5-2016

Identification of Biomarkers for the Overall Survival of Ovarian Cancer Patients

Kristi Mai
University of Arkansas, Fayetteville

Identification of Biomarkers for the Overall Survival of Ovarian Cancer Patients

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Statistics and Analytics

by

Kristi Mai
University of Belize
Bachelor of Science in Mathematics, 2012

May 2016
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

Dr. Qingyang Zhang
Thesis Director

Dr. Giovanni Petris
Committee Member

Dr. Avishek Chakraborty
Committee Member

ABSTRACT

Rapid advance in sequencing technology has led to genome-wide analysis of genetic and epigenetic features simultaneously, making it possible to understand the biological mechanisms underlying cancer initiation and progression. However, how to identify important prognostic features poses a great challenge for both statistical modeling and computing. In this thesis, a network-based approach is applied to the Cancer Genome Atlas (TCGA) ovarian cancer data to identify important genes related to the overall survival of ovarian cancer patients. In the first step, a stepwise correlation-based selector is used to reduce the dimensionality of TCGA data, by filtering out a large number of unrelated genes. Second, we employ the graphical lasso to construct a sparse gene-gene co-expression network. The undirected network allows us to classify genes into groups based on gene-gene interaction. We fit a cox proportional hazard model with a sparse group lasso penalty for further variable selection and identify 232 genes, which are prognostic for ovarian cancer survival. Of these 232 genes, many were reported to be associated with cancer initiation or progression in the literature. The Kaplan-Meier curves based on the identified genes show clear separation among different groups of patients based on different gene expression levels.

ACKNOWLEDGEMENTS

Foremost, I am highly grateful to God for His abundant blessings that continue to inspire my life. I thank our Almighty God for giving me the courage and the determination, as well as the guidance in conducting this research study, despite all difficulties. It is only through His grace that I have accomplished this much.

I would like to express my sincere gratitude to my thesis advisor, Dr. Qingyang Zhang, who has mentored me during my research. It is only with his patience, encouragement, and valuable insights in both statistics and biology, that this research and thesis were made possible. I am truly honored and fortunate to be his student.

I also want to thank the remaining members in my thesis committee: Dr. Avishek Chakraborty and Giovanni Petris, for imparting their knowledge during the last two years and providing me with the tools necessary to complete this research. With them, I learned the importance of statistics.

Last but not least, I would like to thank my family for all their love and encouragement throughout these years. For my parents who have always believed in me, motivated me and supported me in all my pursuits. For my sister who has always been there for me through it all. For Avjinder who has who has supported me throughout this process and has constantly encouraged me when the tasks seemed arduous and insurmountable. I love you all dearly.

DEDICATION

To my parents, Alfredo and Patricia Mai, for their unconditional love and endless support. To my sister, Keila, my confidant, my best friend. To Avjinder Kaler, for his self-less help and love. To the loving memory of my paternal grandparents, I know you would have both been so proud.

TABLE OF CONTENTS

1. INTRODUCTION

Ovarian cancer is one of the most common gynecologic cancers, ranking fifth as the cause for cancer-related deaths among women in the United States. According to The American Cancer Society, it is estimated that about 22, 280 women will receive a new diagnosis of ovarian cancer and about 14, 240 women will die from it in 2016. About 70% of most deaths occur in patients with advanced-stage, high-grade serous ovarian cancer.

The standard treatment for these patients is usually surgery, followed by platinum-taxane chemotherapy. Platinum-resistant cancer often recurs within six months in about 25% of patients and there is an overall five-year survival probability of 31%. Approximately 13% of high-grade serous ovarian cancer can be attributed to germline mutations in BRCA1 and BRCA2 and a smaller percentage can be accounted for by other germline mutations (The Cancer Genome Atlas Research Network [8]).

Due to the rapid advances in next-generation sequencing technology, it is now possible to simultaneougly perform genome-wide analysis of genetic and epigenetic features (Zhang et al. [47]). The Cancer Genome Atlas (TCGA) project provides the most extensive genomic data resource for more than 30 types of cancers (http://cancergenome.nih.gov/). For instance, the ovarian cancer data from the TCGA contain both clinical and molecular profiles from 586 tumor samples. The clinical profile includes records on recurrence, survival, and treatment resistance. The molecular profile includes copy number variation (CNV), DNA methylation, exon expression, gene expression (microarray), gene expression (RNA-seq), genotype (SNP), MicroRNA expression (microarray), MicroRNA-seq, protein expression, and somatic mutation.

These high-dimensionality datasets have motivated the study of molecular mechanisms of cancer through computational approaches.

A crucial step in the construction of a regression model when there are tens of thousands of features present in the dataset is feature selection. The purpose of feature selection is to select a subset of the original features so that the feature space is optimally reduced based on a certain evaluation criterion. As the years progress, the dimensionality of data keeps increasing in both the number of instances as well as the number of features in various applications. This high-dimensionality leads to problems such as scalability and learning performance of many machine learning algorithms. For instance, high-dimensional data such as a gene expression dataset with hundreds or thousands of genes can have large amount of irrelevant and redundant features which may significantly reduce the performance of machine learning algorithms. Through feature selection, we are able to remove irrelevant or redundant features which increases computational efficiency and estimation accuracy.

Feature selection algorithms are divided into two categories which include the filter model and the wrapper model. Using the filter model, certain features are selected based on general characteristics of the training data without the use of any learning algorithm. On the other hand, the wrapper model uses the performance of a predetermined learning algorithm to evaluate and select the features. The wrapper model has a superior learning performance than the filter model since it selects features which are more suited to the predetermined learning algorithm; however, it tends to be more computationally expensive than the filer model. So, the filter model is often preferred due to its computational efficiency when dealing with a large number of features.

Zhang et al. [49] proposed a novel stepwise correlation-based selector (SCBS) which imitates the hierarchy of the Bayesian network model for feature selection. This approach was applied to the TCGA ovarian cancer data and several interesting results were obtained which provided insight on the genetic/epigenetic mechanisms of ovarian cancer.

In this paper, we identify biomarkers which play a crucial role in the overall survival of the ovarian cancer patients. The data we are going to analyze is the ovarian cancer data, which was retrieved from the TCGA portal. The ovarian cancer data from TCGA includes 586 samples with gene expression profiles containing level 3 UNC Agilent G4502A_07 microarrays. The data contains gene expression level for 17,814 genes. Due to the high-dimensionality of the data, we use the stepwise correlation-based selector (SCBS) proposed by Zhang et al. [49] and select a subset of 603 genes from the 17, 814 genes. With these 603 genes, we will construct an undirected network using the graphical lasso model proposed by Friedman et al. [11]. This will allow for the identification of gene clusters, which will be used in fitting a cox proportional hazard model using a sparse group lasso penalty (Friedman et al. [12]).

The rest of paper is organized as follows: in Chapter 2, we provide some background information through the revision of papers based on the sparse inverse covariance estimation with the graphical lasso and sparse group lasso. In Chapter 3, we study statistical methods such as the stepwise correlation-based selector, graphical lasso, cox proportional hazard model with sparse group lasso penalty, and Kaplan-Meier curves. In Chapter 4, we interpret the results obtained from the analysis. Conclusions are given in Chapter 5.

2. BACKGROUND/LITERATURE REVIEW

2.1 Sparse inverse covariance estimation with the graphical lasso

Several authors have proposed the method of l_1 (lasso) regularization as a form of estimating sparse undirected graphical models. The underlying assumption for this basic model is that the observations follow a multivariate Gaussian distribution with mean μ and covariance matrix Σ. Given other variables, variables i and j are said to be conditionally independent if the ij^{th}component of Σ^{-1} is zero. For this reason, an l_1 penalty is imposed when estimating Σ^{-1} under sparsity assumption.

Different methods for the optimization of the exact log-likelihood have been proposed by several researchers (Yuan and Lin [46]; Banerjee et al. [3]; Friedman et al. [11]). Given n multivariate normal observations of dimension p, with mean μ and covariance matrix Σ, we want to maximize the penalized log-likelihood

$$\mathrm{l}(\Theta) = \log|\Theta| - tr(S\Theta) - \lambda\|\Theta\|_1 \qquad (\ 2.1.1)$$

where S represents the sample covariance matrix, $\Theta = \Sigma^{-1}$, and $\|\Theta\|_1 = \sum_{i,j}|\Theta_{\mathrm{ij}}|$.

According to Banerjee et al. [3], the maximization of equation (2.1.1) is equivalent to solving the dual problem

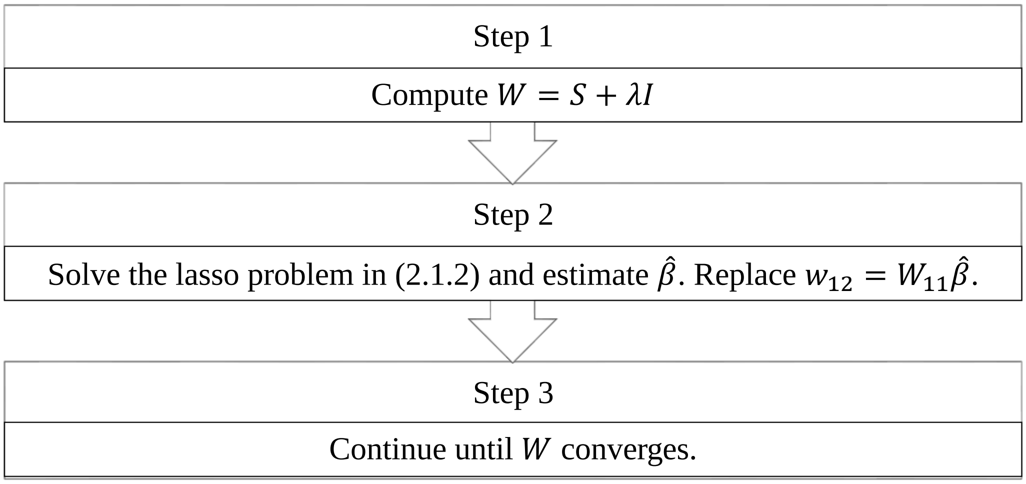
$$\min_{\beta}\left\{\frac{1}{2}\left\|W_{11}^{\frac{1}{2}}\beta - W_{11}^{-\frac{1}{2}}s_{12}\right\|^2 + \lambda\|\beta\|_1\right\} \qquad (\ 2.1.2)$$

where

$$W = \begin{bmatrix} W_{11} & w_{12} \\ w_{12}^{T} & w_{22} \end{bmatrix} \tag{ 2.1.3}$$

$$S = \begin{bmatrix} S_{11} & s_{12} \\ s_{12}^{T} & s_{22} \end{bmatrix} \tag{ 2.1.4}$$

Suppose we let $\beta = W_{11}^{-1} w_{12}$, then the problem becomes much easier due to the equivalence between (2.1.1) and (2.1.2). This lasso problem can be solved using a coordinate descent procedure. Friedman et al. [13] developed a simple algorithm known as the graphical lasso, which is extremely fast. This algorithm is able to solve a 1000-node problem within a minute and is 3000 times faster than other competing algorithms. The graphical lasso algorithm can be implemented as follows:



Step 1

Compute $W = S + \lambda I$

Step 2

Solve the lasso problem in (2.1.2) and estimate $\hat{\beta}$. Replace $w_{12} = W_{11}\hat{\beta}$.

Step 3

Continue until W converges.

Since the graphical lasso algorithm is simple and fast in estimating a sparse inverse covariance matrix using the l_1 penalty, it should aid in the application of sparse inverse covariance procedures involving large datasets, which contain thousands of parameters.

2.2 A sparse-group lasso

For problems where there are grouped covariates, which can have sparse effects on a group as well as within group level, a regularized model for linear regression is introduced with l_1 and l_2 penalties. Let us begin by examining the usual linear regression model. We have a dataset which consists of an n response vector y, and an n by p matrix of features, X. In recent times, we have been presented with applications in which $p \gg n$. For such applications, standard regression fails. To overcome this problem, Tibshirani [41] developed the lasso approach, which regularizes the problem by bounding the l_1 norm of the solution. The lasso approach minimizes

$$\frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1 \tag{ 2.2.1}$$

and computes a solution with a small number of nonzero entries in β. Suppose our data contains predictor variables which are divided into m different groups. An example of this is gene expression data which may contain groups for gene pathways or factor level indicators in categorical data.

The objective is to find a solution which uses only a few of the groups, in addition to achieving sparsity in β. To solve this problem, Yuan and Lin [46] proposed the group lasso criterion. The problem is as follows

$$\min_{\beta}\frac{1}{2}\left\|y - \sum_{l=1}^{m} X^{(l)}\beta^{(l)}\right\|^2 + \lambda\sum_{l=1}^{m}\sqrt{p_l}\|\beta^{(l)}\| \tag{ 2.2.2}$$

where $X^{(l)}$ is a submatrix of X with columns corresponding to the predictors in group l, $\beta^{(l)}$ is the coefficient vector corresponding to that group and p_l is the length of $\beta^{(l)}$. The magnitude of the tuning parameter λ determines the sparsity of the solution. Note that if each group size is 1, the result is a regular lasso solution.

The group lasso model yields a sparse set of groups; however, the presence of a group in the model results in all nonzero coefficients in the group. Suppose we want to achieve both sparsity of groups and within each group. To do this, we use the sparse group lasso, which uses the formula

$$\min_{\beta} \frac{1}{2n}\left\| y - \sum_{l=1}^{m} X^{(l)}\beta^{(l)} \right\|^2 + (1-\alpha)\lambda \sum_{l=1}^{m} \sqrt{p_l}\left\|\beta^{(l)}\right\| + \alpha\,\lambda\|\beta\|_1 \qquad (\ 2.2.1)$$

where $\alpha \in [0,1]$. The mixing parameter, α, is a convex combination of the lasso and group lasso penalties since $\alpha = 0$ produces a group lasso fit and $\alpha = 1$ produces a lasso fit.

The sparse group lasso model is often used for regression problems involving categorical predictors. For predictors with a large number of levels, many of the levels for the predictors included are sometimes not very informative so the sparse group lasso accounts for this by replacing the coefficients with zero for many levels even in the nonzero groups. The sparse group lasso is sometimes useful for analyzing gene expression data as it is able to find interesting pathways from which driving genes are selected. In addition, the model also reduces the estimated effects of driving genes within a group toward one another (Simon et al. [34]).

For comparison purposes, all three models (sparse group lasso, group lasso, and lasso) were applied on two real data examples involving gene expression data, the colitis data and breast cancer data. In the colitis data, the lasso outperformed the group lasso and the sparse group lasso while in the breast cancer data, the sparse group lasso outperformed the lasso and group lasso. The difference in these results is due to the fact that group information in the cancer data is critical for classification and the grouping provides us with insights into the biological mechanisms while the group information in the colitis data simply increases model variance. Although the sparse group lasso may not be applicable to all grouped data, it can sometimes be useful as in the case of the cancer data.

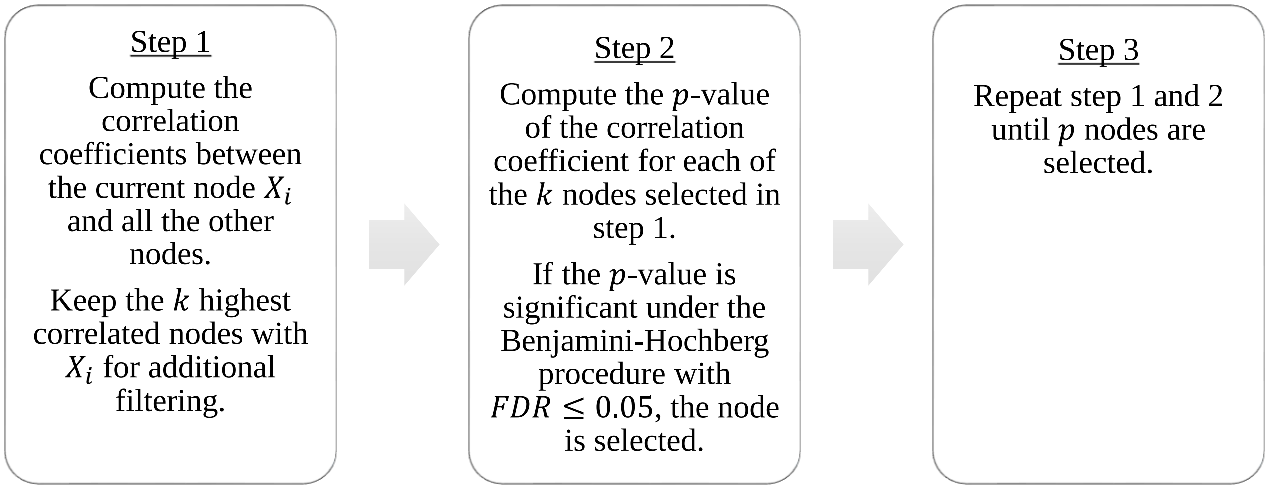
3. METHODOLOGY

3.1 Feature Selection

As the years progress, the dimensionality of data keeps increasing in both the number of instances as well as the number of features in various applications. This high-dimensionality leads to problems such as scalability and learning performance of many machine learning algorithms. For instance, high-dimensional data such as the TCGA ovarian cancer data with 17,814 genes can have a large number of irrelevant and redundant genes, which may significantly reduce the performance of machine learning algorithms. As the dimensionality of a dataset increases, there is an increasing difficulty in proving the result statistically significant due to the sparsity of the meaningful data in the dataset in question. With an increase in dimensionality also comes an increase in computational cost which is usually exponentially. To overcome this problem we use feature selection methods to reduce the number of features in consideration.

Feature selection is a very essential requirement when dealing with high-dimensional data so that data overfitting is avoided and further analysis is possible. Feature selection algorithms are divided into two categories which include the filter model and the wrapper model. Using the filter model, certain features are selected based on general characteristics of the training data without the use of any learning algorithm. On the other hand, the wrapper model uses the performance of a predetermined learning algorithm to evaluate and select the features. The wrapper model has a superior learning performance than the filter model since it selects features which are more suited to the predetermined learning algorithm; however, it tends to be more computationally expensive than the filer model. So, the filter model is often preferred due to its computational efficiency when dealing with a large number of features.

In this paper, the feature selection method that is applied to the TCGA ovarian cancer data is a stepwise correlation-based selector (SCBS). The underlying assumption we make from a biological perspective is that cancer phenotype is directly associated with gene expression. The 17,814 genes from our TCGA ovarian cancer data are fed into the stepwise correlation-based selector (SCBS) and the selection process begins. We begin by computing the correlation between the genes and survival time. At this step, we detect those genes which are significantly correlated with survival time and these genes are selected to be a part of our subset. In the next step, we select those genes which are correlated with the genes that were selected in the first step. We continue in this manner of progressively selecting genes that correlate with the selected genes until a subset with the desired number of genes is obtained. Using this stepwise correlation-based selector, we select 603 genes from the total 17,814 genes. The SCBS algorithm can be implemented as follows:



Step 1

Compute the correlation coefficients between the current node X_i and all the other nodes.

Keep the k highest correlated nodes with X_i for additional filtering.

Step 2

Compute the p-value of the correlation coefficient for each of the k nodes selected in step 1.

If the p-value is significant under the Benjamini-Hochberg procedure with $FDR \leq 0.05$, the node is selected.

Step 3

Repeat step 1 and 2 until p nodes are selected.

The correlation coefficients are computed using Pearson's correlation method. To perform the hypothesis test, the correlations are transformed using Fisher's z-transformation, which is a function of r whose sampling distribution of the transformed value is close to normal. Fisher's z-transformation is given by

$$Z = \frac{1}{2} \ln\left(\frac{1 + r}{1 - r}\right) \tag{3.2.1}$$

where r is the sample correlation, Z is the transformed value of r, and ln is the natural logarithm. Using the fact that Z approximately follows a standard normal distribution, $Z \sim N(0,1)$, we compute the p-values of the correlation coefficients. Note that in the implementation of the SCBS algorithm, k is set to four. The value of k is selected based on previous studies which suggest that k should be four, five, or six. A small value of k fails to capture weakly connected nodes and a large value of k tends to capture more false positives.

When compared to single-round filtering methods, the stepwise correlation-based selector appears to be more effective in selecting those with features, which are associated with the phenotype-related pathways but are indirectly associated with the cancer phenotype. To better understand this, let us consider the following scenario. Assume there is a casual relationship such as $A \rightarrow B \rightarrow Cancer$. Although there is a strong correlation between A and B or B and cancer, the correlation between A and cancer could decay significantly to the extent of being undetectable.

3.2 Graphical Lasso

Recently, the estimation of the inverse covariance in a high-dimensional setting where the number of features p is greater than the number of observations n has gained much interest. Even more so, the estimation of a sparse inverse covariance matrix has gained more spotlight. This is because it involves the estimation of the inverse covariance matrix which has some elements equal to zero. For instance, in an $n \times p$ data matrix X with independent rows which are distributed $N(0, \Sigma)$, a zero in an off-diagonal element of Σ^{-1} would be due to a pair of variables which are conditionally independent. To this end, if we assume a multivariate Gaussian distribution then we can estimate a graphical model for the data using the estimation of the sparse inverse covariate matrix.

In the graphical model, each node represents a feature and the edge between the corresponding pair of nodes represents the nonzero off-diagonal element in the inverse covariance matrix. Usually, Σ^{-1} is estimated by maximizing the log-likelihood of the data. Using the Gaussian model, we can represent the log-likelihood as

$$\log \det \Sigma^{-1} - tr(\mathrm{S}\Sigma^{-1}) \qquad (3.2.2)$$

where $S = \frac{X^T X}{n}$ is the estimated covariance matrix of the data. Let $\Theta = \Sigma^{-1}$. Then we can denote the maximum likelihood estimate of (3.2.2) by $\widehat{\Theta} = S^{-1}$. Generally, this estimate does not contain any elements equal to zero. In addition, having more features than observations in our data, that is, $p \gg n$, will produce an S which is singular so we would not be able to compute the maximum likelihood estimate.

Yuan and Lin [46] proposed an alternative to this, which involves maximizing the penalized log-likelihood over nonnegative definite matrices Θ, instead of simply maximizing the log-likelihood. The penalized log-likelihood is

$$\log\det\Theta - tr(S\Theta) - \lambda\|\Theta\|_1 \tag{3.2.3}$$

where λ is a nonnegative tuning parameter. This problem is referred to as the graphical lasso (Friedman, Hastie, and Tibshirani [11]).

There are two main advantages of using a penalized log-likelihood rather than the simple log-likelihood. First, regardless if S is singular, the solution will always be positive definite for all $\lambda > 0$. Second, for a sufficiently large λ, the estimated $\widehat{\Theta}$ will be sparse because of the lasso-type penalty, which has been applied to the elements of Θ (Tibshirani [41]).

In order for the solution to the graphical lasso problem to be block diagonal with blocks $C_1, C_2, \dots, C_K$, that is, for a set of nodes to form a connected component in the graphical model, a necessary and sufficient condition is required. The condition is that $|S_{ii'}| \leq \lambda$ for all $i \in C_k$, $i' \in C_{k'}$, $k \neq k'$. This condition was discovered by Mazumder and Hastie [26] and can be implemented prior to solving equation (3.2.3) so that large computational gain is achieved.

The R package for graphical lasso with version glasso1.7 uses the above condition to estimate a sparse inverse covariance matrix using a lasso (l_1) penalty. The general idea behind the algorithm implemented in this package is that for a specified value of the tuning parameter, if the solution to the graphical lasso problem will be block diagonal, then the graphical lasso algorithm is applied to each block separately. Using a block diagonal screening decreases computation time significantly.

The covariance matrix, S, which is a symmetric $p \times p$ matrix is computed from our $n \times p$ data matrix X. Note that n is the number of samples at risk of death which is 296 and p is the number of genes which is 603. The glasso function is applied to the covariance matrix S and the value of lambda, the regularization parameter for lasso is set equal to 0.1. A smaller value of λ is always preferred to a larger value of λ. This is due to the fact a smaller value of λ yields less sparse Θ which fits the data well while a larger value of λ yields a sparser Θ which fits the data less well.

The output from the glasso function includes: w which is the estimated covariance matrix, w_i which is the estimated inverse covariance matrix, $loglik$ which is the value of the maximized log-likelihood penalty, del which is the change in the parameter value at convergence, $niter$ which is the number of iterations of the outer loop used by the algorithm, $approx$, and $errflag$.

Butts et al. [6] developed the network package in R, which provides a general framework for encoding complex relational structures composed of a vertex set along with a combination of edges. The tools in this package allow us to create, access, and modify network class objects which facilitate the representation of more complex structures from adjacency matrices. In addition, it also allows us to efficiently handle large sparse networks.

Let G denote a network, a relational structure on a given vertex set (V) and an edge, such that T is the "tail set" of the edge and H is the corresponding "head set" belonging to the ordered pair (T, H) with the property that $T, H \subseteq V(G)$. The cardinality of the vertex set and corresponding edge set are denoted by $|V(G)| = n$ and $|E(G)| = m$, respectively. In an undirected network, the head and tail sets of an edge are interchangeable, meaning that i is adjacent to j if there exists an edge such that $i \in T, j \in H$ or $i \in H, j \in T$.

Using the inverse covariance matrix w_i which was previously estimated using the glasso function, we construct our adjacency matrix. The network function uses the adjacency matrix to create an undirected network object. The object is plotted and a two-dimensional plot of the undirected network is obtained.

3.3 Cox proportional hazard model with sparse group lasso penalty

The advantage of using sparse group lasso over lasso and group lasso is that it generates a solution, which is both between and within group sparsity. Using the SGL package in R, which was developed by Simon et al. [35], we fit a cox proportional hazard model via a penalized maximum likelihood, which is a combination of a lasso and group lasso regularization. This package contains four functions, two of which we use; cvSGL and SGL. The cvSGL function is used to fit and cross-validate a cox model via the penalized maximum likelihood.

The arguments specified in the function are: data, index, type, nlam, nfold, and alpha. The argument 'data' is a list which consists of an $n \times p$ input matrix X, an n-vector time which corresponds failure/censor times, and an n-vector status which indicates failure (1) or censoring (0). In our case, X is a 568×603 matrix with gene expression levels with n being the total number of samples, and p being the number of genes selected using SCBS. The argument 'index' is a p-vector which indicates group membership of each covariate. To construct the index vector, we use the estimated inverse covariance matrix w_i generated using the glasso function since this was used for the estimation of the undirected network in the network function. All genes belonging to the cluster ($\Sigma_{ij}^{-1} \neq 0$) are assigned to group 1 and those genes not belonging to the cluster ($\Sigma_{ij}^{-1} = 0$) are each assigned a different group number. Type corresponds to the model type; in our case, the cox model. The argument ‘nlam’ corresponds to the number of lambdas to use in the regularization path which we set equal 10 and ‘nfold’ corresponds to the number of folds of the cross-validation loop which is set equal to 5. The mixing parameter, α, determines how much weight should be given to either the lasso or group lasso regression. In our case, we set the mixing parameter, alpha equal to 0.95 which indicates that more weight is given to the lasso than the group lasso. Note that choosing a value of α which is close to 1 eliminates any degeneracies and problematic behavior caused by extreme correlations.

The cvSGL function runs a total of $(nfold + 1)$ times. In the first run, the sequence of lambda is generated. The cross-validated error rate and its standard deviation are computed in the consecutive runs. The output values of the cvSGL function include: lldiff which is an nlam vector of cross-validated log-likelihoods, llSD which is an nlam vector of approximated standard deviations of lldiff, lambdas which is a list of the values of lambda used in the regularization path, type which is the response type, and fit which is a model fit object created.

The sparse group penalty model can be extended to other models. The two most common cases in which this model is implemented include logistic regression and the cox model for survival data. In a cox regression model, the data is a covariate matrix, X, which is divided into sub-matrices based on the groups, an n-vector y which contains failure censoring times, and an n-vector δ which indicates failure or censoring for each observation. Note that $\delta_i = 1$ indicates that observation i failed and $\delta_i = 0$ indicates the observation i was censored. Under this model, the sparse group lasso is expressed by

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}\left[\log\left(\sum_{i\in D}\left(\sum_{j\in R_i}\exp\left(x_j^T\beta\right) - x_i^T\beta\right)\right)\right] + (1-\alpha)\lambda\sum_{l=1}^{m}\sqrt{p_l}\left\|\beta^{(l)}\right\| + \alpha\lambda\|\beta\|_1 \qquad (3.3.1)$$

where D is the set of failure indices, and R_i is the set of indices, j, such that we have $y_j \geq y_i$ which denotes those patients still at risk at failure time i.

3.4 Kaplan-Meier Curves

In 1958, Edward L. Kaplan and Paul Meier developed a way of dealing with incomplete observations and as a result, Kaplan-Meier curves and estimates of survival data have become useful in dealing with differing survival times such as times-to-event in which some of the subjects do not continue in the study. Time-to-event can be defined as a clinical duration variable for each subject in the study. It may begin at the point in time when the subject becomes a part of a study or when the subject begins receiving treatement and ends when the subject reaches the event of interest or is censored from the study.

Kaplan-Meier survival analysis requires three variables for each of the subjects in the study. These variables include the survival time (time-to-death), their status at the end of the study (event occurrence or censored), and the group they belong to. Censoring occurs when the total survival time for a subject cannot be correctly determined due to reasons such as the subject dropping out from the study or the subject survives until the end of the study (Rich et al. [30]).

The Kaplan-Meier estimate is the simplest way of estimating a population survival curve from a sample as it allows us to compute the survival over time regardless of the difficulties associated with subjects or situations. In estimating the survival curve, we compute the probabilities of the occurrence of an event at a certain point of time and multiply these successive probabilities by any previously computed probabilities to get the final estimate. The Kaplan-Meier estimator of the survival function at time t is

$$\hat{S}(t) = \prod_{t_{(i)} \leq t} \frac{n_i - d_i}{n_i} \tag{3.4.1}$$

where n_i denotes the number at risk of dying at $t_{(i)}$ and d_i denotes the observed number of deaths. Note that $\hat{S}(t) = 1$ if $t < t_{(1)}$. The survival probability is calculated by dividing the number of subjects surviving by the number of patients at risk. Subjects at risk do not include subjects who have died, dropped out of the study, or have been censored (Goel et al. [18]).

The cox proportional hazard model is useful in identifying variables, which may be of prognostic importance. In theory, the number of variables which can be included in the cox model are infinite. For a regression model with k-variables, the hazard function is

$$h(t, x_1, x_2, \dots, x_k, \beta) = h_0(t) \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k) \tag{3.4.2}$$

where $h_0(t)$ is the baseline hazard function, β_0 is the intercept, and $\beta_1, \beta_2, \dots, \beta_k$ are the corresponding regression coefficients estimated in the modelling process.

We can express the above equation as a log-hazard function in the form

$$\ln\left[\frac{h(t, x_1, x_2, \dots, x_k, \beta)}{h_0(t)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \qquad (\ 3.4.3)$$

Although it is possible to include an infinite number of variables in the model, there are practical constraints in the estimation of the regression coefficients. For this reason, the number of variables included in the model cannot be greater than the number of events available for the analysis.

To calculate the confidence interval (CI), $\hat{S}(t)$ is transformed using a scale which approximately follows a Normal distribution. This is commonly achieved using a logarithmic transformation of $\hat{S}(t)$. Using this transformed scale, the endpoints of a $100(1-\alpha)$ percent confidence interval for the log-log survival function are given by the expression

$$\ln\left[-\ln\left(\hat{S}(t)\right)\right] \pm z_{1-\alpha/2}\widehat{SE}\left\{\ln\left[-\ln\left(\hat{S}(t)\right)\right]\right\} \qquad (\ 3.4.5)$$

where $z_{1-\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution.

Taking the antilog of the lower and upper values of the CI in (3.4.5) allows us to return to the untransformed scale. The lower and upper endpoints of the confidence interval for the survival function are, respectively

$$\exp[-\exp(\ \hat{c}_u\ )] \quad \text{and} \quad \exp[-\exp(\ \hat{c}_l\ )] \qquad (\ 3.4.6)$$

Note that since $\hat{S}(t)$ always has values ranging from 0 to 1, the CI computed with (3.4.6) will always be in the range of 0 to 1.

When interpreting K-M curves, we look for gaps in these curves in a horizontal or vertical direction. A horizontal gap indicates that a particular group took longer to experience a certain fraction of deaths. A vertical gap indicates that at a specific point in time, a particular group had a greater fraction of subjects surviving. The Kaplan-Meier survival analysis is a convenient method of estimating survival times as it allows us to use the information from subjects who are censored up to the time when they are censored (Machin et al. [24]).

4. RESULTS AND DISCUSSION

4.1 Results

Using the ‘data matrix’ tool available in TGCA data portal, the data was extracted. This data set contains the expression values of 17,814 genes. Table 4.1.1 presents a summary of TCGA ovarian cancer data, which includes the data types we incorporated in the analysis and the number of available cases for each data type.

Data type	Platform	Cases
Gene expression	Agilent G4502A_07	583
Clinical information	N/A	585

Table 4.1.1. Summary of TCGA ovarian cancer data.

Using the stepwise correlation-based selector (SCBS) approach for feature selection, a subset of 603 genes was selected from the total 17,814 genes. The sparse inverse covariance matrix was estimated using the blockwise coordinate descent algorithm for penalized maximum likelihood estimation which is employed in the glasso package in R. The undirected network was constructed using the network package in R. The predicted network contains 589 nodes within the cluster and the remaining 14 nodes are not connected.

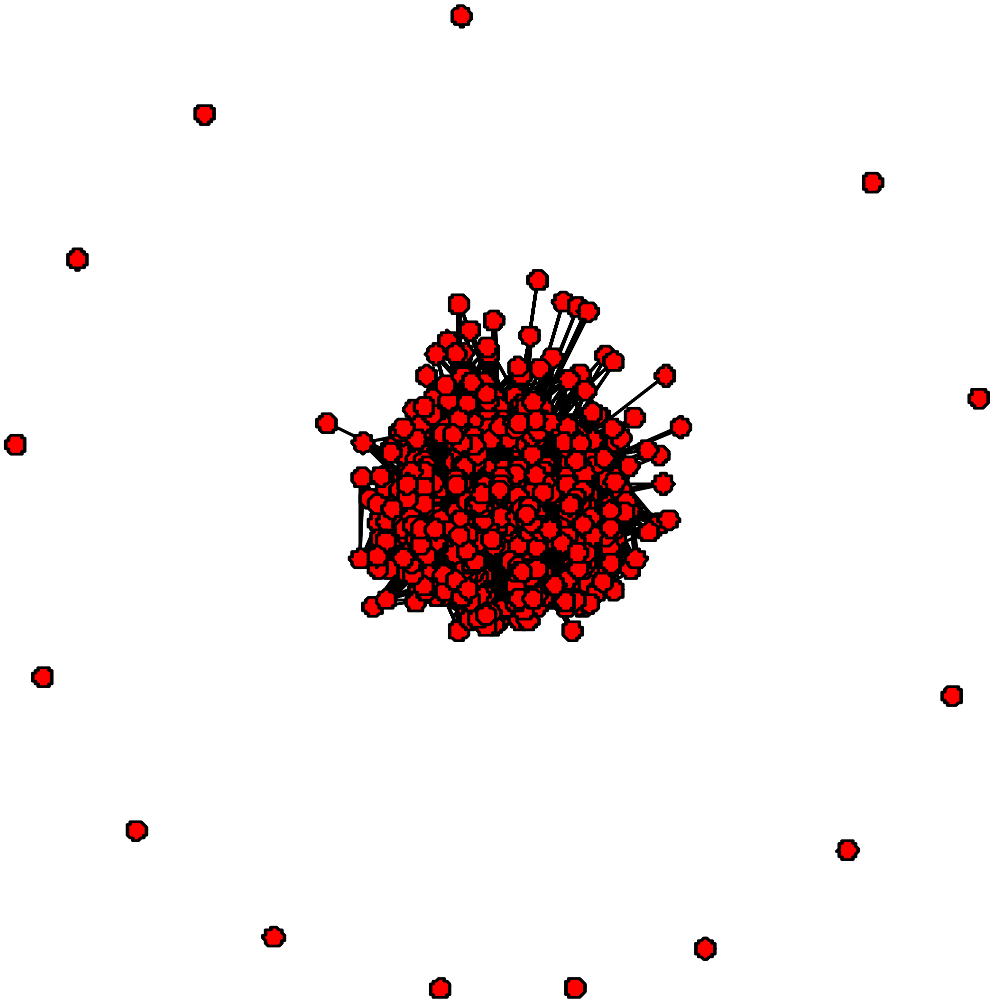


Figure 4.1.1. Undirected network generated using the graphical lasso algorithm with a regularization parameter $\boldsymbol{\lambda = 0.1}$.

We applied the cox proportional hazard model with sparse group lasso penalty to fit a survival model to our data. The 603 gene expression levels for all 568 samples along with the clinical information for these samples were fed into the sparse group lasso (SGL) algorithm which is implemented in R. The survival time is given in days and it is defined as the time between diagnosis and death. The death risk (status) is treated as a binary variable which represent failure as 1 and censoring as 0. The index for all 603 genes contains the group membership of genes. The 589 genes which are in the same cluster are assigned to group 1 and the remaining 14 genes which do not belong to the cluster are each assigned to a different group from group 2 to group 15.

A cox proportional hazard model is fit to the data using 10 lambdas in the regularization path and 5 folds for the cross-validation loop. Using the log-likelihoods along with the lambda values used in the regularization path from the output, we construct a plot.

Number	Lambda	Log likelihood
1	0.003507721	2149.039
2	0.002514584	2220.560
3	0.001802633	2362.001
4	0.001292256	2613.528
5	0.000926381	3013.686
6	0.000664096	3694.887
7	0.000476071	4995.810
8	0.000341282	8315.698
9	0.000244655	14627.111
10	0.000175386	25431.210

Table 4.1.2. Lambda values with their corresponding log-likelihood computed by cv.SGL.

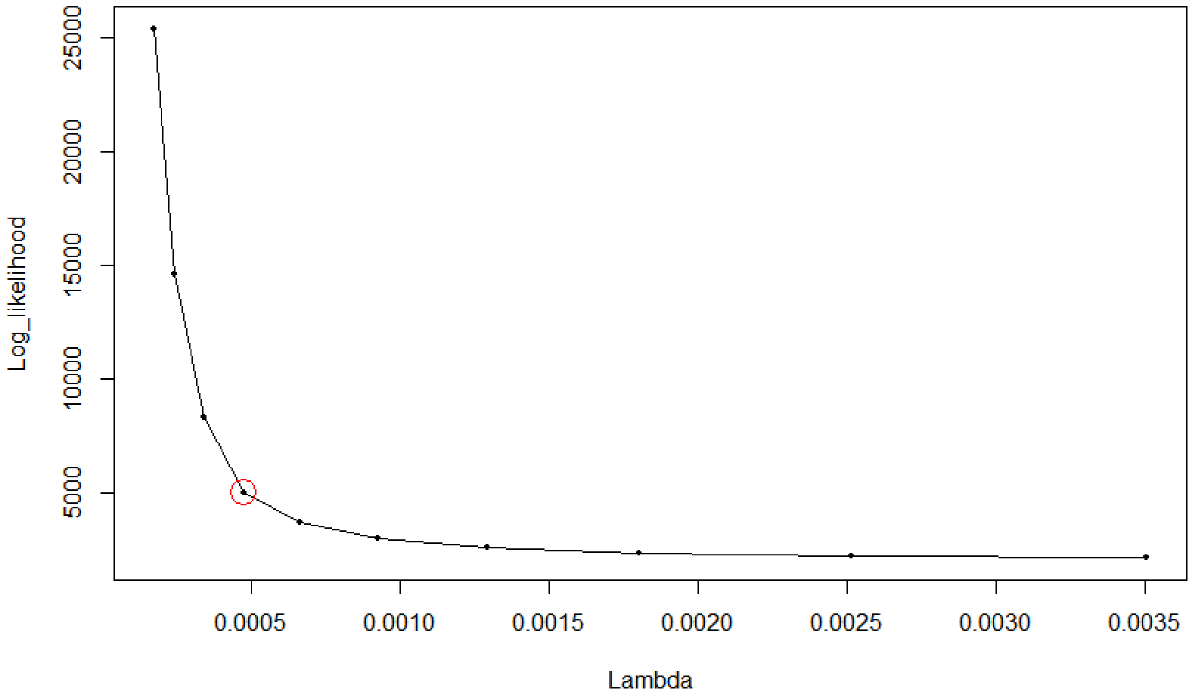


Figure 4.1.2. Plot of the log-likelihoods against their corresponding lambda values.

A common method for selecting the tuning parameter λ is to use cross-validation to select the optimal λ (Sun et al. [39]; Wasserman and Roeder [42]; Sofer et al. [37]). The problem with using the cross-validation method is that it yields large number of false positives in the sparse network problem (Fu and Zhou [14]). A method which has shown to be more effective in indentifying the optimal λ is the “change point” method. The “change point” method uses the change in the log likelihood for different values of λ. Based on this method, the optimal λ corresponds to the change point at which increasing λ does not yield a significant decrease in the log likelihood value. The optimal lambda selected is lambda 7 with a value of λ=0.000476071.

Using the optimal λ that was selected, we fit a cox proportional hazard model with a combination of lasso and group lasso regularization. The input matrix, survival time, status and index all remain the same as what was used in the cv.SGL function. The difference in using the SGL function is that the optimal λ, lambda 7, is used in fitting the cox model to the data. The mixing parameter, α, is set equal to 0.95. The beta cofficients for all 603 genes were estimated using the cox model in the SGL function. After fitting the regression model to the data, those genes for which the null hypothesis $(H_0: \beta_i = 0)$ is rejected, are kept in the model and are termed prognostic. The remaining genes which are not statistically different from zero are removed from the model and are not considered prognostic for the outcome.

The total number of genes with nonzero beta coefficients is 232 genes. Using the gene expression level for these 232 genes along with their estimated beta coefficients, we will compute the survival rate for all 568 samples. The survival estimates are computed by

$$S_i = \beta_0 + \beta_1 g_{1i} + \beta_2 g_{2i} + \cdots + \beta_{232} g_{232i} \qquad (4.1.1)$$

where $\beta_1, \beta_2, \ldots, \beta_{232}$ are the corresponding regression coefficients estimated in the modelling process and i is the index for the sample where $i = 1, 2, \ldots, 568$.

After computing these survival estimates, we will sort these estimates in ascending order. We evenly divide the survival estimates into 2 groups where group 1 includes the first 284 estimates and group 2 includes the remaining 284 estimates. Similarly, we divide the survival estimates into 3 groups while sorted in ascending order. The survival package in R allows us to construct survival curves from a fitted cox model using the survfit function. Kaplan-Meier curves will be plotted using the survival time and status for the 2 groups (low risk and high risk). The procedure is also repeated for the case of 3 groups (low risk, medium risk, and high risk).

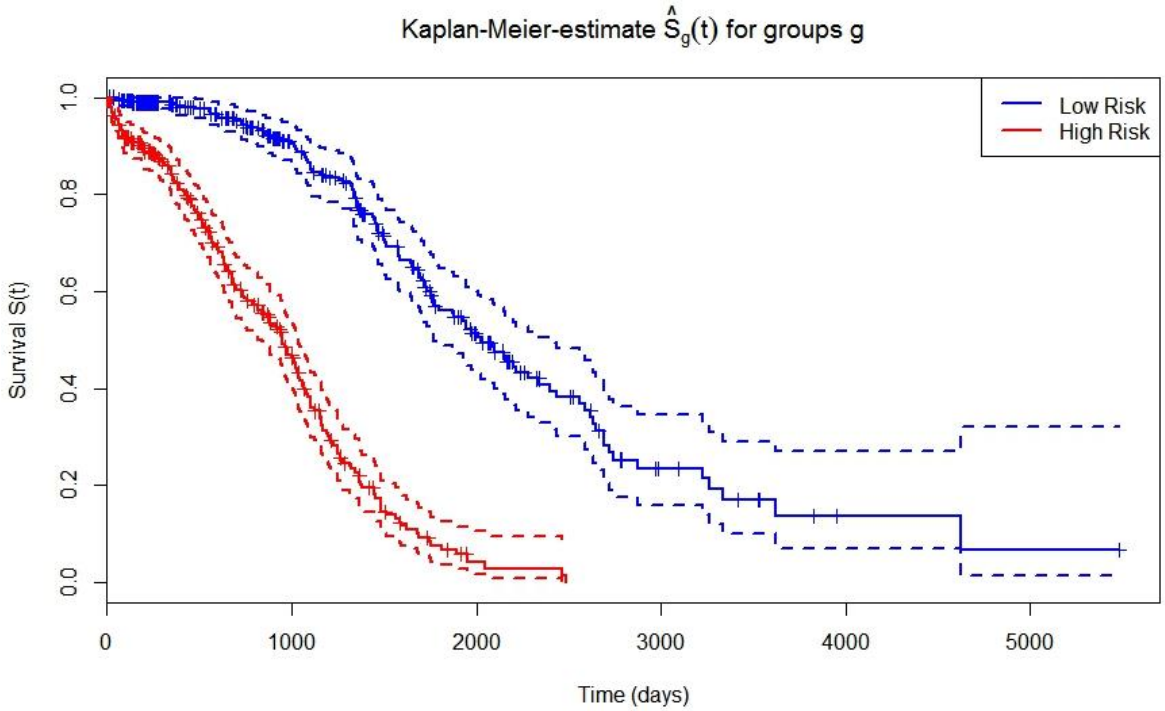


Figure 4.1.3. Kaplan-Meier curves with 95% confidence intervals for low and high-risk groups.

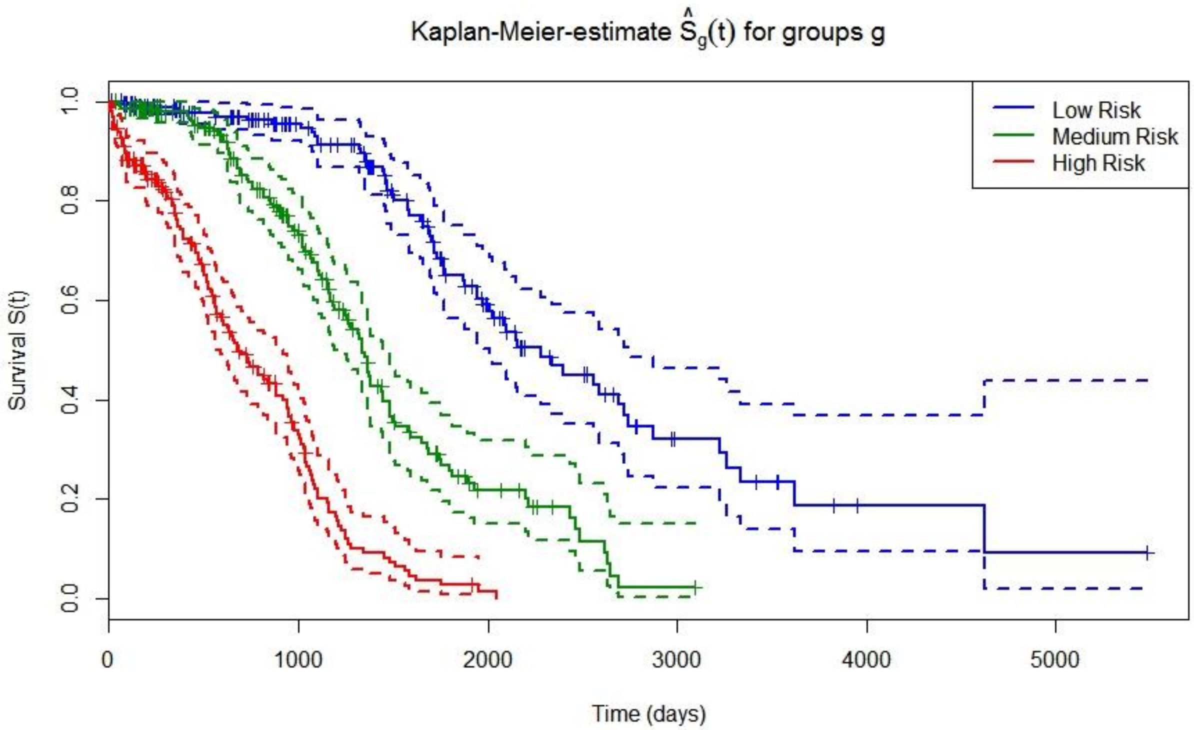


Figure 4.1.4. Kaplan-Meier curves with 95% confidence intervals for low, medium, and high-risk groups.

4.2 Discussion

The K-M estimates of the survival curves are given by the graph of $S(t)$ against time in days in figures 4.1.3 and 4.1.4. $S(t)$ begins at 1, where all patients in the study are alive, since $S(0) = 1$ and then progressively decline towards 0 where all patients have died with time. Since the estimated survival curve remains at a plateau between successive patient death times, the graph of $S(t)$ is plotted as a step function. At each time of death, there is an instantaneous drop to a new level. The graph only attains a value of 0 if the patient with the longest observed survival time dies. In the event that the patient is still alive, the K-M curve has a plateau which begins at the time of the last death and continues until the censored survival time of this longest surviving patient. The censored survival times are marked on the curve with bold vertical lines cutting the curve.

Since we are estimating the difference between 2 groups and 3 groups depending on the potential risk, it is useful to calculate confidence intervals (CI) for the estimates. The survival estimates were partitioned into 2 and 3 groups based on low risk and high risk and low risk, medium risk, and high risk, respectively. The corresponding survival curves were estimated using the samples that fall into these groups. As a measure of the reliability of the estimates at key points along the K-M survival curves, we computed the 95% CI for $S(t)$ at time t.

The survival curves show a better outcome for low risk patients than the high risk patients in figure 4.1.3. As expected, the survival curves indicate a gradient of survival differences between the two groups. Since the K-M curves for the different risk groups are adequately separated, these groups can be used for prognosis. It can be noted from figure 4.1.4 that even though there is a clear difference between low, medium, and high risk groups, the separation between the three groups is not as pronounced as that of figure 4.1.3. For instance, the medium and high risk groups have ‘shrunk’ closer to each other while the low risk group appears to have a relatively similar prognosis.

From the 232 genes which were termed prognostic for ovarian cancer survival, we found 10 of those genes which are directly related to cancer. Protein ubiquitination (CCNB1IP1) is important for many cellular processes as it is able to regulate protein degradation and signal mechanisms. Alterations of the ubiquitination mechanism have become evident in human cancers. Levels of UB ligases have been found to be significantly correlated with relevant prognostic factors as well as with the clinical outcome (Confalonieri et al. [7]). CDK5RAP2 is necessary for spindle checkpoint function (Zhang et al. [51]). The expression of COL2A1 has also shown useful in predicting tumor recurrence in high-grade serous ovarian cancer (Ganapathi et al. [16]). COL8A1 in hepatocarcinoma cells has shown to be correlated with increased tumor cell proliferation (Ma et al. [23]). Over-expression of EIF6 has shown to increase the motility and invasiveness of cancer cells by controlling the expression of a critical subset of membrane-bound proteins (Pinzaglia et al. [28]). GATA6 promotes colon cancer cell invasion through the regulation of urokinase plasminogen activator (uPA) gene expression. It contributes to colorectal tumorigenesis and tumor invasion (Belaguli et al. [4]). Splice variants (SVs) of receptors for the growth hormone-releasing hormone (GHRH) have been detected in several human cancers and cancer cell lines. Antagonists of GHRH have shown to inhibit growth of various human cancers (Garcia-Fernandez et al. [15]). The expression of interferon regulatory factor-1 (IRF-1) is a nuclear transcription factor which mediates interferon and other cytokine effects. IRF-1 appears to have antitumor activity in vitro and in vivo in cancer cells (Kim et al. [20]). The expression NLRX1 acts as a potential tumor suppressor through the regulation of the TNF-α induced apoptosis (cell death) and metabolism in cancer cells (Singh et al. [36]). The expression level of presenilin 1 (PSEN1) has shown to be negatively correlated with chemoresistance. A minor interference of the RNA mediated repression in the PSEN1 gene has shown to suppress cell apoptosis, the multi-chemoresistance of bladder cancer (Deng et al. [10]).

Gene Symbol	Gene Name	Resource
CCNB1IP1	Cyclin B1 Interacting Protein 1 , E3 Ubiquitin Protein Ligase	http://www.ncbi.nlm.nih.gov/pubmed/19543318
CDK5RAP2	CDK5 Regulatory Subunit Associated Protein 2	http://www.ncbi.nlm.nih.gov/pubmed/19282672
COL2A1	Collagen, Type II, Alpha 1	http://www.ncbi.nlm.nih.gov/pubmed/26311224
COL8A1	Collagen, Type VIII, Alpha 1	http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3501407
EIF6	Eukaryotic Translation Initiation Factor 6	http://bmccancer.biomedcentral.com/articles/10.1186/s12885-015-1106-3
GATA6	GATA Binding Protein 6	http://www.ncbi.nlm.nih.gov/pubmed/21076612
GHRH	Growth Hormone Releasing Hormone	http://www.ncbi.nlm.nih.gov/pubmed/12602901
IRF1	Interferon Regulatory Factor 1	http://www.nature.com/onc/journal/v23/n5/full/1207023a.html
NLRX1	NLR Family Member X1	http://www.ncbi.nlm.nih.gov/pubmed/25639646
PSEN1	Presenilin 1	http://www.ncbi.nlm.nih.gov/pubmed/25542424

Table 4.2.1. Ten cancer-related genes, which were found to be prognostic for ovarian cancer survival, based on the cox proportional hazard model with sparse group lasso penalty.

Gene Symbol	Function
CCNB1IP1	Functions in progression of the cell cycle through G(2)/M
CDK5RAP2	Potential regulator of CDK5 activity via its interaction with CDK5R1
COL2A1	Essential for the normal embryonic development of the skeleton, for linear growth and for the ability of cartilage to resist compressive forces
COL8A1	Necessary for migration and proliferation of vascular smooth muscle cells and thus, has a potential role in the maintenance of vessel wall integrity and structure
EIF6	Binds to the 60S ribosomal subunit and prevents its association with the 40S ribosomal subunit to form the 80S initiation complex in the cytoplasm.
GATA6	Transcriptional activator that regulates SEMA3C and PLXNA2
GHRH	essential for normal expansion of the somatotrope lineage during pituitary development
IRF1	Plays roles in the immune response, regulating apoptosis, DNA damage and tumor suppression
NLRX1	Participates in antiviral signaling. Acts as a negative regulator of MAVS-mediated antiviral responses, through the inhibition of the virus-induced RLH (RIG-like helicase)-MAVS interaction
PSEN1	Plays a role in intracellular signaling and gene expression or in linking chromatin to the nuclear membrane

Table 4.2.1. Functions of the ten cancer-related genes, which were found to be prognostic for ovarian cancer survival, based on the cox proportional hazard model with sparse group lasso penalty.

Based on the presence of these ten cancer-related genes in our cox proportional hazard model for cancer survival, we have shown that ovarian cancer shares common genes with other cancer types due to the pathological similarity. These findings suggest that certain genes could play essential and common roles across different cancer types.

5. CONCLUSIONS

5.1 Summary

The stepwise correlation-based selector was used in selecting relevant genes for ovarian cancer survival. Out of the 17,814 genes, a subset of 603 genes was selected using SCBS. These 603 genes were then used to estimate the sparse inverse covariance matrix through the graphical lasso algorithm and an undirected network of these genes was constructed. Genes belonging to the same cluster were assigned to the same group and genes outside of the cluster were each assigned a different group number. A cox proportional hazard model with sparse group lasso penalty was fit to our data. The model determined 232 genes which are prognostic in cancer survival. Survival estimates were calculated using the gene expression levels and the estimated beta coefficients for these 232 genes. Based on these estimates, we divided the samples into 2 and 3 groups based on low risk, medium risk, and high risk. The K-M curves for the different risk groups were adequately separated which may suggest that these groups can be used for prognosis. Of these 232 genes, many were reported to be associated with cancer initiation or progression in the literature. Based on these findings it appears that certain genes share common roles across different types of cancer.

5.2 Future Work

In this paper, we considered gene expression levels as prognostic biomarkers in ovarian cancer survival. Although the results presented here have demonstrated the effectiveness of identifying biomarkers important in cancer survival, it could be further developed in a number of ways. Future extensions to this research could include: incorporation of more genomic profiles, use of Bayesian network modeling, extension of the graphical lasso model for nonparanormal distribution, use of cross-validation to select an optimal value for the mixing parameter (α), and using a smaller value for the regularization parameter, λ, along with community detection to partition the network structure into more clusters.

Carcinogenesis involves multi-level dysregulations, which include genomics, DNA methylomics, and transcriptomics (An et al. [1]). With recent advances in rapid high-throughput genetic and genomic analysis, we are now able to identify a plethora of alterations which can possibly serve as new cancer biomarkers. Each distinct data type such as copy number variations, gene and microRNAs expression, CpG island methylation provides us with a different, somewhat independent, and complementary view of the entire genome (Sokolova et al. [38]). To understand a gene function, it is necessary to analyze more than one single type of data. For us to be able to uncover the intricate underlying mechanisms, we must go beyond simply understanding one molecular level of cancer.

Data type	Platform	Cases
Gene expression	Agilent 244K	583 (8 organ-specific controls)
Somatic mutation	Agilent 415K	587 (8 organ-specific controls)
DNA methylation	Illumina 27K	592 (8 organ-specific controls)
Copy number variation	Agilent 1M	587 (8 organ-specific controls)
Clinical information	N/A	585

Table 5.2.1. Summary of TCGA ovarian cancer data including data types, platform, and the number of available cases.

Bayesian network (BN) is a probabilistic model consisting of a directed acyclic graph (DAG) and an underlying joint probability distribution which uses the prior probability in the prediction of dependent variables. With the use of Bayesian network, we are able to model a multidimensional probability distribution in a sparse way while at the same time searching for independency relations in the data. Compared to the undirected network model, directed networks models such as the Bayesian network are more informative since we are able to visualize the influences and relations of genes as well as describe hidden dependencies among genes. Bayesian network is of great interest in bioinformatics since the probabilistic inference provides a passage for clinical decision making through the intuitive encapsulation of causal links, which exist between diagnostic and prognostic factors (Gevaert et al. [17]; Sesen et al. [32]).

The Gaussian graphical model is the standard parametric model used for continuous data; however, its distributional assumptions are generally unrealistic. For real-valued data in high-dimensional situations, the estimation of sparse undirected graphs relies heavily of the assumption of normality. Assuming normality is not always realistic, especially in a practical setting. Both the nonparanormal and Gaussian graphical models can be used in graph estimation and construction; however, they yield different graphs over a wide range of regularization parameters, which suggests the possibility of having different biological conclusions (Lafferty et al. [21]). Fitting a high-dimensional nonparanormal model can also be achieved using the graphical lasso approach and is no more computationally difficult than estimating a multivariate Gaussian model.

Cross-validation is often performed to aid in model selection through the choice of an optimal value of a penalty parameter. To select the optimal parameter value, a 10-fold cross-validation (CV) is commonly used. The optimal parameter value is that value for which the 10-fold cross-validated penalized (partial) log-likelihood deviance of the model is minimal (Sill et al. [33]). In this research, the change point method was used to select the optimal value of the regularization parameter, λ; however, the mixing parameter α was set to 0.95. Since we want to achieve both sparsity between and within groups, using cross-validation to select an optimal value of the mixing parameter α will be more useful (Ritter [31]).

Community structure is the division of networks into communities (clusters), which are densely connected among their members, and sparsely connected with the rest of the network (Pizzuti [29]). It is an interesting property to investigate as it can reveal abundant hidden information about complex networks, which cannot be not easily detected by simple observation (Liu et al. [22]). One of the main problems in network and data sciences is community detection (Abbe [1]). Detecting communities within a network can provide useful insights on the general structure of the network so that we may further understand specific gene functions in these complex biological networks. Common algorithms used for community detection include Infomap, LPA, Fastgreedy and Walktrap. In this research, using a smaller value for the regularization parameter, λ, would yield a network with more clusters. We could then use a community detection algorithm to detect communities within the undirected network based on similar characteristics and gene functions. This would allow us to have more groups when fitting a cox proportional hazard model with a sparse group lasso penalty.

There is still a lot more work to be done before we can fully understand the prognostic biomarkers in ovarian cancer survival. Employing different network models, relaxing the normality assumption, using cross-validation to select an optimal value for the mixing parameter (α), as well as using a smaller value for the regularization parameter, λ, to partition the network structure into more clusters, along with community detection in the analysis of different genomic profiles could potentially lead to the identification of new biomarkers.

REFERENCES

1 Abbe, E. (2016). Community detection and the stochastic block model.

2 An, N., Yang, X., Cheng, S., Wang, G., & Zhang, K. (2015). Developmental genes significantly afflicted by aberrant promoter methylation and somatic mutation predict overall survival of late-stage colorectal cancer. *Scientific reports*, 5.

3 Banerjee, O., El Ghaoui, L., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, *9*, 485-516.

4 Belaguli, N. S., Aftab, M., Rigi, M., Zhang, M., Albo, D., & Berger, D. H. (2010). GATA6 promotes colon cancer cell invasion by regulating urokinase plasminogen activator gene expression. *Neoplasia*, *12*(11), 856-IN1.

5 Butts, C. T. (2008). network: a Package for Managing Relational Data in R. *Journal of Statistical Software*, *24*(2), 1-36.

6 Butts, C. T. (2015). network: Classes for Relational Data. R package version 1.12.0. http://CRAN.R-project.org/package=network

7 Confalonieri, S., Quarto, M., Goisis, G., Nuciforo, P., Donzelli, M., Jodice, G., & Di Fiore, P. P. (2009). Alterations of ubiquitin ligases in human cancer and their association with the natural history of the tumor. *Oncogene*, *28*(33), 2959-2968.

8 Cancer Genome Atlas Research Network. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, *474*(7353), 609-615.

9 Chen, L., Xuan, J., Gu, J., Wang, Y., Zhang, Z., WANG, T. L., & SHIH, I. M. (2012). Integrative network analysis to identify aberrant pathway networks in ovarian cancer. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (p. 31). NIH Public Access.

10 Deng, H., Lv, L., Li, Y., Zhang, C., Meng, F., Pu, Y., & Zhang, D. (2015). The miR-193a-3p regulated PSEN1 gene suppresses the multi-chemoresistance of bladder cancer. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1852*(3), 520-528.

11 Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432-441.

12 Friedman, J., Hastie, T., & Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*.

13 Friedman, J., Hastie, T., & Tibshirani, R. (2014). glasso: Graphical lasso- estimation of Gaussian graphical models. R package version 1.8. http://CRAN.R-project.org/package=glasso

14 Fu, F., & Zhou, Q. (2013). Learning sparse causal Gaussian networks with experimental intervention: regularization and coordinate descent. *Journal of the American Statistical Association, 108*(501), 288-300.

15 Garcia-Fernandez, M. O., Schally, A. V., Varga, J. L., Groot, K., & Busto, R. (2003). The expression of growth hormone-releasing hormone (GHRH) and its receptor splice variants in human breast cancer lines; the evaluation of signaling mechanisms in the stimulation of cell proliferation. *Breast cancer research and treatment*, *77*(1), 15-26.

16 Ganapathi, M. K., Jones, W. D., Sehouli, J., Michener, C. M., Braicu, I. E., Norris, E. J., & Ganapathi, R. N. (2016). Expression profile of COL2A1 and the pseudogene SLC6A10P predicts tumor recurrence in high-grade serous ovarian cancer. *International Journal of Cancer, 138*(3), 679-688.

17 Gevaert, O., De Smet, F., Timmerman, D., Moreau, Y., & De Moor, B. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*, *22*(14), e184-e190.

18 Goel, M., Khanna, P., & Kishore, J. (2010). Understanding survival analysis: Kaplan-Meier estimate. *International journal of Ayurveda research*, *1*(4), 274.

19 Hira, Z. M., & Gillies, D. F. (2015). A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Advances in bioinformatics*, *2015*.

20 Kim, P. K., Armstrong, M., Liu, Y., Yan, P., Bucher, B., Zuckerbraun, B. S., & Yim, J. H. (2004). IRF-1 expression induces apoptosis and inhibits tumor growth in mouse mammary cancer cells in vitro and in vivo. *Oncogene*, *23*(5), 1125-1135.

21 Lafferty, J., Liu, H., & Wasserman, L. (2012). Sparse nonparametric graphical models. *Statistical Science*, *27*(4), 519-537.

22 Liu, W., Pellegrini, M., & Wang, X. (2014). Detecting communities based on network topology. *Scientific reports*, *4*.

23 Ma, Z. H., Ma, J. H., Jia, L., & Zhao, Y. F. (2012). Effect of enhanced expression of COL8A1 on lymphatic metastasis of hepatocellular carcinoma in mice. *Experimental and therapeutic medicine*, *4*(4), 621-626.

24 Machin, D., Cheung, Y. B., & Parmar, M. (2006). *Survival analysis: a practical approach*. John Wiley & Sons.

25 Mazumder, R., & Hastie, T. (2012). Exact covariance thresholding into connected components for large-scale graphical lasso. *The Journal of Machine Learning Research*, *13*(1), 781-794.

26 Mazumder, R., & Hastie, T. (2012). The graphical lasso: New insights and alternatives. *Electronic journal of statistics*, *6*, 2125.

27 Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 1436-1462.

28 Pinzaglia, M., Montaldo, C., Polinari, D., Simone, M., La Teana, A., Tripodi, M., ... & Benelli, D. (2015). eIF6 over-expression increases the motility and invasiveness of cancer cells by modulating the expression of a critical subset of membrane-bound proteins. *BMC cancer*, *15*(1), 1.

29 Pizzuti, C. (2008). Ga-net: A genetic algorithm for community detection in social networks. In *Parallel Problem Solving from Nature–PPSN X* (pp. 1081-1090). Springer Berlin Heidelberg.

30 Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C., Nussenbaum, B., & Wang, E. W. (2010). A practical guide to understanding Kaplan-Meier curves. *Otolaryngology-Head and Neck Surgery*, *143*(3), 331-336.

31 Ritter, S. J. (2013). Software for prediction and estimation with applications to high-dimensional genomic and epidemiologic data (Doctoral dissertation, University of California, Berkeley).

32 Sesen, M. B., Nicholson, A. E., Banares-Alcantara, R., Kadir, T., & Brady, M. (2013). Bayesian networks for clinical decision support in lung cancer care. *PloS one*, *8*(12), e82349.

33 Sill, M., Hielscher, T., Becker, N., & Zucknick, M. (2014). c060: Extended inference with lasso and elastic-net regularized Cox and generalized linear models. *Journal of Statistical Software*, *62*(5), 1-22.

34 Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, *22*(2), 231-245.

35 Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). SGL: Fit a GLM (or cox model) with a combination of lasso and group lasso regularization. R package version 1.1. http://CRAN.R-project.org/package=SGL

36 Singh, K., Poteryakhina, A., Zheltukhin, A., Bhatelia, K., Prajapati, P., Sripada, L., & Singh, R. (2015). NLRX1 acts as tumor suppressor by regulating TNF-α induced

apoptosis and metabolism in cancer cells. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, *1853*(5), 1073-1086.

37 Sofer, T., Dicker, L., & Lin, X. (2012). Variable selection for high-dimensional multivariate outcomes. *Statistica Sinica, 22(4)*. 1633-54

38 Sokolova, V., Crippa, E., & Gariboldi, M. (2016). Integration of genome scale data for identifying new players in colorectal cancer. *World J Gastroenterol*, *22*(2), 534-545.

39 Sun, W., Wang, J., & Fang, Y. (2013). Consistent selection of tuning parameters via variable selection stability. *The Journal of Machine Learning Research*, *14*(1), 3419-3440.

40 Therneau, T. (2014). A Package for Survival Analysis in S. survival: Survival Analysis. R package version 2.37-7. http://cran.r-project.org/web/packages/survival

41 Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

42 Wasserman, L., & Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, *37*(5A), 2178.

43 Witten, D. M., Friedman, J. H., & Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, *20*(4), 892-900.

44 Ye, J., & Liu, J. (2012). Sparse methods for biomedical data. *ACM SIGKDD Explorations Newsletter*, *14*(1), 4-15.

45 Yu, L., & Liu, H. (2003, August). Feature selection for high-dimensional data: A fast correlation-based filter solution. *In ICML* (Vol. 3, pp. 856-863).

46 Yuan, M., & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, *94*(1), 19-35.

47 Zhang, Q. (2015). Learning Sparse Bayesian Network with Mixed Variables and its Application to Cancer Systems Biology (Doctoral dissertation, NORTHWESTERN UNIVERSITY).

48 Zhang, Q., Burdette, J. E., & Wang, J. P. (2014). Integrative network analysis of TCGA data for ovarian cancer. *BMC systems biology*, *8*(1), 1.

49 Zhang, Q., & Wang, J. P. (2015). A Bayesian network approach for modeling mixed features in TCGA ovarian cancer data. *bioRxiv*, 033332.

50 Zhang, S., Lu, Z., Unruh, A. K., Ivan, C., Baggerly, K. A., Calin, G. A., ... & Le, X. F. (2015). Clinically relevant microRNAs in ovarian cancer. *Molecular Cancer Research*, 13(3), 393-401.

51 Zhang, X., Liu, D., Lv, S., Wang, H., Zhong, X., Liu, B., & Xu, X. (2009). CDK5RAP2 is required for spindle checkpoint function. *Cell cycle*, *8*(8), 1206-1216.