DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

PARTIAL LEAST SQUARES METHOD FOR THE ANALYSIS OF GENE EXPRESSION DATA

by
Ayça ÖLMEZ

July, 2018
İZMİR

PARTIAL LEAST SQUARES METHOD FOR THE ANALYSIS OF GENE EXPRESSION DATA

A Thesis Submitted to the
Graduate School of Natural And Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Master of
Science in Statistics, Statistics Program

by
Ayça ÖLMEZ

July, 2018
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled “**PARTIAL LEAST SQUARES METHOD FOR THE ANALYSIS OF GENE EXPRESSION DATA**” completed by **AYÇA ÖLMEZ** under supervision of **PROF. DR. AYLİN ALIN** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

....

Prof. Dr. Aylin ALIN

Supervisor

....	....
Assoc. Prof Dr Neslihan Demirel	Assist. Prof Dr. Aslı SUNER KARAKÜLAH
Jury Member	Jury Member

L. Salum

Prof. Dr. Latif SALUM
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I owe my deepest gratitude to my supervisor Prof. Dr. Aylin ALIN who made this thesis possible and gave the opportunity to continue my statistical education up to master level. Her constant support, guidance and motivation allowed me to improve myself, my capabilities and studies. Thank you for sparing your priceless and precious time for me.

I am also very grateful to Assist. Prof. Dr. Gökhan KARAKÜLAH and Assist. Prof. Dr. Aslı SUNER KARAKÜLAH for endless support, answering all my numerous questions insistently and always sharing their outstanding knowledge with me.

Words cannot describe how lucky I feel for meeting and being beyond a friend with Aylin GÖÇOĞLU. She always motivated me, listened to me, made me stronger, understand me even without speaking, and most importantly made me laugh with her unique sense of humor even the stressful times. Also, I would like to thank Yusuf Can SEVİL and Tolga YAMUT for being there for me and their friendships. Furthermore, dear Atlas thank you for lying on my computer and supporting me when every time I had to study. I “meow” you too.

Last but absolutely not least, I want to express my deepest love and thankfulness to my parents and my sisters for my whole life. I could never succeed in any accomplishment without their love, hugs, smiles, and patience.

Ayça ÖLMEZ

PARTIAL LEAST SQUARES METHOD FOR THE ANALYSIS OF GENE EXPRESSION DATA

ABSTRACT

Partial Least Squares Regression (PLSR) is an unsupervised machine learning technique to modeling associations between variables through orthogonal latent variables. Using these latent variables, PLSR can make an inference from huge and computationally complex datasets that have missing values, noise and a numerous number of variables relativity more than the number of observations. The classical and standard algorithm of the PLSR is the Nonlinear Iterative Partial Least Squares Regression (NIPALS). The NIPALS is proposed for regression, classification and dimension reduction. The NIPALS and other PLSR algorithms have been used frequently for various bioinformatic studies.

In high-throughput gene expression data research, one of the important goals is to investigate gene-gene or their products interactions. To measure the level of association between these genes or their products, a highly recommended method can be used which is calculated by the variable weights and loadings based on PLSR, called Connectivity Scores.

In this thesis, PLSR was used for computing connectivity scores to construct gene networks for three brain region of a developing mouse brain in the embryonic period. Statistical analysis is performed using R statistical language and Cytoscape software is used to visualize gene networks.

Keywords: Gene networks, high throughput gene expression data, machine learning, partial least squares regression

GEN EKSPRESYON VERİLERİNİN ANALİZİ İÇİN KISMİ EN KÜÇÜK KARELER YÖNTEMİ

ÖZ

Kısmi En Küçük Kareler Regresyonu (KEKKR) değişkenler arasındaki ilişkileri ortogonal gizli değişkenler aracılığıyla modellemek için kullanılan denetimsiz bir makine öğrenme tekniğidir. KEKKR bu gizli değişkenleri kullanarak, kayıp gözlemlere, gürültüye ve değişken sayısına oranla çok sayıda gözlem değerine sahip olan büyük ve karmaşık veri kümelerinden bir çıkarım yapabilir. KEKKR'nin klasik ve standart algoritması Doğrusal Olmayan Iteratif Kısmi En Küçük Kareler Regresyonu'dur (NIPALS). NIPALS, regresyon, sınıflandırma ve boyut küçültme için önerilmiştir. NIPALS ve diğer KEKKR algoritmaları çeşitli biyoinformatik çalışmaları için sıklıkla kullanılmaktadır.

Yüksek çıktılı gen ifadesi veri araştırmalarında, önemli hedeflerden biri gen-gen veya gen ürünlerinin etkileşimlerini araştırmaktır. Bu genler veya gen ürünleri arasındaki ilişki seviyesini ölçmek için oldukça tavsiye edilen bir yöntem olan ve KEKKR metoduna göre hesaplanan değişken ağırlıklari ve yükleri ile bulunan Bağlantı Skorlar kullanılabilir.

Bu tezde, embriyonik dönemde gelişen fare beynine ait üç beyin bölgesinin gen ağları oluşturmak için bağlantı puanları KEKKR kullanılarak hesaplanmıştır. İstatistiksel analizler için R istatistik dili ve gen ağlarını görselleştirmek için Cytoscape yazılımı kullanılmıştır.

Anahtar kelimeler: Gen ağları, yüksek çıktılı gen ifade verileri, makine öğrenmesi, kısmi en küçük kareler regresyonu

CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

Machine learning, a convenient approach including algorithms and methods, is used for understanding and learning statistical relationships concerning large data by constructing computer programs that improve with past experiences. Machine learning methods are divided into two groups as Supervised Learning and Unsupervised Learning methods. Supervised learning is based on creating models for estimating output by using one or more inputs. With this approach, the outputs are known, and the algorithm can be controlled with these known outputs. On the other hand, there are no known outputs in unsupervised learning. The purpose of the aforementioned method is to learn from the structure of the data.

One of the supervised machine learning techniques, namely the Partial Least Squares (PLS), is widely used for classification, dimension reduction or modelling relationships between two matrices, response variable(s) Y and predictor variables X by creating new components which is called latent variables. While building a model for collinear data, Partial Least Squares Regression (PLSR) can remove multicollinearity in both independent and dependent variables using these latent variables. In addition, PLSR can handle the classical regression problems such as noise, missing values and complex problems. Moreover, PLSR is recommended for creating a model in which the number of variables much bigger than the number of observations. It was developed by Wold (1966) for modelling economic paths. The first type of PLSR algorithm is used for calculation of principal components. Shortly afterwards, this algorithm was improved a version of calculation of canonical correlation. In the same year, Wold (1975) published the classical PLSR procedure Nonlinear Iterative Least Square (NILES) algorithm building a model based on the loadings and weights of the variables and later changed the name as Nonlinear Iterative Partial Least Square (NIPALS) algorithm.

PLSR is also a very popular in other scientific areas. The first application of PLSR was in chemometrics, and it is still very popular in the field. In many chemometric researches, collinearity problem among the variables causes miscalculation of regression coefficients. Wold et al. (1984), applied PLSR and some additional methods including Principal Component Regression (PCR), James Stein shrunk estimate (JS), Ridge Regression (RR) and Total Least Squares (TLS) to evaluate a chemical data and PLSR has shown to be a proper method. Geladi & Kowalski (1986) showed the weaknesses of Multiple Linear Regression (MLR) and Principal Component Regression (PCR) and claimed that PLSR is a better alternative. Höskuldsson (1988) investigated statistical and mathematical structure of PLSR. The superiority of the PLSR over other regression methods is explained by the fact that it chooses the minimum number of variables with the maximal reduction. Helland (1988) demonstrated the mechanism of PLSR and emphasized that PLSR obtained smaller mean square error using fewer components than the PCR by a simulation study. Helland (1990) showed that PLSR algorithms have some advantages, such as usages of projections and geometric intuition. De Jong (1993) introduced an algorithm, SIMPLS, where PLS factors are determined to maximize covariance criteria between X and Y latent variables and obey certain orthogonality and normalization restrictions. Wold et al. (2001) used PLSR algorithm on Quantitative Structure – Activity Relationship (QSAR) and Quantitative Structure – Property Relationships (QSPR) modelling. Also, they reported PLSR results better than MLR. Besides these, there are many PLSR algorithms in the field of chemometrics in the literature. (See e.g. Hubert & Branden (2003), Branden & Hubert (2004), Serneels et al. (2005), Alin & Agostinelli (2017), among others.)

Recent years, it has also been a highly preferred method in bioinformatics for analysis of high-throughput gene expression data. High-throughput gene expression profiling has become an important research area with advent of low-cost microarray technology. This profiling is used to discover the relationships or associations between gene expression levels and it helps to understand the mechanism of biological systems such as budding yeast or common diseases. For this profiling process, researchers generally use network-based approaches to investigate, represent and analyze any biological situation and the causes, biomarkers, alternative therapy strategies, pathways, pathogenesis and even gene associations of various disease. There are many biological network types such as protein-protein interaction, cell signaling, metabolic, gene regulatory, gene co expression etc. Networks of DNA, neurons, RNA-sequence, proteins and LinkRNA data, provide a better understanding each kind of relationship between and within species, diseases and genetic mechanisms. Datta (2001) proposed PLSR for gene co expression network construction and shown that classification into temporal groups using expression levels during sporulation of budding yeast. The magnitudes of the regression coefficients indicate which genes are important in each sporulation phases. Also, coexpression analysis based on PLSR has been used in other budding yeast researches. Johansson et al. (2003) showed that cell cycle behavior with cyclically expressed genes of budding yeast. Carter et al. (2004) provided a new perspective to molecular structure of cellular state. They showed mechanisms of different cellular states and selective genetic elements of yeast and human medulloblastoma networks. After that, Boulesteix & Strimmer (2005) used PLSR to estimate and understand the complex regulatory mechanisms in cells. Therefore, they build transcription factor activity (TFA) networks for budding yeast datasets. Bras & Menezes (2006) applied PLS-based methods which have stronger dimension reduction performance to predict missing values using a pattern of original dataset that is non-time series cell cycle regulated genes in yeast. Pihur et al. (2008) compared reconstruction of genetic networks based on PLS and other alternative network inference methods on yeast DNA-damage data. Yeast DNA genetic associations network illustrated by Cytoscape program. Moreover, PLS-based network exposed seventeen out of 118 associations that was the highest matched the existing associations. Tenenhaus et al. (2010) presented an investigation of sensitivity and specificity of PLS and other 4 methods on budding yeast microarray data. According to the results, PLS has provided worthful information about gene associations. Lately, Mehmood et al. (2011b) and Mehmood et al. (2011a) suggested PLS-based criteria for mapping relationships of genotype-phenotype on budding yeast. The derived results were consistent with known yeast phylogeny.

PLSR is also one of the most popular methods for disease research studies. Alaiya et al. (2000) used PLS and PCA to classify three types (benign, borderline and malignant) of ovarian tumors. Using the PLS loadings, the most related variables for distribution of three tumor types were determined. Analysis of a learning set selected randomly from original data revealed 18 cancer types 11 of which was correct classification. Nguyen & Rocke (2002) used PLS and PC for dimension reduction on five different type of cancer to distinguish between normal and cancer tissue, and types of tumors. They showed PLS components were superior to PC components and had better true classification rate. Pérez-Enciso & Tenenhaus (2003) categorized before and after chemotherapy treatment in case control study, estrogen receptor positive and negative tumors and tumor classification on breast cancer data. Also they showed the satisfactory performance of PLS for classification problem. Huang & Pan (2003) proposed a penalized version of PLSR, that removed genes which have low prediction power. In his study, a PPLSR model was used to estimate the support time of LVAD (left mechanical ventricular assist device), which is a substitution therapy for heart failure patients, using gene expression levels. Compared with Random Forest, PLS-based method displayed better results. Boulesteix (2004) examined gene classification by PLS dimension reduction for gene selection for tumor diagnosis on nine different cancer types. Musumarra et al. (2004) studied on the genes, associated with cancer development via PLS. They suggested new diagnostic tools for colon cancer. Huang et al. (2005) applied 5 different statistical methods including PLS to determine prognosis, find alternative therapy and differentiate between ischemic and non-ischemic heart failure. Boulesteix & Strimmer (2006) examined bioinformatics applications of PLS analysis for high dimensional data. They focused on several advantages of PSL such as fast, efficient and availability for classification, survival analysis and transcription factors activities analysis. Abdi (2010) explained PLSR models for especially brain imaging datasets (because of the multidimensional structure of brain imaging data) and applied bootstrap and jackknife methods to measure the quality of predictions. Land Jr et al. (2011) claimed PLS can be used to discover biomarkers of colon cancer. In other studies that used different network methods on the same colon cancer microarray data, many important prognostic indicator genes were common. Huang et al. (2013) investigated multi-classification toxicology to show some prognostic significance of breast cancer. PLSR was proved to be an effective and practicable alternative to categorize two or more classes. Wang et al. (2014) introduced a new interpretation for the mechanism of epilepsy and potential targets for new alternative treatments with PLS regression analysis. Most recently, Ding (2014) used PLS to identify prognostic genes in end-stage renal failure patients. The PLS latent variables network was constructed to show key molecules. The results of this study gave the molecular mechanism of underlying renal failure.

There are some studies on using PLSR for an analysis of brain data. Dreissig et al. (2009) presented an application of PLSR to analysis lipid brain tissue for diagnostics of tumor type and grade. Ramírez et al. (2010) proposed an early diagnosis system for Alzheimer's disease by using PLSR classification of the regional cerebral blood flow with single means of photon emission computerized tomography (SPECT). They compared PLSR and Random Forest methods and found that PLSR had higher sensitivity, specificity and accuracy rate. Faria et al. (2011) investigated brain tumor classes according their biochemical changes and patterns. The metabolic patterns detected by PLS based method and showed PLS was an efficient method to chemical classification of brain tumors. Shokri-Kojori et al. (2017) used functional connectivity density (FCD) mapping to expose alcohol and resting brain activity relationships. The PLSR comparison between heavy drinkers and control group showed that, heavy drinkers had higher FCD in cerebellum and control group had higher FCD in visual and prefrontal cortices and thalamus. Biomarkers of transitioning from light to heavy drinking were presented.

In this thesis, we will build a network for mouse brain regions, fore-, mid- and hindbrain, to better understanding the functions of developing mouse brain in embryotic process. This thesis contains four chapters. In chapter two, some biological background is given, a brief information about biological network is mentioned and focused on PLSR methodology. In chapter three, the investigated mouse brain data sets, the brain properties, and functions are discussed. Furthermore, the results of the classification of the data, PLSR analysis, and the gene networks are given. Finally, some future works and conclusions are given in chapter four.

CHAPTER TWO
GENE NETWORKS BY PARTIAL LEAST SQUARE REGRESSION

2.1 Biological Background

In living organisms and many viruses, the basic biological units of heredity are DNA, RNA, proteins, and metabolites. DNA, also known as deoxyribonucleic acid, is a molecule that carries the genetic information of biological compounds such as RNA and proteins during reproduction. This information determines the function of genes thereby the function of cells. Each cell contains chromosomes that composed of many genes. A single strand of DNA has millions of nucleotides which labeled as adenine (A), thymine (T), cytosine (C) and guanine (G). The nucleotides in one strand bound the other nucleotides in another strand, according to base pairing rules (A with T and C with G). In this way, the structure of DNA becomes a double helix. Figure 2.1 shows that, the first photograph of DNA's double helix structure in electron microscope by taken physicist Prof. Enzo Di Fabrizio (MacKinnon, 2012). The red arrows show the strands.

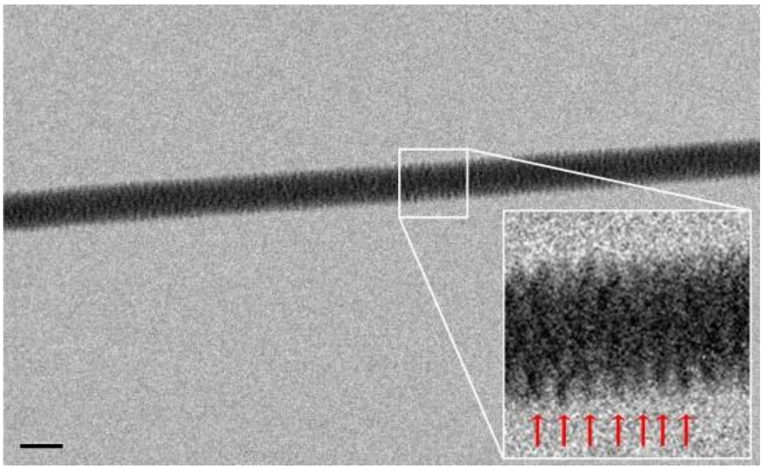

Figure 2.1 The first photograph of DNA (MacKinnon, 2012)

A gene is a special stretch of the double helix. Genes determine the characteristics of a whole and of a function of the cells. All these genes are ingredients a part of a tissue. These tissues make up an organ which is a part of an organism. Figure 2.2 shows the gene and chromosome (NLM, 2018).

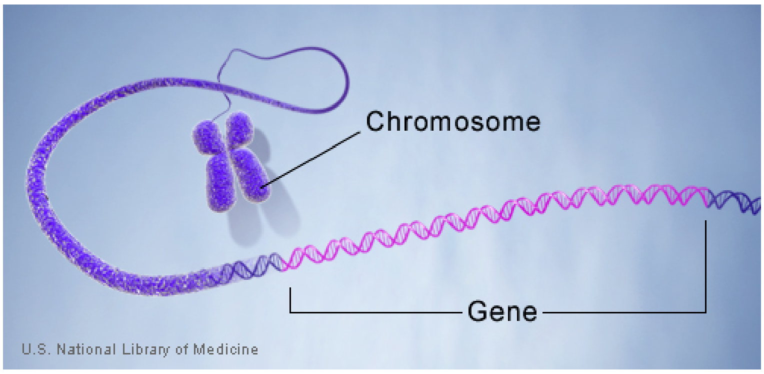

Figure 2.2 Gene structure. By U.S. National Library of Medicine (NLM, 2018)

2.1.1 Biological Networks

In organisms, interested cases, complex processes or genetic abnormalities caused by mutation, evolution, etc., can be found by investigation of genetic relationships. Network approaches have been used to explore these observed associations or relationships. Network methods are developing day by day and their importance and popularity are increasing. Biological networks provide a mathematical representation of connections found in ecological, evolutionary and physiological research. A biological network basically consists of nodes and edges. Nodes represent metabolites, genes or their products, while edges represent the association between them. Edges can be direct, that means high betweenness or high degree, or indirect. These nodes are used to represent any organism, gene, protein or neuron.

Types of biological networks are protein-protein interaction networks, gene regulatory networks (DNA- protein interaction networks), metabolic networks, neural networks, signaling networks, ecological networks and gene co-expression networks (transcript-transcript association networks). In this thesis, analysis of mouse brain regions is made by gene co-expression network.

Along with the development of high throughput genomics technologies including microarray and RNA sequencing, genome-wide expression analysis has become a major study area. Gene co-expression networks (GCN) describes the associations or associations between high throughput expression patterns of genes or their products such as DNA, RNA and long non-coding RNA (lincRNA). These associations show the similarities or differences between gene expression patterns from same or different biological conditions.

Different statistical methods have been used to measure the level of similarities or differences including Pearson's and Spearman' s correlations, Bayesian and ARACNE approach. Alternatively, partial least squares regression can be used.

2.2 Partial Least Squares Regression

High-throughput gene expression data are generally characterized by the number of observations (*n*) that are much smaller than the number of genes (*k*). Under this circumstance, the probability of occurrence of multicollinearity problem increases. Multicollinearity problem causes the estimators to have a large amount of variance and the estimated values may be much different than their true values. Therefore, the traditional statistical methodology for such data cannot be used directly. In that case, dimension reduction or latent variable methods such as PC and PLS regression can be applicable.

PLS regression is originally proposed for modelling the relation between response and explanatory variables. It is also proposed for building gene co-expression networks by Datta (2001). PLSR is a latent variable based method where orthogonal latent variables are obtained as to maximize covariance between them.

The goal of PLSR is to predict matrix of response variables, $\boldsymbol{Y}$ is a (nxm) using matrix of explanatory variables, $\boldsymbol{X}$ is a (nxk) through the linear relation given in (2.1). Here, n is number of observations, m is number of $\boldsymbol{Y}$ variables, k is number of $\boldsymbol{X}$ variables. PLSR decomposes $\boldsymbol{X}$ and $\boldsymbol{Y}$ matrices according to the linear models in (2.2) and (2.3). $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)$ is *iid* normally distributed error with $E(\boldsymbol{\varepsilon}_i) = 0$ and $Var(\boldsymbol{\varepsilon}_i) = \sigma^2$. The rest of the this thesis $(')$ will be used to represent the matrix transpose.

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon \tag{2.1}$$

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}' + \boldsymbol{E} \tag{2.2}$$

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{C}' + \boldsymbol{G} \tag{2.3}$$

$\boldsymbol{T}$ (nxA) is the score matrix of $\boldsymbol{X}$, and $\boldsymbol{P}$ (kxA) is the loading matrix of $\boldsymbol{X}$ and while, $\boldsymbol{U}$ (nxA) is a score matrix of $\boldsymbol{Y}$, and $\boldsymbol{C}$ (mxA) is the loading matrix of $\boldsymbol{Y}$. $\boldsymbol{E}$ (nxk) and $\boldsymbol{F}$ (nxm) are residual matrices for $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. $\boldsymbol{A}$ defines the number of latent variables.

For gene network construction PLSR is used to obtain connectivity scores which will be introduced in the next section. In this context, each gene is modeled on the rest of the genes. So, our response will be a vector of genes, let's say x_i, and $\boldsymbol{X}_{(i)}$ matrix contains the remaining genes. Orthogonal score matrix $\boldsymbol{T}_{(i)} = (\boldsymbol{t}^{1}_{(i)}, \boldsymbol{t}^{2}_{(i)}, ..., \boldsymbol{t}^{A}_{(i)})$ is obtained as in (2.4).

$$\boldsymbol{t}^{a}_{(i)} = \sum_{k} \boldsymbol{w}^{a}_{(i)j}\boldsymbol{x}_{(i)j} \qquad (\boldsymbol{T}_{(i)} = \boldsymbol{X}_{(i)}\boldsymbol{W}_{(i)}) \tag{2.4}$$

$\boldsymbol{W}_{(i)}$ with size of $((k-1)xA)$ is the $\boldsymbol{X}$ weight matrix, when i^{th} gene is taken as response. $\boldsymbol{w}^{a}_{(i)j}$ represents the contribution of i^{th} variable on j^{th} variable at the a^{th} component.$(i = 1, 2, ..., m$ and $j = 1, 2, ..., k)$ $\boldsymbol{x}_{(i)j}$ describes j^{th} variable and $\boldsymbol{t}^{a}_{(i)} = (\boldsymbol{t}^{a}_{(i)1}, \boldsymbol{t}^{a}_{(i)2}, ..., \boldsymbol{t}^{a}_{(i)n})^{\boldsymbol{T}}$ is the score vector for a^{th} component when x_i is our response variable.

Using these scores, the j^{th} explanatory variable an response variable x_i can be rewritten as in (2.5) and (2.6).

$$\boldsymbol{x}_{(i)j} = \sum_{a} \boldsymbol{t}^{a}_{(i)}\boldsymbol{p}^{a}_{(i)j} + \boldsymbol{e}_{(i)j} \tag{2.5}$$

$$\boldsymbol{x_i} = \sum_{\boldsymbol{a}} \boldsymbol{c^a_{(i)}} \boldsymbol{t^a_{(i)}} + \boldsymbol{f_{(i)}} \tag{2.6}$$

$\boldsymbol{c^a_{(i)}}$ is the loading of the response x_i on a^{th} component, and $\boldsymbol{p^a_{(i)j}}$ is the loading of $\boldsymbol{x_{(i)j}}$ on a^{th} component.

Coefficients for the model of x_i on the rest of $(k-1)$ x variables are then calculated by the formula in (2.7).

$$\boldsymbol{b_{ik}} = \sum_{\boldsymbol{a}} \boldsymbol{c_{ia} w_{ka}} \qquad , j \neq i \tag{2.7}$$

PLSR is an iterative method. After all computations, $\boldsymbol{X_{(i)}}$ matrix is deflated as in (2.8) and all calculations are repeated with deflated $\boldsymbol{X^a_{(i)}}$. Iterations last until all of A components are obtained. By deflating $\boldsymbol{X_{(i)}}$, we guarantee orthogonal component scores.

$$\boldsymbol{X^a_{(i)}} = \boldsymbol{X^{(a-1)}_{(i)}} - \boldsymbol{t^a_{(i)} p^a_j} \tag{2.8}$$

These components have a new vector space which is a subspace of the original data. PLSR generates these orthogonal latent variables that have maximum covariance. When A equals k, PLSR and MLR give same results.

The process of generating new latent variables at each iteration, depends on the algorithm that used. Several algorithms have been proposed for PLSR in the literature, starting with Nonlinear Iterative Partial Least Square (NIPALS) which is used in this thesis. The historical evolution of some of the well-known algorithms is given below (Kondylis (2006)).

- NIPALS algorithm (Wold (1966) and Wold (1975))
- SIMPLS algorithm (De Jong (1993))

- SAMPLS algorithm (Bush & Nachbar (1993))
- KERNEL algorithm (Rosipal & Trejo (2001))
- PPLSR algorithm (Bastien et al. (2005))

2.2.1 NIPALS Algorithm

The classical and standard algorithm for computing PLSR components is NIPALS. It has been used for PCA first and later used for PLSR. It effectively tolerates the missing values and uses predictions from the model for handling these the missing values. These predictions have zero residuals and therefore there is no influence on the scores and loadings.

NIPALS algorithm, which works with centralized and scaled $\boldsymbol{X}$ and $\boldsymbol{Y}$ variables, consists of the following steps (Wold et al. (2001)). Here, we assume only one response. For the sake of simplicity we refer our response as y. But for network construction y will be the i^{th} gene taken as a response variable.

Step 1: A single $\boldsymbol{Y}$ variable column is assigned to $\boldsymbol{u}$, $\boldsymbol{u} = \boldsymbol{y}$.

Step 2: The $\boldsymbol{w}$ weight vector is obtained with $\boldsymbol{u}$, $\boldsymbol{w} = \boldsymbol{X'u}/\boldsymbol{u'u}$.

Step 3: $\boldsymbol{w}$ is scaled to have unit length, divided by the euclidean norm $\boldsymbol{w}$, $\boldsymbol{w} = \boldsymbol{w}/||\boldsymbol{w}||$.

Step 4: The $\boldsymbol{t}$ vector, $\boldsymbol{X}$ scores, is calculated, $\boldsymbol{t} = \boldsymbol{Xw}$.

Step 5: The $\boldsymbol{c}$ loading vector, for $\boldsymbol{y}$, is calculated, $\boldsymbol{c} = \boldsymbol{yy'}/\boldsymbol{t't}$.

Step 6: $\boldsymbol{c}$ is scaled to have unit length, divided by the euclidean norm $\boldsymbol{c}$, $\boldsymbol{c} = \boldsymbol{c}/||\boldsymbol{c}||$.

Step 7: The $\boldsymbol{u}$ vector is updated for next iteration, $\boldsymbol{u} = \boldsymbol{yc}/\boldsymbol{c'c}$.

Step 8: The loadings of $\boldsymbol{X}$, $\boldsymbol{p}$ vector, is computed, $\boldsymbol{p} = \boldsymbol{X't}/\boldsymbol{t't}$.

Step 9: The deflation of $\boldsymbol{X}$ variables, $\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{tp'}$.

Step 10: The deflation of $\boldsymbol{Y}$ variables, $\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{t}\boldsymbol{c}'$.

After the first component is obtained, both $\boldsymbol{X}$ and $\boldsymbol{Y}$ matrixes can be reduced in order to calculate the parameters for next component. The deflation in Step 8 is optional, the results are same whether $\boldsymbol{Y}$ deflated or not. These steps are repeated as many as the determined number of components, A.

There is no certain method for choosing the proper number of component, which is one of the most important decision, in PLSR procedure. Since PCA is known with a strong classification accuracy in many situations, in this thesis PCA is used as a classification method on mouse brain data to choose the number of components.

Finally, after determining the optimal number of components, PLSR parameters are obtained from the NIPALS process and connectivity scores are calculated to show the relation between each pair of cases, and to build a network.

2.2.2 Connectivity Scores

Connectivity scores is a useful tool for exploring associations in gene network structure (Pihur et al. (2008)). If there is an edge between two nodes (i^{th} and j^{th} genes), this edge is formed by statistically significant connectivity score of i^{th} and j^{th} genes. This connectivity score is obtained by association score between i^{th} and j^{th} genes and the symmetric association score between j^{th} and i^{th} genes, in presence of other genes.

$$\hat{\boldsymbol{s}}_{\boldsymbol{ij}} = \frac{\sum_{\boldsymbol{a}=1}^{\boldsymbol{A}} \boldsymbol{c}^{\boldsymbol{a}}_{(\boldsymbol{i})} \boldsymbol{w}^{\boldsymbol{a}}_{(\boldsymbol{i})\boldsymbol{j}} + \sum_{\boldsymbol{a}=1}^{\boldsymbol{A}} \boldsymbol{c}^{\boldsymbol{a}}_{(\boldsymbol{j})} \boldsymbol{w}^{\boldsymbol{a}}_{(\boldsymbol{j})\boldsymbol{i}}}{\boldsymbol{2}} \tag{2.9}$$

Equation 2.9 shows the computation of connectivity score for i^{th} and j^{th} genes, $\hat{\boldsymbol{s}}_{\boldsymbol{ij}}$. In $\sum_{\boldsymbol{a}=1}^{\boldsymbol{A}} \boldsymbol{c}^{\boldsymbol{a}}_{(\boldsymbol{i})} \boldsymbol{w}^{\boldsymbol{a}}_{(\boldsymbol{i})\boldsymbol{j}}$, gene i is the response variable. $\boldsymbol{c}^{\boldsymbol{a}}_{(\boldsymbol{i})}$ is the loading of the i^{th} gene on the a^{th} component and $\boldsymbol{w}^{\boldsymbol{a}}_{(\boldsymbol{i})\boldsymbol{j}}$ is the contribution of j^{th} gene on the a^{th} component when the i^{th} gene is modeled by other genes. In $\sum_{a=1}^{A} \boldsymbol{c}^{a}_{(j)} \boldsymbol{w}^{a}_{(j)i}$, the symmetric calculation, gene k is the response variable. $\boldsymbol{c}^{a}_{(j)}$ is the loading space j^{th} gene on the a^{th} component and $\boldsymbol{w}^{a}_{(j)i}$ is the contribution of the i^{th} gene on the a^{th} component when the j^{th} gene is modeled by other genes. Once the connectivity scores are calculated for each gene pair, the gene network can be constructed the significant scores.

To decide if a connectivity score is significant, all scores are normalized from -1 to 1. The following equation(2.10) is used for all i^{th} and j^{th} gene pairs normalization calculations.

$$\boldsymbol{\hat{s}^{new}_{ij}} = \frac{\boldsymbol{2(\hat{s}_{ij} - min(\hat{s}_{ij}))}}{\boldsymbol{(max(\hat{s}_{ij}) - min(\hat{s}_{ij})) - 1}} \tag{2.10}$$

Then, the significance of the normalized connectivity score is determined using some threshold values, $\boldsymbol{\epsilon}$. The modular structure and the sensitivity of the gene network changes with choice of $\boldsymbol{\epsilon}$ which can be $\boldsymbol{\epsilon} \in \{0.35, 0.4, 0.45, 0.5, 0.55\}$ (Gill et al. (2010)).

CHAPTER THREE
NUMERICAL RESULTS

In this chapter, the PLSR in gene networks has been examined in the developing mouse brain. A brief information about brain and their functions,mouse brain and their development have been given. The variables and observations of the investigated mouse brain datasets have been mentioned. The classifications and filtering according to the biological properties have been explained. Filtered data PLSR analysis and the gene networks that built by using the predicted parameters have been performed. R 3.3.1 is used for statistical analysis and Cytoscape 3.6.1 is used for visualizing gene networks.

3.1 Data Sets

Brain, which is a mass of nerve tissue, is the most complex organ that controls all functions of the body. The mammalian brain has three major regions; fore-, mid- and hindbrain. The developmentally oldest portion of the brain is the reptilian brain, which is also known as hindbrain (rhombencephalon). The hindbrain consists of the pons, medulla oblongata and cerebellum. It coordinates vital functions such as heartrate, balance, breathing, reflex actions and body temperature. The second one is the midbrain (mesencephalon) that connects the forebrain to the hindbrain. The midbrain formed by cerebral peduncles, corpora quadrigeminal and cerebral aqueduct. It associated with hearing, eye movement, emotions, learning and memory. The rostral-most region of the brain is the forebrain (prosencephalon). It is the largest brain division and includes cerebrum, thalamus and hypothalamus. The forebrain provides imagination, language, creativity and logic.

In Figure 3.1, the mouse brain parts are shown; the purple part is the hindbrain, the red part is the midbrain and the beige part is the forebrain. In this thesis, brain development in the embryonic period is observed using gene expression levels from the three parts of the mouse prenatal brain. In each brain parts, genes belonging to different regions datasets are modeled by PLSR, individually. And gene networks are created for each brain part.

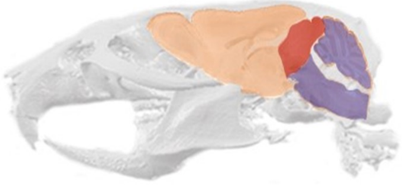

Figure 3.1 The mouse brain parts

The pregnancy process of mouse is approximately 21 days. The rapid brain growth and development of these regions continues until the tenth day of the postnatal period. The pregnancy can be examined in three-time section as the first seven days, the second seven days and the third seven days. In the first seven days, brain tissue and central nervous system occur. Somites, which originate the vertebral, the spinal muscle and the spinal dermis, form on the eighth day. On day nine, gray matter ingenerates in the brainstem, hindbrain. Spinal cord, Purkinje cells, and medulla oblongata begin to form in the tenth day. In the eleventh day, the brain development is accelerated and cerebral cortex, globus pallidus which regulate voluntary movements, and thalamus development seen. On the twelfth day, the formation of the brain partitions such as the amygdala, optic axons, mitral cells and cochlear nucleoids is observed. On the thirteenth day, thalamus, hypothalamus and optic axons continue to develop, and pons cells that control motor activities begin to develop. White matter and the development of retinal cells (cones, and amacrine cells) begin to appear on the fourteenth day. Hypothalamus and retinal cell development continue in the fifteenth and sixteenth day. On the eighteenth day, the development of the cerebral cortex continues. The formation of another retinal cell, rods occurs on the nineteenth day (Çoban (2014)). Table 3.1 shows the daily process of brain development (Finlay & Darlington (1995)).

Table 3.1 Mouse brain development timeline

Day	Active Parts
9	Hindbrain
10	Hindbrain
10.5	Forebrain
11	Fore-Mid-Hindbrain
11.5	Forebrain
12	Fore-Mid-Hindbrain
12.5	Fore-Midbrain
13	Fore-Midbrain
13.5	Fore-Mid-Hindbrain
14	Forebrain
15	Fore-Midbrain
16	Fore-Midbrain
17	Forebrain
19	Forebrain

We have three separate data set; forebrain, midbrain and hindbrain parts data. This data sets are taken from the Sequence Read Archive, which is an open source for high-throughput gene datasets. All data sets consist of 47,415 gene, 121,400 gene transcript, and 26 samples, except the forebrain data with 25 samples. There are four gene expression measurements for each tenth (E10.5), eleventh (E11.5) (three measurements in forebrain data), twelfth (E12.5), fourteenth (E14.5) prenatal days, two gene expression measurements for each thirteenth (E13.5), fifteenth (E15.5) and sixteenth (E16.5) prenatal days, and four gene expression measurements for postnatal day (P0).

3.2 Results

3.2.1 Clustering

In high-throughput expression data and biological network studies, clustering is a preprocess which allows to see the most important features of a cluster. PCA is a widely used unsupervised statistical method which aims to divide a data into clusters or classes and build new significant variables by using dimension reduction. PCA converts a set of data of possibly multicollinear variables into a set of values of uncorrelated variables called principal components. In this thesis, PCA is used to choose proper number of components before building the networks with PLSR. In addition, hierarchical clustering, another classification method that supports PCA results, is used.

All mouse datasets collected at 8 different time points. For forebrain data, 8 of 25 samples selected using PCA based on the structure of the data. The first principal component is the most correlated with 24 of the original variables, that has maximum variance explained ratio with 94.94%, shown in Table 3.2.

Table 3.2 Cumulative proportion of variance explained for forebrain

Number of Componets	Cumulative Explained Variances
1	0.9494
2	0.9835
3	0.9876
4	0.9904
5	0.9921
6	0.9931
7	0.9938
8	0.9944
9	0.9949
10	0.9954

Alternatively, the scree plot of proportion of variance explained (Figure 3.2) shows that cumulative variance of components from smallest variance to largest. First two components explain 98.35% of the variation in the data. Despite three components have an adequate amount of variation explained it is more convenient to work with 8 components for this data set. The 8^{th} component has 99.44% variance explained ratio.

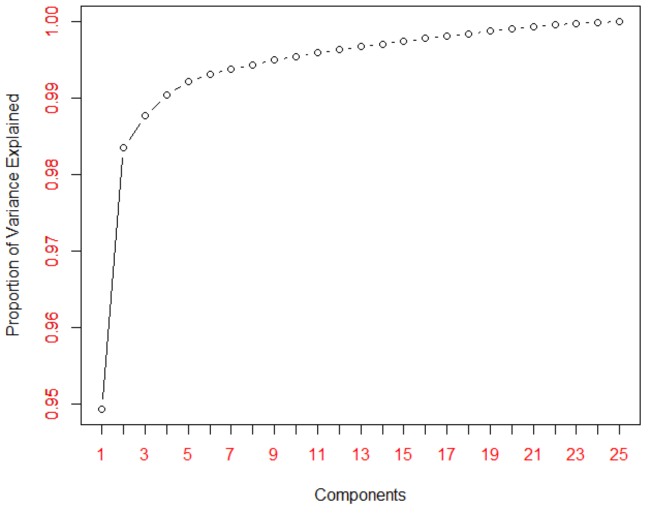

Figure 3.2 Proportion of variance explained plot for the forebrain data

The plot of classification for the first and the second component, Figure 3.3 shows that a projection of all samples. Here, the 25 samples divided into 8 classes, pointed out with red rectangles, compatible with the time points in our data. Each class represented by eight different color and shape.

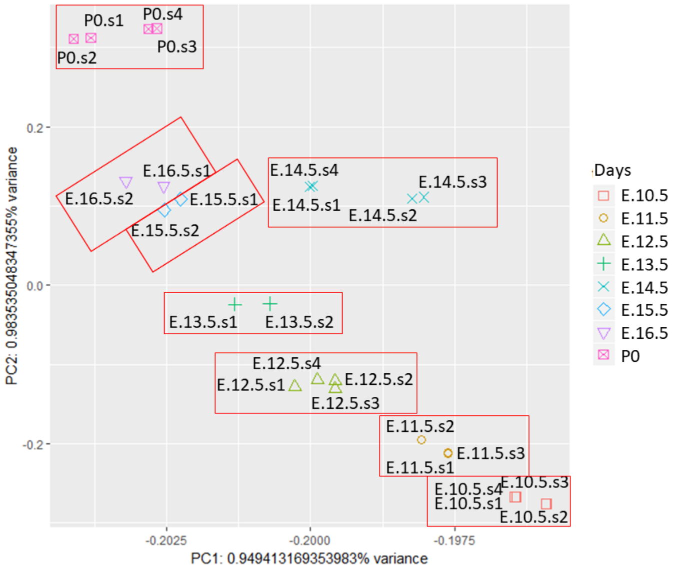

Figure 3.3 Classification of PC1 and PC2 for days on the forebrain data

In hierarchical clustering, another way to represent the data, a dendrogram (a cluster tree) that is the plot for getting meaningful classification projections. Figure 3.4 shows the 8 classes, shown by different colors, for 25 time points that are the same as PCA classification.

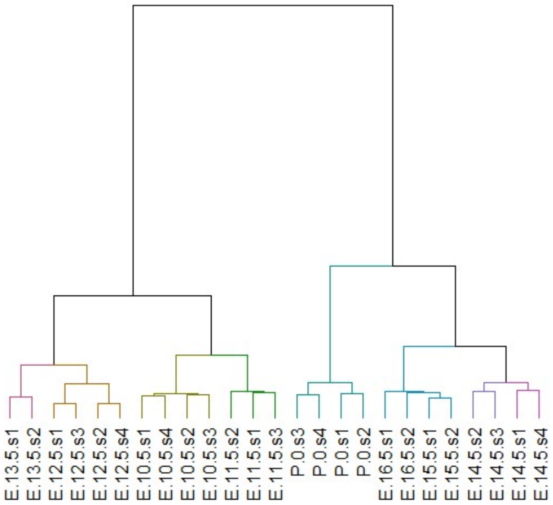

Figure 3.4 Hierarchical classification for the forebrain data

For the second data set, the midbrain data have 8 classes for 26 samples selected using PCA based on the structure of the data. The first principal component explains the 94.96% of the variance (Table 3.3).

Table 3.3 Cumulative proportion of variance explained for midbrain

Number of Componets	Cumulative Explained Variances
1	0.9496
2	0.9848
3	0.9882
4	0.9906
5	0.9921
6	0.9932
7	0.9939
8	0.9945
9	0.9950
10	0.9955

Figure 3.5 illustrates the cumulative proportion of variances explained by components. Although first two components explain 98.48% of the variation in the data, it is better to use 8 components for this data set. The 8th component has 99.45% variance explained ratio.

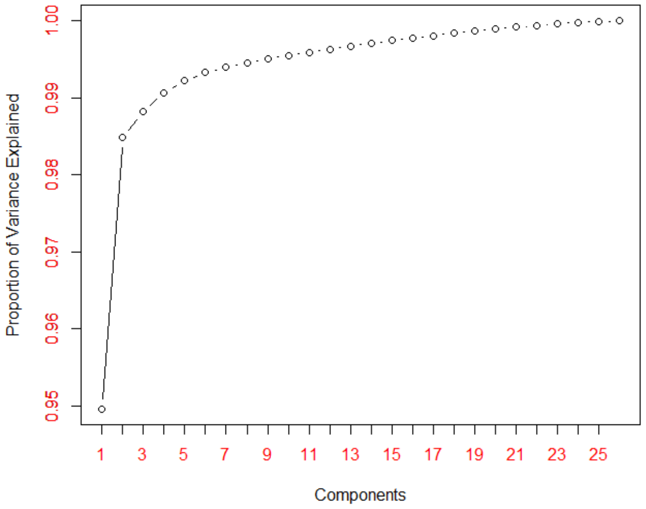

Figure 3.5 Proportion of variance explained plot for the mibrain data

The visualization of subsets of the first and the second components, Figure 3.6 shows that the 26 samples divided into 8 classes.

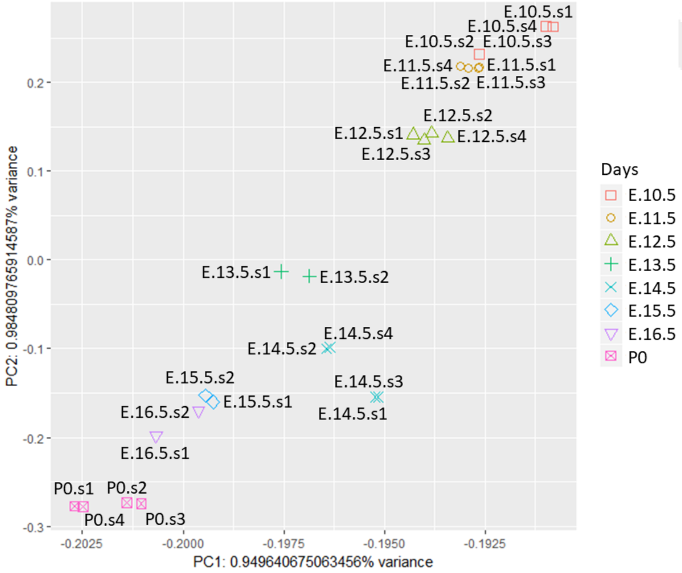

Figure 3.6 Classification of PC1 and PC2 for days on the midbrain data

The hierarchical method shows the 8 classes which fundamentally carries the same results with PCA clusters. (Figure 3.7)

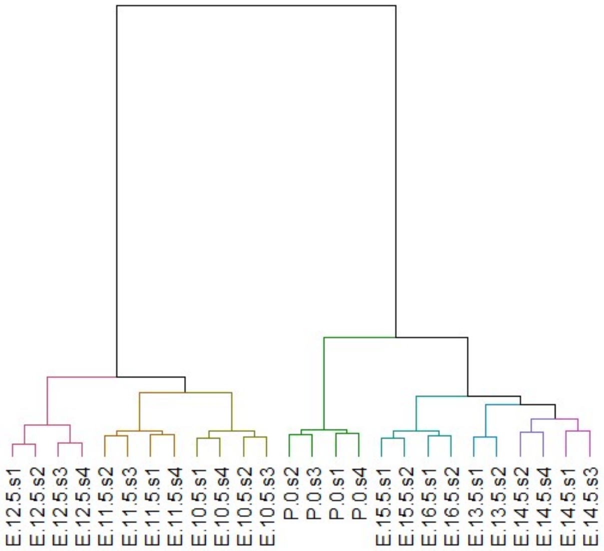

Figure 3.7 Hierarchical classification for the midbrain data

In the hindbrain data, which is similar to the midbrain data, it has been shown that 8 days in 26 samples according to PCA results. The variance explained ratio of the first component is 95.46%. The cumulative variance explained ratios of all variables are given in Table 3.4.

Table 3.4 Cumulative proportion of variance explained for hindbrain

Number of Componets	Cumulative Explained Variances
1	0.9546
2	0.9851
3	0.9884
4	0.9904
5	0.9922
6	0.9931
7	0.9938
8	0.9944
9	0.9949
10	0.9954

Figure 3.8 supports previous table and shows the cumulative variances of components. As it is expected, first two components explain 98.51% which is most of the variation in the data. The 8^{th} component has 99.44% variance explained ratio.

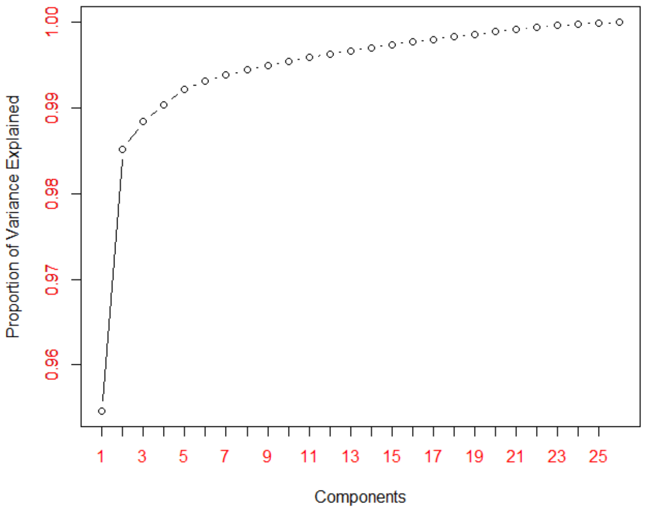

Figure 3.8 Proportion of variance explained plot for the hindrain data

The scree plot of the first and the second component scores, Figure 3.9 displays that the 26 samples divided into 8 classes.

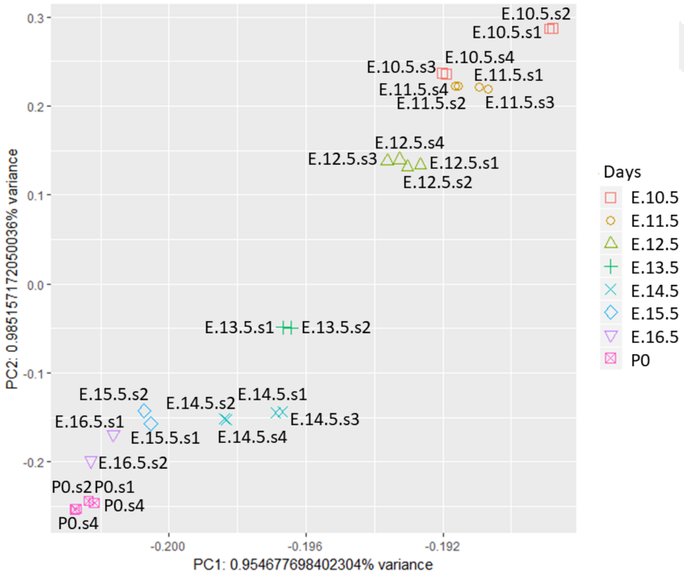

Figure 3.9 Classification of PC1 and PC2 for days on the hindbrain data

The dendrogram of hindbrain data illustrates the 8 classes by different colors in Figure 3.10.

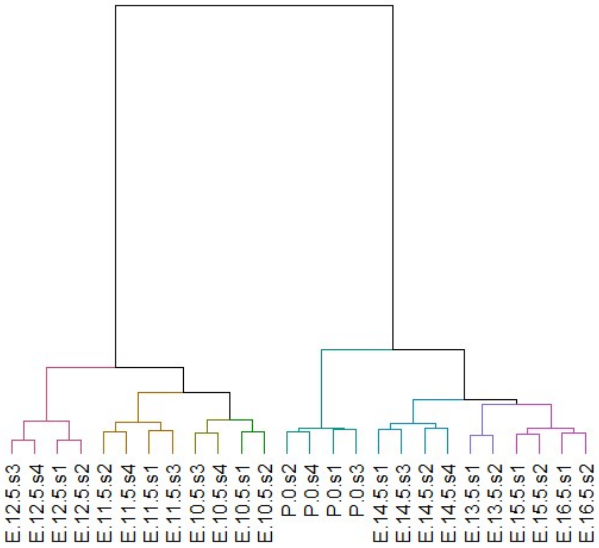

Figure 3.10 Hierarchical classification for the hindbrain data.

3.2.2 Connectivity Scores with Partial Least Square Regression

Before performing analysis by PLSR, a few filtering steps have been used to show possible biologically significant genes. First, expression levels that are smaller than 1 based on the daily time points (transcripts per million - TPM), are eliminated from the data. TPM is a normalization method for RNA-seq expression, that provides a digital measure of the abundance of transcripts. TPM is necessary to remove technical biases in sequenced data. If there is a larger number of zero expression values than the half of the number of samples of a day, the gene is extracted because it is not biologically meaningful. In case the number of expression values which are greater than 1 TPM, is more than half of the number of samples for any day out of 8 days, the investigated gene should remain in the dataset. For example, the gene named “Tbx2” with the code “ENSMUSG00000000093” has the expression values that are zero on all days except day 10. This is enough to remain the interested gene in the dataset. Another example is the “Pbsn” gene with the code “ENSMUSG00000000003”, that has been removed since it has expression values that are zero for all samples at all days. In addition, Differential Expression Analysis was used as the second filter.

As a result of the filtering on the forebrain data, $12,746$ genes out of $47,415$, and $25,418$ gene transcripts out of $121,400$ was obtained. The midbrain data filtering gave $12,879$ genes out of $47,415$, and $25,554$ gene transcripts out of $121,400$. Similarly, the filtering on the hindbrain data results presented $12,618$ genes out of $47,415$, and $24,318$ gene transcripts out of $121,400$.

Because the analysis using whole datasets computationally expensive, the PLSR and the network analysis were performed with 10% of the data, which is randomly selected. This reduction led to 2288 genes and 2543 transcripts for the forebrain data, 2286 genes and 2555 transcripts for the midbrain data, and 2234 genes and 2432 transcripts for the hindbrain data.

$3,232,153$ connectivity score was calculated for 2543 transcripts in the reduced forebrain data set. Table 3.5 shows the forebrain data connectivity scores between first 25 transcript id and 3 transcript id, using PLSR. The connectivity score can be interpreted similarly to the correlation coefficient. For instance, the score for the same two transcript id are equal to 1, as seen in ”ENSMUST00000123809.1” and ”ENSMUST00000123809.1”. This means that, as in the correlation interpretation, there is the perfect relation between them. Besides this, there are different connectivity scores between different transcripts, such as -0.05508792 for ”ENSMUST00000123809.1” and ”ENSMUST00000046950.12”. There is a poor relation between these two transcripts.

Table 3.5 Connectivity scores for a part of forebrain data

Transcript ID	$ENSMUST$00000189259.1	$ENSMUST$00000046950.12	$ENSMUST$00000123809.1
ENSMUST00000189259.1	1	−0.055908792	0.019843849
ENSMUST00000046950.12	−0.055908792	1	0.145207725
ENSMUST00000123809.1	0.019843849	0.145207725	1
ENSMUST00000149344.7	−0.112757438	0.099835833	−0.089340209
ENSMUST00000167868.7	−0.130669106	0.005473841	−0.022489436
ENSMUST00000189509.1	0.137113441	−0.086750946	−0.066904227
ENSMUST00000119197.7	0.106424016	−0.053223473	−0.062112685
ENSMUST00000109290.1	0.125259372	0.02178721	0.08581417
ENSMUST00000149039.1	0.000939426	−0.029122427	0.091181666
ENSMUST00000146506.1	−0.062944873	0.144164886	0.001261276
ENSMUST00000066646.10	−0.00203935	0.077093814	0.038430567
ENSMUST00000054462.10	0.069951374	−0.002403063	0.061446709
ENSMUST00000143587.1	0.098155928	−0.021073061	0.010444539
ENSMUST00000165007.7	0.065714151	−0.054344048	−0.169778656
ENSMUST00000076587.4	0.009950895	0.109715279	0.251859546
ENSMUST00000169433.2	−0.02691602	−0.030667344	−0.017149811
ENSMUST00000117721.7	0.127431119	0.003147644	0.086900793
ENSMUST00000172082.1	0.130047248	−0.074826788	−0.036869253
ENSMUST00000098566.4	−0.200171475	0.0669632	0.046264033
ENSMUST00000132069.1	0.014420096	0.076583141	0.070644103
ENSMUST00000214893.1	0.080510788	−0.058849397	0.096746841
ENSMUST00000120912.7	−0.008869985	0.019437347	0.045674396
ENSMUST00000136120.1	−0.102641455	0.039155616	0.036417927
ENSMUST00000044352.6	−0.048739562	0.010524215	0.005903374
ENSMUST00000105507.4	−0.01591187	−0.054766826	−0.062871316

The statistical significance of these scores is determined by comparing the ϵ threshold value. In this thesis, ϵ was taken as 0.4. The absolute values of all scores were compared with 0.4, and the transcripts with significant association were considered. 8 component PLSR showed associations between 619 genes out of 2288 gene. For the first transcript ”ENSMUST00000149344.7”, there was a relation between 10 other transcripts, according to the connectivity scores equal to or greater than 0.4 (Table 3.6).

Table 3.6 Connectivity scores ≥ 0.4 for a part of forebrain data

Transcript ID	$TranscriptID$	$ConnectivityScores$
ENSMUST00000149344.7	$ENSMUST00000136252.7$	0.788720287
ENSMUST00000149344.7	$ENSMUST00000156967.7$	0.438031448
ENSMUST00000149344.7	$ENSMUST00000174661.8$	0.589502431
ENSMUST00000149344.7	$ENSMUST00000200480.1$	0.496122416
ENSMUST00000149344.7	$ENSMUST00000156232.1$	0.504289268
ENSMUST00000149344.7	$ENSMUST00000123869.7$	0.410540942
ENSMUST00000149344.7	$ENSMUST00000176717.1$	0.441532964
ENSMUST00000149344.7	$ENSMUST00000055408.12$	0.450291922
ENSMUST00000149344.7	$ENSMUST00000147268.7$	0.507600116
ENSMUST00000149344.7	$ENSMUST00000180465.7$	0.426338591

In the reduced midbrain data set, the $3,262,735$ connectivity score for 2555 transcripts was computed. The connectivity scores for 3 transcripts with the first 25 transcripts are shown in Table 3.7.

Table 3.7 Connectivity scores for a part of midbrain data

Transcript ID	$ENSMUST$00000170167.7	$ENSMUST$00000181821.7	$ENSMUST$00000127116.6
ENSMUST00000170167.7	1	−0.006072448	0.231886263
ENSMUST00000181821.7	−0.006072448	1	−0.063622277
ENSMUST00000127116.6	0.231886263	−0.063622277	1
ENSMUST00000193907.5	−0.124461086	0.099264254	0.097582449
ENSMUST00000146165.7	−0.30688227	−0.067562387	−0.324209242
ENSMUST00000128831.1	0.068870263	0.093454677	0.038999954
ENSMUST00000055754.7	0.008346625	−0.372046418	−0.002650352
ENSMUST00000153951.1	0.017143382	0.039443539	−0.022113878
ENSMUST00000049150.7	0.062686207	−0.042277869	0.040629481
ENSMUST00000109431.9	0.043194004	0.151454843	0.013722276
ENSMUST00000142984.1	0.060035375	0.151325826	−0.175359019
ENSMUST00000081982.11	−0.043131516	−0.025504114	−0.042866955
ENSMUST00000126117.1	−0.027027322	0.092563002	0.020236553
ENSMUST00000183638.7	0.012483043	−0.019917872	0.121762503
ENSMUST00000165856.2	−0.214871781	0.204773702	−0.138251734
ENSMUST00000187055.1	0.169648759	0.019032932	0.21858333
ENSMUST00000128028.1	0.11056939	−0.092357566	0.13576049
ENSMUST00000210543.1	0.117652154	0.110541271	−0.050815579
ENSMUST00000160009.1	0.071511176	−0.141425807	−0.07221787
ENSMUST00000077119.7	0.116444819	0.188963545	−0.063909401
ENSMUST00000118875.7	0.008811168	−0.067307579	0.108778897
ENSMUST00000064635.11	0.404250888	−0.142220168	0.234567626
ENSMUST00000028102.13	0.030960018	0.007008734	−0.07114548
ENSMUST00000116440.8	0.019539134	−0.055554789	0.068152946
ENSMUST00000112979.2	−0.087330347	0.004359345	−0.062015419

To build a network of genes with the statistically significant associations, the absolute values of the scores were compared with $\epsilon = 0.4$. Table 3.8 illustrates the significant transcript ids and their scores. There are associations between 684 genes out of 2286 gene. The first transcript ”ENSMUST00000170167.7” was associated with 13 different transcripts, in midbrain data.

Table 3.8 Connectivity scores ≥ 0.4 for a part of midbrain data

Transcript ID	$TranscriptID$	$ConnectivityScores$
ENSMUST00000170167.7	$ENSMUST00000064635.11$	0.404250888
ENSMUST00000170167.7	$ENSMUST00000142182.7$	0.457933312
ENSMUST00000170167.7	$ENSMUST00000184550.7$	0.445709491
ENSMUST00000170167.7	$ENSMUST00000199377.1$	0.436250601
ENSMUST00000170167.7	$ENSMUST00000179001.7$	0.407556632
ENSMUST00000170167.7	$ENSMUST00000192355.5$	0.418907943
ENSMUST00000170167.7	$ENSMUST00000169754.7$	0.468947117
ENSMUST00000170167.7	$ENSMUST00000119603.1$	0.55767594
ENSMUST00000170167.7	$ENSMUST00000206462.1$	0.4254855
ENSMUST00000170167.7	$ENSMUST00000151973.1$	0.433882114
ENSMUST00000170167.7	$ENSMUST00000054310.3$	0.416900156
ENSMUST00000170167.7	$ENSMUST00000141589.1$	0.538911152
ENSMUST00000170167.7	$ENSMUST00000145956.1$	0.444156303

Lastly, $2,956,096$ connectivity scores for 2432 transcripts were found in the reduced hindbrain data set. The results of 25 transcripts between 3 transcripts are given in Table 3.9.

Table 3.9 Connectivity scores for a part of hindbrain data

Transcript ID	$ENSMUST$00000073388.12	$ENSMUST$00000085248.11	$ENSMUST$00000198527.1
ENSMUST00000073388.12	1	0.079893051	−0.05695058
ENSMUST00000085248.11	0.079893051	1	−0.117364113
ENSMUST00000198527.1	−0.05695058	−0.117364113	1
ENSMUST00000043722.9	0.118068306	0.07080645	−0.063171369
ENSMUST00000156581.1	−0.005280443	0.0393527	−0.001950395
ENSMUST00000169053.1	0.015972268	−0.013581895	0.067293174
ENSMUST00000108480.1	0.030713706	0.128882077	−0.103555013
ENSMUST00000182802.7	−0.037814443	0.013137091	0.035477227
ENSMUST00000129726.2	0.086322617	−0.050930127	0.1294122
ENSMUST00000182002.1	−0.01780755	0.005478253	0.004562857
ENSMUST00000170036.7	−0.192267736	0.030109105	0.040555545
ENSMUST00000133405.7	0.083848737	−0.079491843	−0.10876217
ENSMUST00000067043.4	−0.082764953	−0.025669855	0.102746839
ENSMUST00000051094.8	−0.050927519	0.027033528	−0.03729811
ENSMUST00000152223.1	0.028593952	−0.028643921	0.105279517
ENSMUST00000153474.8	0.169528033	0.102939673	−0.126207147
ENSMUST00000070004.3	0.155125529	0.108863702	−0.047329185
ENSMUST00000121454.1	0.023816446	−0.014927865	0.02256159
ENSMUST00000131155.7	0.059232493	0.016647596	−0.072559874
ENSMUST00000197400.4	−0.029414609	−0.025517075	0.182363596
ENSMUST00000149611.1	0.210367062	0.052164062	0.050968793
ENSMUST00000140111.7	0.023061071	−0.004342424	0.138058565
ENSMUST00000129808.1	0.004377881	−0.045237753	−0.012183707
ENSMUST00000054384.5	−0.05425471	−0.038752553	0.163011528
ENSMUST00000210051.1	0.107443646	−0.041596533	0.031468293

For all reduced hindbrain data, PLSR showed associations between 675 genes out of 2234 gene, using 8 components. The results of comparing the scores of the first transcript with 0.4 showed the significant relationships with 3 transcripts.

Table 3.10 Connectivity scores ≥ 0.4 for a part of hindbrain data

Transcript ID	$TranscriptID$	$ConnectivityScores$
ENSMUST00000169053.1	$ENSMUST00000173493.7$	0.409291638
ENSMUST00000169053.1	$ENSMUST00000143425.1$	0.406130761
ENSMUST00000169053.1	$ENSMUST00000143361.1$	0.433458687

3.2.3 Gene Networks

Gene networks are established after the significant connectivity scores are obtained. The network is visualized with the Cytoscape software. Forebrain network consists of 619 nodes (genes) with $3,232,153$ associations between them (Figure 3.11 and closer look on Figure 3.12). Midbrain network have 684 nodes with $3,262,735$ associations (Figure 3.13 and closer look on Figure 3.14). Hindbrain network contains 675 nodes with $2,956,096$ associations (Figure 3.15 and closer look on Figure 3.16).

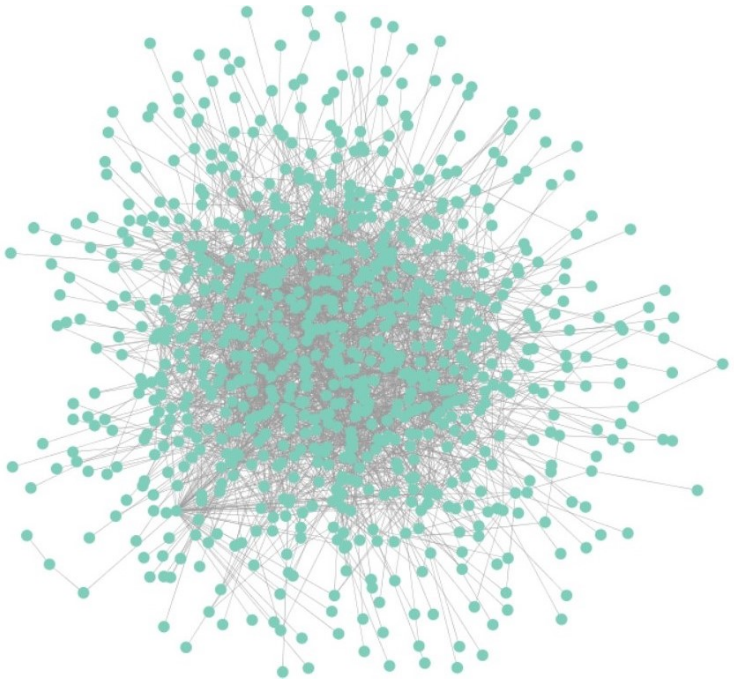

Figure 3.11 Constructed association gene networks using the PLSR for the forebrain data (a)

ENSMUST00000155391.7
ENSMUST00000144304.1
ENSMUST00000151384.1
ENSMUST00000029944.12
ENSMUST00000132069.1
ENSMUST00000066804.4
ENSMUST00000202912.1
ENSMUST00000155205.1
ENSMUST00000134057.7
ENSMUST00000106535.1
ENSMUST00000202604.1
ENSMUST00000161258.7
ENSMUST00000097634.3
ENSMUST00000045366.9
ENSMUST00000111646.7
ENSMUST00000164900.7
ENSMUST00000156796.2
ENSMUST00000150420.1
ENSMUST00000147356.7
ENSMUST00000118599.7
ENSMUST00000185567.1
ENSMUST00000163739.2
ENSMUST00000136814.1
ENSMUST00000212625.1
ENSMUST00000058771.12
ENSMUST00000124640.7
ENSMUST00000191695.1
ENSMUST00000192003.1

Figure 3.12 Constructed association gene networks using the PLSR for the forebrain data (b)

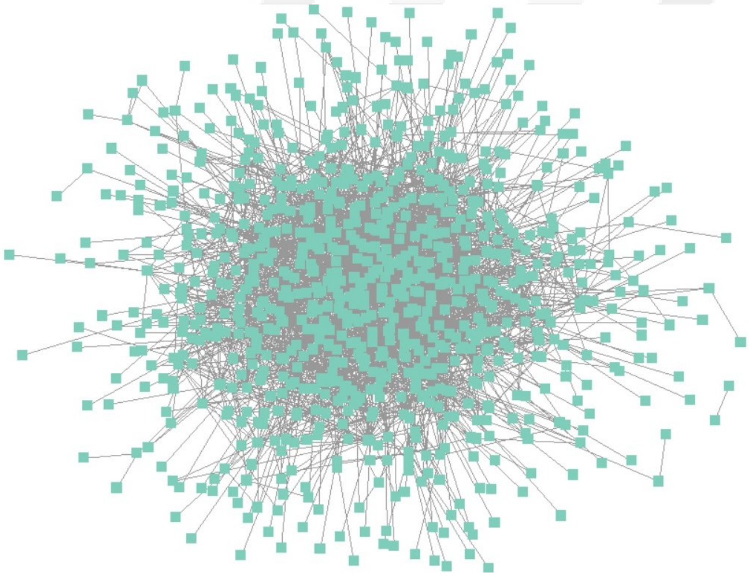

Figure 3.13 Constructed association gene networks using the PLSR for the midbrain data (a)

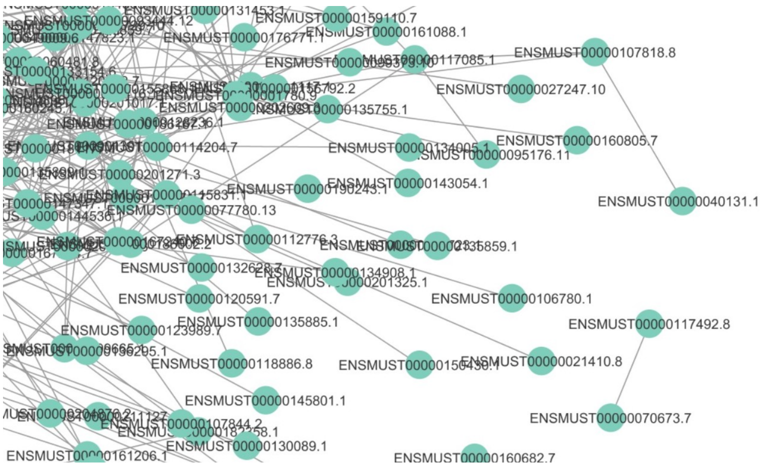

Figure 3.14 Constructed association gene networks using the PLSR for the midbrain data (b)

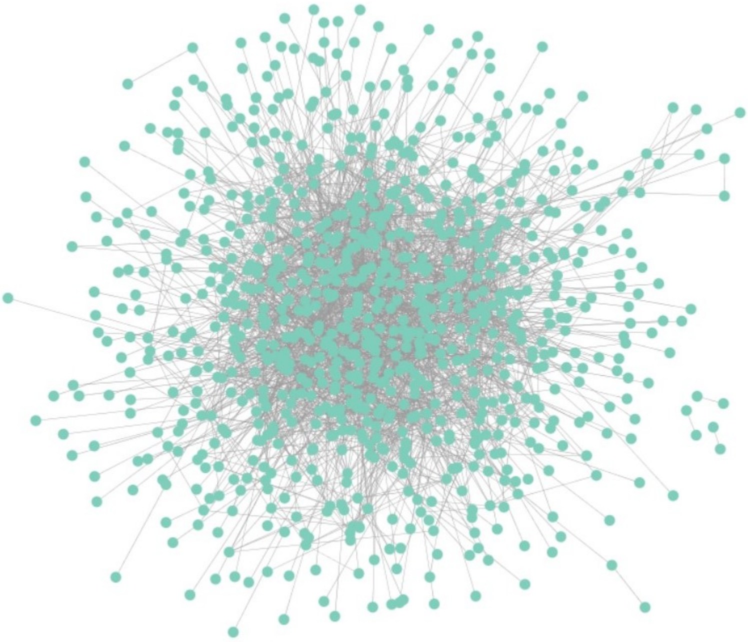

Figure 3.15 Constructed association gene networks using the PLSR for the hindbrain data (a)

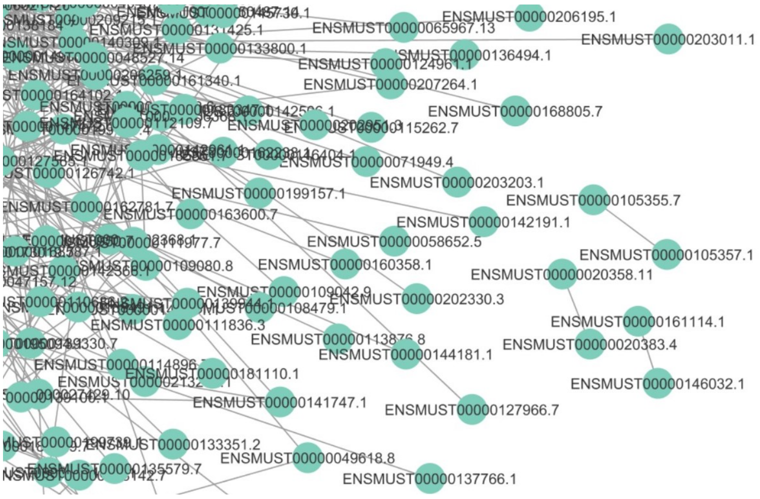

Figure 3.16 Constructed association gene networks using the PLSR for the hindbrain data (b)

CHAPTER FOUR
CONCLUSION

The goal of this thesis is built gene co expression network by connectivity scores based on PLSR. Because PLSR can be applied to complex data sets in many ways, it has various application areas such as chemistry, economics and biology. The first suggestion for using PLSR in gene networks was made by Datta. And since then, it has been among the preferred methods. The PLSR method is shown as a powerful tool for discovering associations between genes since it can tolerate missing values, noise, the cases that the large number of variables relative to the number of observations, and most importantly multicollinearity problem within the response ($\boldsymbol{Y}$ variables) and explanatory variables ($\boldsymbol{X}$ variables). The decision of how many components are used is necessary before computing the PLSR parameters. For this purpose, one of the dimension reduction or latent variable based methods, such as PCA, must use. The PCA method, which is frequently used in high-throughput gene expression data and in many biological studies, generates the new independent variables using the dimension reduction when the problem of multicollinearity problem is seen in the datasets. After the number of components is determined by PCA, PLSR parameters which are loadings and weights of $\boldsymbol{X}$ and $\boldsymbol{Y}$ variables, are calculated for each gene pair. Several algorithms for PLSR, can be found in the literature. To find the associations between genes for each pair, the connectivity scores are calculated by using one of the PLSR algorithms, NIPALS. The statistical significance of the connectivity score, which is a practical tool for exploring the associations in gene networks, is determined with comparing a threshold value, ϵ. If the absolute value of the score for two genes is greater than the threshold value ϵ, it can be said that there is a meaningful relationship between them. The gene networks are established by using significant gene associations obtained from the comparison of each gene pair's connectivity scores with ϵ.

In this thesis, PLSR is used to observe the development of the mouse brain in the embryonic period on three brain parts; those are fore-, mid- and hindbrain. There are 26 samples (except the forebrain data with 25 samples) that spread of 8 days in the datasets. Classification of the all three datasets by 8 days and the decision of the 8 number of component were shown by PC analysis. PLSR was performed with 8 components on the reduced datasets. According to PLSR weights and loadings, the connectivity scores for each gene pair was calculated and compared with the threshold, $\epsilon = 0.4$. For the forebrain data, 619 genes from 2288 genes were found to be significant association. 684 genes out of 2286 genes were found to be significant association in the midbrain data. Lastly, 675 genes from 2234 genes were shown to associate in the hindbrain data. Furthermore, instead of 10% of the datasets, all calculations will be renewed and gene networks will be created for whole datasets in future works.

REFERENCES

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (pls regression). *Wiley interdisciplinary reviews: computational statistics*, *2*(1), 97–106.

Alaiya, A. A., Franzén, B., Hagman, A., Silfverswärd, C., Moberger, B., Linder, S., & Auer, G. (2000). Classification of human ovarian tumors using multivariate data analysis of polypeptide expression patterns. *International journal of cancer*, *86*(5), 731–736.

Alin, A., & Agostinelli, C. (2017). Robust iteratively reweighted simpls. *Journal of Chemometrics*, *31*(3), e2881.

Bastien, P., Vinzi, V. E., & Tenenhaus, M. (2005). Pls generalised linear regression. *Computational Statistics & data analysis*, *48*(1), 17–46.

Boulesteix, A.-L. (2004). Pls dimension reduction for classification with microarray data. *Statistical applications in genetics and molecular biology*, *3*(1), 1–30.

Boulesteix, A.-L., & Strimmer, K. (2005). Predicting transcription factor activities from combined analysis of microarray and chip data: a partial least squares approach. *Theoretical Biology and Medical Modelling*, *2*(1), 23.

Boulesteix, A.-L., & Strimmer, K. (2006). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in bioinformatics*, *8*(1), 32–44.

Branden, K. V., & Hubert, M. (2004). Robustness properties of a robust partial least squares regression method. *Analytica Chimica Acta*, *515*(1), 229–241.

Bras, L., & Menezes, J. (2006). Dealing with gene expression missing data. *IEE Proceedings-Systems Biology*, *153*(3), 105–119.

Bush, B. L., & Nachbar, R. B. (1993). Sample-distance partial least squares: Pls optimized for many variables, with application to comfa. *Journal of computer-aided molecular design*, *7*(5), 587–619.

Carter, S. L., Brechbühler, C. M., Griffin, M., & Bond, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, *20*(14), 2242–2250.

Çoban, Z. (2014). *Fare gelişiminde rolü olan nöronal faktörlerin gen ekspresyon düzeylerinin belirlenmesi.* Doktora Tezi. Gülhane Askeri Tıp Akademisi, Ankara.

Datta, S. (2001). Exploring relationships in gene expressions: a partial least squares approach. *Gene expression*, *9*(6), 249–255.

De Jong, S. (1993). Simpls: an alternative approach to partial least squares regression. *Chemometrics and intelligent laboratory systems*, *18*(3), 251–263.

Ding, Shuang, X. Y. H. T. M. P. (2014). Partial least square based gene expression analysis in renal failure. *Diagnotic Pathology*, *9*.

Dreissig, I., Machill, S., Salzer, R., & Krafft, C. (2009). Quantification of brain lipids by ftir spectroscopy and partial least squares regression. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *71*(5), 2069–2075.

Faria, A., Macedo Jr, F., Marsaioli, A., Ferreira, M., & Cendes, F. (2011). Classification of brain tumor extracts by high resolution ^{1}h mrs using partial least squares discriminant analysis. *Brazilian Journal of Medical and Biological Research*, *44*(2), 149–164.

Finlay, B. L., & Darlington, R. B. (1995). Linked regularities in the development and evolution of mammalian brains. *Science*, *268*(5217), 1578–1584.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica chimica acta*, *185*, 1–17.

Gill, R., Datta, S., & Datta, S. (2010). A statistical framework for differential network analysis from microarray data. *BMC bioinformatics*, *11*(1), 95.

Helland, I. S. (1988). On the structure of partial least squares regression. *Communications in statistics-Simulation and Computation*, *17*(2), 581–607.

Helland, I. S. (1990). Partial least squares regression and statistical models. *Scandinavian Journal of Statistics*, 97–114.

Höskuldsson, A. (1988). Pls regression methods. *Journal of chemometrics*, *2*(3), 211–228.

Huang, C.-C., Tu, S.-H., Huang, C.-S., Lien, H.-H., Lai, L.-C., & Chuang, E. Y. (2013). Multiclass prediction with partial least square regression for gene expression data: applications in breast cancer intrinsic taxonomy. *BioMed research international*, *2013*.

Huang, X., & Pan, W. (2003). Linear regression and two-class classification with gene expression data. *Bioinformatics*, *19*(16), 2072–2078.

Huang, X., Pan, W., Grindle, S., Han, X., Chen, Y., Park, S. J., Miller, L. W., & Hall, J. (2005). A comparative study of discriminating human heart failure etiology using gene expression profiles. *BMC bioinformatics*, *6*(1), 205.

Hubert, M., & Branden, K. V. (2003). Robust methods for partial least squares regression. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *17*(10), 537–549.

Johansson, D., Lindgren, P., & Berglund, A. (2003). A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. *Bioinformatics*, *19*(4), 467–473.

Kondylis, A. (2006). *PLS methods in regression*. PhD thesis, Université de Neuchâtel.

Land Jr, W. H., Ford, W., Park, J.-W., Mathur, R., Hotchkiss, N., Heine, J., Eschrich, S., Qiao, X., & Yeatman, T. (2011). Partial least squares (pls) applied to medical bioinformatics. *Procedia Computer Science*, *6*, 273–278.

MacKinnon, E. (2012). *Dna direclty photographed for first time.* https://www.livescience.com/25163-dna-directly-photographed-for-first-time.html.

Mehmood, T., Martens, H., Sæbø, S., Warringer, J., & Snipen, L. (2011a). Mining for genotype-phenotype relations in saccharomyces using partial least squares. *BMC bioinformatics*, *12*(1), 318.

Mehmood, T., Martens, H., Sæbø, S., Warringer, J., & Snipen, L. (2011b). A partial least squares based algorithm for parsimonious variable selection. *Algorithms for Molecular Biology*, *6*(1), 27.

Musumarra, G., Barresi, V., Condorelli, D., Fortuna, C., & Scire, S. (2004). Potentialities of multivariate approaches in genome-based cancer research: identification of candidate genes for new diagnostics by pls discriminant analysis. *Journal of Chemometrics*, *18*(3-4), 125–132.

Nguyen, D. V., & Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, *18*(1), 39–50.

NLM, U. (2018). *What is a gene?* https://ghr.nlm.nih.gov/primer/basics/gene.

Pérez-Enciso, M., & Tenenhaus, M. (2003). Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (pls-da) approach. *Human genetics*, *112*(5-6), 581–592.

Pihur, V., Datta, S., & Datta, S. (2008). Reconstruction of genetic association networks from microarray data: a partial least squares approach. *Bioinformatics*, *24*(4), 561–568.

Ramírez, J., Górriz, J., Segovia, F., Chaves, R., Salas-Gonzalez, D., López, M., Álvarez, I., & Padilla, P. (2010). Computer aided diagnosis system for the alzheimer's disease based on partial least squares and random forest spect image classification. *Neuroscience letters*, *472*(2), 99–103.

Rosipal, R., & Trejo, L. J. (2001). Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of machine learning research*, *2*(Dec), 97–123.

Serneels, S., Croux, C., Filzmoser, P., & Van Espen, P. J. (2005). Partial robust m-regression. *Chemometrics and Intelligent Laboratory Systems*, *79*(1-2), 55–64.

Shokri-Kojori, E., Tomasi, D., Wiers, C. E., Wang, G.-J., & Volkow, N. D. (2017). Alcohol affects brain functional connectivity and its coupling with behavior: greater effects in male heavy drinkers. *Molecular psychiatry*, *22*(8), 1185.

Tenenhaus, A., Guillemot, V., Gidrol, X., & Frouin, V. (2010). Gene association networks from microarray data using a regularized estimation of partial correlation based on pls regression. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, *7*(2), 251–262.

Wang, D., Song, X., Wang, Y., Li, X., Jia, S., & Wang, Z. (2014). Gene expression profile analysis in epilepsy by using the partial least squares method. *The Scientific World Journal*, *2014*.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate analysis*, 391–420.

Wold, H. (1975). Soft modelling by latent variables: the non-linear iterative partial least squares (nipals) approach. *Journal of Applied Probability*, *12*(S1), 117–142.

Wold, S., Ruhe, A., Wold, H., & Dunn, III, W. (1984). The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, *5*(3), 735–743.

Wold, S., Sjöström, M., & Eriksson, L. (2001). Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, *58*(2), 109–130.