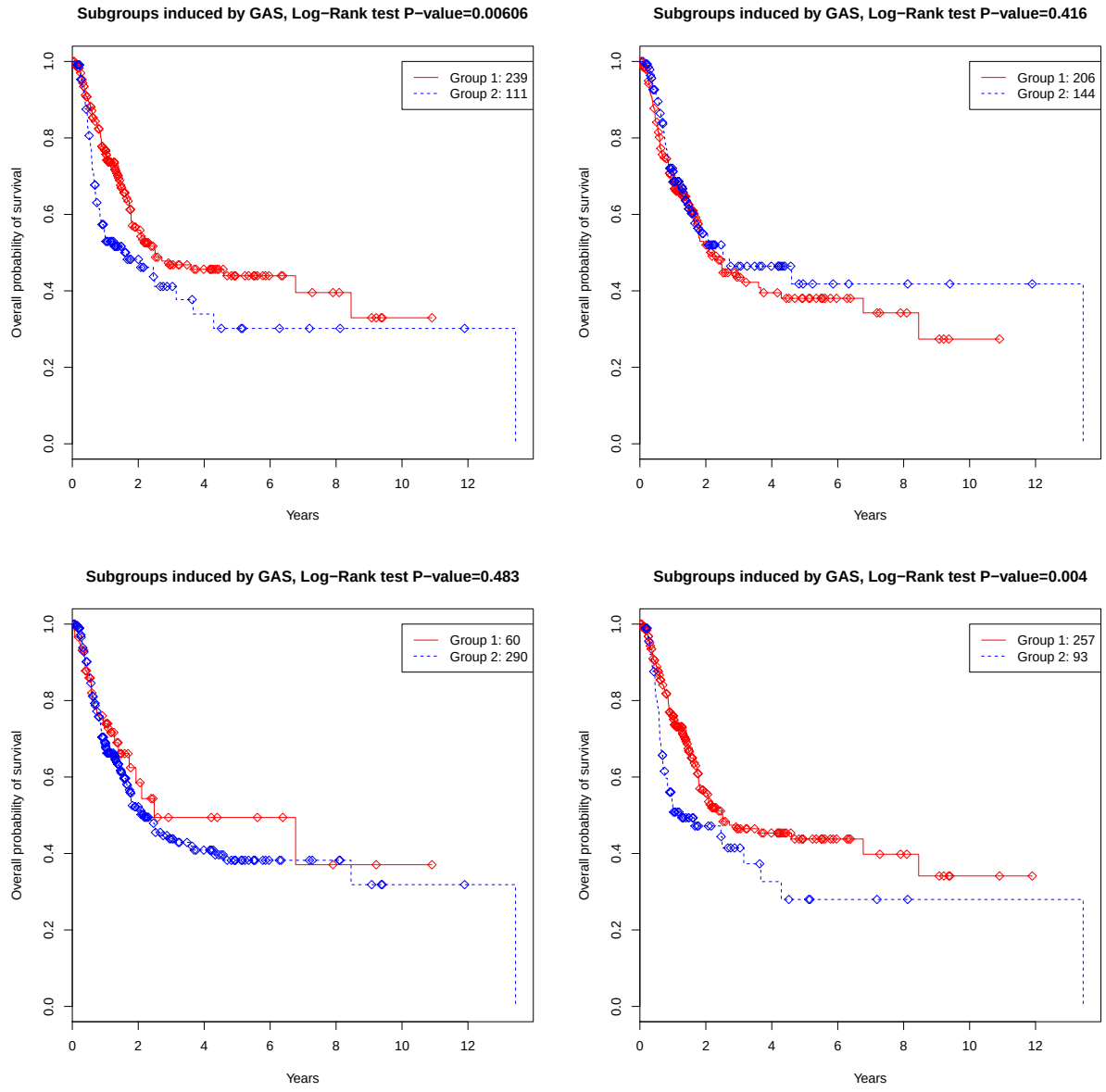

Figure 4.7: Kaplan-Meier plot for progression-free interval for bladder cancer. Top left: clustered by joint signals. Top right: clustered by individual signals of CIBERSORT. Bottom left: clustered by CIBERSORT proportions. Bottom right: clustered by DeMixT proportions.

4.3.3 Colorectal cancer

We consider colorectal cancer data in TCGA with CIBERSORT and DeMixT proportions. We keep immune and stroma proportions of DeMixT proportions as in Section 4.3.1. The resulting data contains 420 samples, 20 and 2 proportions for CIBERSORT and DeMixT proportions, re-

Table 4.1: Number of patients in different groups clustered by common sigals or DeMixT proportions.

	Clustered by common signals	
Clustered by DeMixT	group 1	group 2
group 1	238	19
group 2	1	92

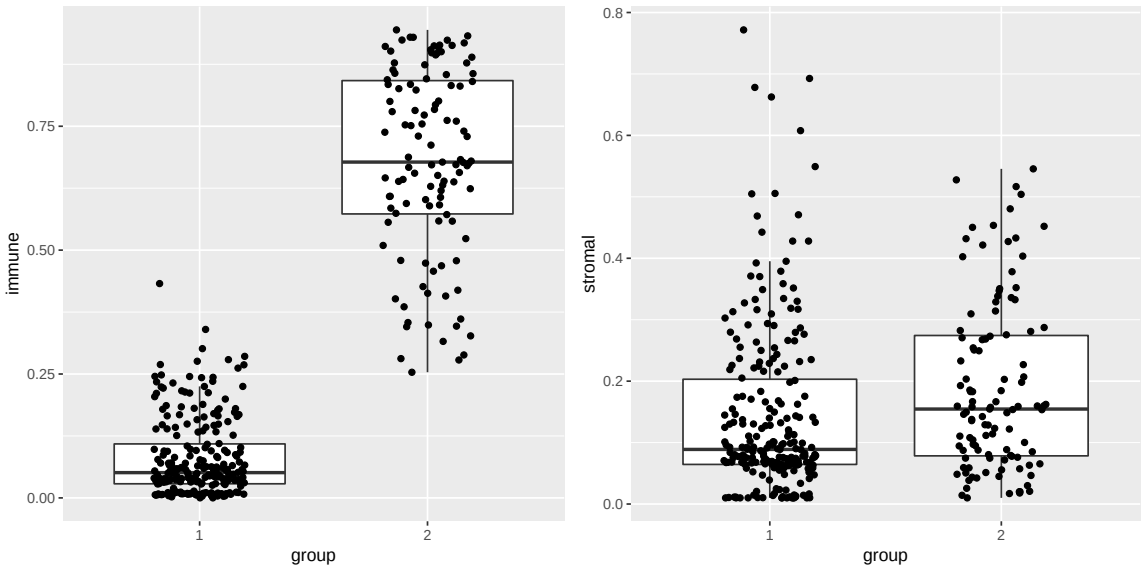

Figure 4.8: The boxplots of DeMixT proportions. Patients were clustered by common signals.

spectively.

The rank selection results are presented in Figure 4.9. The rank of the joint structure is equal to 2, and the ranks of individual structures are equal to 4 and 0 for CIBERSORT and DeMixT proportions, respectively. Similar to BLCA, this implies that all the signal in DeMixT proportions are shared by CIBERSORT proportions, however there is additional information in CIBERSORT proportions. The association coefficient is found to be 0.347. The association coefficient is highly significant based on permutation tests (p-value ≤ 0.001), thus indicating moderate associations between CIBERSORT and DeMixT proportions.

We investigated the relationship that the estimated common and individual signals have with

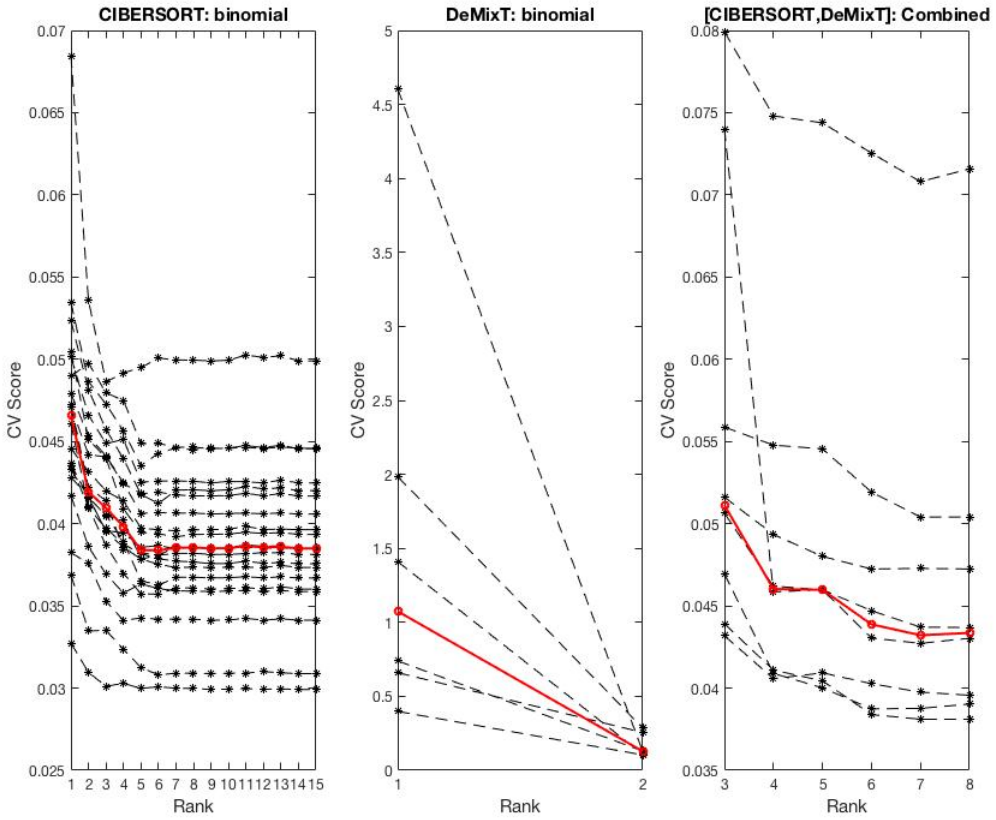

Figure 4.9: Cross-validation scores for CIBERSORT, DeMixT and (CIBERSORT, DeMixT) of COLON, respectively. The red solid lines indicate the median of CV scores.

the survival, using the survival time. The subjects are clustered into 3 groups based on the joint/individual scores or CIBERSORT/DeMixT estimates. We do not cluster the subjects into 2 groups since the clustering results are imbalanced. Figure 4.10 shows the results of survival analysis. In contrast to PROD and BLCA, only the clusters based on individual signals of CIBERSORT have significantly different survival probability (log-rank test p-value $= 0.038$). This implies that the additional information in CIBERSORT proportions are predictive to cancer prognosis.

4.4 Discussion

In this chapter, we compare different cellular subtype estimations for Pan-cancer data. In particular, we identified a rather strong association between CIBERSORT and DeMixT estimations

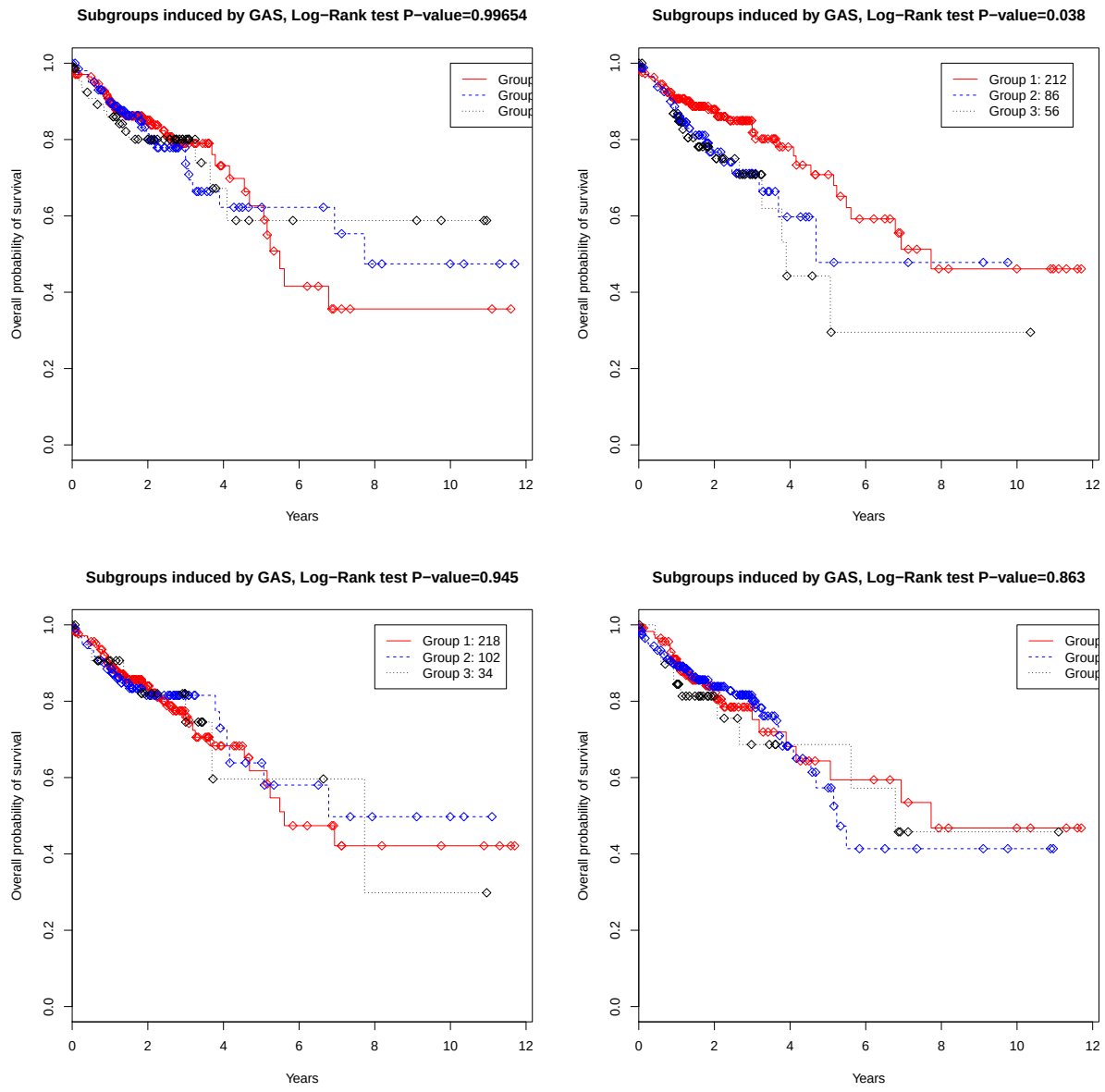

Figure 4.10: Kaplan-Meier plot for progression-free interval for colorectal cancer. Top left: clustered by joint signals. Top right: clustered by individual signals of CIBERSORT. Bottom left: clustered by CIBERSORT proportions. Bottom right: clustered by DeMixT proportions.

for BLCA, and moderate associations between TIMER and DeMixT estimations for PROD, and between CIBERSORT and DeMixT estimations for COLON. Based on the survival results, we observe that immune proportions are predictive of the prognosis of BLCA and PROD, but have the opposite relationship with the progression-free interval. On the one hand, patients with higher

immune proportions tend to have longer PFI for PROD, but have shorter PFI for BLCA. On the other hand, for COLON, immune proportions do not show a clear relationship with survival. Nevertheless, the additional information in CIBERSORT proportions after removing the shared signals with DeMixT estimations are predictive to colorectal cancer prognosis.

There are a few works to be done. First, how immune proportion estimations is related to PFI is still not clear. A thorough investigation will not only help us understand the mechanism of cancer, but also shed lights on personalized medicine and anticancer therapies. Second, the individual structures extracted by GAS still contain shared information, since they are not orthogonal to each other. By further decompose individual structures, we may gain more knowledge regarding the cellular heterogeneity problem of different tissues. We will look into this problem in the next chapter.

5. LOW-RANK CANONICAL CORRELATION ANALYSIS

5.1 Introduction

With the advancement of biomedical technologies, the complexity and number of sources of tumor data has been growing rapidly, and multi-view data becomes common in the downstream analysis of cancer data. Often, multi-view data comes from different platforms are processed by different tools, thus the measurements have heterogeneous types in practice. One major challenge of this type of data is how to conduct association analysis. For example, the Pan-cancer dataset discussed in Section 4.2 consists of compartment-specific gene expression data for tumor samples. Different views are estimated proportions or abundances by various deconvolution methods proposed by the biomedical community (Newman et al., 2015; Li et al., 2017; Wang et al., 2018). Since the ranges of the data are usually bounded between zero and one instead of real-valued, the standard CCA method is no longer an appropriate tool in this scenario. The reason is two folded. First, CCA implicitly assumes that the input variables are real-valued, and Bach and Jordan (2005) provided a probabilistic interpretation of CCA under the Gaussian assumption by using a factor model. However, this assumption may be inappropriate if data is proportion. Secondly, correlations are usually not well-defined for proportion data. Therefore the objective function of CCA lacks a straightforward interpretation when it is applied to non-Gaussian data.

In order to address this challenge, a typical approach is to disentangle the common and individual signals of two views (Lock et al., 2013; OConnell and Lock, 2016; Shu et al., 2019). These methods factorize the common and individual signals of two data matrices, and then use them to further perform integrative or discriminative analysis. Nevertheless, these methods are proposed for Gaussian data and cannot be applied to the non-Gaussian cases. Recently, several methods have been proposed to extend this idea to handle non-Gaussian data by connecting to the exponential family. Such methods include Exponential Family CCA (Klami et al., 2010), discrete CCA (Podosinnikova et al., 2016) and the Generalized association study framework (Li and Gaynanova,

2018). However, these decomposition-based methods assume the shared factors or score matrix are identical between two views. Therefore, the extracted common parts usually do not coincide with the desired canonical variables, even when applied to the Gaussian cases. In addition, with the exception of D-CCA (Shu et al., 2019), most methods only enforces the column-space orthogonality between the common and individual structures, and hence do not guarantee the orthogonality between two individual parts of two data matrices. In other words, although each dataset is factored into common and individual parts, there is still shared information embedded in the individual parts.

In this chapter, we propose a decomposition based method to conduct association analysis for both Gaussian and non-Gaussian data, such as non-negative, binary or count data. We call it low-rank CCA. Throughout the chapter, we assume that the variables in each dataset follows the exponential family distribution conditioned on the underlying natural parameters. Our model decomposes the corresponding natural parameter matrices into a low-rank joint structure and individual structure, and use the joint structure to capture the shared information between the views. In contrast to the existing decomposition methods, a unique characteristic of our approach is that we allow the joint scores of two views to take different values. In addition, our method imposes the orthogonality between two individual parts, hence guarantees that no shared information retained in the individual parts.

The proposed optimization problem for our method is not convex and therefore the traditional gradient descent algorithm cannot be used. For implementation, we derive an alternating algorithm that optimize over each component alternatively. Within each step, the optimization problem can be formulated as a convex question with a quadratic equality constraint, and we propose to modify the splitting orthogonality constraints method (Lai and Osher, 2014) to solve it. Although the overall global convergence is not guaranteed, our numerical experiments show that the proposed method has promising performance.

The rest of the chapter is organized as follows. Section 5.2 introduces the low-rank CCA model under a low-rank frame and connect it to the exponential family. We also discuss the corre-

sponding regularity conditions. Section 5.3 describes an alternating algorithm to fit the model. In Section 5.4, we demonstrate the effectiveness of the proposed method based on simulation studies. Section 5.5 summarizes the major conclusions and discusses future extensions.

5.2 Proposed methodology

In this section, we first review the natural exponential family. Then we introduce the proposed method by connecting the structural decomposition method with the exponential family.

5.2.1 Natural exponential family

Assume a random variable x follows from an exponential family distribution, then the distribution function given the parameter θ has the form

$$f(x|\theta) = c(x)\exp\{x\theta - b(\theta)\},$$

where θ is called the canonical natural parameter, $b(\cdot)$ is a real-valued convex function that has different forms for members in the exponential family, and $c(\cdot)$ ensures that the probability function $f(x|\theta)$ is normalized. We further note that $\mu = \mathbb{E}(x|\theta) = b'(\theta) = \partial b(\theta)/\partial\theta$ and $\mathrm{Var}(x|\theta) = b''(\theta) = \partial^2 b(\theta)/\partial\theta^2$. Define the canonical link function $g(\cdot)$ such that $g(b'(\theta)) = g(\partial b(\theta)/\partial\theta) = \theta$. Therefore $g(\mu) = b'^{-1}(\mu)$ holds.

5.2.2 The normal case

In this subsection, we reformulate CCA problem under normal distribution and propose a low-rank method to decompose the matrices. Without loss of generality, let $\boldsymbol{X}_1 \in \mathbb{R}^{n\times p_1}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{n\times p_2}$ be two column-centered data matrices with rank $r_1 < p_1$ and $r_2 < p_2$, respectively. Assume the covariance matrix of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ has rank r_0 with $0 \le r_0 \le \min(r_1, r_2)$. Define ℓ-th pair of canonical variables u_l, v_l as

$$u_l, v_l = \arg\max_{u,v} \mathrm{Cor}(u, v), \text{ subject to } \mathrm{Var}(u) = \mathrm{Var}(v) = 1,$$
$$\text{and } u \in Col(\boldsymbol{X}_1)/\text{span}(\{u_i\}_{i=1}^{l-1}), v \in Col(\boldsymbol{X}_2)/\text{span}\left(\{v_i\}_{i=1}^{l-1}\right),$$

where $Col(\boldsymbol{X}_k)$ is the column space of the matrix $\boldsymbol{X}_k$, and span(u) is the space spanned by the vector u. Here $\mathrm{Cor}(u_l, v_l) = \rho_l$ is ℓ-th canonical correlation.

Remark 4. *Let $\{u_i\}_{i=r_0+1}^{r_1}$ be a set of arbitrary orthogonal basis of $Col(\boldsymbol{X}_1)$/span$(\{u_i\}_{i=1}^{r_0})$, and let $\{v_i\}_{i=r_0+1}^{r_2}$ be a set of arbitrary orthogonal basis of $Col(\boldsymbol{X}_2)$/span$(\{u_i\}_{i=1}^{r_0})$. Then $\{u_i\}_{i=1}^{r_1}$ is a set of orthogonal basis of the column space of $\boldsymbol{X}_1$, and $\{v_i\}_{i=1}^{r_2}$ is a set of orthogonal basis of the column space of $\boldsymbol{X}_2$.*

Theorem 5. *Define $\boldsymbol{U}_1 = [u_1, \cdots, u_{r_0}] \in \mathbb{R}^{n\times r_0}$, $\boldsymbol{U}_2 = [v_1, \cdots, v_{r_0}] \in \mathbb{R}^{n\times r_0}$. Further define $\boldsymbol{Z}_1 = [u_{r_0+1}, \cdots, u_{r_1}] \in \mathbb{R}^{n\times(r_1-r_0)}$ and $\boldsymbol{Z}_2 = [v_{r_0+1}, \cdots, v_{r_2}] \in \mathbb{R}^{n\times(r_2-r_0)}$. Then the covariance matrices of $\boldsymbol{J}_1 = \begin{pmatrix}\boldsymbol{U}_1 & \boldsymbol{Z}_1\end{pmatrix}$, $\boldsymbol{J}_2 = \begin{pmatrix}\boldsymbol{U}_2 & \boldsymbol{Z}_2\end{pmatrix}$ are*

$$\mathrm{Cov}(\boldsymbol{J}_1) = \boldsymbol{I}_{r_1}, \mathrm{Cov}(\boldsymbol{J}_2) = \boldsymbol{I}_{r_2} \text{ and}$$
$$\mathrm{Cov}(\boldsymbol{J}_1, \boldsymbol{J}_2) = \begin{pmatrix}\boldsymbol{\Lambda}_{r_0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{(\mathbf{r_1}-\mathbf{r_0})\times(\mathbf{r_2}-\mathbf{r_0})}\end{pmatrix},$$

where $\mathbf{0}_{\mathbf{p}\times\mathbf{q}}$ is a $p \times q$ zero-valued matrix, and $\boldsymbol{\Lambda}_{r_0} = \mathrm{diag}(\lambda_1, \cdots, \lambda_{r_0})$ is a diagonal matrix.

The rigorous proof of Theorem 5 is similar to Theorem 1 in Shu et al. (2019). We restate this theorem here for completeness. Theorem 5 indicates that the correlations between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are captured by $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$, and $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are orthogonal to each other. Based on the construction, $\boldsymbol{J}_1$ is a set of basis of $Col(\boldsymbol{X}_1)$, and $\boldsymbol{J}_2$ is a set of basis of $Col(\boldsymbol{X}_2)$. This motivates us to characterize the joint and individual parts by a low-rank plus noise structure.

In general, when $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are not column-centered, we propose to model them as

$$\begin{cases}\boldsymbol{X}_1 = \mathbf{1}_n\boldsymbol{\mu}_1^\top + \boldsymbol{U}_1\boldsymbol{V}_1^\top + \boldsymbol{Z}_1\boldsymbol{A}_1^\top \\ \boldsymbol{X}_2 = \mathbf{1}_n\boldsymbol{\mu}_2^\top + \boldsymbol{U}_2\boldsymbol{V}_2^\top + \boldsymbol{Z}_2\boldsymbol{A}_2^\top\end{cases}. \tag{5.1}$$

where $\mathbf{1}_n$ is a vector of all ones with length n. We call (5.1) low-rank CCA. Essentially, each data matrix is decomposed into three parts: the intercept(the first term), the joint structure (the

second term) and the individual structure (the third term). We call $\boldsymbol{\mu}_1 \in \mathbb{R}^{p_1}$ and $\boldsymbol{\mu}_2 \in \mathbb{R}^{p_2}$ the intercept vectors. Denote $\boldsymbol{U}_1 \in \mathbb{R}^{n\times r_0}$ and $\boldsymbol{U}_2 \in \mathbb{R}^{n\times r_0}$ to be shared score matrices and $\boldsymbol{V}_1 \in \mathbb{R}^{r_0\times p_1}$ and $\boldsymbol{V}_2 \in \mathbb{R}^{r_0\times p_2}$ to be corresponding loading matrices. Denote $\boldsymbol{Z}_1 \in \mathbb{R}^{n\times(r_1-r_0)}$ and $\boldsymbol{Z}_2 \in \mathbb{R}^{n\times(r_2-r_0)}$ to be individual score matrices and $\boldsymbol{V}_1 \in \mathbb{R}^{(r_1-r_0)\times p_1}$ and $\boldsymbol{V}_2 \in \mathbb{R}^{(r_2-r_0)\times p_2}$ to be corresponding loading matrices. In contrast to GAS and D-CCA, we allow $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ to take different values. They can be considered as common latent factors, and we assume any shared information are bared in them.

To ensure the identifiability of model (5.1) and maintain the desired constraints in Theorem 5, we consider the following regularity conditions

- The rank of $\boldsymbol{U}_k$ or $\boldsymbol{V}_k$ is r_0. The rank of $\boldsymbol{Z}_1$ or $\boldsymbol{A}_1$ is r_1. The rank of $\boldsymbol{Z}_2$ or $\boldsymbol{A}_2$ is r_2.
- The intercept vectors μ_k are independent from the columns of joint and individual loading matrices ($\boldsymbol{V}_k$ and $\boldsymbol{A}_k$, for $k = 1, 2$).
- The score matrices ($\boldsymbol{U}_k$ and $\boldsymbol{Z}_k$) are column centered and have orthonormal columns.
- The column spaces of $\boldsymbol{U}_k$ and $\boldsymbol{Z}_l$ are orthogonal, and the column spaces of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are also orthogonal, that is,

$$\boldsymbol{U}_1^\top \boldsymbol{U}_2 = \Lambda, \boldsymbol{Z}_1^\top \boldsymbol{Z}_2 = \mathbf{0}, \boldsymbol{U}_\mathbf{k}^\top \boldsymbol{Z}_\mathbf{l}, \text{ for } \mathbf{k}, \mathbf{l} = \mathbf{1}, \mathbf{2}.$$

The first two conditions guarantee that the joint and individual score matrices are not ill-conditioned, and the ranks are correctly specified. Therefore, the model cannot be further reduced. The third condition assures the score matrices are unique up to orthogonal transformation. The last condition states that there is no shared information between joint and individual structures, and we additionally enforce that no common structure is retained in the individual structures.

5.2.3 Exponential CCA

Suppose we observe two data matrices $\boldsymbol{X}_1 \in \mathbb{R}^{n\times p_1}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{n\times p_2}$, where rows are n matched samples, and each variable comes from an exponential family distribution with different

natural parameters. Therefore, each data matrix $\boldsymbol{X}_k$, $k = 1, 2$ corresponds to a natural parameter matrix $\boldsymbol{\Theta}_k \in \mathbb{R}^{n \times p_k}$. We assume that each variable of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is independent from each other, given the natural parameters. This assumption is also used in Li and Gaynanova (2018) and Landgraf and Lee (2015).

We consider decomposing natural parameter matrix $\boldsymbol{\Theta}_k$ instead of data matrices as in Section 5.2.1. In other words, we assume that the natural parameter matrices admit a low-rank structure which describes the associations among the variables. We propose to model the natural parameter matrices as

$$\begin{cases} \boldsymbol{\Theta}_1 = \mathbf{1}_n \boldsymbol{\mu}_1^\top + \boldsymbol{U}_1 \boldsymbol{V}_1^\top + \boldsymbol{Z}_1 \boldsymbol{A}_1^\top \\ \boldsymbol{\Theta}_2 = \mathbf{1}_n \boldsymbol{\mu}_2^\top + \boldsymbol{U}_2 \boldsymbol{V}_2^\top + \boldsymbol{Z}_2 \boldsymbol{A}_2^\top \end{cases}, \tag{5.2}$$

with the same regularity conditions in Section 5.2.1. A nice structure of this method is that in the normal distribution, the natural parameter are $\theta = x$, thus model (5.2) reduce to (5.1).

5.3 Estimation of parameters

We propose to use an alternating algorithm to estimate the score and loading matrices in (5.2), based on prefixed ranks of joint and individual structures. Essentially, this minimization problem is not convex overall the parameters. However, this problem can be formulated as a set of convex problems or convex problems with a quadratic equality constraint.

We consider the joint negative log-likelihood of the data matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ as the loss function. Since the variables are conditionally independent given natural parameters, the loss function takes the form

$$\begin{aligned} \min_{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2} L &= \min_{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2} L(\boldsymbol{X}_1 | \boldsymbol{\Theta}_1) + L(\boldsymbol{X}_2 | \boldsymbol{\Theta}_2) \\ &= \min_{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{p_1} (-x_{1,ij} \theta_{1,ij} + b_1(\theta_{1,ij})) + \sum_{i=1}^{n} \sum_{j=1}^{p_2} (-x_{2,ij} \theta_{2,ij} + b_2(\theta_{2,ij})) \right\}, \end{aligned} \tag{5.3}$$

where $x_{k,ij}$ and $\theta_{k,ij}$ are i, jth element of $\boldsymbol{X}_k$ and $\boldsymbol{\Theta}_k$, for $k = 1, 2$. The model parameters in (5.3)

include intercept $\boldsymbol{\mu}_k$, joint score matrices $\boldsymbol{U}_k$, individual score matrices $\boldsymbol{Z}_k$ and loading matrices $\boldsymbol{V}_k$ and $\boldsymbol{A}_k$. Since the minimization problem (5.3) is not convex in $\boldsymbol{u}$, $\boldsymbol{v}$, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ together, we adopt an alternating method that estimate each parameter matrix alternatively until the whole algorithm converges. Although the global convergence is not guaranteed, our simulation provide evidence of its effectiveness if the initial points are chosen appropriately.

Consider solving (5.3) in terms of $\boldsymbol{\mu}_1$ and $\boldsymbol{A}_1$ with other parameters fixed. We only need to consider the first term of (5.3), which is $L(\boldsymbol{X}_1|\boldsymbol{\Theta}_1)$, since the second term doesn't contain $\boldsymbol{\mu}_1$ and $\boldsymbol{A}_1$. Let $\mu_{1,i}$ be ith element of $\boldsymbol{\mu}_1$, $\boldsymbol{V}_{1,i:}$ be the ith row vector of $\boldsymbol{V}_1$ and $\boldsymbol{A}_{1,i:}$ be the ith row vector of $\boldsymbol{A}_1$. Then the ith column of $\boldsymbol{\Theta}_1$ can be formulated as

$$\theta_{1,:i} = \mu_{1,i}\mathbf{1}_n + \boldsymbol{U}_1\boldsymbol{V}_{1,i:} + \boldsymbol{Z}_1\boldsymbol{A}_{1,i:},$$

where the second term ($\boldsymbol{U}_1\boldsymbol{V}_{1,i:}$) is fixed. Therefore, the optimisation problem can be further separated into p_1 convex problems:

$$\min_{\mu_{1,j},\boldsymbol{A}_{1,j:}} \sum_{i=1}^{n} \left[-x_{1,ij}\theta_{1,ij}(\mu_{1,j}, \boldsymbol{A}_{1,j:}) + b_1(\theta_{1,ij}(\mu_{1,j}, \boldsymbol{A}_{1,j:}))\right], \text{ for } j = 1, \cdots, p_1.$$

We use the damped Newton's method to estimate $\mu_{1,j}$ and $\boldsymbol{A}_{1,j:}$, and choose the step size that satisfies Armijo-Wolfe conditions (Fletcher, 2013; Nocedal and Wright, 2006). Therefore, by choosing the step size carefully, the convergence of damped Newton's method is guaranteed. What's more, these p_1 convex problems can be solved in parallel, which accelerate the speed of the algorithm.

Next, we estimate $\boldsymbol{\mu}_1$ and $\boldsymbol{A}_1$ in a similar manner by separating them into p_2 convex problems. The damped Newton's method with proper step size is again used to solve the optimization problems. At this step, we update the intercepts and the individual loading matrices.

Now we estimate the joint score matrices $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$. With other parameters fixed, we formu-

late the estimation problem as

$$\begin{aligned}\min_{\boldsymbol{Z}_1,\boldsymbol{Z}_2} L &= \min_{\boldsymbol{Z}_1,\boldsymbol{Z}_2} L(\boldsymbol{X}_1|\boldsymbol{\Theta}_1) + L(\boldsymbol{X}_2|\boldsymbol{\Theta}_2) \\ &\text{subject to } \begin{pmatrix}\mathbf{1}_n & \boldsymbol{U}_1 & \boldsymbol{U}_2\end{pmatrix}^\top \begin{pmatrix}\boldsymbol{Z}_1 & \boldsymbol{Z}_2\end{pmatrix} = \mathbf{0} \text{ and } \begin{pmatrix}\boldsymbol{Z}_1 & \boldsymbol{Z}_2\end{pmatrix}^\top \begin{pmatrix}\boldsymbol{Z}_1 & \boldsymbol{Z}_2\end{pmatrix} = \mathbf{I},\end{aligned} \tag{5.4}$$

where the constraints are inherited from the regularity conditions. We remark that the problem is actually convex with orthogonality and linear constraints. In general, this type of problems are challenging due to the non-convex constraints, and may have several different local minimizers. In the literature, several methods have been proposed to convex problems with only orthogonality constraint (Lai and Osher, 2014; Wen and Yin, 2013). Inspired by the idea of method of splitting orthogonality constraints (SOC) and Bregman iteration method (Yin et al., 2008; Lai and Osher, 2014), we propose a new algorithm to solve (5.4).

We introduce two auxiliary variables $\boldsymbol{P}_1 = \boldsymbol{Z}_1$ and $\boldsymbol{P}_2 = \boldsymbol{Z}_2$ to separate the original constraints into an orthogonal constrained problem with an analytical solution and an unconstrained one. Hence the minimization problem (5.4) becomes

$$\begin{aligned}\min_{\boldsymbol{Z}_1,\boldsymbol{Z}_2,\boldsymbol{P}_1,\boldsymbol{P}_2} L &= \min_{\boldsymbol{Z}_1,\boldsymbol{Z}_2,\boldsymbol{P}_1,\boldsymbol{P}_2} L(\boldsymbol{X}_1|\boldsymbol{\Theta}_1) + L(\boldsymbol{X}_2|\boldsymbol{\Theta}_2) \\ &\text{subject to } \begin{pmatrix}\boldsymbol{Z}_1 & \boldsymbol{Z}_2\end{pmatrix} = \begin{pmatrix}\boldsymbol{P}_1 & \boldsymbol{P}_2\end{pmatrix}, \begin{pmatrix}\mathbf{1}_n & \boldsymbol{U}_1 & \boldsymbol{U}_2\end{pmatrix}^\top \begin{pmatrix}\boldsymbol{P}_1 & \boldsymbol{P}_2\end{pmatrix} = \mathbf{0} \\ &\text{and } \begin{pmatrix}\boldsymbol{P}_1 & \boldsymbol{P}_2\end{pmatrix}^\top \begin{pmatrix}\boldsymbol{P}_1 & \boldsymbol{P}_2\end{pmatrix} = \mathbf{I}.\end{aligned}$$

Solving the above problem by adding Bregman penalties leads to an iteration algorithm that solves

$$
\begin{cases}
\boldsymbol{Z}_1^k, \boldsymbol{Z}_2^k, \boldsymbol{P}_1^k, \boldsymbol{P}_2^k = \min_{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{P}_1, \boldsymbol{P}_2} L + \frac{\gamma}{2}\|\boldsymbol{Z}_1 - \boldsymbol{P}_1 + \boldsymbol{B}_1^k\|_F^2 + \frac{\gamma}{2}\|\boldsymbol{Z}_2 - \boldsymbol{P}_2 + \boldsymbol{B}_2^k\|_F^2, \\
\qquad\qquad \text{subject to } \begin{pmatrix} \mathbf{1}_n & \boldsymbol{U}_1 & \boldsymbol{U}_2 \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{P}_1 & \boldsymbol{P}_2 \end{pmatrix} = \mathbf{0} \text{ and } \begin{pmatrix} \boldsymbol{P}_1 & \boldsymbol{P}_2 \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{P}_1 & \boldsymbol{P}_2 \end{pmatrix} = \mathbf{I}, \\
\begin{pmatrix} \boldsymbol{B}_1^k \\ \boldsymbol{B}_2^k \end{pmatrix} = \begin{pmatrix} \boldsymbol{B}_1^{k-1} \\ \boldsymbol{B}_2^{k-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{Z}_1^k \\ \boldsymbol{Z}_2^k \end{pmatrix} - \begin{pmatrix} \boldsymbol{P}_1^k \\ \boldsymbol{P}_2^k \end{pmatrix},
\end{cases}
$$

where γ is a positive tuning parameter. Noticing that the first optimization problem is separable and can be solved by iteratively updating $\boldsymbol{Z}_k$ and $\boldsymbol{P}_k$, $k = 1, 2$, and the iteration algorithm can be further formulated as

$$
\begin{cases}
\boldsymbol{Z}_1^k & = \min_{\boldsymbol{Z}_1} L(\boldsymbol{X}_1|\boldsymbol{\Theta}_1) + \frac{\gamma}{2}\|\boldsymbol{Z}_1 - \boldsymbol{P}_1 + \boldsymbol{B}_1^k\|_F^2 \\
\boldsymbol{Z}_2^k & = \min_{\boldsymbol{Z}_2} L(\boldsymbol{X}_2|\boldsymbol{\Theta}_2) + \frac{\gamma}{2}\|\boldsymbol{Z}_2 - \boldsymbol{P}_2 + \boldsymbol{B}_2^k\|_F^2 \\
\boldsymbol{P}_1^k, \boldsymbol{P}_2^k & = \min_{\boldsymbol{P}_1, \boldsymbol{P}_2} \frac{\gamma}{2}\|\boldsymbol{Z}_1 - \boldsymbol{P}_1 + \boldsymbol{B}_1^k\|_F^2 + \frac{\gamma}{2}\|\boldsymbol{Z}_2 - \boldsymbol{P}_2 + \boldsymbol{B}_2^k\|_F^2, \text{ subject to} \\
& \qquad \begin{pmatrix} \boldsymbol{P}_1 & \boldsymbol{P}_2 \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{P}_1 & \boldsymbol{P}_2 \end{pmatrix} = \mathbf{I} \text{ and } \begin{pmatrix} \mathbf{1}_n & \boldsymbol{U}_1 & \boldsymbol{U}_2 \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{P}_1 & \boldsymbol{P}_2 \end{pmatrix} = \mathbf{0}, \\
\begin{pmatrix} \boldsymbol{B}_1^k \\ \boldsymbol{B}_2^k \end{pmatrix} & = \begin{pmatrix} \boldsymbol{B}_1^{k-1} \\ \boldsymbol{B}_2^{k-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{Z}_1^k \\ \boldsymbol{Z}_2^k \end{pmatrix} - \begin{pmatrix} \boldsymbol{P}_1^k \\ \boldsymbol{P}_2^k \end{pmatrix}.
\end{cases}
$$

The first iteration is convex and can be solved similarly as individual loading matrices by using the damped Newton's method with a proper step size. The second constrained problem has a closed-form solution illustrated in theorem 6.

Theorem 6. *Let $\boldsymbol{U} \in \mathbb{R}^{n \times r}$ be an orthogonal matrix and $\boldsymbol{C} \in \mathbb{R}^n \times p$ be a full-rank matrix. Then the constrained quadratic problem:*

$$
\boldsymbol{P}^* = \arg\min_{\boldsymbol{P}} \|\boldsymbol{P} - \boldsymbol{C}\|_F^2, \textit{ s.t. } \boldsymbol{U}^\top \boldsymbol{P} = \mathbf{0} \ \& \ \boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{I}.
$$

has the following closed-form solution:

$$\boldsymbol{P}^* = \boldsymbol{M}\boldsymbol{I}_{n\times p}\boldsymbol{N}^\top,$$

where $\boldsymbol{M}$ *and* $\boldsymbol{N}$ *are two orthogonal matrices and* $\boldsymbol{D} \in \mathbb{R}^{n\times p}$ *is a diagonal matrix satisfying the SVD factorization* $(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{C} = \boldsymbol{M}\boldsymbol{D}\boldsymbol{N}^\top$.

Proof. We first decompose $\boldsymbol{C}$ into $\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{C}$ and $(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{C}$. Assume the constraint $\boldsymbol{U}^\top\boldsymbol{P} = \boldsymbol{0}$ holds, then the objective function becomes

$$\begin{aligned}\|\boldsymbol{P} - \boldsymbol{C}\|_F^2 &= \left\|\boldsymbol{P} - \left[\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{C} + (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{C}\right]\right\|_F^2 \\ &= \left\|\boldsymbol{P} - (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{C}\right\|_F^2 + \left\|\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{C}\right\|_F^2\end{aligned}$$

Therefore, the constrained quadratic problem is equivalent to the following one

$$\boldsymbol{P}^* = \arg\min_{\boldsymbol{P}} \|\boldsymbol{P} - (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{C}\|_F^2, \text{ s.t. } \boldsymbol{U}^\top\boldsymbol{P} = \boldsymbol{0} \ \ \& \ \ \boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}.$$

Note that the above problem can be relaxed to the following Orthogonal Procrustes problem

$$\widetilde{\boldsymbol{P}} = \arg\min_{\boldsymbol{P}} \|\boldsymbol{P} - (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{C}\|_F^2, \text{ s.t. } \boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}.$$

By the results from Theorem 1 in Lai and Osher (2014) and Manton (2002), we have $\widetilde{\boldsymbol{P}} = \boldsymbol{M}\boldsymbol{I}_{n\times p}\boldsymbol{N}^\top$. Since $\boldsymbol{U}^\top\widetilde{\boldsymbol{P}} = 0$, we have $\boldsymbol{P}^* = \widetilde{\boldsymbol{P}} = \boldsymbol{M}\boldsymbol{I}_{n\times p}\boldsymbol{N}^\top$.

□

Based on the above theorem, the iterating updates leads to Algorithm 2.

We next estimate the joint structures with the individual parts fixed. We again use the damped Newton's method to update $\boldsymbol{\mu}_1, \boldsymbol{V}_1$ and $\boldsymbol{\mu}_2, \boldsymbol{V}_2$. The estimation of $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ is separable, and each sub-problem can be again solved by the Splitting orthogonal constraint algorithm described in Algorithm 2.

Algorithm 2 Splitting orthogonal constraint algorithm for (5.4)

Given: $k = 0$, $\boldsymbol{Z}_1^0$, $\boldsymbol{Z}_2^0$, $\boldsymbol{U} = \begin{pmatrix} \mathbf{1}_n, & \boldsymbol{U}_1, & \boldsymbol{U}_2 \end{pmatrix}$;
$\boldsymbol{P}_1^0 = \boldsymbol{Z}_1^0, \boldsymbol{P}_2^0 = \boldsymbol{Z}_2^0, \boldsymbol{B}_1^0 = \mathbf{0}, \boldsymbol{B}_2^0 = \mathbf{0}$;
while $k \neq k_{max}$ *and 'not converge'* **do**
  $k \leftarrow k + 1$;
  $\boldsymbol{Z}_1^k \leftarrow \min_{\boldsymbol{Z}_1} L(\boldsymbol{X}_1|\boldsymbol{\Theta}_1) + \frac{\gamma}{2}\|\boldsymbol{Z}_1 - \boldsymbol{P}_1^{k-1} + \boldsymbol{B}_1^{k-1}\|_F^2$.
  $\boldsymbol{Z}_2^k \leftarrow \min_{\boldsymbol{Z}_2} L(\boldsymbol{X}_2|\boldsymbol{\Theta}_2) + \frac{\gamma}{2}\|\boldsymbol{Z}_2 - \boldsymbol{P}_2^{k-1} + \boldsymbol{B}_2^{k-1}\|_F^2$.
  Compute SVD of $(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\begin{pmatrix} \boldsymbol{Z}_1^k + \boldsymbol{B}_1^{k-1}, & \boldsymbol{Z}_2^k + \boldsymbol{B}_2^{k-1} \end{pmatrix} = \boldsymbol{M}\boldsymbol{D}\boldsymbol{N}^\top$.
  $\begin{pmatrix} \boldsymbol{P}_1^k, & \boldsymbol{P}_2^k \end{pmatrix} \leftarrow \boldsymbol{M}\boldsymbol{I}\boldsymbol{N}^\top$.
  $\boldsymbol{B}_1^k \leftarrow \boldsymbol{B}_1^{k-1} + \boldsymbol{Z}_1^k - \boldsymbol{P}_1^k$.
  $\boldsymbol{B}_2^k \leftarrow \boldsymbol{B}_2^{k-1} + \boldsymbol{Z}_2^k - \boldsymbol{P}_2^k$.
end

However, the estimated $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ may not satisfy the last regularity condition in Section 5.2.2. Therefore, we further normalize them such that $\boldsymbol{U}_1^\top \boldsymbol{U}_2$ is a diagonal matrix. Denote the estimated $\boldsymbol{U}_k$ and $\boldsymbol{V}_k$ as $\widetilde{\boldsymbol{U}}_k$ and $\widetilde{\boldsymbol{V}}_k$, for $k = 1, 2$. Assume the SVD of $\widetilde{\boldsymbol{U}}_k \widetilde{\boldsymbol{V}}_k^\top$ is

$$\widetilde{\boldsymbol{U}}_k \widetilde{\boldsymbol{V}}_k^\top = \boldsymbol{U}_{\theta k} \boldsymbol{\Lambda}_k \boldsymbol{V}_{\theta k}^\top,$$

where $\boldsymbol{U}_{\theta k}$ and $\boldsymbol{V}_{\theta k}$ are orthogonal matrices. Let the SVD of $\boldsymbol{U}_{\theta 1}^\top \boldsymbol{U}_{\theta 2}$ be

$$\boldsymbol{U}_{\theta 1}^\top \boldsymbol{U}_{\theta 2} = \boldsymbol{\Gamma}_1 \boldsymbol{\Lambda}_{12} \boldsymbol{\Gamma}_2^\top,$$

where $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$ are orthogonal matrices. Further we let

$$\widehat{\boldsymbol{U}}_k = \boldsymbol{U}_{\theta k} \boldsymbol{\Gamma}_k, \quad \widehat{\boldsymbol{V}}_k = \boldsymbol{V}_{\theta k} \boldsymbol{\Lambda}_k \boldsymbol{\Gamma}_k.$$

Since $\widehat{\boldsymbol{U}}_k^\top \widehat{\boldsymbol{U}}_k = \boldsymbol{I}$ and $\widehat{\boldsymbol{U}}_1^\top \widehat{\boldsymbol{U}}_2 = \boldsymbol{\Lambda}_{12}$, the normalized estimators $\widehat{\boldsymbol{U}}_k$ satisfy all the regularity conditions and the likelihood stays the same after the normalization.

In summary, we estimate the joint and individual structures using an alternating method. In each iteration, we either solve a convex problem with the damped Newton's method, or solve a convex problem with orthogonal and linear constraints using the method of splitting orthogonality

Algorithm 3 The alternating method.

Given: $k = 0$, $\boldsymbol{\mu}_1^{(0)}, \boldsymbol{\mu}_2^{(0)}, \boldsymbol{U}_1^{(0)}, \boldsymbol{U}_2^{(0)}, \boldsymbol{V}_1^{(0)}, \boldsymbol{V}_2^{(0)}, \boldsymbol{Z}_1^{(0)}, \boldsymbol{Z}_2^{(0)}, \boldsymbol{A}_1^{(0)}, \boldsymbol{A}_2^{(0)}$;
while *'not converge' and* $k \leq k_{max}$ **do**
  $k \leftarrow k + 1$;
  Update $\boldsymbol{\mu}_1$ and $\boldsymbol{A}_1$ via the damped Newton's method
  Update $\boldsymbol{\mu}_2$ and $\boldsymbol{A}_2$ via the damped Newton's method
  Update $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ via SOC
  Update $\boldsymbol{\mu}_1$ and $\boldsymbol{V}_1$ via the damped Newton's method
  Update $\boldsymbol{\mu}_2$ and $\boldsymbol{V}_2$ via the damped Newton's method
  Update $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ via SOC
  Normalize $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ such that $\boldsymbol{U}_1^\top \boldsymbol{U}_2$ is diagonal.
end

constraints. The detailed estimation approach is given in Algorithm 3. In each iteration, the negative log-likelihood is non-increasing. Since the negative log-likelihood is bounded from below, the algorithm will always converge (including converging to the infinity).

5.4 Simulation studies

In this section, we demonstrate the effectiveness of our model by comparing the performance of low-rank CCA with existing ones in different settings. We consider the following methods: (i) Low-rank CCA, the proposed model; (ii) an ad hoc exponential CCA method, where we first compute the saturated natural parameters, then apply CCA to the estimated natural parameter matrices. We still denote this method by CCA since it reduces to standard CCA in the normal settings. All simulations are implemented using R.

The SOC algorithm in low-rank CCA requires a tuning parameter γ. Intuitively, the user should always choose a large gamma to ensure the equality constraint holds. To avoid complication, we set $\gamma = 10$ in our experiments. In addition, we set the joint and individual ranks (r_0, r_1 and r_2) to be the true ranks. For CCA, we set the ranks of natural parameter matrices to the true ranks by keeping the leading r_1 or r_2 pairs of canonical variables and the corresponding loading vectors only. Consequently, low-rank CCA and CCA yields natural parameter estimations with the same ranks.

5.4.1 Data generation

We generate the data using model (5.2) with the corresponding regularity conditions. More specifically, we set the joint rank $r_0 = 2$ and total ranks $r_1 = r_2 = 4$. We set sample size $n = 150$ and the dimensions of both data matrices to be $p_1 = p_2 = 10$. All elements in $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are generated from a uniform distribution between $(-2, -0.5) \cup (0.5, 2)$. Each matrix is then centered and normalized such that the regularity conditions are satisfied. $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are further normalized such that the canonical correlations are $\rho_1 = 0.8$ and $\rho_2 = 0.6$, that is $\boldsymbol{U}_1^\top \boldsymbol{U}_2 = \text{diag}(0.8, 0.6)$. Next, we generate $\boldsymbol{\mu}_k, \boldsymbol{V}_k$ and $\boldsymbol{A}_k$ in a similar manner for $k = 1, 2$. We consider the following settings for the experiments and repeat the simulations 100 times.

1. Gaussian distribution. We generate the natural parameter matrices as described above, and further add white noise with mean zero to each element. The standard deviation of noises is set to be 0.05 or 0.1. Note that the resulting data sets has standard deviation close to 0.5.

2. Binomial distribution. The natural parameter matrices are generated similarly to the Gaussian case. The size of the binomial distribution is set to 50 or 100, and the observed data matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are generated based on natural parameters using corresponding distributions. The columns of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are further standardized by its size.

5.4.2 Result

We compare the low-rank CCA and ad hoc CCA in terms of estimation accuracy. To compare the estimation accuracy of the joint scores, we consider the subspace difference (Ye and Lim, 2016) between the estimated and the true joint scores ($\boldsymbol{U}_1$ and $\boldsymbol{U}_2$). In particular, the subspace distance is defined as

$$\left\| \boldsymbol{U}_k \boldsymbol{U}_k^\top - \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \right\|_2, k = 1, 2,$$

where $\|\cdot\|_2$ denotes the matrix $2-$norm; $\boldsymbol{U}_k$ denotes the true joint scores and $\widehat{\boldsymbol{U}}_k$ denotes the estimated joint scores. To compare the overall estimation accuracy, we consider the relative error

of the reconstructed natural parameter matrices denoted as

$$\text{relative error} = \frac{\|\boldsymbol{\Theta}_k - \widehat{\boldsymbol{\Theta}}_k\|_F^2}{\|\boldsymbol{\Theta}_k\|_F^2},$$

where $\widehat{\boldsymbol{\Theta}}_k$ represents the estimated natural parameter matrices and $\boldsymbol{\Theta}_k$ denotes the true natural parameter matrices.

Figure 5.1 shows the results where two datasets are both generated from Gaussian distribution. In this case, low-rank CCA is generally better than standard CCA. More specifically, low-rank CCA gives lower subspace distance between true and estimated joint scores, and also has significantly lower relative error with smaller variance. This is not surprising, since the proposed method utilizes the low-rank assumption and explicitly assumes the orthogonality between the individual structures. Consequently, the estimation of individual structures is more accurate.

Figure 5.2 reports the results where distributions are both binomial. The low-rank CCA and ad hoc CCA have similar performance for the estimation of the joint score matrices $\boldsymbol{U}_k$, although the subspace distances are greater than those of the Gaussian case. We conjecture this is because binomial data is in general difficult to model, since the binomial data can be seen as the summation of multiple Bernoulli data, and Bernoulli distribution is known to have convergence issues (Li and Gaynanova, 2018; Collins et al., 2001). On the other hand, the low-rank CCA still shows significantly smaller relative error, which indicates the proposed method outperforms ad hoc CCA in terms of estimating the individual parts.

5.5 Discussion

In this Chapter, we present a low-rank CCA model for the association analysis of two datasets. A unique characteristic of the proposed model is that it imposes orthogonal constraints on the individual score matrices to guarantee no more shared information is rendered in the individual structures. The simulation studies suggest that the proposed method outperforms the classical CCA in the cases of both Gaussian and binomial distributions.

There are many possible future studies for the current proposal. First, the current model does

not consider the similarity between the joint score matrices $\boldsymbol{U}_k$. One possible extension is to allow the user to control the magnitude of the estimated canonical correlations, which are the diagonal elements of $\boldsymbol{U}_1^\top \boldsymbol{U}_2$. This goal can be achieved by considering additional penalties on $\boldsymbol{U}_k$. Specifically, we consider the objective function (5.5) with the same regularity conditions

$$\min_{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2} L = \min_{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2} L(\boldsymbol{X}_1|\boldsymbol{\Theta}_1) + L(\boldsymbol{X}_2|\boldsymbol{\Theta}_2) + \rho \|\boldsymbol{U}_1 - \boldsymbol{U}_2\|_F^2, \tag{5.5}$$

where $\rho > 0$ is a tuning parameter to control the weight of the penalty. We remark that since $\boldsymbol{U}_k$ are orthogonal matrices, adding $\|\boldsymbol{U}_1 - \boldsymbol{U}_2\|_F^2$ to the objective function (5.3) is the same to add $-\operatorname{Tr}(\boldsymbol{U}_1^\top \boldsymbol{U}_2)$.

Another possible extension is to allow the ranks r_0, r_1 and r_2 to be determined in a data-driven way. We adopt a two-step strategy similar to the method introduced in Li and Gaynanova (2018). We first use cross-validation to determine the ranks of the natural parameter matrices of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, which are r_1 and r_2, and then iterate r_0 from 1 to $\min(r_1, r_2)$ to determine its best value. We refer to Li and Gaynanova (2018) for the cross-validation method to determine the rank of a matrix.

Third, the application of the proposed method to the Pan-cancer deconvolution data is promising. We consider clustering the patients based on estimated joint score matrices, and then compare the overall survival probability differences among different clusters of patients. The results will be compared with those by the GAS model.

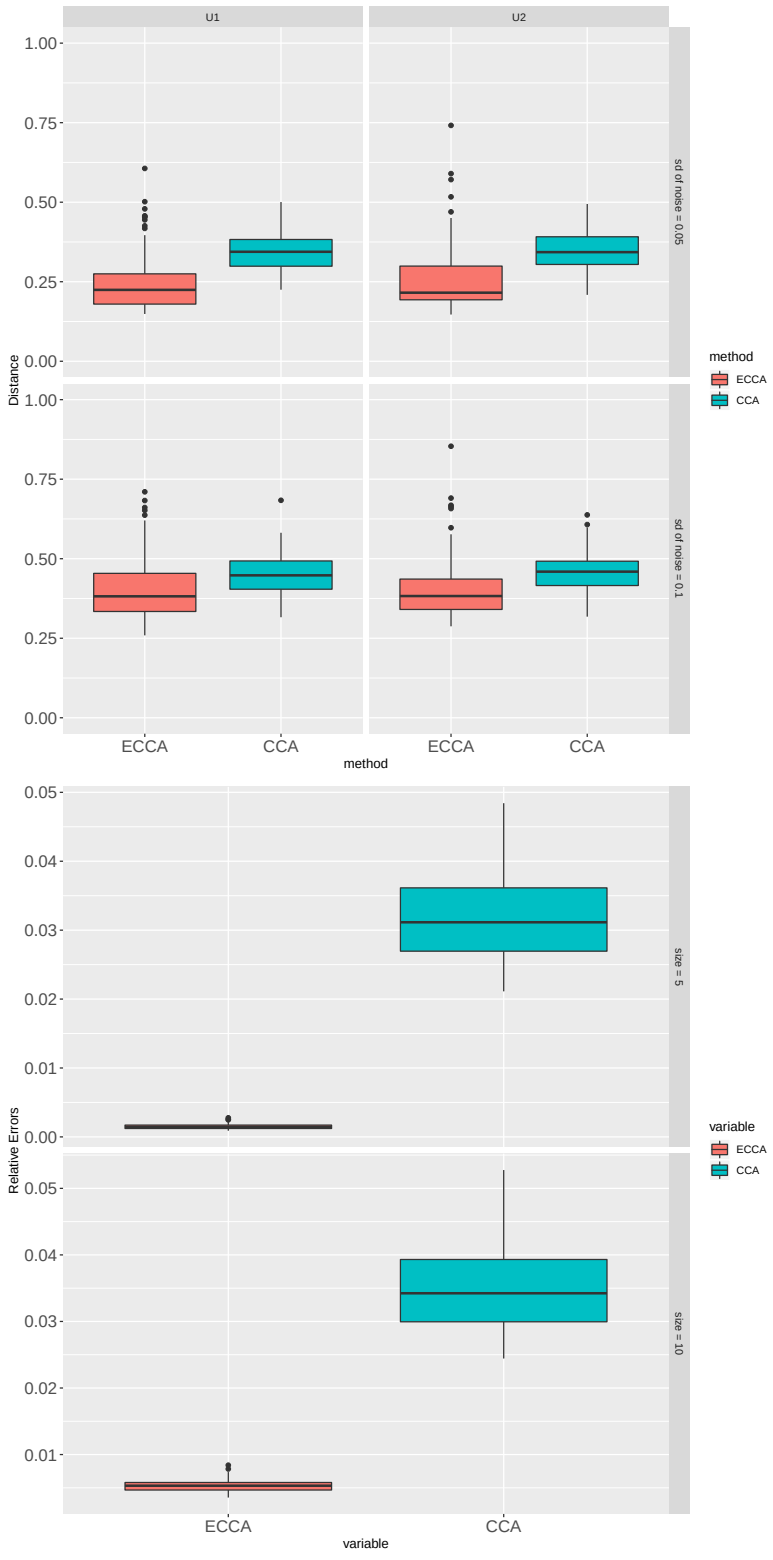

Figure 5.1: Simulation results under the Gaussian setup based on 100 replications. Top: Comparison of subspace difference of joint scores between Low-rank CCA and CCA. Bottom: Comparison of relative error between Low-rank CCA and CCA.

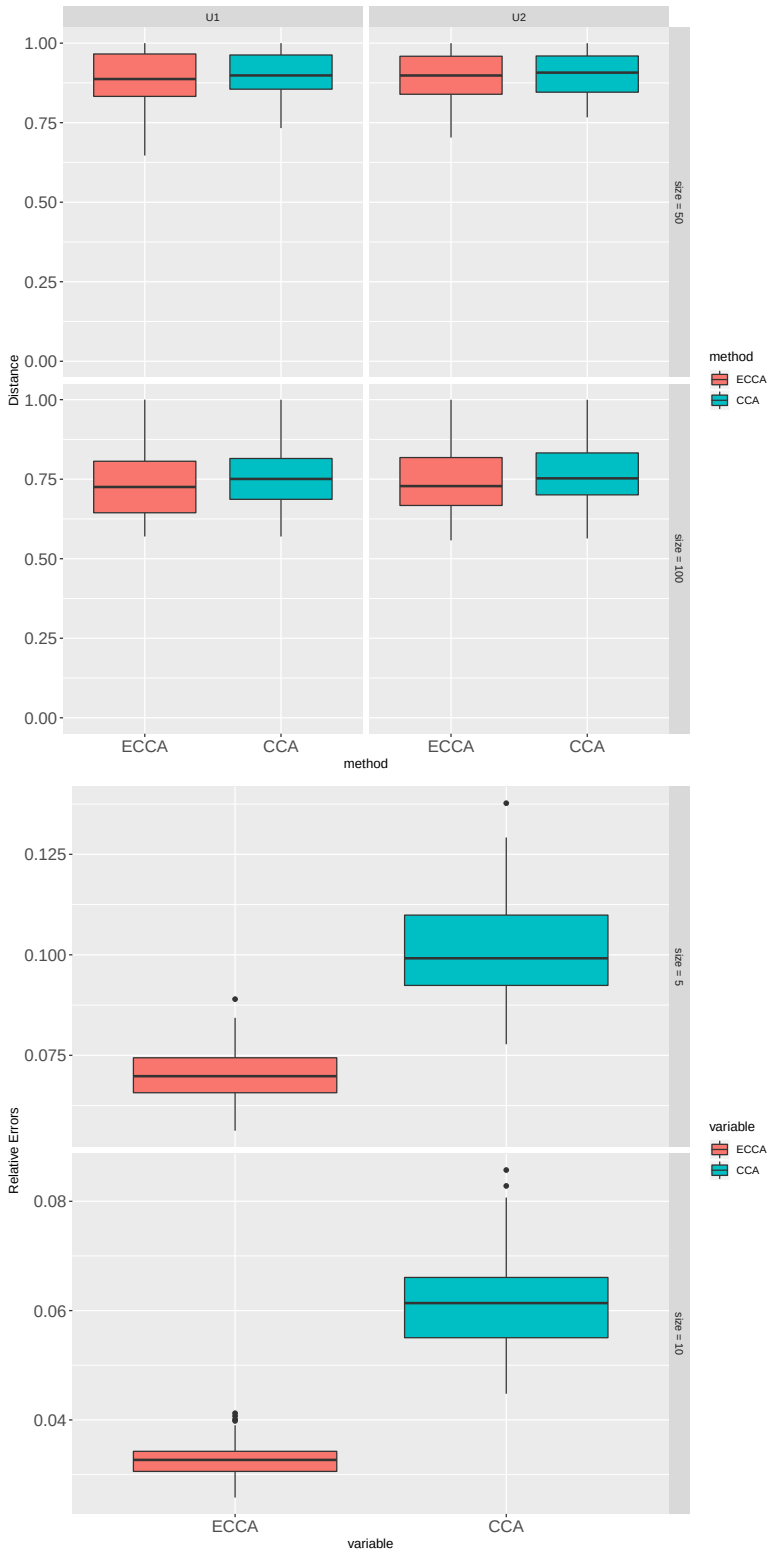

Figure 5.2: Simulation results under the binomial setup based on 100 replications. Top: Comparison of subspace difference of joint scores between Low-rank CCA and CCA. Bottom: Comparison of relative error between Low-rank CCA and CCA.

6. SUMMARY

Modern datasets, such as multi-view data, have become more common with the advancement of scientific technologies. Meanwhile, it also introduces new challenges to statistical inference because the cross-platform datasets are intrinsically correlated. In this dissertation, we investigate the challenges of association analysis of multi-view data under different settings. First, we address the problem of simultaneous classification and association problems for multi-view data. Then we conduct an association analysis between different cellular composition estimations using Pan-Cancer data, and further identify the critical cellular proportions that are predictive to the cancer prognosis. Finally, a low-rank CCA method is proposed to handle both Gaussian and non-Gaussian data.

In Chapter 2, we develop a joint framework for classification and association analysis of multi-view data by exploring the connections between linear discriminant analysis and canonical correlation analysis. An efficient algorithm using block-coordinate descent method is proposed to fit the model, and a corresponding R package is available on GitHub. We support the methodology with numerical comparisons with existing methods. Nevertheless, there are several parts of the method that requires further investigation. First, the trade-off between classification and association criteria in (2.7) is controlled by the parameter α. While we fix $\alpha = 1/2$ for the analysis, it would be of interest to investigate whether there is the optimal value, both from empirical and theoretical perspectives. Secondly, we treat all views equally within our framework; however, in practice, some views may have stronger associations with class membership as well as with each other. This scenario can be addressed by adding view-specific weights within (2.7), however, it is unclear how to choose the weights in practice. Finally, we focused on row-sparse structure via a group-lasso penalty to perform variable selection. However, the method could be used with other structured penalties depending on the problem of interest.

In Chapter 3, we provide theoretical guarantees for estimation consistency in high-dimensional settings for JACA, which are absent from recent joint learning approaches. We demonstrate that

the estimation error of discriminant vectors converges to zero in the same rates as that of group-lass. In addition, we show that a particular advantage of our approach is that it allows us to use both samples with missing class labels and samples with missing subsets of views. And such advantage is demonstrated based on simulation studies and colorectal cancer data from The Cancer Genome Atlas project. However, throughout the chapter, we have assumed that the views are normally distributed conditioned on the responses. One possible extension of JACA is to allow non-Gaussian views, such as composition and binary data. This goal can be achieved by considering the exponential family distributions. Similar ideas have been explored in Collins et al. (2001) and Landgraf and Lee (2019) to extend PCA to non-Gaussian data.

In Chapter 4, we study the associations among different cellular composition estimates by various tools for Pan-cancer data. We compute the association coefficients to assess the strength of association. The common and individual signals are extracted from the estimates for each cancer type, and then being used in clustering and survival analysis to give further insights of the cancer prognosis. In particular, we found that the common signals are informative for survival in PROD and BLCA, but not informative in COLON. For PROD, patients with higher immune and higher normal proportions tend to have longer PFI, whereas, for BLCA, higher immune proportions were associated with shorter PFI. In contrast to PROD and BLCA, the DeMixT proportions are not informative to the prognosis, but the corresponding individual signals of CIBERSORT proportions are strongly associated with survival for COLON. In the next step, it is of interest to understand the relationship between cancer and immune proportions, and investigate how it affects PFI differently for BLCA and PROD.

In Chapter 5, we propose a decomposition-based CCA in the low-rank framework, and extend it to the non-Gaussian settings. In contrast to the other decomposition methods, the proposed method guarantees that all shared information is rendered in the joint structures. We develop an alternating method to estimate the parameters and propose a splitting orthogonal constraint algorithm to solve the orthogonal constrained sub-problems. There are several ways to extend the proposed formulation. In the case of high-dimensional settings, we consider sparse penalty

on the loadings to perform variable selection. Further, the alternating algorithm does not require substantial changes as the corresponding sub-problem is still convex. Finally, given the simulation results, we also expect to use the method to handle other exponential family distributions, such as Bernoulli and Poisson distribution, which is known to be challenging to fit (Li and Gaynanova, 2018).

REFERENCES

Bach, F. R. and Jordan, M. I. (2005). A probabilistic interpretation of canonical correlation analysis.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge Univ Press, Cambridge.

Chen, M., Gao, C., Ren, Z., and Zhou, H. H. (2013). Sparse cca via precision adjusted iterative thresholding. *arXiv preprint arXiv:1311.6186*.

Collins, M., Dasgupta, S., and Schapire, R. (2001). A generalization of principal component analysis to the exponential family advances in neural information processing systems. *TG Dietterich, S. Becker, and Z. Ghahramani (eds)*, 617632.

Fletcher, R. (2013). *Practical methods of optimization*. John Wiley & Sons.

Gao, C., Ma, Z., and Zhou, H. H. (2017). Sparse CCA: Adaptive estimation and computational barriers. *Annals of Statistics*, 45(5):2074–2101.

Gaynanova, I. (2016). *MGSDA: Multi-Group Sparse Discriminant Analysis*. R package version 1.4.

Gaynanova, I. (2019). Prediction and estimation consistency of sparse multi-class penalized optimal scoring. *Bernoulli*, page accepted.

Gaynanova, I., Booth, J. G., and Wells, M. T. (2016). Simultaneous sparse estimation of canonical vectors in the $p >> N$ setting. *Journal of the American Statistical Association*, 111(514):696–706.

Gaynanova, I. and Kolar, M. (2015). Optimal variable selection in multi-group sparse discriminant analysis. *Electronic Journal of Statistics*, 9(2):2007–2034.

Gaynanova, I. and Li, G. (2019). Structural learning and integrative decomposition of multi-view data. *Biometrics*, page accepted.

Gross, S. M. and Tibshirani, R. J. (2015). Collaborative regression. *Biostatistics*, 16(2):326–338.

Guinney, J., Dienstmann, R., Wang, X., De Reyniès, A., Schlicker, A., Soneson, C., Marisa, L.,

Roepman, P., Nyamundanda, G., Angelino, P., et al. (2015). The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, 21(11):1350–1356.

Hastie, T. J., Tibshirani, R. J., and Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, 89(428):1255–1270.

Hastie, T. J., Tibshirani, R. J., and Wainwright, M. J. (2015). *Statistical Learning with Sparsity*. The Lasso and Generalizations. CRC Press.

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 27(4).

Klami, A., Virtanen, S., and Kaski, S. (2010). Bayesian exponential family projections for coupled data sources. *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (2010)*, pages 286–293.

Lai, R. and Osher, S. (2014). A splitting method for orthogonality constrained problems. *Journal of Scientific Computing*, 58(2):431–449.

Landgraf, A. J. and Lee, Y. (2015). Dimensionality Reduction for Binary Data through the Projection of Natural Parameters. *arXiv preprint arXiv:1510.06112*.

Landgraf, A. J. and Lee, Y. (2019). Generalized principal component analysis: Projection of saturated model parameters. *Technometrics*, pages 1–14.

Li, G. and Gaynanova, I. (2018). A general framework for association analysis of heterogeneous data. *Annals of Applied Statistics*, 12(3):1700–1726.

Li, G. and Jung, S. (2017). Incorporating covariates into integrated factor analysis of multi-view data. *Biometrics*, 73(4):1433–1442.

Li, G., Yang, D., Nobel, A. B., and Shen, H. (2016). Supervised singular value decomposition and its asymptotic properties. *Journal of Multivariate Analysis*, 146:7–17.

Li, T., Fan, J., Wang, B., Traugh, N., Chen, Q., Liu, J. S., Li, B., and Liu, X. S. (2017). Timer: a web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer research*, 77(21):e108–e110.

Li, Y. and Jia, J. (2017). L1 least squares for sparse high-dimensional LDA. *Electronic Journal of Statistics*, 11(1):2499–2518.

Lock, E. F. and Dunson, D. B. (2013). Bayesian consensus clustering. *Bioinformatics*, 29(20):2610–2616.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Annals of Applied Statistics*, 7(1):523–542.

Lounici, K., Pontil, M., Van De Geer, S. A., and Tsybakov, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics*, 39(4):2164–2204.

Luo, C. and Chen, K. (2017). *CVR: Canonical Variate Regression*. R package version 0.1.1.

Luo, C., Liu, J., Dey, D. K., and Chen, K. (2016). Canonical variate regression. *Biostatistics*, 17(3):468–483.

Manton, J. H. (2002). Optimization algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing*, 50(3):635–650.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press Inc.

Nardi, Y. and Rinaldo, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2(0):605–633.

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., and Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453–457.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Obozinski, G., Wainwright, M. J., and Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *Annals of Statistics*, 39(1):1–47.

OConnell, M. J. and Lock, E. F. (2016). R. jive for exploration of multi-source molecular data. *Bioinformatics*, 32(18):2877–2879.

Podosinnikova, A., Bach, F., and Lacoste-Julien, S. (2016). Beyond cca: Moment matching for multi-view models. *arXiv preprint arXiv:1602.09013*.

Robert, P. and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: The rv- coefficient. *Journal of the Royal Statistical Society, Ser. C*, 25(3):257–265.

Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447.

Searle, S. R. (2006). *Linear Models for Unbalanced Data*. Wiley-Interscience.

Shu, H., Wang, X., and Zhu, H. (2019). D-cca: A decomposition-based canonical correlation analysis for high-dimensional datasets. *Journal of the American Statistical Association*, pages 1–29.

The Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61.

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

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, pages 210–268. Cambridge Univ. Press, Cambridge.

Wan, Y. W., Allen, G. I., and Liu, Z. (2015). TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. *Bioinformatics*, 32(6):952–954.

Wang, Z., Cao, S., Morris, J. S., Ahn, J., Liu, R., Tyekucheva, S., Gao, F., Li, B., Lu, W., Tang, X., et al. (2018). Transcriptome deconvolution of heterogeneous tumor samples with immune infiltration. *iScience*, 9:451–460.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Cancer Genome Atlas Research Network, et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113.

Wen, Z. and Yin, W. (2013). A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434.

Witten, D., Tibshirani, R., Gross, S., and Narasimhan, B. (2013). *PMA: Penalized Multivariate Analysis*. R package version 1.0.9.

Witten, D. M. and Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis

with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–27.

Witten, D. M., Tibshirani, R. J., and Hastie, T. J. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.

Ye, K. and Lim, L.-H. (2016). Schubert varieties and distances between subspaces of different dimensions. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1176–1197.

Yin, W., Osher, S., Goldfarb, D., and Darbon, J. (2008). Bregman iterative algorithms for \ell_1-minimization with applications to compressed sensing. *SIAM Journal on Imaging sciences*, 1(1):143–168.

Zhao, J., Xie, X., Xu, X., and Sun, S. (2017). Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54.

Zou, H. and Hastie, T. J. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Ser. B*, 67(2):301–320.