MODEL COMPARISON FOR GYNECOLOGICAL CANCER DATASETS AND SELECTION OF THRESHOLD VALUE

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

BAŞAK BAHÇİVANCI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

SEPTEMBER 2019

Approval of the thesis:

MODEL COMPARISON FOR GYNECOLOGICAL CANCER DATASETS AND SELECTION OF THRESHOLD VALUE

submitted by **BAŞAK BAHÇİVANCI** in partial fulfillment of the requirements for the degree of **Master of Science in Statistics Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Ayşen Akkaya
Head of Department, **Statistics** _____

Prof. Dr. Vilda Purutçuoğlu
Supervisor, **Statistics, METU** _____

Examining Committee Members:

Assoc. Prof. Dr. Meral Ebegil
Statistics, Gazi University _____

Prof. Dr. Vilda Purutçuoğlu
Statistics, METU _____

Prof. Dr. Erdem Karabulut
Biostatistics, Hacettepe University _____

Prof. Dr. Ömür Uğur
Institute of Applied Mathematics, METU _____

Prof. Dr. Barış Sürücü
Statistics, METU _____

Date: 06.09.2019

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Başak Bahçivancı

Signature:

ABSTRACT

MODEL COMPARISON FOR GYNECOLOGICAL CANCER DATASETS AND SELECTION OF THRESHOLD VALUE

Bahçivancı, Başak
Master of Science, Statistics
Supervisor: Prof. Dr. Vilda Purutçuoğlu

September 2019, 75 pages

Cancer is a very common system's disease with its structural and functional complexities caused by high dimension and serious correlation of genes as well as sparsity of gene interactions. Hereby, different mathematical models have been suggested in the literature to unravel these challenges. Among many alternates, in this study we use the Gaussian graphical model, Gaussian copula graphical model and loop-based multivariate adaptive regression splines with/without interaction models due to their advantages over others from simulated datasets. In the first part of the thesis, we apply these models in our quasi-true cancer network by implementing real microarray datasets. The gynecological cancer is the second leading cancer type in women after the breast cancer. But there are less studies about it regarding the breast cancer because of its sociological reasons. Herein, initially, we detect the related literature and generate a list of core genes for this illness. Then, we construct a quasi-true network from these genes. Finally, we infer this network via underlying models and assess their accuracies. Hence we can realistically evaluate the performance of these models in an actual disease's system.

In these analyses, we also observe that the estimates of models highly depend on their threshold values which convert estimated strengths of gene interactions as binary form to construct the graphical network. Thereby, in the second part of the thesis, we

propose a novel approach for the selection of this value by considering the topology of networks and assess our performance via accuracy and computational time.

Keywords: Gaussian Graphical Model, Gaussian Copula Graphical Model, The Birth-and-death Monte Carlo Method, Reverse Jump Markov Chain Monte Carlo Method, Loop-based Multivariate Adaptive Regression, Threshold Selection

ÖZ

JİNEKOLOJİK KANSER VERİ KÜMELERİ İÇİN MODEL KARŞILAŞTIRILMASI VE EŞİK DEĞERİ SEÇİLMESİ

Bahçivancı, Başak
Yüksek Lisans, İstatistik
Tez Danışmanı: Prof. Dr. Vilda Purutçuoğlu

Eylül 2019, 75 sayfa

Günümüzde çok yaygın görülen ve bir sisten hastalığı olan kanser; genler arasındaki ciddi korelasyon, yüksek boyutluluk ve dahası gen etkileşimlerinin seyrekliği sebebiyle, yapısal ve fonksiyonel olarak karmaşık bir yapıdadır. Bu karmaşıklıkların üstesinden gelebilmek için literatürde birçok model ortaya atılmıştır. Bu çalışmada, bu modeler arasından simülasyon veri setleri aracılığıyla diğerlerine göre daha avantajlı olduğu görülen Gaussian grafiksel modeli, Gaussian Coplula grafiksel modeli ve son olarak etkileşimli/etkileşimsiz döngü temelli çok değişkenli uyarlanabilir regresyon modelleri kullanılmıştır. Tezin ilk bölümünde, bu modellerle ve gerçek mikrodizin veri setlerini kullanarak, oluşturduğumuz olası-gerçek jinekolojik kanser ağını modelliyoruz. Jinekolojik kanser kadınlarda, meme kanserinden sonar en sık görülen kanser türüdür. Fakat bu kanser üzerine, meme kanserine göre sosyolojik sebepler yüzünden daha az çalışma bulunmaktadır. Bu nedenle, çalışmada, ilk olarak jinekolojik kanser üzerine literatür çalışması yaparak, bir gen listesi elde etmekteyiz. Daha sonra bu genlerden olası-gerçek bir ağ yapısı oluşturmaktayız. Son olarak, seçilen modeler aracılığıyla ağ yapısını tahmin ederek, model doğruluklarını değerlendirmekteyiz.

Bu analizler esnasında ayrıca, tahmin edilen gen etkileşim miktarlarını iki değişkenli forma dönüştürerek grafiksel model elde edebilmek için kullanılan eşik değerinin ve

dolayısıyla bu değerin seçiminin model tahminlerinde çok önemli olduğunu gördük. Bu sebeple, tezin ikinci bölümünde ağ yapılarının topolojik özelliklerini göz önünde bulundurarak eşik değeri seçimi yapan yeni bir yaklaşım önermekteyiz ve bu yaklaşımın performansı doğruluk ölçütleri ve hesaplama zamanı açısından değerlendirmekteyiz.

Anahtar Kelimeler: Gaussian Grafiksel Model, Gaussian Kopula Grafiksel Model, Doğum ve Ölüm Markov Zinciri Monte Carlo Metodu, Geri Sıçramalı Markov Zinciri Monte Carlo Metodu, Döngü Temelli Çok Değişkenli Uyarlanabilir Regresyon Modeli, Eşik Değeri Seçimi

To my dear family

ACKNOWLEDGEMENTS

Throughout my research I have received a great deal of support. I would first like to thank my supervisor Vilda Purutçuoğlu for her patient guiding, endless support and smiling face. It was really a great chance to have her as my supervisor.

I would also like to express my appeciations to the examining committee members, Prof. Dr. Meral Ebegil, Prof. Dr. Barış Sürücü, Prof. Dr. Erdem Karabulut, Prof. Dr. Ömür Uğur, for their detailed reviews and time.

I would also thank to the Middle East Technical University research grant (no: BAP-08-11-2017-035) for its support.

My family is my greatest supporter. Especially, my parents are my heroes. It would not be enough how much I thank to them. Thank you both for your endless support throughout my education. Thanks mum, for being a great teacher and a role model of mine as well.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

BDMCMC	The Birth-and-Death Markov Chain Monte Carlo
BF	Basis Function
FN	False Negative
FP	False Positive
GGM	Gaussian Graphical Model
glasso	Graphical Lasso
LMARS	Loop-based Multivariate Adaptive Regression Splines
KMT	Kappa Maximized Threshold
MARS	Multivariate Adaptive Regression Splines
MCC	Matthew's Correlation Coefficient
MDT	Minimized Difference Threshold
MST	Maximized Sum Threshold
RJMCMC	Reverse Jump Markov Chain Monte Carlo
TN	True Negative
TP	True Positive

CHAPTER 1

INTRODUCTION

Cancer is a branch of system disease that is caused by mutations in tumor suppressor genes, oncogenes and DNA repair genes (Romero-Garcia et al., 2011). Considering all cancer types, gynecological cancers are the most prevalent among women in developing counties. (Hsu et al., 2017, Small et al., 2017; Iyoke & Ugwu, 2013). Because of sociological reasons, there is an imminent and serious crisis for gynecological cancer types in developing countries (Small et al., 2017). In addition to that, only small proportion of cancer patients respond to the drug prescribed for their treatment (Wijst et al., 2018). Therefore, accurately screening genes and their interactions has become more crucial to prevent gynecological cancers and finding the right cure.

On the other hand, still, there are several challenges to discover gene interactions due to the structural and the functional complexities inherited in biological systems such as high sparsity of the system, high number of genes relative to the number of observations and high correlation between genes. Hence, to be able to obtain reliable interaction network to explain actual system's disease, the choice of mathematical models plays very critical role. For this purpose, in this study, mathematical models are studied carefully to model network systems more accurately in the real network datasets. Because in the literature, majority of the comparative studies are based on either Monte Carlo runs or benchmark data. Herein, we evaluate them comprehensively in order to construct gynecological cancer networks by using real microarray datasets whose core genes are investigated from the associated literature. Since detecting the genes related with the particular cancer is the first step to cure that disease, we aim to compare network models to figure out which model can detect the

cancer related genes better. We also aim to improve one of these models, GGM, by comparing the threshold methods.

Hereby, in the organization of the thesis, we present the following content. In Chapter 2, we give some background information about the networks, graph theory and the topology of biological network. Afterwards, we represent how a true (quasi) network for gynecological cancers can be constructed. For the construction of the underlying network, we check the associated literature and detect eleven genes selected from different review studies. There genes are MAP2K1, MAPK1, CEBPB, CTNNB1, TFAM, TP53, PDIA3, IMP3, ERBB2, CHD4 and MBD3 (The Cancer Genome Atlas Research, 2011; Cancer Genome Atlas Research Network, 2013; Hu et al., 2015).

Accordingly, Chapter 3 continues with the description of different mathematical methods to construct the network. The first method used in this study is the Gaussian graphical model (GGM). This model is one of the fundamental modeling approaches to explain the relationship between two biological entities under the steady-state activation of the system via an undirected graph (Whittaker, 1990). The second approach applied in our analyses is the Gaussian copula graphical model (GCGM) which is the combination of GGM with the Gaussian copula (Green,1995). The major distinction of this model regarding GGM is its inference in the sense that GCGM can be estimated via reverse jump Markov chain Monte Carlo (RJMCMC) approach (Dobra & Lenoski, 2011; Farnoudkia & Purutçuoğlu, 2018) and the birth-and-death Monte Carlo methods (BDMCMC) (Mohammadi & Wit, 2015) under the Bayesian settings. On the other side, GGM is inferred by the penalized likelihood approach or generalized least square methods. Besides these two parametric techniques, we also implement a nonparametric method that is specifically designed for complex biological network. This model is called the loop-based Multivariate Adaptive Regression Splines (LMARS) model.

On the other hand, in the application of all these mathematical models, it is seen that the selection of the threshold value which converts the estimated model parameters to

a binary form has a direct effect in the accuracy of the fitted model. Accordingly, in order to make more reliable decisions about the activations of the complex systems, certain methods about the selection of the optimal threshold value are suggested in the literature (Bassest & Bullmore, 2006). Some of these selections are based on parametric approaches and some are fully nonparametric. For instance, Scheneider et al. (2019) select the threshold in their studies parametrically by taking into account of the distribution of the data. They also adopt an optimization approach based on an estimator, which is the Hill, in order to propose a nonparametric threshold selection method in the univariate extreme value analysis. On the other hand, a nonparametric procedure which depends on the selection of the optimal p-value for the edges' significance via a hypothesis testing is done by comparing existing edges via their randomness is suggested by Aldemuvar (2019). Moreover, Chen et al. (2015) present a parametric method which uses the mixture of distributions in the exponential family. There are also other studies which implement empirical analyses for the underlying system (Liu et al., 2016) or intuitively select 0.5 or 0.6 value as a constant term in order to control biological networks sparsity (Gibson et al., 2013). On conclusion, considering all of the threshold selection methods mentioned above, it is observed that parametric approaches needed detailed information about the data and have restrictions about the underlying distribution. On the contrary, the nonparametric approaches are flexible. However, they can be computationally demanding if their calculations depend on the optimization techniques. On the other side, such methods like using a fix value or accuracy measures to validate the estimated systems 'structure by their (quasi) representations can be more computationally user-friendly. Whereas, current methods in this group do not consider the topology features of the network of interest. Thereby, in this study, we initially evaluate the influence of major threshold selection methods in the literature that are typically applied for the binary construction. Then, we propose an alternative approach in this field and assess the accuracies of estimates under distinct measures and computational time. In this assessment, we apply GGM as it is the fundamental model of many complex approaches. Thus, in Chapter 3, we also describe well-known threshold selection methods which are used

to convert numerical entries of precision matrix into a binary form under an adjacency matrix. The selected threshold selection methods are the kappa maximized threshold criterion (Guisan et al., 1998), minimized difference threshold (Jiménez-Valverde & Lobo, 2007), maximized sum threshold (Manel et al., 2001), and 0.5T criterion (Manel et al., 1999). In addition to these approaches, researchers can also perform either by expert opinions or assigning arbitrary constants for the selection of this value in gene network analysis (Zhao & Duan, 2019; Purutçuoğlu & Seçilmiş, 2019). Accordingly, as a contribution to the field, we introduce a novel nonparametric procedure which considers the topology of the gene network system while imposing a threshold value to the precision matrix.

Additionally, in Chapter 4, firstly, we present how we obtain the real datasets which include all the genes of the true (quasi) gynecological network. To collect microarray data, we use the ArrayExpress database which is a free database for microarray studies. After searching throughout the ArrayExpress which has more than 1000 data for gynecological cancers, only three different types of gynecological data are found which have same the characteristics and worked simultaneously in the same study. Afterwards, the models which are mentioned in Chapter 3 are assessed by comparing their accuracy measures for the underlying three real oncogenic datasets. Once the listed eleven genes are modelled under a network, we also augment the dimensions (i.e., the total number of genes in the system) in the network construction. Furthermore, to estimate adjacency matrix comparable with the true precision matrix, threshold selection methods are applied under GGM. To illustrate the performance of threshold methods, outputs are evaluated via simulated datasets which are created under different Monte Carlo scenarios, as well as real datasets. Finally, we cover the findings, discuss the outputs and suggest some future works.

CHAPTER 2

BACKGROUND

2.1. Networks

Even in this moment we are surrounded by many changing and interconnected networks while signaling of our neurons and our organs for their works. So, since from the beginning till the end, everything has connection with a network. The network science is a rising and very interdisciplinary research area which develops mathematical approaches in order to expand our natural and man-made network understanding (Börner et al., 2007). Hereby, as a description, network is a group of interconnected things which consists of nodes and edges, and has the application in almost all systems. Majority of the mathematical description of networks is based on the graph theory due to its visual simplicity in understanding the complexity of the actual structure and its flexibility in the application. Hence, the graphs which can represent the network can be divided into two groups. These are directed and undirected graphs. Below, we initially describe these two sorts of graphs. Then, we explain the networks which generate graphs based on their topologies. These topological distinctions are originated from the distributions of the interactions in the systems. For the biological networks, we can observe two types of the distribution in the spread of the interactions. They are the random networks and scale-free networks. Hereby, in the following parts, we also introduce both networks' types in details.

Directed and Undirected Graphs

A finite directed graph consists of a set of edges and nodes in such a way that each edge connects a starting node u to a terminal node v. By this way, the direction of the flow can be observable from the graph. On the other hand, an undirected graph also consists of a set of edges and nodes. However, there is no direction between node u and v. For biological networks, the nodes indicate the genes, proteins or other species in the systems and the edges show the physical or functional interactions (or relationships) between them. Figure 2.1 presents the basic representation of both graphs for two nodes via one edge.

Figure 2.1. A Directed (Left) and an Undirected Graph (Right), Consists of Two Nodes and One Edge

For the biological networks, majority of the graphs are presented by the undirected type since there is a limited information from the direction of the flow (i.e., interaction) about each gene in the system. Therefore, in this study, we merely work on the class of undirected graphs for both the mathematical modeling and the threshold selection. Below, we describe the main criterion which can separate the undirected graphs into two parts as random and scale-free. This criterion is called the degree or connectivity of the graph.

The basic characteristic of a node is its degree (or connectivity), k, which represents the number of links that the node has to other nodes. An undirected network is characterized by an average degree

$$\bar{k} = \frac{2L}{N} \tag{2.1}$$

where N is the size of the networks. The network types can be classified by the probability distribution of their links. Below, we present them with their mathematical descriptions.

2.1.1. Random Networks

There are two definitions for the random network. First one is a $G(N, L)$ model which can simply be described as N labeled nodes that are connected with L randomly placed links (Barabási, 2016).

On the other side, the second definition stands for a $G(N, p)$ model. This model connects each pair of N labeled nodes with a probability p. Accordingly, the $G(N, p)$ model fixes the probability, on the contrary, the $G(N, L)$ model fixes the number of links. Since in real networks, the number of links doesn't stay fixed, the second model is given below as a description of random networks.

In a random network, the probability that a selected node having k links is the product of the following steps (Barabási, 2016):

- The probability that the link k has connection is presented by p^k
- The probability that remaining $(N - 1 - k)$ links are missing is computed by $(1 - p)^{N-1-k}$
- The number of possible ways that k links from $N - 1$ potential links a node can have is selected via $\binom{N-1}{k}$

So, the degree distribution p_k can be shown by

$$p_k = \binom{N-1}{k} p^k (1 - p)^{N-1-k} \tag{2.2}$$

However, most real networks indicate a sparse structure (Barabási, 2016), therefore, $\bar{k} \ll N$ where $\bar{k}$ shows the average degree. Hence, the limiting distribution of the degree (2.2) can be approximated by the Poisson distribution as presented in the following expression.

$$p_k = e^{-\bar{k}} \frac{\bar{k}^k}{k!}. \tag{2.3}$$

The density Equation (2.3) is also called the degree distribution of the random network. In Figure 2.2 (a), a simple view of a random network with eleven nodes can be seen.

2.1.2. Scale-free Networks

The degree distribution p_k gives the proportion of a selected node having k links. The probability is calculated by

$$p_k = \frac{N(k)}{N}, \tag{2.4}$$

where $N(k)$ is the the number of nodes with k number of links.

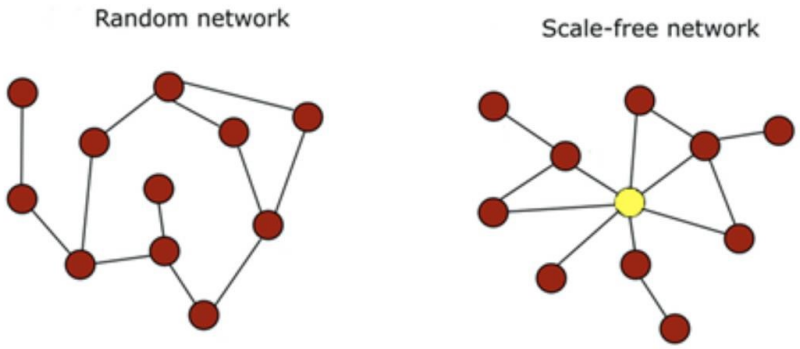
Most biological networks including protein-protein networks are scale-free (Milo et al., 2002; Barabási & Oltvai, 2004). That is, their degree distribution follows a power law distribution as stated below.

$$p_k = k^{-\gamma}\,, \tag{2.5}$$

in which γ denotes the degree exponent. For the biological networks, this value lies from 2 to 3, $2 < \gamma < 3$.

For scale-free networks, although most of the nodes have less number of links, i.e., small-degree, there exists few nodes having very large number of links, i.e., high-degree, in such a way that they are father than the average degree, $\bar{k}$. Those nodes are

called as hubs (Barabási & Oltvai, 2004). In figure 2.2 (b), a simple illustration of a scale-free network with eleven nodes is shown.



(a) (b)

Figure 2.2. The Representation of (a) the Random and (b) the Scale-free Networks with Eleven Genes

2.2. Biological Networks

Although there are many publications suggesting that most of the biological networks are scale-free (Barabási & Oltvai, 2004), there are some recent publications proposing its contrary (Broido & Clauset, 2019). As illustrated in the Figure 2.3, Broido and Clauset (2019) show that a majority of the biological networks do not have any direct and indirect evidence for the structure of scale-free network.

Broido and Clauset (2019) argue that 63% of the biological data that they examine have a lack evidence of scale-freeness. However, they also state that 6% of the biological data, which are mostly metabolic networks, are in the category of strongest, and not all, but most of them, show a direct evidence of the scale-free structure such as some protein-protein interaction networks.

Hereby, in our analyses for the threshold selection, we study via both random and scale-free networks under different Monte Carlo runs. On the other side, we only

consider the scale-free network structure for the real gene networks' data since they are benchmark datasets and are worked as scale-free in distinct studies.

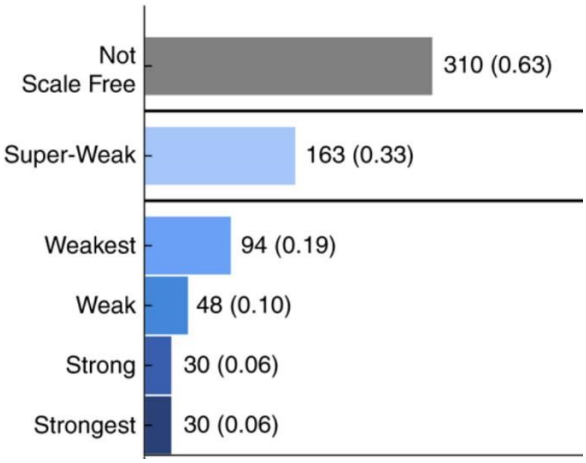


Figure 2.3. The Proportion of Biological Networks by Scale-Free Evidence Category Regarding the Study of Broido and Clauset (2019)

2.2.1. Quasi-Gynecological Network

The gynecological cancer is the second most common cancer type among women after the breast cancer (Iyoke & Ugwu, 2013). There are five main gynecological cancer types seen in the oncogenic literature. These are the vaginal, vulvar, ovarian, cervical and endometrial cancers. In this study, we combine those five subtypes of cancers under a single group and call it as the gynecological cancer while constructing a quasi-gynecological cancer network.

Accordingly, by comprehensive literature review given below, we find that CTNNB1, TFAM, CEBPB, MAP2K1, MAPK1, TP53, PDIA3, IMP3, ERBB2, CHD4 and MBD3 are the genes related with the gynecological cancer by having significant fold changes, and these genes have dense connections within each other so that they can generate complete graph. However, it should be underlined that these genes have

many aliases. For instance, MRPS4, BRMS2, C15orf12 are some aliases of the IMP3 gene symbol and other aliases can be found for the same gene as well. Hence, the researchers can choose each one of the aliases in their works since they all represent the same gene symbol. But this situation creates difficulties in the detection of genes which are related with some specific cancers. Accordingly, in this study, gene aliases are also considered and some of the findings are presented for the eleven genes mentioned above.

CTNNB1

The Ovarian and endometrial carcinomas include many gynecological carcinomas in developed countries (McConechy et al., 2013). It is found that most frequent mutations are observed for ovarian and endometrial endometrioid carcinomas related with CTNNB1 gene (Cho & Shih, 2009; McConechy et al., 2013; Hong et al., 2015). Especially, the endometrial carcinomas are proven to have different types of mutations specific to each subtype by a mutation in CTNNB1 (O'Hara & Bell, 2012).

TFAM

The mitochondrial transcription factor is encoded by the TFAM gene. It has been shown that the unsteady state of Mitochondrial DNA (mtDNA) has a link to metabolic changes, and therefore, contributes to tumorigenesis and increases expression of pro-tumorigenic genes. Hereby, TFAM regulates the energetic metabolism of the glutamine in order to maintain the metabolic needs of the cancer cell. Since highly invasive ovarian cancer are highly dependent to glutamine, TFAM also plays important role on the ovarian cancer cells as well. (Arauja et al., 2018). In another study, Liu et al. (2001) also report that the finding of a high occurrence (60%) of the somatic mtDNA mutations in the human ovarian carcinomas and it is shown that it also has mutations in the endometrial carcinomas (Hong et al.; 2015).

CEBPB

From the studies, it is indicated that a high expression level of the CEBPB gene detected in the endometrial cancers' cells. Since the CEBPB gene is only expressed in the proliferative cancer cells, the expression level of CEBPB plays an important role as a proliferative maker for both cervical and endometrial cancers (Arnett et al., 2003). Moreover, Pan et al. (2010) also report the fact that CEBPB is involved in the cervical cancer.

MAP2K1

The mutation in this gene is found with several types of cancer. The study of carcinoma by He et al. (2015) points out that MAPK1 and MAP2K1 are two of the target genes for the endometrial cancer. Another study states that MAP2K1 (MEK1) is a critical mediator of the pathway which has a significant role in the ovarian carcinogenesis (Miller et al., 2014).

MAPK1

This gene is linked with ovarian, endometrial carcinoma and cervical cancers as presented in the studies of Yiwei et al. (2015) and He et al. (2015). Especially, the studies present that the primary cervical carcinomas are related with the high frequency of mutations in MAPK1. This is also supported by the recent integrated studies of the genomic characterization (Penson et al., 2016). Other studies report that, besides the endometrial cancer, MAPK1 has a crucial role in tumors progression too (He et al., 2015; Chang et al., 2017).

TP53

This gene is the most frequently mutated genes in the human cancer as tumor suppressor gene. Specifically, TP53 is recognized by its association with the ovarian cancer (Köbel et al., 2016; Penson et al., 2016). Moreover, researchers encounter with the TP53 mutations in almost all advanced ovarian carcinomas (Brachova et al., 2014; Mullany et al., 2015). The cancer Genome Atlas also supports the fact that 96% of advanced ovarian carcinomas have the TP53 mutation (Vang et al., 2016).

PDIA3

PDIA3 gene is reported to be highly expressed in serous ovarian cancer (Chay et al, 2010; Takata et al., 2016). PDIA3 is also related with the adenocarcinoma of the uterine cervix (Liao et al, 2011) and the squamous cell carcinoma of the uterine cervix (Chung et al, 2013).

IMP3

IMP3 is reported and validated as biomarker for the ovarian clear cell carcinoma (Köbel et al., 2009). It is also highly expressed in a serous ovarian cancer cell line and is chosen as the best independent biomarker for the uterine serous carcinoma (Mhawech et al., 2010). Lastly, IMP3 is detected to involve in progression of the ovarian cancer and a potential prognostic biomarker (Noske et al., 2009).

ERBB2

ERRB2 is reported to be related with the ovarian cancer, and it is often observed in advanced stages of the cancer (Afify et al., 1999; Fukushi et al., 2001; Wu et al., 2003). Although, its role is controversial, recent studies show that ERBB2 can be used as a prognostic biomarker in the ovarian cancer patients (Luo et al., 2018). Since when ERBB2 is overexpressed, it is reported to be associated with ovarian, endometrial cancer (Koopman et al., 2018; Yang et al., 2019) as well as the uterine serous carcinoma (Zao et al., 2013).

CHD4

CHD4 is also shown as mutated in endometrial endometrioid carcinomas (Le Gallo et al., 2012; Hong et al., 2015). CHD4 is reported as the third most frequently mutated gene in the study of the uterine serous carcinoma (Zao et al., 2013). Another recent study also supports these findings in such a way that CHD4 is a frequently mutated gene in the endometrial cancer (Li et al., 2018).

MBD3

In the pathogenesis of the uterine serous carcinoma, the somatic copy number of the variations plays a major role. Most of the copy number of the deletion is found in the segment of the chromosome 19 which includes 17 genes. Considering those 17 genes, CHD4 and MBD4 are the part of the same complex -the NuRD-. Therefore, MBD3 is also represented as the mutated gene in the uterine serous carcinoma (Zhao et al., 2013).

Table 2.1. Full Names, Aliases and Functions of Our Eleven Core Genes

Gene Symbol/Name	Full Name	Alias	Function
CTNNB1	Catenin Beta 1	NEDSDV, MRD19, EVR7, CTNNB	involves in the production of a protein called beta-catenin which plays an important role in the cell adhesion and in the cell-cell communication
TFAM	Transcription Factor A, Mitochondrial	MTDPS15, TCF6L1, TCF6L2, TCF6 TCF6L3, MTTF1, MTTFA, MtTF1, TCF-6, MtTFA	encodes an important mitochondrial transcription factor in the charge of mitochondrial DNA replication and the repair

CEBPB	CCAAT Enhancer Binding Protein Beta	C/EBP-Beta, C/EBP Beta, IL6DBP, NF-IL6, TCF-5, TCF5, LAP, LIP	encodes a transcription factor which plays important roles in the regulation of genes involved in the immune and the inflammatory responses
MAP2K1	Mitogen-Activated Protein Kinase Kinase 1	EC 2.7.12.2, MAPKK 1, MKK1, PRKMK1, MEK 1, MAPKK1, CFC3, MAPK/ERK Kinase1,	directs the production of the protein known as the MEK1 protein kinase which is a part of a signaling cascade called the RAS/MAPK pathway
MAPK1	Mitogen-Activated Protein Kinase 1	P42MAPK, P41mapk, MAPK 1, MAPK2, ERK, P38, P40, P41, PRKM1, PRKM2, ERK-2, ERK2	encodes a member of the MAP kinase family which is the part of many cellular processes such as the proliferation, differentiation and the transcription regulation

TP53	Tumor Protein P53	Tumor Supressor P53, Tumor Protein 53, BMFS5, TRP53, BCC7, LFS1	encodes a protein called the tumor protein p53 (or p53) which acts as a tumor suppressor (regulates the cell division)
PDIA3	Protein Disulfide Isomerase Family A Member 3	GRP58, ERp57, ERp60, P58, HEL-S-269, HEL-S-93n, HsT17083, PI-PLC, ERp61, GRP57, ERP57, ERP60, ER60	directs the production of a protein of the endoplasmic reticulum that interacts with the lectin chaperones to regulate the folding of newly synthesized glycoproteins
IMP3	IMP U3 Small Nucleolar Ribonucleoprotein 3	C15orf12, MRPS4, BRMS2	involves in the production of the human homolog of the yeast Imp3 protein and interacts with the U3 snoRNP complex

ERBB2	Erb-B2 Receptor Tyrosine Kinase 2	P185erbB2, MLN19, HER2, NGL, NEU, CD340, HER-2, TKR1, HER-2/Neu	encodes a member of the epidermal growth factor (EGF) receptor family of the receptor tyrosine kinases and stabilizes the binding of ligands
CHD4	Chromodomain Helicase DNA Binding Protein 4	Mi2-BETA, Mi2-Beta, EC 3.6.1, SIHIWES, Mi-2b, ATP-Dependent Helicase CHD4	directs the production of the SNF2/RAD54 helicase family which involves in the epigenetic transcriptional repression
MBD3	Methyl-CpG Binding Domain Protein 3	Methyl-CpG Binding Domain Protein 3, Methyl-CpG Binding Domain Protein 3, Methyl-CpG Binding Protein MBD3	the final product of this gene constitutes a multisubunit complex which involves in nucleosome remodeling and histone deacetylase activities

Table 2.2. Sources and Cancer Types of Our Eleven Core Genes

Gene Symbol/Name	Sources	Cancer Types
CTNNB1	Cho & Shih, 2009; Cancer Genome Atlas Research Network, 2011; Cancer Genome Atlas O'Hara & Bell, 2012; McConechy et al., 2013; Research Network, 2013; Hong et al., 2015	Ovarian cancer, endometrial endometrioid carcinomas
TFAM	Liu et al., 2001; Hong et al., 2015; Arauja et al., 2018	Ovarian cancer, endometrial carcinomas
CEBPB	Arnett et al., 2003; Pan et al., 2010	cervical cancer, endometrial carcinomas
MAP2K1	Miller et al., 2014; He et al., 2015	Endometrial carcinomas, ovarian cancer
MAPK1	Cancer Genome Atlas Research Network, 2011; Cancer Genome Atlas Research Network, 2013 He et al., 2015; Yiwei et al., 2015; Penson et al., 2016; Chang et al., 2017;	Ovarian cancer, cervical cancer, endometrial carcinomas, tumor progression

TP53	Cancer Genome Atlas Research Network, 2011; Cancer Genome Atlas Research Network, 2013; Brachova et al., 2014; Mullany et al., 2015; Köbel et al., 2016; Oda et al., 2016; Penson et al., 2016; Vang et al., 2016	Ovarian cancer, tumor progression
PDIA3	Chay et al, 2010; Liao et al, 2011; Chung et al, 2013; Takata et al., 2016	Ovarian cancer, carcinoma of uterine cervix
IMP3	Köbel et al., 2009; Noske et al., 2009; Mhawech et al., 2010	Ovarian cancer, uterine serous carcinoma
ERBB2	Afify et al., 1999; Fukushi et al., 2001; Wu et al., 2003; Cancer Genome Atlas Research Network, 2011; Cancer Genome Atlas Research Network, 2013; Zao et al., 2013; Luo et al., 2018; Koopman et al., 2018; Yang et al., 2018	Ovarian cancer, endometrial cancer, uterine serous carcinoma
CHD4	Le Gallo et al., 2012; Zao et al., 2013; Hong et al., 2015; Li et al., 2018	Endometrial endometrioid carcinomas, uterine serous carcinoma
MBD3	Zhao et al., 2013	Uterine serous carcinoma

CHAPTER 3

METHODS

3.1. Gaussian Graphical Models

The graphical models are an efficient way to represent interactions between two entities of a biological system. They have a set of nodes which represent proteins or genes, and edges between those biological entities as interactions. It is assumed that graphical models have p set of nodes and state vector formalized as $\mathbf{Y} = (Y_1, \dots, Y_p)$. Depending on the application area, Y represents different random variables, such as genes for microarray experiment. As mentioned in Chapter 2, graphical models are divided into two groups such as directed and undirected graphical models. The Gaussian graphical model (GGM) is one of the well-known undirected graphical models for the multivariate continuous data which assumes for the vector $\mathbf{Y}$ to be a multivariate Gaussian distribution via,

$$\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2.6}$$

where $\mu' = (\mu_1, \dots, \mu_p)$ shows the mean vector and $\boldsymbol{\Sigma}$ denotes variance-covariance matrix with a $(p \times p)$ dimension where σ_{ij} is the covariance between Y_i and Y_j when $i \neq j$ and the variances when $i = j$. Furthermore, $\boldsymbol{\Sigma}$ is a symmetric and positive definite matrix.

As the generalization of the univariate normality, the multivariate normally distributed $\boldsymbol{Y}$ has a p dimensional density function as

$$f(y) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} e^{\{-\frac{1}{2}(y-\mu)^T \boldsymbol{\Sigma}^{-1}(y-\mu)\}}, -\infty < y < \infty. \tag{2.7}$$

In GGM, the absence of an edge between two nodes means that these nodes are conditionally independent given all the other nodes. That is, $Y_i \perp Y_j \mid Y_{V\setminus i,j}$ where $V = \{1, 2, \ldots, p\}$ be the set of nodes.

The concept of the independence can be represented indirectly with the variance-covariance matrix. Particularly, it is related with the precision matrix which is the inverse of the variance-covariance matrix, denoted by

$$\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1} = \theta_{ij}\,, \tag{2.8}$$

where $\boldsymbol{\Sigma}$ is invertible since it is symmetric and positive definite matrix. Furthermore, zero in the precision matrix means the conditional independence between the corresponding variables since the precision can is written in terms of the partial covariance. Moreover, the diagonal entries of the precision matrix are simply the inverse of the partial variance as shown below.

$$\theta_{ii} = \frac{1}{var(Y_i \mid Y_{V\setminus i,j})}\,, \tag{2.9}$$

and a minus partial correlation forms the scaled of the diagonal entries via

$$\pi_{ij} = \frac{-\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}\,, \tag{2.10}$$

with π_{ij} the partial correlation as $Y_i \perp Y_j \mid Y_{V\setminus i,j}$.

However, as mentioned above, the high number of genes relative to the number of observations causes the sample variance-covariance estimator, **S**, to be a singular matrix which results in a matrix that is not invertible. Therefore, it is difficult to obtain positive definite and symmetric variance-covariance matrix, $\boldsymbol{\Sigma}$, for the networks such as gene networks because of their distinct nature.

Fortunately, there are different methods to obtain a network's partial correlations. It is a Gaussian graphical model which consists of a set of regression functions with each node against all remaining nodes.

The method can be shown as a joint multivariate Gaussian with vector $\boldsymbol{Y} = (\boldsymbol{Y}_{(-\boldsymbol{p})}, Y_p)$, where $\boldsymbol{Y}_{(-\boldsymbol{p})} = (Y_1, Y_2, \dots, Y_{p-1})$ contains all nodes except p. The conditional distribution of the node Y_p given the remaining node is

$$Y_p \mid \boldsymbol{Y}_{(-\boldsymbol{p})} = y \sim N(\mu_p \big(\boldsymbol{y} - \boldsymbol{\mu}_{-\boldsymbol{p}}\big)^t \boldsymbol{\Sigma}^{-1}_{-\boldsymbol{p},-\boldsymbol{p}} \boldsymbol{\sigma}_{-\boldsymbol{p},\boldsymbol{p}}, \sigma_{p,p} - \boldsymbol{\sigma}^{\boldsymbol{t}}_{-\boldsymbol{p},\boldsymbol{p}} \boldsymbol{\Sigma}^{-1}_{-\boldsymbol{p},-\boldsymbol{p}} \boldsymbol{\sigma}_{-\boldsymbol{p},\boldsymbol{p}}). \quad (2.11)$$

Furthermore, the mean and the covariance matrix can be written by partitioning as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{-\boldsymbol{p}} \\ \mu_p \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{-\boldsymbol{p},-\boldsymbol{p}} & \boldsymbol{\sigma}_{-\boldsymbol{p},\boldsymbol{p}} \\ \boldsymbol{\sigma}^{\boldsymbol{t}}_{-\boldsymbol{p},\boldsymbol{p}} & \sigma_{p,p} \end{bmatrix}. \quad (2.12)$$

Here, $\boldsymbol{\mu}_{-\boldsymbol{p}}$ is the mean vector of all nodes except p and μ_p denotes the mean vector of the node p. The variance-covariance matrix of all nodes except the node p is indicated as $\boldsymbol{\Sigma}_{-\boldsymbol{p},-\boldsymbol{p}}$, and the covariance vector for $\boldsymbol{Y}_{(-\boldsymbol{p})}$ is shown as $\boldsymbol{\sigma}_{-\boldsymbol{p},\boldsymbol{p}}$, and finally, the variance of the node p is represented as $\sigma_{p,p}$. Hereby, the model is presented as

$$y^p = y^{-p}\beta + \varepsilon. \quad (2.13)$$

Here, the conditional independence structure is determined by $\beta = \boldsymbol{\Sigma}^{-1}_{-\boldsymbol{p},-\boldsymbol{p}} \boldsymbol{\sigma}_{-\boldsymbol{p},\boldsymbol{p}}$, that is, by regression coefficients. Since, $\boldsymbol{\Sigma\Sigma^{-1}} = \boldsymbol{I}$, with **I** as the identity matrix, we can get

$$\boldsymbol{\theta}_{-\boldsymbol{p},\boldsymbol{p}} = -\boldsymbol{\theta}_{\boldsymbol{p},\boldsymbol{p}} \boldsymbol{\Sigma}^{-1}_{-\boldsymbol{p},-\boldsymbol{p}} \boldsymbol{\sigma}_{-\boldsymbol{p},\boldsymbol{p}} \quad (2.14)$$

$$= -\boldsymbol{\theta}_{\boldsymbol{p},\boldsymbol{p}} \beta. \quad (2.15)$$

Therefore,

$$\boldsymbol{\beta} = -\frac{\boldsymbol{\theta}_{-\boldsymbol{p},\boldsymbol{p}}}{\boldsymbol{\theta}_{\boldsymbol{p},\boldsymbol{p}}}. \quad (2.16)$$

As seen in Equation (2.14), β and the precision matrix are directly connected with each other. It can be said that Y_i is independent of Y_j given the rest when $\beta_{ij} = 0$ or equivalently $\theta_{i,j} = 0$.

The regression functions are very useful formulations of a Gaussian graphical model when the precision matrix $\boldsymbol{\Theta}$ is desired to estimate via the maximum likelihood technique. The joint density function for observation y_i can be given as

$$f(y_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}} exp\left\{-\frac{1}{2}(y_i - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(y_i - \boldsymbol{\mu})\right\}, \tag{2.17}$$

and likelihood is given by

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2.18}$$

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log\big(L(\boldsymbol{\mu}, \boldsymbol{\Sigma})\big) = -\frac{n}{2} log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(y_i - \boldsymbol{\mu}). \tag{2.19}$$

By replacing the variance-covariance matrix $\boldsymbol{\Sigma}$ with the precision matrix $\boldsymbol{\Theta}$,

$$l(\boldsymbol{\mu}, \boldsymbol{\Theta}) = \frac{n}{2} log|\boldsymbol{\Theta}| - \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{\mu})^t \, \boldsymbol{\Theta} \, (y_i - \boldsymbol{\mu}), \tag{2.20}$$

and also, replacing $\boldsymbol{\mu}$ with its maximum likelihood estimate $\overline{\boldsymbol{y}}$ in Equation (2.20), we get following function:

$$l(\boldsymbol{\mu}, \boldsymbol{\Theta}) = \frac{n}{2} log|\boldsymbol{\Theta}| - \frac{1}{2}\sum_{i=1}^{n}(y_i - \overline{\boldsymbol{y}})^t \, \boldsymbol{\Theta} \, (y_i - \overline{\boldsymbol{y}}). \tag{2.21}$$

Since,

$$\sum_{i=1}^{n}(y_i - \overline{\boldsymbol{y}})^t \, \boldsymbol{\Theta} \, (y_i - \overline{\boldsymbol{y}}) = \sum_{i=1}^{n} tr[(y_i - \overline{\boldsymbol{y}})^t \, \boldsymbol{\Theta} \, (y_i - \overline{\boldsymbol{y}})], \tag{2.22}$$

$$= \sum_{i=1}^{n} tr[\boldsymbol{\Theta} \, (y_i - \overline{\boldsymbol{y}})^t (y_i - \overline{\boldsymbol{y}})], \tag{2.23}$$

Equation (2.21) can be rearranged as following equation:

$$l(\boldsymbol{\Theta}) = \frac{n}{2} log|\boldsymbol{\Theta}| - \frac{1}{2}\sum_{i=1}^{n} tr(\boldsymbol{\Theta}\mathbf{S}), \tag{2.24}$$

where the sample variance covariance is $S = s_{ij}$

$$s_{ij} = \frac{1}{n}\sum_{i=1}^{n}\left(y_k^{(i)} - \bar{\boldsymbol{y}}^{(i)}\right)\left(y_k^{(j)} - \bar{\boldsymbol{y}}^{(j)}\right). \tag{2.25}$$

Finally, maximum likelihood estimator of the precision matrix can be obtained by maximizing Equation (2.24) under the constraints on zero entries. One down side of maximum likelihood estimator technique is that it does not tend to estimate a sparse graph. On the other hand, biological networks have very sparse network structure. Therefore, in order to estimate a sparse and also symmetric precision matrix, graphical lasso (GLASSO) with L_1-penalty can be imposed on the entries of the precision matrix and not on the regression coefficient.

In GLASSO, the optimization equation can be described as

$$\max_{\|\boldsymbol{\Theta}\|_1 \leq \lambda} [log|\boldsymbol{\Theta}| - tr(\boldsymbol{\Theta}\mathbf{S})], \tag{2.26}$$

where $\boldsymbol{\lambda}$ is a non-negative tuning parameter and $\|\boldsymbol{\Theta}\|_1$ is equal to $\sum_{\boldsymbol{ij}}|\theta_{ij}|$. Then the dual form of the Lagrange multiplier is applied for optimization problem via

$$\max_{\boldsymbol{\Theta}}[log|\boldsymbol{\Theta}| - tr(\boldsymbol{\Theta}\mathbf{S}) - \boldsymbol{\lambda}\|\boldsymbol{\Theta}\|_1], \tag{2.27}$$

with the non-negative Lagrange multiplier λ. When λ gets larger, the solution gets sparser which means that precision matrix contains the larger number of zero elements, but the lower the related likelihood. This form of optimization allows us to obtain a symmetric and a sparse estimate of the precision matrix. In the R programming, the "*glasso*" method by the *huge* package is applied to infer the graph. The Huge package offers a couple of model selection criteria that can be listed as BIC, AIC, RIC, and StARS to find an optimal penalty value. In this study, I use the StARS

(Liu et al., 2010) criterion to select the optimal model, resulting in choosing the optimal λ while estimating the precision matrix for GGM via GLASSO.

3.2. Gaussian Copula Graphical Models

In statistical analysis, the copula is used when the assumption of normality does not hold for the data while there are correlated measurements. Thus, the problem can be solved by different approaches and one of the efficient solution can be combining the data in such a way that their joint distribution can be partitioning via a separate Gaussian variance-covariance matrix.

Accordingly, let $\mathbf{Y} = \boldsymbol{Y_v}$, $\boldsymbol{V} = \{1, 2, \dots, p\}$ be the set of nodes and $\boldsymbol{E} \subset \boldsymbol{V} \times \boldsymbol{V}$ be the set of existing edges while $(i, j) \in \boldsymbol{E}$. It is assumed that $\boldsymbol{Y_v}$ can be continuous, count, binary and categorical with ordered categories. For ordinal categorical data and binary data, a continuous latent variable Z is introduced. Furthermore, the observed samples are associated with $\boldsymbol{Y_v}$ which can be denoted by $\{y_v^1, y_v^2, \dots, y_v^n\}$. Then, the observed samples from $\boldsymbol{Z_v}$ are presented by $\{z_v^1, z_v^2, \dots, z_v^n\}$. Here, the relationship between $\boldsymbol{Y_v}$ and its substitute $\boldsymbol{Z_v}$ is shown through some increasing thresholds $\tau_v = (\tau_{v,1}, \tau_{v,2}, \dots, \tau_{v,w_v})$. It is set as

$$y_v^j = \sum_{l=1}^{w_v} l \times 1_{\tau_{v,l-1} < z_v^j < \tau_{v,l}}. \tag{2.28}$$

So the approach follows that the relationship between the latent and the observed samples satisfies the constraints below.

$$y_v^{j_1} < y_v^{j_2} \Rightarrow z_v^{j_1} < z_v^{j_2}, \qquad z_v^{j_1} < z_v^{j_2} \Rightarrow y_v^{j_1} \leq y_v^{j_2}, \tag{2.29}$$

for which $1 \leq j_1 \neq j_2 \leq n$.

Moreover, the latent samples $z^{1:n} = (z^1, z^2, \dots, z^n)$ belong to the following set

$$A(y^{1:n}) = \{z^{1:n} \in \boldsymbol{R}^{n\times p} : L_v^j(z^{1:n}) < z_v^j < U_v^j(z^{1:n}\}, \tag{2.30}$$

where $L_v^j(z^{1:n}) = max\{z_v^k : y_v^k < y_v^j\}$ and $U_v^j(z^{1:n}) = min\{z_v^k : y_v^k < y_v^j\}$.

For the precision matrix $\boldsymbol{\Theta}$, the joint distribution of $\boldsymbol{Y} = \boldsymbol{Y_v}$ is modelled as follows (Hoff, 2007):

$$Z_v \sim N_p(0, \boldsymbol{\Theta^{-1}}), \tag{2.31}$$

$$\tilde{Z}_v = \frac{Z_v}{(\boldsymbol{\Theta^{-1}})_{v,v}^{\frac{1}{2}}}, \quad v \in V, \tag{2.32}$$

$$Y_v = F_v^{-1}\left(\Phi\left(\tilde{Z}_v\right)\right), \quad v \in V. \tag{2.33}$$

Here, $\Phi(\cdot)$ stands for the cumulative distribution function of the stanard normal distribution. In addition, $\boldsymbol{F_v}$ stands for the univariate distribution of $\boldsymbol{Y_v}$ and F_v^{-1} denotes the pseudo inverse of $\boldsymbol{F_v}$. The joint distribution of the latent variables is the multivariate normal with $\tilde{Z} = \tilde{Z}_v \sim N_p(0, \Upsilon(\boldsymbol{\Theta}))$ where $\Upsilon(\boldsymbol{\Theta})$ is a correlation matrix with entries

$$\Upsilon_{v_1,v_2}(\boldsymbol{\Theta}) = \frac{(\boldsymbol{\Theta}^{-1})_{v_1,v_2}}{\sqrt{(\boldsymbol{\Theta}^{-1})_{v_1,v_1}(\boldsymbol{\Theta}^{-1})_{v_2,v_2}}}. \tag{2.34}$$

In Equation (2.34), $\boldsymbol{\Theta}_{v_1,v_1}$ and $\boldsymbol{\Theta}_{v_2,v_2}$ stand for the diagonal entries of v_1 and v_2 nodes, respectively. Correspondingly, $\boldsymbol{\Theta}_{v_1,v_2}$ indicates the precision value between v_1 and v_2 nodes.

Therefore, by standing $C(u_1, \dots, u_p \mid \Upsilon)$ as the Gaussian copula matrix with a $p \times p$ dimensional correlation matrix Υ, we have

$$p(Y_1 \le y_1, \dots, Y_p \le y_p) = C(y_1, \dots, y_p \mid \Upsilon(\boldsymbol{\Theta}), F_1, \dots, F_p) \tag{2.35}$$

$$= C\left(F_1(y_1), \dots, F_p(y_p) \,\middle|\, \Upsilon(\boldsymbol{\Theta})\right). \tag{2.36}$$

To avoid making a formal assumption for the parametric representation of the marginal distributions $\{F_v: v \in V\}$ which can be discouraging task for most of the real datasets, their marginal distribution are treated as nuisance parameters. Accordingly, this method focus on $\tilde{Z}_v$ which is the joint distribution of the latent variables and the joint distribution has a relationship with the observed variables $\boldsymbol{Y_v}$ as shown in Equation (2.33) (Dobra & Lenkoski, 2011).

Hoff (2007) suggests the inference in the space of the latent variable via a substitution of the observed data $y^{1:n}$ with the event $\mathfrak{D} = \{z^{1:n} \in A(y^{1:n})\}$. Then the likelihood function is written as

$$p(\, y^{1:n} \mid \boldsymbol{\Theta}, \{F_v: v \in V\}\,) = p(\, \mathfrak{D} \mid \boldsymbol{\Theta}\,) p(\, y^{1:n} \mid \mathfrak{D}, \Upsilon(\boldsymbol{\Theta}), \{F_v: v \in V\}\,). \tag{2.37}$$

For this decomposition, only the likelihood of the observed data part, which is related with the inference on $\boldsymbol{\Theta}$, is indicated by $p(\, \mathfrak{D} \mid \boldsymbol{\Theta}\,)$ and it does not depend on the marginal distributions $\{F_v: v \in V\}$. For the estimation of $p(\, \mathfrak{D} \mid \boldsymbol{\Theta}\,)$, Hoff (2007) constructs a Gibbs sampler with stationary distribution as below.

$$p(\, \boldsymbol{\Theta} \mid \mathfrak{D}\,) \propto p(\, \mathfrak{D} \mid \boldsymbol{\Theta}\,)\, p(\boldsymbol{\Theta}). \tag{2.38}$$

Here, $\boldsymbol{\Theta}$ has a Wishart prior distribution $W_p(b, D)$.

On the other side, the joint posterior distribution of $\boldsymbol{\Theta} \in P_G$ and the graph G are given by

$$p(\, \boldsymbol{\Theta}, G \mid \mathfrak{D}\,) \propto p(\, \mathfrak{D} \mid \boldsymbol{\Theta}\,)\, p(\, \boldsymbol{\Theta} \mid G\,)\, p(G). \tag{2.39}$$

In this expression, G-Wishart $W_G(b, D)$ is the prior distribution of $\boldsymbol{\Theta}$ conditional on G. and the prior distribution is uniform, i.e., $p(G) \propto 1$.

3.2.1. Gaussian Graphical Model Under Reverse Jump Markov Chain Monte Carlo Method

The reverse jump Markov chain Monte Carlo method uses the Cholesky decomposition in order to obtain a positive definite precision matrix due to its advantageous for the precision matrix which considers the G-Wishart prior $W_G(b, D)$ for $\boldsymbol{\Theta}$ with the following density

$$p(\boldsymbol{\Theta} \mid \boldsymbol{G}) = \frac{1}{I_G(b,D)} (det\boldsymbol{\Theta})^{\frac{b-2}{2}} \exp\left\{-\frac{1}{2} tr(\boldsymbol{\Theta}^{\boldsymbol{T}} D)\right\}, \tag{2.40}$$

where G stands for the graphical structure of the data, $I_G(b, D)$ is the normalizing constant with $b > 2$ and D is set to p-dimensional identity matrix, I_p (Dobra & Lenkoski, 2011). If G is complete graph (i.e., no missing edges) with p nodes, $W_G(b, D)$ becomes the Wishart distribution $W_p(b, D)$ and the normalizing constant can be explicitly calculated. However, when G is not complete graph, then it is not decomposable and therefore, the Monte Carlo method should be applied to numerically approximate its normalizing constant.

Furthermore, Lenkoski (2013) applies the sampling algorithm from the G-Wishart distribution. Therefore, the G-Wishart distribution with parameters $b + n$ and $D + U$ is the posterior distribution of $\boldsymbol{\Theta}$ for the given G. Here, U denotes

$$U = \sum_{j=1}^{n} y_j y_j^T, \tag{2.41}$$

that is, the trace of YY^T.

For the joint distribution given in Equation (2.38), the method describes a Markov chain Monte Carlo sampler. Dobra and Lenkoski (2011) use small values for σ_p and σ_g which are the precision parameters, i.e., $\sigma_p = \sigma_g = 0.1$.

Considering the current state of the chain $(\boldsymbol{\Theta}^s, G^s)$, the next state $(\boldsymbol{\Theta}^{s+1}, G^{s+1})$ is generated by the following steps.

Step 1: Resample the latent data.

The latent variable Z is used instead of the observed variable **Y**, if **Y**'s are not normally distributed. Here, Z is an matrix with a dimension of $(n \times p)$. In this step, for each $v \in V$ and $j \in \{1,2,\dots,n\}$, the latent value z_v^j is updated via sampling from its conditional distribution.

$$Z_{V\setminus\{v\}} = z^j_{V\setminus\{v\}} \sim N(\mu_v, \sigma_v^2), \tag{2.42}$$

truncated to the interval $\left[L_v^j, U_v^j\right]$, where

$$\mu_v = -\sum_{v' \in bd_G(v)} \frac{\boldsymbol{\Theta}^s_{v,v'}}{\boldsymbol{\Theta}^s_{v,v}} z^j_{v'} \ , \tag{2.43}$$

and

$$\sigma_v^2 = \frac{1}{\boldsymbol{\Theta}^s_{v,v}} , \tag{2.44}$$

bound for L and U are given in Equation (2.30).

So, at the end, the new value of z_v^j is obtained via sampling from a truncated normal distribution.

Step 2: Resample the precision matrix.

In this stage, the precision matrix is calculated via using the latent variables which are obtained from Step 1 and here, this method also applies the Cholesky decomposition to the precision matrix.

In this stage, ϕ^s shows the upper triangular of the precision matrix $\boldsymbol{\Theta}$ at the current state, i.e., $\boldsymbol{\Theta^s}$. Thereby, the method applies the Metropolis-Hasting update of $\boldsymbol{\Theta^s}$ related with a diagonal element $\phi^s{}_{v_1,v_1} > 0$ by sampling a value γ from

$\mathrm{N}(\phi^{s}{}_{v_1,v_1}, \sigma_p^2)$ which is truncated below at 0. Afterwards, γ is replaced with the related diagonal elements of ϕ^s, denoted as ϕ', with the acceptance probability $\min\{R_p, 1\}$ via

$$R_p = \frac{\Phi\left(\frac{\phi^{s}{}_{v_1,v_1}}{\sigma_p}\right)}{\frac{\gamma}{\sigma_p}} \left(\frac{\gamma}{\phi^{s}{}_{v_1,v_1}}\right)^{b+n+d_{v_1}^{G^s}-1} R_p' \ , \tag{2.45}$$

where

$$R_p' = exp\left\{-\frac{1}{2} tr(\boldsymbol{\Theta}' - \boldsymbol{\Theta}^s)^T (D + tr(Z^T Z)\right\}. \tag{2.46}$$

In this equation, $\boldsymbol{\Theta}' = (\phi')^T \phi'$ and the candidate matrix $\boldsymbol{\Theta}'$ are accepted with a probability $\min\{R_p', 1\}$. Finally, the precision matrix obtained via performing the Metropolis-Hasting update is taken as $\boldsymbol{\Theta}^{s+1/2} \in P_{G^s}$.

Step 3: Resample the graph.

The Cholesky decomposition for $\boldsymbol{\Theta}^{s+1} = (\phi^{s+1/2})^T \phi^{s+1/2}$, where $\phi^{s+1/2}$ is the upper triangular. If there is no edge between v_1 and v_2 in G^s, one can update the system by adding this edge to the current graph G^s so that a candidate graph G' can be obtained. Accordingly, the candidate precision matrix can be found by $\boldsymbol{\Theta}' = (\phi')^T \phi' \in P_{G'}$. Since the parameter space dimension increases by one, the reverse jump Markov chains methodology by Green (1995) is used. Hence, the update of $(\boldsymbol{\Theta}^{s+1/2}, G^s)$ to $(\boldsymbol{\Theta}', G')$ with a probability $\min\{R_g, 1\}$, where R_g is below, can be perfrmed.

$$\sigma_g \sqrt{2\pi}\, \phi_{v_1,v_1}^{s+1/2} \frac{I_{G^s}(b, \boldsymbol{D})}{I_{G'}(b, \boldsymbol{D})} \times$$
$$\times exp\left\{-\frac{1}{2} tr\left((\boldsymbol{\Theta}' - \boldsymbol{\Theta})^T \left(\boldsymbol{D} + tr(\boldsymbol{Z^T Z})\right)\right) + \frac{(\phi'_{v_1,v_1} - \phi_{v_1,v_1}^{s+1/2})^2}{2\sigma_g^2}\right\}. \tag{2.47}$$

If there is an edge between v_1 and v_2 in G^s, the system is updated by deleting this edge from the current graph G^s so that a candidate graph G'can be obtained. Like above, taking $\boldsymbol{\Theta}' = (\phi')^T\phi' \in P_{G'}$ and increasing the dimension of parameter space by one from $\phi^{s+1/2}$ to ϕ, one can lastly compute the acceptance probability of the update from $(\boldsymbol{\Theta}^{s+1/2}, G^s)$ to $(\boldsymbol{\Theta}', G')$ with a probability $\min\{R'_g, 1\}$ where R'_g is as follows.

$$\left(\sigma_g\sqrt{2\pi}\,\phi_{v_1,v_1}^{s+1/2}\right)^{-1}\frac{I_{G^s}(b,\boldsymbol{D})}{I_{G'}(b,\boldsymbol{D})}\times \\ \times exp\left\{-\frac{1}{2}tr\left((\boldsymbol{\Theta}'-\boldsymbol{\Theta})^T\left(\boldsymbol{D}+tr(\boldsymbol{Z^T Z})\right)\right)+\frac{(\phi'_{v_1,v_1}-\phi_{v_1,v_1}^{s+1/2})^2}{2\sigma_g^2}\right\} \tag{2.48}$$

Finally, the updated graph G^{s+1} and the related precision matrix $\boldsymbol{\Theta}^{s+1}$ are found at the end of Step 3.

3.2.2. Gaussian Graphical Model Under Birth-and-Death Markov Chain Monte Carlo Method

As given in Chapter 3.2, let $V = \{1,2,\dots,p\}$ be the set of nodes and $E \subset V \times V$ be the set of existing edges and $(i,j) \in E$ and $\bar{E}$ as the set of non-existing edges. Accordingly, $G = (V,E)$ denotes an undirected graph. This time, let show the independent and identically distributed sample as $\boldsymbol{y} = \{y^1, y^2, \dots, y^n\}$. Then, the likelihood can also be written as follows (Wit & Mohammadi, 2015).

$$p(\,\boldsymbol{y}\mid\boldsymbol{\Theta},G\,) \propto |\boldsymbol{\Theta}|^{\frac{n}{2}}exp\left\{-\frac{1}{2}tr(\boldsymbol{\Theta S})\right\}, \tag{2.49}$$

where $\boldsymbol{S} = \boldsymbol{y}'\boldsymbol{y}$. The joint distribution is formulated as Equation (2.40). Thus, the posterior distribution of $\boldsymbol{\Theta}$ is

$$p(\boldsymbol{\Theta} \mid \boldsymbol{y}, G) = \frac{1}{I_G(b^*, D^*)} (det\boldsymbol{\Theta})^{\frac{b^*-2}{2}} \exp\left\{-\frac{1}{2} tr(\boldsymbol{\Theta}^{\boldsymbol{T}} D^*)\right\}, \qquad (2.50)$$

where $b^* = b + n$ and $D^* = D + \boldsymbol{S}$, so it is G-Wishart distribution, $W_G(b^*, D^*)$.

Hereby, this joint posterior distribution is described with a trans-dimensional Markov chain Monte Carlo (MCMC) sampler scheme which is the birth-death (MCMC) (BDMCMC) method. This process search over the graph space, *G,* by adding and removing as a birth and death event. The birth and death rates determined by the stationary distribution of the process occurs in a continuous time.

Let the birth and death process at time t be denoted by the state $(G, \boldsymbol{\Theta})$, So the method considers the following continuous time process.

Death:

By a poisson process with a rate of $\delta_e(\boldsymbol{\Theta})$, each edge $e \in E$ dies independently on the others. Therefore, the overall death rate is $\delta(\boldsymbol{\Theta}) = \sum_{e \in E} \delta_e(\boldsymbol{\Theta})$. If a death of an edge occurs, i.e., $e = (i, j) \in E$, the process jumps to a new state which is show as $(G^{-e}, \boldsymbol{\Theta}^{-\boldsymbol{e}})$, where the undirected graph becomes $G^{-e} = (V, E \setminus \{e\})$, and the precision matrix is $\boldsymbol{\Theta}^{-\boldsymbol{e}} \in \boldsymbol{P}_{G^{-e}}$. Only the difference between the precision matrix $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}^{-\boldsymbol{e}}$ is the entries corresponds to $\{(i, j), (j, i), (j, j)\}$ (recall by definition of an edge $i < j$).

Birth:

By a poisson process with a rate of $\beta_e(\boldsymbol{\Theta})$, each edge $e \in \bar{E}$ is born independently on the others. Thus, overall birth rate is $\beta(\boldsymbol{\Theta}) = \sum_{e \in E} \beta_e(\boldsymbol{\Theta})$. If a birth of an edge occurs, i.e., $e = (i, j) \in \bar{E}$, the process jumps to a new state which is show as $(G^{+e}, \boldsymbol{\Theta}^{+\boldsymbol{e}})$ where the undirected graph becomes $G^{+e} = (V, E \cup \{e\})$, and the precision matrix is $\boldsymbol{\Theta}^{+\boldsymbol{e}} \in \boldsymbol{P}_{G^{+e}}$. Only difference between the precision matrix $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}^{-\boldsymbol{e}}$ is the entries corresponds to $\{(i, j), (j, i), (j, j)\}$.

$$P(birth\ for\ the\ edge\ e) = \frac{\beta_e(\boldsymbol{\Theta})}{\beta(\boldsymbol{\Theta}) + \delta(\boldsymbol{\Theta})}, \quad for\ each\ e \in \bar{E}, \qquad (2.51)$$

$$P(death\ for\ the\ edge\ e) = \frac{\delta_e(\boldsymbol{\Theta})}{\beta(\boldsymbol{\Theta}) + \delta(\boldsymbol{\Theta})}, \quad for\ each\ e \in E, \tag{2.52}$$

Wit and Mohammadi (2015) propose a BDMCMC algorithm which is based on the choice of the birth and the death rates specifically. The birth and death rates are considered as follows.

$$\beta_e(\boldsymbol{\Theta}) = \frac{P(G^{+e}, \boldsymbol{\Theta}^{+e} \backslash (k_{ij}, k_{jj}) \mid \boldsymbol{y})}{P(G, \boldsymbol{\Theta} \backslash k_{jj}) \mid \boldsymbol{y})}\ for\ each\ e \in \bar{E}, \tag{2.53}$$

$$\delta_e(\boldsymbol{\Theta}) = \frac{P(G^{-e}, \boldsymbol{\Theta}^{-e} \backslash\ k_{jj} \mid \boldsymbol{y})}{P(G, \boldsymbol{\Theta} \backslash (k_{ij}, k_{jj}) \mid \boldsymbol{y})}\ for\ each\ e \in E. \tag{2.54}$$

Based on the rates given above and a given graph *G* as well as and the precision matrix $\boldsymbol{\Theta}$, BDMCMC algorithm iterates the following steps whose mathematical details are presented in the following part.

Step 1. Process of birth and death

1.1 Calculate the birth rates by Equation (2.53) and $\beta(\boldsymbol{\Theta}) = \sum_{e \in E} \beta_e(\boldsymbol{\Theta})$,

1.2 Calculate the death rates by Equation (2.54) and $\delta(\boldsymbol{\Theta}) = \sum_{e \in E} \delta_e(\boldsymbol{\Theta})$,

1.3 Simulate the jump type (birth or death) by Equation (2.51) and Equation (2.52)

Step 2. According to jump type, sampling from the new precision matrix.

Step 1: Computing the Birth and Death Rates

In this step, the method calculates the rates. However, since both the birth and the death rates are computed with the same manner, only the calculation of the death rates are given as below.

The death rates' numerator (see Equation (2.54)) for each $e \in E$ is

$$P(P(G^{-e}, \boldsymbol{\Theta}^{-e}\backslash(k_{ij}, k_{jj}) \mid \boldsymbol{y})) = \frac{P\left(G^{-e}, \boldsymbol{\Theta}^{-e}\backslash k_{jj} \mid \boldsymbol{y}\right)}{P\left(k_{jj} \mid \boldsymbol{\Theta}^{-e}\backslash k_{jj}, G^{-e}, \boldsymbol{y}\right)}. \quad (2.55)$$

Following is obtained by Wang and Li (2012) and after some simplification,

$$P(G^{-e}, \boldsymbol{\Theta}^{-e}\backslash(k_{ij}, k_{jj}) \mid \boldsymbol{y}) = \frac{P(G)}{P(\boldsymbol{y})} \frac{I\left(b^*, D_{jj}^*\right)}{I_{G^{-e}}} \left|\boldsymbol{\Theta}_{V\backslash j, V\backslash j}^0\right|^{\frac{b^*-2}{2}} exp\left\{-\frac{1}{2} tr(\boldsymbol{\Theta^0} D^*)\right\}, \quad (2.56)$$

where $\boldsymbol{\Theta}^0 = \boldsymbol{\Theta}$ with the exception of an entry 0 in positions *(i, j)* and *(j, i)* and an entry c in the position *(j, j)* and $I\left(b^*, D_{jj}^*\right)$ denotes the normalizing constant of the G- Wishart distribution for $\mathrm{p} = 1$.

For the denominator of Equation (2.54), the following expression is taken.

$$P(G, \boldsymbol{\Theta}\backslash(k_{ij}, k_{jj}) \mid \boldsymbol{y}) = \frac{P\left(G, \boldsymbol{\Theta} \mid \boldsymbol{y}\right)}{P((k_{ij}, k_{jj}) \mid \boldsymbol{\Theta}\backslash(k_{ij}, k_{jj}), G, \boldsymbol{y})}, \quad (2.57)$$

By using the expression obtained via Wang and Li (2012) and after some simplification,

$$P\left(K\backslash(k_{ij}, k_{jj}), g \mid \boldsymbol{y}\right) = \frac{P(G)}{P(\boldsymbol{y})} \frac{J(b^*, D_{ee}^*, \boldsymbol{\Theta})}{I_G(b, D)} \left|\boldsymbol{\Theta}_{V\backslash e, V\backslash e}^1\right|^{\frac{b^*-2}{2}} exp\left\{-\frac{1}{2} tr(\boldsymbol{\Theta^1} D^*)\right\} \quad (2.58)$$

where

$$J(b^*, D_{ee}^*, \boldsymbol{\Theta}) = \left(\frac{2\pi}{D_{jj}^*}\right)^{\frac{1}{2}} I\left(b^*, D_{jj}^*\right)(k_{ii} - k_{ii}^1)^{\frac{b^*-2}{2}} exp\left\{-\frac{1}{2}\left(D_{ii}^* - \frac{{D_{ij}^*}^2}{D_{jj}^*}\right)(k_{ii} - k_{ii}^1)\right\}. \quad (2.59)$$

Herein, $\boldsymbol{\Theta}^1 = \boldsymbol{\Theta}$ with the exception of an entries $\boldsymbol{\Theta}_{e,V\backslash e}(\boldsymbol{\Theta}_{V\backslash e,V\backslash e})^{-1}\boldsymbol{\Theta}_{V\backslash e,e}$ in the positions corresponding to $e = (i, j)$.

By plugging Equation (2.56) and (2.58) into the death rate, and the equation below can be found.

$$\delta_e(\boldsymbol{\Theta}) = \frac{P(G^{-e})}{P(G)} \frac{I_G(b,D)}{I_{G^{-e}}(b,D)} H(\boldsymbol{\Theta}, D^*, e), \tag{2.60}$$

where $H(\cdot)$ stands for

$$H(\boldsymbol{\Theta}, D^*, e) = \left(\frac{D^*_{jj}}{2\pi(k_{ii}-k^1_{ii})}\right)^{1/2} \times \\ \times exp\left\{-\frac{1}{2}\left[tr\left(D^*(\boldsymbol{\Theta}^0 - \boldsymbol{\Theta}^1)\right) - \left(D^*_{ii} - \frac{D^{*\,2}_{ij}}{D^*_{jj}}\right)(k_{ii}-k^1_{ii})\right]\right\}. \tag{2.61}$$

Furthermore, let suppose $(G, \boldsymbol{\Theta})$ is the current state of the algorithm. So, in order to calculate death rates via Equation (2.60), first $\widetilde{\boldsymbol{\Theta}}$ is sampled from $W_G(b,D)$ by algorithm below (see step 2). Then, the death rates are replaced with

$$\delta_e(\boldsymbol{\Theta}) = \frac{P(G^{-e})}{P(G)} \frac{H(\boldsymbol{\Theta}, D^*, e),}{H\left(\widetilde{\boldsymbol{\Theta}}, D, e\right),}, \tag{2.62}$$

Step 2: Direct Sampler from Precision Matrix

Lenkoski (2013) presents an exact sampler method as follows where the precision matrix is $\boldsymbol{\Theta}$ and $\Sigma = \boldsymbol{\Theta}^{-1}$ while the graph $G = (V, E)$.

Step 1. Set $\Omega = \Sigma$.

Step 2. Until the converges, repeat for $i = 1, \ldots, p$

2.1 Let $N_i \subset V$ be the set of node *i* in graph *G*. Solve

$$\hat{\beta}^*_i = \Omega^{-1}_{N_i} \Sigma_{N_i,i} \ , \tag{2.63}$$

by forming $\Sigma_{N_i,i}$ and Ω_{N_i}.

2.2 By padding the $\hat{\beta}^*_i$ elements to the appropriate places and zeroes in those places not connected to i in the graph G, form $\hat{\beta}_i \in R^{p-1}$.

2.3 By $\Omega_{-i,-i}\hat{\beta}_i$, update $\Omega_{i,-i}$ and $\Omega_{-i,i}$.

Step 3. Return $\boldsymbol{\Theta} = \Omega^{-1}$.

3.3. Loop-based Multivariate Adaptive Regression Splines

The multivariate adaptive regression Splines (MARS) is a nonparametric regression modelling technique which does not use an assumption between dependent and independent variables. The method has a procedure to reduce the complexity of nonlinear functions by constructing the piecewise linear functions. Such smoothing has the calculation which implements a two stage procedure, called as the forward stage and the backward stage. In the forward stage, model starts by adding the intercept term to the model and inserts the basis functions (BFs), iteratively. The procedure ends up with the largest model that includes many basis functions. On the other hand, the backward stage reduces the complexity of the model by removing the BFs resulting in a slight increase in the residual sum of squared error.

Suppose the parametric model as follows.

$$y_i = f(\beta, x_i') + \varepsilon_i, \tag{2.64}$$

where β is the parameters of the vector and x_i' $(i = 1, \dots, n)$ denotes the predictors vector for the *i*th case, ε_i $(i = 1, \dots, n)$ stands for the vector of the random error and n is the total number of observations. Finally, y_i shows the associated response vector.

Furthermore, the fundamental element of MARS has a form where function f consists of the piecewise linear BFs given below (for $x \in R$)

$$[x-t]_+ = \begin{cases} x-t, & x > t \\ 0, & otherwise \end{cases}, and \quad [t-x]_+ = \begin{cases} t-x, & x < t \\ 0, & otherwise \end{cases}. \tag{2.65}$$

In the Equation (2.65), t stands for the knot and represents the breaking point of the spline function, as illustrated in Figure 3.1. The goal is obtain the reflected pairs of each x_j with the knots at each observed x_{ij}. Therefore, the set of all basis functions is random variables which can be defined as follows.

$$C = \{ [x - t]_+, [t - x]_+ \mid t \in \{x_{1,j}, x_{2,j}, \dots, x_{N,j}\}, j = 1,2, \dots, p \}, \qquad (2.66)$$

where N represents the number of observations and p is the number of independent variables. Thus, $2Np$ numbers of basis function exists if all the input variables are different.

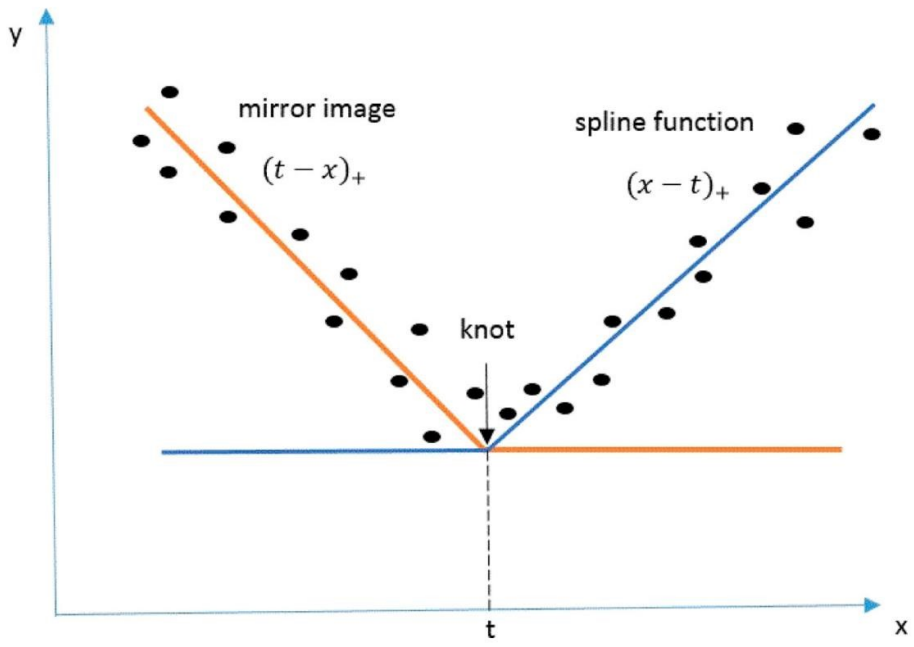


Figure 3.1. A Simple Representation of the MARS via a Knot and Splines

The model building strategy of MARS consists of the forward and backward elimination and the functions from the C set is used, rather than the original variables. Thus, the model can be described as

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(x) + \varepsilon, \tag{2.67}$$

in which $h_m(x)$ refers to the spline basis functions and M is the total number of parameters, and finally β_0 and β_m refer to the intercept term and the regression coefficient, respectively. β_m is estimated via minimizing the residual sum of squares by a linear regression given a choice of h_m. Furthermore, as mentioned beforehand, at the end of the forward stage, a similar large model as in Equation (2.67) is obtained. However, this model may overfit the data. Hence, the backward scheme is applied by removing the term which induces the smallest increase in the residual squares. Then, the best model denoted by $\hat{f}_\lambda$ is obtained with the λ number of terms. The best model in MARS is obtained by getting the optimal λ. To do so, the generalized cross validation value (GCV) is used and GCV is presented as follows

$$GCV(\lambda) = \frac{\sum_{i=1}^{N}(y_i - \hat{f}_\lambda\,(x_i)^2}{\left(1 - \frac{M(\lambda)}{N}\right)^2}, \tag{2.68}$$

where $M(\lambda)$ is the effective number of parameters and it can be obtained with the equation $M(\lambda) = r + cK$ where r represents the number of linearly independent BFs and K stands for the number selected knots during the forward stage. Finally, c indicates the cost for the optimization of the basis function and also the smoothing parameter of the model. Generally, this model is equated to $c = 3$, however when the model is an additive model, it is set to $c = 2$.

Hereby, one can construct the model for MARS similar to the lasso regression as it is used in GGM in Equation (2.13). By doing so, we model each node against all the remaining nodes. By this way, one can detect a selected gene's links with the other remaining genes. In this study, the model which contains only the main effects is the LMARS without interaction and the model with also second-order interaction terms, is named as the LMARS with interaction terms.

3.4. Threshold Selection Methods

To obtain a system's graphical representation and evaluate the network model, the numerical entries of the precision matrix, $\boldsymbol{\Theta}$, should be converted into a binary representation via a threshold value. In this representation, "0" stands for no interaction between two entities of the network system, whereas, "1" implies the functional or physical interaction between those two entities. So as to convert the real-valued $\boldsymbol{\Theta}$ to a binary form, there are different threshold-determining approaches in the literature. In this study, the most common ones which are the kappa maximized threshold criterion, maximized sum threshold, minimized difference threshold and 0.5 value criterion are presented. These methods are known and widely used either because of their high accuracy measures or for their conveniences in the application. Other than these methods, we present the novel nonparametric threshold selection criterion which consider the underlying topology of the network system.

3.4.1. Kappa Maximized Threshold Criterion

The kappa maximization threshold approach (KMT) is an approach which maximizes the kappa statistics. (Guisan et al., 1998).

Kappa maximized threshold criterion (KMT) calculates kappa scores via

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{2.69}$$

where p_o stands for the observed percentage of the agreement (accuracy), and p_e denotes the expected agreement probability due to the chance are calculated by

$$p_o = \frac{TP + TN}{n}, \tag{2.70}$$

$$p_e = \frac{(TP + FN) * (TP + FP) + (TN + FN) * (TN + FP)}{n^2}. \quad (2.71)$$

In these expression, the meaning of abbreviation can be listed as follows:

True Positive (TP): The number of the agreement on the presence of the edge, i.e., 1.

True Negative (TN): The number of the agreement on the absence of the edge, i.e., 0.

False Positive (FP): The number of incorrect prediction of the actually absence of the edge.

False Negative (FN): The number of incorrect prediction of the actually presence of the edge.

In Table 3.1, the meaning of these entries are also shown via confusion matrix.

Furthermore, the procedure which can select the optimal kappa statistic involves series of calculations. The method calculates a kappa value for each 100-threshold value and the one that provides the maximum kappa statistics is accepted as the threshold (Guisan et al., 1998; Thuiller, 2003; Jiménez-Valverde & Lobo, 2007).

For the application of KMT criterion, in this study, the 100 candidate threshold values which are between the minimum and the maximum entries of the precision matrix are calculated via an increment and the increment is calculated with following formula.

$$increment = \frac{max(entries\ of\ S^{-1}) - min(entries\ of\ S^{-1})}{100}. \quad (2.72)$$

Thereon, the kappa statistics are calculated for each 100 candidate thresholds and the one which provides the maximum kappa value is selected as the optimal threshold.

3.4.2. Maximized Sum Threshold Criterion

In maximized sum threshold (MST) (Manel et al., 2001), it first calculates the sum of the specificity and the sensitivity which are calculated as

$$Specificity = \frac{TN}{TN + FP}, \tag{2.73}$$

$$Sensitivity = \frac{TP}{TP + FN}. \tag{2.74}$$

Here, the specificity calculates the proportion of actually non-existing edges that are correctly identified, whereas the sensitivity computes the proportion of actually existing edges that are correctly identified.

After calculating the sum of specificity and sensitivity for each 100 candidate threshold value which is calculated with the same manner as in the KMT criterion, MST selects the one which maximize the sum as the optimal threshold (Jiménez-Valverde & Lobo, 2007).

3.4.3. Minimized Difference Threshold Criterion

The minimized difference threshold (MDT) criterion (Jiménez-Valverde & Lobo, 2007) calculates the difference between the specificity and the sensitivity for again each 100 candidate thresholds which is calculated as in the previous two methods; KMT criterion and MST.

Accordingly, this method selects the one which has the minimum difference between the specificity and the sensitivity (considering their absolute values) as its optimal threshold value (Jiménez-Valverde & Lobo, 2007).

3.4.4. 0.5T Criterion

This method is considered as the subjective approach in the literature since it neither uses a specific index such as kappa nor the trade-off between two properties which are conflicted such as the sensitivity and the specificity (Liu et al., 2005). The optimal threshold is selected as 0.5. Even if this method is widely preferred (Jiménez-Valverde & Lobo, 2007), the outcome is actually biased since it fails to represent the nature of the data (Liu et al., 2005). However, it may be robust if we both consider situations such as having more cancer patients in the data or less cancer patients in the data.

3.4.5. Proposed Threshold Selection Criteria

Rather than using a specific index or a trade-off between two properties, this method makes use of the topology of the network. To do so, in order to select the optimal threshold value, we use a very important feature of the network which is the sparsity level that changes from one network class to another.

The steps of the procedure are as follows:

Step 1: Calculate the sparsity percentages specific to the network classes.

Step 2: Construct GGM model via GLASSO and obtain the precision matrix.

Step 3: Sort the entries of the precision matrix from the smallest to the largest.

Step 4: Find the threshold value by imposing the sparsity level to the vector which has the sorted entries of the precision matrix.

Step 5: Construct the adjacency matrix via imposing the threshold value to the precision matrix.

Step 6: Obtain the accuracy measures by using the adjacency matrix found in the previous step.

For Step 1, by using *huge* package in the R programming, the sparsity percentages of the random and scale-free networks' mean are calculated under 300 Monte Carlo runs via generating random and scale-free networks. On the other hand, from our preliminary analyses, it is observed that the dimension is another network characteristic that can affect the sparsity percentage of the system. Therefore, the dimensions are also considered when calculating the sparsity percentages. In the application part, it needs to be mentioned that the structure of the random network is considered for the gynecological datasets with the dimension of eleven genes since our quasi network has a complete graph structure. On the other side, when considering 1000 genes' system for real gynecological datasets, we consider their network structures as scale-free. For the random network under a dimension of eleven genes, the mean sparsity percentage is found via the Monte Carlo runs and found as 0.745. On the other hand, since it is computationally demanding to calculate a mean sparsity percentage for the dimension of 1000 genes under the scale-free network structure, we merely calculated the mean sparsity percentage under the 300 Monte Carlo runs and obtain as 0.961. Additionally, in the simulation part, we consider the dimension of the system as 20, 50 and 100 for both random and scale-free structure. Therefore, we calculate the mean sparsity percentage of the random network under the 300 Monte Carlo runs for the dimensions of the system 20, 50 and 100 as 0.854, 0.942 and 0.97, respectively. Then, we also calculate the mean sparsity percentage of the scale-free network under the 300 Monte Carlo runs for the dimensions of the system 20, 50 and 100 as 0.905, 0.961 and 0.98, respectively.

In Step 3, we sort the precision matrix entries, which is obtained in Step 2, from smallest to the largest by creating a vector. Then, in Step 4, by multiplying the length of the vector with the sparsity percentage, we get the vector index corresponding to the cut-off place that will satisfy the sparsity percentage of the underlying network. Thus, we get the element of the vector which denotes the vector index. The element

that we obtain from the vector is simply one of the entries of the precision matrix and it also corresponds to our optimal threshold value that satisfies the sparsity percentage of the underlying network structure which imposes to the precision matrix.

Hereby, in Step 5, by using the threshold value which optimize the precision matrix' sparsity level, we convert the precision matrix into the adjacency matrix in the same manner as applied in the literature. That is, the entries which exceed the threshold value are converted into "1" which stands for an interaction between two nodes, and the entries below the threshold value are converted into "0" which represents no interaction between two nodes. In Step 6, the accuracy measures are calculated by using the adjacency matrix.

3.5. Measure of Accuracy

There are different accuracy measures which can be used in the literature in order to evaluate the performances of the models or methods. In this study, accuracy, F-measure and Matthew correlation coefficient (MCC) are used for the evaluation of the performances of the methods and models. When evaluated together, each of the method gives a different point of view about how accurate the performances are. For example, the accuracy considers the correctly classified objects which are true positive (TP) and true negative (TN). Therefore, when the number of true negative is high, the accuracy value tends to be high too. On the other side, the F-measure does not take into account the correctly classified objects with no interactions, i.e., true negatives (TNs). As a result, that helps us to assess threshold criteria with another point of view. Herein, F-measure is between 0 and 1, and if F-measure is closer to 1 is the better. The same evaluation is also valid for the accuracy. Finally, although it is not applicable for some cases in our analysis, MCC value is another measurement that combines the true positive (TP), true negative (TN), false positive (FP), false negative (FN). Therefore,

it is an important accuracy measure as well. When MCC is closer to 1, it is interpreted as high accuracy. By contrast, when it is closer to -1, it is interpreted as low accuracy.

In Table 3.1, the representation of the actual and the predicted classes are given.

Table 3.1. The General Confusion Matrix

		ACTUAL CLASS	
		Positive	**Negative**
PREDICTED CLASS	**Positive**	TP	FP
	Negative	FN	TN

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2.75}$$

$$F - measure = 2\frac{precision \times recall}{precision + recall}, \tag{2.76}$$

where

$$precision = \frac{TP}{TP + FP} \quad and \quad recall = \frac{TP}{TP + FN}, \tag{2.77}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{2.78}$$

CHAPTER 4

APPLICATION

In this chapter, the procedure of the data collection and the application of real data via network models are presented. Accordingly, we first describe the description of each dataset. Then, we evaluate the model performance for eleven core genes by accuracy measures, and afterwards, we extend our study and use 1000 genes with our core eleven genes in each dataset. Within these augmented datasets, the accuracy of the links is still evaluated for the same eleven core genes regardless of the dimension of the system. In the third part, we initially apply threshold methods under 1000 Monte Carlo runs via simulated data. Finally, we implement them in the analyses via real datasets under different dimensions.

4.1. Dataset Collection

At first step of the data collection, the ArrayExpress database was searched by gathering all studies about gynecological cancers as well as ovarian, cervical, endometrial, vulvar and vaginal cancers, separately. Here, our aim was to collect datasets which include our eleven core genes mentioned in Chapter 2. However, there were some constraints to consider when selecting datasets. In this searching process, we applied two major constraints: *i)* Datasets should be normalized with RMA or MAS5.0 techniques so that they can be comparable. *ii)* The datasets should be published between 2008-2018, so that recent studies are considered.

However, in this type of searching, one can face with certain difficulties. Initially, most of published works have the datasets from their own laboratories with their own

gene numbering without any information about corresponding gene symbols. For instance, one study implements an arbitrary number like "7984319" for a gene symbol, he/she cannot present a supplementary text file which gives corresponding probe set or gene symbols related with those arbitrary numbers. Therefore, majority of the data cannot be comparable even though their associated laboratories work on the same genes. On the other hand, some of the datasets do not have any published articles, resulting in no supplementary text files attached to their measurements so that the researcher can follow the condition (i.e., treatment and control group) as well as design of the data.

Another difficulty is that even if the laboratory includes gene symbols or probe set names, it uses gene aliases in the measurements. In other words, there exists more than one symbol for a gene. Therefore, for the datasets that include directly gene symbols in their data files, the researcher has to check all gene aliases related with his/her gene list. Hence, in our work, we checked our core eleven genes with their gene aliases in the studies where this information is available. However, not all datasets include all of our eleven core genes.

Furthermore, for the studies that used arbitrary numbering in their datasets and have probe set names in their supplementary files which have correspondence to these arbitrary numbers, as described above, we first looked for our core genes' probe set names from the literature and found the datasets which cover those eleven core genes through their supplementary files.

As a result, it has been detected that there are a lot of constraints which hinder the researcher to find suitable datasets by using free database. Because of this reason, after our comprehensive searching process in more than 1000 datasets in ArrayExpress, we merely obtained three reliable datasets. Below, application of methods are presented with selected datasets.

4.2. Real Data Application via Network Models

Here, we have three datasets from ArrayExpress database. Datasets' feature such as cancer types which are included in the studies are also given in Table 4.1. In the analyses, we see that, E-GEOD-63678 dataset has higher accuracies compare to other datasets since it includes more types of gynecological cancers.

Table 4.1. Real Datasets

Datasets	Cancer Type	Data Source
E-GEOD-9891	Ovarian tumor	https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-9891/
E-GEOD-14764	Ovarian cancer	https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-14764/
E-GEOD-63678	Cervical, Endometrial, Vulvar	https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-63678/

4.2.1. Application of E-GEOD-9891 Data

Data Description

This dataset has the transcription profiling of 285 human ovarian tumors. The data is a cohort of 285 patients with the epithelial ovarian, primary peritoneal, or fallopian tube cancer, diagnosed between the years 1992 and 2006. They are identified through Australian Ovarian Cancer Study (sample size n = 206), Royal Brisbane Hospital (n = 22), Westmead Hospital (n = 54) and Netherlands Cancer Institute (NKI-AVL; n = 3) (Tothill et al., 2008). In this E-GEOD-9891, the arrays are designed by randomly selected samples from the Australian Ovarian Cancer Study whose expression profiles

on the Affymetrix U133-plus2 platform aim to identify novel subtypes of the ovarian tumor by the gene expression profiling with a linkage to clinical and pathologic features (Tothill et al., 2008).

In Table 4.2, the findings of our accuracy measures from eleven core genes in the E-GEOD-9891 dataset are presented. As seen in Table 4.2, best accuracy is obtained for LMARS with interaction under nonparametric models and GCGM via RJMCMC approach under parametric models. Moreover, GGM has the lowest accuracy to detect the true network system from the E-GEOD-9891 data. The reason for GGM has such lower accuracy compared to GCGM via RJMCMC and GCGM via BDMCMC is that GCGM makes inference by modelling multivariate associations separately from the observed variables' univariate distributions.

Table 4.2. The Accuracy Table for E-GEOD-9891 Data

Methods	**F-Measure**	**Accuracy**
GGM	0.167	0.091
GCGM via RJMCMC	0.846	0.733
GCGM via BDMCMC	0.448	0.298
LMARS without Interaction	0.752	0.603
LMARS with Interaction	0.858	0.752

Table 4.3. The Accuracy Table for E-GEOD-9891 Data under 1000 Genes' System

Methods	**F-Measure**	**Accuracy**
GGM	0.271	0.157
LMARS without Interaction	0.246	0.140
LMARS with Interaction	0.271	0.157

On the other hand, in Table 4.3, we extend our study and use 1000 genes which also includes our core eleven genes and evaluate the accuracy for the eleven genes only. Table 4.3 shows that nonparametric model LMARS with interaction and parametric model GGM have the same accuracies for the detection of network systems when the dimension increases. However, when the dimension is augmented, GCGM via RJMCMC and BDMCMC estimations are discarded since their procedure become computationally infeasible. The underlying analyses take five days.

4.2.2. Application of E-GEOD-63678 Data

Data Description

This dataset presents the gene expression data from vulvar, cervical, endometrial and carcinoma tissues. In this set, 35 samples that are used to identify potential biomarkers and signatures in each type of cancer. Specifically, 18 cancer samples with 5 cervical, 7 endometrial and 6 vulvar cancers, and also 17 normal samples with 5 cervical, 5 endometrial and 7 vulvar cancers are hybridized on the Affymetrix platform in order to identify the common features among cancer types, embryonic stem cells and the newly discovered cell population of the squamocolumnar junction of the cervix, considered to host the early cancer events (Pappa et al., 2015). Moreover, total RNAs are extracted from physiological and cancer patients from cervix, endometrium and vulvar tissues and are hybridized on the Affymetrix HG133-A-2.0 microarray chips corresponding to more than 12.000 uniquely represented genes (Pappa et al., 2015). Also, this data is recently used to investigate the gynecological cancer types in the work of Liu et al (2019). They try to find hub genes that may serve as biological markers for three types of gynecological cancer.

Here, Table 4.4 shows that GCGM via RJMCMC has the best accuracy measures comparing to others. GCGM via BDMCMC also performs well. On the other hand, GGM is still the worst method to detect existing links.

Table 4.4. The Accuracy Table for E-GEOD-63678 Data

Methods	**F-Measure**	**Accuracy**
GGM	0.167	0.091
GCGM via RJMCMC	0.981	0.791
GCGM via BDMCMC	0.964	0.655
LMARS without Interaction	0.726	0.570
LMARS with Interaction	0.778	0.636

Accordingly, Table 4.5 gives the accuracy for eleven genes under 1000 genes' system. From the tabulated values, it is seen that LMARS without interaction produces the best accuracies with respect to other approaches. Lastly, similar to previous analyses, GCGM with both estimation methods cannot be applicable due to the serious computational demands.

Table 4.5. The Accuracy Table for E-GEOD-63678 Data under 1000 Genes' System

Methods	**F-Measure**	**Accuracy**
GGM	0.167	0.091
LMARS without Interaction	0.193	0.107
LMARS with Interaction	0.167	0.091

4.2.3. Application of E-GEOD-14764 Data

Data Description

This dataset describes the prognostic gene expression in the ovarian cancer. The data have a cohort of 80 ovarian carcinomas (TOC cohort) for the development of a predictive model, which is then evaluated in an entirely independent cohort of 118 carcinomas (Duke cohort) (Denkert et al., 2009). In this dataset, RNAs from 80 frozen ovarian cancer samples are hybridized on the Affymetrix Human Genome U133A Array and the collected data contain our eleven core genes. In the data collection, it is aimed to investigate the hypothesis that molecular markers are able to predict outcome of the ovarian cancer independently from classical clinical predictors, and these molecular markers can be validated by using independent datasets (Denkert et al., 2009).

For this datasets, Table 4.6 shows that GCGM gives significantly the best accuracy measures regarding other models and LMARS with interaction has the second best performance. GGM, however, fails to correctly estimate links between eleven core genes compare to other methods.

Table 4.6. The Accuracy Table for E-GEOD-14764 Data

Methods	F-Measure	Accuracy
GGM	0.167	0.091
GCGM via RJMCMC	0.911	0.836
GCGM via BDMCMC	0.226	0.127
LMARS without Interaction	0.193	0.107
LMARS with Interaction	0.752	0.603

On the other hand, Table 4.7 presents the outcomes under 1000 genes' system. From the findings, it is observed that LMARS without interaction has the best performance and the performance of LMARS with interaction is the worst.

Table 4.7. The Accuracy Table for E-GEOD-14764 Data under 1000 Genes' System

Methods	F-Measure	Accuracy
GGM	0.167	0.091
LMARS without Interaction	0.193	0.107
LMARS with Interaction	0.049	0.025

On conclusion, by considering all cases, it is observed that LMARS without interaction is better than its alternates in the construction of networks for high dimensional real datasets. On the other hand, for small dimensional real datasets, among other models, GCGM with the RJMCMC inference approach is more accurate. Since the performance of this model has a strong limitation via dimension of the system, we can recommend it for small or moderate dimensional systems. Lastly, it is observed that GGM has the worst accuracy in all datasets although it is one of the most well-known and common methods in the systems biology.

4.3. Application via Threshold Methods

In the following parts, simulated and real data application results are given, respectively.

4.3.1. Simulated Data Application

In order to evaluate the performances of threshold methods on the data under normality assumption, we generate random and scale-free network data in the R programing language with *huge* package. We use the Gaussian graphical model (GGM) via the graphical lasso (glasso) to estimate precision matrix. To do so, GGM algorithm uses a penalty value, λ, to construct the structure of the network. The higher the value of the λ, results in sparser the network (or precision matrix). On the other hand, the lower λ causes denser network which is not commonly seen in the majority of biological network. Accordingly, as stated in Chapter 3, StARS (Liu et al., 2010) criterion is implemented to calculate optimal λ and to estimate precision matrix for GGM via glasso. However, the estimation of adjacency matrix from the estimated precision matrix has some add-hoc calculation since the selection of the threshold value which converts the precision matrix to the adjacency matrix is not clear although the selection of the threshold has an undeniable impact on accuracy of the model. In the literature, there are certain selection methods which are either computationally very demanding or not very insightful. Therefore, the researcher generally chooses a suitable threshold value after a preliminary study by controlling the associated studies or he/she intuitively assigns a threshold in such a way that the network looks like sparse. Hereby, in the following analyses, we present the results of the methods for different threshold selection in the literature together with the results for our suggested threshold method which takes into account the topology of the underlying network.

In our analyses, the simulated networks are generated under 20, 50 and 100 dimensions with 50 observations per gene. Hence, we generate random and scale-free network data whose states are multivariate normal distribution and we compare the mean accuracy value, F-measure and Matthew's correlation coefficient (MCC) based on 1000 Monte Carlo runs for the dimension of 20 and 50, however, because of its computational demand we compare the mean accuracy value, F-measure and

Matthew's correlation coefficient (MCC) based on 200 Monte Carlo runs for the dimension of 100.

Table 4.8. GGM via Graphical Lasso Results Based on 1000 Monte Carlo Runs under the Dimension of 20 with Random Network

Methods	Accuracy	F-Measure	MCC	User CPU Time
Proposed Method	0.837	0.312	0.215	0.065
MST Method	0.839	0.289	0.212	5.818
MDT Method	0.835	0.283	0.188	5.834
KMT Method	0.839	0.289	0.212	5.663
0.5T Criterion	0.809	-	-0.093	0.062

Table 4.9. GGM via Graphical Lasso Results Based on 1000 Monte Carlo Runs under the Dimension of 20 with Scale-Free Network

Methods	Accuracy	F-Measure	MCC	User CPU Time
Proposed Method	0.878	0.324	0.248	0.066
MST Method	0.881	0.320	0.256	5.709
MDT Method	0.878	0.315	0.24	5.741
KMT Method	0.881	0.320	0.256	5.524
0.5T Criterion	0.855	-	-0.074	0.060

Hereby, the simulated network which is generated under the dimension of 20 for random and scale-free networks are given in Table 4.8 and Table 4.9, respectively. It is seen that MST and KMT method have the best accuracy for both network types. However, KMT method has better computational time than MST method. Furthermore, the proposed method is the best in terms of F-measure, and it has significantly better computational time compared to KMT method. On the top of that,

0.5T Criterion has the worst MCC, and it cannot calculate F-measure since it fails to detect any true positive (any interaction).

Table 4.10. GGM via Graphical Lasso Results Based on 1000 Monte Carlo Runs under the Dimension of 50 with Random Network

Methods	Accuracy	F-Measure	MCC	User CPU Time
Proposed Method	0.932	0.271	0.247	0.538
MST Method	0.930	0.312	0.280	34.271
MDT Method	0.930	0.312	0.280	34.063
KMT Method	0.930	0.312	0.280	35.408
0.5T Criterion	0.921	-	-0.036	0.342

Table 4.11. GGM via Graphical Lasso Results Based on 1000 Monte Carlo Runs under the Dimension of 50 with Scale-free Network

Methods	Accuracy	F-Measure	MCC	User CPU Time
Proposed Method	0.954	0.149	0.180	0.538
MST Method	0.943	0.249	0.220	35.144
MDT Method	0.942	0.245	0.217	33.837
KMT Method	0.943	0.251	0.223	33.725
0.5T Method	0.941	-	-0.029	0.347

Results in the Table 4.10 and Table 4.11 show the results of the Monte Carlo runs that are generated under the dimension of 50 for random and scale-free networks, respectively. It is seen that proposed method is slightly better in terms of the detection of correctly classified interactions, which is presented as the accuracy value in the table. On the other hand, 0.5T criterion is the worst one to predict interactions with its negative MCC and lowest accuracy value compared to other methods, even though it is widely used criteria (Jiménez & Lobo, 2007). Moreover, F- measure of 0.5T is not

applicable since the method is unable to investigate any correctly classified interactions, represented as true positive (TP). On the other side, besides the comparisons via the accuracy of all methods, the second major criterion for the researchers is the computational demand that can be evaluated under the central processing unit (CPU) time. should be considered together to be able to interpret the method performances.

Although MST, MDT and KMT criteria are better than the proposed method in terms of F-measure and MCC, Table 4.10 and Table 4.11 indicate that they are computationally very demanding regarding the suggested approach. MST, MDT and KMT methods require considerably more CPU time with respect to our proposed method and 0.5T criterion even though 0.5T criterion has low accuracy.

Overall, the results in both Table 4.10 and Table 4.11 present us that the proposed method can be a reasonable threshold criterion comparing to other criteria under both network types.

Table 4.12. GGM via Graphical Lasso Results Based on 200 Monte Carlo Runs under the Dimension of 100 with Random Network

Methods	**Accuracy**	**F-Measure**	**MCC**	**User CPU Time**
Proposed Method	0.958	0.292	0.273	1.609
MST Method	0.960	0.273	0.255	154.009
MDT Method	0.960	0.273	0.255	154.066
KMT Method	0.960	0.273	0.255	153.868
0.5T Method	0.960	-	-0.018	1.605

Hereby, Table 4.12 and following Table 4.13 present the Monte Carlo runs for the random and scale-free networks for the dimension of 100, respectively. It is observed that MST, MDT and KMT criteria have the best accuracy, however the proposed

method has a better F-measure, MCC and significantly much better CPU time for random network with the dimension of 100.

On the other hand, in Table 4.13, it is seen that KMT is the best in terms of F-measure and MCC, however it has worse CPU time compared to the proposed method and 0.5T criterion for scale-free network. Furthermore, it is clear that although 0.5T criterion has the best accuracy and CPU time, it dramatically fails to catch network structure since it cannot predict any interaction at all and has the worst MCC. Overall, the proposed method has high accuracy measures as KMT and less computational time as 0.5T criterion.

Table 4.13. GGM via Graphical Lasso Results Based on 200 Monte Carlo Runs under the Dimension of 100 with Scale-free Network

Methods	Accuracy	F-Measure	MCC	User CPU Time
Proposed Method	0.967	0.166	0.149	1.452
MST Method	0.965	0.174	0.158	141.767
MDT Method	0.965	0.173	0.157	141.798
KMT Method	0.966	0.175	0.159	141.746
0.5T Method	0.970	-	-0.014	1.449

4.3.2. Real Data Application

In this part, we present the performance of the proposed method and 0.5T criterion for three gynecological datasets under both eleven and 1000 genes. In this analysis, MST, MDT and KMT methods, as well as MCC values for all approaches are not applicable since the quasi-true network is a complete graph and therefore, TN and FP values are all zero for these datasets. Hereby, to be able to compare all methods via real gene

network datasets, we use two different real gene network datasets which are cell signaling pathway data and the human gene expression data. These datasets are benchmark data that are applied in different studies for comparative analyses and their true network structures are known.

Table 4.14. The Performances of the Proposed Method and 0.5T Criterion for Gynecological Datasets with Eleven Genes

Methods	**Proposed Method**		**0.5T Criterion**	
Datasets	**Accuracy**	**F-Measure**	**Accuracy**	**F-Measure**
E-GEOD-9891	0.207	0.343	0.107	0.193
E-GEOD-14764	0.107	0.193	0.091	0.167
E-GEOD-63678	0.174	0.296	0.091	0.167

Table 4.15. The Performances of the Proposed Method and 0.5T Criterion for Gynecological Datasets with 1000 Genes

Methods	Proposed Method		0.5T Criterion	
Datasets	*Accuracy*	*F-Measure*	*Accuracy*	*F-Measure*
E-GEOD-9891	0.091	0.167	0.091	0.167
E-GEOD-14764	0.091	0.167	0.091	0.167
E-GEOD-63678	0.124	0.221	0.174	0.296

As seen in Table 4.14, the performance of the proposed method is far better than the 0.5T criterion for a small dimensional system.

On the other hand, in Table 4.15, since underlying model which is Gaussian graphical model (GGM) itself is affected directly by the higher dimensionality when making an inference about network system, the performances of both method is affected indirectly.

From the findings of Table 4.15 it is observed that the accuracies drastically decrease for both methods. However, surprisingly, accuracy measures for data E-GEOD-63678 increases under the 0.5T criterion and gives better accuracy regarding the outputs of Table 4.14.

Cell Signaling Pathway

This dataset is obtained from the study of Sachs et al. (2005). The dataset includes the flow cytometry measurements of eleven phospholipids and phosphorylated proteins which are measured on 11,672 red blood cells. These components belong to network of the cellular protein signaling of the human immune system's cells. By doing so, in the data collection, researchers intent to understand the signals of the native state tissue, actions of the complex drug and the dysfunctional signals of the diseased cells (Sachs et al., 2005).

In the Table 4.16, we list the results of different threshold selection methods by considering the true network system as given in the study of Sachs et al. (2005).

Human Gene Expression Data

The Large-scale human gene expression data are described in the works of Bhadra and Mallick (2013), Chen and Chen (2008) and Stranger et al. (2007). In this study, the gene expression of B-lymphocyte cells from the Utah residents with Northern and Western European ancestry sample is included in this study. The genes of 60 unrelated individuals are probed for 100 different transcripts. But, the focus is on the 3125 Single Nucleotide Polymorphisms (SNPs) that are found in the 5 UTR (untranslated region) of mRNA (messenger RNA) with a minor allele frequency greater than 0.1. This system includes 45 biologically validated links whose transcription factor and target genes are known (Bhadra & Mallick, 2013). Therefore, in our analyses, the

performances of all threshold selection methods are compared by considering these 45 validated links.

In the Table 4.16, it is seen that worst measures are obtained from widely used 0.5T criterion for cell signaling pathway dataset. Furthermore, the MST method and the KMT method have the same accuracy measures and are the best methods. On the other hand, our proposed method is the second best method by being very close to MST and KMT methods in terms of accuracy measures. Whereas, from the outputs of Table 4.17, it is seen that the KMT method is the best method by considering the all accuracy measures for the human gene expression dataset. As second best method, our proposed method can be selected since it has better results overall.

Table 4.16. The Performances of Threshold Selection Methods for Cell Signaling Pathway Dataset

Methods	*Accuracy*	*F-Measure*	*MCC*
Proposed Method	0.793	0.615	0.578
MST Method	0.802	0.637	0.596
MDT Method	0.785	0.618	0.542
KMT Method	0.802	0.637	0.596
0.5T Method	0.719	0.392	0.411

Table 4.17. The Performances of Threshold Selection Methods for Human Gene Expression Dataset

Methods	*Accuracy*	*F-Measure*	*MCC*
Proposed Method	0.990	0.752	0.748
MST Method	0.968	0.544	0.599
MDT Method	0.976	0.604	0.645
KMT Method	0.991	0.783	0.781
0.5T Method	0.991	0.689	0.722

CHAPTER 5

CONCLUSION

In this study, we initially aimed to generate a gynecological cancers pathway by a quasi-network and compare the selected parametric and nonparametric network models via real datasets. Secondly, by considering the Gaussian graphical model (GGM) as a fundamental model, we aimed to compare certain threshold selection criteria in the literature and our proposed threshold selection method which considers the underlying network topology in order to construct the adjacency matrix from the precision matrix. Overall, since it is very crucial to detect cancer related genes to cure cancer, our aim is to evaluate certain network models which are already validated by the literature, and to see which model can capture gene interactions better via real data applications. Then, we try to improve one of these models, GGM, by applying and comparing some of the main threshold selection criteria and our proposed nonparametric threshold selection method.

Hereby, in Chapter 2, the background information about the networks classes and the biological networks are given. First of all, the structures of the random and scale-free network are described. Although most of the biological networks are thought to be scale-free, the literature review shows that some of them do not follow the degree distribution, that is, they are not scale-free. On the other hand, majority of the researchers are in agreement to accept the fact that the metabolic networks are scale-free networks, mostly. Accordingly, in the application part, both network classes are considered for the modeling and the application of the threshold selection methods. Lastly, in the Chapter 2, the construction of our quasi-gynecological network is explained via the detailed literature review for the further analyses.

Accordingly, in Chapter 3, the parametric and nonparametric network models are presented. In this part, initially, GGM as an undirected model under the steady-state behaviors of biological system is explained. Then, the copula Gaussian graphical model (GCGM) which is the combination of GGM with the Gaussian copula is described via two different estimation methods which are reverse jump Markov chain Monte Carlo (RJMCMC) approach and the birth-and-death Monte Carlo methods (BDMCMC) under Bayesian settings. GCGM under RJMCMC approach makes use of the univariate marginal distributions of the observed variables by using copulas' theoretical framework which enables the multivariate associations to be modelled from those univariate marginal distributions. Its stationary distribution is the joint posterior distribution of the parameters and the model. On the other hand, GCGM under BDMCMC approach jumps between the models which are always accepted so that the stationary distribution is always the posterior distribution of interest. So, the algorithm of these two MCMC methodology searches over the model space in order to estimate the parameter of interests and to identify the high posterior regions for probability models. In addition to the parametric models, under the nonparametric methods, the loop-based Multivariate Adaptive Regression Splines (LMARS) is presented which is based on lasso regression. Then, LMARS model is constructed in two ways. First of all, we consider only the main effects in the model. And secondly, we also include the second-order interactions to the model. Accordingly, it is seen that converting the precision matrix into the adjacency matrix in the application of these methods is very important since it directly affects the accuracy of the methods. For this reason, we also present certain threshold selection criteria in the literature and also propose a method which considers the topology of the network systems.

Herein, in Chapter 4, the comparative analyses show that GCGM via RJMCMC and BDMCMC are not computationally feasible with the high dimensional data. However, GCGM via RJMCMC performs well for the small dimensions regarding other methods. Considering all of the results and the advantages over GGM and GCGM, LMARS without interaction is better than the other method for the high dimensional

networks. On the other hand, GGM performs very poorly in every case. In addition, threshold selection methods are evaluated via simulated and real datasets under distinct dimensions. We can conclude that the proposed method can efficiently detect the network structure by correctly classifying the true links between genes regarding the other approaches for both random and scale-free networks under the simulated data. Moreover, the proposed method and the 0.5T criterion have the lowest the central processing unit (CPU) time as compared to other methods. However, although the 0.5T criterion is the most common approach in the literature, it has the worst prediction performance to detect the interactions of the network for both network classes. On the other side, real data application under different dimensions via real gynecological datasets and benchmark datasets indicate that the proposed method also performs very well regarding the all other methods when considering both accuracy measures and the computational demand of the procedure under GGM.

As the future study, we consider to select the threshold value via a novel parametric approach by considering the distribution of the underlying network, i.e., the degree distribution of the scale-free networks. Furthermore, similar to GGM model, the LMARS model also uses a threshold value implicitly so that its regression coefficients can be directly converted into a binary form to be able to represent a graph. So, we think that a threshold selection procedure can be also inserted in its computation and the gain in performance of the model can be evaluated. By doing so, we believe that models can detect cancer related genes better. Thus, critical biomarkers can be detected more accurately, and it can support the treatments of cancers in the future (Liu et al, 2019). Finally, datasets which have the same features and are collected in the same manner, can be merged and analyses results can be generalized via data integration.

REFERENCES

Afify, A. M., Werness, B. A., and Mark, H. F. (1999). HER-2/neu oncogene amplification in stage I and stage III ovarian papillary serous carcinoma. *Exp. Mol. Pathol*. 66,163–69

Almudevar, A. (2009). Selection of statistical thresholds in graphical models. *EURASIP Journal on Bioinformatics and Systems Biology*, 1-13, article ID 878013.

Araujo, L. F., Siena, A. D. D., Plaça, J. R., Brotto, D. B., Barros, I. I., Muys, B. R.,Silva, W. A. et al. (2018). Mitochondrial transcription factor A (TFAM) shapes metabolic and invasion gene signatures in melanoma. *Scientific Reports*, 8, 1.

Arnett, B., Soisson, P., Ducatman, B. S., and Zhang, P. (2003). Expression of CAAT enhancer binding protein beta (C/EBP beta) in cervix and endometrium. *Molecular cancer*, 2, 21.

Barabási, A. L., and Oltvai, Z. N. 2004. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5, 2, 101–113.

Barabási, A. L. (2016). *Network Science*. Cambridge Press.

Bassett, D. S., and Bullmore, E. (2006). *Small-World Brain Networks*. The Neuroscientist, 12, 6, 512–523.

Börner, K., Sanyal, S., and Vespignani, A. (2007). *Network science.* Annual Review of Information Science and Technology, *41(1), 537–607.*

Brachova, P., Mueting, S. R., Devor, E. J. and Leslie, K. K. (2014). Oncomorphic TP53 Mutations in Gynecologic Cancers Lose the Normal Protein:Protein Interactions with the microRNA Microprocessing Complex. *Journal of cancer therapy*, 5(6), 506–516.

Bhadra, A., and Mallick, B. K. (2013). Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics*, 69, 2, 447–457.

Cancer Genome Atlas Research Network. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, 497, 7447, 67–73.

Chay, D., Cho, H., Lim, B. J., Kang, E. S., Oh, Y. J., Choi, S. M., Kim, B. W., Kim, Y. T., and Kim, J. H. (2010). ER-60 (PDIA3) is highly expressed in a newly established serous ovarian cancer cell line, YDOV-139. *International Journal of Oncology*, 37, 2.

Chen, J., and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika Trust*, 95, 3, 759–771.

Chen, S., Witten, D. M., and Shojaie, A. (2015). Selection and estimation for mixed graphical models. *Biometrika*, 102, 1, 47-64.

Cho, K. R., and Shih, I. (2009). Ovarian cancer. *Annual review of pathology*, 4, 287–313.

Chang, L., Zhang, D., Shi, H., Bian, Y., and Guo, R. (2017). MiR-143 inhibits endometrial cancer cell proliferation and metastasis by targeting MAPK1. *Oncotarget*, 8, 48, 84384–84395.

Chung, H., Cho, H., Perry, C., Song, J., Ylaya, K., Lee, H., and Kim, J. H. 2013. Downregulation of ERp57 expression is associated with poor prognosis in early-stage cervical cancer. *Biomarkers*, 18, 573–579.

Denkert, C., Budczies, J., Darb-Esfahani, S., Györffy, B., Sehouli, J., Könsgen, D., Zeillinger, R., Weichert, W., Noske, A., Buckendahl, A. C., Müller, B. M., and Dietel, M. (2009). A prognostic gene expression index in ovarian cancer - validation across different independent data sets. *Journal of Pathology*, 218, 2, 273-280.

Dobra, A., and Lenkoski, A. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*, 5(2A), 969–993.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 711–32.

Fan, Y., Wang, X., and Peng, Q. (2017). Inference of Gene Regulatory Networks Using Bayesian Nonparametric Regression and Topology *Information. Computational and mathematical methods in medicine 2017*, 8307530.

Farnoudkia, H., & Purutçuoğlu, V. (2019). Copula Gaussian graphical modelling of biological networks and Bayesian inference of model parameters. *Scientia Iranica*, 26, 4, 2495-2505.

Fukushi, Y., Sato, S., Yokoyama, Y., Kudo, K., Maruyama, H., and Saito, Y. (2001). Detection of numerical aberration in chromosome 17 and c-erbB2 gene amplification in epithelial ovarian cancer using recently estab- lished dual color FISH. *Eur J Gynaecol Oncol*, 22, 23–25.

Gibson, S., M., Ficklin, S. P., Isaacson, S., Luo, F., Fletus, F. A., and Smith, M. C. (2013). Massive-scale gene co-expression network construction and robustness testing using random matrix theory. *PLoS One*, 8, 2, e55871.

Guisan, A., Theurillat, J.-P., and Kienast, F. (1998). Predicting the potential distribution of plant species in an alpine environment. *Journal of Vegetation Science*, 9, 1, 65–74.

He, S., Zeng, S., Zhou, Z. W., He, Z. X., and Zhou, S. F. (2015). Hsa-microRNA-181a is a regulator of a number of cancer genes and a biomarker for endometrial carcinoma in patients: a bioinformatic and clinical study and the therapeutic implication. *Drug design, development and therapy*, 9, 1103–1175.

Hoff, P. (2007). Extending the rank likelihood for semiparametric copula estimation. *Annals of Applied Statistics*, 1, 265–283.

Hong, B., Le Gallo, M., & Bell, D. W. (2015). The mutational landscape of endometrial cancer. *Current Opinion in Genetics and Development*, 30, 25–31.

Hsu, H. C., Tsai, S. Y., Wu, S. L., Jeang, S. R., Ho, M. Y., Liou, W. S., et al. (2017). Longitudinal perceptions of the side effects of chemotherapy in patients with gynecological cancer. *Supportive Care in Cancer*, 25(11), 3457–3464.

Iyoke, C.A., and Ugwu, G.O. (2013). Burden of Gynaecological Cancers in Developing Countries. *World Journal of Obstetrics and Gynecology*, 2, 1-7.

Jiménez-Valverde, A., and Lobo, J. M. (2007). Threshold criteria for conversion of probability of species presence to either or presence absence. *Acta Oceologica*, 31, 3, 361–369.

Koopman, T., Vegt, V. D. B., Dijkstra, M., Bart, J., Duiker, E., Wisman, G. B. A., Bock, G. sH., and Hollema, H. (2018). HER2 immunohistochemistry in endometrial and ovarian clear cell carcinoma: discordance between antibodies and within situ hybridization. *Histopathology*,73, 5.

Kovtun, I. V., Piyan, Z.,Harris, F. R., Hou, X., Weroha, J. S., Vasmatzis, G. (2017). Targeting ERBB2 pathway in ovarian cancer. *In Proceedings of the AACR-NCI-EORTC International Conference: Molecular Targets and Cancer Therapeutics.*

Köbel, M., Piskorz, A. M., Lee, S., Lui, S., LePage, C., Marass, F., Brenton, J. D. et al. (2016). Optimized p53 immunohistochemistry is an accurate predictor of TP53mutation in ovarian carcinoma. *The Journal of Pathology: Clinical Research*, 2, 4, 247–258.

Köbel, M., Xu, H., Bourne, P. A., Spaulding, B. O., Shih, I. M., Mao, T. L., Soslow, R. A., Ewanowich, C. A., Kalloger, S. E., Mehl, E., Lee, C. H., Huntsman, D., and Gilks, C. B. (2009). IGF2BP3 (IMP3) expression is a marker of unfavorable prognosis in ovarian carcinoma of clear cell subtype. *Modern Pathology*, 22, 3, 469–475.

Le Gallo, M., O'Hara, A. J., Rudd, M. L., Urick, M. E., Hansen, N. F., O'Neil, N. J., Bell, D. W. et al. (2012). Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. *Nature genetics*, 44, 12, 1310–1315.

Lenkoski, A. (2013). ADirect Sampler for G-Wishart Variates. arXiv: 1304.1350

Li, Y., Liu, Q., McGrail, D. J., Dai, H., Li, K., and Lin, S. Y. (2018). CHD4 mutations promote endometrial cancer stemness by activating TGF-beta signaling. *American journal of cancer research*, 8, 5, 903–914.

Liao, C.J., Wu, T.I., Huang, Y. H., Chang, T. C., Wang, C. S., Tsai, M. M., Lai, C. H., Liang, Y., Jung, S. M., and Lin, K. H. (2011). Glucose-regulated protein 58 modulates cell invasiveness and serves as a prognostic marker for cervical cancer. *Cancer Science*, 102, 2255–2263.

Liu, C., Berry, P. M., Dawson, T. P., and Pearson, R. G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, 28, 385–393.

Liu, F., Zhang, S. W., Guo, W. F., Wei, Z. G., & Chen, L. (2016). Inference of Gene Regulatory Network Based on Local Bayesian Networks. *PLoS computational biology*, 12, 8, e1005024.

Liu, H., Roeder, K. and Wasserman, L. (2010). Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. arXiv: 1006.3316v1

Liu, V. W., Shi, H. H., Cheung, A. N., Chiu, P. M., Leung, T. W., Nagley, P., Wong, L. C., Ngan, H. Y. (2001). High incidence of somatic mitochondrial DNA mutations in human ovarian carcinomas. *Cancer Research*, 61(16), 5998-6001.

Liu, Y., Yi, Y., Wu, W., Wu, K., and Zhang, W. (2019). Bioinformatics prediction and analysis of hub genes and pathways of three types of gynecological cancer. *Oncology Letters*, 18, 617-628.

Luo, H., Xu, X., Ye, M., Sheng, B., and Zhu, X. (2018). The prognostic value of HER2 in ovarian cancer: A meta-analysis of observational studies. *PloS one*, 13, 1, e0191972.

Manel, S., Williams, H. C., and Ormerod, S. J. (2001). Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, 38, 921–931.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298, 5594, 824–827.

McConechy, M. K., Ding, J., Senz, J., Yang, W., Melnyk, N., Tone, A. A., Huntsman, D. G. et al. (2013). Ovarian and endometrial endometrioid carcinomas have distinct CTNNB1 and PTEN mutation profiles. *Modern Pathology*, 27, 1, 128–134.

Manel, S., Dias, J.M., Buckton, S.T., Ormerod, S.J. (1999). Alternative methods for predicting species distributions: an illustration with Himalayan river birds. *Journal of Applied Ecology*, 36, 734–747.

Manel, S., Williams, H. C., and Ormerod, S. J. (2001). Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology,* 38, 921–931.

Mohammadi, A., and Wit, E. C. (2015). Bayesian Structure Learning in Sparse Gaussian Graphical Models. Bayesian analysis, 10, 1, 109-138.

Miller, C. R., Oliver, K. E., and Farley, J. H. (2014). MEK1/2 inhibitors in the treatment of gynecologic malignancies. Gynecologic Oncology, 133, 1, 128–137.

Mhawech-Fauceglia, P., Herrmann, F. R., Rai, H., Tchabo, N., Lele, S., Izevbaye, I., Odunsi, K., and Cheney, R. T. (2010). IMP3 distinguishes uterine serous carcinoma from endometrial endometrioid adenocarcinoma. *American Journal of Clinical Pathology*, 133, 899–908.

Noske, A., Faggad, A., Wirtz, R., Darb-Esfahani, S., Sehouli, J., Sinn, B., Denkert, C. et al. (2009). IMP3 Expression in Human Ovarian Cancer is Associated With Improved Survival. *International Journal of Gynecological Pathology*, 28, 3, 203–210.

Rosenfeld, N., Mes-Masson, A. M., Brenton, J. D., Mullany, L. K., Wong, K. K., Marciano, D. C., Katsonis, P., King-Crane, E. R., Ren, Y. A., Lichtarge, O.,

and Richards, J. S. (2015). Specific TP53 Mutants Overrepresented in Ovarian Cancer Impact CNV, TP53 Activity, Responses to Nutlin-3a, and Cell Survival. *Neoplasia*, 17, 10, 789–803.

Pappa, K. I., Polyzos, A., Jacob-Hirsch, J., Amariglio, N., Vlachos, G. D., Loutradis, D., and Anagnou, N. P. (2015). Profiling of discrete gynecological cancers reveals novel transcriptional modules and common features shared by other cancer types and embryonic stem cells. *PLoS One*, 10, 11, 1-20.

Pan, Z., Chen, S., Pan, X., Wang, Z., Han, H., Zheng, W., Shao, R. et al. (2010). Differential gene expression identified in Uigur women cervical squamous cell carcinoma by suppression subtractive hybridization. Neoplasma, 57, 2, 123–128.

Penson, R. T., Sales, E., Sullivan, L., Borger, D. R., Krasner, C. N., Goodman, A. K., Birrer, M. J. et al. (2016). A SNaPshot of potentially personalized care: Molecular diagnostics in gynecologic cancer. Gynecologic Oncology, 141, 1, 108–112.

Purutçuoğlu, V., and Seçilmiş, Deniz. (2019). Modeling of Biochemical Networks via Classification and Regression Tree Methods. Mathematical Methods in Engineering. In K. Taş, D. Baleanu, & J.A.T. Machado (Eds), *Mathematical Methods in Engineering: Applications in Dynamics of Complex Systems*. Cham, Switzerland: Springer.

O'Hara, A. J., and Bell, D. W. (2012). The genomics and genetics of endometrial cancer. *Adv Genomics Genet*, 2, 33–47.

Romero-Garcia, S., Lopez-Gonzalez, J. S., Báez-Viveros, J. L., Aguilar-Cazares, D., and Prado-Garcia, H. (2011). Tumor cell metabolism: an integral view. *Cancer biology & therapy*, 12, 11, 939–948.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D., and Nolan, G. (2005). Casual protein signaling networks derived from single cell data. *Science*, 308, 5721, 523–529.

Schneider, L. F., Krajina, A., Krivobokova, T. (2019). Threshold selection in univariate extreme value analysis. 1-36, arXiv: 1903,02517.

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., et al., Population genomics of human gene expression. *Nature Genetics*, 39, 10, 1217–1224, 2007.

Small, W., Bacon, M. A., Bajaj, A., Chuang, L. T., Fisher, B. J., Harkenrider, M. M., Gaffney, D. K. et al. (2017). Cervical cancer: A global health crisis. *Cance*r, 123, 13, 2404–2412.

Takata, H., Kudo, M., Yamamoto, T., Ueda, J., Ishino, K., Peng, W. X., Wada, R., Taiai, N., Yoshida, H., Uchida, E., and Naito, Z. (2016). Increased expression of PDIA3 and its association with cancer cell proliferation and poor prognosis in hepatocellular carcinoma. *Oncology letters*, 12, 6, 4896–4904.

Thuiller, W., Lavorel, S., Araujo, M. B., Sykes, M. T., and Prentice, I. C. (2005). Climate change threats to plant diversity in Europe. Proceedings of the National Academy of Sciences, USA 102, 8245–8250.

Tothill, R., Tinker, A., George, J., Brown, R., Fox, S., Johnson, D., et al. (2008). Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical Cancer Research*, 14, 16, 198-208.

The Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. Nature, 474(7353), 609–615. doi:10.1038/nature10166

Vang, R., Levine, D. A., Soslow, R. A., Zaloudek, C., Shih, I., and Kurman, R. J. (2016). Molecular Alterations of TP53 are a Defining Feature of Ovarian High-Grade Serous Carcinoma: A Rereview of Cases Lacking TP53 Mutations in The Cancer Genome Atlas Ovarian Study. *International journal of gynecological pathology: official journal of the International Society of Gynecological Pathologists*, 35, 1, 48–55.

Wang, H. and Li, S. (2012). Efficient Gaussian graphical model determination under G-Wishart prior distributions. *Electronic Journal of Statistics*, 6: 168–198.

Wijst, V. D. M., de Vries, D. H., Brugge, H., Westra, H. J., and Franke, L. (2018). An integrative approach for building personalized gene regulatory networks for precision medicine. *Genome Medicine*, 10, 1, 96.

Whittaker, J. (2009). Graphical Models in Applied Multivariate Statistics. Wiley Publishing.

Yang, H. L., Lin, R. W., Rajendran, P., Mathew, D. C., Thigarajan, V., Lee, C. C., Hsu, C. J., SHseu, Y. C. (2019). Antrodia salmonea-induced oxidative stress abrogates HER-2 signaling cascade and enhanced apoptosis in ovarian carcinoma cells. *Journal of Cellular Physiology*, 234,3, 3029-3042.

Yiwei, T., Hua, H., Hui, G., Mao, M., & Xiang, L. (2015). HOTAIR Interacting with MAPK1 Regulates Ovarian Cancer skov3 Cell Proliferation, Migration, and Invasion. *Medical science monitor: international medical journal of experimental and clinical research*, 21, 1856–1863.

Wu, R., Lin, L., Beer, D. G., Ellenson, L. H., Lamb, B. J., Rouillard, J. M., Kuick, R., Hanash, S., Schwartz, D. R., and Cho, K. R. (2003). Amplification and Overexpression of the L-MYC Proto-Oncogene in Ovarian Carcinomas. *The American Journal of Pathology*, 162, 5, 1603–1610.

Zhao, H., and Duan, Z. H. (2019). Cancer Genetic Network Inference Using Gaussian Graphical Models. *Bioinformatics and Biology Insights*, 13, 1177932219839402.

Zhao, S., Choi, M., Overton, J. D., Bellone, S., Roque, D. M., Cocco, E., et al. (2013). Landscape of somatic single-nucleotide and copy-number mutations in uterine serous carcinoma. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 8, 2916–2921.

Z. Hu, D. Zhu, W. Wang, W. Li, W. Jia, X. Zeng, and et al. (2015). Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration. *Nature Genetics*, 47, 158-163.