Canonical Correlation Analysis to relate a Genomic Dataset with a Neuroimage Dataset.

Augustine Annan
(10551764)

THIS THESIS IS SUBMITTED TO THE UNIVERSITY OF GHANA, LEGON IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF MPHIL MATHEMATICS DEGREE

July, 2016

DECLARATION

This thesis was written in the Department of Mathematics, University of Ghana, Legon from September 2015 to July 2016 in partial fulfillment of the requirements for the award of Master of Philosophy degree in Mathematics under the supervision of Dr. Margaret McIntyre, Dr. Douglas Adu-Gyamfi, and Dr. Eyram Schwinger of the University of Ghana

I hereby declare that except where due acknowledgement is made, this work has never been presented wholly or in part for the award of a degree at the University of Ghana or any other University.

Signature: .....
Student: Augustine Annan

Signature: .....
Dr. Margaret McIntyre

Signature: .....
Dr. Douglas Adu-Gyamfi

DEDICATION

I dedicate my research project to my family. A special feeling of gratitude to my loving mother, Agnes Esuon whose words of encouragement and push for tenacity ring in my ears. My brothers Stephen and Humphrey, my sister Faustina and my friend Ansbertha have never left my side and are very special.

ACKNOWLEDGEMENTS

My warmest appreciation goes to my supervisors, Dr. Margaret McIntyre and Dr. Alessandro Crimi, for the patience, motivation, immense knowledge and continuous support and guidance he offered me throughout this project. Also to my other supervisors Dr. Douglas Adu-Gyamfi and Dr. Eyram Schwinger, I show great appreciation for taking much time to assist me in this work with so much patience.

I want to appreciate the African Institute for Mathematical Sciences (AIMS-Ghana), for supporting this research financially.

To the Head of Department, Dr. Margaret McIntyre; and all the lecturers, I say a big thank you for giving me such a great opportunity to step up my goals in academia.

To my mother, and siblings, I am grateful for your unconditional love, support and encouragement. My sincere, heartfelt gratitude goes to all my colleagues for all their encouragement and fun moments.

To God be the glory.

ABSTRACT

This thesis investigates the relationship between copy number variations and neuro-image features of Glioblastoma patients. Canonical correlation analysis was employed to elicit these relationships. This thesis highlights some of the concepts of the technique which enabled us to obtain our main results. We found three pairs of significant canonical variates with correlations of $0.6704, 0.6347$ and 0.5552 respectively, which was used to identify genes and neuro-image features related to Glioblastoma.

Contents

Declaration i

Dedication ii

Acknowledgements iii

Abstract iv

1 Introduction 1

1.1 Organisation of the Study 6

2 Definitions 7

2.1 Definitions of statistical and mathematical terms 7

3 Methodology 12

3.1 Canonical Correlation Analysis (CCA) 12

3.1.1 Canonical Correlation 12

3.1.2 Mathematical Formulation 14

3.1.3 Formulation and Derivation of the Canonical Variables 14

3.1.5 Properties of the Canonical Variable Pairs 22

3.1.6 Canonical correlation coefficient under the non-singular transformation 24

3.1.7 Correlation Coefficient Between Canonical Variables and the Original Variables . . . 26
3.1.8 Computation of Canonical Correlation Coefficient Using Standardized Variables . . . 28
3.1.9 Assessing Overall Model Fit and Canonical Dimension Reduction . 30
3.2 Example: Computation of Canonical variables and Canonical Coefficients . 35

4 Results **40**

4.1 Data . . . 40
4.1.1 Patient Features . . . 40
4.2 Preliminaries . . . 43
4.3 Main Results . . . 45
4.3.1 Correlation matrix of variables . . . 45
4.3.2 Assessment of Overall Model Fit . . . 51
4.3.3 Interpreting Canonical Variate Pairs . . . 54
4.3.4 Interpretation of Canonical Variate Using Canonical Weights . . . 55
4.3.5 Interpretation of Canonical Variate Using Canonical Loadings . . . 58
4.3.6 Cross Validation . . . 60
4.3.7 CCA on Sub-Sample A . . . 60
4.3.8 CCA on Sub-Sample B . . . 63
4.4 Summary . . . 65

5 Conclusion **67**

References **71**

List of Tables

4.1 Description of Neuro-image Features Used 42

4.2 Copy Number Variation Variables (Genes) 43

4.3 Sex and Survival Status Distribution of Patients 44

4.4 Age and Overall Survival Time of Patients 44

4.5 Frequency Distribution of Expression Subtype 45

4.6 Correlations for Variable Set 1 46

4.7 Correlations for the Copy Number Variation Variables 47

4.8 Correlations for the Copy Number Variation Variables 48

4.9 Correlations between Variable Set 1 and Variable Set 2 49

4.10 Raw Coefficients for the Neuro-image features 50

4.11 Raw Coefficients for the Copy Number Variation Variables 51

4.12 Test of Significance of all Canonical Correlations 52

4.13 Test of Significance of each Canonical Correlation 53

4.14 Canonical Correlations and Eigenvalues 53

4.15 Canonical redundancy analysis for Canonical Correlations 54

4.16 Standardized Coefficients for the Neuro-image features 56

4.17 Standardized Coefficients for the Copy Number Variation Variables 57

4.18 Summary of Important Related Variables 58

4.19 Canonical Loadings for the Neuro-image features 58

4.20 Canonical Loadings for the Copy Number Variation Variables 59

4.21 Summary of Important Related Variables 60
4.22 Test of Significance of each Canonical Correlation 61
4.23 Canonical Loadings for the Neuro-image features 62
4.24 Canonical Loadings for the Copy Number Variation Variables 62
4.25 Test of Significance of each Canonical Correlation 63
4.26 Canonical Loadings for the Neuro-image features 64
4.27 Canonical Loadings for the Copy Number Variation Variables 65

List of Figures

1.1 [The gene amplification has created a copy number variation.]The chromosome now has two copies of this section of DNA, rather than one [34]. . . . 3

1.2 [Magnetic Resonance Imaging (MRI) images of patients with GBM][37, 13] 4

1.3 [Fully automated Segmentation and VASARI Feature Extraction:]necrotic core/contrast enhancing tumor(right) and edema(left) [37] 5

Chapter 1

Introduction

Many complex diseases result from the interplay of genetics and neuroimage features. As such understanding the underlying biological mechanism of such datasets are very important. As a result of the emergence of increasing development of a wide range of genome-wide assays, it is now possible for multiple measures of genomic markers from various platforms for a particular subject such as single nucleotide polymorphism, gene expression, copy number variation and so on. These measurements relay information about variations of genome. Putting together two or more types of data does not only help in the diagnosis of diseases but it does enhance comprehension of the biological mechanisms and consequently could improve treatment strategies. So there is a high demand for integrative approaches for use in large-scale genomic data analysis. Therefore, investigating the associations between such entities is of great use.

Glioma is the most common type of primary brain tumor which arises from glial cells. It is considered responsible for approximately 13000 deaths in the United States and more than 14000 in Europe each year [35]. Gliomas are heterogeneous and they can be classified in accord with their grade: low-grade glioma, anaplastic glioma, and glioblastoma. The most common type of glioma in adults is glioblastoma (GBM). It is generally diagnosed at an average age of 55 years, and gives the affected patient an average survival time of only 10 to 18 months. Lower grade glioma can occur at younger ages [35]. The underlying tumor pathology and biological function can be identified by imaging and genetic biomarkers. In the context of clinical routing, if imaging phenotypes of GBM from magnetic resonance imaging (MRI) can be easily associated with specific gene expression signatures, they will serve as a non-invasive alternative to biopsy, providing important information for diagnosis, prognosis and personalized treatment. Therefore this thesis seeks to investigate the corre-

spondence between genetic data, in particular the copy number variations and the imaging phenotypes of the GBM.

One of the most important means of acquiring the relationships between two or more entities or objects is to take measurements of pertinent relationships. A measure of a relationship depicts the strength of the relationship or association between the objects. So we introduce the term correlation to mean any broad class of statistical relationships depicting dependence. The degree of correlation can be measured by the use of correlation coefficients, denoted by ρ or r. The most used coefficient is the measure developed by Karl Pearson which is the Pearson correlation coefficient. The core of the project is to present the idea of canonical correlation analysis and use it to investigate the relationship between the copy number variations and neuroimage features. The main highlights of the technique that helps to elicit the relationship between the datasets will be discussed. In the next two paragraphs we introduce copy number variations and the neuroimage features of tumors.

Copy number variation (CNV) can be defined as alterations of the deoxyribonucleic acid (DNA) of a genome that makes the cell have an abnormal repetitions and deletions of one or more sections of the DNA [10]. The number of repetitions of such sections differs between individuals in the human population [23]. It is a kind of structural variation, precisely a kind of duplication event that highly affects a number of base pairs [34]. Human beings differ in the number of copies of each gene and this leads to the idea of copy number invariants. Recent research has shown that about two thirds of the entire human genome comprises of repeats [36] and also about $4.75 - 9.46\%$ of the entire genome can be described as copy number variations [39]. CNVs play a very notable role in producing the necessary variation in the population and also in disease phenotype [23].

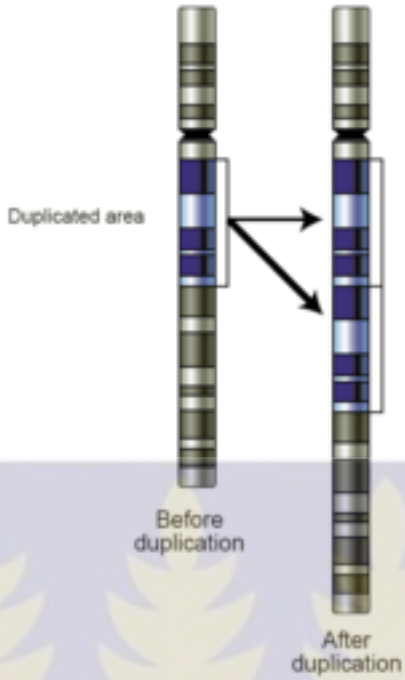

Figure 1.1: [The gene amplification has created a copy number variation.]The chromosome now has two copies of this section of DNA, rather than one [34].

Humans have two copies of most genes, one from the mother's chromosome and the other from the father's chromosome. Some alterations in the chromosome may cause either a loss or a gain of one copy. Duplications and deletions of more than 1000 nucleotides are referred to as copy number variants [3]. It is considered to be a very notable risk factor for cancer and constitutes a wide spectrum of the total genomic variation [38]. There has been an identification of recurrent copy number variations that demonstrate that various chromosome regions are present. Also, as a result of cancer being an acquired disease and also because inherited factors play a major role in its occurrence, there have been comparisons of the early constitutional copy number alterations with the copy number variations present in tumor biopsy [12].

GBM is an aggressive tumor with poor prognosis. Despite the introduction of new strategies to treat the disease, the median survival is less than one year [12]. In recent studies, important features have been identified. The pediatric primary GBM is different from the adult GBM, considering both the genetic profiling and mean commulative survival [29, 28, 9, 30]. Pediatric GBM and adult GBMs have varying pathways of tumorigenesis [30]. In $35-50\%$ of the time, a primary adult patient forms present amplification of

the epidermal growth factor receptro (EGRF) gene and inactivation of the phosphatase and tensin homolog (PTEN) gene [26, 8]. However, in the secondary adult GBM patients that may evolve from low-grade lesions, normally have no alterations of gene PTEN and no EGFR duplications but most often have TP53 mutations [33]. Studies have shown that there are differences in CNV between the adult GBMs and childhood GBMs. In pediatric GBMs, heterozygous deletions are more common while duplications are more frequent in adult GBMs [32].

Analyzing imaging features has revealed interesting relationships between the imaging features and survival of patients. Considering patients with malignant gliomas, some tumor imaging features and clinical data such as age, perioperative karnofsky performance status and tumor resection have been established to correlate with survival [31]. The image features include necrosis and edema. According to Pope et. al [31], edema, noncontrast-enhancing tumor (nCET) and multifocality were the significant features related to survival and these features could be classified as prognostic indicators.

There have been several studies on the relationship between imaging features and survival. Consequently, there are reports that, the level of edema and the degree of necrosis are correlated with survival negatively [27, 21, 16].

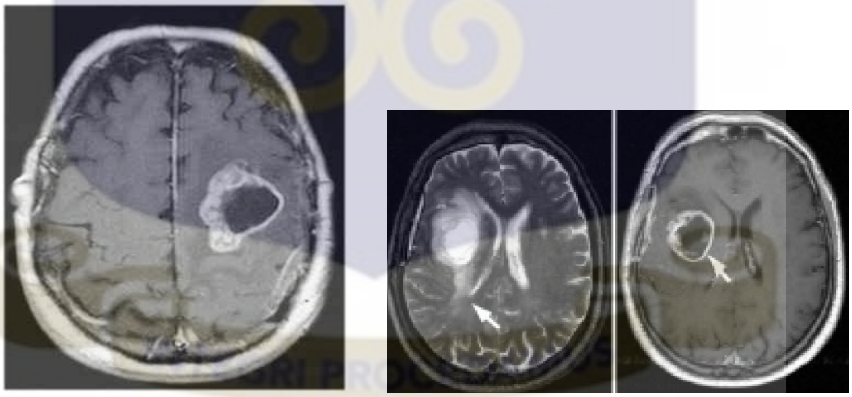

Figure 1.2: [Magnetic Resonance Imaging (MRI) images of patients with GBM][37, 13]

The importance of imaging has made it necessary for the availability of accurate informative quantities. The Visually AcceSAble Rembrandt Images (VASARI) feature set presents actual standards by which a numeric score can be associated to a feature that will enable the description of the degree of tumor features. It is a standard imaging feature consisting of 30 features describing the size, location and the appearance of the MRI image set. The

image presents the global view of the tumor. A small tumor in the frontal lobe has a vastly different outcome to a small tumor adjacent to motor area, for instance the eloquent cortex [13]. For more accurate results, the Columbia University Medical Center [37], designed a fully automated computer algorithm to score glioma tumors based on the available feature set.

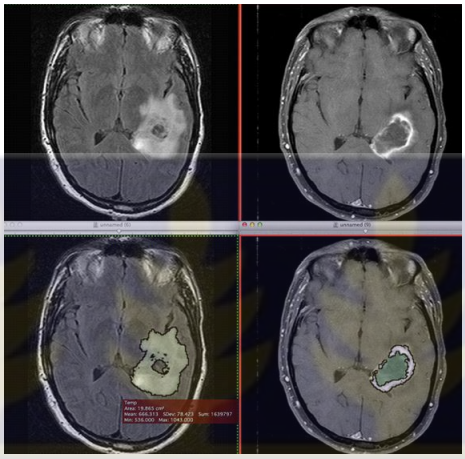

Figure 1.3: [Fully automated Segmentation and VASARI Feature Extraction:]necrotic core/contrast enhancing tumor(right) and edema(left) [37]

Image features have also been used for exploratory radiogenomic analysis [11]. Gevaert et. al obtained quantitative image features from MR images that characterize the radiographic phenotype of GBM lesions. They also constructed radiogenomic maps relating the features with particular molecular data [11]. Even after the consideration of clinical variables, imaging features provide notable prognostic information. Currently, qualitative work suggests an association between imaging phenotypes and genotypes [13].

Dongdong Lin et al (2013) [22] investigated the correspondence between single nucleotide polymorphism (SNP) and brain activity measured by functional magnetic resonance imaging (fMRI) to understand how genetic variation influences the brain activity. They developed a group sparse canonical correlation analysis method to explore the relationship between these two datasets. They found two pairs of significant canonical variates with average correlations of 0.4527 and 0.4292 respectively, which were used to identify genes and voxels associated with schizophrenia.

1.1 Organisation of the Study

Chapter 2 will present brief definitions of some of the mathematical and statistical terms that will be used in this work. The review of the main technique to be employed to investigate the relationships will be discussed in Chapter 3. In Chapter 4, results from the analysis of the data will be presented and discussion will follow in chapter 4. Chapter 5 will contain the conclusions and recommendations and a brief discussion of possible directions for the future work.

Chapter 2

Definitions

Prior to the presentation and discussion of the existing technique and methodology, this chapter will present some definitions of concepts, terms and theorems to be used in the sequel.

2.1 Definitions of statistical and mathematical terms

Definition 2.1.1.
Supposing we have a square matrix, A, of size m, then the $m \times 1$ vector k is a *right eigenvector* for A and $\lambda \geq 0$ is the corresponding *eigenvalue* if $Ak = \lambda k$. Also, a *left eigenvector* n can be defined as satisfying $nA = \lambda n$.

Definition 2.1.2.
Given an $m \times m$ matrix B, a matrix M for which $M^2 = B$ is called the *square root of the matrix B*.

Several studies have examined the computation of matrix square roots [17, 6, 7, 18, 4]. Here we find the square root of an $m \times m$ matrix by the diagonalization method [4].

An $m \times m$ matrix B is diagonalizable if we have a diagonal matrix D and an invertible matrix K such that $B = KDK^{-1}$. The diagonal matrix is made up of the eigenvalues of B and the columns of K are the m eigenvectors of B. The square root of B is given as

$$B^{\frac{1}{2}} = K\sqrt{D}K^{-1}$$

Example 2.1.3.

Given a matrix $B = \begin{pmatrix} 18 & 12 \\ 12 & 28 \end{pmatrix}$, we find $B^{\frac{1}{2}}$ as follows.

The eigenvalues of B are $10, 36$ and eigenvectors are $(-3,2),(2,3)$, so B eigendecomposes to

$$B = \begin{pmatrix} -3 & 2 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} 10 & 0 \\ 0 & 36 \end{pmatrix} \begin{pmatrix} -3 & 2 \\ 2 & 3 \end{pmatrix}^{-1}$$

So we have the form $B = KDK^{-1}$. Since from Definition 2.1.2, $M^2 = B$, then there is an M of the form $K\sqrt{D}K^{-1}$

$$\begin{aligned} M &= \begin{pmatrix} -3 & 2 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} \sqrt{10} & 0 \\ 0 & \sqrt{36} \end{pmatrix} \begin{pmatrix} -3 & 2 \\ 2 & 3 \end{pmatrix}^{-1} \\ \sqrt{B} &= \begin{pmatrix} 4.035 & 1.310 \\ 1.310 & 5.127 \end{pmatrix} \end{aligned}$$

Definition 2.1.4.

Let $X_1, \ldots, X_p$ be a set of $n \times 1$ vectors. Then we have that the $n \times 1$ vector l_x is a *linear combination of these vectors* if $l_x = a_1X_1 + \ldots + a_pX_p$ for some real constants $a_1, \ldots a_p$ which are usually called loadings.

Singular Value Decomposition

Let A be a $p \times q$ real matrix. Then it can be represented as $A = UDV'$ where U is a $p \times p$ orthogonal matrix, V is a $q \times q$ orthogonal matrix and D is a $p \times q$ diagonal matrix with non-negative diagonal elements $\lambda_i, i = 1, \ldots, min(p,q)$. The first $min(p,q)$ columns of U and V are left and right singular vectors, respectively, and $\lambda_i, i = 1, ..., min(p,q)$ are the corresponding singular values. Note that left singular vectors for A are the eigenvectors for AA' while the right singular vectors are the eigenvectors for $A'A$. The eigenvalues are equal for AA' and $A'A$ and they are equal to the squared singular values of A.

Lemma 2.1.5. *(The Cauchy-Schwartz Inequality)*
Let H be a Hilbert space over $\mathbb{C}$. We have that

$$| \langle x,y \rangle |^2 \leq \langle x,x \rangle \langle y,y \rangle ,$$

$\forall x, y \in H$.

Proof. If $y = 0$, then $\langle x, 0\rangle = 0$ and the inequality is true. Assume $y \neq 0$ and that

$$a = -\frac{\langle x, y\rangle}{\langle y, y\rangle}.$$

Clearly a is a complex number since $\langle x, y\rangle$ is a complex number and $\langle y, y\rangle$ is a real number. Then we have,

$$\begin{aligned}
0 \leq \langle x+ay, x+ay\rangle &= \langle x, x+ay\rangle + \langle ay, x+ay\rangle \\
&= \langle x,x\rangle + \langle x,ay\rangle + \langle ay,x\rangle + \langle ay,ay\rangle \\
&= \langle x,x\rangle + \bar{a}\langle x,y\rangle + a\langle y,x\rangle + a\langle y,ay\rangle \\
&= \langle x,x\rangle + \bar{a}\langle x,y\rangle + a\langle y,x\rangle + a\overline{\langle ay,y\rangle} \\
&= \langle x,x\rangle + \bar{a}\langle x,y\rangle + a\langle y,x\rangle + a\bar{a}\langle y,y\rangle \\
&= \langle x,x\rangle + \bar{a}\langle x,y\rangle + a\langle y,x\rangle + |a|^2\langle y,y\rangle \\
&= \langle x,x\rangle - \frac{\overline{\langle x,y\rangle}}{\langle y,y\rangle}\langle x,y\rangle - \frac{\langle x,y\rangle}{\langle y,y\rangle}\overline{\langle x,y\rangle} + \left|-\frac{\langle x,y\rangle}{\langle y,y\rangle}\right|^2\langle y,y\rangle \\
&= \langle x,x\rangle - \frac{2\overline{\langle x,y\rangle}}{\langle y,y\rangle}\langle x,y\rangle + \frac{|\langle x,y\rangle|^2}{\langle y,y\rangle} \\
&= \langle x,x\rangle - \frac{2|\langle x,y\rangle|^2}{\langle y,y\rangle} + \frac{|\langle x,y\rangle|^2}{\langle y,y\rangle} \\
&= \langle x,x\rangle - \frac{|\langle x,y\rangle|^2}{\langle y,y\rangle}.
\end{aligned}$$

Hence,

$$\begin{aligned}
0 &\leq \langle x,x\rangle - \frac{|\langle x,y\rangle|^2}{\langle y,y\rangle} \\
|\langle x,y\rangle|^2 &\leq \langle x,x\rangle\langle y,y\rangle \\
|\langle x,y\rangle| &\leq \sqrt{\langle x,x\rangle}\sqrt{\langle y,y\rangle} \\
|\langle x,y\rangle|^2 &\leq \langle x,x\rangle\langle y,y\rangle \quad \text{as desired.}
\end{aligned}$$

□

Definition of Statistical Terms

Definition 2.1.6.
Variance measures the spread or dispersion or compactness of a set of data. It is computed as the average of the squared deviations from the mean score of the data set.

Definition 2.1.7.
Covariance is a measure of how much or the degree at which two variables change together. The *covariance matrix* is a matrix which has the covariance of the ith and jth elements of the variables in the position of the ijth position . All covariance matrices are symmetric and positive semi-definite.

The following definitions are adapted from the supplement to Hair et. al's textbook [14].

Definition 2.1.8.
A canonical variate also known as *a linear compound* or *a linear composite* is a linear combination that constitutes the weighted sum of two or more variables. Thus a canonical variate can be defined for either set of variables.

Definition 2.1.9.
A Canonical function depicts the relationship between two canonical variates (linear composites). For each canonical function, there are two canonical variates, one variate for one set of variables and another variate for the other set of variables. The degree of the relationship is the canonical correlation.

Definition 2.1.10.
The canonical roots are the squared canonical correlations. They are also known as eigenvalues. The canonical roots provide the estimation of the shared variance between the weighted canonical variates of the two set of variables.

Definition 2.1.11.
Orthogonality here is a mathematical constraint which specifies that canonical functions are not dependent of one another. Put differently, to arrive at statistical independence of the canonical functions we derive the functions so that each function is perpendicular to all others when it is being plotted in a space (multivariate).

Definition 2.1.12.
The canonical loading is the measure of correlation between the original variables and their canonical variates.

Definition 2.1.13.

The redundancy index is the measure of the amount of variance explained between a canonical variate pair in a canonical function.

Chapter 3

Methodology

In this chapter, we present the idea of Canonical Correlation Analysis. The technique seeks to identify the relationships between two datasets. The canonical correlation analysis will be presented in Section 1 and an example will be illustrated in section 2. The discussion of the technique will be skewed towards the datasets involved for this thesis. The main references used for this chapter are [20, 15, 24].

3.1 Canonical Correlation Analysis (CCA)

3.1.1 Canonical Correlation

Canonical correlation analysis is a technique that measures the relationship between two multidimensional variables. It seeks to find two bases in which the correlation matrix between the variables is diagonal and the correlations on the diagonal are maximized.

CCA was first introduced by H. Hotelling in 1936 [19]. Canonical correlation is invariant with respect to affine transformations of the variables. This property differentiates it from the normal correlation analysis. Adopting CCA helps to summarize relationships while preserving main features. CCA enables us to summarize the relationships into fewer number of statistics while preserving the main facets of the relationships.

We begin with the following notation:

we define two vectors X and Y as two sets of variables, where X consists of p variables and Y consists of q variables. We select X and Y depending on the number of variables in each set so that $p \leq q$ for computational reasons and convenience.

So

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \quad \text{and} \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{pmatrix} \tag{3.1}$$

We define a set of linear combinations, M and N. M will consist of linear combinations of variables X_i in X, and N will consist of linear combinations of variables Y_j in Y. We have

$$\begin{aligned} M_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ M_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \\ M_p &= a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p = a'X \\ \\ N_1 &= b_{11}Y_1 + b_{12}Y_2 + \cdots + b_{1q}Y_q \\ N_2 &= b_{21}Y_1 + b_{22}Y_2 + \cdots + b_{2q}Y_q \\ &\vdots \\ N_p &= b_{p1}Y_1 + b_{p2}Y_2 + \cdots + b_{pq}Y_q = b'Y. \end{aligned}$$

We also define (M_i, N_i) as the i^{th} canonical variate pair. So (M_1, N_1) is the first canonical variate pair, and (M_2, N_2) is the second canonical variate pair and so on. There are p canonical variate pairs.

We seek to find linear combinations that maximize the correlations between the members of each canonical variate pair.

The correlation $\text{corr}(M_i, N_j)$ between M_i and N_j is then calculated using (3.2):

$$\text{corr}(M_i, N_j) = \frac{\text{cov}(M_i, N_j)}{\sqrt{\text{var}(M_i)\text{var}(N_j)}}, \tag{3.2}$$

where $\text{cov}(M_i, N_j)$ is the covariance between M_i and N_j and $\text{var}(M_i)$ and $\text{var}(N_j)$ are the variances of M_i and N_j respectively. The canonical correlation for the i^{th} canonical variate pair is simply the correlation between M_i and N_i:

$$\rho_i = \frac{\text{cov}(M_i, N_i)}{\sqrt{\text{var}(M_i)\text{var}(N_i)}}. \tag{3.3}$$

The quantity in (3.3) is to be maximized, thus we find linear combinations of the $X_i's$ and linear combinations of the $Y_j's$ that maximize the above correlation.

So the **main purpose of canonical correlation analysis** is to explain the covariance structure or correlations structure between two sets of random vectors in terms of fewer linear combinations.

3.1.2 Mathematical Formulation

The p-dimensional random vector X and q-dimensional vector Y, are such that $cov(X,X), cov(Y,Y)$ and $cov(X,Y)$ are denoted by Σ_{11}, Σ_{22} and Σ_{12} respectively. So, the covariance structure of X and Y is given as

$$cov\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Considering the linear combinations $a'X$ and $b'Y$, we have that

$$cov(a'X, b'Y) = a' \sum_{12} b.$$

This implies that the canonical correlation of X and Y is

$$\rho(a'X, b'Y) = \frac{a' \Sigma_{12} b}{\sqrt{a' \Sigma_{11} a \times b' \Sigma_{22} b}}.$$

3.1.3 Formulation and Derivation of the Canonical Variables

The canonical variables and associated correlation coefficients are defined iteratively.

1st Pair of Canonical Variables:

Definition: Consider $M_1 = a'X$ and $N_1 = b'Y$ such that

- $var(M_1) = var(N_1) = 1$ and

- $\rho(M_1, N_1) = \max_{a,b} \quad \rho(a'X, b'Y)$,

then (M_1, N_1) is the 1st pair of canonical variables (canonical variate) and

$\rho_1 = \max_{a,b} \quad \rho(a'X, b'Y)$ is the *1st canonical correlation coefficient*.

2nd pair of Canonical Variables:

Definition: Consider linear combinations $a'X$ and $b'Y$ such that

- $cov(a'X, M_1) = 0 = cov(b'Y, N_1)$, that is M_1 is uncorrelated with the linear combinations $a'X$ and N_1 is uncorrelated with $b'Y$ and

- $var(a'X) = var(b'Y) = 1$

Then maximize the correlations between $a'X$ and $b'Y$ such that the above is satisfied. The maximizing $a'X$ and $b'Y$ are called the second pair of canonical variates. The correlation coefficient that maximizes the correlation of the second canonical variate pairs is the *second canonical correlation coefficient*.

K^{th} pair of Canonical Variables:

Definition: The K^{th} pair of canonical variables are the linear combinations (M_k, N_k) having unit variance which maximize the correlation among all possible linear combinations uncorrelated with the previous $(k-1)$ canonical variate pairs.

The following statements will help us in the derivation of the canonical variables.

$$\begin{aligned} cov(X, X) &= \sum_{11} > 0, \\ cov(Y, Y) &= \sum_{22} > 0. \end{aligned}$$

The covariance structure is positive definite. Now we consider a $p \times q$ matrix, A such that

$$A = \sum_{11}^{-\frac{1}{2}} \sum_{12} \sum_{22}^{-\frac{1}{2}}$$

and we now consider the following matrices

$$AA' = \sum_{11}^{-\frac{1}{2}} \sum_{12} \sum_{22}^{-1} \sum_{21} \sum_{11}^{-\frac{1}{2}} (p \times p)$$
$$A'A = \sum_{22}^{-\frac{1}{2}} \sum_{21} \sum_{11}^{-1} \sum_{12} \sum_{22}^{-\frac{1}{2}} (q \times q)$$

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, be the eigenvalues of AA' and let $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_q$, be the eigenvalues of $A'A$.

We have that,
(i) $A'A$ and AA' are positive semi definite implies that $\lambda_i \geq 0$ and $\gamma_j \geq 0 \quad \forall i, j$.
(ii)Non-zero eigenvalues of AA' are same as the non-zero eigenvalues of $A'A$ and the eigenvalue 0 has different multiplicities in AA' and $A'A$ if $q < p$.

Theorem 3.1.4. *[20] We suppose that $p \leq q$ and cov* $\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$.

Considering the linear combinations $M = a'X$ and $N = b'Y$, we have that

$$\max_{a,b} \quad \rho(a'X, b'Y) = \rho_1$$

is attained by the linear combination

$$M_1 = e_1' \sum_{11}^{-\frac{1}{2}} X \quad \text{and} \quad N_1 = f_1' \sum_{22}^{-\frac{1}{2}} Y.$$

M_1 and N_1 are the first pair of canonical variables and

$$\max_{a,b} \quad \rho(a'X, b'Y) = \rho_2$$

is attained by the linear combination

$$M_2 = e_2' \sum_{11}^{-\frac{1}{2}} X \quad \text{and} \quad N_2 = f_2' \sum_{22}^{-\frac{1}{2}} Y.$$

M_2 and N_2 are the second pair of canonical variables.

In general

$$\max_{a,b} \quad \rho(a'X, b'Y) = \rho_k$$

is attained by the linear combination

$$M_k = e_k' \sum_{11}^{-\frac{1}{2}} X \quad \text{and} \quad N_k = f_k' \sum_{22}^{-\frac{1}{2}} Y.$$

Now $(\rho_1)^2 \geq (\rho_2)^2 \geq \ldots \geq (\rho_p)^2$ *are the eigenvalues of the matrix* $\sum_{11}^{-\frac{1}{2}} \sum_{12} \sum_{22}^{-\frac{1}{2}} \sum_{21} \sum_{11}^{-\frac{1}{2}}$ *matrix and* $e_1, e_2, \ldots, e_p$ *are the orthonormalized eigenvectors corresponding to* $(\rho_1)^2, \ldots (\rho_p)^2$.

The values $(\rho_1)^2, (\rho_2)^2, \ldots \geq (\rho_p)^2$ *are the p largest eigenvalues of the matrix*

$$\sum_{22}^{-\frac{1}{2}} \sum_{21} \sum_{11}^{-1} \sum_{12} \sum_{22}^{-\frac{1}{2}}$$

with eigenvectors $f_1, f_2, \ldots, f_p$, *where each* f_i *is proportional to* $\sum_{22}^{-\frac{1}{2}} \sum_{21} \sum_{11}^{-\frac{1}{2}} e_i$.

Derivation of the 1st pair of canonical variables

Proof. From the definitions, we have that

$$\rho(a'X, b'Y) = \frac{a' \sum_{12} b}{(a' \sum_{11} a b' \sum_{22} b)^{\frac{1}{2}}}. \tag{3.4}$$

$$\text{We let} \quad \sum_{11}^{\frac{1}{2}} a = u \implies a = \sum_{11}^{-\frac{1}{2}} u$$
$$\text{and let} \quad \sum_{22}^{\frac{1}{2}} b = v \implies b = \sum_{22}^{-\frac{1}{2}} v.$$

So, equation 3.4 becomes

$$\rho(a'X, b'Y) = \frac{u' \sum_{11}^{-\frac{1}{2}} \sum_{12} \sum_{22}^{-\frac{1}{2}} v}{((u'u)(v'v))^{\frac{1}{2}}}.$$

By applying the Cauchy Schwartz inequality, we have that

$$u'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}v \leq \left(u'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}u\right)^{\frac{1}{2}}\left(v'v\right)^{\frac{1}{2}}. \tag{3.5}$$

We make use of the following result to find an upper bound of the expression on the right.

From matrix theory, if $C(p\times p)$ is a real symmetric matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ and eigenvectors orthornormalised at $e_1,\ldots,e_p$, then we have the following result

$$\max_d \quad \frac{d'Cd}{d'd} = \lambda_1,$$

where λ_1 is the largest eigenvalue of the real symmetric matrix C and d is a vector. The maximum is attained at $d = e_1$, where e_1 the orthonormalised eigenvector corresponding to the largest eigenvalue λ_1.

This implies that

$$(d'Cd) \leq \lambda_1 d'd.$$

So we have that

$$\left(u'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}u\right) \leq (\rho_1)^2 u'u. \tag{3.6}$$

In equation 3.6 equality holds at $u = e_1$ and in equation 3.5 equality is attained if $v = \Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e_1$.

That is,

$$u = \Sigma_{11}^{-\frac{1}{2}}a, \text{ so } a = \Sigma_{11}^{-\frac{1}{2}}e_1 \text{ and } b = \Sigma_{12}^{-\frac{1}{2}}\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e_1.$$

$$\begin{aligned}\rho(a'X,b'Y) &\leq \frac{\left[(u'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}u)(v'v)\right]^{\frac{1}{2}}}{(u'u\cdot v'v)^{\frac{1}{2}}} \\ &= \left(\frac{u'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}u}{u'u}\right)^{\frac{1}{2}} \\ &\leq \left(\frac{(\rho_1)^2u'u}{u'u}\right)^{\frac{1}{2}} = \rho_1.\end{aligned}$$

This implies that

$$\max_{a,b} \quad \rho(a'X,b'Y) = \rho_1$$

and

$$\begin{aligned}\rho(e_1'\Sigma_{11}^{-\frac{1}{2}}X, f_1'\Sigma_{22}^{-\frac{1}{2}}Y) &= \frac{cov(e_1'\Sigma_{11}^{-\frac{1}{2}}X, f_1'\Sigma_{22}^{-\frac{1}{2}}Y)}{\left(var(e_1'\Sigma_{11}^{-\frac{1}{2}}X)var(f_1'\Sigma_{22}^{-\frac{1}{2}}Y)\right)^{\frac{1}{2}}} \\ &= \rho_1.\end{aligned}$$

This implies that, the first pair of canonical variables is given by $M_1 = e_1'\Sigma_{11}^{-\frac{1}{2}}X$ and $N_1 = f_1'\Sigma_{22}^{-\frac{1}{2}}Y$.

So we now have that

$$\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e_1 = \lambda_1 e_1 (\lambda_1 = \rho_1). \tag{3.7}$$

We multiply both sides of equation 3.7 by the matrix $\left(\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}\right)$ to obtain

$$\left(\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}\right)\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e_1 = \lambda_1\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e_1.$$

That is,

$$\sum_{22}^{-\frac{1}{2}}\sum_{21}\sum_{11}^{-1}\sum_{12}\sum_{22}^{-\frac{1}{2}}\left(\sum_{22}^{-\frac{1}{2}}\sum_{21}\sum_{11}^{-\frac{1}{2}}e_1\right) = \lambda_1\left(\sum_{22}^{-\frac{1}{2}}\sum_{21}\sum_{11}^{-\frac{1}{2}}e_1\right).$$

Since f_1 is proportional to $\sum_{22}^{-\frac{1}{2}}\sum_{21}\sum_{11}^{-\frac{1}{2}}e_1$, we have that

$$\sum_{22}^{-\frac{1}{2}}\sum_{21}\sum_{11}^{-1}\sum_{12}\sum_{22}^{-\frac{1}{2}}f_1 = \lambda_1 f_1.$$

Thus we conclude that if (λ_1, e_1) is the eigenvalue-eigenvector pair of $\sum_{11}^{-\frac{1}{2}}\sum_{12}\sum_{22}^{-1}\sum_{21}\sum_{11}^{-\frac{1}{2}}$, then (λ_1, f_1) is the eigenvalue-eigenvector pair of $\sum_{22}^{-\frac{1}{2}}\sum_{21}\sum_{11}^{-1}\sum_{12}\sum_{22}^{-\frac{1}{2}}$.

□

Derivation of the second canonical variables

M_1 and any linear combinations of Xs' say given by

$$a_2'X, u_2'\sum_{11}^{-\frac{1}{2}}X,$$

where $\sum_{11}^{\frac{1}{2}}a_2 = u_2$ are uncorrelated if

$$\begin{aligned} cov(M_1, u_2'\sum_{11}^{-\frac{1}{2}}X) = cov(e_1'\sum_{11}^{-\frac{1}{2}}, u_2'\sum_{11}^{-\frac{1}{2}}X) &= 0 \\ = e_1'\sum_{11}^{-\frac{1}{2}}\sum_{11}\sum_{11}^{-\frac{1}{2}}u_2 &= 0 \\ = e_1'u_2 &= 0. \end{aligned}$$

So, u_2 is to be determined such that it is orthogonal to e_1.

We want to find

$$\begin{aligned} \rho(a_2'X, b_2'Y) &= \frac{cov(a_2'X, b_2'Y)}{\left(var(a_2'X)\cdot var(b_2'Y)\right)} \\ &= \frac{a_2'\sum_{12}b_2}{\left((a_2'\sum_{11}a_2)(b_2'\sum_{22}b_2)\right)^{\frac{1}{2}}}. \end{aligned}$$

$$\text{We let } \sum_{11}^{\frac{1}{2}}a_2 = u_2 \implies a_2 = \sum_{11}^{-\frac{1}{2}}u_2$$
$$\text{and let } \sum_{22}^{\frac{1}{2}}b_2 = v_2 \implies b_2 = \sum_{22}^{-\frac{1}{2}}v_2.$$

So we have that

$$\rho(a_2'X, b_2'Y) = \frac{u_2'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}v_2}{(u_2'u_2 \cdot v_2'v_2)^{\frac{1}{2}}}.$$

We apply the Cauchy Schwartz inequality to the numerator and have that

$$\left(u_2'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{\frac{1}{2}}v_2\right) \leq \left(u_2'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}u_2\right)^{\frac{1}{2}}(v_2v_2')^{\frac{1}{2}}. \tag{3.8}$$

So we concentrate on the expression $u_2'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}u_2$ and try to see what can be given as an upper bound of this particular expression.

In order to get that, we again recall a result from matrix theory that states that for a real symmetric matrix $C_{p\times p}$ with eigenvalue-eigenvector pairs (λ_i, e_i); $i = 1, 2, \ldots p$ such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, we have that

$$\max_{d\perp e_1} \frac{d'Cd}{d'd} = \lambda_2 \implies d'Cd \leq \lambda_2 d'd \tag{3.9}$$

and

$$\max_{d\perp e_1, e_2, \ldots e_k} \frac{d'Cd}{d'd} = \lambda_{k+1} \implies d'Cd \leq \lambda_{k+1} d'd. \tag{3.10}$$

In equation 3.9, equality holds if $d = e_2$ and for equation 3.10, equality holds if $d = e_{k+1}$.

From 3.9, we have that

$$\left(u_2'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}u_2\right) \leq \lambda_2(u_2'u_2) \quad \text{with equality at} \quad u_2 = e_2.$$

In equation 3.8 equality is attained if

$$\begin{aligned} v_2 = \Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e_2 \implies b_2 &= \Sigma_{22}^{-\frac{1}{2}}\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e_2 \\ b_2 &= \Sigma_{22}^{-\frac{1}{2}}f_2. \end{aligned}$$

So now we have that

$$\begin{aligned}\rho(a_2'X, b_2'Y) &\leq \frac{\left[\left(u_2{}'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}u_2\right)(v_2{}'v_2)\right]^{\frac{1}{2}}}{(u_2{}'u_2 \cdot v_2{}'v_2)^{\frac{1}{2}}} \\ &= \left(\frac{u_2'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}u_2}{u_2{}'u_2}\right)^{\frac{1}{2}} \\ &\leq \left(\frac{(\rho_2)^2 u_2'u_2}{u_2'u_2}\right)^{\frac{1}{2}} = \rho_2.\end{aligned}$$

Thus

$$corr(a_2'X, b_2'Y) \leq \rho_2 \quad \text{with equality at} \quad u_2 = e_2$$
$$\implies a_2 = \Sigma_{11}^{-\frac{1}{2}} e_2.$$

The Second Canonical Variable pairs are $M_2 = e_2'\Sigma_{11}^{-\frac{1}{2}}X$ and $N_2 = f_2'\Sigma_{22}^{-\frac{1}{2}}Y$.

The second canonical correlation coefficient is ρ_2 as required.

3.1.5 Properties of the Canonical Variable Pairs

(i) $var(M_k) = var(N_k) = 1$.

Proof.

$$var(M_k) = var(e_k'\Sigma_{11}^{-\frac{1}{2}}X) = e_k'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{11}\Sigma_{11}^{-\frac{1}{2}}e_k = e_k'e_k = 1.$$

Similarly,

$$var(N_k) = f_k'\Sigma_{22}^{-\frac{1}{2}}\Sigma_{22}\Sigma_{22}^{-\frac{1}{2}}f_k = f_k'f_k = 1.$$

□

(ii) $cov(M_k, M_t) = corr(M_k, M_t) = 0, \quad \forall k \neq t.$

Proof.

$$\begin{aligned} cov(M_k, M_t) &= cov(e_k' \sum_{11}^{-\frac{1}{2}} X, e_t' \sum_{11}^{-\frac{1}{2}} X) \\ &= e_k' \sum_{11}^{-\frac{1}{2}} \sum_{11} \sum_{11}^{-\frac{1}{2}} e_t \\ &= e_k' e_t = 0 \quad \forall \quad k \neq t \quad \text{since} \quad e_k \quad \text{and} \quad e_t \quad \text{are orthogonal.} \end{aligned}$$

□

(iii) $cov(N_k, N_t) = corr(N_k, N_t) = 0, \quad \forall \ k \neq t.$

Proof.

$$\begin{aligned} cov(N_k, N_t) &= cov(f_k' \sum_{22}^{-\frac{1}{2}} Y, f_t' \sum_{22}^{-\frac{1}{2}} Y) \\ &= f_k' \sum_{22}^{-\frac{1}{2}} \sum_{22} \sum_{22}^{-\frac{1}{2}} f_t. \end{aligned}$$

Also, because of the orthogonality of f_k and f_t,

$$cov(N_k, N_l) = f_k' f_t = 0 \quad \forall k \neq l.$$

□

(iv) $cov(M_k, N_t) = corr(M_k, N_t) = 0, \quad \forall \ k \neq t.$

Proof.

$$cov(M_k, N_t) = cov(e_k' \sum_{11}^{-\frac{1}{2}} X, f_t' \sum_{22}^{-\frac{1}{2}} Y) = e_k' \sum_{11}^{-\frac{1}{2}} \sum_{12} \sum_{22}^{-\frac{1}{2}} f_t. \tag{3.11}$$

We recall that f_k is proportional to $\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e_k$ and so

$$cov(M_k,N_t) = Qf_k'f_t = 0, \quad \forall\ k \neq t \quad \text{since} \quad f_k \perp f_t \quad \text{where} \quad Q \text{ is a constant.}$$

□

3.1.6 Canonical correlation coefficient under the non-singular transformation

In this section we seek to find the canonical correlations if the vectors, X and Y are being transformed. We will also demonstrate that we can compute the canonical correlation coefficients either from the covariance matrix or from the correlation matrix. We derive the canonical correlation coefficient under the transformation.

$$X_{p\times 1} \to CX \quad \text{and}$$
$$Y_{q\times 1} \to DY,$$

where C and D are non-singular matrices. We have

$$cov\begin{pmatrix} CX \\ DY \end{pmatrix} = \begin{pmatrix} C\Sigma_{11}C' & C\Sigma_{12}D' \\ D\Sigma_{21}C' & D\Sigma_{22}D' \end{pmatrix}.$$

We have seen that $\rho_1,\rho_2,\ldots,\rho_p$ are the canonical correlation coefficients for the $\begin{pmatrix} X \\ Y \end{pmatrix}$ set up. Also, $(\rho_1)^2,(\rho_2)^2,\ldots,(\rho_p)^2$ are the eigenvalues of $\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}$. Hence, $(\rho_1)^2,(\rho_2)^2,\ldots,(\rho_p)^2$ are the roots of

$$\left|\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}} - \lambda I\right| = 0.$$

So we now pre and post multiply by the matrix $\sum_{11}^{\frac{1}{2}}$ and $\sum_{11}^{-\frac{1}{2}}$ to get

$$\left|\sum_{11}^{\frac{1}{2}}\sum_{11}^{-\frac{1}{2}}\sum_{12}\sum_{22}^{-1}\sum_{21}\sum_{11}^{-\frac{1}{2}}\sum_{11}^{-\frac{1}{2}}-\lambda I\right|=0$$
$$\left|\sum_{12}\sum_{22}^{-1}\sum_{21}\sum_{11}^{-1}-\lambda I\right|=0.$$

The matrix $\sum_{12}\sum_{22}^{-1}\sum_{21}\sum_{11}^{-1}$ can be transformed under C and D as

$$\begin{aligned}\sum_{12}\sum_{22}^{-1}\sum_{21}\sum_{11}^{-1} &\xrightarrow{C,D} \left((C\sum_{12}D')(D\sum_{22}D')^{-1}(D\sum_{21}C')(C\sum_{11}C')^{-1}\right)\\ &= C\sum_{12}\sum_{22}^{-1}\sum_{21}\sum_{11}^{-1}C^{-1}.\end{aligned}$$

We have that, the non-zero eigenvalues of $C\sum_{12}\sum_{22}^{-1}\sum_{21}\sum_{11}^{-1}C^{-1}$ are the same as the non-zero eigenvalues of $C^{-1}C\sum_{12}\sum_{22}^{-1}\sum_{21}\sum_{11}^{-1}=\sum_{12}\sum_{22}^{-1}\sum_{21}\sum_{11}^{-1}$.

Hence we conclude that the canonical correlation coefficient under the non-singular transformation C,D are the same.

We now take a special case of such a transformation by defining C and D as follows;

$$C=N_{11}^{-\frac{1}{2}} \quad \text{where } N_{11}=\text{ diag }\left(\sum_{11}\right) \quad \text{and} \quad D=N_{22}^{-\frac{1}{2}} \quad \text{where } N_{22}=\text{ diag }\left(\sum_{22}\right).$$

So we transform the vectors X and Y under the given transformation and compute the covariance of X and Y under the transformation.

$$X\to CX=N_{11}^{-\frac{1}{2}}X\to cov(N_{11}^{-\frac{1}{2}}X)=N_{11}^{-\frac{1}{2}}\sum_{11}N_{11}^{-\frac{1}{2}}=\rho_{11},$$
$$Y\to DY=N_{22}^{-\frac{1}{2}}Y\to cov(N_{22}^{-\frac{1}{2}}Y)=N_{22}^{-\frac{1}{2}}\sum_{22}N_{22}^{-\frac{1}{2}}=\rho_{22}.$$

This implies that the eigenvalues of $\sum_{11}^{\frac{1}{2}}\sum_{12}\sum_{22}^{-1}\sum_{21}\sum_{11}^{-\frac{1}{2}}$ are identical to the eigenvalues of $\rho_{11}^{-\frac{1}{2}}\rho_{12}\rho_{22}^{-1}\rho_{21}\rho_{11}^{-\frac{1}{2}}$.

Therefore, computing the canonical correlation coefficients from either the covariance matrix or the correlation matrix will yield the same values.

3.1.7 Correlation Coefficient Between Canonical Variables and the Original Variables

We now derive the correlation coefficient between the canonical variables, (M_i and N_i) where $i = 1, 2, \ldots, p$ and the original variables X and Y.

The pth canonical variate pairs are defined as follows

$$M_p = e_p' \sum_{11}^{-\frac{1}{2}} X \quad \text{and} \quad N_p = f_p' \sum_{22}^{-\frac{1}{2}} Y.$$

$$\underbrace{M}_{p\times 1} = \begin{pmatrix} M_1 \\ M_2 \\ \vdots \\ M_p \end{pmatrix} = \begin{pmatrix} e_1' \\ e_2' \\ \vdots \\ e_p' \end{pmatrix} \sum_{11}^{-\frac{1}{2}} X = CX \quad \text{and} \quad C = \begin{pmatrix} e_1' \\ e_2' \\ \vdots \\ e_p' \end{pmatrix} \sum_{11}^{-\frac{1}{2}}.$$

$$\underbrace{N}_{q\times 1} = \begin{pmatrix} N_1 \\ N_2 \\ \vdots \\ N_p \end{pmatrix} = \begin{pmatrix} f_1' \\ f_2' \\ \vdots \\ f_q' \end{pmatrix} \sum_{22}^{-\frac{1}{2}} Y = DY \quad \text{and} \quad D = \begin{pmatrix} f_1' \\ f_2' \\ \vdots \\ f_q' \end{pmatrix} \sum_{22}^{-\frac{1}{2}}.$$

$$cov(M,X) = cov(CX,X) = C\sum_{11} = \begin{pmatrix} e_1' \\ e_2' \\ \vdots \\ e_p' \end{pmatrix} \sum_{11}^{\frac{1}{2}} \quad \text{and}$$

$$cov(N,Y) = cov(DY,Y) = B\sum_{22} = \begin{pmatrix} f_1' \\ f_2' \\ \vdots \\ f_q' \end{pmatrix} \sum_{22}^{\frac{1}{2}}.$$

This implies that

$$\begin{aligned}
corr(M_i, X_k) &= \frac{cov(M_i, X_k)}{\sigma_{kk}^{\frac{1}{2}}} \quad ; \quad (var(X_k) = \sigma_{kk}) \\
&= cov(M_i, \sigma_{kk}^{-\frac{1}{2}} X_k) \\
corr(M, X) &= cov(M, N_{11}^{-\frac{1}{2}} X) \quad \text{where} \quad N_{11} = \text{diag}\left(\sum_{11}\right) = \text{diag}(\sigma_{11}, \ldots, \sigma_{pp}) \\
&= cov(CX, N_{11}^{-\frac{1}{2}} X) \\
&= C \sum_{11} N_{11}^{-\frac{1}{2}} \\
&= C N_{11}^{\frac{1}{2}} N_{11}^{-\frac{1}{2}} \sum_{11} N_{11}^{-\frac{1}{2}} \\
&= C N_{11}^{\frac{1}{2}} \rho_{11}. \qquad (3.12)
\end{aligned}$$

Similarly,

$$\begin{aligned}
corr(M, Y) &= cov(\underset{\sim}{M}, N_{22}^{-\frac{1}{2}} Y) \\
&= cov(CX, N_{22}^{-\frac{1}{2}} Y) \quad \text{where} \quad N_{22} = \text{diag}\left(\sum_{22}\right) = \text{diag}\left(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{qq}\right) \\
&= C \sum_{12} N_{22}^{-\frac{1}{2}} = C N_{22}^{\frac{1}{2}} \rho_{12}. \qquad (3.13)
\end{aligned}$$

And

$$\begin{aligned}
corr(N, X) &= cov(N, N_{11}^{-\frac{1}{2}} X) \\
&= Cov(DY, N_{11}^{-\frac{1}{2}} X) \\
&= D \sum_{21} N_{11}^{-\frac{1}{2}} = D N_{11}^{\frac{1}{2}} \rho_{21}. \qquad (3.14)
\end{aligned}$$

Finally,

$$\begin{aligned}
corr(N, Y) &= cov(N, N_{12}^{-\frac{1}{2}} Y) \\
&= cov(BY, N_{22}^{-\frac{1}{2}} Y) \\
&= D \sum_{22} N_{22}^{-\frac{1}{2}} = D N_{22}^{\frac{1}{2}} \rho_{22}. \qquad (3.15)
\end{aligned}$$

Equations 3.12, 3.13, 3.14 and 3.15 are the derived canonical coefficients between the canonical variate pairs and the original variables.

3.1.8 Computation of Canonical Correlation Coefficient Using Standardized Variables

Here, we seek to derive the canonical coefficient by standardizing the original variables. We denote the standardized variables are follows

$$\begin{aligned} Z^{(X)} &= (X - M_X)N_{11}^{-\frac{1}{2}} \quad \text{and} \\ Z^{(Y)} &= (Y - M_Y)N_{22}^{-\frac{1}{2}}. \end{aligned}$$

So the covariance matrix of the standardized variables is given by

$$cov\begin{pmatrix} Z^{(X)} \\ Z^{(Y)} \end{pmatrix} = \begin{pmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{pmatrix}.$$

From the correlation matrix, the derived canonical variables are

$$\begin{aligned} M_{Z_k} &= e_k' \sum_{11}^{-\frac{1}{2}} N_{11}^{\frac{1}{2}} Z^{(X)} \quad \text{and} \\ N_{Z_k} &= f_k' \sum_{22}^{-\frac{1}{2}} N_{22}^{\frac{1}{2}} Z^{(Y)}. \end{aligned}$$

$$M_Z = \begin{pmatrix} M_{Z_1} \\ \vdots \\ M_{Z_p} \end{pmatrix} = \begin{pmatrix} e_1' \\ \vdots \\ e_p' \end{pmatrix} \sum_{11}^{-\frac{1}{2}} N_{11}^{\frac{1}{2}} Z^{(X)} = C_Z Z^{(X)} \tag{3.16}$$

and

$$N_Z = \begin{pmatrix} N_{Z_1} \\ \vdots \\ N_{Z_q} \end{pmatrix} = \begin{pmatrix} f_1' \\ \vdots \\ f_q' \end{pmatrix} \sum_{22}^{-\frac{1}{2}} N_{22}^{\frac{1}{2}} Z^{(Y)} = D_Z Z^{(Y)}. \tag{3.17}$$

Now we compute the correlation between the canonical variables obtained from the correlation matrix and the standardized variables. We have

$$\begin{aligned} \rho(M_Z, Z^{(X)}) = cov(M_Z, Z^{(X)}) &= cov(C_Z Z^{(X)}, Z^{(X)}) \\ &= C_Z \rho_{11}. \end{aligned} \tag{3.18}$$

$$\begin{aligned}
\rho(N_Z, Z^{(Y)}) = cov(N_Z, Z^{(Y)}) &= cov(D_Z Z^{(Y)}, Z^{(Y)}) \\
&= D_Z \rho_{22}. \qquad (3.19)\\
\rho(M_Z, Z^{(Y)}) = cov(C_Z Z^{(X)}, Z^{(Y)}) &= C_Z \rho_{12}. \qquad (3.20)\\
\rho(N_Z, Z^{(X)}) = cov(D_Z Z^{(Y)}, Z^{(X)}) &= D_Z \rho_{21}. \qquad (3.21)
\end{aligned}$$

From equations 3.16 and 3.17, we have that

$$\begin{aligned}
C_Z &= \begin{pmatrix} e_1' \\ \vdots \\ e_p' \end{pmatrix} \sum_{11}^{-\frac{1}{2}} N_{11}^{\frac{1}{2}} \quad \text{and} \\
D_Z &= \begin{pmatrix} f_1' \\ \vdots \\ f_p' \end{pmatrix} \sum_{22}^{-\frac{1}{2}} N_{22}^{\frac{1}{2}}.
\end{aligned}$$

This gives

$$\begin{aligned}
\rho(M,X) &= CN_{11}^{\frac{1}{2}}\rho_{11} = \begin{pmatrix} e_1' \\ \vdots \\ e_p' \end{pmatrix} \sum_{11}^{-\frac{1}{2}} N_{11}^{\frac{1}{2}}\rho_{11} = C_Z\rho_{11} = \rho(M_Z, Z^{(1)}),\\
\rho(M,Y) &= CN_{22}^{\frac{1}{2}}\rho_{12} = \begin{pmatrix} e_1' \\ \vdots \\ e_p' \end{pmatrix} \sum_{12}^{-\frac{1}{2}} N_{22}^{\frac{1}{2}}\rho_{12} = C_Z\rho_{11} = \rho(M_Z, Z^{(X)}),\\
\rho(\underset{\sim}{N},X) &= DN_{11}^{\frac{1}{2}}\rho_{21} = \begin{pmatrix} f_1' \\ \vdots \\ f_p' \end{pmatrix} \sum_{21}^{-\frac{1}{2}} N_{11}^{\frac{1}{2}}\rho_{21} = D_Z\rho_{21} = \rho(N_Z, Z^{(X)}),\\
\rho(N,Y) &= DN_{11}^{\frac{1}{2}}\rho_{22} = \begin{pmatrix} f_1' \\ \vdots \\ f_p' \end{pmatrix} \sum_{22}^{-\frac{1}{2}} N_{22}^{\frac{1}{2}}\rho_{22} = D_Z\rho_{22} = \rho(N_Z, Z^{(Y)}).
\end{aligned}$$

We then conclude that, computing correlations by standardizing the variables has no effect.

3.1.9 Assessing Overall Model Fit and Canonical Dimension Reduction

Under this section, two techniques will be discussed to explore the possibility that interpreting fewer canonical dimensions or canonical variate pairs can be enough to capture sufficient covariance or correlation structure. It is known that not all canonical functions are important. Evidently, the strength of the canonical correlation coefficient can suggest the importance of the canonical variate pairs [2]. We are ultimately interested in the significant canonical coefficients to make informed decisions. The first technique involves the use of Wilk's lambda and it's corresponding F-tests to test the null hypothesis that all canonical functions have canonical correlation coefficients to be zero at a 5% significance level. Wilk's lambda evaluates each canonical function against the null hypothesis that the canonical coefficient is zero. The second technique seeks to ascertain if choosing $k < p$ canonical variate pairs is enough to capture the covariance structure.

Technique I

For each canonical correlation coefficient, there exists an eigenvalue that is related to the Wilk's lambda. The eigenvalue for each coefficient in relation to the Wilk's lamda is calculated as

$$\lambda_i = \frac{\rho_i}{(1-\rho_i)^2}$$

and Wilk's lamda is computed as

$$\Lambda = \frac{1}{\prod(1-\lambda_i)}.$$

The F-test value is calculated as

$$F = \frac{1-\Lambda^{\frac{1}{w}}}{\Lambda^{\frac{1}{w}}} \left(\frac{\text{degrees of freedom1}}{\text{degrees of freedom2}} \right).$$

$$\begin{aligned}
\text{Degrees of Freedom1} &= p \times q. \\
\text{Degrees of Freedom2} &= vw - \frac{pq}{2} + 1. \\
v &= n - \frac{3}{2} - \frac{p+q}{2}, \quad n \text{ is the sample size.} \\
w &= \left(\frac{p^2q^2 - p}{p^2 + q^2 - q} \right)^{\frac{1}{2}}.
\end{aligned} \tag{3.22}$$

The computation of w in equation 3.22 is iterative. We begin with the initial values of p and q and repeatedly subtract one from p and q until either p or q has been reduced to one.

We now compute the p-value or the critical value to make the final decision. The critical value is a value that the computed F value must exceed to reject the test hypothesis. The critical value is computed from the F-distribution table using the two degrees of freedom and the level of significance (5%).

The p-value is computed using the F value and the two degrees of freedom values. If the p-value is less than 0.05, then we reject the null hypothesis, otherwise we fail to reject the null hypothesis.

Technique II

We have that

$$M = \begin{pmatrix} M_1 \\ \vdots \\ M_p \end{pmatrix} = SX \quad \text{and so} \quad X = S^{-1}M, \quad \text{where} \quad S = \begin{pmatrix} e_1' \\ \vdots \\ e_p' \end{pmatrix} \sum_{11}^{-\frac{1}{2}}$$

and

$$N = \begin{pmatrix} N_1 \\ \vdots \\ N_q \end{pmatrix} = TY \quad \text{thus} \quad Y = T^{-1}N, \quad \text{and} \quad T = \begin{pmatrix} f_1' \\ \vdots \\ f_q' \end{pmatrix} \sum_{22}^{-\frac{1}{2}}.$$

Clearly,

$$S^{-1} = \sum_{11}^{\frac{1}{2}}(e_1, \ldots, e_p) \text{ and } \quad T^{-1} = \sum_{22}^{\frac{1}{2}}(f_1, \ldots, f_q).$$

So writing S^{-1} and T^{-1} in the form below eases the computation.
We write

$$\begin{aligned} S^{-1} &= \left(s^{(1)},\ldots,s^{(p)}\right), \quad \text{where} \\ s^{(i)} &= \textstyle\sum_{11}^{\frac{1}{2}} e_i \;; \quad i=1,2,\ldots,p \qquad \text{and} \qquad (3.23) \\ T^{-1} &= \left(t^{(1)},\ldots,t^{(q)}\right), \quad \text{where} \\ t^{(i)} &= \textstyle\sum_{22}^{\frac{1}{2}} f_i; \quad i=1,2,\ldots,q. \qquad (3.24) \end{aligned}$$

Using this we rewrite X and Y as

$$\begin{aligned} X &= \left(s^{(1)},\ldots,s^{(p)}\right) M \\ &= \sum_{i=1}^{p} s^{(i)} M \quad \text{and} \qquad (3.25) \\ Y &= \left(t^{(1)},\ldots,t^{(q)}\right) N \\ &= \sum_{i=1}^{q} t^{(i)} N. \qquad (3.26) \end{aligned}$$

We can then compute the covariance of X and Y as

$$\begin{aligned} cov(X) &= cov\left(\sum_{i=1}^{p} s^{(i)} M_i\right) = \sum_{i=1}^{p} s^{(i)} s^{(i)'} \quad \text{and} \\ cov(Y) &= cov\left(\sum_{i=1}^{p} t^{(i)} N_i\right) = \sum_{i=1}^{q} t^{(i)} t^{(i)'}. \end{aligned}$$

So considering the first k canonical variables, we have that

$$\begin{aligned} X^* &= \sum_{i=1}^{k} a^{(i)} M_i \quad \text{and} \quad Y^* = \sum_{i=1}^{k} b^{(i)} N_i, \quad \text{thus} \\ cov(X^*) &= \sum_{i=1}^{k} s^{(i)} s^{(i)'} \quad \text{and} \quad cov(Y^*) = \sum_{i=1}^{k} t^{(i)} t^{(i)'}. \end{aligned}$$

We then compute the covariance between X and Y as

$$cov(X,Y) = cov(S^{-1}M, T^{-1}N) = S^{-1}\begin{pmatrix} \rho_1 & 0 & 0 \\ & \ddots & 0 \\ 0 & & \rho_p \end{pmatrix}\left(T^{-1}\right)'$$

$$\text{and so} \quad cov(X,Y) = (s^{(1)},\ldots,s^{(1)})\begin{pmatrix} \rho_1 & 0 & & & 0 \\ 0 & 0 & \rho_2 & 0 & 0 \\ & & & \ddots & 0 \\ 0 & & 0 & & \rho_p \end{pmatrix}\begin{pmatrix} t^{(1)'} \\ \vdots \\ t_q^{(1)'} \end{pmatrix}$$

$$= \sum_{i=1}^{p} \rho_i^* s^{(i)} t^{(i)'}.$$

Therefore,

$$cov(X^*,Y^*) = \sum_{i=1}^{k} \rho_i s^{(i)} t^{(i)'}.$$

So having the covariance structure for the first k canonical variables, we now seek to find out the closeness to a null matrix of the three matrices.

$$\sum_{i=k+1}^{p} s^{(i)} s^{(i)'}, \quad \sum_{i=k+1}^{q} t^{(i)} t^{(i)'} \quad \text{and} \quad \sum_{i=k+1}^{p} \rho_i s^{(i)} t^{(i)'}.$$

We make three observations.

(1) Since we usually choose k such that ρ_{k+1} and hence $\rho_{k+2},\ldots,\rho_p$ are negligible, $\sum_{i=k+1}^{p} \rho_i s^{(i)} t^{(i)'}$ will be closer to a null matrix than $\sum_{i=k+1}^{p} s^{(i)} s^{(i)'}$ and $\sum_{i=k+1}^{q} t^{(i)} t^{(i)'}$.

(2)

$$cov\,(X,M) = cov\left(S^{-1}M, M\right) = S^{-1} = \left(s^{(1)},\ldots,s^{(p)}\right)$$

$$= \begin{pmatrix} cov(X_1,M_1) & \ldots & cov(X_1,M_p) \\ \vdots & & \vdots \\ cov(X_p,M_1) & \ldots & cov(X_p,M_p) \end{pmatrix}.$$

(3) Considering $k < p$ canonical variables, $M_1,\ldots,M_k$, the proportion of total variance X

explained by $M_1, \ldots, M_k$ is given as

$$\frac{tr(cov(X^*))}{tr(cov(X)} = \frac{tr\left(\sum_{i=1}^{k} s^{(i)} a^{(i)'}\right)}{tr\sum_{11}}.$$

where tr is the trace of the matrices in question.

In addition

$$\begin{aligned} S^{-1} &= (s^{(1)}, \ldots, s^{(p)}) = cov(X, M) \\ \text{and} \quad s^{(i)} &= \begin{pmatrix} cov(X_1, M_i \\ \vdots \\ cov(X_p, M_i) \end{pmatrix} \quad i = 1, \ldots, p \\ \text{thus} \quad s^{(i)'} s^{(i)} &= \sum_{j=1}^{p} cov(X_j, M_i)^2 \quad \text{and} \\ \sum_{i=1}^{k} s^{(i)'} s^{(i)} &= \sum_{i=1}^{k} \sum_{j=1}^{p} cov(X_j, M_i)^2. \end{aligned}$$

Thus

$$\frac{tr\left(\sum_{i=1}^{k} s^{(i)} s^{(i)'}\right)}{tr\sum_{11}} = \frac{\sum_{i=1}^{k} tr(s^{(i)} s^{(i)'})}{\sum_{i=1}^{p} tr(s^{(i)} s^{(i)'})} = \frac{\sum_{i=1}^{k} tr(s^{(i)'} s^{(i)})}{\sum_{i=1}^{p} tr(s^{(i)} s^{(i)'})}$$

Since $(s^{(i)'} s^{(i)})$ is a scalar quantity, we have that

$$\begin{aligned} \frac{\sum_{i=1}^{k} tr(s^{(i)'} s^{(i)})}{\sum_{i=1}^{p} tr(s^{(i)} s^{(i)'})} &= \frac{\sum_{i=1}^{k} s^{(i)} s^{(i)'}}{\sum_{i=1}^{p} s^{(i)} s^{(i)'}} \\ &= \frac{\sum_{i=1}^{k} \sum_{j=1}^{p} cov(X_j, M_i)^2}{\sum_{i=1}^{p} \sum_{j=1}^{p} cov(X_j, M_i)^2}. \end{aligned}$$

Similarly, the proportion of total variance of Y explained by $N_1,\ldots,N_k$, is given by

$$\frac{tr(\sum_{i=1}^{k} t^{(i)}t^{(i)'})}{tr\Sigma_{22}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{q} cov(Y_j,N_i)^2}{\sum_{i=1}^{q}\sum_{j=1}^{q} cov(Y_j,N_i)^2}.$$

If the proportion of total variance is close to 1 or 100%, then the k dimensions are retained.

3.2 Example: Computation of Canonical variables and Canonical Coefficients

Here we use the derived formulas obtained in this chapter to compute the canonical variable pairs and the canonical coefficients of the covariance structure below. We consider a Z standardized vector with variables standardized. It is divided into two.

$$Z_{q\times 1} = \begin{pmatrix} Z^{(1)} \\ \\ Z^2 \end{pmatrix}.$$

The $Z^{(X)}$ and $Z^{(Y)}$ are standardized variables (2×1).
Suppose we are given

$$cov(Z) = cov\begin{pmatrix} Z^{(1)} \\ \\ Z^{(2)} \end{pmatrix} = \begin{pmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} 1.00 & 0.40 \\ 0.40 & 1.00 \end{pmatrix} & \begin{pmatrix} 0.50 & 0.60 \\ 0.30 & 0.40 \end{pmatrix} \\ \begin{pmatrix} 0.60 & 0.40 \\ 0.50 & 0.30 \end{pmatrix} & \begin{pmatrix} 1.00 & 0.20 \\ 0.20 & 1.00 \end{pmatrix} \end{pmatrix}.$$

We begin by calculating $\rho_{11}^{-\frac{1}{2}}$ and ρ_{22}^{-1} as

$$\rho_{11}^{-\frac{1}{2}} = \begin{pmatrix} 1.068 & -0.223 \\ -0.223 & 1.068 \end{pmatrix}$$

$$\text{and} \quad \rho_{22}^{-1} = \begin{pmatrix} 1.042 & -0.208 \\ -0.208 & 1.042 \end{pmatrix}.$$

so

$$\rho_{11}^{-\frac{1}{2}}\rho_{12}\rho_{22}^{-1}\rho_{21}\rho_{11}^{-\frac{1}{2}} = \begin{pmatrix} 0.437 & 0.218 \\ 0.218 & 0.120 \end{pmatrix}.$$

Now we seek to ascertain the eigenvalues of the matrix $\rho_{11}^{-\frac{1}{2}}\rho_{12}\rho_{22}^{-1}\rho_{21}\rho_{11}^{-\frac{1}{2}}$. The eigenvalues ρ_1^2, ρ_2^2 are as follows

$$\rho_1^2 = 0.548 \quad \text{and} \quad \rho_2^2 = 0.0090,$$

hence,

$$\rho_1 = 0.740 \quad \text{and} \quad \rho_2 = 0.030.$$

The eigenvector, e_1 associated to ρ_1^2 is obtained as

$$e_1 = \begin{pmatrix} 0.8911 \\ 0.4538 \end{pmatrix}.$$

This implies that the coefficient vector for $M_1 : \rho_{11}^{-\frac{1}{2}} e_1 = a_1 = \begin{pmatrix} 0.856 \\ 0.278 \end{pmatrix}$. So

$$M_1 = e_1'\rho_{11}^{-\frac{1}{2}} Z^{(X)} = 0.856 Z_1^{(X)} + 0.278 Z_2^{(X)}. \tag{3.27}$$

We find the coefficient vector, b, for N_1.

We have that f_1 is proportional to $\rho_{22}^{-\frac{1}{2}}\rho_{21}\rho_{11}^{-\frac{1}{2}} e_1$ and $b_1 = \rho_{22}^{-\frac{1}{2}} f_1$. Thus f_1 is proportional to $\rho_{22}^{-\frac{1}{2}}\rho_{21} a_1$. The constant of proportionality $= 1$ since b_1 is such that $var(b_1' Z^{(Y)}) = var(N_1) = b_1'\rho_{22} b_1 = 1$.

$$
\begin{aligned}
b_1\rho_{22}^{\frac{1}{2}} &\propto \rho_{22}^{-\frac{1}{2}}\rho_{21}a_1 \\
b_1 &\propto \rho_{22}^{-\frac{1}{2}}\rho_{22}^{-\frac{1}{2}}\rho_{21}a_1 \\
b_1 &\propto \rho_{22}^{-1}\rho_{21}a_1 \\
\rho_{22}^{-1}\rho_{21}a_1 &= \begin{pmatrix} 0.403 \\ 0.544 \end{pmatrix}.
\end{aligned}
$$

We orthonormalize $\rho_{22}^{-1}\rho_{21}a_1$

$$
\begin{aligned}
b_1'\rho_{22}b_1 &= 0.546 \\
b_1 &= \frac{1}{\sqrt{0.546}}\begin{pmatrix} 0.403 \\ 0.544 \end{pmatrix} \\
N_1 &= b_1 Z^{(Y)} = \frac{0.403}{\sqrt{0.546}}Z_1^{(Y)} + \frac{0.544}{\sqrt{0.546}}Z_2^{(Y)}.
\end{aligned}
$$

The second canonical correlation coefficient is too small and hence further calculations will not be done. We later show why only one canonical coefficient was enough.

We now compute the correlations between the original set of variables(standardized) and the canonical variates M_1 and N_1.

For the first canonical variable pair, we have that

$$
\begin{aligned}
C_Z' &= (0.86, 0.28) \quad \text{and} \\
D_Z' &= (0.54, 0.74).
\end{aligned}
$$

The correlation between M_1 and $Z^{(X)}$ is

$$
\begin{aligned}
\rho(M_1, Z^{(X)}) &= C_Z\rho_{11} = (0.97, 0.62) \\
\text{Similarly,} \quad \rho(N_1, Z^{(Y)}) &= D_Z\rho_{22} = (0.69, 0.85), \\
\rho(M_1, Z^{(Y)}) &= C_Z\rho_{12} = (0.51, 0.63) \quad \text{and} \\
\rho(N_1, Z^{(X)}) &= D_Z\rho_{21} = (0.71, 0.46).
\end{aligned}
$$

We now show that only one canonical variable was sufficient to capture the correlation structure.

For $k = 1$, the canonical functions are as follows

$$\begin{aligned} M_1 &= 0.86X_1 + 0.28X_2 \\ N_1 &= 0.54Y_1 + 0.74Y_2. \\ \text{So take} \quad a_1' &= (0.86, 0.28) \quad \text{and} \quad b_1' = (0.54, 0.74). \end{aligned}$$

Now

$$\begin{aligned} cov(X_1, M_1) &= 0.86cov(X_1, X_1) + 0.28cov(X_1, X_2) = 0.97, \\ cov(Y_1, N_1) &= 0.54cov(Y_1, Y_1) + 0.74cov(Y_1, Y_2) = 0.69, \\ cov(X_2, M_1) &= 0.86cov(X_1, X_2) + 0.28cov(X_2, X_2) = 0.62, \\ cov(Y_2, N_2) &= 0.54cov(Y_1, Y_2) + 0.74cov(Y_2, Y_2) = 0.85. \end{aligned}$$

From the covariances computed above, we have that

$$\begin{aligned} s^{(1)} &= \begin{pmatrix} 0.97 \\ 0.62 \end{pmatrix} \quad \text{and} \quad t^{(1)} = \begin{pmatrix} 0.69 \\ 0.85 \end{pmatrix} \\ s^{(1)}s^{(1)'} &= \begin{pmatrix} 0.95 & 0.61 \\ 0.61 & 0.4 \end{pmatrix} \quad \text{and} \quad t^{(1)}t^{(1)'} = \begin{pmatrix} 0.47 & 0.58 \\ 0.58 & 0.72 \end{pmatrix} \\ \rho_1 s^{(1)}t^{(1)'} &= \begin{pmatrix} 0.5 & 0.61 \\ 0.31 & 0.39 \end{pmatrix}. \end{aligned}$$

Thus if considering only 1 canonical variate pair (M_1, N_1), we check to see whether $s^{(1)}s^{(1)'}$, $t^{(1)}t^{(1)'}$, $\rho_1 s^{(1)}t^{(1)'}$ approximate ρ_{11}, ρ_{22} and ρ_{12} respectively.

From our computations, we have

$$\begin{pmatrix} 0.5 & 0.61 \\ 0.31 & 0.39 \end{pmatrix} \approx \begin{pmatrix} 0.5 & 0.6 \\ 0.3 & 0.4 \end{pmatrix}.$$

We observe that of the three matrices only $\rho_1 s^{(1)}t^{(1)'}$ has a reasonable approximation to ρ_{12}. This result conforms to the note presented above stating that, $\sum_{i=k+1}^{p} \rho_i s^{(i)}t^{(i)'}$ is very close to the null matrix.

We calculate the proportion of total variance explained by both M_1 and N_1.

$$\begin{aligned} \frac{tr(s^{(1)}s^{(1)'})}{tr\sum_{11}} &= \frac{0.95+0.4}{2} \simeq 68\% \\ \frac{tr(t^{(1)}t^{(1)'})}{tr\sum_{22}} &= \frac{0.47+0.72}{2} \simeq 60\% \end{aligned}$$

M_1 explains 68% of the total variation in X and N_1 explains 60% of variation in Y. This shows that the first canonical variate pairs is enough to capture sufficient covariance structure of the sets of variables.

Chapter 4

Results

This chapter presents the results and discussion of the analysis of the available data set. The chapter is sub-divided into four sections. The first section gives a brief description of the data and the variables used. The second section describes the characteristics of the glioblastoma patients and the third section will present the main results of the analysis. The final section presents a summary of the results obtained from the analysis.

4.1 Data

The data set consist of thirty-two (32) variables. The neuroimage features are explored using six (6) variables while the copy number variations of patients contain 26 variables. We define the neuroimage features variables as set M and the copy number variation variables as set N. Five hundred and twenty-seven (527) GBM patients were involved in this analysis. Out of the 527 patients, only 267 patients had a corresponding MRI of their tumor available. Hence for the main analysis, 267 patients were involved.

4.1.1 Patient Features

The VASARI lexicon for magnetic resonance imaging annotation contains several imaging descriptors based on different magnetic resonance imaging modalities [13]. The cardinal image features as presented by Gutman et al [13] in their paper are edema, necrosis, non Contrast-enhancing tumor (nCet) and enhancing. We added two more features, the major axis length and minor axis length of the tumor to the cardinal features. So the following magnetic resonance imaging features of Gliobastoma patients available on the Can-

cer Imaging Archive (TCIA) were used for the analysis: edema, necrosis, non Contrast-enhancing tumor, enhancing tumor, major axis length and minor axis length. Table 4.1 lists each image feature with its description.

The copy number variations of the Glioblastoma patients was obtained from the The Cancer Genome Atlas (TCGA). The variables under the copy number variations are measured as homozygous deletion, hemizygous deletion, neutral/no change, gain and high level amplification. Further information about the patients was acquired from TCGA to assess some characteristic features of the patients. Table 4.2 gives the variables (genes) in the copy number variation for the patients.

Table 4.1: Description of Neuro-image Features Used

Variable Name	Description
Edema	What proportion of the abnormality is vasogenic edema? It is an accumulation of fluid in the brain that happens when the blood-brain barrier is broken. Edema should be greater in signal than nCET and somewhat lower in signal than CSF. (Pseudopods are characteristic of edema)
Proportion Necrosis	Defined as the region within the tumor that does not enhance or shows markedly diminished enhancement, is high on T2W and proton density images, is low on T1W images, and has an irregular border
Proportion Enhancing	Proportion of tumor that is enhancing. (Assuming that the entire abnormality may be comprised of: (1) an enhancing component, (2) a nonenhancing component, (3) a necrotic component and (4) an edema component.)
Proportion nCet	Defined as the regions of T2W hyperintensity (less than the intensity of cerebrospinal fluid, with corresponding T1W hypointensity) that are associated with mass effect and architectural distortion, including blurring of the gray-white interface.(Assuming that the the entire abnormality may be comprised of: (1) an enhancing component, (2) a nonenhancing component, (3) a necrotic 9= Indeterminate component and (4) an edema component.)
Major Axis	Largest perpendicular$(x-y)$ cross-sectional diameter of T2 signal abnormality measured on a single axial image only
Minor Axis	Smallest perpendicular$(x-y)$ cross-sectional diameter of T2 signal abnormality measured on a single axial image only

Table 4.2: Copy Number Variation Variables (Genes)

Variables	Label
akt1	AKT serine/threonline kinase 1
akt2	AKT serine/threonline kinase 2
akt3	AKT serine/threonline kinase 3
ccnd2	cyclin D2
cdk4	cyclin dependent kinase 4
cdk6	cyclin dependent kinase 6
cdk2na	cyclin dependent kinase inhibitor 2A
cdkn2c	cyclin dependent kinase inhibitor 2C
egfr	epidermal growth factor receptor
erbb2	erb-b2 receptor tyrosine kinase 2
foxo1	forkhead box C1
foxo3	forkhead box C3
hras	HRas proto-oncogene, GTPase
kras	KRAS proto-oncogene, GTPase
mdm2	MDM2 proto-oncogene
mdm4	MDM4 proto-oncogene
met	MET proto-oncogene, receptor tyrosine kinase
nf1	neurofibromin 1
nras	neuroblastoma RAS viral oncogene homolog
pdgfra	platelet derived growth factor receptor alpha
pik3ca	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
pik3r1	phosphoinositide-3-kinase regulatory subunit 1
pten	phosphatase and tensin homolog
rb1	RB transcriptional corepressor 1
spry2	sprouty RTK signaling antagonist 2
tp53	tumor protein p53

4.2 Preliminaries

This section seeks to describe some notable characteristics of the Glioblastoma patients. The characteristics range from sex, age of diagnosis, survival status (Deceased or Living), the expression subtype and overall survival status of patient after diagnosis (Length of time from diagnosis to death). Frequencies and descriptives of these variables will be presented and discussed.

Observations from table 4.3 are that, of the 527 GBM patients, the majority (61.5%) are males. Also, about seven out of every ten (77%) of the patients are deceased as at March

2016. The mean survival time from time of diagnosis to death was recorded to be 15 months with a standard deviation of 16.53. The mean age of diagnosis was obtained as 58 (Table 4.4). The survival time and age of diagnosis from our data set conforms to the cancer statistics in 2012 [35] which stated that GBM is generally diagnosed at an average age of 55 years, and gives the affected patient an average survival time of only 10 to 18 months.

Table 4.3: Sex and Survival Status Distribution of Patients

Characteristic		Frequency	Percentage
Sex:	Male	324	61.5
	Female	203	38.5
Survival Status:	Deceased	406	77.0
	Living	121	23.0

Table 4.4: Age and Overall Survival Time of Patients

Variable	Minimum	Maximum	Mean	SD
Age (in years)	10	89	58.23	14.31
Survival time (in months)	0	128	15.10	16.54

The Cancer Genome Atlas (TCGA) in 2011 indicated four distinct expression subtypes of GBM [1]. The four subtypes were Classical, Proneural, Neural and Mesenchymal. The Classical GBM tumors are always characterized by extremely high levels of EGFR. However, the abnormality of the EGFR gene occur a lower rate in the three subtypes. Furthermore, there is no mutation of the most mutated gene tumor protein p53(TP 53) in GBM in the Classical GBM tumors. The TP53 is however significantly mutated in the Proneural tumors. Only Proneural tumors have abnormally high levels of mutations of PDGFRA. The most frequent number of mutations in the tumor suppressor gene NF1 can be found in the Mesenchymal group. Also, tumor suppressor genes such as TP53 and PTEN have frequent mutations in this group. For the Neural group, there is no stand out gene that exists in abnormally higher or lower mutation rate [1].

There has also been an identification of a CpG Island Methylator Phenotype (G-CIMP) that also presents a distinct subgroup of GBM [25].

Table 4.5 shows that majority (26.5%) of the GBM patients in our dataset have the Mesenchymal subtype, followed by the Classical subtype (25.1%).

Table 4.5: Frequency Distribution of Expression Subtype

Subtype	Frequency	Percent
Classical	144	25.1
G-CIMP	38	6.6
Mesenchymal	152	26.5
Neural	83	14.5
Proneural	97	16.9
Not Available	13	2.3

4.3 Main Results

4.3.1 Correlation matrix of variables

Canonical correlation analysis demands that there exist no high correlations within each of the sets of variables. So we checked for correlations among the sets of variables.

Tables 4.6 and 4.7 lists the correlation coefficients between each variable set. Variable set 1 is the VASARI neuroimage features whereas variable set 2 is the copy number variations variables. Table 4.6 shows the correlations between the VASARI neuroimage features and Table 4.7 presents the correlations between the copy number variation variables. Among the VASARI features, observations showed that the farthest correlation coefficient from zero that existed was -0.6443, which is the correlation between the enhancing and edema. This depicts that, as the proportion of edema increases, then the proportion of enhancing diminishes and vice versa. The major axis of the tumor has a positive relationship with the minor axis and with nCET. However, the major axis showed a negative relationship with necrosis, edema and enhancing. Edema is negatively correlated with all other features. nCET recorded a positive relationship with the major axis, the minor axis and necrosis.

Moreover, for the copy number variations, the farthest correlation coefficient from zero

recorded among the 26 variables was 0.8962 (Table 4.7). This relationship existed between the foxo1 gene and rb1 gene. This relationship shows that the foxo1 gene and rb1 gene has a strong direct positive relationship, hence an amplification of a patient's foxo1 gene will result in the amplification of the patient's rb1 gene and vice versa.

Table 4.6: Correlations for Variable Set 1

	Major Axis	Minor Axis	Necrosis	Edema	nCET	Enhancing
Major Axis	1.0000					
Minor Axis	0.4828	1.0000				
Necrosis	-0.0152	0.1356	1.0000			
Edema	-0.0752	-0.3974	-0.2578	1.0000		
nCET	0.1168	0.1724	0.0160	-0.2488	1.0000	
Enhancing	-0.0203	0.2519	-0.0034	**-0.6443**	-0.1208	1.0000

Table 4.7: Correlations for the Copy Number Variation Variables

Variables	akt1	akt2	akt3	ccnd2	cdk4	cdk6	cdk2na	cdk2nc	egfr	erbb2	foxo1	foxo3
Akt1	1.00											
Akt2	0.3072	1.00										
Akt3	0.0244	0.2297	1.00									
ccnd2	-0.0585	0.0806	-0.0109	1.00								
cdk4	-0.1009	0.0722	0.0887	0.4062	1.00							
cdk6	0.0921	0.1925	0.0311	0.0562	0.0524	1.0000						
cdk2na	0.0569	-0.1651	0.0470	0.0066	0.2427	-0.1943	1.0000					
cdkn2c	0.1730	0.0992	0.4207	-0.0629	0.0176	-0.0088	0.1245	1.0000				
egfr	0.2634	0.2630	-0.0267	0.0035	0.1480	0.5317	-0.1641	0.0125	1.0000			
erbb2	0.0487	-0.0836	0.0751	-0.2019	0.0057	-0.1688	-0.0588	-0.0746	-0.0058	1.0000		
foxo1	0.1566	-0.0414	-0.1745	0.0001	0.0118	-0.0048	-0.1429	-0.1155	0.2134	0.2780	1.0000	
foxo3	0.4316	0.1600	0.0577	-0.2336	-0.1191	-0.1039	0.1712	0.1527	0.0310	0.2872	0.0328	1.0000
hras	0.0757	-0.0546	-0.1156	-0.0973	-0.0315	0.0575	-0.1985	-0.0313	0.2057	0.0090	-0.0644	0.0314
kras	-0.1672	0.0850	0.1797	0.6658	0.4086	-0.0853	-0.0583	0.0909	-0.0645	-0.1893	-0.0974	-0.2586
mdm2	-0.0172	0.1293	0.2369	0.3919	0.6349	0.1119	0.0436	0.1182	0.2073	-0.0677	-0.0115	-0.1872
mdm4	0.0080	0.0050	0.4332	0.0200	0.0799	-0.0355	0.0477	0.2399	0.0619	0.0623	-0.2043	0.0954
met	0.1184	0.1034	-0.0171	0.0509	0.0340	0.6795	-0.0528	-0.0016	0.3681	-0.2262	-0.0306	-0.0205
nf1	0.0560	-0.0036	0.0588	-0.1925	0.0587	-0.1221	0.0258	-0.0979	0.0446	0.8323	0.2801	0.2985
nras	-0.0228	0.0153	0.5424	-0.0938	0.0172	-0.0099	0.1558	0.4750	-0.0369	0.1158	-0.1737	0.2416
pdgfra	-0.1186	-0.0684	0.0142	-0.1621	0.0568	-0.1763	0.0188	-0.0579	-0.1502	0.0691	0.0276	0.0687
pik3ca	-0.1534	-0.1360	0.0782	-0.2153	-0.1913	-0.0298	-0.0921	0.0439	-0.0283	0.1617	-0.1309	-0.0823
pik3r1	-0.0123	0.1168	-0.1214	0.0781	0.0712	0.3163	-0.1567	0.1627	0.1767	-0.1206	0.1065	-0.0631
pten	-0.0678	-0.0984	-0.0462	0.1408	-0.0412	-0.3242	0.2137	0.0579	-0.1502	0.0691	0.1718	0.0655
rb1	0.2131	0.0598	-0.2147	0.0782	0.0633	0.346	-0.1346	-0.1344	0.2282	0.2064	**0.8962**	-0.0277
spry2	0.0444	-0.0568	-0.0835	-0.0375	0.0012	0.0258	-0.0870	-0.0031	0.1697	0.2115	0.7361	0.0089
tp53	0.1030	0.0385	-0.2054	-0.0149	0.1742	-0.0396	-0.1314	-0.0800	0.0831	0.5962	0.3445	0.1749

Table 4.8: Correlations for the Copy Number Variation Variables

Variables	hras	kras	mdm2	mdm4	met	nf1	nras	pddfra	pik3ca	pik3r1	pten	rb1	spry2
hras	1.0000												
kras	-0.0013	1.0000											
mdm2	0.0796	0.4919	1.0000										
mdm4	0.169	0.0642	-0.0025	1.0000									
met	0.0723	-0.0967	0.0742	-0.0179	1.0000								
nf1	0.0468	-0.1841	-0.0940	0.0228	-0.1840	1.0000							
nras	-0.0541	0.0751	0.0380	0.3072	0.0079	0.0950	1.0000						
pdgfra	-0.2558	-0.1153	0.0211	0.1107	-0.0817	0.1272	-0.0608	1.0000					
pik3ca	-0.0818	-0.1110	-0.1209	0.0405	-0.0035	0.0448	0.0536	-0.0132	1.0000				
pik3r1	0.0034	0.0180	0.0826	-0.0591	0.3858	-0.1868	0.0174	-0.0731	-0.1010	1.0000			
pten	-0.0028	0.0675	-0.0102	-0.0363	-0.2668	0.0790	0.1167	-0.0271	-0.0257	0.0200	1.0000		
rb1	-0.0856	-0.1199	0.0180	-0.2309	0.0040	0.2060	-2139	-0.0022	-0.1577	0.1360	0.1456	1.0000	
spry2	-0.0330	-0.1979	-0.0595	-0.0160	0.0002	0.1884	0.0044	-0.0293	-0.1076	0.1065	0.2381	0.7404	1.0000
tp53	0.1641	-0.1804	0.0810	-0.1645	-0.1200	0.6250	-0.1912	0.1097	-0.2035	0.1164	0.0355	0.3393	0.2742

Table 4.9: Correlations between Variable Set 1 and Variable Set 2

Variables	Major Axis	Minor Axis	Necrosis	Edema	nCET	Enhancing
Akt1	-0.0289	-0.2096	0.1527	0.0281	-0.2009	-0.0587
Akt2	0.0271	-0.0422	-0.0062	-0.0024	-0.1415	0.0594
Akt3	0.1124	0.1127	0.0157	0.1128	-0.0196	0.0094
ccnd2	-0.0192	-0.1031	0.0303	-0.0629	0.0140	0.0103
cdk4	0.0141	-0.0816	0.1420	0.0031	-0.3819	0.0483
cdk6	-0.0236	-0.0024	-0.0659	-0.0416	-0.2304	0.0700
cdk2na	-0.0265	-0.1238	0.0559	0.1030	0.1351	-0.2265
cdkn2c	-0.1915	-0.0106	0.1025	-0.0262	0.1272	0.0155
egfr	0.0772	0.0195	-0.0626	0.0062	-0.1012	0.0867
erbb2	0.1583	0.1254	-0.0789	0.1716	0.0138	-0.1357
foxo1	0.0979	0.3049	-0.1601	-0.1051	-0.0991	0.1337
foxo3	0.0067	0.0066	0.1963	0.0420	-0.0899	-0.0708
hras	-0.0282	-0.0790	-0.1048	0.1182	0.0174	-0.0265
kras	0.0288	-0.0641	0.0527	-0.0821	0.0586	0.0724
mdm2	-0.0116	0.0226	0.1213	-0.0244	-0.0498	0.0478
mdm4	0.0430	0.0116	0.0812	-0.0212	0.0375	0.0729
met	-0.0707	0.0033	0.0124	-0.1086	-0.0251	0.0212
nf1	0.1517	0.2630	-0.0658	0.0481	0.0132	-0.0629
nras	-0.1572	0.1336	0.1472	0.1360	0.0256	-0.0295
pdgfra	0.2463	0.2697	0.7260	-0.1559	-0.0174	0.0033
pik3ca	-0.0046	0.0441	-0.0637	0.1438	-0.0432	-0.1074
pik3r1	-0.0621	0.0953	-0.0721	-0.0738	-0.1124	0.0897
pten	-0.3747	-0.2399	0.2036	0.0265	0.0331	-0.0672
rb1	0.0462	0.0101	-0.2429	-0.1075	-0.1338	0.1660
spry2	-0.4289	-0.0147	-0.4163	-0.0987	0.4001	0.1351
tp53	0.4475	0.0736	-0.4066	-0.0740	0.3999	-0.0047

The correlations between the copy number variation variables and the image features are presented in table 4.9. There are both negative and positive relationships between the variable sets. The highest correlation coefficient (0.7260) existed between pdgfra and necrosis. There are relatively low correlations between the two variable sets. Moderate correlations (-0.4066,-0.4163) existed between spry2, tp53 and necrosis respectively. Also, moderate correlations (0.4475,-0.4289,-0.3747) existed between tp53, spry2, pten and major axis respectively. Moreover, nCET was also moderately correlated with cdk4 (-0.3819), spry2 (0.4001), tp53 (0.3999). These bivariate correlations seem to suggest a relationship between some of the features and genes in the study.

The raw canonical coefficients are the weights of the M-variables and the N-variables,

maximizing the correlation among the sets of variables. The coefficients are interpreted the same way as the regression coefficients. So from Table 4.10, for the variate M_1, a unit increase in the proportion of necrosis leads to a 1.6797 increment on the first canonical variate of the N-variable set, with all other variables to be held constant.

Table 4.10: Raw Coefficients for the Neuro-image features

	1	2	3	4	5	6
Major Axis	0.4264	0.1005	0.6760	0.2438	0.2857	-0.2874
Minor Axis	0.1240	-0.6065	-0.4383	-0.2898	0.0600	0.0756
Necrosis	1.6797	2.2665	-2.3086	0.3454	1.3586	-1.6622
Edema	0.6631	-0.6532	-1.1169	1.4566	-0.7277	-1.1151
nCET	-1.2893	-0.3184	-0.3771	0.8520	0.5133	-1.5226
Enhancing	0.2989	0.0219	-0.2357	0.2624	-1.1305	-1.6793

Table 4.11: Raw Coefficients for the Copy Number Variation Variables

Variables	1	2	3	4	5	6
Akt1	0.1207	0.9216	-0.0343	0.2824	0.2217	0.4739
Akt2	-0.0602	-0.1999	0.0780	0.1185	-0.1724	0.4816
Akt3	0.7972	-0.7069	1.1209	0.6279	-0.0815	-0.3316
ccnd2	0.2203	-0.6280	-0.2054	0.1681	0.4499	0.7753
cdk4	0.7373	0.7428	-0.0870	0.1356	-0.5273	0.3419
cdk6	0.3429	0.3914	0.2704	-0.0327	-0.4198	0.7079
cdk2na	-0.3741	-0.3616	0.3788	0.5088	0.4692	0.2825
cdkn2c	-0.6646	0.2594	-0.3989	-0.1759	0.1041	-0.3262
egfr	0.1092	-0.3501	0.2107	0.2607	0.1746	-0.6444
erbb2	0.3651	-0.2714	-0.0902	2.4166	0.1850	0.0921
foxo1	0.6733	-1.2847	0.0458	-0.5207	0.7877	1.0521
foxo3	0.5249	0.3438	-0.0674	-0.3994	-0.0492	0.0241
hras	0.3176	-0.4606	0.1801	0.7627	-0.2045	0.2671
kras	-0.9254	0.6619	1.1563	-0.0723	0.3101	-1.0851
mdm2	-0.1203	-0.2112	-0.6641	-0.4173	0.0846	-0.1012
mdm4	-0.1421	0.3140	0.1615	-0.2607	-0.0856	-0.6363
met	-0.8376	0.4377	-0.1256	-0.4054	0.6902	0.0058
nf1	0.1457	-0.3427	-0.0013	-1.6359	0.0401	0.2830
nras	0.1661	-0.5333	-1.9065	-0.2347	-0.1279	-0.2008
pdgfra	0.4226	-0.0704	0.0563	-0.3877	0.7704	0.0288
pik3ca	0.0430	-0.1970	-0.0217	-0.0556	-0.0546	0.8620
pik3r1	0.6607	-0.8831	0.1641	-0.4134	-0.4106	0.8242
pten	0.0186	1.3438	-0.4755	0.1518	0.0121	0.1979
rb1	0.5788	0.9265	0.4969	0.0324	-1.0624	-1.5470
spry2	-1.4400	-0.0932	0.0854	-0.1515	-0.6050	0.6273
tp53	-1.4510	-0.0840	0.3839	-0.4220	0.6305	-0.4147

4.3.2 Assessment of Overall Model Fit

We now present results on the overall statistical fit of the entire model. The multivariate F-tests and its corresponding Wilk's lambda evaluate the hypothesis below.

H_0 : The canonical correlation coefficient for all functions are zero.

H_1 : The canonical correlation coefficient for at least one function is not zero.

Again, we check against the null hypothesis that each of the canonical functions' canonical correlation coefficient is zero.

From Table 4.12, we have that the null hypothesis for the entire model is rejected at 0.05 significance level, hence we can conclude that at least one canonical function has a non-zero canonical correlation coefficient. Also, we confirm from Table 4.13 that the first three canonical correlation coefficients are statistically significant at a significance level of 0.05. This means that the null hypothesis, which states that the canonical correlation coefficient of each of the the first three canonical function is zero is rejected. The remaining three correlation coefficients are not significant based on the multivariate F-tests and Wilk's lambda. This means that the remaining coefficients will not be subjected to interpretations.

Table 4.12: Test of Significance of all Canonical Correlations

	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	0.127081	156	1386.69	3.7459	0.0000

Table 4.13: Test of Significance of each Canonical Correlation

Test of Canonical Correlation 1					
	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	0.127081	156	1386.69	3.7459	0.0000

Test of Canonical Correlation 2					
	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	0.230809	125	1166.35	3.2384	0.0000

Test of Canonical Correlation 3					
	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	0.38655	96	941.39	2.6591	0.0001

Test of Canonical Correlation 4					
	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	0.730162	69	350.64	1.330	0.0514

Test of Canonical Correlation 5					
	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	812344	44	248.03	1.2001	0.1957

Test of Canonical Correlation 6					
	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	0.894344	21	160.12	1.0831	0.3711

The canonical correlation coefficient and eigenvalues or canonical roots for each of the functions are shown in Table 4.14. The magnitude of the relationship occurring between the variate pairs is given by the canonical correlation coefficient.

Table 4.14: Canonical Correlations and Eigenvalues

Coefficients	0.6704	0.6347	0.5552	0.4844	0.4285	0.3250
Eigenvalues	0.4494	0.4028	0.3082	0.2346	0.1836	0.1056

Table 4.15 presents the canonical redundancy index for the canonical correlations. In the

first canonical function, the redundancy for the M-variables is 0.2012 and the redundancy for the N-variables is 0.2101. The values obtained depict that each variate explains almost the same amount of variance in the opposite set of variables in the canonical function. Considering the second function, the redundancy measure for the M and N variables are 0.1876 and 0.1501. This means that the variate for the N-variables explains less variance in the M-variables in the first function than the variate for the M- variables explains in the set of N-variables.

Table 4.15: Canonical redundancy analysis for Canonical Correlations

Canonical redundancy analysis for Canonical Correlation 1		
Canonical Correlation Coefficient		0.6704
Squared Canonical Correlation Coefficient		0.4494
Proportion of standardized variance	O.V	OP.V
of M variables with	0.3001	0.2101
of N variables with	0.3121	0.2112
Canonical redundancy analysis for Canonical Correlation 2		
Canonical Correlation Coefficient		0.6347
Squared Canonical Correlation Coefficient		0.4028
Proportion of standardized variance	O.V	OP.V
of M variables with	0.4212	0.0.1501
of N variables with	0.3212	0.1876
Canonical redundancy analysis for Canonical Correlation 3		
Canonical Correlation Coefficient		0.5552
Squared Canonical Correlation Coefficient		0.3052
Proportion of standardized variance	O.V	OP.V
of M variables with	0.3992	0.1001
of N variables with	0.3685	0.1019

O.V = Own Variate, OP.V= Opposite Variate

4.3.3 Interpreting Canonical Variate Pairs

Based on the F-test and the Wilk's lambda, we have concluded that only three canonical coefficients are significant, so we can can interpret and report the contribution of each of the variables (original) that is in the canonical function. We would then resort to the stan-

dardized canonical coefficients and or canonical loadings to elicit the relative contributions of the variables.

The canonical functions can be interpreted by observing the magnitude and sign of the standardized canonical correlation coefficient or the canonical loadings that is assigned to each original variable in its canonical variate. Variables that have higher coefficients have a higher contribution to the variate. We set a coefficient threshold of $|0.5|$ and above to depict the most important variable in the canonical function. Moreover, original variables that have coefficients with opposite signs depict an inverse association with one another. Again, original variables with coefficients that have the same sign depict a direct association. However, because the interpretation of the contribution of original variables by its canonical coefficient faces the same problems that are associated to the interpretation of beta values in the regression model, caution is taken in the interpretation of the results in canonical analysis [2]. One of the problems faced is that, the weights or the coefficients are subjected to considerable variability from a sample to the other. Therefore, the canonical loadings will also be used to assess the contribution of the original variables.

Hence, if the findings from using the standardized coefficients and the canonical loadings are similar or the same, then there is evidence for accuracy of the results.

4.3.4 Interpretation of Canonical Variate Using Canonical Weights

Here, we present the standardized coefficients and interpret them. The standardized coefficients always enable for easier comparisons among variables when the variables have varying standard deviations. So because the canonical coefficients are standardized, then we can make comparisons using their weights. The proportion of canonical correlation weights for a set of canonical roots is their relative significance for the given impact [2].

The standardized canonical coefficients for the significant functions are shown in Table 4.16. Considering the first set of variables(Neuro image features) and the first canonical function, the nCET is the most important, followed by major axis then edema and necrosis. A one standard deviation increase in proportion of necrosis leads to a 0.4280 standard deviation increase in the score on the first canonical variate in the second variable set when the other variables all held constant. Also, a one standard deviation increase in nCET leads to 0.6407 decrease in the score on the first canonical variate in the second variable set with other variables held constant. With the second canonical function, the most important features are minor axis, necrosis and edema. The third canonical function has high coefficient

values for major axis, minor axis, necrosis and edema.

Considering standardized coefficients of the copy number variations from Table 4.17, spry2, tp53, cdk4, foxo1, met, pdgfra, rb1, cdk2na, cdk2nc and akt3 are more closely related to the first canonical function since their coefficients are greater than $|0.3|$ whilst foxo1, cdk4, akt1, pten, rb1, akt3, ccnd2, cdk2na, pik3r1 and kras are most closely related to the second canonical function. For the third canonical function, nras, kras ,akt3, mdm2 and cdk2na are also more closely related to it. Table 4.18 below summarize the most important features and genes for each function based on the magnitude of the canonical loadings with a threshold of $|0.5|$ and above.

Table 4.16: Standardized Coefficients for the Neuro-image features

	1	2	3
Major Axis	0.5317	0.1253	0.8430
Minor Axis	0.1914	-0.9363	-0.6766
Necrosis	0.4280	0.5774	-0.5882
Edema	0.4327	-0.4263	-0.7288
nCET	-0.6407	-0.1582	-0.1874
Enhancing	0.2125	0.0156	-0.1675

Table 4.17: Standardized Coefficients for the Copy Number Variation Variables

Variables	1	2	3
Akt1	0.0735	0.5615	-0.0209
Akt2	-0.0355	-0.1178	0.0459
Akt3	0.3587	-0.3181	0.5040
ccnd2	0.1246	-0.3551	-0.1162
cdk4	0.6223	0.6269	-0.0734
cdk6	0.1661	0.1896	0.1310
cdk2na	-0.3365	-0.3253	0.3407
cdkn2c	-0.3274	0.1278	-0.1965
egfr	0.0755	-0.2418	0.1455
erbb2	0.1871	-0.1390	-0.0462
foxo1	0.3640	-0.6945	0.0247
foxo3	0.2905	0.1903	-0.0373
hras	0.1534	-0.2224	0.0870
kras	-0.0145	0.3219	0.5623
mdm2	-0.0895	-0.1571	-0.4939
mdm4	-0.0929	0.2052	0.1055
met	-0.4111	0.2148	-0.0617
nf1	0.0759	-0.1785	-0.0007
nras	0.0772	-0.2479	-0.8862
pdgfra	0.3257	-0.0542	0.0434
pik3ca	0.0235	-0.1076	-0.0118
pik3r1	0.2777	-0.3712	0.0690
pten	0.0077	0.5549	-0.1963
rb1	0.3224	0.5161	0.2768
spry2	-0.7778	-0.0503	0.0461
tp53	-0.6892	-0.0399	0.1823

Table 4.18: Summary of Important Related Variables

1		2		3	
Image features	Coeff.	Image features	Coeff.	Image features	Coeff.
nCET	-0.6407	Minor Axis	-0.9363	Major Axis	0.8430
Major axis	0.5317	Necrosis	0.5774	Edema	0.7288
				Minor Axis	-0.6766
				Necrosis	-0.5882
CNV		CNV		CNV	
spry2	-0.7778	foxo1	-0.6945	nras	-0.8862
tp53	-0.6892	cdk4	0.6269	kras	0.5623
cdk4	0.6223	Akt1	0.5615	Akt3	0.5040
		pten	0.5549	cdk2na	0.5001
		rb1	0.5161		

4.3.5 Interpretation of Canonical Variate Using Canonical Loadings

Observations from Table 4.19 show that major axis, nCET and necrosis were most closely related to the first canonical function since their coefficients were greater than |0.3|. The second canonical function is closely related to minor axis, necrosis and major axis. The third function is most related to major axis and necrosis.

From table 4.20, tp53, spry2 cdk4, pdgfra and cdk2na are closely related to the first function while akt1, pten, foxo1, akt3, cdk4, nf1, erbb2 and rb1 are closely related to the second function. Also, nras, cdkn2c,cdkn2a, foxo1, mdm2, rb1, akt3 and kras are closely related to the third. Table 4.21 below summarizes the most important features and genes for each function based on the magnitude of the canonical loadings with a threshold of |0.5| and above.

Table 4.19: Canonical Loadings for the Neuro-image features

	1	2	3
Major Axis	0.5059	-0.3222	0.5615
Minor Axis	0.2772	-0.6514	-0.1343
Necrosis	0.3233	0.5559	-0.5072
Edema	0.2289	-0.1832	-0.2172
nCET	-0.6721	-0.1915	-0.0134
Enhancing	0.0470	0.0689	0.1391

Table 4.20: Canonical Loadings for the Copy Number Variation Variables

Variables	1	2	3
Akt1	0.1107	0.5473	0.0647
Akt2	0.1580	0.1002	0.1321
Akt3	0.2259	-0.4004	-0.5277
ccnd2	-0.0760	0.2148	0.1390
cdk4	0.6696	0.5968	0.0133
cdk6	0.0584	0.0010	0.1143
cdk2na	-0.3552	0.2199	-0.5610
cdkn2c	-0.2230	0.0574	-0.3996
egfr	0.1550	-0.0473	0.1596
erbb2	0.1655	-0.3476	-0.0178
foxo1	0.2746	-0.6825	0.3215
foxo3	0.2231	0.1626	-0.2092
hras	-0.0606	-0.0688	0.0115
kras	-0.0525	0.2893	0.5230
mdm2	0.1230	0.1033	-0.3419
mdm4	0.0629	0.0719	-0.0418
met	-0.0866	0.0721	0.0202
nf1	0.1233	-0.3074	0.0529
nras	0.0614	-0.1925	-0.8356
pdgfra	0.3336	-0.0345	0.0154
pik3ca	0.0684	-0.2123	-0.1349
pik3r1	0.0202	-0.1385	-0.0263
pten	-0.1482	0.6384	-0.2301
rb1	0.0594	-0.5262	0.3453
spry2	-0.7745	-0.1333	0.1369
tp53	-0.6364	-0.1668	0.1664

Table 4.21: Summary of Important Related Variables

1		2		3	
Image features	Loading	Image features	Loading	Image features	Loading
nCET	-0.6721	Minor Axis	-0.6514	Major Axis	0.5615
Major axis	0.5059	Necrosis	0.5559	Necrosis	0.5072
CNV		CNV		CNV	
spry2	-0.7745	foxo1	-0.6825	nras	-0.8356
cdk4	-0.6696	pten	0.6384	cdk2na	-0.5615
tp53	-0.6364	cdk4	0.5968	Akt3	-0.5277
		Akt1	0.5473	kras	0.5230
		rb1	-0.5262		

Since the two methods of interpretation, using the standardized coefficients and canonical loadings, resulted in the similar conclusions, we are more confident in our findings and hence move on to conduct model validation in the next section of the thesis.

4.3.6 Cross Validation

In this section, we subject our model to validation. There are various approaches in model validation. We validate our model by using the sample splitting approach. The entire sample (267) is divided into two sub-samples and the canonical correlation analysis is conducted separately on each of the sub-samples. We then compare the results obtained from each of the analyses.

4.3.7 CCA on Sub-Sample A

The first sub-sample contains 134 patients. From the six canonical functions, only two of the functions were significant from the F-tests and Wilk's lambda observations (see Table 4.22). Hence we present results on the canonical loadings of each of the variable set for only the significant functions. Table 4.23 and 4.24 shows the contributions of each variable in the each of the canonical functions. The significant canonical correlation coefficients for the new sample were found to be 0.6601 and 0.6372.

Table 4.22: Test of Significance of each Canonical Correlation

Test of Canonical Correlation 1-6					
	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	0.118385	156	606.447	1.7054	0.000

Test of Canonical Correlation 2-6					
	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	0.226496	125	511.824	1.4423	0.0033

Test of Canonical Correlation 3-6					
	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	0.383356	96	414.513	1.1824	0.1366

Test of Canonical Correlation 4-6					
	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	0.564471	69	314.54	0.9617	0.5660

Test of Canonical Correlation 5-6					
	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	0.735015	44	212	0.8018	0.8069

Test of Canonical Correlation 6					
	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	0.902262	21	107	0.5519	0.9408

Major axis and nCET were the most important variables in the first function since their coefficients were equal to or greater than |0.5| while minor axis and necrosis were the most important variables in the second function. In table 4.24, we observed that spry2, tp53 and cdk4 were the most important variables in the first function. Akt1, cdk4, pten, rb1 and foxo1 were the most important variables in the second canonical function.

Table 4.23: Canonical Loadings for the Neuro-image features

	1	2
Major Axis	0.5091	-0.3300
Minor Axis	0.2534	-0.6565
Necrosis	0.3137	0.5473
Edema	0.2633	-0.1854
nCET	-0.7192	-0.1742
Enhancing	0.0638	0.0534

Table 4.24: Canonical Loadings for the Copy Number Variation Variables

Variables	1	2
Akt1	0.0866	0.5843
Akt2	0.1538	0.1053
Akt3	0.2341	-0.2273
ccnd2	-0.1114	0.2097
cdk4	0.5836	0.5938
cdk6	0.0724	0.0135
cdk2na	-0.2017	0.0954
cdkn2c	-0.2398	0.0821
egfr	0.1729	-0.0423
erbb2	0.1869	-0.2338
foxo1	0.1791	-0.5588
foxo3	0.2398	0.1746
hras	-0.0432	-0.0606
kras	-0.0989	0.1855
mdm2	0.1222	0.0973
mdm4	0.1207	0.0428
met	-0.1068	0.0827
nf1	0.1457	-0.2965
nras	0.0809	-0.2096
pdgfra	0.3263	-0.0160
pik3ca	0.0965	-0.2645
pik3r1	0.0369	-0.1067
pten	-0.1087	0.5057
rb1	0.0566	-0.5162
spry2	-0.6601	-0.1130
tp53	-0.5210	-0.1153

4.3.8 CCA on Sub-Sample B

Sub-sample B contains 133 patients. Also, only two of the functions were significant from the F-tests and Wilk's lambda observations (see Table 4.25). Therefore only the results from the significant functions will be presented and interpreted. Tables 4.26 and 4.27 shows the contributions of each variable in each of the canonical functions. The significant canonical correlation coefficients for this analysis were obtained as 0.6543 and 0.6338.

Table 4.25: Test of Significance of each Canonical Correlation

Test of Canonical Correlation 1-6					
	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	0.128719	156	600.581	1.6104	0.0000

Test of Canonical Correlation 2-6					
	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	0.225069	125	506.903	1.4354	0.0037

Test of Canonical Correlation 3-6					
	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	0.376206	96	410.551	1.1970	0.1202

Test of Canonical Correlation 4-6					
	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	0.539844	69	311.552	1.0348	0.4120

Test of Canonical Correlation 5-6					
	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	0.721821	44	210	0.8449	0.7430

Test of Canonical Correlation 6					
	Statistic	df1	df2	F	Prob>F
Wilk's Lambda	0.880579	21	106	0.6845	0.8398

Major axis and nCET were the most important variables in the first function since their coefficients were equal to or greater than $|0.5|$ while minor axis and necrosis were the most important variables in the second function. Observations from Table 4.27 revealed that

spry2, tp53 and cdk4 were the most important variables in the first function. Akt1, cdk4, pten, rb1 and foxo1 were the most important variables in the second canonical function.

Table 4.26: Canonical Loadings for the Neuro-image features

	1	2
Major Axis	0.5889	0.3072
Minor Axis	0.2899	0.6673
Necrosis	0.4356	-0.5342
Edema	0.1722	0.1771
nCET	-0.6209	0.1970
Enhancing	0.0388	-0.0818

Table 4.27: Canonical Loadings for the Copy Number Variation Variables

Variables	1	2
Akt1	0.1421	-0.6099
Akt2	0.1683	-0.0947
Akt3	0.2054	0.1766
ccnd2	-0.0259	-0.2243
cdk4	0.5612	-0.5921
cdk6	0.0490	0.0116
cdk2na	-0.1022	-0.1496
cdkn2c	-0.2065	-0.0216
egfr	0.1347	0.0505
erbb2	0.1261	0.3605
foxo1	0.0707	0.5014
foxo3	0.1995	-0.1374
hras	-0.0823	0.0706
kras	0.0111	-0.1975
mdm2	0.1303	-0.1008
mdm4	-0.0023	-0.0977
met	-0.0541	0.0548
nf1	0.0878	0.3180
nras	0.0196	0.1945
pdgfra	0.3428	0.0677
pik3ca	0.0244	0.1641
pik3r1	-0.0001	0.1765
pten	-0.1991	-0.5676
rb1	0.0668	0.5293
spry2	-0.6229	0.1449
tp53	-0.5531	0.2149

4.4 Summary

The study investigated a model that links some neuroimage features (six features) with copy number variations (26 genes) of Glioblastoma patients.

Wilk's lambda and F-tests were employed to evaluate the null hypothesis that canonical correlation coefficients for all the canonical functions are zero. From our model, only the first three canonical correlation coefficients are statistically significant, thus with a p-value less than 0.05. The other three functions were not significant and hence was not interpreted.

With our 3 significant canonical variate pairs, the strength of the relationship was depicted

by the canonical correlation coefficient. The first pair of canonical variates (first canonical function) had a coefficient of 0.6704. The second canonical function had a coefficient of 0.6347 and the third pair of variate had a canonical correlation coefficient of 0.5552 Squaring the canonical correlation coefficients shows the proportion of variance accounted between the two optimally weighted variates.

The redundancy index measured the proportion of variance of the M-set of variables that is predicted from the linear combination of the N-set of variables. The redundancy index can only be equal to 1 if the the squared canonical coefficient (eigenvalue) is 1 and the variables for the canonical function amount to all the variations of every variable in the set. The M-variables in the first function had redundancy index to be 0.2012, and N-variables had redundancy index to be 0.2101. The second function had a redundancy measure of 0.1876 for the $M-$variables and 0.1501 for the N-variables. For the third function, redundancy index was equal to 0.1001 and 0.1019 for the M-variables and N-variables respectively.

The canonical loadings and standardized canonical coefficients were employed to evaluate the importance of the variables in the function. A coefficient threshold of $|0.5|$ and above were used to select the important variables in each function. The standardized canonical coefficients showed that, for the first function, major axis, nCET, spry2, tp53 and cdk4 were the most important variables. Minor axis, necrosis, foxo1, rb1, pten, cdk4 and are the most important variables in the second function. For the third function, major axis, edema, minor axis, necrosis, nras,cdk2na, kras and akt3 are the most important variables.

Using the canonical loadings, we obtained that for the first function, the most important variables were nCET, major axis, spry2, cdk4 and tp53. The important contributing variables in the second function were minor axis, necrosis, foxo1, pten, cdk4, akt1 and rb1. For the third function, major axis, necrosis, nras, cdk2na, akt3 and kras were the most important variables.

We performed cross validations to check if the results were influenced by the number of samples. So the 267 sample was divided into two and the canonical correlation analysis was performed on both samples. Results from both samples indicated that only two functions were significant and hence should be interpreted. For sample A, the first canonical variate pair had a canonical coefficient of 0.6601 while the second variate pair had a canonical correlation coefficient as 0.6372 Considering the first function, nCET, major axis, spry2, cdk4, tp53 are most closely related and are most important. With the second function, akt1, cdk4, foxo1 ,pten and rb1 was the most important variables. For sample B, the canonical correlation coefficients were obtained to be 0.6543 and 0.6338. The same set of variables from the first sample were found to be important in the second sample.

Chapter 5

Conclusion

Canonical correlation analysis is a very powerful and important technique for investigating the relationship between multiple independent and dependent variables. Although the technique is fundamentally descriptive, it can also be employed for predictive purposes. This thesis provided a review of canonical correlation analysis and applied it in exploring the relationship between the copy number variations and neuro-image features of Glioblastoma patients.

Canonical correlation coefficients under a non-singular transformation are unchanged and the canonical correlation coefficients either from the correlation matrix or the covariance matrix yield the same values. Also, computing correlations by standardizing the original variables has no effect on the correlations.

We obtained from the data that mean survival status for Glioblastoma is 15 months and mean age of diagnosis is 55 years.

The two set of multiple variables were related in three ways. We obtained three pairs of significant canonical variates with correlations of $0.6704, 0.6347$ and 0.5552 respectively, which were used to identify genes and features related to Glioblastoma. The important genes and features forming these relationships are as follows. The major axis of the tumor, the non-contrast enhancing tumor, the sprouty RTK signaling antagonist 2, the tumor protein p53 and cyclin dependent kinase 4 are very much related. Also, minor axis of the tumor, proportion of necrosis, forkhead box C1, phosphatase and tensin homolog, RB transcriptional corepressor 1, AKT serine/ threonline kinase 1 and cyclin dependent kinase 4 are also very much related. Finally, we also obtained that major axis, proportion of necrosis, neuroblastoma RAS viral oncogene homolog, cyclin dependent kinase inhibitor 2A, AKT serine/threonline kinase 3 and KRAS prott-oncogene, GTPase are highly related.

References

[1] Bartek, J., Ng, K., Fischer, W., Carter, B., and Chen, C. C. (2012). *Key concepts in glioblastoma therapy*. Journal of Neurology, Neurosurgery & Psychiatry, 83(7):753–760.

[2] Cliff, N. and Krus, D. J. (1976). *Interpretation of canonical analysis: Rotated vs. unrotated solutions*. Psychometrika, 41(1):35–42.

[3] CNV (Accessed March 2016). *Copy number variants*. DNA Learning Center, http://www.dnalc.org/view/552-Copy-Number-Variants.html.

[4] Davies, E. B. (2007). *Approximate diagonalization*. SIAM Journal on Matrix Analysis and Applications, 29(4):1051–1064.

[5] de Koning, A. J., Gu, W., Castoe, T. A., Batzer, M. A., and Pollock, D. D. (2011). *Repetitive elements may comprise over two-thirds of the human genome*. PLoS Genet, 7(12):e1002384.

[6] Denman, E. D. (1981). *Roots of real matrices*. Linear Algebra and its Applications, 36:133–139.

[7] Denman, E. D. and Beavers, A. N. (1976). *The matrix sign function and computations in systems*. Applied mathematics and Computation, 2(1):63–94.

[8] Duerr, E.-M., Rollbrocker, B., Hayashi, Y., Peters, N., Meyer-Puttlitz, B., Louis, D. N., Schramm, J., Wiestler, O. D., Parsons, R., Eng, C., et al. (1998). *PTEN mutations in gliomas and glioneuronal tumors.* Oncogene, 16(17).

[9] Ganigi, P., Santosh, V., Anandh, B., Chandramouli, B., and Sastry Kolluri, V. (2005). *Expression of p53, EGFR, pRb and bcl-2 proteins in pediatric glioblastoma multiforme: a study of 54 patients*. Pediatric neurosurgery, 41(6):292–299.

[10] Genetic Variability (Accessed May 2016). *Copy Number Variations*. Pathway detail - flipper e nuvola http://flipper.diff.org/app/pathways/3685.

[11] Gevaert, O., Mitchell, L. A., Achrol, A. S., Xu, J., Echegaray, S., Steinberg, G. K., Cheshier, S. H., Napel, S., Zaharchuk, G., and Plevritis, S. K. (2014). *Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features*. Radiology, 273(1):168–174.

[12] Giunti, L., Pantaleo, M., Sardi, I., Provenzano, A., Magi, A., Cardellicchio, S., Castiglione, F., Tattini, L., Novara, F., Buccoliero, A. M., et al. (2014). *Genome-wide copy number analysis in pediatric glioblastoma multiforme*. Am J Cancer Res, 4:293–303.

[13] Gutman, D. A., Cooper, L. A., Hwang, S. N., Holder, C. A., Gao, J., Aurora, T. D., Dunn Jr, W. D., Scarpace, L., Mikkelsen, T., Jain, R., et al. (2013). *MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set*. Radiology, 267(2):560–569.

[14] Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L., et al. (2006a). *Canonical Correlation Analysis: A Supplement to Multivariate Data Analysis*, volume 6. Pearson Prentice Hall Upper Saddle River, NJ.

[15] Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L., et al. (2006b). *Multivariate data analysis*, volume 6. Pearson Prentice Hall Upper Saddle River, NJ.

[16] Hammoud, M. A., Sawaya, R., Shi, W., Thall, P. F., and Leeds, N. E. (1996). *Prognostic significance of preoperative MRI scans in glioblastoma multiforme*. Journal of neuro-oncology, 27(1):65–73.

[17] Higham, N. J. (1987). *Computing real square roots of a real matrix*. Linear Algebra and its applications, 88:405–430.

[18] Hoskins, W. and Walton, D. (1978). *A faster method of computing the square root of a matrix*. Automatic Control, IEEE Transactions on, 23(3):494–495.

[19] Hotelling, H. (1936). *Relations between two sets of variates*. Biometrika, 28(3/4):321–377.

[20] Johnson, R. A., Wichern, D. W., et al. (2002). *Applied multivariate statistical analysis*, volume 5. Prentice hall Upper Saddle River, NJ.

[21] Lacroix, M., Abi-Said, D., Fourney, D. R., Gokaslan, Z. L., Shi, W., DeMonte, F., Lang, F. F., McCutcheon, I. E., Hassenbusch, S. J., Holland, E., et al. (2001). *A multivariate analysis of 416 patients with glioblastoma multiforme: prognosis, extent of resection, and survival*. Journal of neurosurgery, 95(2):190–198.

[22] Lin, D., Calhoun, V. D., and Wang, Y.-P. (2014). *Correspondence between fMRI and SNP data by group sparse canonical correlation analysis*. Medical image analysis, 18(6):891–902.

[23] McCarroll, S. A. and Altshuler, D. M. (2007). *Copy-number variation and association studies of human disease*. Nature genetics, 39:S37–S42.

[24] Multivariate Analysis (Accessed March 2016). *Multivariate Analysis*. Philender, http://www.philender.com/courses/multivariate/notes2/can1.html.

[25] Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., Pan, F., Pelloski, C. E., Sulman, E. P., Bhat, K. P., et al. (2010). *Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma*. Cancer cell, 17(5):510–522.

[26] Ohgaki, H., Dessen, P., Jourde, B., Horstmann, S., Nishikawa, T., Di Patre, P.-L., Burkhard, C., Schüler, D., Probst-Hensch, N. M., Maiorka, P. C., et al. (2004). *Genetic Pathways to Glioblastoma A Population-Based Study*. Cancer research, 64(19):6892–6899.

[27] Pierallini, A., Bonamini, M., Pantano, P., Palmeggiani, F., Raguso, M., Osti, M., Anaveri, G., and Bozzao, L. (1998). *Radiological assessment of necrosis in glioblastoma: variability and prognostic value*. Neuroradiology, 40(3):150–153.

[28] Pollack, I. F., Boyett, J. M., Yates, A. J., Burger, P. C., Gilles, F. H., Davis, R. L., Finlay, J. L., Group, C. C., et al. (2003). *The influence of central review on outcome associations in childhood malignant gliomas: results from the CCG-945 experience*. Neuro-oncology, 5(3):197–207.

[29] Pollack, I. F., Finkelstein, S. D., Woods, J., Burnham, J., Holmes, E. J., Hamilton, R. L., Yates, A. J., Boyett, J. M., Finlay, J. L., and Sposto, R. (2002). *Expression of p53 and prognosis in children with malignant gliomas*. New England Journal of Medicine, 346(6):420–427.

[30] Pollack, I. F., Hamilton, R. L., James, C. D., Finkelstein, S. D., Burnham, J., Yates, A. J., Holmes, E. J., Zhou, T., and Finlay, J. L. (2006). *Rarity of PTEN deletions and EGFR amplification in malignant gliomas of childhood: results from the Children's Cancer Group 945 cohort*. Journal of Neurosurgery: Pediatrics, 105(5):418–424.

[31] Pope, W. B., Sayre, J., Perlina, A., Villablanca, J. P., Mischel, P. S., and Cloughesy, T. F. (2005). *MR imaging correlates of survival in patients with high-grade gliomas*. American Journal of Neuroradiology, 26(10):2466–2474.

[32] Qu, H.-Q., Jacob, K., Fatet, S., Ge, B., Barnett, D., Delattre, O., Faury, D., Montpetit, A., Solomon, L., Hauser, P., et al. (2010). *Genome-wide profiling using single-nucleotide polymorphism arrays identifies novel chromosomal imbalances in pediatric glioblastomas*. Neuro-oncology, 12(2):153–163.

[33] Reifenberger, G. and Collins, V. P. (2004). *Pathology and molecular genetics of astrocytic gliomas*. Journal of molecular medicine, 82(10):656–670.

[34] Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., Pertz, L. M., Clark, R. A., Schwartz, S., Segraves, R., et al. (2005). *Segmental duplications and copy-number variation in the human genome*. The American Journal of Human Genetics, 77(1):78–88.

[35] Siegel, R., Naishadham, D., and Jemal, A. (2012). *Cancer statistics, 2012*. CA: a cancer journal for clinicians, 62(1):10–29.

[36] Taniguchi, Y., Choi, P. J., Li, G.-W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X. S. (2010). *Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells*. Science, 329(5991):533–538.

[37] Velazquez, E. R., Meier, R., Dunn Jr, W. D., Alexander, B., Wiest, R., Bauer, S., Gutman, D. A., Reyes, M., and Aerts, H. J. (2015). *Fully automatic GBM segmentation in the TCGA-GBM dataset: Prognosis and correlation with VASARI features*. Scientific reports, 5.

[38] Xiong, M., Dong, H., Siu, H., Peng, G., Wang, Y., and Jin, L. (2010). *Genome-Wide Association Studies of Copy Number Variation in Glioblastoma*. In *Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on*, pages 1–4. IEEE.

[39] Zarrei, M., MacDonald, J. R., Merico, D., and Scherer, S. W. (2015). *A copy number variation map of the human genome*. Nature Reviews Genetics, 16(3):172–183.