

Sparse Canonical Correlation Analysis (SCCA):
A Comparative Study

SPARSE CANONICAL CORRELATION ANALYSIS (SCCA):
A COMPARATIVE STUDY

BY
SATHISH CHANDRA PICHKA, M.Sc.

A THESIS
SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS
AND THE SCHOOL OF GRADUATE STUDIES
OF MCMASTER UNIVERSITY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

© Copyright by Sathish Chandra Pichika, December 2011

All Rights Reserved

Master of Science (2011)
(Mathematics & Statistics)

McMaster University
Hamilton, Ontario, Canada

TITLE: Sparse Canonical Correlation Analysis (SCCA):
A Comparative Study

AUTHOR: Sathish Chandra Pichika
M.Sc., (Statistics)
Loyola College, India

SUPERVISOR: Dr. Joseph Beyene

NUMBER OF PAGES: xii, 63

I dedicate my thesis to my parents

Abstract

Canonical Correlation Analysis (CCA) is one of the multivariate statistical methods that can be used to find relationship between two sets of variables. I highlighted challenges in analyzing high-dimensional data with CCA. Recently, Sparse CCA (SCCA) methods have been proposed to identify sparse linear combinations of two sets of variables with maximal correlation in the context of high-dimensional data. In my thesis, I compared three different SCCA approaches. I evaluated the three approaches as well as the classical CCA on simulated datasets and illustrated the methods with publicly available genomic and proteomic datasets.

Acknowledgements

I thank the many people who made this thesis possible.

It is difficult to overstate my gratitude to my supervisor, Dr. Joseph Beyene. With his enthusiasm, his inspiration, and his great efforts to explain things clearly and simply, he helped to make statistics fun for me. Throughout my thesis-writing period, he provided encouragement, sound advice, good teaching, good company, and lots of good ideas. I would have been lost without him. More importantly, I like to thank Dr. N. Balakrishnan for his constant mentoring and providing me with sound advice about my thesis.

I would like to thank the many people who have taught me mathematics and statistics: my high school math teachers (especially Karuppaiya), my undergraduate teachers at Chennai (especially B. Chandraseker, Leo Alexander, Jerome Stanley Davis, V. Vaidyanadhan, S. Santharaman, Thobias, and my every lovable Academic advisor Martin Luther Williams), and my graduate teachers (especially Angelo Canty, Fred Hoppe, Sonia Anand, and Aaron Childs). For their kind assistance with writing letters, giving wise advice, helping with various applications, and so on, I wish to thank in addition Yoan Gerrard and Deepika Desai.

I am indebted to my many student colleagues for providing a stimulating and fun environment in which to learn and grow. I am especially grateful to Ram, Karthikeyan at Chennai, and to Ashley Bonner, Xiao de yang, Hong, Tracey, Suvra, Debanjan at McMaster. Angelo Canty was particularly helpful mathematically and computationally, patiently teaching me the R coding and running simulation studies.

I wish to thank my best friend since my childhood (Macharla Thireesha), my best friend as a undergraduate student (Ganesh), and my best friend as a graduate student (Ashley Bonner), for helping me get through the difficult times, and for all the emotional support, camaraderie, entertainment, and caring they provided.

I am grateful to the secretaries in the math departments of McMaster, for helping the departments to run smoothly and for assisting me in many different ways.

I wish to thank my entire extended family for providing a loving environment for me. My sister, my half-siblings, my brother-in-common-law, some uncles, and some first-cousins-once-removed were particularly supportive.

Lastly, and most importantly, I wish to thank my parents, Pichika Sachindrapal and Padmaja. They bore me, raised me, supported me, taught me, and loved me. To them I dedicate this thesis.

Contents

Abstract	iv
Acknowledgements	v
1 Introduction	1
1.1 Recent Advances in High Dimensional Data	1
1.2 Organization of the thesis	4
2 Methods	5
2.1 PCA and its Extensions	5
2.2 Canonical Correlation Analysis	6
2.3 Relationship with other Linear Methods	10
3 SCCA Literature Review	12
3.1 Sparse Canonical Correlation Analysis	13
3.1.1 SCCA Approach of Parkhomenko et al. (2009)	15
3.1.2 SCCA Approach of Witten et al. (2009)	19
3.1.3 SCCA Approach of Lee et al. (2011)	22

4	Simulation Studies	26
4.1	Introduction	26
4.2	Simulation Strategies	31
4.3	Simulation Results	33
4.4	Key Finding	45
5	Real Data Application of SCCA	47
5.1	Background	47
5.2	Data Description	48
5.3	Normalization	49
5.4	Results	50
5.4.1	Result 1	50
5.4.2	Result 2	53
6	Discussion and Future directions	55
6.1	Discussion	55
6.1.1	Simulation studies limitations	56
6.2	Future directions	57

List of Tables

4.1	Summary of Parameter in Simulation Studies	32
4.2	$n = 100, p = 300, q = 200, r = (10, 20, 40), \sigma_\mu = 1.8, \sigma_e = 0.1$, WT = Witten et al. (2009), LT = Lee et al. (2011), PT = Parkhomenko et al. (2009), CCA = Classical Canonical Correlation Analysis. FPN = Number of false positives and FNN = Number of false negatives . . .	34
4.3	$n = 200, p = 300, q = 200, r = (10, 20, 40), \sigma_\mu = 1.8, \sigma_e = 0.1$	35
4.4	$n = 100, p = 1000, q = 800, r = (10, 20, 40), \sigma_\mu = 1.8, \sigma_e = 0.1$	36
4.5	$n = 200, p = 1000, q = 800, r = (10, 20, 40), \sigma_\mu = 1.8, \sigma_e = 0.1$	37
4.6	$n = 100, p = 200, q = 150, r = (10, 20), \sigma_\mu = 2, \sigma_e = 0.1$	38
4.7	$n = 200, p = 200, q = 150, r = (10, 20), \sigma_\mu = 2, \sigma_e = 0.1$	38
4.8	$n = 30, p = 300, q = 200, r = 20, \sigma_\mu = 1.8, 2, \sigma_e = 0.1$	38
4.9	$n = (30, 100, 500), p = 50, q = 40, r = 5, \sigma_\mu = 2, \sigma_e = 0.1$	39
4.10	$n = 100, p = 50, q = 30, r = 5, \sigma_\mu = 1.8, 2.5, 3, \sigma_e = 0.3$	40
4.11	$n = 100, p = 500, q = 300, r = 15, \sigma_\mu = 1.8, 2, \sigma_e = 0.3$	40
4.12	$n = 100, p = 300, q = 200, r = (5, 15), \sigma_\mu = 3, \sigma_e = 0.4$	43
4.13	$n = 300, p = 100, q = 50, r = (5, 15), \sigma_\mu = 3, \sigma_e = 0.4$	43

4.14	$n = 50, p = 100, q = 80, r = (5, 15), \sigma_\mu = 3, \sigma_e = 0.5$	43
4.15	$n = 50, p = 100, q = 80, r = (5, 15), \sigma_\mu = 4, \sigma_e = 0.5$	44
4.16	$n = 100, p = 2000, q = 1500, r = 40, \sigma_\mu = 3, \sigma_e = 0.4$	44
4.17	$n = 500, p = 150, q = 100, r = (5, 15), \sigma_\mu = (3, 4), \sigma_e = 0.5$	44
5.1	Overview of mRNA expression and protein data sets	49
5.2	Summary of results of three different approaches of SCCA; CCA - Classical Canonical Correlation Analysis, LT - Lee et al. (2011) method, WT - Witten et al. (2009), PT - Parkhomenko et al. (2009)		51
5.3	Summary of results of three different approaches of SCCA	53

List of Figures

4.1	Plotting number of false positives of both parameters with different σ_μ and sample size = 100, $p = 50$, $q = 30$, $r = 5$ and $\sigma_e = 0.3$. with all the methods	39
4.2	Plotting number of false negatives of both parameters with different σ_μ and sample size = 100, $p = 50$, $q = 30$, $r = 5$ and $\sigma_e = 0.3$. considering Witten et al. Approach to SCCA	41
4.3	Plotting number of false negatives of both parameters with different sample sizes between 30 and 500, $p = 50$, $q = 40$, $r = 5$, $\sigma_\mu = 2.5$ and $\sigma_e = 0.1$. considering only Witten et al. Approach to SCCA	41
4.4	Plotting number of false positives of both parameters with varying nuisance standard deviation, $p = 50$, $q = 40$, $r = 5$, $\sigma_\mu = 2$ and $n = 100$. consider all Approaches to SCCA	42
4.5	Plotting number of false negatives of both parameters with varying nuisance standard deviation, $p = 50$, $q = 40$, $r = 5$, $\sigma_\mu = 2$ and $n = 100$. consider all Approaches to SCCA	45

4.6	Plotting distance between estimates and true value of estimates of both parameters with varying nuisance standard deviation, $p = 50$, $q = 40$, $r = 5$, $\sigma_\mu = 2$ and $n = 100$. consider all Approaches to SCCA	45
5.1	Central Drogma of Cell Biology	48
5.2	The cross-validated sample canonical correlation and the number zeros considering LT penalty function of the tuning parameters, λ_a ranges from 100 to 500	51
5.3	The canonical correlation of WT penalty function of the tuning parameters, λ_a and λ_a ranging from 0 and 0.7	52
5.4	The cross-validated sample canonical correlation and the number zeros considering LT penalty function of the tuning parameters, λ_a ranging from 100 and 500	53
5.5	The canonical correlation of WT penalty function of the tuning parameters, λ_a and λ_a ranging from 0 and 0.7	54

Chapter 1

Introduction

1.1 Recent Advances in High Dimensional Data

Due to advances in understanding complexities arising from high-dimensional data over the past few decades, more robust statistical and computational methods have been proposed to analyse such data. Novel modern approaches help us in more efficient ways of interpreting the data and one such approach is sparseness. Let us start with review of some classic multivariate statistical methods that can be used with high-dimensional data.

Partial Least Square (PLS) is a method that combines and generalizes features from multiple regression and Principal Component Analysis (PCA). It is most commonly used when we wish to predict a set of dependent variables from a large set of independent variables. It was introduced by H. Wold (1966) in social science studies.

Principal Component Analysis (PCA) is a multivariate statistical method that

helps in reducing the dimension of data. It produces a smaller number of uncorrelated variables known as principal components which are obtained from transforming a number of possible correlated variables. It was proposed by Karl Pearson (1901).

Linear Discriminant Analysis (LDA) is a commonly used method for data classification and reduction and is closely related to PCA. This method maximizes the ratio of between-class variance to the within-class variance in any particular data set, thereby guaranteeing maximal separability (Fisher 1936).

Canonical Correlation Analysis (CCA) is one of the classical methods used to find correlation between linear combinations of two sets of variables. It was proposed by H. Hotelling (1936). It helps in finding pairs of linear combinations between two sets of variables that are maximally correlated where the two sets of variables come from same or different measurement.

In some of the recent literatures in genomic studies the main emphasis is on finding relationships between two or more view on the same data as the analyses motivate researchers to attain good understanding of their biological relationship (Le Cao et al. 2009). The relationship between two sets of variables can be studied one at a time but it leads to high false-discovery rates and is biologically not interpretable (Parkhomenko et al. 2007).

The linear combinations in PCA involve all of the variables in a dataset. In high-dimensional data, where we have more than a thousand variables in a set, interpreting each principal component which is linear combination of all variables is very difficult. Therefore a better interpretation of data is required via sparseness where we will force some of the variables to zero that don't have any significant

effect in the solution and end up considering only a few variables. Zou et al. (2006) first introduced Sparse Principal Component Analysis (SPCA). It considers all of the variables in a dataset but forces some to 0, therefore resulting in a smaller set of variables. SPCA produces principal components with fewer variables which helps in better biological interpret-ability.

For cases when we have two sets of variables we cannot use any of the above methods. CCA enables us to find linear combinations of two variable sets say X and Y that are maximally correlated with each other. However, in the case of high-dimensional data there is an issue of interpreting the canonical variate. The classical CCA is a linear combinations of all the variables present in X and Y . Researchers might be interested in only fewer variables in each set that have an important effect in obtaining those linear combinations which result in maximal correlation.

Let us consider that there are 1000 variables in X and 1500 variables in Y , which is a common case when looking at gene expression and proteomic data. To find the canonical variates is not very difficult but to interpret them is not so easy as some of the variables present in each set might be attributed to noise. Hence a modified approach to Canonical Correlation Analysis is necessary to analyse the data. Recently, Parkhomenko et al. (2009) proposed a novel way of interpreting data with the above problem, namely Sparse Canonical Correlation Analysis (SCCA). Later on Witten et al. (2009) provided a method which helps us to find canonical correlations for two or more sets of variables. Lee et al. (2011) introduced another approach to SCCA with a different penalty function.

In recent years, there have been a few important methodological advances in

CCA that helps in addressing the key issues that have been outlined above. But only limited comparative work has been done to evaluate some of the approaches to SCCA methods. Hence in my thesis I was interested in comparing three novel methods with different penalty and optimization algorithms, proposed by Parkhomenko et al. (2009), Witten et al. (2009), and Lee et al. (2011). I present simulation studies to compare these methods and I provide a real data application to gene expression and proteomic data. There have been few other SCCA proposed algorithms and penalty functions in the field of machine learning, for example Hardoon et al. (2007) but I will focus on comparing the above three methods.

1.2 Organization of the thesis

In Chapter 2, I discuss principal component analysis and canonical correlation analysis: How is CCA more valuable than ordinary correlation and PCA?

In Chapter 3, I begin by reviewing a few approaches to SCCA: How sparseness makes a significant effect in the overall results. I constructed a literature review of the methods proposed by Parkhomenko et al. (2009), Witten et al. (2009) and Lee et al. (2011).

In Chapter 4, I evaluate how effective the different methods work under different scenarios and their algorithms using simulation studies.

In Chapter 5, I use a real dataset to illustrate the three approaches.

The thesis finishes with discussion about the SCCA methods in brief and my future research interests.

Chapter 2

Methods

This chapter starts with an introduction to Principal Component Analysis (PCA) and proceeds with some background to classical Canonical Correlation Analysis (CCA) and discusses key issues when we have high dimensional data.

2.1 PCA and its Extensions

One of the classical methods that are available for dimension reduction of a dataset is Principal Component Analysis (PCA), in which we obtain linear combinations of the variables using variance-covariance structure that have maximum variance. But there are few restriction when we use PCA. One of them is it deals with only one set of variables. The most important drawback of PCA is that each principal component is a linear combination of all variables in the dataset which ends up being very difficult to interpret.

In recent literature, PCA has become popular as a method of analysing high dimensional genomic data but interpreting the principal components which are linear combinations of all the variables in the dataset is very difficult. There were few methods proposed to address this issues. One among them is Sparse Principal Component Analysis (SPCA) where only a sparse set of parameters are included in each principal component for better interpretation. Elastic net (Zou and Hastie 2005) is one approach by which sparse principal component can be obtained. The main idea of these methods is to view the original problem as a regression problem and use a penalty function to find sparseness.

In genomic studies, for example, biologists may be interested in comparing two sets of variables or finding a relation between two datasets coming from the same or different subjects (e.g. Parkhomenko et al. (2009)). Hence using PCA or SPCA will not help us in finding the relationship between two sets of variables. Instead, Canonical correlation analysis (CCA) can be used to identify and quantify the linear relationship between two sets of variables.

2.2 Canonical Correlation Analysis

CCA helps in finding the relationship between two sets of variables. Despite lacking popularity in early introduction, the rapid growth in the statistical software packages has improved the use of CCA rather more effectively and efficiently. To explain CCA, Hotelling (1936) provided an example of relating arithmetic speed and arithmetic power with respect to reading speed and reading power of seventh-grade

children received on their four tests. One of the main properties of CCA is that the affine transformation of variables is invariant (Borga et al. 1997). This property differentiates CCA from ordinary correlation analysis where the latter highly depends on the variables that are specified.

Compared to ordinary correlation, we can see that canonical correlation squared is the percent of variance between two set of variables. Adding to that, Canonical correlation tells us how strong the relationship is between two sets of variables. It also helps in finding how many dimensions are required for the relationship (Bhatia 2007). The main aim of CCA is to find those pairs of linear combinations which lead to the highest correlation. The pairs of linear combinations are called the canonical variables, and the correlation between them are called canonical correlations. The linear combination is extracted and it will be repeated with the residual data to find the second linear combination that is uncorrelated with the first set of variables. As many linear combinations are produced until an extracted linear combination is not significant.

Canonical Correlation Analysis is part of the multiple general linear hypothesis family and also shares most of the assumptions of multiple regression such as linearity of relationships, interval or near interval data, homoscedasticity, lack of multicollinearity, proper specification of model, and multivariate normality for purposes of hypothesis testing (Bhatia 2007). Often it can be noticed that researchers come across problems due to high dimensional data where they are unable to understand the dependency structure, how to do dimension reduction, or how to construct a subset of good predictors from the variables.

Our interest is to find association between two different sets of variables from same observations. Let there be n observations for both sets of variables. The first set has p variables and it is represented by a $(n \times p)$ matrix, X . The second set has q variables and it is represented by a $(n \times q)$ matrix, Y . For convenience and computational purpose, let us assume that $p \leq q$.

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \quad Y = \begin{pmatrix} y_{11} & \cdots & y_{1q} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nq} \end{pmatrix},$$

where the mean and variance of X and Y are written as μ_x , Σ_{xx} and μ_y , Σ_{yy} , respectively. The total covariance matrix between X and Y can be written as follows:

$$Cov(X, Y) = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix},$$

where μ_x and μ_y are assumed to be 0, and Σ_{xx} and Σ_{yy} are variance-covariance matrices of X and Y , respectively, and $\Sigma_{xy} = \Sigma_{yx}^T$ is the covariance matrix between X and Y . Let A and B be the corresponding linear combinations of X and Y , given by

$$A = Xa, \quad B = Yb,$$

where a and b are some pair of coefficient vectors which are $(p \times 1)$ and $(q \times 1)$ respectively. Also $Var(A) = a^T \Sigma_{xx} a$, $Var(B) = b^T \Sigma_{yy} b$ and $Cov(A, B) = a^T \Sigma_{xy} b$. We choose a and b such that $max_{a,b} Cor(A, B) = a^T \Sigma_{xy} b = \rho_1$ subject to constraints $a^T \Sigma_{xx} a = 1$ and $b^T \Sigma_{yy} b = 1$ (Johnson and Wichern (2007)).

The canonical correlation between X and Y is obtained by solving the eigenvalue equations

$$\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}a = \rho^2a \quad (2.1)$$

and

$$\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}b = \rho^2b \quad (2.2)$$

where eigenvalues ρ^2 are the squared canonical correlation coefficients and the eigenvectors a and b are normalized basis vector. Here, we can notice that the two matrices $\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$ and $\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$ will have same eigenvalues but different eigenvectors. The total number of non-zero solutions to the above mentioned equations will be less than or equal to the smallest dimensions of X and Y i.e., if there are 4 variables in X and 5 variables in Y then the maximum number of canonical correlation coefficients is 4.

It is enough to solve either equation (2.1) or (2.2) as they are related to each other by

$$\Sigma_{xy}b = \rho\lambda_x\Sigma_{xx}a \quad (2.3)$$

$$\Sigma_{yx}a = \rho\lambda_y\Sigma_{yy}b \quad (2.4)$$

where $\lambda_x = \sqrt{\frac{b^T\Sigma_{yy}b}{a^T\Sigma_{xx}a}}$ and $\lambda_y = \lambda_x^{-1}$ (Borga et al. 1997).

The covariance between variables in one set with variables in the other set is represented by Σ_{xy} or Σ_{yx} , i.e., Σ_{xy} represents the relationship between X and Y which contains pq elements which is replaced by $\min(p, q)$ canonical correlations that concisely explains the relationship between X and Y . It can be noticed that when p

and q are large it is very difficult to interpret or to compute Σ_{xy} . Also Σ_{xx}^{-1} , Σ_{yy}^{-1} may not exist, This situations arise when the number of variables in each set is larger than the number of observations and when there is multicollinearity.

Due to the high volume of data available in recent years, using the classical methods for analysing such data is not appropriate and interpreting the data is biologically not feasible. Hence modified approaches to existing classical methods are highly recommended to best use the data. In my thesis, I compare three modified CCA approaches.

2.3 Relationship with other Linear Methods

The two eigenvalue equations (2.3) and (2.4) can be written together as a one eigenvalue equation:

$$C^{-1}D\hat{e} = \rho\hat{e},$$

$$\text{where } C = \begin{bmatrix} \Sigma_{xx} & 0 \\ 0 & \Sigma_{yy} \end{bmatrix}, D = \begin{bmatrix} 0 & \Sigma_{xy} \\ \Sigma_{yx} & 0 \end{bmatrix} \text{ and } \hat{e} = \begin{bmatrix} \lambda_x a \\ \lambda_y b \end{bmatrix} \text{ (Borga et al. 1997).}$$

When we solve for the eigenvalue in the above equation with slightly different matrices, we will get solution to PCA, PLS and multivariate linear regression (MLR).

The matrices are as follows:

Table 2.1: Matrices C and D for PCA, PLS, CCA and MLR

	D	C
PCA	Σ_{xx}	I
PLS	$\begin{pmatrix} 0 & \Sigma_{xy} \\ \Sigma_{yx} & 0 \end{pmatrix}$	$\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$
CCA	$\begin{pmatrix} 0 & \Sigma_{xy} \\ \Sigma_{yx} & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & \Sigma_{xy} \\ \Sigma_{yx} & 0 \end{pmatrix}$
MLR	$\begin{pmatrix} 0 & \Sigma_{xy} \\ \Sigma_{yx} & 0 \end{pmatrix}$	$\begin{pmatrix} \Sigma_{xx} & 0 \\ 0 & I \end{pmatrix}$

Chapter 3

SCCA Literature Review

As the number of statistical softwares increased, the usage and capability of some classical multivariate methods such as Discriminant Analysis, Principle Component Analysis (PCA) and Canonical Correlation Analysis (CCA) have become more simple and efficient. In analysing high dimensional data with one set of variables we could use PCA, but it is limited when we have $p \gg n$. This problem has been addressed with Sparse Principle Component Analysis [Zou et al. (2006), Johnstone Lu (2004)].

With a growing number of large scale genomic data the focus these days have been in finding the relationship between two or more sets of variables. One of the classical methods that can be used in cases when we have two set of variables from the same subject is CCA but it lacks biological interpretation for situations in which each set of variables has more than thousands of variables. This issue was first addressed by Parkhomenko et al. (2009) who proposed a novel method for Sparse Canonical Correlation Analysis (SCCA).

Recently there are a few other proposed methods to find relationship between two sets of variables based on different penalty functions but there are very few comparative studies that have been done so far. A few of the proposed methods are Waaijenborg et al. (2008) who used SCCA to find relationships between the effect of copy number alterations on gene expression and progression of glioma, Witten and Tibshirani (2009) used SCCA to find association between gene expression and array comparative genome hybridization (CGH) measurements, Parkhomenko et al. (2009) and Waaijenborg et al. (2009) used SCCA technique to find correlation between Single-nucleotide polymorphism (SNP) and gene expression data, and Lee et al. (2011) used SCCA approach to find association between gene expression and proteomic data.

3.1 Sparse Canonical Correlation Analysis

SCCA approach is an extension to classical canonical correlation analysis. Despite both methods focusing on finding correlation between two sets of variables, classical CCA involves all the variables in both the data sets, whereas SCCA results in fewer highly significant variables. Hence, the results from SCCA are more robust compared to that of CCA. Also the classical methods which contain the entire variables from both sets lack biological interpretation, and are statistically prone to high false-discovery rate (Parkhomenko et al. 2007).

SCCA was first introduced by Parkhomenko et al. (2009) in which a sparseness parameter controls how many variables will be included from each data set. Later it

was extended by others by adjusting the penalty function; some of these extensions are explained later in this chapter.

Let us consider that there are p and q variables in the sets X and Y , where X and Y are measured on n individuals.

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \quad Y = \begin{pmatrix} y_{11} & \cdots & y_{1q} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nq} \end{pmatrix}$$

Consider the linear combinations $A = Xa$ and $B = Yb$ for some a and b which are p and q elements respectively, then we could write $Var(A) = a^T \Sigma_{xx} a$ and $Var(B) = b^T \Sigma_{yy} b$ and $Cov(A, B) = a^T \Sigma_{xy} b$. We choose a and b in such a way that

$$\max_{a,b} a^T \Sigma_{xy} b \text{ subject to constraint } a^T \Sigma_{xx} a = 1 \text{ and } b^T \Sigma_{yy} b = 1$$

which is the CCA proposed by H. Hotelling (1936). The above criteria requires finding eigenvalues and eigenvectors which has been explained in the previous chapter. In high-dimensional data it is not feasible to use a and b as they are not sparse and, when $p(q) \gg n$, the eigenvectors are not unique. In genomic data, where $p(q) \gg n$ we are interested in only few variables to be included for better understanding of the data. Hence the SCCA criterion is obtained by adding some penalty function as follows.

$$\max_{a,b} a^T \Sigma_{xy} b \text{ subject to constraints } a^T \Sigma_{xx} a \leq 1, b^T \Sigma_{yy} b \leq 1, P_1(a) \leq c_1, P_2(b) \leq c_2,$$

where P_1 and P_2 are penalty functions and they are chosen such that sparse a and b are found.

3.1.1 SCCA Approach of Parkhomenko et al. (2009)

Just like the classical CCA, technique SCCA works with singular value decomposition(SVD) approach. The solution to the SCCA criterion is obtained by solving the Singular Value Decomposition (SVD) of matrix, say K ,

$$K = \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2} = U D V^T, \quad (3.1)$$

where $U = (u_1, u_2, \dots, u_k)$ and $V = (v_1, v_2, \dots, v_k)$, k is the rank of the matrix K . The solution is the rank 1 approximation to $(p \times q)$ matrix K . The canonical vectors, which is the linear combinations of the two sets of variables that have the largest correlation are given by

$$a = \Sigma_{xx}^{-1/2} u_1 \text{ and } b = \Sigma_{yy}^{-1/2} v_1,$$

where u_1 and v_1 are the first canonical variate pair. Thus, the variables obtained $A = Xa$ and $B = Yb$ based on n observations of X and Y are called canonical variables (Mardia et al. 1979), or latent variables (Wegelin 2000).

Instead of considering one pair at a time from two sets of variables Parkhomenko et al. (2009) came up with a more enlightening method. They considered the original variables as composite latent variables corresponding to different sets of measurements and found relationships between the latent variables. For example, the latent variables show joint functionality of the specified essential genes in genomic studies. The sparse linear combinations are obtained from the sparse singular value decomposition of the matrix K , as mentioned earlier, thus the canonical vectors a

and b mentioned above have sparse loadings.

In addition, they proposed an algorithm that performs variable selection before applying SCCA and were able to find approximations to the left and right singular vectors of the SVD using iterative soft-threshold for feature selection. The approach provided by Parkhomenko et al. (2009) is similar to that of SVD algorithm of I.J. Good (1969), PLS by H. Wold (1985, 1982) and Sparse Principle Component analysis developed by Zou et al. (2006). The soft-threshold parameter λ_a and λ_b are obtained from X and Y and controls the number of variables to be included from each set. This approach is very similar to that of the lasso (Tibshirani 1996). Also, the penalties used are of Lagrangian form which lead to extra normalization due to some computational problems.

The first canonical weight is chosen based on the following algorithm

1. Select sparseness parameters λ_a and λ_b

2. Select initial values a^0 and b^0 set $i = 0$.

3. Update a :

(a) $a^{i+1} \leftarrow Kb^i$

(b) Normalize: $a^{i+1} \leftarrow \frac{a^{i+1}}{\|a^{i+1}\|}$

(c) Apply soft-thresholding to obtain sparse solution: $u_j^{i+1} \leftarrow (|u_j^{i+1}| - \frac{1}{2} \frac{\lambda_a}{|a_{SVD}|^\gamma}) + \text{Sign}(a_j^{i+1})$ for $j = 1, \dots, p$

(d) Normalize: $a^{i+1} \leftarrow \frac{a^{i+1}}{\|a^{i+1}\|}$

4. Update b :

- (a) $b^{i+1} \leftarrow K^T a^{i+1}$
 - (b) Normalize: $b^{i+1} \leftarrow \frac{b^{i+1}}{\|b^{i+1}\|}$
 - (c) Apply soft-thresholding to obtain sparse solution: $b_j^{i+1} \leftarrow (|b_j^{i+1}| - \frac{1}{2} \frac{\lambda_b}{|b_{SV D}|^\gamma}) + \text{Sign}(b_j^{i+1})$ for $j = 1, \dots, q$
 - (d) Normalize: $b^{i+1} \leftarrow \frac{b^{i+1}}{\|b^{i+1}\|}$
5. $i \leftarrow i + 1$
 6. Repeat steps 3, 4 and 5 until convergence.

where γ is defined by user.

Since the computation of the matrix K in equation (3.1) requires Σ_{xx}^{-1} and Σ_{yy}^{-1} which do not exist when the total number of variables is greater than the number of observation and it may not exist in case of collinearity, the two problems in finding inverse have been solved by replacing Σ_{xx} and Σ_{yy} by $\text{diag}(\Sigma_{xx})$ and $\text{diag}(\Sigma_{yy})$.

The algorithm is closely related to that of Le Cao et al. (2009) and Waaijenborg et al. (2008). The optimal combination of sparseness parameters are chosen by using k-fold cross-validation (CV) and the criteria that controls how many variables that has to be included in each soft-threshold steps is given by

$$\Delta_{cor} = \frac{1}{k} \sum_{j=1}^k |\text{cor}(X_j \hat{a}^{-j}, Y_j \hat{b}^{-j})|, \quad (3.2)$$

which maximizes the test sample correlation, where k is the number of steps in CV, \hat{a}^{-j} and \hat{b}^{-j} are the weights estimated for the training sets X_{-j} and Y_{-j} , in which the subset j has been removed.

As the training sample sizes approaches to infinity the criterion is same as Waaijenborg et al. (2008). Hence, the optimal combination of \hat{a}^{-j} and \hat{b}^{-j} chosen from the k-fold CV hence results in the highest average test sample correlation. The left and right boundaries of the sparseness parameters have been set to 0 and 2. When sparseness parameters are selected to be 0 then the SCCA approach provides the full SVD solution and when sparseness parameters is set to 2, it results in none of the variables included in the solution. As we increase the soft-thresholding parameters from 0 to 2 smaller set of variables from each variables are included.

After the optimal combinations of sparseness parameters is found, subsets of variables X and Y can be selected by applying SCCA to the whole data and we obtain the loading for singular vectors a and b . For better variable selection they have extended SCCA to adaptive SCCA and it filters bias in sparseness parameters. This approach is analogous to the adaptive lasso method (Zou 2006) which includes additional weights in the lasso constraint, where the weight is defined as $\hat{w} = \frac{1}{|\hat{\beta}|^\gamma}$, $\hat{\beta}$ is consistent estimator of β , and $\gamma > 0$ is a pre-specified parameter.

The main idea of adaptive lasso is that increasing sample size makes OLS estimates more precise and, most importantly, the estimates of zero-coefficients will converge to 0. Hence, it can be noticed that the weights of zero-coefficients converge to 0 and the weight of non zero-coefficients converges to some value. In the case of adaptive SCCA, parameters with the largest OLS value will be assigned a smaller penalty function and less shrinkage. The connection between SVD and OLS has been illustrated by I.J. Good (Good, 1969). In the algorithm above it can be noted that a_{SVD} and b_{SVD} denote the first singular vectors, have unit length, and are found

from full SVD approach of Parkhomenko et al. (2009) of K . The performance of SCCA will be examined by simulation studies in the next chapter.

3.1.2 SCCA Approach of Witten et al. (2009)

The columns of X and Y are standardized to have mean zero and standard deviation one and $a^T X^T X a \leq 1$, $b^T Y^T Y b \leq 1$ are replaced by $\|a\|^2$ and $\|b\|^2$ i.e., the $X^T X$ and $Y^T Y$ are replaced by their respective identity matrices. The SCCA criterion proposed by Witten et al. (2009) is

$$\max_{a,b} a^T X^T Y b \text{ subject to constraints } \|a\|^2 \leq 1, \|b\|^2 \leq 1, P_1(a) \leq c_1, P_2(b) \leq c_2,$$

where the penalty function $P_1(a) = \|a\|_1$ and $P_2(b) = \|b\|_1$ are lasso penalty functions. The penalty provides sparse a and b for a chosen c_1 and c_2 where $1 \leq c_1 \leq \sqrt{p}$ and $1 \leq c_2 \leq \sqrt{q}$. Most importantly both the penalty functions are viewed as lasso penalty function and the values of c_1 and c_2 are chosen by cross validation, where the corresponding value are chosen by grid search such that $Cor(Xa, Yb)$ attains maximum. Also we can choose c_1 and c_2 such that a preferred quantity of sparseness parameter are selected. They proposed a new approach to SVD named penalized matrix decomposition (PMD). For some matrix Z , with n rows and p columns, let the rank- k approximation be

$$\hat{Z} = ADB^T = \sum_{i=1}^k d_i a_i b_i^T, \quad A^T A = I_n, \quad B^T B = I_p, \quad d_1 \geq d_2 \geq \dots \geq d_k > 0,$$

where a_i and b_i are unit vectors respectively denoting the i^{th} elements of A and B , and d_i is the non-negative constant denoting the i^{th} diagonal element of the diagonal matrix D . The estimations of a_i and b_i are subject to a penalty on their elements.

The algorithm proposed by Witten et al. (2009) for calculating the canonical covariate of the SCCA is as follows:

1. Initialize b to have L_1 norm 1.
2. Iterate the following two steps until convergence:
 - (a) $a \leftarrow \arg \max_a a^T X^T Y b$ subject to $\|a\|^2 \leq 1, P_1(a) \leq c_1$.
 - (b) $b \leftarrow \arg \max_b a^T X^T Y b$ subject to $\|b\|^2 \leq 1, P_2(b) \leq c_1$.

where P_1 is an L_1 or lasso penalty and the update has the form

$a \leftarrow \frac{S(X^T Y b, \Delta_1)}{\|S(X^T Y b, \Delta_1)\|^2}$, where $\Delta_1 = 0$ then $\|a\|_1 \leq c_1$; otherwise, $\Delta_1 > 0$ is chosen so that $\|a\|_1 = c_1$. Here $S(\cdot)$ is a soft-threshold operator; that is, $S(a, c) = \text{sign}(a)(|a| - c)_+$.

The algorithm proposed by Witten et al. (2009) for computing Sparse CCA is similar to that of Waaijenborg et al. (2008). Waaijenborg et al. (2008) penalized the classical CCA as an iterative regression and then applied an elastic net penalty to find the canonical vectors. The elastic net is a combination of ridge regression and lasso. For more detail about ridge regression, see Hoerl (1962). Most importantly it can be noticed that unlike other approach to SCCA, Witten et al. (2009) use bounded form of the penalty function and the canonical vectors are fixed one at a

time such that the objective function of the biconvex criterion increases by each step of the iterative algorithm.

The tuning parameter values are selected with the following algorithm that he proposed

1. For each tuning parameter value T_j being considered:
 - (a) Compute the canonical vectors a and b using the data X and Y and tuning parameter T_j . Compute $d_j = Cov(Xa, Yb)$.
 - (b) For $i \in 1, \dots, N$, where N is some large number of permutations:
 - i. Permute the rows of X to obtain the matrix X^i , and compute canonical vectors a^i and b^i using data X^i and Y and tuning parameter T_j .
 - ii. Compute the $d_j^i = Cor(X^i a^i, Y b^i)$.
 - (c) Calculate the p-value $p_j = \frac{1}{N} \sum_{i=1}^N 1_{d_j^i \geq d_j}$.
2. Choose the tuning parameter T_j corresponding to the smallest p_j . Alternatively, one can choose the tuning parameter T_j for which $\frac{(d_j - \frac{1}{N} \sum_{i=1}^N d_j^i)}{sd(d_j^i)}$ is largest, where $sd(d_j^i)$ indicates the standard deviation of $d_j^1, d_j^2, \dots, d_j^N$. The resulting p -value is p_j . It also to be noted that since we have multiple p -values (p_j) we should use a strict stop value to avoid multiple testing problems linked to it.

LASSO

Lasso (Tibshirani 1996) is a shrinkage and selection of variables for linear regression method. Lasso minimizes the error sum of square subject to a constraint on the bound of the absolute values of the regression coefficient. Consider an independent variable Y which is a $(1 \times p)$ and set of dependent variables says X_1, X_2, \dots, X_p . The linear model fit can be given by $\hat{Y}_i = \beta_1 + \beta_2 X_1 + \dots + \beta_p X_p$ and the criterion as follows:

$$\text{Minimize } \sum_{i=1}^p (Y_i - \hat{Y}_i)^2 \text{ subject to } \sum_{i=1}^p |\beta_i| \leq c,$$

where the bound c is the tuning parameter. Based on the value chosen for c , as many regression coefficients are can be added in the model.

3.1.3 SCCA Approach of Lee et al. (2011)

Let X and Y contain p and q variables obtained from the same set of n observations each. Also let us define $u_j = Xa_j$ and $v_j = Yb_j$ to be the linear combinations of X and Y , where a_j and b_j are some pairs of canonical vectors. The sample cross-covariance matrix is denoted by $\hat{\Sigma} = \frac{X^T Y}{(n-1)}$, where X and Y are centred across the columns. The CCA approach involves finding a_j and b_j that maximize $a_j^T \hat{\Sigma} b_j$ subject to constraint $a_j^T a_j = b_j^T b_j = 1$, $a_j \perp a_h$, and $b_j \perp b_h$ for all $h < j$.

Hence a_k and b_k are the optimal canonical weight vectors that produce maximally correlated linear combinations between X and Y under orthogonality constraints. In high-dimensional data, such as genomic data, p and q are very large and hence a_j and b_j contain many noise weight vectors involving thing that are very difficult to

interpret. Hence it is required that the canonical weight vectors a_j and b_j must be sparse for easy interpretation. The SVD of $\hat{\Sigma}$ is a factorization of the form

$$\hat{\Sigma} = ADB^T$$

where D is a diagonal matrix with positive values arranged in decreasing order, A and B are orthogonal matrices. Let d_{ii} denote the i^{th} diagonal element of D which is a singular values of $\hat{\Sigma}$ and let A_i and B_i denote the i^{th} columns of A and B which are left and right singular vectors of $\hat{\Sigma}$. We can see that

$$a_1^T \hat{\Sigma} b_1 = \sum_i d_{ii} a_1^T A_i B_i b_1$$

which suggests that $a_1^T \hat{\Sigma} b_1$ is maximized when $a_1 = A_1$ and $b_1 = B_1$ and its value is d_{11} . Similar arguments can be followed for $a_j = A_j$ and $b_j = B_j$. Hence the CCA estimates can be obtained from the SVD on $\hat{\Sigma}$. Since p and q are very large in practice all the variables are not necessarily required in computing the CCA estimates. Hence a different approach to CCA estimates are necessary for obtaining the largest eigenvalue and its associated eigenvector of a matrix, this is achieved by non-linear iterative partial least-squares(NIPALS) algorithm.

The NIPALS algorithm computes a pair of singular vectors at a time which is effective if we wish to obtain only a few singular vectors. The approach to find the sparseness parameter is pretty much the same as that of Witten et al. (2009) but the penalized criterion is viewed as a random-effect model and includes a different penalty to it. The penalty function in random-effect model is a convex penalty

function but is allowed to be non-convex. Let us assume the regression model

$$Y_i = X_i^T \beta + \epsilon_i, \quad i = 1, 2, \dots, n,$$

where β is a $(p \times 1)$ vector of fixed unknown parameters and ϵ is some noise parameter with mean 0 and variance σ^2 . Sparse parameters are obtained by

$$\text{minimizing } \frac{1}{2} \sum_{i=1}^n (Y_i - X_i^T \beta)^2 + \sum_{j=1}^p p_\lambda(|\beta_j|), \quad (3.3)$$

where $p_\lambda(\cdot)$ is a penalty function. The derivative of the penalty function is given by

$$p_{\lambda_x}(|x|) = \frac{\lambda_x |x|}{\{w\{(2/w - 1) + k\}/4\}},$$

where $k = (\sqrt{8x_i^2/(w\theta) + (2/w - 1)^2})$ and when we choose $w = 2$, the above penalty function becomes the same as the lasso penalty function.

The modified NIPALS algorithm proposed by Lee et al. (2011) for calculating the canonical covariate of the SCCA is as follows:

Here we consider $X_1 = X$ and $Y_1 = Y$,

1. Set b_1 as the first column of X_1
2. Compute $a_1 = (\nabla_{b_1} + \lambda_a W_{a,\delta})^{-1} X_1^T b_1$, scale $a_1^* = \frac{a_1}{\sqrt{a_1^T a_1}}$.
3. Compute $u_1 = X_1 a_1^*$.
4. Compute $b_1 = (\nabla_{u_1} + \lambda_b W_{b,\delta})^{-1} Y_1^T u_1$, scale $b_1^* = \frac{b_1}{\sqrt{b_1^T b_1}}$.
5. Compute $v_1 = Y_1 b_1^*$.

6. If they converges then go to next step or else return to 2.
7. Obtain $\rho = u_1^T v_1$.
8. Compute residual matrices $X_2 = X_1 - u_1 a_1^{*T}$ and $Y_2 = Y_1 - v_1 b_1^{*T}$ for the subsequent pair of sparse singular vectors. At the start we replace X with X_2 and Y with Y_2 and so on.

The criterion to select the tuning parameter is the same as the one specified in Parkhomenko et al. (2009) SCCA approach. The optimal values of the tuning parameter corresponds to the highest test sample correlation.

Chapter 4

Simulation Studies

The behaviour of different approaches to sparse CCA is studied with simulation studies. Simulation studies help us to investigate the performance of the various SCCA methods in different scenarios, and to check the robustness of the methods to certain constraints before using them on a real data problem.

4.1 Introduction

The objective of SCCA is to find associations between two sets of variables coming from same object where, for computational purpose, we assume that a subset of variable in X is correlated to a subset of variables in Y . For comparing the three approaches that have been discussed in Chapter 3 using simulation, we generate two datasets, each with same size n . The total number of variables in each dataset may vary. Moreover, while simulating data we should make sure that only few variables in X are truly correlated with a few variables in Y while the rest are uncorrelated

among each other so that it helps us determine how well different methods are able to find a true sample correlation. The steps involved in simulating the data are as follows:

1. Let X contain p variables and Y contain some q variables.
2. The total number of samples from X and Y be n .
3. Let us assume that a subset of variables in X is correlated with a subset of variables in Y based on some model. The rest of variables in X are uncorrelated with other variables in Y .
4. Also let us assume that only the first few variables say r in X are highly correlated with the first few variables say r in Y .

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1r} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nr} & \cdots & x_{np} \end{bmatrix} \quad Y = \begin{bmatrix} y_{11} & \cdots & y_{1r} & \cdots & y_{1q} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nr} & \cdots & y_{nq} \end{bmatrix}$$

Let some r be the number of variables that are correlated between X and Y . Using the same approach as Parkhomenko et al. (2009), we generate a latent random vector say $\mu = [\mu_1, \mu_2, \dots, \mu_n]^T$ from $N(0, \sigma_\mu^2 I_n)$ where the first r variables in each set are generated as follows:

$$x_{ij} = \alpha_i \mu_i + e_{X_{ij}} \text{ for some } i = 1, 2, \dots, n \text{ and } j = 1, 2, \dots, r$$

$$y_{ij} = \beta_i \mu_i + e_{Y_{ij}} \text{ for some } i = 1, 2, \dots, n \text{ and } j = 1, 2, \dots, r,$$

where $\alpha_i \sim N(1, 0.1^2)$, $\beta_i \sim N(1, 0.1^2)$, $i = 1, 2, \dots, r$, $e_{X_{ij}} \sim N(0, \sigma_e^2)$, and $e_{Y_{ij}} \sim N(0, \sigma_e^2)$.

The vectors $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_r)^T$ and $\beta = (\beta_1, \beta_2, \dots, \beta_r)^T$ are $r \times 1$ random vector and are normalized to have $\|\alpha\|_2 = \|\beta\|_2 = 1$.

Let the uncorrelated variables in each set be assumed as follows:

$$x_{ij} = e_{X_{ij}} \text{ for some } i = 1, \dots, n \text{ and } j = r + 1, \dots, p$$

$$y_{ij} = e_{Y_{ij}} \text{ for some } i = 1, \dots, n \text{ and } j = r + 1, \dots, q,$$

where $e_{X_{ij}} \sim N(0, \sigma_e^2)$ and $e_{Y_{ij}} \sim N(0, \sigma_e^2)$ for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, p(q)$

Let us define $\mu_x = \mu + \sum_{i=1}^r e_{X_i}$ and $\mu_y = \mu + \sum_{i=1}^r e_{Y_i}$ where $e_{X_i} \sim N(0, \sigma_e^2)$ and $e_{Y_i} \sim N(0, \sigma_e^2)$ for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, r$

The maximum correlation between the linear combinations of correlated variables from both set say X and Y is given by:

$$\text{cor}(\mu_x, \mu_y) = \frac{\sigma_\mu^2}{\sqrt{\sigma_{\mu_x}^2 + r \cdot \sigma_e^2} \sqrt{\sigma_{\mu_y}^2 + r \cdot \sigma_e^2}}.$$

For example, let there be 50 variables that are correlated between the datasets X and Y. Let $\sigma_\mu^2 = 2$ and $\sigma_e^2 = 0.04$ will provide with maximum correlation between the linear combination of $\text{cor}(\mu_x, \mu_y) = 0.67$ whereas $\sigma_\mu^2 = 2$ and $\sigma_e^2 = 0.01$ will result in $\text{corr}(\mu_x, \mu_y) = 0.89$

Hence, the covariance matrix between variables in set X and Y will be

$$\begin{array}{c}
 \\
 \\
 x_1 \\
 x_2 \\
 \\
 x_r \\
 x_{r+1} \\
 \vdots \\
 \\
 x_p
 \end{array}
 \begin{pmatrix}
 & y_1 & y_2 & \cdots & y_r & y_{r+1} & \cdots & y_q \\
 \alpha_1\beta_1\sigma_\mu^2 & \alpha_1\beta_2\sigma_\mu^2 & \cdots & \alpha_1\beta_r\sigma_\mu^2 & 0 & \cdots & 0 \\
 \alpha_2\beta_1\sigma_\mu^2 & \alpha_2\beta_2\sigma_\mu^2 & \cdots & \alpha_2\beta_r\sigma_\mu^2 & 0 & \cdots & 0 \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
 \alpha_r\beta_1\sigma_\mu^2 & \alpha_r\beta_2\sigma_\mu^2 & \cdots & \alpha_r\beta_r\sigma_\mu^2 & 0 & \cdots & 0 \\
 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
 0 & 0 & \cdots & 0 & 0 & \cdots & 0
 \end{pmatrix}$$

The above matrix is the sparse representation of singular vectors where few set (say, r) of correlated variables for each X and Y being considered, and the rest of variables are uncorrelated. Also it can be noted that both X and Y were centred across columns. Hence, we have only one pair of canonical variables and $\Sigma_{xy} = dab^T$ where d is a singular eigenvalues and a and b are singular canonical vectors. Then the corresponding singular canonical vectors is given by,

$$a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_r \\ o \\ \vdots \\ 0 \end{pmatrix} \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_r \\ o \\ \vdots \\ 0 \end{pmatrix}$$

and its corresponding covariance matrix is as follows

$$\Sigma_{xy} = \begin{pmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_r & 0 & \cdots & 0 \\ a_2b_1 & a_2b_2 & \cdots & a_2b_r & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_rb_1 & a_rb_2 & \cdots & a_rb_r & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}$$

To select the tuning parameters we set,

1. Lee et al: 27 points at equal intervals from 0.2 to 10 for both left and right tuning parameters and the number of cross validation is set to be 10.
2. Parkhomenko et al: 10 points at equal intervals from 0 to 0.2 for both tuning

parameters and the number of cross validation is set to 10.

3. Witten et al: Computed by R package PMA with function `CCA.permute(.)` for both tuning parameters and the number of permutation is set to 10.
4. Classical CCA: Computed by R package CCA with function `rcc`.

The R program provided by Lee et al. (2011) produces Sparse canonical covariance analysis, hence sparse canonical correlation for the approach is obtained by:

$$\rho_1 = |\text{cor}(Xa, Yb)|$$

where a and b are corresponding sparse canonical vectors. Also a and b are $(p \times 1)$ and $(q \times 1)$ with non-sparse variables are set to 0.

4.2 Simulation Strategies

Three different approaches to SCCA and the Classical CCA have been compared with simulation and a summary of the parameters that are varied during simulation is presented in Table 4.1. The simulation studies were performed based on 100 replications and the averaged results are presented in Tables 4.2 - 4.19.

The approach that was used in Lee et al. (2011) for testing the performance of the estimates of the canonical variates was used in this thesis, i.e., the sine values of the angles between the true canonical variates and their estimates as the measure of

Table 4.1: Summary of Parameter in Simulation Studies

Varied Parameter	Description	Values between
n	Sample Size	30 and 500
p	Number of variables in X	50 and 2000
q	Number of Variables in Y	30 and 1500
r	Number of Correlated	5 and 40
σ_μ	Standard Deviation of Latent variable (μ)	1.8 and 4
σ_e	Standard Deviation of nuisance variable	0.1 and 0.5

closeness (Johnstone and Lu (2009)) which is given by

$$dist(a_1, \hat{a}_1) = \sin \angle(a_1, \hat{a}_1) = \sqrt{1 - (a_1^T \hat{a}_1)^2}$$

When the estimates are close to each other, the distance between a_1 and \hat{a}_1 approaches 0. From the table of results obtained from various simulations studies, we could see that all approach except the Classical CCA has better estimation, and all the estimates are almost close to zero. I have used the standard deviation for simulation of latent variable μ , 1.8, 2, 2.5, 3 and 4 and the standard deviation of nuisance variables is considered to be 0.1, 0.2, 0.3, 0.4 and 0.5 so that the maximum canonical correlation between two sets of variables ranges between 0.90 and 0.99, which are obtained by average of 100 simulation run.

The precision of the model selection is calculated by discordance measure, which corresponds to wrongly selected variables. The false positive is the number of nuisance variables with non-zero loadings in the resulted vector and the false negative is the number of correlated variables with zero loadings in the resulted vector. The number of false negatives and false positives are averaged over the 100 simulations

studies as well.

4.3 Simulation Results

All the methods performs well in some situation, but Witten et al. (2009) performs poorly on consideration of prediction and number of false negatives, Parkhomenko et al. (2009) and Lee et al. (2011) shows uniformly better prediction power compared with Witten et al. (2009). When we consider both performance and prediction accuracy, Parkhomenko et al. (2009) performs better than other proposed methods. It can also be notice that as we increase the variance of nuisance parameter (σ_μ) then number of false positive in Lee et al. (2011) increases despite providing with better prediction power.

From Table 4.2 - 4.17, we observe that all the methods performs well in a certain situation. For example, when we consider $\sigma_e = 0.1$, we could see from Tables 4.2 and 4.9 that Lee et al. approach performs better than the other two methods and Parkhomenko et al. (2009) method performs better for larger p and q . Also it can be noticed that as we increase r i.e., the number of correlated variable between X and Y , results in increasing false negative in Witten et al. approach to SCCA compared to the other methods. Also when r is large, Parkhomenko et al. (2009) method performs poorly but for large p say $p = 1000$ it performs better than other methods.

Looking at Table 4.2, it can be noticed that all the methods perform very well as the true correlation becomes very high, also the classical CCA gives rise to more false positives as compared to other methods. From Table 4.2 - 4.9, in situations

Table 4.2: $n = 100$, $p = 300$, $q = 200$, $r = (10, 20, 40)$, $\sigma_\mu = 1.8$, $\sigma_e = 0.1$, WT = Witten et al. (2009), LT = Lee et al. (2011), PT = Parkhomenko et al. (2009), CCA = Classical Canonical Correlation Analysis. FPN = Number of false positives and FNN = Number of false negatives

Method	r	Test Corr.	Distance (a_1)	Distance (b_1)	FPN in a_1	FPN in b_1	FNN in a_1	FNN in b_1
CCA	10	0.9530	0.84	0.84	290	190	0	0
WT		0.9933	0.67	0.79	1.14	0.26	3.94	5.38
PT		0.9968	0.09	0.09	1.79	1.75	0	0
LT		0.9968	0.02	0.02	0.42	0.55	0	0
CCA	20	0.953	0.84	0.84	280	180	0	0
WT		0.9895	0.76	0.84	3.17	0	11.35	13.34
PT		0.9968	0.09	0.09	5.98	5.75	0	0
LT		0.9968	0.03	0.03	0.27	0.55	0	0
CCA	40	0.9530	0.84	0.84	260	160	0	0
WT		0.9802	0.9	0.93	1.22	0	31.27	33.71
PT		0.997	0.09	0.09	15.38	13.44	0	0
LT		0.9969	0.04	0.04	0.49	0.47	0	0

where a high correlation is considered between a subset of variables in X and Y , Lee et al. algorithm performs better than other methods in terms of the number of false negatives and The algorithm proposed by Lee et al. (2011) provide with close to true correlation and better estimates compared to the other methods.

Let us take into account where a larger subset of variables in each sets is correlated, i.e., $r = 40$, Table 4.2 - 4.5, Lee et al. (2011) method outperforms other methods when we consider performance of estimates and the accuracy of model selection concurrently. It can be noticed that the estimates of Parkhomenko et al. (2009) and Lee et al. (2011) are similar in all the tables. The performance of estimates of canonical CCA and Witten et al. (2009) perform poorly. In addition, we notice that the mean of false negative for Lee et al. (2011) and Parkhomenko et al. (2009) in

Table 4.3: $n = 200, p = 300, q = 200, r = (10, 20, 40), \sigma_\mu = 1.8, \sigma_e = 0.1$

Method	r	Test Corr.	Dist(a_1)	Dist(b_1)	FPN in a_1	FPN in b_1	FNN in a_1	FNN in b_1
CCA	10	0.9532	0.84	0.84	290	190	0	0
WT		0.9935	0.59	0.75	1.5	0.24	3.08	4.81
PT		0.9969	0.09	0.09	1.4	0.92	0	0
LT		0.9969	0.06	0.01	0.28	0.95	0	0
CCA	20	0.953	0.84	0.84	280	180	0	0
WT		0.9907	0.66	0.76	8.41	0	8.71	11.09
PT		0.9969	0.09	0.09	3.07	5.07	0	0
LT		0.9967	0.02	0.02	0.26	1.04	0	0
CCA	40	0.9533	0.84	0.84	260	160	0	0
WT		0.9795	0.87	0.91	1.63	0	29.02	32.19
PT		0.9969	0.08	0.08	11.24	6.36	0	0
LT		0.9969	0.02	0.03	0.31	0.72	0	0

this case is close to 0 whereas the number of false negative for Witten et al. (2009) is high.

Let us consider cases with a sufficiently large number of variables where ($p \geq 1000$) in each sets, i.e., Table 4.4, 4.5, and 4.16, where in Table 4.4 and 4.5 the true correlation had been set to high values whereas in Table 4.15, the true correlation had been set to moderately high value, 0.8333. It can be seen that when the true correlation has been set to low value, Parkhomenko et al. (2009) performs better to attain true correlation and the performance of estimates. When we set $\sigma_e = 0.1$, the performance of estimates of Lee et al. and Parkhomenko et al. (2009) is very similar but the performance of false positives by Witten et al. (2009) performs better than other methods as r increases. There is not much difference between Table 4.4 and 4.5 as we increase sample size.

Consider a situation with a small subset of variables correlated, i.e., $r = 5, 10$. In

Table 4.4: $n = 100, p = 1000, q = 800, r = (10, 20, 40), \sigma_\mu = 1.8, \sigma_e = 0.1$

Method	r	Test Corr	Dist(a_1)	Dist(b_1)	FPN in a_1	FPN in b_1	FNN in a_1	FNN in b_1
CCA	10	0.9548	0.84	0.85	990	790	0	0
WT		0.9968	0.09	0.39	1	0	0	0.36
PT		0.9968	0.09	0.09	2.28	2.37	0	0
LT		0.997	0.07	0.06	35.04	37.8	0	0
CCA	20	0.9546	0.84	0.85	980	780	0	0
WT		0.9956	0.6	0.65	14.98	7.96	5.45	7.45
PT		0.9968	0.09	0.09	5.31	5.7	0	0
LT		0.997	0.06	0.06	18.88	37.91	0	0
CCA	40	0.9552	0.84	0.85	960	760	0	0
WT		0.9934	0.83	0.87	6.81	1.92	19.29	22.59
PT		0.997	0.08	0.08	15.57	25.8	0	0
LT		0.997	0.05	0.07	9.11	42	0	0

these situations, we could see that most of the Table representing such a situation had number of false negative close to 0 for all the methods except Witten et al. (2009), Lee et al. (2011) method performs poorly when $\sigma_e \geq 0.2$. Moreover, the performance of estimates of Lee et al. (2011) seems to be better than other methods. The performance of estimates of classical CCA and Witten et al. (2009) is poor. Furthermore, Lee et al. (2011) method has better accuracy towards true correlation in comparison with other methods. The overall performance of Lee et al. method is better than other methods when $\sigma_e = 0.1$ and Witten et al. (2009) performs better when $\sigma_e \geq 0.2$. Overall for small r , Parkhomenko et al. (2009) method performs better than the rest, but performs poorly when sample size is very small.

Let us consider when we have extremely small sample size situation, i.e., Table 4.8, 4.9, 4.14, and 4.15 represents this simulation scenario. The results can be split into two when $\sigma_e = 0.1$, the average number of false negative in all the methods

Table 4.5: $n = 200$, $p = 1000$, $q = 800$, $r = (10, 20, 40)$, $\sigma_\mu = 1.8$, $\sigma_e = 0.1$

Method	r	Test Corr	Dist(a_1)	Dist(b_1)	FPN in a_1	FPN in b_1	FNN in a_1	FNN in b_1
CCA	10	0.9543	0.84	0.85	990	790	0	0
WT		0.9967	0.09	0.38	1.01	0	0	0.42
PT		0.9969	0.09	0.09	1.14	1.16	0	0
LT		0.9969	0.05	0.04	36	46	0	0
CCA	20	0.9543	0.84	0.84	980	780	0	0
WT		0.9945	0.55	0.6	21.79	6.82	4.75	6.76
PT		0.9969	0.09	0.09	3.62	4.61	0	0
LT		0.9969	0.12	0.04	20	46	0	0
CCA	40	0.9545	0.84	0.85	960	760	0	0
WT		0.9933	0.66	0.71	21.04	6.41	16.53	19.67
PT		0.9969	0.07	0.08	8.72	13.19	0	0
LT		0.9969	0.04	0.05	9.87	50	0	0

except Witten et al. (2009) is 0. Lee et al. (2011) performs better than other methods. The prediction accuracy of Lee et al. (2011) for this case is better than other methods. Despite providing with poor false negatives, Witten et al. (2009) method provides with better false positives. When $\sigma_e \geq 0.2$ then Witten et al. (2009) method outperforms other methods. None of the method provides with 0 false negatives in any situation. Also when r increases results in larger value of false negatives for Witten et al. (2009) method compared to others. Both Lee et al. (2011) and Parkhomenko et al. (2009) methods perform poorly in this situation but Lee et al. method performs well with better prediction accuracy.

Different simulation settings have been considered where the sample size is greater than the number of variables in set X . It can be seen from Table 4.9, 4.10, 4.13, and 4.17 that the classical CCA and Witten et al. (2009) estimates perform poorly in these situations where $n \gg p$. Also the estimates given by Parkhomenko et al.

Table 4.6: $n = 100, p = 200, q = 150, r = (10, 20), \sigma_\mu = 2, \sigma_e = 0.1$

Method	r	Test Corr	Dist(a_1)	Dist(b_1)	FPN in a_1	FPN in b_1	FNN in a_1	FNN in b_1
CCA	10	0.9613	0.87	0.87	190	140	0	0
WT		0.9944	0.62	0.73	2.1	0	3.75	4.7
PT		0.9974	0.09	0.09	1.15	0.91	0	0
LT		0.9974	0.02	0.02	0.18	0.4	0	0
CCA	20	0.9611	0.87	0.87	180	130	0	0
WT		0.9895	0.82	0.86	1.1	0.04	12.46	14.11
PT		0.9974	0.09	0.09	5.02	3.66	0	0
LT		0.9974	0.02	0.02	0.29	0.5	0	0

Table 4.7: $n = 200, p = 200, q = 150, r = (10, 20), \sigma_\mu = 2, \sigma_e = 0.1$

Method	r	Test Corr	Dist(a_1)	Dist(b_1)	FPN in a_1	FPN in b_1	FNN in a_1	FNN in b_1
CCA	10	0.9616	0.87	0.87	190	140	0	0
WT		0.9943	0.61	0.72	2.63	0	3.49	4.61
PT		0.9975	0.09	0.09	1.05	0.65	0	0
LT		0.9975	0.01	0.01	0.19	1.1	0	0
CCA	20	0.9615	0.87	0.87	180	130	0	0
WT		0.9895	0.78	0.84	2.37	0.77	11.91	13.46
PT		0.9975	0.09	0.09	1.83	2.43	0	0
LT		0.9975	0.02	0.02	0.19	1.1	0	0

Table 4.8: $n = 30, p = 300, q = 200, r = 20, \sigma_\mu = 1.8, 2, \sigma_e = 0.1$

Method	σ_μ	Test Corr	Dist(a_1)	Dist(b_1)	FPN in a_1	FPN in b_1	FNN in a_1	FNN in b_1
CCA	1.8	0.9522	0.83	0.84	280	180	0	0
WT		0.9932	0.8	0.86	1.86	0.25	11.7	13.98
PT		0.997	0.18	0.16	42.99	23.35	0	0
LT		0.9967	0.05	0.05	0.26	0.7	0	0
CCA	2	0.9607	0.86	0.87	280	180	0	0
WT		0.9440	0.81	0.87	1.51	0.26	11.99	14.16
PT		0.9975	0.18	0.15	41.19	22.34	0	0
LT		0.9975	0.04	0.05	0.18	0.44	0	0

Table 4.9: $n = (30, 100, 500)$, $p = 50$, $q = 40$, $r = 5$, $\sigma_\mu = 2$, $\sigma_e = 0.1$

Method	n	Test Corr	Dist(a_1)	Dist(b_1)	FPN in a_1	FPN in b_1	FNN in a_1	FNN in b_1
WT	30	0.9942	0.76	0.81	0.09	0.03	2.91	3.07
PT		0.9974	0.22	0.2	10.42	8.52	0	0
LT		0.9974	0.03	0.03	0.21	0.3	0	0
WT	100	0.9936	0.66	0.73	0.59	0.2	2.39	2.53
PT		0.9974	0.09	0.08	1.04	0.97	0	0
LT		0.9974	0.01	0.01	0.22	0.31	0	0
WT	500	0.9956	0.45	0.6	0.59	0.15	1.04	1.37
PT		0.9975	0.08	0.08	0.29	0.23	0	0
LT		0.9975	0.01	0.01	0.12	0.59	0	0

(2009) and Lee et al. (2011) in these cases perform much better, notably Lee et al. (2011) method outperformed other methods when we consider the discordance measure but performs poorly on considering number of false positives when $\sigma_e \geq 0.2$. Overall in these situations, Parkhomenko et al. (2009) method outperformed other methods.

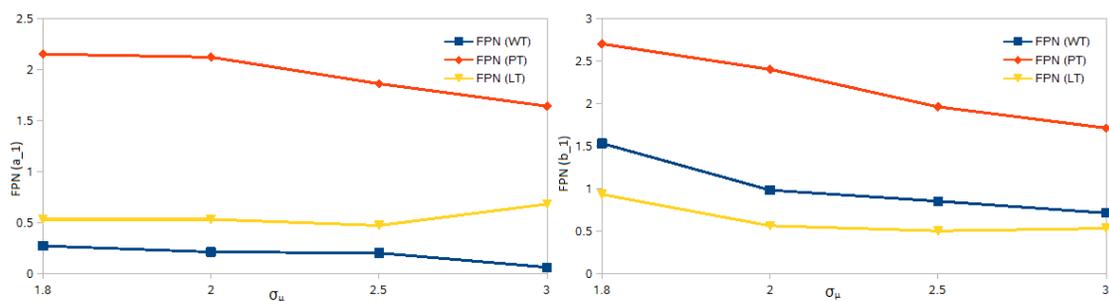


Figure 4.1: Plotting number of false positives of both parameters with different σ_μ and sample size = 100, $p = 50$, $q = 30$, $r = 5$ and $\sigma_e = 0.3$. with all the methods

Looking at Figure 4.1, it can be noticed that as we increase σ_μ , the number of false positives of Witten et al. (2009) reaches close to zero where as we could see

Table 4.10: $n = 100, p = 50, q = 30, r = 5, \sigma_\mu = 1.8, 2.5, 3, \sigma_e = 0.3$

Method	σ_μ	Test Corr	Dist(a_1)	Dist(b_1)	FPN in a_1	FPN in b_1	FNN in a_1	FNN in b_1
CCA	1.8	0.9364	0.84	0.84	45	25	0	0
WT		0.9484	0.55	0.68	1.53	0.27	2.11	2.84
PT		0.9713	0.2	0.2	2.7	2.15	0	0.01
LT		0.9725	0.12	0.04	32.06	0.91	0	0.01
CCA	2.5	0.9657	0.92	0.92	45	25	0	0
WT		0.9707	0.6	0.72	0.85	0.2	2.35	3.13
PT		0.9845	0.22	0.21	1.96	1.86	0	0
LT		0.9854	0.08	0.03	32.62	0.95	0	0
CCA	3	0.9759	0.94	0.94	45	25	0	0
WT		0.9789	0.6	0.71	0.71	0.06	2.46	3.15
PT		0.9892	0.23	0.22	1.71	1.64	0	0
LT		0.9898	0.07	0.03	33.15	0.91	0	0

Table 4.11: $n = 100, p = 500, q = 300, r = 15, \sigma_\mu = 1.8, 2, \sigma_e = 0.3$

Method	σ_μ	Test Corr	Dist(a_1)	Dist(b_1)	FPN in a_1	FPN in b_1	FNN in a_1	FNN in b_1
CCA	1.8	0.9556	0.88	0.87	485	285	0	0
WT		0.9576	0.57	0.67	6.83	1.85	5.1	8.03
PT		0.973	0.15	0.15	10.55	6.92	0.05	0.05
LT		0.977	0.27	0.22	113.81	72.8	0.01	0
CCA	2	0.9630	0.9	0.89	485	285	0	0
WT		0.9649	0.5	0.68	4.46	1.42	5.13	8.15
PT		0.9779	0.15	0.15	5.24	5.84	0.05	0.04
LT		0.9806	0.25	0.19	117.75	73.1	0.01	0

that number of false positives for Lee et al. (2011) method tends to increase. Also other than σ_μ , the rest of the parameters are fixed. The false negative of Witten et al. (2009) method has been shown in Figure 4.2, increasing σ_μ increases number of false negative of both the parameters.

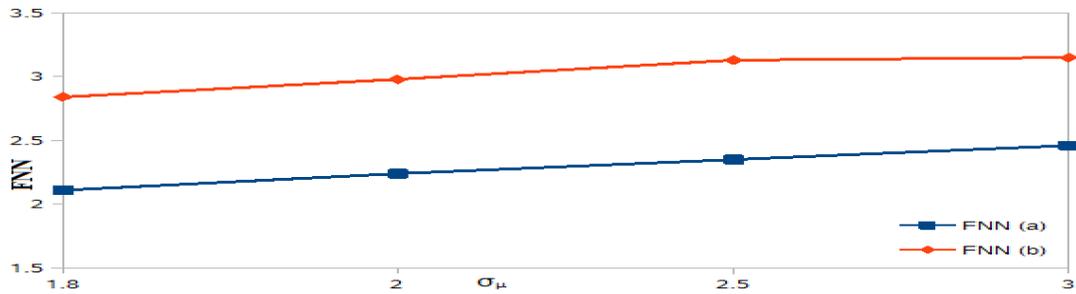


Figure 4.2: Plotting number of false negatives of both parameters with different σ_μ and sample size = 100, $p = 50$, $q = 30$, $r = 5$ and $\sigma_e = 0.3$. considering Witten et al. Approach to SCCA

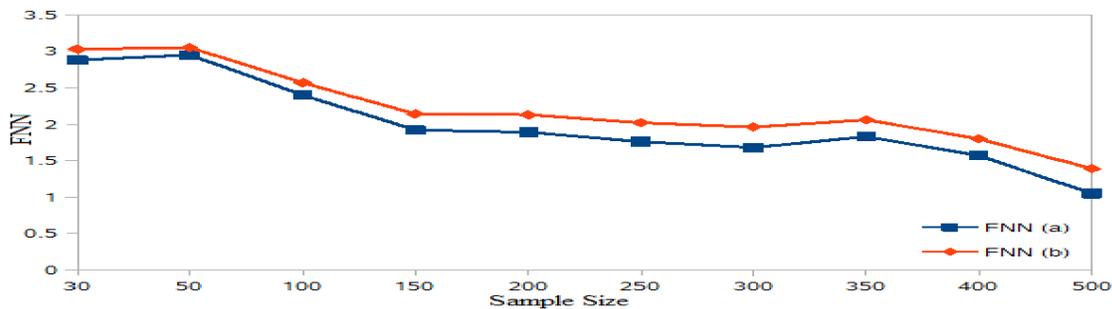


Figure 4.3: Plotting number of false negatives of both parameters with different sample sizes between 30 and 500, $p = 50$, $q = 40$, $r = 5$, $\sigma_\mu = 2.5$ and $\sigma_e = 0.1$. considering only Witten et al. Approach to SCCA

Figure 4.3 shows what happens when we vary σ_e from 0.1 to 0.5. The number of false positives increases for all the methods but for both Lee et al. (2011) and Parkhomenko et al. (2009) method it increases rapidly.

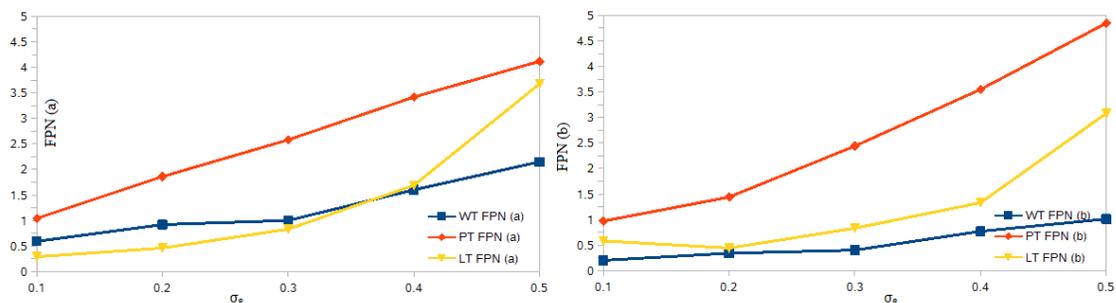


Figure 4.4: Plotting number of false positives of both parameters with varying nuisance standard deviation, $p = 50$, $q = 40$, $r = 5$, $\sigma_\mu = 2$ and $n = 100$. consider all Approaches to SCCA

Figure 4.4 represents what happens when sample size increases in Witten et al. (2009) method. It can be noticed that as sample size increases the average number of false negatives approaches 0. Also in Table 4.16 it can be noticed that the simulation run have been averaged and median has been considered whereas for all other tables the simulation results have been averaged over 100 simulation run. The reason for considering median in Table 4.16 is to show that there was not much of variation between mean and median in other tables whereas in Table 4.16 there were high variation between number of false positives in both parameters.

Figure 4.5 shows that varying σ_e shows constant change to number of false negatives in all the method. Figure 4.6 identifies the performance accuracy of both the parameters and it can be noticed that increasing σ_e results in increasing estimates of Lee et al. and Parkhomenko et al. (2009) method whereas Witten et al. (2009) methods decreases gradually but provide with poor estimates.

Table 4.12: $n = 100, p = 300, q = 200, r = (5, 15), \sigma_\mu = 3, \sigma_e = 0.4$

Method	r	Test Corr	Dist(a_1)	Dist(b_1)	FPN in a_1	FPN in b_1	FNN in a_1	FNN in b_1
WT	5	0.9768	0.39	0.57	0.91	0.23	0.9	1.85
PT		0.9809	0.26	0.26	3.81	4.7	0.01	0.01
LT		0.9835	0.19	0.15	95.05	58.27	0	0
WT	15	0.9668	0.62	0.73	1.4	0.21	7.86	9.69
PT		0.9818	0.2	0.2	5.71	5.59	0.2	0.28
LT		0.9834	0.18	0.15	77.41	57.12	0.06	0.06

Table 4.13: $n = 300, p = 100, q = 50, r = (5, 15), \sigma_\mu = 3, \sigma_e = 0.4$

Method	r	Test Corr	Dist(a_1)	Dist(b_1)	FPN in a_1	FPN in b_1	FNN in a_1	FNN in b_1
WT	5	0.9688	0.51	0.65	4.4	0.2	2.19	2.9
PT		0.9812	0.26	0.25	0.36	0.75	0.01	0.01
LT		0.9825	0.04	0.03	11.6	2.7	0	0
WT	15	0.9497	0.65	0.81	0.4	0	8.93	11.48
PT		0.9819	0.19	0.18	1.39	0.35	0.19	0.29
LT		0.9825	0.05	0.04	8.61	3.7	0.02	0.08

Table 4.14: $n = 50, p = 100, q = 80, r = (5, 15), \sigma_\mu = 3, \sigma_e = 0.5$

Method	r	Test Corr	Dist(a_1)	Dist(b_1)	FPN in a_1	FPN in b_1	FNN in a_1	FNN in b_1
WT	5	0.9543	0.57	0.61	0.83	0.14	2.72	2.84
PT		0.9719	0.31	0.31	11.52	9.16	0.07	0.04
LT		0.9737	0.16	0.15	17	14	0.07	0.03
WT	15	0.9462	0.67	0.73	0.21	0	9.52	10.39
PT		0.9741	0.24	0.23	12.84	9.97	0.77	0.71
LT		0.9743	0.18	0.16	15.9	14	0.76	0.31

Table 4.15: $n = 50, p = 100, q = 80, r = (5, 15), \sigma_\mu = 4, \sigma_e = 0.5$

Method	r	Test Corr	Dist(a_1)	Dist(b_1)	FPN in a_1	FPN in b_1	FNN in a_1	FNN in b_1
WT	5	0.9719	0.61	0.65	0.35	0	2.98	3.09
PT		0.9831	0.33	0.33	10.11	8.28	0.04	0.01
LT		0.9845	0.12	0.11	18	14	0.05	0.02
WT	15	0.9634	0.7	0.75	0.5	0.11	10.07	10.83
PT		0.9634	0.25	0.25	9.62	9.06	0.59	0.5
LT		0.9848	0.13	0.12	15	14	0.48	0.38

Table 4.16: $n = 100, p = 2000, q = 1500, r = 40, \sigma_\mu = 3, \sigma_e = 0.4$

Method		Test Corr	Dist(a_1)	Dist(b_1)	FPN in a_1	FPN in b_1	FNN in a_1	FNN in b_1
WT	mean	0.9718	0.4	0.53	4.44	2.1	13.53	19.28
PT		0.9761	0.24	0.26	19.7	36.18	6.63	6.48
LT		0.9851	0.47	0.46	346.92	432.04	3.39	1.21
WT	median	0.9719	0.40	0.60	0	0	14	20
PT		0.9767	0.22	0.23	1	0	6	5.5
LT		0.9856	0.46	0.46	347	446	2	1

Table 4.17: $n = 500, p = 150, q = 100, r = (5, 15), \sigma_\mu = (3, 4), \sigma_e = 0.5$

Method	σ_μ	r	Test Corr	Dist(a_1)	Dist(b_1)	FPN in a_1	FPN in b_1	FNN in a_1	FNN in b_1
WT	3	5	0.9622	0.34	0.51	0.85	0.1	1.09	2.38
PT			0.9710	0.26	0.26	0.2	0.65	0.08	0.06
LT			0.9733	0.07	0.05	38	24	0.02	0.01
WT	4	5	0.9769	0.39	0.56	0.15	0	1.38	2.7
PT			0.9831	0.29	0.29	0.19	0.6	0.04	0.04
LT			0.9847	0.05	0.04	39	24	0	0.01
WT	3	15	0.9573	0.47	0.62	1.65	0.08	6.55	8.81
PT			0.9723	0.17	0.17	0	0	0.72	0.92
LT			0.9733	0.07	0.06	30	24	0.12	0.12
WT	4	15	0.9727	0.5	0.65	0.74	0.05	7.18	9.35
PT			0.9839	0.22	0.22	0	0	0.6	0.71
LT			0.9847	0.05	0.05	30	25	0.08	0.08

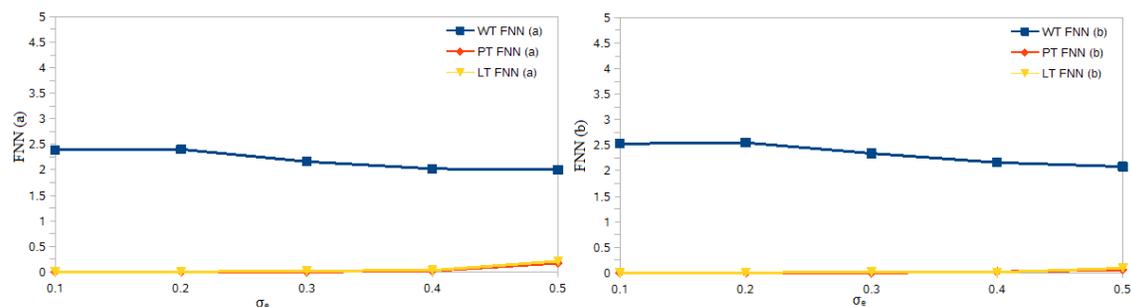


Figure 4.5: Plotting number of false negatives of both parameters with varying nuisance standard deviation, $p = 50$, $q = 40$, $r = 5$, $\sigma_\mu = 2$ and $n = 100$. consider all Approaches to SCCA

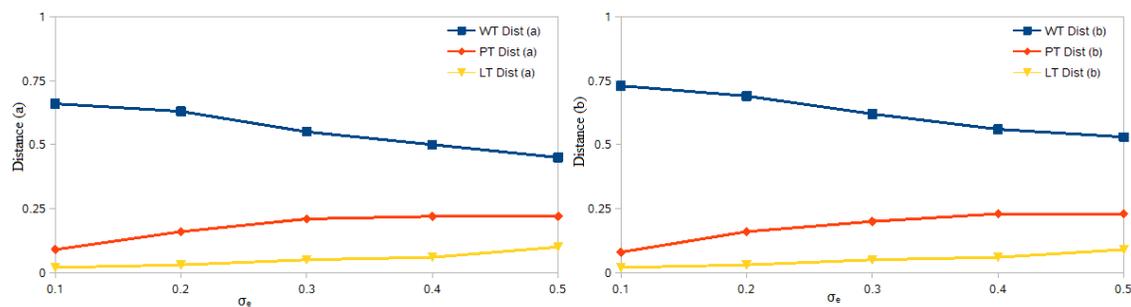


Figure 4.6: Plotting distance between estimates and true value of estimates of both parameters with varying nuisance standard deviation, $p = 50$, $q = 40$, $r = 5$, $\sigma_\mu = 2$ and $n = 100$. consider all Approaches to SCCA

4.4 Key Finding

Classical CCA outperforms other methods if all the variables are highly significant in both sets of variables as SCCA methods provides with high false negatives and false positives but give poor estimates.

Witten et al.(2009) performs well when there is high variance in nuisance parameters but provides with poor parameter estimates throughout all simulations. Furthermore, as sample size increases the method leads to better estimation. In

addition, when we increase the number of correlated variable between two sets of variable's increases false negative simultaneously.

Parkhomenko et al. (2009) works well when the number of variables in each sets is very large but performs poorly when sample size is very small. Despite giving low false negatives as we increase the number of correlated variables between X and Y leads to increase in false positives. The method prediction accuracy of estimates is better than Classical CCA and Witten et al. (2009). Furthermore, Parkhomenko et al. (2009) method cannot be used when sample size less than 30.

Lee et al. (2011) gives better prediction accuracy but performs poorly as variance of nuisance parameter increases. The method performs poorly as the number of variables in each sets is very large. Moreover, the method give rise to high false positives when variance in nuisance parameter is ≥ 0.2 . The method works wonderfully when $\sigma_e = 0.1$ and the numbers of variables in each sets are less than or equal to 500.

Chapter 5

Real Data Application of SCCA

The chapter begins with introduction to the dataset and method used to normalize the data, and followed by interpretation of results based on the analysis of data on three different approaches to SCCA. Lee et al. (2011) have used the same dataset for their real data analysis as well.

5.1 Background

In the human body, the genome consists of 23 pairs of chromosomes. Either one of each pair is inherited from the mother or father. Each chromosome is made of chains of DNA; DNA are bundled around each other in a structure known as a double helix. The central dogma of cell biology of DNA is shown in Figure 5.1. Genes are parts of the DNA structure. Generally speaking, a gene is a segment of DNA that defines a single trait encryption a particular pattern, about 20,500 (Clamp et al. (2007)) of which exists in humans.

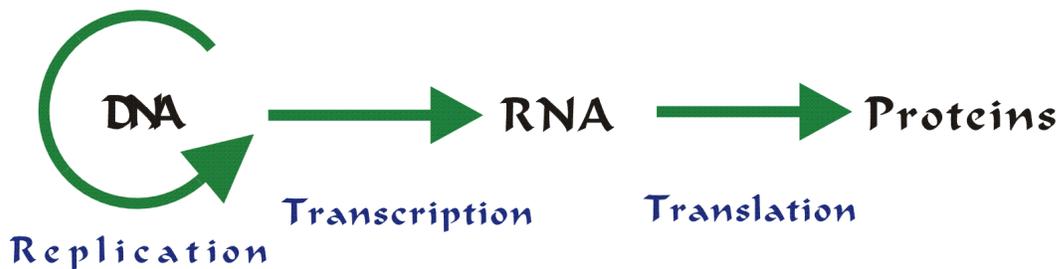


Figure 5.1: Central Dogma of Cell Biology

The persistence of genes is to perform as a prototype in the constitution of proteins. The information contained in the genes is transcribed into a messenger ribonucleic acid (mRNA) strand and a mRNA molecule departs from the nucleus of the cell where it is transcribed into a protein expression in a process translation. This is performed by ribosomes, which read the code carried by mRNA molecules from the cell nucleus and create proteins. These proteins are the structural block of the human body. Translating genes into a functional product is known as gene expression. (Allison 2007)

5.2 Data Description

The Microarray and proteomic data used in the following analysis is obtained from the National Cancer Institute <http://discover.nci.nih.gov/cellminer/>. The microarray data contains 60 humans cancer cell lines that include a variety of cancer tissues of origins such as leukemias, lymphomas, and carcinomas of ovarian, renal, breast, prostate, colon, lung, and CNS origin.

For the micro array, I used mRNA expression database on the NCI-60 from

Table 5.1: Overview of mRNA expression and protein data sets

Data set	No. feature sets or clones (genes)	Reference
mRNA expression Affymetrix HG-U133A	22,283 feature sets	Shankavaram et al. (2007)
Protein expression	162 Antibodies	Shankavaram et al. (2007)

Affymetrix HG-U133A. Furthermore, in the analysis, 59 of the total 60 human cancer cell lines were used and a cell line with missing microarray information has been removed. The following table provides with the information about mRNA and protein expression data sets and their references.

5.3 Normalization

The normalization of Affymetrix HG-U133A chip has been done by Gene Chip Robust Multiarray Averaging(GCRMA) method and for protein expression data, reverse-phase protein lysate arrays (RPLA) have been used to obtain 89 proteins expression values. GCRMA is a method for normalizing and summarizing probe-level intensity measurements from Affymetrix GeneChips. Starting with the probe-level data from a set of GeneChips, the perfect-match (PM) values are background-corrected, normalized and finally summarized resulting in a set of expression measures.

The protein values were adjusted to 25% 'dose interpolation' (DI25) algorithm (Nishizuka et al. (2003)). After normalization of mRNA expression, Genes with Inter quartile range (IQR) lesser than 0.5 across sample expressions has been removed from the dataset before applying SCCA approaches. After filtration, 11141 genes are left

for further analysis. No filtration has been done in protein expression dataset. The approach to normalization of mRNA expression of Affymetrix HG-U133A chips that I have discussed above is very similar to that of Lee et al. (2011) except that I filtered those genes with $IQR < 0.5$ compared to $IQR < 0.25$ chosen by them.

5.4 Results

The real data analysis have been split into two different analysis, one analysis had been done after removing those genes without their geneid and second analysis had been done including those genes without their gene names. It can noted that in the first analysis the total number of genes used for analysis is 10485 which is obtained after normalization from the gene expression data (21,069), whereas in the second analysis the total number of gene expressions used for analysis is 11,141 which is obtained after normalization from the gene expression data (22,283).

5.4.1 Result 1

After normalization of data and filtration has been completed, the data is ready to apply SCCA methods. For Lee et al. (2011) method the optimal tuning parameters were selected by 5-fold cross validation whereas for Parkhomenkho et al. (2009) was selected based on 10-fold cross validation. As mentioned earlier the number of variables in gene expression data was 10,485 and there is no change to proteomic data. Table 5.2 shows the results of analysis where correlation and number of non-zero in both gene expression and protein expression data were displayed.

Table 5.2: Summary of results of three different approaches of SCCA; CCA - Classical Canonical Correlation Analysis, LT - Lee et al. (2011) method, WT - Witten et al. (2009), PT - Parkhomenko et al. (2009)

Method	Non-zeros in gene data	Non-zeros in protein data	Canonical correlation
CCA	10,423	89	0.8236
LT	3078	7	0.8569
WT	3201	24	0.9509
PT	340	34	0.9546

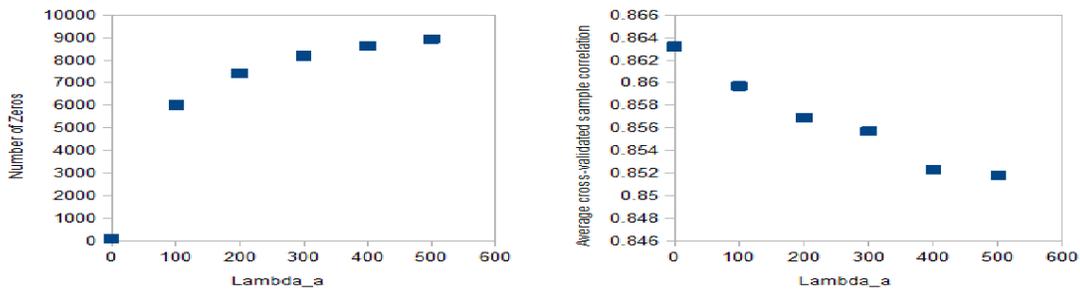


Figure 5.2: The cross-validated sample canonical correlation and the number zeros considering LT penalty function of the tuning parameters, λ_a ranges from 100 to 500

From Table 5.2, it's shown that all SCCA methods performance better than classical CCA method. When comparing three different approach to SCCA it can be noticed that the SCCA approach by Parkhomenkho et al. (2009) method provides with sparse gene expression and protein expression compared to other methods. Furthermore, Parkhomenkho et al. (2009) proposed method performs better in providing maximum canonical correlation among all the methods. Lee et al. (2011) method provides with sparse protein expression whereas Parkhomenkho et al. (2009) method performs better in terms of providing with sparse mRNA expression. Lee et al. (2011) method provides with poor cross validated canonical correlation.

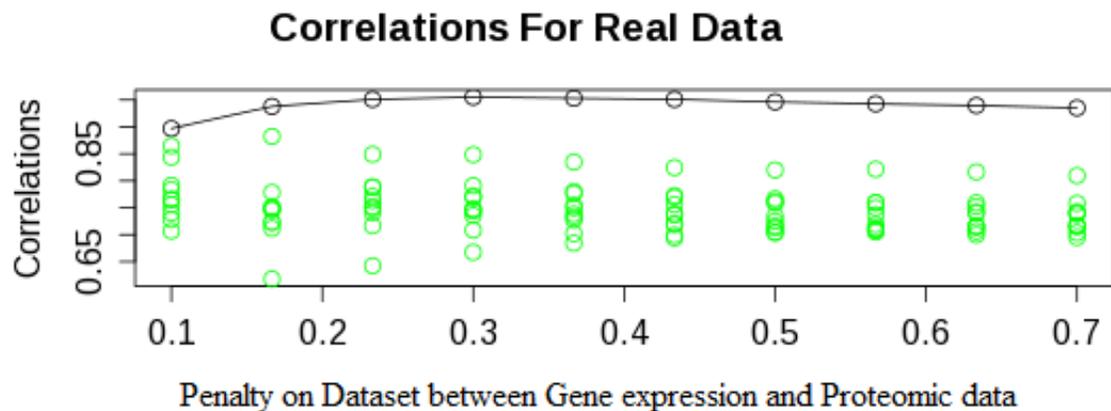


Figure 5.3: The canonical correlation of WT penalty function of the tuning parameters, λ_a and λ_a ranging from 0 and 0.7

Figure 5.2 shows the number of zeros and cross-validated canonical correlation of Lee et al. (2011) method as a function of λ_a . Moreover, it can be noted that as specified by Lee et al. (2011), λ_b have been fixed to be 300, which corresponds to 7 protein expression. From the right scattered plot in Figure 5.2, corresponds to $\lambda_a = 200$ and from the left scatter plot we have that 7407 for the gene pattern $(\lambda_a, \lambda_b) = (200, 300)$. Thus, the Lee et al. (2011) method can also have sparse gene pattern comparable to Parkhomenko et al. (2009) method at same cross-validated correlation, while having sparse protein expression. Furthermore, the left tuning parameter has been set between 100 and 500. In addition, Figure 5.3 represents with canonical correlation at different levels of tuning parameters and the optimal parameters have been selected based on high correlation obtained for Witten et al. (2009) method.

Table 5.3: Summary of results of three different approaches of SCCA

Method	Non-zeros in gene data	Non-zeros in protein data	Canonical correlation
CCA	10,823	89	0.8426
LT	3232	7	0.8579
WT	3418	24	0.9516
PT	310	34	0.9559

5.4.2 Result 2

Similar to Section 5.4.1, the optimal tuning parameters for Lee et al. (2011) and Parkhomenko et al. (2009) methods are selected based on 5 and 10 cross-validation respectively. The total number of genes available in gene expression after cleaning is 11,141 and protein expressions were unchanged. Table 5.3 displays the results of the statistical analysis done on the classical CCA and SCCA methods.

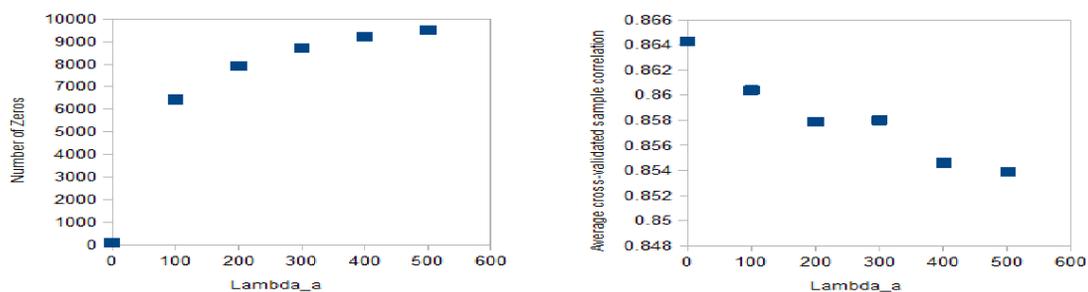


Figure 5.4: The cross-validated sample canonical correlation and the number zeros considering LT penalty function of the tuning parameters, λ_a ranging from 100 and 500

From Table 5.3, we could see that adding those genes without gene names have contributed to better correlation as compared to results thus obtained from Table 5.2. Despite increase in correlation, adding those gene names had resulted in poor sparse

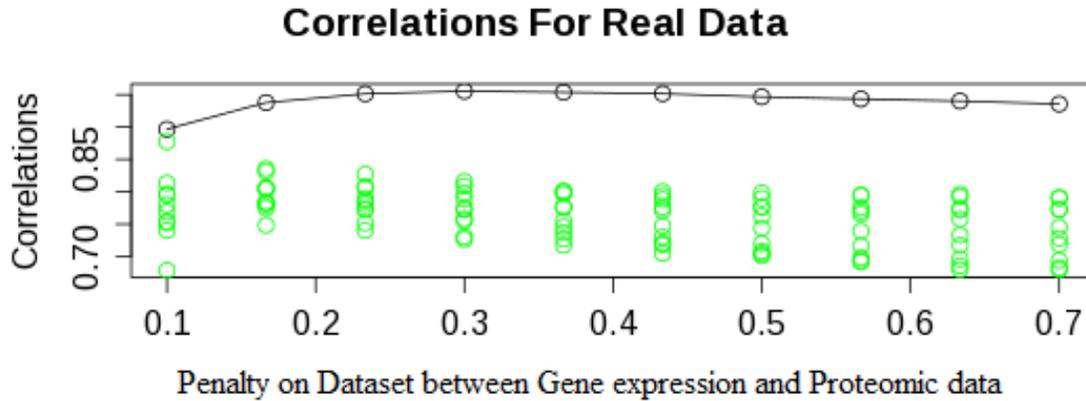


Figure 5.5: The canonical correlation of WT penalty function of the tuning parameters, λ_a and λ_b ranging from 0 and 0.7

variables whereas in Parkhomenko et al. (2009) approaches able to find better sparse solution than other methods. Lee et al. (2011) and Witten et al. (2009) method gives better sparseness for protein expressions. Parkhomenko et al. (2009) method provides with better sparseness to gene expressions. It can noticed that when λ_a is set to be 200 and optimum tuning parameter is found with 5-fold cross validation in Lee et al. (2011) method. The explanation of Figure 5.4 is similar to Section 5.4.1 and the optimal tuning parameter $(\lambda_a, \lambda_b) = (200, 300)$ that results with sample correlation of 0.8579 and number of zeros in gene pattern is 7909. For Witten et al. (2009) method at different levels of tuning parameters, the optimal parameters is selected based on high correlation obtained, which is shown in Figure 5.5.

Chapter 6

Discussion and Future directions

6.1 Discussion

CCA is one of the most important multivariate methods, which is commonly used for data visualization and dimension reduction. In high dimensional genomic data, the number of variables exceeds tens of thousands but very few samples are typically available, which is quite a challenge for statistician to find and interpret canonical correlation coefficients between two datasets. It is not biologically feasible to include all the variables in the solution as the variables may contain noise parameters, high-false positives and lack interpret-ability (Parkhomenko et al. 2007). This problem can be solved when we use sparse canonical correlation analysis (SCCA).

The main focus of this thesis is to compare different approaches available in finding sparse canonical correlation analysis (SCCA); I compared Parkhomenkho et al. (2009), Witten et al. (2009), and Lee et al. (2011) methods. SCCA performs

simultaneous analysis of two dataset at the same time to find a relationship between them. Using SCCA we can find linear combination of variables that are maximally correlated. In addition, we have seen in simulation results that when we applied these methods to the simulated data set only a small subset of the variables are included from each data set, which makes very easier to interpret. Simulation studies show that sparse CCA methods outperform the classical CCA method. It can be noticed that SCCA identifies that subset from the entire variables that are highly significant from both datasets in finding correlation between them. From various simulation studies, it can be found that the three methods that have been discussed outperform classical CCA in estimation and variable selection. In real data application, we have seen that Parkhomenko et al. (2009) method produces better sparse solution compared to other methods.

6.1.1 Simulation studies limitations

In my simulation studies, I used only one latent variable and limited myself with first r variables in each sets as highly correlated variables. In real data several such groups will be present in the dataset which requires several latent variables to be involved in the data. In my simulation result, it can be noticed that I consider only the first canonical correlation and for the above-mentioned situation which requires more than one pair of canonical vectors, which involves in several linear combinations. Moreover, calculating all the canonical correlation coefficient are very complex.

For example, in high dimensional genomic data, the number of variables involved will usually be tens of thousands whereas my simulation studies investigated only

up to 2000 variables. In addition, my sample size involves up to 500 samples as compared to the real data I have used containing only 59 samples.

6.2 Future directions

SCCA may not identify the key information of the data because of linearity, which can be identified with the help of Kernel CCA, it projects the data into a higher-dimensional vector space. It provides with high flexibility as the kernels can be generated from other kernels. In addition, in kernels, the data appears only through Gram matrix, which provides a greater advantage because the tuning parameters and updating time taken for them does not depend on the number of variables being considered (Haroon et al. 2003). There are limited work done in KCCA of high dimensional genome data, one among them is (Haroon et al. 2009) and not enough work has been considered in adding sparseness to KCCA, which is one of my interests for further work. Furthermore, there have not been much comparative work done in extending SCCA to more than two datasets.

Bibliography

- [1] Allison LA (2007). *Fundamental Molecular Biology*, First Edition. Wiley-Blackwell, Massachusetts.
- [2] Borga M, Landelius T, Knutsson H (1997). *A Unified Approach to PCA, PLS, MLR and CCA*. Report LiTH-ISY-R-1992, ISY, SE-581 83 Linköping, Sweden.
- [3] Bhatia VK (2007). *Canonical Correlation Analysis*. URL [iasri.res.in/ebook/EBADAT/4-Applications of Multivariate Techniques/3-canonical correlation.pdf](http://iasri.res.in/ebook/EBADAT/4-Applications_of_Multivariate_Techniques/3-canonical_correlation.pdf).
- [4] Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the USA*. ;104:19428–19433.
- [5] Fisher RA (1936): The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7: 179–188.

- [6] Gonzalez I and Djean S (2009). CCA: Canonical correlation analysis. R package version 1.2. <http://CRAN.R-project.org/package=CCA>.
- [7] Good IJ (1969). Some applications of the singular decomposition of a matrix. *Technometrics*, 11(4):828–831.
- [8] Haroon D, Szedmak S, and Shawe-Taylor J (2003). Canonical correlation analysis; an overview with application to learning methods, Technical Report, Royal Holloway University of London.
- [9] Haroon D, and Shawe-Taylor J (2007) . Sparse canonical correlation analysis (Technical report). UK: University College London.
- [10] Haroon D, Ettinger U, Mouro-Miranda J, Antonova E, Collier D, Kumari V, Williams S, Brammer M (2009). Correlation-based multivariate analysis of genetic influence on brain volume, *Neuroscience Letters*, Volume 450, Issue 3, , Pages 281-286.
- [11] Hoerl AE (1962). Application of Ridge Analysis to Regression Problems. *Chemical Engineering Progress*. 58, 54-59.
- [12] Hotelling H (1936). Relations between two sets of variates. *Biometrika*, 28:312–377.
- [13] Johnson RA and Wichern DW (2007). *Applied Multivariate Statistical Analysis*, Sixth Edition. Pearson Prentice Hall, New Jersey.
- [14] Johnstone IM and Lu AY (2004). Sparse principal component analysis. (Unpublished manuscript).

- [15] Le Cao KA, Martin PG, Robert-Granie C, and Besse P (2009). Sparse canonical methods for biological data integration: application to a cross platform study. *BMC Bioinformatics* 10, 10-34.
- [16] Lee W, Lee D, Lee Y., and Pawitan Y (2011). Sparse Canonical Covariance Analysis for High-throughput Data. *Statistical Applications in Genetics and Molecular Biology*: Vol. 10: Iss. 1, Article 30.
- [17] Lee W, Lee D, Lee Y, and Pawitan Y (2011). Sparse canonical covariance analysis. R package <http://www.meb.ki.se/~yudpaw/>.
- [18] Mardia KV, Kent JT, and Bibby JM (1979). *Multivariate analysis*. New York: Academic Press.
- [19] Nishizuka S. et al. (2003): Proteomic profiling of the NCI-60 Cancer Cell Lines Using New High-density Reverse-phase Lysate Microarray. *PNAS* 100, 14229-14234.
- [20] Parkhomenko E, Trichtler D, and Beyene J (2007). Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC proceedings*, 1:S119.
- [21] Parkhomenko E, Tritchler D, and Beyene J (2009). Sparse Canonical Correlation Analysis with Application to Genomic Data Integration. *Statistical Applications in Genetics and Molecular Biology*: Vol. 8: Iss. 1, Article 1.

- [22] Pearson K (1901). On lines and planes of closest fit to systems of points in space, *Philosophical Magazine, Series 6*, vol. 2, no. 11, pp. 559-572.
- [23] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- [24] Shankavaram U.T. et al. (2007): Transcript and protein expression profiles of the NCI-60 cancer cell panel:an intergenomic microarray study. *Molecular Cancer Therapeutics*. 6, 820-832.
- [25] Tibshirani R (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, 58, 267-288.
- [26] Waaijenborg S and Zwinderman AH (2009). Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks. *BMC Bioinformatics*, 10:315.
- [27] Waaijenborg S, Verselewe de Witt Hamer PC, and Zwinderman AH (2008). Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis. *Statistical Applications in Genetics and Molecular Biology: Volume 7: Issue 1*, Article 3.
- [28] Wegelin JA (2000). A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. Technical Report 371 , Department of Statistics, University of Washington, Seattle.

- [29] Witten DM and Tibshirani R (2009) Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. *Statistical Applications in Genetics and Molecular Biology*: Vol. 8: Iss. 1, Article 28.
- [30] Witten DM, Tibshirani R, and Hastie T (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10(3): 515-534.
- [31] Witten D M, Tibshirani R, and Gross S (2011). PMA: Penalized Multivariate Analysis. R package version 1.0.7.1. <http://CRAN.R-project.org/package=PMA>.
- [32] Wold H (1966). Nonlinear estimation by iterative least squares procedures, in David, F. N. (ed.), *Research Papers in Statistics, Festschrift for J. Neyman*, Wiley, New York, 411-444.
- [33] Wold H (1982). Soft modeling: the basic design and some extensions. In H.Wold K.G. Joreskog, editor, *Systems under indirect observation: causality, structure, prediction, Part II*, number 139 in *Proceedings of the Conference on Systems Under Indirect Observation*, pages 1–54, Cartigny, Switzerland, North Holland.
- [34] Wold H (1985). Partial Least Squares, *Encyclopedia of the statistical sciences*. Wiley, 581–591.
- [35] Zou H and Hastie T (2005). Regularization and variable selection via

the elastic net. *Journal of the Royal Statistical Society, Series B*, 567(2):301–320.

[36] Zou H, Hastie T, and Tibshirani R (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15, pp. 265–286.

[37] Zou H (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476),1418–1429, 2006.