NOVEL MODEL SELECTION CRITERIA ON HIGH DIMENSIONAL BIOLOGICAL NETWORKS

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

BY

GÜL BAHAR BÜLBÜL

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

JUNE 2019

Approval of the thesis:

NOVEL MODEL SELECTION CRITERIA ON HIGH DIMENSIONAL BIOLOGICAL NETWORKS

submitted by **GÜL BAHAR BÜLBÜL** 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 Dener Akkaya
Head of Department, **Statistics** _____

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

Examining Committee Members:

Prof. Dr. Olcay Arslan
Department of Statistics, Ankara University _____

Prof. Dr. Vilda Purutçuoğlu
Department of Statistics, METU _____

Prof. Dr. Serpil Aktaş Altunay
Department of Statistics, Hacettepe University _____

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

Assoc. Prof. Dr. Ceren Vardar Acar
Department of Statistics, METU _____

Date:

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: Gül Bahar Bülbül

Signature :

ABSTRACT

NOVEL MODEL SELECTION CRITERIA ON HIGH DIMENSIONAL BIOLOGICAL NETWORKS

Bülbül, Gül Bahar
M.S., Department of Statistics
Supervisor: Prof. Dr. Vilda Purutçuoğlu

June 2019, 129 pages

Gaussian graphical model (GGM) is an useful tool to describe the undirected associations among the genes in the sparse biological network. To infer such high dimensional biological networks, the l_1-penalized maximum-likelihood estimation method is used. This approach performs a variable selection procedure by using a regularization parameter which controls the sparsity in the network. Thus, a selection of the regularization parameter becomes crucial to define the true interactions in the biological networks. In this sense, we suggest to combine some information-theoretic measures such as CAIC, CAICF and ICOMP with a penalized likelihood approach in order to yield the true graph. Also, loop-based multivariate adaptive regression splines (LMARS) can be presented as a nonparametric modelling technique which is good at dealing with the problem of nonlinearity and collinearity in the data which the problems arise from high-dimensional networks. In this study, we interfere the model selection procedure of LMARS by applying our measures to find the correct structure, while it has been originally introduced with generalized cross validation as a model selection technique.

Keywords: Gaussian graphical model, l_1-penalized estimation, Loop-based multivariate adaptive regression splines, High dimensional model selection criteria, Information-theoretic measures.

ÖZ

YÜKSEK BOYUTLU BİYOLOJİK AĞLAR İÇİN YENİ MODEL SEÇME TEKNİKLERİ

Bülbül, Gül Bahar
Yüksek Lisans, İstatistik Bölümü
Tez Yöneticisi: Prof. Dr. Vilda Purutçuoğlu

Haziran 2019 , 129 sayfa

Gaussian grafiksel modeli seyrek biyolojik ağlarda genler arasındaki yönsüz ilişkileri gösterirken kullanılan, kullanışlı bir parametrik metotdur. Yüksek boyutlu biyolojik ağların tahmininde, l_1-cezalandırmalı tahmin metodu olan grafiksel lasso kullanılmaktadır. Grafiksel lasso metodu değişken seçme prosedürü uygular ve ağdaki seyrekliği belirlemek için düzenlileştirme parametresi kullanılmaktadır. Bu sebeple, bir biyolojik ağda bulunan doğru ilişkileri belirlemek için düzenlileştirme parametresi seçimi büyük önem kazanmaktadır. Bu bağlamda, biz doğru grafiği elde etmek için cezalandırmalı olabilirlik yaklaşımı ile birlikte bilgi kuramsal metotlar olan CAIC, CAICF ve ICOMP kullanmayı önermekteyiz. Ayrıca, çok yönlü uyarlanabilir regresyon çizgileri modeli, verideki doğrusal olmama ve ağların yüksek boyutundan kaynaklanan kolinerlik problemlerini çözmede başarılı olan parametrik olmayan bir modelleme tekniği olarak sunulabilir. Döngü tabanlı çok değişkenli uyarlanabilir regresyon splineları orjinal olarak genelleştirilmiş çapraz geçerlilik ölçütünü model seçme tekniği olarak kullanırken, biz bu çalışmada model seçme prosedürüne müdahale ederek bizim önerdiğimiz ölçüm kriterlerini kullanıp, doğru ağ yapısını bulmayı amaçlamak-

tayız.

Anahtar Kelimeler: Gaussian grafiksel modelleri, l_1-cezalandırmalı tahmin metodu, Çok boyutlu uyarlanabilir regresyon uzanımları, Yüksek boyutlu model seçme kriterleri, bilgi-kuramsal kriterleri.

To my lovely dear family

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Dr. Vilda Purutçuoğlu for her everlasting support, guidance and encouragement during this thesis study. Her guidance, support, and feedbacks have turned this study to an immeasurable learning. With her love and respect to her profession and disciplined stance, she has always been a role model for me in my academic career. It was a great fortune to be her student and it has been a great honor for me to work with her.

I would like to present my grateful thanks to my examining committee members, Prof. Dr. Olcay Arslan, Prof. Dr. Serpil Aktaş Altunay, Prof. Dr. Ömür Uğur and and Assoc. Prof. Dr. Ceren Vardar Acar for their detailed reviews and their constructive comments.

I also owe my special thanks to Zeynep Bahar, İldeniz Gevişen and Fatma Özcan for their warm friendship.

I also owe my special thanks to my cousins Veli Biçer, Utku Karakoç, Işık Biçer and Erdem Karakoç for their support and guidance during my thesis. I feel great appreciation to my lovely aunts Gülten Biçer, Nuriye Çalış, Hülya Bülbül, Yıldız Karakoç for their unconditional love and support throughout my life. I would like to convey my thanks to my grandparents: Bahar Bülbül, Türkan Ertuğrul, Musa Bülbül and Hüseyin Hüsnü Ertuğrul.

Finally, my deepest gratitude are for my lovely parents, Muhterem and Ünal for their endless support and sacrifices. I am forever indebted to my parents. Also, I would like to express my thanks to my sister Ege Türkü for her intense and priceless support. They loved me unconditionally through every stage of my life. I would never achieve to be here without their support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

acc	Accuracy
AIC	Akaike's Information Criterion
AL	Adaptive Lasso
ANOVA	Analysis of Variance
BMARS	Bootstrapping MARS
BCMARS	Bootstrapping Conic MARS
BF	Basis Functions
BIC	Bayesian Information Criterion
CART	Classification and Regression Trees
CAIC	Consistent AIC
CAICF	Consistent AIC with Fisher Information matrix
CQP	Conic Quadratic Programming
CMARS	Conic MARS
CRLB	Cramer-Rao Lower Bound
CV	Cross-Validation
F	F-measure
FLM	Fuzzy Logic Model
FP	False Positive
FN	False Negative
GCV	Generalized Cross Validation
GGM	Gaussian Graphical Model
GLM	Generalized Additive Models
Glasso	Graphical Lasso

Huge	High Dimensional Undirected Graph Estimation
ICOMP	Information Complexity Criterion
K-L	Kullback-Leibler Distance
LARS	Least Angle Regression
Lasso	Least Absolute and Shrinkage Operator
LSE	Least Square Estimation
MARS	Multivariate Adaptive Regression Splines
MB	Meinshausen-Bühlmann regression model in huge
MLE	Maximum Likelihood Estimate or Estimator
Negentropy	Negative Entropy
pre	Precision
RCMARS	Robust CMARS
rec	Recall
RF	Random Forest
RIC	Rotational Information Criterion
RSS	Residual Sum of Squares
SCAD	Smoothly clipped absolute deviation
StARS	Stability Approach to Regularization Selection for High Dimensional Graphical Models
TGD	Treshold Gradient Descent
TSCGM	Time Series Chain Graphical Models
TP	True Positive
TN	True Negative

CHAPTER 1

INTRODUCTION

1.1 Motivation

The massive amount of genomics data accelerated the process of understanding the relations between genes and disease in the higher dimensions. In a small scale, scientists in the field of computational biology try to solve the discoveries of the biological gene networks by using statistical approaches. In this sense, to model high dimensional biological networks, Gaussian graphical model (GGM) is considered as an useful tool to describe the undirected associations among the genes. Also, such networks are known for their sparsity where their patterns are determined by the inverse covariance matrix under the Gaussian assumptions. In higher dimensions, the statistical inference of the graphical models can be achieved via several penalized maximum-likelihood estimation methods. They perform a variable selection procedure by applying one or more regularization parameter which controls the sparsity of the network in a data-dependent way. Therefore, a selection of the regularization parameter plays a pivotal role to define the true interactions in the biological networks. In this context, our motivation arises from the selection of the best representation for such sparse biological networks under high dimensional setting. In this sense, we suggest to combine some information-theoretic measures such as CAIC, CAICF and ICOMP with a penalized likelihood approach, namely, graphical lasso algorithm, in order to yield the true network structure. Then, we compare them with the state-of-art model selection criteria, AIC and BIC, and the others where there have been already used for biological networks under higher dimensions, StARS, RIC, EBIC. In addition to the GGM, MARS, a nonparametric counterpart of the GGM, are used in the context of this thesis since they are able to handle nonlinearities and collinear-

ities in the data where the problems arise from high-dimensional networks. While it was introduced with its computer-intensive method GCV, we insert our information-theoretic measures to find the true structure of the biological network at the model selection stage. Finally, we aim to capture the true underlying mechanism of a biological network by selecting the true modelling strategy with a criterion among a collection of model selection criteria that we compare. Hereby, via this motivation, we list the aim of our study in the following part.

1.2 Aim of the Study

Under Gaussian Graphical Network Setting

- How can we insert information-theoretic measures into penalized likelihood method, namely graphical lasso algorithm, in order to penalize the covariance matrix when selecting the best biological network structure under higher dimensions?

- Which model selection method achieves to correctly specify the true biological network structure by correctly defining the regularization parameter ?

- Which variable selection procedure outperforms the others when the dimension of the networks is increases ?

Under MARS Setting

- How can we combine information-theoretic measures on the model selection stage with nonparametric modelling approach so called LMARS?

- Which model selection procedure works efficiently as the dimension increases ?

- Which model selection criterion provides an efficient approach in order to capture the true structure of the high dimensional sparse biological networks?

Overall

- Which modelling approach, GGM or MARS, with which model selection technique achieves to reflect the true nature of the sparse biological networks by correctly defining the regularization parameter under high dimensional setting?

- Which network structure (scale-free or random) can be seen as our main objective to fit our novel applications under both GGM and LMARS setting?

1.3 Thesis Overview

The twenty first century is known as the Information Era and its name suggests that the availability of the information offers to reach easily the big data. Specifically, after Human Genome Project was completed in 2003 which provides the universal access for complete genome sequences, the studies on bioinformatics have been substantially raised.

With the notable advances in biotechnology, rendering powerful high-throughput techniques such as chips and screens to obtain microarray data, the feasibility of such sequences triggers the investigations on high-throughput genomic data in the molecular form such as DNA, RNA and protein. Although the vast use of such biotechnological methods delivers the huge amounts of large-scale genomic data, uncovering the complex mechanism in the sparse cellular network entails the challenges in higher dimensions. Thereby, the understanding of the functionality of such cellular networks requires the worldwide interdisciplinary efforts, encompassing mainly statistics [13, 38] molecular biology and computer science. By this way, this worldwide task has aimed to unravel the complex interactions among molecular constituents which include DNA, RNA and protein in the cell. In fact, there are various attempts to embark on a quest interpret, organize and present the high-throughput data in a standardized way by means of complex networks in several fields that includes from natural sciences such as biology [13] and physics [15] to social sciences [140] such as economics [108] and communication [43]. In addition to these areas, graphical models have already been implemented to address some problems of the complex networks on the other fields of studies which range from graph theory [62] to artificial intelligence and machine learning [4]. In the sense of molecular biology, to examine complex molecular networks by means of high throughput data, graphical models are preferred as the useful exploratory tools that facilitate the systematic use of complete genomic sequences. Also, statisticians become a part of this universal effort when analyzing such complex sparse biological networks in a mathematical way by taking the randomness into account. From the perspective of system biology, graphical models pave the way not only for the understanding of the complex cellular networks but also for the determination of the relations among system components

in molecular level. Its several applications in microarray data [8, 61, 107, 108] enables to account for complex interactions among biological entities in the complex networks with the help of graphical models.

Table 1.1: Types of gene networks with respect to scale, presence of edges and topologies.

Types of Networks		
Scale	**Presence/Absence of Edge**	**Network Topologies**
i.Networks on the microscopic sclae		
a)Transcption regulatory networks		i.Homogeneous Network
b)Signal transduction networks		a) Random
c)Protein interaction networks	i.Directed	ii.Non-homogeneous Networks
ii.Networks on the macroscopic scale	ii.Undirected	a)Scale-free
a)Neural Networks		b)Hierarchical
b)Food webs		c)Modular
c) Phylogenetic networks		

In the sense of system biology, type of networks based on scaling criteria can be clustered into two basic parts as the microscopic and macroscopic in Table 1.1. In this sense, we can list the previous group as networks on the microscopic scale, or intra-cellular networks in which they are comprised of transcription regulation and signal transduction networks. On the other side, neural networks, ecological networks and phylogenetic networks appear on the macroscopic scale. As an example of intra-cellular networks, protein-protein interaction networks and metabolic networks take part in the macroscopic side. In the scope of this thesis, we will deal with genes, specifically, gene regulatory networks under the intra-cellular network concept. These networks whose entities DNA, RNA and protein are constructed to determine the gene expression levels of mRNA and protein in the molecular level. As well as the those constituents, links among them can be represented by means of such networks.

In addition to scaling criterion, the classification of the networks is achieved according to whether the direction of links exists or not in the network. Lastly, we put links into groups according to how they are distributed like in Table 1.1. The manner of ei-

ther presence of direction or distribution of the links will be mentioned the following chapters.

Intra-cellular networks whose not only sparse but also complex nature implies that the number of genes are far greater than the number of observations in the network. In the following chapters of the thesis, it will be mentioned that there are various ways how to handle such sparse data structures in higher dimensions.

The concept of *learning* from graphical models which is used in expert systems and artificial intelligence theory [109] is basically associated with the processes of fitting the graphical models in the field of statistics. The learning procedure mainly includes two overlapping steps called the structure learning and the parameter learning, in order to both accurately represent the biological map and truly analyze biological interactions of the entities. The first step is related to finding the proper graphical structure with sticking to underlying assumptions if there are available. Therefore, this stage is applied for model selection purposes and is known as the network structure or the structure learning [80, 81, 109]. On the other hand, the second step is called parameter learning. This task is used to estimate model parameters according to the properties of the model, as well as assumptions. Since the large-scale genomic data include the number of genes that far exceeds the number of observations in the model, we need to avoid the ubiquitous problem so-called *the curse of higher dimension* in order to estimate model parameters. It also implies small n, large p problem for sparse genomic data that is caused by ill-suited matrices. Thereby, we need to deal with solving such matrices transforming into optimization problem. Therefore, inferring parameter estimates in higher dimensions has already been a hot area topic [27, 75, 107, 109]. In this thesis context, either structure learning or parameter learning methods will be discussed in the following chapters in a detailed way.

In terms of structure learning, based on modeling approaches, three main methods are classified as Boolean, deterministic and stochastic approaches, each is directly connected to distinct point of views in the modeling sense. The first modeling approach is the Boolean network, the most primitive way to explain the relations between molecular constituents in complex cellular networks. This method is designed

to determine next possible status of the genes in the network based on current status. Also, this model reduces the complexity of the network, but it is unable to reflect true interactions in the molecular level [8, 126].

The most widely used techniques to represent large scale gene networks are based on deterministic and stochastic approaches. Deterministic types offer an effective ways of modeling for the sparse biological networks and are motivated by detecting steady-state behaviour of these large-scale genomic data with the aim of specifying relations among genes in the system. Since these types of models make use of differential equations to identify associations in the complex networks, deterministic models are so-called differential equation models in the literature. In the sense of modeling, deterministic models are ability to reveal random nature of the biological networks, as well as interactions among the entities in the networks so that they can embrace both parametric and nonparametric approaches to detect such randomness in the networks. In the statistical literature, Gaussian graphical models and its derivatives such time-series graphical chain(TSCGM) [1] are introduced as essential ways to generate gene networks parametrically. GGM produces undirected graphical representations by capturing linear interactions among biological components in the network under some assumptions. In the literature, to relax its linearity assumption, copula GGM has been proposed [39, 40] and the other modified versions have been implemented [128]. On the other hand, random forest(RF) [23], neural networks(NN) [76], fuzzy logic models(FLM) [48, 85, 131] classification and regression trees (CART) [21], as well as multivariate adaptive regression splines (MARS) [59] and its derivatives so-called conic MARS (CMARS) [6, 117, 124], robust CMARS [100] are applicable as nonparametric ways to model the cellular network by taking the steady-state behavior of the system into account. This nonparametric ways to do so, can be also grouped under the generalized additive models in the statistical learning [69]. While deterministic models are designed to steady-state behaviour of these large-scale genomic data [16], the stochastic models are specifically adjusted for explaining dynamic behaviour of the biological network, in turn, dynamic relations between genes in the networks [65, 127]. In the sense of high-dimensional modeling, deterministic and stochastic networks dominate the Boolean network in order to capture true nature of the system, reflecting both its sparsity and complexity of the network. Also, deterministic mod-

els surpass the stochastic models in terms of interpretation by the way providing the simpler models than the latter. Therefore, deterministic tools contribute to correctly the steady-state nature of such sparse biological networks, mimicking successfully the random behavior of the interactions in the networks.

This thesis covers the both GGM and MARS as linear modelling methods in higher dimensions, so these will be elaborated in the chapter 2.1.1 and 2.1.3, respectively. In the literature, GGM, one of the reliable deterministic way to construct complex biological networks in higher dimensions, so it facilitates model-based analysis of cellular networks. In sense of deterministic modeling of such sparse biological networks, MARS is known as nonparametric analog of GGM. It was introduced by Jerome Friedman, well-known pioneer statistician, in 1991 as a flexible linear regression modeling that is designed for high dimensional data. The idea of MARS is motivated by handling nonlinearities and multicollinearities in higher dimensions.

After specifying network structures under deterministic modeling approaches as GGM and MARS, respectively parametric and nonparametric methods, we continue with the parameter learning procedure for each method separately. In the learning parameter procedure, GGM requires to infer inverse covariance matrices when dealing with large-scale genomic data [16, 109]. GGM achieves to capture the linear interactions graphical model, as well as assumes the normality in higher dimensions for microarray data so that it prefers to apply extension and modification of the least squares regression, replacing with the plain least squares regression [45, 69, 75] in order to estimate the inverse of the covariance matrices. In the literature, these adaptive procedures based on linear regression ranges from the most primitive one known as subset selection to dimension reduction algorithm, is named as piecewise linear component. In the scope of this thesis, we will stress out two well accepted approaches in the case of inference of GGM, graphical lasso [56] and lasso-based graphical regression model [95]. While the first offers an exact solution, the latter approach gives an approximate way to do so. However, they are able to tackle intractability problem of the covariance matrices by applying particular optimization procedure. In addition to these two well-known lasso based approaches, there have been several methods for inferring inverse of the covariance matrices in higher dimensions, including the other

optimization procedures based on gradient descent [88], coordinate descent [56] and the derivatives of lasso including fused lasso [118, 122], group-lasso [135], elastic net [142], adaptive lasso [143], block descent algorithms [11]. Also, nonnegative garotte [22] and LARS [46] procedures are served as alternative regression methods, rather than simple linear regression approach. In addition, there exist the other methods which have mathematical impetus for improving ill-posed estimator in higher dimensions such as [86], Dantzig Selector [26], compressing sensing,[41, 42], Bernstein polynomials [102].

In the Section 2.2, data-dependent model selection procedures will be examined where the motivation behind the information based model selection methods lies in information theory. In the literature on model selection, among all penalized regression method, sole the lasso procedure simultaneously operates the feature selection [69] by assuming $||\beta||_p,\ p \geq 1$. Also, it requires a penalty parameter in order to penalize the covariance matrix, in turn, coefficients in the gene network. In our thesis context, the final gene network whose sparsity and interactions are directly affected by choice of penalty parameter. So, the selection of such penalty parameter is so important that several model selection procedures have been proposed for higher dimensions such as EBIC [53], StARS [66] and RIC [90] in order to determine the optimal sparse networks. To select optimal model among a collection of the candidates, in the statistical literature, AIC [2] and BIC [110] are the well-known state-of-art approaches. However, any combination of a penalized regression method with a model selection criteria dominates the others for all the time. Therefore, in this study, in addition to these model selection methods, CAIC, CAICF and ICOMP are proposed as alternative model selection criteria in order to choose the proper model in higher dimensions [17, 18, 19, 20], so they are presented by combining them with graphical lasso procedure under GGM setting. In the sense of the MARS modeling technique, it is originally designed to use GCV as a model selection criterion in order to compile model selection procedure. However, in this thesis, it is aimed to compete its performance in terms of some accuracy measures with the suggested procedures, namely, CAIC, CAICF and ICOMP, as well as AIC and BIC. Thus, we modify the model selection procedure of the original MARS by replacing GCV with our proposed data-dependent techniques by constructing separate MARS models for each genes which

we call them LMARS models [7].

In the application part of this thesis, real data implementations and simulations obtained under distinct topologies and dimensions will be presented for both modelling approaches GGM and LMARS in Section 3.4 and 3.5, respectively.

CHAPTER 2

EXISTING METHODOLOGIES AND PROPOSED METHODS

2.1 Modelling Gene Networks

This chapter provides essential information about the existing methodology of modeling high dimensional networks: GGM and MARS, respectively, their inference methods such as *graphical lasso* and GCV methods will be introduced with the motivation behind them.

2.1.1 Gaussian Graphical Models

In systems biology graphical models have a crucial role to exhibit the structure of the network at the molecular level by identifying quantifiable pattern of associations among these molecular entities in the network.

Generally, graphical models are composed of a collection of vertices, or nodes and a catalogue of edges among the nodes. In the sense of determining the strength of interactions at the complex intra-cellular web, or gene expression network when nodes are represented by genes and its products such as proteins, DNA and RNA, edges among the genes refer to links or interactions among them. In Figure 2.1, p_1, p_2 and p_3 refer the genes in the simple network structure and links are defined among them. In the representation of biological systems, GGM can be defined as an useful tool in the high dimensional setting. Thereby, GGM aims to reflect true underlying mechanism of the network under the steady-state assumption. It can be classified as deterministic modeling approach, as well as taking part into model-based approach.

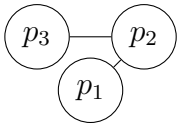


Figure 2.1: A simple representation of a network with three nodes via an undirected graph.

From a statistical point of view, GGM aims to build multivariate model which includes random variables are represented by genes in the map, simultaneously links among components in the network that imply correlations between a pair of node. Totally, its construction includes *p* random variables associated with nodes, or genes in the network, so the vector for random variables can be denoted by $\mathbf{Y} = (Y_1, \cdots, Y_p)$, each $Y_{(i)}$ could be interpreted as gene expression level of gene **i** (*i*=1, $\cdots$, *p*).

That is, vector $\mathbf{Y}$ has a multivariate Gaussian distribution and is represented as in the following form

$$\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

with mean vector $\boldsymbol{\mu}$=$(\mu_1, \cdots, \mu_p)$ and $(p \times p)$-dimensional variance-covariance matrix $\boldsymbol{\Sigma}$= $(\sigma_{ij})_{ij}$.

Since GGM is classified as model-based or parametric approach, and as its name *Gaussian* suggests that it posses the oracle properties of the Gaussian normality assumption. It also inherits the linearity concept so that the sole linear dependencies among a pair of nodes can be assigned by means of GGM.

When grouping model-based approaches in terms of presence/absence of edge, GGM is belong to the undirected part in Table 1.1. In other words, unlike Bayesian graphical model, it attempts to form undirected edges among genes [61, 77, 101]. Lacking

Table 2.1: The classification of the gene networks: Gene clustering, revelance and association networks.

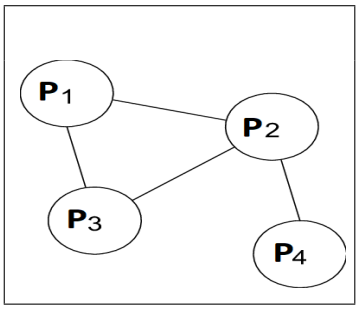
Gene Networks		
1.Gene clustering networks (Eisen et al., 1998)	**2. Gene Relevance Networks (i.e.,correlation)** (Butte et al., 2000)	**3. Gene association networks and Covariance Selection Networks also Concentration Graph** (Dobra et al., 2004) (Schäfer and Strimmer, 2005) (Sctuari and Strimmer, 2010) (Barabási and Oltvai, 2004)

of direction can be interpreted in a biological way like that the information based on the direction of the activation or inhibition cannot be learned from this type of models. Although undirected graphical models distinguish form directed ones by ignoring the direction of the edges, the latter can be expressed in terms of the other one under the Markov properties [61, 109].

Under closer inspection of the gene networks in Table 2.1, the analysis based on genomic data was begun with clustering technique [47]. Then, it evolved to correlation based examination [24]. Finally, GGM have been formed by the way undergoing the change in dependence assumption. The sole difference exists between gene relevance networks and gene association networks in the sense of representing independence assumption by the fact that while gene relevance network assumes marginal independence structure to show interactions among entities, gene association network refers conditional dependencies for relations [107, 109]. Thereby, GGM was presented as a covariance selection model [36] and concentration graph [126].

From the GGM scope of view, ignoring of edge is coincided with the conditional independence assumption, that is, an absence of the link between a pair of genes implies that conditional independence of the corresponding genes exists given all other genes. It is expressed in a mathematical formula such that $Y_1 \perp Y_4 \mid$ rest shown in Figure 2.2.

Figure 2.2: The simple representation of the conditional independence between the node 1 and the node 4 for the given node (2,3).



There is a direct approach that conditional dependence, or partial correlations among genes are inferred from the inverse of the covariance matrix so-called the *precision matrix* or concentration matrix. It is denoted by $\Theta = \Sigma^{-1} = \Theta_{(ij)}$. In this sense, the covariance matrix has a key role for explaining conditional independence under Gaussian assumption. For example, zeros in the covariance matrix describe conditional independence between two specific genes given the rest. Also, in the precision matrix, the inverse of the partial variance, located in the diagonal, is expressed as $\Theta_{(ii)}$=$1/var(Y_i|rest)$.

In addition to partial variances, the strength of the partial correlations can be extracted from the precision matrix, which is denoted as

$$\pi_{ij} = \frac{-\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}, \tag{2}$$

where π_{ij} represents the partial correlation between Y_i and Y_j given all the other variables.

The precision matrix was announced as an original way to derive partial correlations

[126]. Despite understanding the vitality of the precision matrix in terms of partial correlations in the gene network, in the high dimensional setting, the challenge for gene association networks arises in to obtain reliable estimate for the population covariance matrix. This inevitable problem stems from the constructional nature of the genetic networks, which includes a huge number of variables, but relatively few samples. Therefore, the empirical covariance matrix, **S**, cannot be served as an unbiased estimate to infer population covariance matrix. In the small *n*, large *p* setting in higher dimensions, sample covariance matrix suffer from two characteristic problems related to invertibility and positive-definiteness. In the statistical sense, these two terms are seamlessly connected to accurate estimates or approximations. Thereby, it is expected from a reliable estimate to be both invertible and positive-definite. The term **invertible** implies the well-conditioned matrix where the ratio between its minimum and maximum singular value is not to be too large, so it has full-rank. Also, an estimate needs to be the one with non-zero variance to obtain accurate positive-definite covariance matrix.

In the statistical sense, it is aimed to derive partial correlations by the way GGM can be performed as a set of regression functions, regressing each nodes against the all other nodes. This procedure refers to construct conditional distributions for each separated nodes given the rest. Under multivarite normality assumption, this offers to make use of the remarkable properties of the normal distribution and $\mathbf{Y}$ indicates joint multivariate vector as $\mathbf{Y} = (Y_{-p}, Y_p)$, $\mathbf{Y}_{-p} = (Y_1, \cdots, Y_{p-1})$ representing that the vector includes all nodes, but not the last one. Thereby, the conditional distribution for Y_p is formed as

$$Y_p|Y_{-p} = y \sim N(\boldsymbol{\mu_p} + (y - \boldsymbol{\mu_{(-p)}})^t \boldsymbol{\Sigma}^{-1}_{(-p,-p)} \sigma_{(-p,p)}, \sigma_{(p,p)} - \sigma^t_{(-p,p)} \boldsymbol{\Sigma}^{-1}_{(-p,-p)} \sigma_{(-p,p)}), \tag{3}$$

where $\boldsymbol{\Sigma_{-p,-p}}$ refers to $((p-1) \times (p-1))$-dimensional variance-covariance matrix except the last nodes and $\boldsymbol{\sigma}_{(-p,p)}$ and $\boldsymbol{\sigma_{(p,p)}}$ indicates $((p-1) \times 1)$-dimensional covariance vector associated with Y_{-p} and Y_p, and variance for Y_p, respectively. Also, the mean and variance decomposition associated with the $\mathbf{Y}$ are obtained like

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_{(-p)} \\ \mu_{(p)} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{(-p,-p)} & \sigma_{(-p,p)} \\ \sigma_{(p,-p)} & \sigma_{(p,p)} \end{pmatrix}. \tag{4}$$

This regression scheme yields regression coefficients $\boldsymbol{\beta}$=$\boldsymbol{\Sigma}^{-1}_{(-p,-p)}$ $\sigma_{(-p,p)}$, which ensures the conditional independence property. So $\beta_j = 0$ intrinsically indicates that Y_p and Y_j are conditionally independent given all the other nodes. This notion can be also expressible via precision matrix and $\boldsymbol{\beta}$ = $-\theta_{(-p,p)}/\theta_{(p,p)}$ is correspondence with the partial correlations in it.

As a standard procedure to infer $\boldsymbol{\Theta}$, maximum likelihood approach is exploited in higher dimensions to attain sample covariance matrix. This procedure is achieved under the consideration that vector $\mathbf{Y}$ follows the multivarite normal distribution and the corresponding joint density function establishes as

$$f(y_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\Big\{ -\frac{1}{2}(y_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (y_i - \boldsymbol{\mu}) \Big\} \tag{5}$$

and likelihood is defined in terms of two unknown parameters, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ via

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{6}$$

Thus, the log-likelihood forms as

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log(L(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = -\frac{n}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^T \boldsymbol{\Sigma}^{-1} (y_i - \boldsymbol{\mu}). \tag{7}$$

The Equation (7) can be written also by means of precision matrix $\boldsymbol{\Theta}$ as

$$l(\boldsymbol{\mu}, \boldsymbol{\Theta}) = \frac{n}{2} \log |\boldsymbol{\Theta}| - \frac{1}{2} \sum_{i=1}^{n} (y_i - \boldsymbol{\mu})^T \boldsymbol{\Theta} (y_i - \boldsymbol{\mu}). \tag{8}$$

Substituting $\boldsymbol{\mu}$ by its likelihood estimate $\bar{y}$, so it forms

$$l(\boldsymbol{\Theta}) = \frac{n}{2} \log |\boldsymbol{\Theta}| - \frac{n}{2} \text{Trace}(\mathbf{S}\boldsymbol{\Theta}) \tag{9}$$

with sample covariance matrix $\mathbf{S}=(\boldsymbol{s_{ij}})_{\boldsymbol{ij}}$ defined as

$$s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (y_{(i)k} - \bar{y}_{(i)})(y_{(j)k} - \bar{y}_{(j)}), \tag{10}$$

where $\bar{y}_{(i)} = \frac{1}{n} \sum_{k=1}^{n}$ and $y_{(i)k}$ represents the *k*-th observation of the variable Y_i. In all equations from (5) to (8), $(.)^T$ denotes the transpose of the given statement.

Hence, the strength of the interactions can be measured via maximum likelihood procedure, that is sample covariance estimate, $\hat{\boldsymbol{\Sigma}}=\mathbf{S}$ and simultaneously, maximum likelihood estimate can be determined by $\hat{\boldsymbol{\Theta}}=\mathbf{S}^{-1}$.

In small *n*, large *p* setting data setting, maximum likelihood approach can generally results in a poor estimate so-called empirical covariance matrix, or **S** which is incapable of satisfying the desirable characteristics such as well-conditioned and positive-definiteness. In other words, MLE estimate that suffers from infinite variance turns into non-invertible matrix, so no unique least square coefficient estimate can be acquired by following this procedure. Also, even if the MLE estimate is extracted as an invertible matrix, it tends to produce a fully connected network rather than a sparse network that it is suitable for the representations of gene association network in high dimensional setting [8, 56, 57]. As the number of variables in the model increases, regardless of whether the feature is significant or not, the residual sum of squares decreases. So, the variable selection procedure based on least squares, in the high dimensional setting, leads to complex networks, including as much as variables.

In this sense, there are various methods have been proposed in order to infer $\boldsymbol{\Theta}$ by

overcoming such difficulties encountered in higher dimensions under the normality assumption.

In Section 2.1.2, it will be discussed why the residual sum of squares does not make sense in the sense of fitting the large *p* and small *n* setting [125] and the reason behind why we need to use penalized likelihood approaches [54] in higher dimensions will be examined. Firstly, this section provides a brief overview of existing methodology along with the advantages and disadvantages of suggested approaches, i.e., ridge [72], lasso [121] and the derivatives of lasso.

Then, in Section 2.1.2, we will discuss the inference procedures for GGM, both makes use of the lasso properties. In this thesis context, the most accepted structure learning approaches for GGM, lasso-based regression model and l_1-penalized likelihood, respectively, *neighborhood selection* [95] and *graphical lasso* [56] will be highlighted.

2.1.2 Inference of GGM

In higher dimensional setting, least square regression does not make sense [69]. So, some regularization methods and alternative regression procedures have been proposed. In this sense, the ridge and lasso regressions are represented as counterparts of the least squares with some oracle properties. Therefore, they aim to shrink the coefficients towards zero by applying a regularization parameter as a constraint. By this way, they achieves to prune the complexity of the network as we expected from $p \gg n$ dimensional setting.

In linear regression, the plain of least squares fitting is designed to obtain estimates $\beta_0, \beta_1, \cdots, \beta_p$ according to minimization of the residual sum of squares and it can be expressed as below

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2, \tag{11}$$

while y_i denotes the ith response and x_{ij} shows the predictor for the ith observation and for the jth random variable.

RSS could not be served as a reliable tool in terms of bias and interpretability of the model in higher dimensional setting [75]. While the least square estimation faces with high variability problem in $p \gg n$ case, resulting in poor predictions, as well as is not capable of finding unique coefficient estimate in the case of infinite variance, shrinkage methods appear that they use RSS in a constraint or a regularized form to restrain the common problems of RSS in higher dimensions. Also, LSE tends to obtain complex models that include the irrelevant variables, but shrinkage methods provide an efficient way to establish more interpretable models, with the goal of determining the importance of the variables. Thereby, their notion of shrinking RSS allows to control bias-variance trade-off by lowering the variance as well as to yield more simpler models, unlike RSS.

To tackle common problems, shrinkage methods include not only the RSS part, but also the regularization criterion based on the shrinkage penalty, allowing to transform the way which we obtain the estimates or the coefficients. For example, *ridge regression* uses ${\beta_j}^2$ to restrict the estimates, enforcing the coefficients to reduce towards almost zero. Its formulae can be written as

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_i x_{ij})^2 + \lambda \sum_{j=1}^{p}\beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p}\beta_j^2 \tag{12}$$

with $\lambda \geq 0$ that is represented as shrinkage penalty or tuning parameter.

From the Equation (12), it can be concluded that when $\lambda = 0$, ridge procedure transforms the simple least square regression, so it uncontrollably will end up with the estimates like the ones produced by the latter. Associatively, as $\lambda \to \infty$, while λ raises up, the more ridge coefficients $\hat{\beta}_\lambda^R$ turn out to be almost zero.

The way the ridge regression shrinks the coefficients towards zero justifies the use of ridge, however, ridge regression procedure will result in the model which includes

all p variables. So, it aims to shrinks the coefficients, but not to reduce towards exactly zero. In the sense of model complexity, it cannot be used for discarding the variables from the full model. To remedy for diminishing the complexity, *lasso regression* achieves both variable selection and shrinkage, whereas ridge regression is only designed to shrink the coefficients. Lasso offers a more efficient way to delete irrelevant predictors than the ridge since $|b|$ is much bigger than b^2 with $0 \leq |b| \leq 1$ [130]. Therefore, lasso and ridge shares a similar approach like that both operates the sum of squares in a penalized form, but lasso differs from the ridge with an attractive feature selection procedure. So, lasso is represented by the following expression

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{i=1}^{p}\beta_i x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda \sum_{j=1}^{p}|\beta_j|. \tag{13}$$

The l_1-constraint is denoted as $\sum_{j=1}^{p} \leq t$ and represents the sum of the absolute value of the coefficients, leading to the coefficients exact zero. Therefore, it utilizes l_1-norm which is represented by $||\beta||_1 = \sum |\beta_j|$ penalty for this purpose, rather than l_2-norm that ridge regression prefers. Comparing to ridge, lasso operates such procedures that subset selection and shrinkage. Also, unlike ridge regression, lasso is equipped with the convexity property [69]. The other desired feature is that the lasso procedure leads to sparse models, performing variable selection [75].

The ridge regression and the lasso form are re-expressed as

$$\begin{aligned} &\underset{\beta}{\text{minimize}} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{i=1}^{p}\beta_j x_{ij})^2 \\ &\text{subject to } \textstyle\sum_{j=1}^{p}|\beta_j| \leq s, \end{aligned} \tag{14}$$

$$\begin{aligned} &\underset{\beta}{\text{minimize}} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{i=1}^{p}\beta_j x_{ij})^2 \\ &\text{subject to } \textstyle\sum_{j=1}^{p}\beta_j^2 \leq s. \end{aligned} \tag{15}$$

In Equation(14) and (15), s represents the constraint for both an absolute value of the sum coefficients and the sum of squared value for the coefficients in an absolute form when $(j = 1, \cdots, p)$ and $(i = 1, \cdots, n)$.

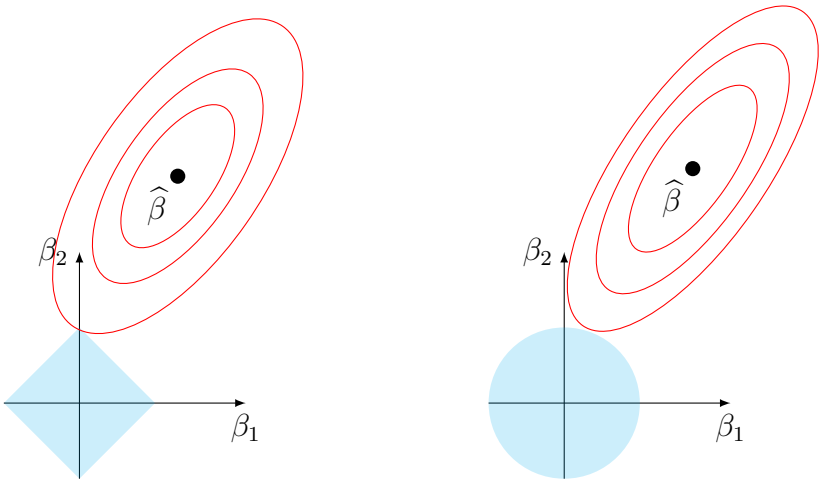


Figure 2.3: From left to right: lasso and ridge estimation procedures differ from the usual RSS.

The red circles in Figure 2.3. show the RSS contour whose $\hat{\beta}$ refers to the usual least square coefficient. Furthermore, it indicates how lasso performs feature selection, in contrast to ridge regression, that is, while the constraint for lasso involves the corners, making some coefficients exactly zero, the ridge constraint is represented by a circle, not necessarily leading to zero for them in high dimensional setting. With this remarkable feature, lasso prevails the ridge due to the fact that lasso operates the variable selection. Therefore, lasso procedure results in simpler models, so both sparser and interpretable models in the $p \gg n$ setting.

Focusing on lasso, it can be re-exppressed as a Lagrangian form like that

$$\underset{\beta \in \Re^p}{\text{minimize}} \frac{1}{2} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} ||\beta_j||, \tag{16}$$

where y_i and x_{ij} take place in a standardized version, represented by $\frac{1}{n}\sum_i y_i = 0$ and $\frac{1}{n}\sum_i^n x_{ij} = 0$ and $\lambda \sum_i^n x_{ij}^2 = 1$, implies that the intercept term, β_0 will be omitted.

Although the lasso was introduced by a well-known statistician Rob Tibshriani in 1996 [121], inspired from nonnegative garrotte [21], the discovery of LARS (least angle regression), or *homotopy* [46, 98] allows lasso to broad its horizon in such a way that it seeks for entire path to construct piecewise linear paths by following a

sequential scheme [122]. This provides a different perspective for lasso that the lasso solution emerges as a forward stagewise regression. Also, in the context of signal processing, l_1-penalty approach was introduced as *basis pursuit* by Chen et al. [29].

The lasso problem served with an appealing convexity property, Lagrangian form can be solved by the fact that a quadratic programming (QP) method is utilized to compute convex constraint problem for lasso [69]. That is, *coordinate descent* procedure is provided that Lagrangian form offer feasibility in terms of numerical computation. [56, 57, 130]. Because of finding global optimum, it is achieved via cyclical coordinate descent scheme in such a way that it operates minimization of the convex function for each coordinate simultaneously, leading to converge global optimum [69]. Also, in contrast to least square estimation, lasso produce unique solution in the high dimensional setting $p \gg n$. Otherwise, coordinate descent procedure has ability to perform lasso regression under sparsity with its attractive properties.

Despite convexity, for fixed *p*, Zho [142, 143] proved that lasso generally does not exhibit the consistency in terms of feature selection. The lasso investigated under the case when letting the number of variables goes to infinity with a higher rate than the number of observations, *n*. Then, he determined the appropriate conditions, where the lasso reflects the consistency, to conduct variable selection [95, 138]. Therefore, it is discovered that lasso is consistent provided that *p* could not reach the value, with $\exp(n^a)$ with $n \geq 1$ and normally distributed errors. It can be said that lasso does not seem to be an efficient approach so as to estimate nonzero coefficients since it has a tendency to over-shrink the coefficients to zero [74]. Also, Zhang and Huang [137] demonstrated that the lasso attributes to right order of sparsity so that the strategy is capable of conducting the variable selection under certain conditions. So, all of the coefficients are greater than $(\lambda/n)\sqrt{k_n}$, where $\sqrt{k_n}$ is the number of nonzero coefficients.

To enjoy oracle properties of an estimator that is associated with the detection of the appropriate number of the nonzero coefficients with the probability that converges to one, as well as the nonzero coefficients that are required to be asymptotically normal

with the same mean and covariance, there have been many adjustments of lasso proposed in the literature, by replacing the original loss function, namely, l_1-norm with the novel one.

While all variants make use of the desirable properties of the lasso, they are specialized differently in order to overcome distinct problems such as correlated structure, seen in microarray studies or to strength the procedure in terms of feature selection. Some examples for these cases will be mentioned in a nutshell [69].

2.1.2.1 Fused Lasso

The fused lasso [118, 122] is the one variant of the lasso that is designed for grouping parameters by taking the time into account to tackle large correlations in the data in which coordinate descent and LARS make no sense in this case. Its least square analogous formula can be written as

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda_n^{(1)} \sum_{j=1}^{p} ||\beta_j|| + \lambda_n^{(2)} \sum_{j=2}^{p} ||\beta_j - \beta_{j-1}||, \tag{17}$$

where $\boldsymbol{\beta}$=$(\beta_0, \cdots, \beta_p)$ and $\boldsymbol{x_i}$=$(x_{i1}, \cdots, x_{ip})^T$ and $\lambda^{(1)}$ as well as $\lambda^{(2)}$ refer to Lagrange multipliers.

This approach establishes two constraints with the goal of restricting the adjacent coefficients with the additional constraint as $\lambda^{(2)}$ on neighbouring coefficients. Here, the coordinate descent procedure is not applicable because of non-separability of this method [118]. So, there have been proposed some alternative algorithms in the sense of fitting [56, 57, 69].

2.1.2.2 Elastic Net

Elastic net approach [142] is another derivative of the lasso and is compromises between the ridge and lasso penalty as a loss function. Comparing lasso, it aims to manage highly correlated variables altogether via the following expression

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + (1-\lambda)||\beta||_2^2 + \lambda||\beta||_1, \tag{18}$$

where $||\beta||_2^2$ refers to l_2-norm regularization and $||\beta||_1$ denotes l_1-norm regularization. It is noted that when $\lambda = 1$, the *elastic net* regression turns into lasso. On the other hand, if $\lambda = 0$ case, it becomes as a ridge regression. By this way, it can differ from the original lasso by means of the unpenalized intercept model [69, 122]. Also, Zou and Hastie [142] launched the LARS-EN algorithm. This algorithm is adapted from its precursor, LARS [46], to manage the feature selection problem that lasso suffers in the high dimensional setting with the aim of learning structure.

2.1.2.3 Group Lasso

It has an ability to operate with a group of variables, rather than individual covariates, determining the coefficients whether they are zero or nonzero simultaneously. When qualitative factors exist as our predictors, a set of dummy variables are served to manage group of variables [135].

In the linear regression settings, group lasso procedure is able to govern group of covariates, in turn construct a model involving *J* group of variables, $(j = 1, \cdots, J)$. Here, collection of covariates is represented in a vectoral form $(\boldsymbol{Z_1}, \cdots, \boldsymbol{Z_j})$, where each is belong to *j*th group $\boldsymbol{Z_j} \in R^{p_j}$. In order to predict response Y, linear regression model can be expressed as $\theta_0 + \sum_{j=1}^{J} Z_j^T\theta_j$ with a group of p_j regression coefficients $\boldsymbol{\theta_j} \in \Re^{p_j}$. The model forms with a set of N samples $(y_i, z_{i,1}, z_{i,2}, \cdots, z_{i,J})$,

$$\sum_{i=1}^{n}(y_i - \theta_0 - z_{ij}^T \boldsymbol{\theta_j})^2 + \lambda \sum_{j=1}^{J} ||\boldsymbol{\theta_j}||_2, \tag{19}$$

with the Euclidean norm of the vector $\boldsymbol{\theta_j}$ denoted as $||\boldsymbol{\theta_j}||_2$.

Group lasso is applied in the case of gene-expression arrays when they have a set of highly correlated genes from the same biological pathway [69, 135]. In the literature, in order to find the coefficient path for a group lasso [49] apply a such procedure with the aim of detecting splice-site on some genomic data where the data includes human DNA with each observation encompassing seven basis $(A, G, C, T)^7$. Furthermore, Yuan and Lin [135] proved that there exists a connection between LARS and group lasso, as well as non-negative garotte [22] and lasso [121, 122] [135].

2.1.2.4 Adaptive Lasso

It can be considered as weighted lasso procedure and expressed as

$$\sum_{i=1}^{n}(y_i - \sum_{i=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \boldsymbol{w_j}||\boldsymbol{\beta_j}||, \tag{20}$$

where $\boldsymbol{w_j}$ represents known weights for estimates.

It can be evaluated as a convex problem l_1-penalty. So, the adaptive lasso possesses the oracle properties of the lasso such as consistency and uniqueness, for fixed *n* when $p \to \infty$ as $n \to \infty$ [143]. Also, it is showed that for p setting, it exhibits the consistency and efficiency in the sense of variable selection [143]. Moreover, it is designed to apply the relatively higher penalty for zero coefficients, on the other hand, the lower penalty for nonzero coefficients with the weighted l_1-norm. Thereby, it can be solved with the LARS[46] procedure and it can be used as a tool for feature selection [69, 74, 143].

2.1.2.5 Smoothly Clipped Absolute Deviation Lasso penalty

The penalized log-likelihood procedure with SCAD penalty can be written as

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \text{SCAD}_{\lambda,\alpha}||\beta_j||. \tag{21}$$

The SCAD penalty can be reformed by its continuous differentible penalty function and is given by

$$\text{SCAD}'_{\lambda,\alpha}(x) = \mathbf{I}(\beta \geq \lambda) + \frac{(\alpha\lambda - \beta)_+}{(\alpha - 1)\lambda}\mathbf{I}(\beta > \lambda), \tag{22}$$

for some $\alpha > 2$ and $\beta > 0$. $\mathbf{I}$ refers an indicator function. Here, $(\alpha\lambda - \beta)_+$ shows a quadratic spline function with knots at λ and $\alpha\lambda$. If $\alpha = \infty$, it turns out to lasso penalty [141].

This penalty is known as smoothly clipped absolute deviation penalty [50]. Correspondingly, it is solved as a quadratic spline function with two knots such as λ and $\alpha\lambda$. Also, SCAD procedure performs hard tresholding penalty function where it does not excessively facilitate to shrink the large values of the β. It offers continuous solution and is given by

$$\hat{\beta} = \begin{cases} \text{sign}(x)(|x| - \lambda)_+, & |x| \leq 2\lambda. \\ ((\alpha - 1)x - \text{sign}(x)\alpha\lambda)/(\alpha - 2), & 2\lambda < |x| \leq \alpha\lambda. \\ x, & |x| \leq \alpha\lambda. \end{cases}$$

For (λ, α), cross validation and generalized cross-validation [32] are originally suggested to find two unknown parameters [50]. In practice, α is chosen as 3.7.

From the high dimensional framework, the adaptive lasso and SCAD penalties share the three requirable properties of an estimator where they yields sparse estimates, consistent model selection and unbiased estimates for large coefficients [141]. However, at the end of SCAD procedure, it results in denser estimates than AL procedure.

Up to know, it have been mentioned the validity of the versatile penalized likelihood methods in the linear regression setup from the lasso perspective in high-dimensional inference problems rather than least square regression. Without loss of generality, they share the similarity where they exploits some filtering methods such as hard and soft tresholding [42]. Also, these penalized likelihood procedures can be served as a tool of variable selection. On closer inspection, to tackle common problems in the sense of high-dimensional sparse networks, specifically for GGM, inference, or structure learning, is achieved via Bayesian variable selection to deal with thousands of variables by means of stochastic algorithm [38] or principle components analysis, rather than LSE can be preferable in order to infer sparse networks by applying l_1-norm penalty. Also, ridge regression can be implemented as a penalized regression [72]. Another method, known as shrinkage approach, can be exploited to infer $\mathbf{\Theta}$ [16, 107] by controlling bias-variance trade-off. The other approach is based on limited order of the partial correlations such as second correlations [34, 92, 128]. For example, Castelo and Roverato [27] use the only first-order partial correlations only, after this is generalized to q setting and this can be a valid approach for the gene relevance networks, instead of gene association networks. In addition to the various inference strategies in higher dimensions, lasso regression surpasses them since it is capable of governing both inference and feature selection, unlike the others. The general principle behind the lasso that enforces the coefficients toward exact zero, satisfies the partial correlation assumption that GGM dictates.

In the sense of achieving high dimensional inference of GGM, the most welcomed procedures are based on the lasso approach. In the scope of this thesis, neighbourhood selection [95] and graphical lasso (GLASSO) [56] will be mentioned as tools for inference of GGM.

Firstly, we will examine the penalized regression method, then will continue with penalized maximum likelihood estimation, respectively. Both is directly related to the feature selection, and fitting procedures so that it is possible to learn structure models in addition to fitting procedure [93]. So, at the end of lasso procedure, it is not required to apply statistical test to infer final network.

So far it have been mentioned that the difficulties of estimating sparse networks from the data, but the notion of sparsity is also coincided with the biological expectations from the gene association networks under the high-dimensional setting. This dictates not only a few number of edges in the network, but also many zeros in the precision matrix. To provide sparser networks, there exists an efficient approach so that covariance selection [36] procedure is used for this purpose under the assumption of the conditional independence.

2.1.2.6 Neighbourhood Selection

Meinshausen and Bühlmann [95] introduced an idea of obtaining sparse networks under a procedure that graphical model can be defined in terms of a set of regression models. In other words, this method which is driven by regression methodology results in a group of regression models in which each node is regressed against the other nodes in the graph. This procedure is so-called neighbourhood selection with Lasso in order to achieve covariance selection in higher dimensions [16].

To detect conditional independence assumption, the usage of precision matrix and edges in the graph are served as usual way [85]. In addition, regression coefficients can be viewed as an indicator of the conditional independence since the neighbourhood selection justifies the usage of coefficients. Therefore, it achieves to control the number of parameters in the network, as well as enabling to reduce many coefficients to zero. Also, the graph can be interpreted in terms of edges and nodes such that

$\mathbf{G} = (\mathbf{\Gamma}, \mathbf{E})$, while $\mathbf{\Gamma}$ ($\Gamma = 1, \cdots, p$) represents the set of nodes, $\mathbf{E}$ denotes the set of edges in the graph. In accordance with the edge set *E* contains a pair of nodes such as *(a,b)*, it implies that X_a and X_b are conditionally dependent given the rest $X_{\Gamma\ a,b}$. On the other hand, a pair of nodes does not included in the edge set which refers conditional independence between the corresponding pair given all the other nodes, and , simultaneously, it can be also observed as a zero entry by means of precision matrix, or inverse covariance matrix.

So, the regression model based approach is motivated by lasso regression rather than simple linear regression. The model is formed as

$$\underset{\beta}{\text{minimize}}[||Y_p - Y_{-p}\beta||_2^2 + \lambda_p||\beta||_1], \tag{23}$$

where λ_p is a regularization or tuning parameter and l_1-lasso regularization refers to $||\beta||_1 = \sum_i |\beta_{ip}| < \lambda$.

Here, λ_p is increased, the number of lasso coefficient β which becomes zero is also increased, accordingly. Therefore, the lasso scheme contributes to the desired sparsity by imposing a lasso constraint.

With the neighbour selection purposes, the lasso regression procedure is adjusted for predicting a variable or one node against the others in order to obtain lasso estimate $\hat{\beta}^{a,\lambda}$ for each β^a by using the following expression.

$$\hat{\beta}^{a,\lambda} = \underset{\beta_a=0}{\text{minimize}}(n^{-1}||\mathbf{Y}_a - Y\beta||_2^2 + \lambda||\beta||_1), \tag{24}$$

where $||\beta||_1 = \sum_{b\in\Gamma(n)} |\beta^b|$ is the l_1-norm of the coefficient vector with $\mathbf{Y}_a$ corresponding under $a \in \Gamma(n)$. This indicating the vector of *n* observations.

Equation (24) shows that the each node is regressed against the all other variables. With the aim of identifying the set of neighbours for each node, it is formalized in

such way that nonzero entries corresponding with specific node in the inverse covariance matrix is attained as a collection of neighbours. It was proven that Equation (24) neighbourhood coefficient for each node a in the graph asymptotically equals to the lasso coefficient. So, it offers asymptotic solution, rather than exact solution. It is noted that Meinshausen and Bühlmann [95] demonstrated even if uniqueness property fails for the Equation (24), the collection of the solutions still serves a feasible convexity. Also, they proved that, comparing with the other penalties regarding with l_p-norm, l_1-norm is the one that operates the feature selection, so it inherits oracle properties in the high dimensional setting when keeping $p \geq 1$ [54].

The lasso estimate extracted from a set of regression is an appealing way to infer partial correlations, as well as managing the structure learning procedure. In other words, the absence of edges in the graph accounts for also zero regression coefficient associated with response and the corresponding variable. Since the lasso procedure achieves to infer final network, in this sense, statistical tests to learn the structure of the network are not required. However, a symmetricity problem lies behind this method, implies lasso-regression approach is not need to result in symmetric covariance matrix. In contrast, conditional independence assumption enforces the symmetric covariance matrix in higher dimensions. To tackle this problem, Meinshausen and Bühlmann [95] suggests two alternative approaches such as AND and OR rule. While AND rule performs such a way that links are set up when both regression coefficients result in zero, OR rule facilitates that one zero entry in the covariance matrix is enough to attain the missing edge in the network. So, a network derived by OR rule turns in more sparser networks than that by AND.

$$\lambda_i = 2\sqrt{\frac{\mathbf{S}_{ii}}{n}\Phi^{-1}(1 - \frac{\alpha}{2p^2})}, \tag{25}$$

where $\mathbf{S}_{ii}$ represents the sample variance for the node $Y_{(i)}$ and Φ refers to the cumulative distribution function of the standard normal.

One way of formalizing the notion of the restriction is achieved via Meinshausen and Bühlmann [95]'s idea that suffices that the probability of false positives is less

than λ. In such a way, the larger variability in data requires to apply larger penalty parameter.

Since the penalized regression method results in an approximate solution to the exact problem, an exact solution have been proposed by adapting interior point optimization [115, 135]. Also, the other exact way was developed to solve the penalized likelihood in a coordinate descent framework [11, 56]. Thereby, the other section will be mentioned mainly about GLASSO, served as an exact solution to penalized likelihood equation in order to infer sparse high dimensional network under the GGM setting.

2.1.2.7 Graphical Lasso

Inheriting the oracle convexity property of the lasso, graphical lasso was introduced as a powerful regression based method for graph selection in higher dimensions. In this sense, to construct continuous model, it is assumed that *N* multivarite observations follows the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Under higher dimensions, GGM dictates the conditional independence assumption, means i and j seems conditionally independent provided that *ij*th component in the $\boldsymbol{\Sigma^{-1}}$ equals to zero. In a multivarite Gaussian framework, lasso implementation is aimed to estimate sparse undirected Gaussian graph, so the l_1-norm is needed to be modified as a penalty on the inverse covariance matrix or the *precision*, expressed as a optimization procedure. Hence, under normality assumption,

$$\underset{||\boldsymbol{\Theta}||_1 \leq \rho}{\text{maximize}} [\log |\boldsymbol{\Theta}| - \text{Trace}(\mathbf{S}\boldsymbol{\Theta})] \tag{26}$$

with $||\boldsymbol{\Theta}||_1 = \sum_{i,j} ||\boldsymbol{\Theta}_{ij}||$ and ρ refers to non-negative tuning parameter.

So, graphical lasso [56] imposes l_1-constraint on $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Theta}$, rather on the regression coefficient, β.
With the help of remarkable convexity property of $\boldsymbol{\Theta}$, the constraint region, as well

as negative log-likelihood are convex. It can be re-written in a Lagrangian dual form, in turn, re-expressed as a penalized log-likelihood optimization problem

$$\underset{\Theta}{\text{maximize}}[\log|\mathbf{\Theta}| - \text{Trace}(\mathbf{S\Theta}) - \lambda||\mathbf{\Theta}||_1], \quad (27)$$

where λ refers to the non-negative Lagrange multiplier and $\mathbf{S}$ represents the empirical covariance matrix.

Banerjee et al. [11] demonstrated that the dual form of $\mathbf{\Theta}$ refers to convex optimization problem in a penalized log-likelihood form with putting λ as a constraint.

While the optimal λ value is closer to 0, the penalized optimization problem turns into usual maximum likelihood procedure. On the other hand, as λ is increased, the sparser network we obtain, relating to penalized likelihood scheme, rather than its usual version. Thus, the optimization problem emerges as an estimation for the precision matrix that has a leading role when determining the sparsity of the network.

Yuan and Lin [135] proposed a novel penalized likelihood method in order to infer concentration matrix so-called precision matrix. In higher dimensional setting, they attempts to transform the convex optimization problem into maximization-determination problem with the interior point algorithm, motivated from Vandenberge et al. [115]. Also, threshold gradient descent (TGD) regularization [58] procedure that exploits negative log likelihood function as a loss function was suggested with the aim of estimating sparse Gaussian networks and was examined its implementation on the cencored data in the scope of pharmagenomics [88].

The l_1-norm penalized regression can be viewed as a procedure with the combination of two overlapping operations. While the first step emerges as a determination procedure to minimize the objective function or loss function, the second step encompasses the selection of the tuning parameter, referring lambda. Although the natural answer for the latter step is cross-validation, the first is hard to answer since, in higher dimensions, matrix operations require matrix operations such as either diagonalization or inversion referring the reasons behind the issue related to the first remain obscure.

In the sense of estimating precision matrix, Banerjee et al. [11] proposed to perform blockwise coordinate descent procedure by solving the optimization problem for each column on $\mathbf{\Sigma}$ since, in contrast to *graphical lasso* [57], they preferred to estimation of $\mathbf{\Sigma}$, instead of $\mathbf{\Sigma^{-1}}$ with the help of convexity of Lagrangian form. By following this way, they proved that the proper solution for both variance-covariance and precision matrix must hold some attractive properties, including symmetricity and invertibility. So, it is possible to obtain a sensible estimate for the precision matrix in higher dimensions where the number of variables exceeds the number of observations.

glasso is motivated by coordinate-wise algorithm by pursuing fast coordinate descent procedure that is suggested by Friedman et al. [56, 57]. They deliver this algorithm as a competing approach for LARS (homotopy) when dealing with the lasso problems. The idea behind the coordinate descent procedure is that the penalized log-likelihood is maximized as an iterative fashion for each node by re-expressing the problem that resembles lasso regression problem. This coordinate descent procedure is available to implement in the *glasso* [57] and *huge* [139] package in the R programming language. This procedure offers to specify distinct amounts of penalty values for each variable, in return, each inverse covariance is penalized differently. So, we can implement to maximize the penalized log-likelihood in such a form

$$\underset{\Theta}{\text{maximize}}[\log|\mathbf{\Theta}| - \text{Trace}(\mathbf{S\Theta}) - ||\mathbf{\Theta} * \mathbf{\Lambda}||_1], \tag{28}$$

where $\lambda = (\lambda_{ij})_{ij}$, with $\lambda_{ij} = \lambda_{ji}$ and (*) denotes component-wise multiplication. So, this procedure can be interpreted in such a way that different amounts of regularization can be attained to each entry of the precision matrix. Thereby, the optimization of the penalty parameter is achieved via this coordinate-wise scheme by determining λ. In relation to, the larger value of λ leads to sparser networks, on the other hand, the smaller value implies to more connected networks.

In order to determine the sensible λ value, Friedman et al. [56] proposed to apply k-fold cross-validation by taking into either prediction error or likelihood account. Associatively, the large data sets result in smaller CV with fully connected graph,

whereas the smaller data sets obtain the higher CV and correspondingly sparse graph. So, the larger sets indicate overly connected graph that includes much more links or edges between pair of observations, that does not reflect true number of interactions. The validity of the performance of cross-validation in the sense of determining penalty parameter was argued by Meinshausen and Bühlmann [95], suggesting competing alternative for this purpose. Also, it is aimed to penalize likelihood for glasso problem that Banerjee et al. [11] proposed via the following formulae.

$$\lambda(a) = \underset{i \leq j, \sqrt{S_{ii}S_{jj}}}{\text{maximize}} \frac{t_{n-2(a/2p^2)}}{\sqrt{n-2+t^2_{n-2}(a/2p^2)}}. \tag{29}$$

where $t_{n-2}(\tilde{a})$ denotes the $(100-\tilde{a})$ % of a student-t distribution with $(n-2)$ degrees of freedom, with n the sample size, p is the number of variables, and $\boldsymbol{S_{ii}}$ is the estimated variance of the ith variable. a can be replaced with $a/2p^2$ in the case of obtaining more restrictive λ value where the number of variables higher than the number of observations. The notion of the idea proposed by Banerjee[11] is promoted to control the λ value according to false positive rate, so λ should surpass the false positive rate. This procedure highlights the usage of Banerjee's type of determination of regularization parameter.

Under GGM setting of the thesis, we aim to use graphical lasso procedure [56] as a penalized regression technique by combining with our suggested data-dependent techniques, CAIC, CAICF and ICOMP [17, 18, 19, 20] as variable selection criteria in order to correctly represent biological networks. Our three criteria are based on K-L divergence will be examined in Section 2.2.

2.1.3 Multivariate Adaptive Regression Splines and Extensions

GGM is constructed under the parametric modeling approach by holding some assumptions such as normality and the conditional independence to discover interactions among the genes in the biological network. In the field of statistics, the competing modelling method to parametric one is so-called nonparametric approach which

offers the relatively flexible way to represent the relations in this sense. The non-parametric models include various set of models and some of the major methods can be listed as classification and regression trees [21], random forest [23], generalized additive models [68] and multivariate adaptive regression splines [59].

Since MARS is regarded as a powerful regression technique, its various applications have been conducted in the distinct field of statistics, encompassing time series analysis [136], sensitivity analysis of ODE models [87], survival analysis [82]. In addition, it have been used for subset selection purpose [64]. Moreover, MARS have been considered from distinct modeling perspectives that while Denison et al. [37] suggested a Bayesian algorithm for MARS, Banks et al. [12] compared it with LS with polynomials. Finally, it was applied as classification tool for either in the Bayesian setting [73] or in the boosted regression tree construction [35].

In the scope of this thesis, we focus on MARS modeling approach as a nonparametric regression method with the aim of capturing the true network structure. In this thesis, we aim to interfere the model selection procedure of the original MARS at the backward procedure by changing its original criteria with our three data-dependent suggestions. In the Section 3, its adaptive version, namely, LMARS model is implemented to compare the results with that of GGM with respect to numerous criteria. In the statistical literature, distinct derivatives of the MARS modeling technique have been proposed to advance the plain MARS approach in terms of backward elimination procedure, namely, Conic MARS (CMARS) [124], Bootstrapping CMARS [132] and Robust CMARS (RCMARS) [117]. Thereby, in this section, it will be mentioned about the alternative fitting approaches, as well as the method which pure MARS apply. In the context of the thesis, it is preferred to construct separate MARS models for each gene where they are designed to compete with lasso procedure which we use in GGM setting. This perspective leads to Loop-based MARS models, in turn, a set of models for each gene in the network by taking only the main effects of the genes [7, 8]. So, we build models for each gene separately in the network rather than the joint distribution, including all genes in it. Also, the novelity of the use of Loop-based MARS will be mentioned in the Section 3.5.

2.1.3.1 MARS

Mars was launched by Friedman [59] as a flexible modeling technique for the high dimensional data. In place of global parametric modelling, it provides an effective nonparametric local modelling technique, not requiring either any distributional assumption or the relationship between dependent and independent variables. Thus, it is capable of handling with nonlinearities in the high dimensional data by dividing whole region into the subregions. In this sense, it achieves to transform non-differentiable problems to differentiable form by following such a way that it represents the model including nonlinear functions by means of piecewise linear models [6]. It was developed from the two precursor ideas that lies into both recursive partitioning and the projection pursuit, respectively, which suffers from the discontinuity and the additivity in higher dimensions. In contrast, MARS achieves to facilitate the continuous modelling strategy with the help of continuous derivatives, as well as additivity into account in higher dimensional setting [59].

Moreover, MARS pursues two-stage iterative scheme in order to build MARS models, involving forward and backward steps. In the forward step, the model is constructed with as possible as much variables, resulting in deliberately over-fitting the model. The second stage is motivated by discarding the terms which do not lead to considerable increase in the residual sum of squares, in turn model. By following such a way, it aims to choose the simpler models, as well as achieving a good fit to the data. The generic nonparametric regression model is described by

$$y_i = f(\boldsymbol{\beta}, \boldsymbol{x}_i^{'}) + \boldsymbol{\varepsilon_i}. \tag{30}$$

Here, $\boldsymbol{\beta}$ refers to the unknown parameter vector and $\boldsymbol{x_i'}$ $(i = 1, \cdots, n)$ indicates the vector of predictors for the ith case where $\mathbf{y}_i$ is corresponding vector of responses. Also, $\boldsymbol{\varepsilon_i}$ $(i = 1, \cdots, n)$ stands for the vector of random error terms and n denotes the total number of observations. As it is stated beforehand, MARS model does not

state any assumption on the relation of $f(\boldsymbol{\beta}, \boldsymbol{x_i'})$. In the sense of high dimensional modeling, MARS prefers linear basis functions rather than original predictors. These linear basis functions (BFs) are obtained by making some amendments on original predictors and represented as separately by two reflected pairs via

$$(x-t)_+ = \begin{cases} x-t & \text{if } x \leq t \\ 0 & \text{otherwise} \end{cases}, \qquad (t-x)_+ = \begin{cases} t-x & \text{if } x \geq t \\ 0 & \text{otherwise} \end{cases}, \tag{31}$$

where $(x-t)_+$ and $(t-x)_+$ indicate the basis linear functions and t denotes the knot for a pair of such linear functions where each pair is positioned both as positive and as negative side with the help of the two stated equations. When basis linear functions are used for smoothing the nonlinearities in the data, the related knot places as an intersection of them. Therefore, MARS achieves to construct a model with random variables, consisting of a set of basis linear functions and the corresponding knots. Thus, its particular parameter space is expressed as below

$$\boldsymbol{\varphi} = (\boldsymbol{x}-t)_+, (t-\boldsymbol{x})_+ | t \in x_{1,j}, x_{2,j}, \cdots, x_{N,j}, (j = 1, 2, \cdots, p), \tag{32}$$

where *p* represents the number of independent variables. Here, *N* denotes the number of observations. Each pair of a linear functions in the parameter space is the reflected pair of each other. In other words, reflected pairs are defined with a tensor product of the univarite functions. The basis linear functions are represented as

$$\boldsymbol{B}_m(X^m) = \prod_{k=1}^{K_m} [S_{Km}(x_{Km} - t_{Km})]_+. \tag{33}$$

While X_{Km} indicates the dependent variables in the kth truncated linear in the mth basis linear functions, K_m stands for the number of truncated functions and t_{Km} refers to the corresponding knot value in it. Here, S_{K_m} can be 1 or -1. So, the MARS model is stated as

$$y = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m \boldsymbol{B}_m(x^m) + \varepsilon, \tag{34}$$

where $\boldsymbol{B_m}$ refers to mth BF and M stands for the number of BFs in the final model. In the sense of fitting the model parameters, the coefficients for β_m are achieved with the linear regression by minimizing the residual sum of squares.

As we mentioned, MARS model selection is facilitated as a stepwise framework including both forward and backward steps. The forward strategy starts with the intercept β_0 and continues with considering BFs as a set of candidate BFs in order to add them iteratively according to their contribution in the model such a way that a basis function is included in a model if it results in the most amount of reduction in RSS. After reaching the maximum prespecified number of terms, forward step is concluded with the overfitting model. Iteratively, the backward strategy attempts to trim the largest model with the aim of restricting the overfitting. This step works to delete a term from the model if its deletion gives rise to the least amount of reduction in the RSS. Following such a two-stage iterative way, the final MARS model is produced as a model with an appropriate size of variables. In the MARS modeling sense, the best model is associated with the proper number of BFs, as well as the suitable locations for their corresponding knots. For the model selection purpose, MARS applies the *generalized cross validation* (GCV) [32] where it seems an equivalent to the lack-of-fit criterion, or the distance function. The lack-of-fit formulae is given as

$$[\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})]^2. \tag{35}$$

LOF can be viewed as squared error loss, so it is needed to be minimum. In this sense, GCV can be expressed as

$$\text{LOF}(\hat{f}_M) = \text{GCV}(M) = \frac{1}{N}\sum_{i=1}^{N}[y_i - \hat{f}_M(x_i)]^2 \Bigg/ \left[1 - \frac{C(M)}{N}\right]^2 \tag{36}$$

in which y_i is the observed response value, $\hat{f}_M(x_i)$ represents the fitted response value associated with the *i*th observed predictor vector as $\boldsymbol{x_i}$= $(x_{i1}, x_{i2}, \cdots, x_{ip}), i = (1, \cdots, N)$. Also, $C(M)$ is expressed as where M stands for the maximum number of BFs in the model

$$C(M) = \text{Trace}(\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T) + 1. \quad (37)$$

Equation (37) represents the cost complexity measure where **B** is refers to the $(M \times N)$-dimensional data matrix of the ***M*** basis functions. While $(.)^T$ denotes the transpose as used beforehand. So, ***C(M)*** indicates the effective number of parameters which is imposed as a penalty measure for complexity. An adjusted form of ***C(M)*** is implemented by making some changes so that the adaptive cost complexity function can be restated as $C(M)^* = C(M) + dM$ in which ***M*** is the number of non-constant basis functions in the MARS model. Here, *d* in the expression is attributed as a cost for each basis function and refers a smoothing parameter of the procedure. The smaller ***C(M)*** is calculated results in the largest model with too many BFs. Contradictly, a MARS model selection procedure tends to choose a relatively smaller model when the larger cost function is implemented, along with minimum GCV. In the additive MARS modeling concept, the reason that *d* is preferable as "2" that it suffices the expected decrease in the average-squared residual by means of a single knot to offer a piece-wise linear model. This allows MARS to build an additive model by determining the upper limit of the $K_m = 1$ [60]. As taken *d* as a parameter, it can be specified, or estimated by applying either bootstrapping [44] or cross-validation methods [32] so as to control degree of smoothness, as well as bias-variance trade-off. In this context, the ANOVA (analysis of variances) decomposition is established to exhibit the additivity concept of MARS by collecting all combinations of the basis linear functions. It is examined as

$$\hat{f} = \beta_0 + \sum_{k_m=1} \beta_m B_m(x_i) + \sum_{k_m=2} \beta_m B_m(x_i, x_j) + \sum_{k_m=3} \beta_m B_m(x_i, x_j, x_k) + \cdots . \quad (38)$$

In this representation, while the first sum reflects solely the main effects, the second and the third terms take account for the interaction effects over all set of a complete basis functions, respectively, as two-variable and three-variable [78].

James et al. [75] discuss the smoothing splines and presented the minimization of

the penalized RSS as

$$\text{PRSS} = \sum_{i=1}^{N}(y_i - f(\boldsymbol{\beta}, \tilde{\boldsymbol{x}}_{\boldsymbol{i}}))^2 + \sum_{j=1}^{p}\lambda_j \int f_j^{''}(t_j)^2 d_{t_j}, \tag{39}$$

where λ_j is nonnegative tuning parameter. In the PRSS form, $f_j^{''}$ shows the second derivative of the basis functions or piecewise linear functions and it penalizes the variability in f as a penalty. Here, while the term $\sum_i^N (y_i - f(\boldsymbol{\beta}, \tilde{\boldsymbol{x}}_{\boldsymbol{i}}))^2$ is presented as a loss function, the second component in the PRSS accounts for complexity measure. Also, $\boldsymbol{t}_{\text{j}}$ denotes the set of the predictors for *j*th basis function, so *t* contains corresponding knot locations and signs, as well as number of factors in that basis function.

It attempts to control λ that has a key role for determining the number of free parameters, so the level of the flexibility in the model and this smoothing spline representation evaluates the cost for each fit by applying the formulae below.

$$\text{RSS}_{cv}(\lambda) = \sum_{i=1}^{n}(y_i - \hat{g}_{\lambda}^{(-i)}(\boldsymbol{x}_{\boldsymbol{i}}))^2 = \sum_{i=1}^{n}\left[\frac{y_i - \hat{g}_{\lambda}(\boldsymbol{x}_{\boldsymbol{i}})}{1 - \boldsymbol{S}_{\boldsymbol{\lambda} ii}}\right]^2. \tag{40}$$

In Equation (40), $\hat{g}_{\lambda}^{(-i)}(x_i)$ refers to the fitted value for a smoothing spline at x_i, where fit uses all of the training data points, excepting the *i*th (x_i, y_i), the other side $\hat{\boldsymbol{g}}_{\boldsymbol{\lambda}}(\boldsymbol{x}_{\boldsymbol{i}})$ takes all data points into account when fitting the observations. So, the latter notation is evaluated for each of *leave-one-out cross validation* fits, exploiting all data points.

With keeping the analogy to the penalized RSS, in the following subsections, a powerful data mining method will be mentioned as an alternative counterpart of the MARS method.

2.1.3.2 Extensions

CMARS

In the high dimensional context, the overfitting problem is needed to be achieved

via the sensitive optimization procedures. In this sense, in the statistical literature, CMARS is introduced as a model-based alternative of MARS by Weber et al. [124] to compete with it by making the adjustments on the backward [134]. This new procedure is named as CMARS. Here, the "C" is associated with the "Conic", "Convex" as well as "Continuous" altogether, reflecting the characteristics of the procedure [124]. When CMARS operates the forward step to create BFs, like in the that of MARS, it differs from the original MARS in that way it enhances the backward step with applying a conic quadratic optimization technique (CQP) [117, 124, 133] in order to avoid the overfitting. With this regard, it carries out penalized RSS (PRSS) rather than plain RSS that MARS performs. PRSS is involved with its two components : lack-of-fit and the complexity and it is basically presented as below.

$$\begin{aligned}\text{PRSS} = &\sum_{i=1}^{N}(\mathbf{y}_i - f(\beta, \tilde{\mathbf{x}}_i))^2 + \sum_{j=1}^{p}\lambda_m \\ &+ \sum_{m=1}^{M_{max}}\lambda_m \sum_{\substack{|\alpha|=1 \\ \alpha=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V(m)}} \int_{\Phi_m} \theta_m^2 [D_{r,s}^{\alpha}\mathbf{B}_m(z^m)]^2 dz^m, \end{aligned} \tag{41}$$

where $(y_i, \tilde{\boldsymbol{x}}_{\boldsymbol{i}})$ $(i = 1, 2, \cdots, N)$ is the *N*-dimensional vector of the values encompassing both dependent and independent parameters where $\tilde{\boldsymbol{x}}_i$=$(\tilde{x}_{i1}, \tilde{x}_i, \cdots, N)$ $(i = 1, 2, \cdots, N)$ and $(y_1, y_2, \cdots, y_{\text{N}})$ are indicated as dependent and independent ones, respectively. Here, M_{max} represents the prespecified number of BFs, and is needed to be accumulated in the model after the forward step. Moreover, the set of the parameters for *m*th BF such as the number of factors and the associated with the knot locations is re-expressed in a form where $\boldsymbol{V}_{(\boldsymbol{m})}$=$K_j^m|\mathbf{j} = 1, 2, \cdots, K_m$ for *m*th knot and BFs are represented by a vector $\boldsymbol{z}^{\boldsymbol{m}} = (z^{m_1}, z^{m_2}, \cdots, z^{m_{K_m}})^T$. In Equation (41), the notation, $\int_{\Phi^m}$ refers to $\boldsymbol{K}_{\boldsymbol{m}}$-dimensional parallel integration. $\boldsymbol{\lambda}_{\boldsymbol{m}}$ is considered as a nonnegative penalty parameters $(\mathbf{m} = 1, 2, \cdots, M_{\text{max}})$. The partial derivative that contributes to the *m*th BF is established by $D_{r,s}^{\alpha}B_m(z^m) = \frac{\partial^{|\alpha|}B_m}{\partial^{\alpha_1}V_r^m\partial^{\alpha_2}Z_m^s}(Z^m)$ where $\boldsymbol{\alpha}$=(α_1, α_2), $|\alpha| = (\alpha_1, \alpha_2)$ and $\alpha_1, \alpha_2 \in 0, 1$.

We are attempting to discretize the integrals by the reason that the higher dimensional integrals cause difficulty in terms of computation [117, 124, 134]. After applying the discretization procedure, PRSS can be recasted into an open form

$$\text{PRSS} \approx \sum_{i=1}^{N}(Y_i - \boldsymbol{\theta}\boldsymbol{B}(\tilde{\boldsymbol{d}}_{\boldsymbol{i}}))^2 + \sum_{m=1}^{M_{max}} \lambda_m \theta_m^2 \sum_{i=1}^{(N+1)^{K_m}} \left(\sum_{\substack{|\alpha|=1 \\ \alpha=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V(m)}} \int_{\Phi_m} [D_{r,s}^{\alpha} B_m(\tilde{x}_i^m)]^2 \right) \Delta \tilde{x}_i^m, \quad (42)$$

where

$$\Delta \tilde{\mathbf{x}}_i^m = \prod_{j=1}^{K_m} \left(\tilde{\mathbf{x}}_{\substack{l_{\sigma^{k_j}}^{k_j^m}+1, k_j^m}} - \tilde{\mathbf{x}}_{\substack{l_{\sigma^{k_j}}^{k_j^m}, k_j^m}} \right), \quad (43)$$

In Equation (42), $\boldsymbol{B}(\tilde{\boldsymbol{d}}_{\boldsymbol{i}})$=$(1, B_1(\tilde{x}^1), \cdots, B_1(\tilde{x}^M), \cdots, B_{M_{\max}}, (\tilde{x}^{M_{\max}}))^T$ is expressed with $(N \times (M_{\max+1}))$-dimensional matrix where it includes $\tilde{\boldsymbol{d}}$ and $\tilde{x}^1, \tilde{x}^2, \tilde{x}^{M_{\max}}$ as its predictors. Here, $\tilde{\boldsymbol{d}}$ denotes the the vector of predictors, containing the elements as $(\tilde{\boldsymbol{x}}^{\mathbf{1}}, \cdots, \tilde{\boldsymbol{x}}^{\boldsymbol{M}}, \tilde{\boldsymbol{x}}^{\boldsymbol{M+1}}, \cdots, \tilde{\boldsymbol{x}}^{\boldsymbol{M_{\max}}})^{\boldsymbol{T}}$, each is separately connected to the *m*th BF $(m = 1, 2, .., M_{max})$. In addition, $\boldsymbol{\theta}$=$(\theta_0, \theta_1, \cdots, \theta_{M_{\max}})^T$ points is associated with *m*th BF. In the first component in Equation (44), $||.||_2$ refers Euclidean norm.

Here, Equation (43) refers to the coefficients are associated with σ and *l*. Respectively, they represent the dimension and coordinate and are presented together as $l_\sigma^j = l_\sigma (j = 1, 2, \cdots, p)$ for coefficient $\tilde{x}_{l_{N+1}^j, j}$.

Furthermore, $\boldsymbol{\Phi}$ includes all predictors with $\Phi = \bigcup_{\substack{N \\ \sigma^j=0}} \prod_{j=1}^{p} \left[\tilde{x}_{l_{\sigma^j+1}^j}, \tilde{x}_{l_{\sigma^{j+1}+1}^j} \right]$.

Also, we can reformulate PRSS as a form

$$\text{PRSS} \approx \sum_{i=1}^{N} ||Y - \mathbf{B}(\tilde{\mathbf{d}})\theta^2||_2^2 + \sum_{m=1}^{M_{max}} \lambda_m \sum_{i=1}^{(N+1)_{K_m}} L_{im}^2 \theta_m^2, \quad (44)$$

where

$$L_{im} = \left[\left(\sum_{\substack{|\alpha|=1 \\ \alpha=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V(m)}} \int [D_{r,s}^{\alpha} \mathbf{B}_m(\tilde{\mathbf{x}}_{\mathbf{i}}^{\mathbf{m}})]^2 \right) \triangle \tilde{\boldsymbol{x}}_{\boldsymbol{i}}^{\boldsymbol{m}} \right]^{1/2}. \quad (45)$$

Here, $\mathbf{L}$ represents the $((M_{\max} + 1) \times (M_{\max} + 1))$-dimensional matrix as

$$\mathbf{L}=\begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & \mathrm{L}_1 & \dots & 0 \\ \dots & \dots & \dots & 0 \\ \dots & \dots & \dots & \mathrm{L}_{M_{\max}} \end{bmatrix},$$

with $\boldsymbol{L_m}$=$(L_{1m}, L_{2m}, \cdots, L_{(N+1)m}^{K_m})^T$ $(m = 1, 2, \cdots, M_{\max})$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_{M_{\max}})$. Applying the uniform penalization on Equation (44) problem turns into problem in Equation (46). So, PRSS has ability to transform the usual LSE to the Tikhonov regularization problem by combining both the loss and the penalty [5]. It performs a continuous optimization technique so-called conic quadratic programming (CQP) in its the backward elimination procedure with the aim of parameter estimation [117, 124, 133]. Here, λ_{m} conducts a trade-off between bias and variance for each basis function. It can be rearranged in a Tikhonov regularization as

$$\sum_{i=1}^{N} ||y - \boldsymbol{\theta B}(\tilde{\boldsymbol{d}})||_2^2 + \lambda ||\mathbf{L}\boldsymbol{\theta}||_2^2. \tag{46}$$

It is induced that Tikhonov regularization is achieved via the uniform penalty that means λ value is chosen to be the same for each derivative term in the Equation (46) [117, 124, 133, 134]. Here, $\mathbf{L}$ presents the second derivative of *f*, as well as $||.||_2$ denoted Euclidean norm in Equation (46). So, the CQP technique is conducted as a continuous optimization and presented as a ridge procedure [67]. In the sense of conic quadratic programming problem, Equation (46) can be re-exprresed as

$$\underset{t,\theta}{\text{minimize}}\, t, \;\; \text{subject to} ||y - \boldsymbol{\theta B}(\tilde{\boldsymbol{d}})||_2 \geq t, \;\; ||\mathbf{L}\boldsymbol{\theta}||_2 \geq \sqrt{\tilde{M}}. \tag{47}$$

By transforming to two-function optimization problem, it is needed to decide the appropriate bound $\tilde{M}$ which is denoted by $\sqrt{\tilde{M}}$, referring the optimum corner point for "L" shape curve.

A bunch of real life analysis to compare CMARS with MARS have been done, us-

ing different datasets. This concludes that no one dominates the other by giving superior results for all selected criteria. In the case of simulation studies, MARS and CMARS models are generated under two alternative interactions setting relying on two-factor interactions and four-factor interactions, and no statistically significant difference exists between theirs performance on stability, but paired-t test analysis shows that CMARS gives better results in terms of some performance measures such as PRESS, MAE and R^2 [133, 134]. CMARS, implemented on simulated data under both three different sizes and scales results in models that as complex as that of MARS, provides better performance in terms of method free performance measures under medium to large training samples, excepting MSE.

However, MARS produces better results in the medium to small setting. Also, MARS operates better than CMARS in terms of stability. Morever, it does not operate a computationally efficient method since the run time for CMARS takes at least three times higher than that of MARS [124, 133]. A comprehensive applications of CMARS has been conducted and compared with the other data mining methods by Yerlikaya-Özkurt et al. [133].

BCMARS

Bootstrapping CMARS [132, 133, 134] is designed to handle the complexity issue in the final CMARS model. Although, in similar to CMARS, it establishes the models by performing forward selection procedure up to the maximum number of BFs, BCMARS procedure attempts to surpass CMARS in order to select less complex model as a final model. With this regard, it employs bootstrapping, so statistical sampling tool that has ability both to detect the variable as significant or not and to delete parameters which may not considerable contribution to the final model. Also, this computer-intensive method is based on empirical distributions of the parameters which are obtained with collecting samples from the data set with replacement. The motivation behind the use of bootstrapping method grounds for avoiding the data dependency of CMARS with the aim of attaining the most suitable corner point in the sense of conic quadratic optimization [132, 133, 134]. Therefore, it is offered as a useful tool either to achieve trade-off between loss function and the complexity or to

find the closest place to the corner of the L-shape curve for $\sqrt{\tilde{M}}$, representing the maximum point on the curvature.

In conclusion, while BCMARS performs better in terms of accuracy and simplicity of the final models for medium scale datasets, it shows less efficiency computationally since it requires the computer intensive technique, so-called bootstrap. The results are coincided with that of the previous studies, for example, MARS dominates CMARS with respect to stability under the relatively smaller dimensions [12, 133]. On the one side, the bootstrap is that driven to obtain more accurate estimators, is used for assessing accuracy for performance criteria. However, the reason behind it surpasses MARS and CMARS in terms of accuracy is suspicious since the bootstrapping technique suffers from the overfitting problem [75].

RCMARS

The Robust CMARS [99, 100] is facilitated with the purpose of robustfying CMARS so as to reduce the variance of the estimators, simultaneously by decreasing the effects of the parameters in the final model [100]. In the forward step, it is motivated for constructing the model that contains as much as large number of variables so-called M_{max}. So, this procedure ends up with the model that implausibly overfits the data set, sharing similarity in the other counterparts. However, in the backward elimination, it applies the reformed RSS rather than that of CMARS uses, while keeping the complexity part of the PRSS as the same in the CMARS. Hence, it entails the adjustments in both input and output domain on the MARS the reason why it treats either input and output variables as random variables. In other words, RCMARS deals with two particular uncertainty sets [55] which are represented with Confidence Intervals (CI) [99, 100]. In the context of the RCMARS modelling procedure, it is performed with the model which includes uncertainities as below

$$y = f(\breve{\boldsymbol{x}}) + \varepsilon, \tag{48}$$

where y indicates the response variable and $\breve{\boldsymbol{x}}$=$(\breve{x}_1, \breve{x}_1, \cdots, \breve{x}_1)$ refers a vector of

predictors, and ε is an error term with zero mean and finite variance. To build each reflected pairs for the input variables, each $\breve{x}_j$ under the normality assumption takes the form below

$$\breve{\boldsymbol{x}}_j = \bar{x} + \boldsymbol{\xi_j},\ \ (j = 1, 2, \cdots, p). \tag{49}$$

In a similar way, the *m* robustified BFs $(m = 1, 2, \cdots, M_{max})$ are expressed in particularly based on $(\breve{\breve{\boldsymbol{x}}}_{\boldsymbol{i}}, \breve{\breve{\boldsymbol{y}}}_{\boldsymbol{i}})$ $(i = 1, 2, \cdots, N)$ in place of the $(\breve{x}, \breve{y})$ with two corresponding uncertainty sets $U_1 \subseteq \Re^{N \times M_{max}}$ and $U_2 \subseteq \Re^N$. In Equation (49), $(\bar{x})$ induces the mean of the input data or average. With respect to analogy between $(\breve{\breve{\boldsymbol{x}}}_{\boldsymbol{i}}, \breve{\breve{\boldsymbol{y}}}_{\boldsymbol{i}})$ and $(\breve{\boldsymbol{x}}, \breve{\boldsymbol{y}})$, the uncertainty sets are associated with either input data or output data, respectively.

$$\begin{aligned} \breve{\boldsymbol{x}}_{\boldsymbol{ij}} \rightarrow \breve{\breve{\boldsymbol{x}}}_{\boldsymbol{ij}}; \breve{\breve{x}}_{ij} = \bar{x}_j + \Delta_{ij}, |\delta_{ij}| \leq \rho_{ij};\ \ (j = 1, 2, \cdots, p; i = 1, 2, \cdots, N). \\ \breve{\boldsymbol{y}}_{\boldsymbol{i}} \rightarrow \breve{\breve{\boldsymbol{y}}}_{\boldsymbol{i}}; \breve{\breve{y}} = \bar{y} + H_i, |H_i| \leq v_i;\ \ (i = 1, 2, \cdots, N). \end{aligned} \tag{50}$$

Here, $\boldsymbol{\Delta}$ and $\mathbf{H}$ represent the uncertainty sets for either input data or output data, respectively. In addition, ρ_{ij} and v_i show the restriction on the amount of perturbation in each dimension.

BFs are presented with $(\breve{\breve{\boldsymbol{x}}}_{\boldsymbol{i}}, \breve{\breve{\boldsymbol{y}}}_{\boldsymbol{i}})$ as data points, as well as corresponding knot values $\boldsymbol{\tau} = (\tau_{i1}, \tau_{i2}, \cdots, \tau_{ip})$. In the form of the piecewise linear expansion on uncertainty sets with

$$(\breve{\breve{x}} - \tau)_+ = \begin{cases} \breve{\breve{x}} - \tau & \text{if } \breve{\breve{x}} \geq \tau \\ 0 & \text{otherwise} \end{cases}, \qquad (\tau - \breve{\breve{x}})_+ = \begin{cases} \tau - \breve{\breve{x}} & \text{if } \breve{\breve{x}} \leq \tau \\ 0 & \text{otherwise} \end{cases}. \tag{51}$$

The *m*th BF can be rearranged into a multiplicative form via

$$\Psi_m(\breve{\breve{\boldsymbol{x}}}_{\boldsymbol{i}}{}^m) = \prod_{j=1}^{K_m} \left[(\breve{\breve{x}}_{iK_j^m}{}^m - \tau_{k_j^m}) \right]_{-,+}, \tag{52}$$

where Ψ indicates the BF encompassing $\breve{\breve{x}}$ and τ. After applying some modifications

on both uncertainty sets and BFs, PRSS takes the form

$$\text{PRSS} = \sum_{i=1}^{N}(\breve{\breve{y}}_i - f(\breve{\breve{\boldsymbol{x}}}_{\boldsymbol{i}}))^2 + \sum_{m=1}^{M_{max}} \lambda_m \sum_{\substack{|\alpha|=1 \\ \alpha=(\alpha_1,\alpha_2)^T}}^{2} \sum_{r<s}^{r,s\in V(m)} \int_{Q_m} \theta_m^2 [D_{r,s}^{\alpha} \Psi_m(t^m)]^2 dt^m. \tag{53}$$

Here, $\int_{Q_m}$ stands for multidimensional integral. After discretization process is done, PRSS that is incorporated with uncertainty will be defined as

$$\text{PRSS} \approx ||\breve{\breve{y}} - \Psi(\breve{\breve{\boldsymbol{d}}})\boldsymbol{\theta})||_2^2 + \lambda||\boldsymbol{L\theta}||_2^2. \tag{54}$$

In Equation (54), the form of PRSS including two-objective functions uses the analogy with the Tikhonov regularization $\lambda \geq 0$, i.e, $\lambda = \phi^2$. So, it can be solved by CQP [124]. While it achieves the conformity with $\lambda||\boldsymbol{L\theta}||_2^2$ as the part of CMARS, $||\breve{\breve{y}} - \Psi((\breve{\breve{\boldsymbol{d}}})\boldsymbol{\theta})||_2^2$ differs from its close alternatives, dealing with uncertainty sets. To exemplify a robust application of CMARS on the ellipsoidal uncertainty sets, the robust Tikhonov regularization problem is presented by

$$\min_{\alpha}\left[\underset{\substack{v\in U_2 \\ u\in U_1}}{\text{maximize}}||\breve{y} + \mathbf{v} - (\Psi(\breve{\boldsymbol{b}}) + \mathbf{U})\alpha||_2^2 + \phi||\boldsymbol{L\alpha}||_2^2\right], \tag{55}$$

where $\Psi(\breve{\breve{\boldsymbol{b}}}) = \Psi(\breve{\boldsymbol{b}}) + \mathbf{U}$ and $\breve{\breve{\boldsymbol{y}}} = \breve{\boldsymbol{y}} + \boldsymbol{v}$ shows the known uncertainty sets with two constraints $\mathbf{U} \in U_1$, $\mathbf{v} \in U_2$. Here, $\mathbf{U} \in U_1$ and $\mathbf{U} \in U_2$ indicate ellipsoidal uncertainty sets where $\mathbf{U} \in U_1$ =$P^{1/2}u|\mathbf{u} \in R^{N.M_{max}}, ||\mathbf{u}||_2 \leq \rho$ and $\mathbf{v} \in U_2$ = $\mathbf{Q}^{1/2}\mathbf{v'}|\mathbf{v'} \in \Re^N \ ||\mathbf{v'}||^2 \leq \mathbf{v}$, $\mathbf{P}$ and $\mathbf{Q}$ stand for symmetric nonnegative matrices. Also, *u* represents the vectoral form of $\mathbf{U}$ on a ellipsoidal bounded sets.

Robust approximation [14, 63] on the robust CMARS emerges as a durable counterpart of CMARS under the worst-case scenarios, for example, polyhedral or ellipsoidal uncertainty sets [98, 100]. To exemplify the use of RCMARS on a time series data, it is used for forecasting electricity prices of Turkey [133].

Up to know, we have been examined two parametric and nonparametric modelling approaches, namely, GGM and MARS with its corresponding inference procedures and model selection procedures, respectively. In this context, we aim to capture true biological network structure by correctly defining its interactions and the number of genes in the network under two modelling setups. Thereby, we are searching for an optimal model selection or variable selection criteria for higher dimensional context. Under GGM setting, we aim to find an model selection criterion to correctly determine the sparsity of the network and it is presented with glasso procedure as a penalized regression method. Also, under MARS setting, we aim to interfere the model selection procedure of the original MARS at its backward step by replacing GCV with the alternative criteria.

2.2 Model Selection For Biological Networks

In this section, it will be focused on the concept of the statistical model selection. In the sense of identifying the optimal model among the collection of competing models, it is aimed to find the *best approximating* model to the *true* model. In higher dimensional network setting, the selection of the appropriate model is associated with capturing the correct representation for the graphical network, along with specifying the true connections between its attributes. Without any loss of the generality, there are two pivotal characteristics which are imputed to so-called best model, where we always take the interpretability and the accuracy of the predictions into account [114]. While we keep in mind accuracy, it is intended to minimize the difference between the true model and the fitted model [89]. Also, we desire the model as simple as possible with regarding the interpretability issue. According to Occam's Razor, a simpler, i.e., more parsimonious model, is preferable to the others, in turn, not only reducing the complexity, but also providing the highest gain of information [17]. Furthermore, these two characteristics directly entail the determination of the number of variables in the statistical model. Therefore, this results in variable selection procedure where the irrelevant variables should be discarded from the model in order to increase the accuracy while keeping the model as simple as possible. In the literature, the distinct variable selection methods which are operated on the linear regression models relies

on either information criteria, penalized likelihood methods, or background knowledge have been proposed [71]. In manner the need for variable selection arises from the bias-variance trade-off where they are known as the two elements of the squared error, while effecting the model interpretability directly [90]. For example, [75] stated that the complex models lead to the decrease in the bias, as well as the increase in the variance.

In an analogy with the network estimation in higher dimensions, the structure learning procedure contains two indispensable components, including an estimation method with a model selection criterion in order to reflect the true interactions among the correctly defined elements in the network [90]. It is achieved via several selection criteria by specifying the regularization parameter whose aim at controlling the sparsity pattern in the network [141]. In this sense, the choice of the regularization parameter plays a vital role in determining the number of nonzero coefficients in the network. While the fewer number of nonzero coefficients indicates the sparsity of the undirected network, the higher numbers represent the dense networks.

In this setup, we will discuss model selection or variable selection criteria in the high dimensional setting, encompassing the state-of-art techniques based on information criteria and the Bayesian paradigm, namely, AIC (Akaike's Information Criterion) [2] and BIC (Bayesian Information Criterion) [110], respectively. Although they perform well in lower dimensions, they suffer from overfitting in higher dimensions. Despite traditional methods, the others such as RIC [90], StaRS [66], EBIC [53] whose depend on penalized likelihood methods was presented as innovative methods that perform well in the case of higher dimensions [50]. The last but not the least, it will be proposed the applications of the three versatile procedures based on information theory, whose names are CAIC, CAICF and ICOMP, which are served as model selection criteria in the sense of network selection. Although they were coined by Bozdogan [17, 18, 19, 20], they have not been attempted to apply on the network estimation under the $n \leq p$ case where the number of variables, p, far exceeds the number of observations n.

Under more closer inspection of the variable selection, in the context of linear regres-

sion, the challenge arises from the use of the sole RSS when comparing the candidate models. By the fact that the addition of the new parameter will always result in a model with a less RSS, correspondingly, the better model, even if the new variable has a weak correlation with the response, so it is not seen as a sensible criterion. In this sense, the variable selection techniques are in need of a penalty for the additional variables in the model so that they play a paramount role in terms of determining the choice of penalty for each additional terms [90]. The model selection procedures are accompanied with distinct choices of regularization parameter in order to detect the proper number in conjuction with how many variables should be included in the model.

In the linear regression framework, the RSS representation with a constant regularization parameter can be expressed as

$$\beta^*_{k,\lambda} = \underset{\beta}{\text{minimize}} \frac{(\boldsymbol{y} - \boldsymbol{x}\beta)^T(\boldsymbol{y} - \boldsymbol{x}\beta)}{\sigma^2} + \lambda k, \tag{56}$$

where *k* defines the number of nonzero predictors while $(.)^T$ is the transpose and λ stands for the penalty. Here, β represents the coefficient for the specific variable. This expression can be reformed as

$$\frac{\text{RSS}_k}{\sigma^2} + \lambda k. \tag{57}$$

The inclusion of the new variable in a model is justified under the case

$$\frac{\text{RSS}_{k+1}}{\sigma^2} + \lambda(k+1) \leq \frac{\text{RSS}_k}{\sigma^2} + \lambda k. \tag{58}$$

Arranging the equation and putting an estimate $\hat{\sigma}^2$ in place of the population variance, we get

$$\frac{\text{RSS}_{k+1} - \text{RSS}_k}{\hat{\sigma}^2}. \tag{59}$$

Under this representation, model selection or variable selection methods will be clas-

sified into three groups: constant , λ as a function of p, and λ as a combination of both p and k [90]. For instance, AIC takes place in the first with its constant as 2, also BIC is in this group with $\log(n)$, where n is fixed. On the other side, RIC is in the second class since it uses $2\log(p)$ as a regularization parameter. In the literature, there exists the modern methods under the third group that has a tuning parameter as asymptotically equals to $2\log(p/k)$ and twice $2\log(p/k)$, respectively, derived by Foster and Stine [5] and Tibshriani and Knight [120].

However, in the content of this thesis, the differences among the variable selection methods will be discussed with respect to their underlying theories. In this sense, AIC is considered as an information theoretic measure, also the proposed methods, including CAIC, CAICF and ICOMP, are based on the information theory. On the other hand, BIC lies on the Bayesian paradigm, that requires the background knowledge, i.e., a prior information. Furthermore, RIC is viewed as a risk function, which is also based on information theory. Also, StARS whose stability approach is served as a resampling procedure. Finally, EBIC makes BIC adaptive to high dimensional case, in turn, it is based on the Bayesian paradigm . It is important to note that all those are the data-driven variable selection methods on the basis of linear regression setting[49].

2.2.1 AIC

Akaike identified as the information-theoretic, or entropic AIC criterion, accordingly, AIC (Akaike's Information Criterion) takes place under the information-criterion since it seems as an extension of information-theoretic interpretation of the maximum likelihood [2]. Although AIC stems from the need for model selection of the time series modelling, it is considered as well-accepted model selection criterion since it is applicable on very diverse fields, including engineering, operational research and medical research [123], as well as statistics [49, 50]. So, it is formed to select an optimal model among the set of candidate models [17, 19].

The extraction of the information about the true unknown parameter is inferred from the underlying distribution. It is aimed that the difference between the true model and the selected model is small as possible, leading to good inference from the selected model to the true one. As an objective measure of the such difference, the distance between such two models are used for inferential purposes, regarding to the gain of information and the closeness between them.

Boltzmann's [28] generalized entropy, or Kullback-Leibler [83] *information quantity* are two pioneering ideas introduced as a measure of either the information or the distance. While generalized entropy is driven by entropy maximization principle, conversely, it is attempted to minimize the K-L information quantity [17].

Assume that there is a continuous random vector $\mathbf{X}$ whose probability density function represented as $f(\mathbf{x}|\boldsymbol{\theta})$ by k-dimensional random vector $\boldsymbol{\theta} = \boldsymbol{\theta_K} = (\theta_1, \theta_2, \cdots, \theta_k)$ $\boldsymbol{\theta_K} \in \Re^K$. It is supposed that there exists a true parameter vector $\boldsymbol{\theta}^*$ of $\boldsymbol{\theta}$ whose probability density indicated as $f(\mathbf{x}|\boldsymbol{\theta}^*)$. Here, it is aimed that $\boldsymbol{\theta}$ should be as "closest" as to the true parameter vector $\boldsymbol{\theta}^*$. Therefore, it is measurable and defined as "closeness", or "goodness-of-fit" between $f(\mathbf{x}|\boldsymbol{\theta}^*)$ and $f(\mathbf{x}|\boldsymbol{\theta})$ by using generalized entropy B of Boltzmann [28], or K-L [83] information quantity I:

$$B(\boldsymbol{\theta}^*; \boldsymbol{\theta}) = -I(\boldsymbol{\theta}^*; \boldsymbol{\theta}). \tag{60}$$

B is re-expressed as below

$$\begin{aligned} \boldsymbol{B(\theta^*; \theta)} &= E[\log f(\boldsymbol{X}|\boldsymbol{\theta}) - \log f(\boldsymbol{X}|\boldsymbol{\theta}^*)] && (61)\\ &= \int f(\boldsymbol{x}|\boldsymbol{\theta}^*) \log f(\boldsymbol{x}|\boldsymbol{\theta}) dx - \int f(x|\boldsymbol{\theta}^*) \log f(\boldsymbol{x}|\boldsymbol{\theta}^*) dx && (62)\\ &= \boldsymbol{H(\theta^*; \theta)} - \boldsymbol{H(\theta^*; \theta^*)}, && (63) \end{aligned}$$

where E stands for the expectation according to the true distribution $f(\boldsymbol{x}|\boldsymbol{\theta}^*)$ of $\mathbf{x}$ with "log" is the natural logarithm. $\boldsymbol{H(\theta^*; \theta)}$= $\int f(\boldsymbol{x}|\boldsymbol{\theta}^*) \log f(\boldsymbol{x}|\boldsymbol{\theta})$ dx indicates the cross-entropy whose ability to detect the goodness of fit of $f(\boldsymbol{X}|\boldsymbol{\theta})$ to $f(\boldsymbol{x}|\boldsymbol{\theta}^*)$,

$\boldsymbol{H}(\boldsymbol{\theta}^*;\boldsymbol{\theta}^*) \equiv \boldsymbol{H}(\boldsymbol{\theta}^*)$ refers to constant for a given $\log f(\boldsymbol{x}|\boldsymbol{\theta}^*)$, so-called usual Shannon negative entropy.

Keeping an analogy between entropy and K-L quantity, it is preferable to minimize K-L information rather than maximizing the entropy:

$$\boldsymbol{I}(\boldsymbol{\theta}^*;\boldsymbol{\theta}) = -\boldsymbol{B}(\boldsymbol{\theta}^*;\boldsymbol{\theta}) \tag{64}$$

$$= \boldsymbol{H}(\boldsymbol{\theta}^*;\boldsymbol{\theta}^*) - \boldsymbol{H}(\boldsymbol{\theta}^*;\boldsymbol{\theta}). \tag{65}$$

The reason we only need to estimate the cross-entropy or the expected log likelihood is that $\boldsymbol{H}(\boldsymbol{\theta}^*;\boldsymbol{\theta}^*) \equiv \boldsymbol{H}(\boldsymbol{\theta}^*)$ is constant.

It is assumed that $f(\mathbf{x}|\boldsymbol{\theta})$ satisfies the regularity conditions by carrying out the first and second partial derivatives for $\boldsymbol{\theta} \in \Re^K$, resulting in $\boldsymbol{H}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ is twice differentiable at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ $\boldsymbol{H}'(\boldsymbol{\theta}^*,\boldsymbol{\theta})$=0,$\boldsymbol{H}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)$=$-\boldsymbol{J}(\boldsymbol{\theta}^*)$ where $\boldsymbol{J}(\boldsymbol{\theta}^*)$ represents the amount of information coming from $f(\mathbf{x}|\boldsymbol{\theta}^*)$ based on $\boldsymbol{\theta}^*$. In other words, Fisher information is nothing but the second derivative of the K-L information quantity, so analytically $\boldsymbol{H}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ in charge of measuring the curvature at its maximum value, $\boldsymbol{\theta}$=$\boldsymbol{\theta}^*$. In this sense of justifying $\boldsymbol{I}(\boldsymbol{\theta}^*;\boldsymbol{\theta})$ with its paramount properties in its use of statistical information theory will be listed as below:

1. $I(\boldsymbol{\theta}^*;\boldsymbol{\theta}) \geq 0$ whenever $f(\mathbf{x}|\boldsymbol{\theta}^*) \neq f(\mathbf{x}|\boldsymbol{\theta})$.
2. $\boldsymbol{I}(\boldsymbol{\theta}^*;\boldsymbol{\theta}) = 0$ if and only if $f(\mathbf{x}|\boldsymbol{\theta}^*) \neq f(\mathbf{x}|\boldsymbol{\theta})$ i.e., under the condition that the model is true with the possible ranges of $\mathbf{x}$.
3. $\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}$ are independent identically distributed.

K-L information quantity for whole sample $\boldsymbol{I_n}(\boldsymbol{\theta}^*;\boldsymbol{\theta})$=n$\boldsymbol{I}(\boldsymbol{\theta}^*;\boldsymbol{\theta})$. The last property implies the additivity of the K-L information. These properties makes the K-L information useful in terms of attaining the information about the true distribution [17, 18, 19, 20].

The fact that K-L quantity is not observable, but it can be consistently estimable from the observed data. Thereby, mean log likelihood is intended for the usage as a measure of the goodness of fit of a model in order to minimize K-L quantity [20].

$$\boldsymbol{H}(\boldsymbol{\theta}^*;\boldsymbol{\theta}) = E[\log f(\mathbf{X}|\boldsymbol{\theta})], \tag{66}$$

$$= \int f(\mathbf{x}|\boldsymbol{\theta}^*) \log f(\mathbf{x}|\boldsymbol{\theta}) dx. \tag{67}$$

It is supposed that the data are generated from n independent observations from the probability density $f(\mathbf{x}|\boldsymbol{\theta})$ which is stated as k-dimensional random vector $\boldsymbol{\theta}$ = $\boldsymbol{\theta_k}$ =$(\theta_1, \theta_2, \cdots, \theta_k)$. Also, the $\log$ likelihood function with respect to the data via

$$L(\boldsymbol{\theta}) = f(x_1, \cdots, x_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}), \tag{68}$$

$$\ln(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i|\theta), \tag{69}$$

$$\frac{1}{n}\ln(\boldsymbol{\theta}) = \frac{1}{n}\log L(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta}) = \ln_n \boldsymbol{\theta}, \tag{70}$$

where $L(\boldsymbol{\theta})$ refers to $\mathbf{ln}\,(\boldsymbol{\theta})$ called the natural logarithm function and can be rearranged as sum of i.i.d. random variables $\log f(x_i|\boldsymbol{\theta})$ $(i = 1, 2, \cdots, n)$. It is concluded with the average or mean log likelihood of the sample represented with $\ln_n \boldsymbol{\theta}$ which is described as an estimator of the "distance" between the true probability density $f(\mathbf{x}|\boldsymbol{\theta}^*)$ and the model $f(\boldsymbol{x}|\boldsymbol{\theta})$. Aiming to infer $\boldsymbol{I}(\boldsymbol{\theta}^*;\boldsymbol{\theta})$, $\tilde{\boldsymbol{I}}(\boldsymbol{\theta}^*;\boldsymbol{\theta})$ is used as an estimator for this purpose. The equation takes the form

$$\tilde{\boldsymbol{I}}(\boldsymbol{\theta}^*;\boldsymbol{\theta}) = \tilde{\boldsymbol{H}}(\boldsymbol{\theta}^*;\boldsymbol{\theta}) - \tilde{\boldsymbol{H}}(\boldsymbol{\theta}^*;\boldsymbol{\theta}^*). \tag{71}$$

Also, it can be rearranged as

$$\tilde{\boldsymbol{H}}(\boldsymbol{\theta}^*;\boldsymbol{\theta}) = -\tilde{\boldsymbol{I}}(\boldsymbol{\theta}^*;\boldsymbol{\theta}) + \tilde{\boldsymbol{H}}(\boldsymbol{\theta}^*;\boldsymbol{\theta}^*). \tag{72}$$

It is provided that maximizing the expected log likelihood $\tilde{\boldsymbol{H}}(\boldsymbol{\theta}^*;\boldsymbol{\theta})$ asymptotically equals to minimizing the K-L information quantity, $\tilde{\boldsymbol{I}}(\boldsymbol{\theta}^*;\boldsymbol{\theta})$ since $\tilde{\boldsymbol{H}}(\boldsymbol{\theta}^*;\boldsymbol{\theta}^*)$ $\equiv \tilde{\boldsymbol{H}}(\boldsymbol{\theta}^*)$ is ignorable as a constant. Let $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{x})$ represent an estimate of $\boldsymbol{\theta}$ provided by a sample of n observations $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$. The mean log likelihood in $\log(n)$ is stated as a natural estimator of the expected likelihood, $\tilde{\boldsymbol{H}}(\boldsymbol{\theta}^*;\boldsymbol{\theta})$.

$$I(\boldsymbol{\theta}^*;\hat{\boldsymbol{\theta}}) = -B(\boldsymbol{\theta}^*;\hat{\boldsymbol{\theta}}) = \int \log\left[\frac{f(x|\boldsymbol{\theta}^*)}{f(x_g|\hat{\boldsymbol{\theta}})}\right] f(x|\boldsymbol{\theta}^*)dx. \tag{73}$$

In this sense, the risk function is formed as below

$$E_x[I(\boldsymbol{\theta}^*;\hat{\boldsymbol{\theta}})] = \int I(\boldsymbol{\theta}^*;\hat{\boldsymbol{\theta}})\boldsymbol{f}(\boldsymbol{x}|\boldsymbol{\theta}^*)dx. \tag{74}$$

$$2I(\boldsymbol{\theta}^*;\boldsymbol{\theta}_{\boldsymbol{k}}) \cong 2I(\boldsymbol{\theta}^*;\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}), \tag{75}$$

$$2I(\boldsymbol{\theta}^*;\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) \cong ||\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}||_J^2, \tag{76}$$

$$\cong ||\boldsymbol{\theta}^* - \boldsymbol{\theta}_{\boldsymbol{k}}^*||_J^2 + ||\boldsymbol{\theta}_{\boldsymbol{k}}^* - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}||_J^2. \tag{77}$$

Taking the expectation of Equation (77) and multiplying by *n*, we obtain

$$2nE(I(\boldsymbol{\theta}^*;\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})) \cong E[n||\boldsymbol{\theta}^* - \boldsymbol{\theta}_{\boldsymbol{k}}^*||_J^2 + n||\boldsymbol{\theta}_{\boldsymbol{k}}^* - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}||_J^2], \tag{78}$$

$$= n||\boldsymbol{\theta}^* - \boldsymbol{\theta}_{\boldsymbol{k}}^*||_J^2 + [n||\boldsymbol{\theta}_{\boldsymbol{k}}^* - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}||_J^2]. \tag{79}$$

The first term in Equation (79) refers to the bias an the second term is adjusted for measuring the variance of random error $(\boldsymbol{\theta_k^*} - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}})$. In case of sufficiently large n, we get

$$n||\boldsymbol{\theta_k^*} - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}||_J^2 = ||(n)^{1/2}\boldsymbol{\theta_k^*} - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}||_J^2 \overset{a.d.}{\sim} \chi_k^2, \tag{80}$$

where χ^2 represents the chi-square distribution with *k* degrees of freedom. Also, a.d. is referred to the asymptotically distributed. This implies that $E(\chi^2) = k$ and for large *n*, we states the overall risk of a statistical model with its two components as

$$2nE(I(\boldsymbol{\theta^*}; \hat{\boldsymbol{\theta}}_k)) \cong n||\boldsymbol{\theta^*} - \boldsymbol{\theta_k^*}||_J^2 + k \tag{81}$$

$$\cong \delta + k. \tag{82}$$

In Equation (82), the overall risk is designed to measure the deviations from the true parameter vector

$$k^{\eta}K = -2\log\lambda \overset{a.d.}{\cong} E[-2\log\lambda] = E[\chi_v^{'2}(\delta)] = \delta + k. \tag{83}$$

with $E[\chi_v^{'2}(\delta)]$ and δ is obtained in such a form that

$$\delta = n||\boldsymbol{\theta^*} - \boldsymbol{\theta_k^*}||_J^2 \cong -2\log\lambda - v. \tag{84}$$

$$= -2\log\lambda - (K - k). \tag{85}$$

It can be represented as

$$-2nE[B(\boldsymbol{\theta^*}; \hat{\boldsymbol{\theta}})] = 2nE[I(\boldsymbol{\theta^*}; \hat{\boldsymbol{\theta}})] \cong -2\log\lambda - (K - k) + k. \tag{86}$$

The right hand side obviously take this form

$$-2\log\lambda-(K-k)+k=-2\log\frac{L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})}{L(\hat{\boldsymbol{\theta}}_{\boldsymbol{K}})}-(K-k)+k \tag{87}$$

$$-2nE[B(\boldsymbol{\theta}^{*};\hat{\boldsymbol{\theta}})]=2nE[I(\boldsymbol{\theta}^{*};\hat{\boldsymbol{\theta}})] \tag{88}$$

$$\cong -2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})+2k+2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{K}})-K \tag{89}$$

$$\cong k^{\eta}K+2k-K. \tag{90}$$

Rather than taking K-L quantity $I(\boldsymbol{\theta}^{*},\hat{\boldsymbol{\theta}})$ as the loss function, the expected loss function, or $E[I(\boldsymbol{\theta}^{*},\hat{\boldsymbol{\theta}})]$, renders a reasonable estimate for the risk function. The fact that it is valid for at least n is sufficiently large, and *K* and *k* are relatively large integers. The risk function, *R* is stated as

$$R(\tilde{\boldsymbol{\theta}}_{\boldsymbol{K}};\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})=\frac{1}{n}(-2\log\lambda+2k-K), \tag{91}$$

$$k^{v}K=k^{\delta}K=-2\sum_{i=1}^{n}\log\frac{f(x_i|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})}{f(x_i|\hat{\boldsymbol{\theta}}_{\boldsymbol{K}})}+2k. \tag{92}$$

Here, it is aimed to search for $\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}$ that makes the $\boldsymbol{R}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{K}};\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})$ in accordance with $2nE[I(\boldsymbol{\theta}^{*},\hat{\boldsymbol{\theta}})]$ the minimum. In this sense, it is enough to evaluate $k^{v}K$, simultaneously, ignoring the constant term in Equation (101), so the simpler form can be obtained pertaining AIC via

$$\text{AIC}(k)=-2\sum_{i=1}^{n}\log f(x_i|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})+2k \tag{93}$$

$$=-2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})+2k \tag{94}$$

AIC serves as a methodological way when compare *k* models by selecting the model with minimum AIC over $k=1,2,\cdots,K$. AIC(k) presents an *unbiased estimator*

of minus twice the mean expected log likelihood. Correspondingly, $-\frac{1}{2}\text{AIC}(k)$ refers asymptotically an unbiased estimator of the mean expected log likelihood and its likelihood model is $L(k) = -\frac{1}{2}\text{AIC}(k)$.

2.2.2 CAIC

In the statistical information theory, there has been discussed the virtue of the consistency known as a large sample property. Although the AIC scheme does not considered as consistent, it achieves to estimate true distibution by means of minimizing negentropy, or expected log likelihood for large *n* [84]. In particular, AIC aims at controlling the both overfitting and underfitting risks when dealing with the bias arising from log likelihood ratio of maximum likelihood estimates. In this section Consistent Akaike Criterion [17] will be mentioned that CAIC is motivated by enhancing the AIC with two amendments. CAIC was developed by Bozdogan [17] in such a way that it is analytically extended to make AIC consistent and to penalize overfitting more harshly. In other words, it designed to obey the desirable AIC principles, as well as advancing it in two ways. As we introduced before, mean log likelihood encompasses the bias causing from maximum likelihood estimates of the paramaters. In the concept of AIC, the bias is associated with the noncentrality parameter δ known as unknown but deterministic constant. It is obvious that it can be affected by the distributional choices, as well as sample size. Furthermore, it plays a leading role in order to specify the model among the set of them by determining the power of the test that we apply. Thereby, the bias can be reduced somehow by adding some correction factor, like in AIC. From the AIC framework, the test statistic is taken as noncentral chi-squre distribution and the degrees of freedom is defined as an increasing function of the sample size *n* [79].

$$v = a(n)(K - k). \tag{95}$$

with *a(n)* represents the penalty parameter depending on an increasing function of n. It is easy to see that AIC choose *a(n)*=1. When selecting *a(n)*, two issues are needed

to be taken into account [33]. It is aimed to determine the correct dimension should be high in probability, or the correct order for finite samples. In addition, consistency is required as a remarkable property that makes the use of the procedure valid for large *n*. To achieve these two, CAIC exploits *a(n)*=$\log(n)$ along with $v = (K - k)\log(n)$, instead of $a(n) = 1$.

$$\text{CAIC}(k) = -2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) + k[(\log(n)) + 1]. \tag{96}$$

To justify the use of CAIC, the proof is given below starting from Equation (75) in AIC

$$2nE[I(\boldsymbol{\theta}^{*}; \hat{\boldsymbol{\theta}})] \cong \delta + k. \tag{97}$$

The noncentrality parameter differs from the AIC by taking $v = (K-k)\log(n)$ rather than $v = (K - k)$

$$\delta \cong -2\log\lambda - v. \tag{98}$$

$$\delta \cong -2\log\lambda - (K - k)\log(n). \tag{99}$$

Rewriting the Equation (89) in terms of CAIC, we get

$$2nE[I(\boldsymbol{\theta}^{*}; \hat{\boldsymbol{\theta}})] \cong -2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) + k\log(n) + k + 2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{K}}) - K\log(n). \tag{100}$$

$$\text{CAIC(k)} = -2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) + k\log(n) + k + 2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{K}}) - K\log(n). \tag{101}$$

CAIC seems as a consistent estimator of the $2nE[I]$, or twice the expected K-L information and after dropping the constants from Equation (101), the simpler form is obtained as below

$$\text{CAIC}(k) = -2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) + k[(\log(n)) + 1]. \tag{102}$$

2.2.3 CAICF

Under the class of the variable selection relying on statistical information theory, like the others including AIC, CAIC, CAICF was produced to estimate minus twice the expected entropy by Bozdogan [17] and is motivated from the intent to penalize the overparametrization more strongly, as well as achieving the consistency, especially for large n, but at the same time satisfying the Akaike's original principles. Aiming at fulfill these purposes, CAICF differs from its analogues by the way it preserve the term $2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{K}})$ in Equation (112). It behaves $f(\boldsymbol{x}|\boldsymbol{\theta}^*)$ as the true density, rather than $f(\boldsymbol{x}|\boldsymbol{\theta}^*_{\boldsymbol{k}})$. In this sense, unlike AIC, it suggests a different approximation for $L(\hat{\boldsymbol{\theta}}_{\boldsymbol{K}})$ that is used for inferring the likelihood function $L(\boldsymbol{\theta}^*)$ of the true model. By applying maximum likelihood estimates, it is capable of estimating the unknown parameters of both the true and approximate models.

$$\text{CAICF}(k) = -2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) + k[\log(n) + 2] + \log|\boldsymbol{J}(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})|, \tag{103}$$

$$= \text{AIC}(k) + k\log(n) + \log|\boldsymbol{J}(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})|. \tag{104}$$

Suppose that the true parameter vector $\boldsymbol{\theta}^*$ fulfills the restricted model, reducing the number of parameters.

$$\text{MODEL}(k) : \boldsymbol{\theta}_{\boldsymbol{k}} = (\theta_1, \theta_2, \cdots, \theta_k, 0, \cdots, 0), \tag{105}$$

where $\mathcal{O}$ refers the "order of" with $\boldsymbol{\theta}^*_{\boldsymbol{k}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}} = O(n^{-1/2})$ Maximum likelihood estimates, under regularity conditions, MLE $\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}$ of $\boldsymbol{\theta}^*_{\boldsymbol{k}}$ is a sufficient statistic for $\boldsymbol{\theta}^*_{\boldsymbol{k}}$, at least asymptotically. This is obviously proved by the factorization theorem of the likelihood can substantiate the sufficiency [31].

Theorem 1: A maximum likelihood estimator $\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}$ is asysmptotically distributed as

multivarite normal vector $\boldsymbol{\theta_k^*}$ and the covariance matrix $n(\boldsymbol{J})^{-1}$.

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} \overset{a.d.}{\equiv} N_k(\boldsymbol{\theta_k^*}; (n\boldsymbol{J}(\boldsymbol{\theta_k^*}))^{-1}) \tag{106}$$

in accordance with the asymptotic multivariate normal density of

$$g(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) = \frac{|n\boldsymbol{J}(\boldsymbol{\theta_k^*})|^{1/2}}{((2\pi)^{k/2}} \exp\left\{\frac{-1}{2}(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \boldsymbol{\theta_k^*})' n\boldsymbol{J}(\boldsymbol{\theta_k^*})(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \boldsymbol{\theta_k^*})\right\}. \tag{107}$$

So, Equation (108) is presented with $\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}$ as its corresponding the maximum likelihood estimate of $\boldsymbol{\theta_k^*}$ and the inverse covariance matrix, $\boldsymbol{C}(\boldsymbol{\theta_k^*})$, is shown by

$$\boldsymbol{C}(\boldsymbol{\theta_k^*}) = \left[-\frac{\partial^2 \log L(\boldsymbol{\theta_k^*})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}}\right] \cong n\boldsymbol{J}(\boldsymbol{\theta_k^*}), \tag{108}$$

where $\boldsymbol{J}(\boldsymbol{\theta}_k^*)$ stands for the Fisher information matrix at $\boldsymbol{\theta_k^*}$ according to one observation. Particularly for large samples, the likelihood of true density, $f(\mathbf{x}|\boldsymbol{\theta}^*)$ at $f(\mathbf{x}|\boldsymbol{\theta_k^*})$ can be approximated by $L(\boldsymbol{\theta_k^*}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) = g(\boldsymbol{\theta_k^*})L(\boldsymbol{\theta_k^*}) = g(\boldsymbol{\theta_k^*}) \exp \log L(\boldsymbol{\theta_k^*})$ in such a way that the product of Taylor series expansions of $g(\boldsymbol{\theta_k^*})$ and $\exp \log L(\boldsymbol{\theta_k^*})$ around the $\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}$ provides the $L(\boldsymbol{\theta_k^*}; \hat{\boldsymbol{\theta}}_{\boldsymbol{k}})$ as

$$L(\boldsymbol{\theta_k^*}; \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) = \frac{(n^k|\boldsymbol{J}(\boldsymbol{\theta_k^*})|)^{1/2}}{(2\pi)^{k/2}} \exp\left\{\frac{-1}{2}(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \boldsymbol{\theta_k^*})' n\boldsymbol{J}(\boldsymbol{\theta_k^*})(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \boldsymbol{\theta_k^*})\right\}[1 + \mathcal{O}(n^{-1/2})]. \tag{109}$$

The asymptotic sufficiency is obviously indicated by the factorization criterion like that

$$L(\boldsymbol{\theta_k^*}) = h(\mathbf{x})L(\boldsymbol{\theta_k^*}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}), \tag{110}$$

where $h(\mathbf{x})$ is free of the particular parameter vector $\boldsymbol{\theta}$, and $L(\boldsymbol{\theta_k^*}; \hat{\boldsymbol{\theta}}_{\boldsymbol{k}})$ form a basis for

sufficient $\boldsymbol{\theta}_{\boldsymbol{k}}^{*}$, relying on x only via $\hat{\boldsymbol{\theta}}_k$=$\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}(\boldsymbol{x})$. Substituting the $L(\hat{\boldsymbol{\theta}}_{\boldsymbol{K}})$ with $L(\boldsymbol{\theta}_{\boldsymbol{k}}^{*})$ by the way of supposing the $\boldsymbol{\theta}^{*}$ as true parameter vector, along with $\boldsymbol{\theta}_{\boldsymbol{k}}^{*}$ as the restricted or pseudotrue parameter vector and we obtain

$$k^{\eta}K = -2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) + 2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}^{*}). \tag{111}$$

From Equations (110) and (111), we get

$$\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}^{*}) = \log h(\mathbf{x}) + \log L(\boldsymbol{\theta}_{\boldsymbol{k}}^{*}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}), \tag{112}$$

After modifying second term in Equation (112), we obtain

$$\begin{aligned}\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}^{*}) = \log h(\mathbf{x}) + \frac{k}{2}\log(n) + \frac{1}{2}\log|\boldsymbol{J}(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})| - \frac{k}{2}\log(2\pi) \\ -\frac{1}{2}(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \boldsymbol{\theta}_{\boldsymbol{k}}^{*})^{'} n\boldsymbol{J}(\boldsymbol{\theta}_{\boldsymbol{k}}^{*})(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \boldsymbol{\theta}_{\boldsymbol{k}}^{*}) + \log[1 + \mathcal{O}(n^{-1/2})].\end{aligned} \tag{113}$$

By multiplying both sides by 2 and simplifying the equation by the way of using the fact that $\log[1+\mathcal{O}(n^{-1/2})]$ is order $n^{-1/2}$ and $(\boldsymbol{\theta}_{\boldsymbol{k}}^{*}-\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})$ refers of order $n^{-1/2}$, $n\boldsymbol{J}(\boldsymbol{\theta}_{\boldsymbol{k}}^{*})$ is order of n, resulting in $(\boldsymbol{\theta}_{\boldsymbol{k}}^{*})^{'} n\boldsymbol{J}(\boldsymbol{\theta}_{\boldsymbol{k}}^{*})(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \boldsymbol{\theta}_{\boldsymbol{k}}^{*})$ is of order $\mathcal{O}(\infty)$, we attain

$$2\log L(\boldsymbol{\theta}_{\boldsymbol{k}}^{*}) = 2\log h(\mathbf{x}) + k\log(n) + \log|\boldsymbol{J}(\boldsymbol{\theta}_{\boldsymbol{k}}^{*})| - k\log(2\pi) + \mathcal{O}(n^{-1/2}). \tag{114}$$

By ignoring the constant terms such as $h(\mathbf{x})$ and $\mathcal{O}(n^{-1/2})$, we get

$$2\log L(\boldsymbol{\theta}_{\boldsymbol{k}}^{*}) = k\log(n) + \log|\boldsymbol{J}(\boldsymbol{\theta}_{\boldsymbol{k}}^{*})|. \tag{115}$$

Simplifying Equation (111), we get

$$k^{\eta}K = -2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) + (n) + \log|\boldsymbol{J}(\boldsymbol{\theta}_{\boldsymbol{k}}^{*})| \tag{116}$$

Here, we are interested in estimating $\boldsymbol{J}(\boldsymbol{\theta}_{\boldsymbol{k}}^{*})$ by $\boldsymbol{J}(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})$ in Equation (116) since $\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}$ is the MLE of $\boldsymbol{\theta}_{\boldsymbol{k}}^{*}$. In place of the result in Equation (89), we have

$$2nE[I(\boldsymbol{\theta}^{*};\hat{\boldsymbol{\theta}})] \cong -2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) + k\log(n) + \log|\boldsymbol{J}(\boldsymbol{\theta}_{\boldsymbol{k}}^{*})||\boldsymbol{J}(\boldsymbol{\theta}_{\boldsymbol{k}}^{*})| + 2k - K \quad (117)$$

$$\cong -2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) + k[\log(n) + 2] + \log|\boldsymbol{J}(\boldsymbol{\theta}_{\boldsymbol{k}}^{*})||\boldsymbol{J}(\boldsymbol{\theta}_{\boldsymbol{k}}^{*})| - K \quad (118)$$

Reducing this equation, we get

$$\text{CAICF}(k) = -2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) + \text{k}[\log(n) + 2] + \log|\boldsymbol{J}(\boldsymbol{\theta}_{\text{k}}^{*})| \quad (119)$$

$$= \text{AIC(k)} + k\log(n) + \log|\boldsymbol{J}(\boldsymbol{\theta}_{\boldsymbol{k}}^{*})| \quad (120)$$

It is easy to show that the first two factors in Equation (120) is the analogue of the CAIC, and also Schwarz's criterion [110]. It is obvious to indicate that CAICF has ability to make AIC analytically consistent, rather than heuristically depending on the arbitrary choice of the correct model in AIC. It is aimed at penalize more strongly than its counterparts do, especially in large samples. However, like the analogues of the model criteria in the statistical information theory, it attempts to infer minus twice the expected entropy. It is important to note that it does not motivated by the Bayesian approach [17]. The Fisher information matrix, $\boldsymbol{J}(\boldsymbol{\theta}_{\boldsymbol{k}}^{*})$, rendered as the penalty term from CAICF, plays a pivotal role in terms of either theory and application of the CAICF. In this manner, the determination of the correct probability model does not seem as necessary condition to fit the model when implementing the model selection criteria. No matter the true distribution is normal or nonnormal, maximum likelihood procedure produces such consistent estimates for both mean and variance that they obey the normality assumption. That is, the consistency of the mean log likelihood is guaranteed to yield the robust estimation that induces as specification tests. Here, we are interested in checking whether the true probability model is misspecified or not. In this sense, a basic test of the information matrix equivalence is conducted to check the model misspecification. To achieve this purpose, two matrices are defined

as below

$$\boldsymbol{J_n(\theta_k^*)} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(x_i|\boldsymbol{\theta_k^*})}{\partial \boldsymbol{\theta_r}\partial \boldsymbol{\theta_s}}. \tag{121}$$

$$\boldsymbol{R_n(\theta_k^*)} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(x_i|\boldsymbol{\theta_k^*})}{\partial \boldsymbol{\theta_r}}.\frac{\partial \log f(x_i|\boldsymbol{\theta_k^*})}{\partial \boldsymbol{\theta_s}}. \tag{122}$$

The expectations take the following form

$$\boldsymbol{J(\theta_k^*)} = -E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(x_i|\boldsymbol{\theta_k^*})}{\partial \boldsymbol{\theta_r}\partial \boldsymbol{\theta_s}}\right], \tag{123}$$

$$R(\boldsymbol{\theta_k^*}) = E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(x_i|\boldsymbol{\theta_k^*})}{\partial \boldsymbol{\theta_r}}.\frac{\partial \log f(x_i|\boldsymbol{\theta_k^*})}{\partial \boldsymbol{\theta_s}}\right] \tag{124}$$

with $r, s = 1, 2, \cdots, k$. When

$$\boldsymbol{C_n(\theta_k^*)} = \boldsymbol{J_n(\theta_k^*)^{-1}R_n(\theta_k^*)J_n(\theta_k^*)^{-1}} \equiv \boldsymbol{J_n^{-1}R_nJ_n^{-1}}, \tag{125}$$

$$\boldsymbol{C(\theta_k^*)} = \boldsymbol{J(\theta_k^*)^{-1}R(\theta_k^*)J(\theta_k^*)^{-1}} \equiv \boldsymbol{J^{-1}RJ^{-1}} \tag{126}$$

with positive-definite Fisher Information matrix, $\boldsymbol{J(\theta_k^*)}$ and the covariance matrix, $\boldsymbol{C(\theta_k^*)}$. [17] indicates the following fact provided that the model is correctly specified.

Theorem 2(Information matrix equivalence): $f(\mathbf{x}) \equiv (\mathbf{x}|\boldsymbol{\theta}^*)$ for $\boldsymbol{\theta}^*$ in Ω_K, then we obtain $\boldsymbol{\theta}^* = \boldsymbol{\theta_k^*}$ and $\boldsymbol{J(\theta_k^*)}$=$\boldsymbol{R(\theta_k^*)}$ with its covariance $\boldsymbol{C(\theta_k^*)}$ = $\boldsymbol{J(\theta_k^*)^{-1}}$ = $\boldsymbol{R(\theta_k^*)^{-1}}$. If the model is not correctly specified, $\boldsymbol{J(\theta_k^*)^{-1}}$ is not hold to be equal to $\boldsymbol{R(\theta_k^*)^{-1}}$. On the other hand, it is supposed to be correctly defined model, the Hessian matrix can be presented in two such ways that Hessian form and outer product form are represented by $\boldsymbol{J(\theta_k^*)^{-1}}$, $\boldsymbol{R(\theta_k^*)^{-1}}$, respectively. This leads to $\boldsymbol{J(\theta_k^*)^{-1}} - \boldsymbol{R(\theta_k^*)^{-1}} = 0$, but its observable analogues referring its consistent estimators, $\boldsymbol{J(\hat{\theta}_k)^{-1}} - \boldsymbol{R(\hat{\theta}_k)^{-1}} = 0$, are used to employ a test statistic to control whether the model misspecification exist or not. In the statistical information theory, model misspecifications can cause many considerable results, producing inconsistent

estimates for parameters. In this manner, a test of the inconsistency of the estimates for either parameter or covariance matrix can be checked by the same way of conducting a test for a model specification.

The problem in finding the $\boldsymbol{J}(\hat{\boldsymbol{\theta}}_k)$ can be caused from singularity in the matrix, or indefinite, resulting no unique value that makes the K-L information quantity minimum at $\boldsymbol{\theta}^*$, the true parameter vector. In accordance, the expected mean log likelihood can be able to obtain more than one optimal parameters to satisfy the equality $\boldsymbol{z}(\boldsymbol{\theta}) = \boldsymbol{z}(\boldsymbol{\theta}^*)$. Therefore, the inability of the identifiability of $\boldsymbol{\theta}$ concludes with the information matrix that suffer from singularity problem [113]. Furthermore, it is expected that we obtain the estimators with not only the higher variance, but also the lower accuracy. Thereby, the CAICF is offered as versatile model selection criterion since it manage the model selection procedure, simultaneously, operate the control for $\boldsymbol{J}(\hat{\boldsymbol{\theta}}_k)$. After holding the nonsingularity condition of the Fisher information matrix, it will be implemented to produce unique estimates.

2.2.4 ICOMP

In the context of information theory, ICOMP (Information Complexity Criterion) was proposed by Bozdogan [17, 19] in order to select the best approximating model among the set of candidate models given a finite data set. It is aimed that this model selection procedure should end up with the the best model that makes the criterion optimal.

Tracing back to the article Akaike [2] and Bozdogan [17, 18, 19, 20], AIC, under the information-theoretic decision theory, was served as an approximately unbiased estimator of the mean expected log-likelihood of a model [106]. Also, BIC was introduced by Gideon Schwarz [110] as an approximation to the posterior probability of the model by the way the best model refers the smallest one containing the true parameter vector [70]. Furthermore, "leave-one-out" cross-validation method has been indicated to be equal to the AIC [112, 114].

Unlike the other criteria that focus either the only on complexity or information, ICOMP has ability to be comprised of two such concepts as information and complexity theory to evaluate the statistical model, simultaneously. In a statistical model, the amount of information that includes some complexity accounts for not only the nature but also the degree of the intricate connections among the model components. In this sense, the complexity take into the interconnections among the model components account. In the statistical sense, the information-based model selection criteria is aimed at predicting the true model by using its best observable analogue that ensures us the optimal fit to the true one. In this manner, I and COMP in ICOMP, stand for Information and complexity, respectively. Although AIC, a generic example of the information based criterion,is composed of sole the lack of fit and the lack of parsimony, ICOMP, being inspired from information-based complexity index van Emden [119], is designed to embrace the lack of fit, the lack of parsimony, as well as the profusion of complexity. In Equation (86), as we have already stated that the first term, or $-2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})$, refers to the the lack of fit and the other, or $2k$, known as the lack of parsimony, penalizes the number of free parameter. However, keeping the lack of parsimony in the loss function, ICOMP encompasses an additional penalty, so-called profusion of complexity, preferring to penalize covariance complexity rather than the number of parameters directly in AIC.

It is given as

$$\text{ICOMP(IFIM)} = -2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) + 2C(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Model}}) \tag{127}$$

where $-2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})$ is the maximized log-likelihood function with the maximum likelihood estimate of the parameter vector, $\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}$, and C refers a real-valued complexity measure. In this sense, the complexity is assessed by the estimated covariance matrix of the parameter vector of the model, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Model}} = \hat{\boldsymbol{cov}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})$. The covariance matrix in ICOMP loss function is employed by Cramer-Rao lower bound (CRLB) matrix that represented by inverse Fisher information matrix (IFIM), $\hat{\boldsymbol{F}}^{\boldsymbol{-1}}$. Also, it is expressed as below

$$\hat{F}^{-1} = \left\{ - E\left(\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right)_{\hat{\theta}} \right\}^{-1} \tag{128}$$

Here, the matrix of second partial derivatives of the log-likelihood function of the fitted model is computed at the maximum likelihood estimators $\hat{\boldsymbol{\theta}}$.
ICOMP, whose ability of managing the whole parameter space of the model, ensures not only the accuracy but also the optimal precision of the parameter estimates given a statistical data. That is, the estimated variances and covariances are denoted in the diagonal and off-diagonal elements of IFIM, respectively.

$$\text{ICOMP(IFIM)} = -2 \log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) + 2\boldsymbol{C_1}(\hat{\boldsymbol{F}}^{\boldsymbol{-1}}), \tag{129}$$

No matter the model is linear/nonlinear, the above expression pertaining ICOMP is valid for either multivariate or univariate case with

$$\boldsymbol{C_1}(\hat{\boldsymbol{F}}^{\boldsymbol{-1}}) = \frac{s}{2} \log \left[\frac{\text{Trace}\hat{\boldsymbol{F}}^{\boldsymbol{-1}}}{\mathbf{s}} \right] \frac{-1}{2} \log |\hat{\boldsymbol{F}}^{\boldsymbol{-1}}|, \tag{130}$$

where $\boldsymbol{C_1}$ indicates the maximal information complexity of the estimated Fisher information matrix with $\boldsymbol{s}$ =dim$(\hat{\boldsymbol{F}}^{\boldsymbol{-1}})$ = rank$(\hat{\boldsymbol{F}}^{\boldsymbol{-1}})$.

ICOMP is adjusted to give new insights into the functionality of the entropic data-adaptive penalty by not exploiting the fixed constant that AIC and its counterparts' use.

Since ICOMP relies on the information theory, it makes use of two additive terms: one is the lack of fit and the other refers the complexity of the parameter estimates of a model in conjunction with the inference uncertainty and the parametric uncertainty, respectively. This way it is motivated from the entropy maximization principle, or minimizing its negative, it can be represented as an approximation to the sum of two K-L distances [83].

$$\text{K-L}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \sum_{i=1}^{n} \int \log \left[f_i(\boldsymbol{Y_i}, \boldsymbol{\theta}^*) f_i(Y_i, \boldsymbol{\theta}^*) \right] dY_i - \sum_{i=1}^{n} \int \log \left[f_i(\boldsymbol{Y_i}, \boldsymbol{\theta}) f_i(\boldsymbol{Y_i}, \boldsymbol{\theta}^*) \right] dY_i. \tag{131}$$

where a vector $\mathbf{Y}$ of independent observations $Y_1, Y_2, \cdots, Y_n$ is assumed to be produced from the usual multiple linear regression with $Y = X\beta + \varepsilon$.

Supposed that $\boldsymbol{\theta}^*$ refers the vector of $(\boldsymbol{\beta}, \boldsymbol{\sigma^2})$ of the true parameters and $\boldsymbol{\theta}$ denotes the vector of parameters corresponding with any value. The $K(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ represents the the distance between the two densities $f(\mathbf{Y}, \boldsymbol{\theta}^*)$ and $f(\mathbf{Y}, \boldsymbol{\theta})$. The KL term can be expressed by using just its second term, ignoring the its first constant term. Therefore, it turns out the problem that the second term,$-\sum_i^n \boldsymbol{E}[(\boldsymbol{Y_i}, \boldsymbol{\theta})]$ to be unbiasedly estimated through $-\sum_i^n \log \boldsymbol{f_i}(\boldsymbol{Y_i}, \boldsymbol{\theta})$, i.e., minus the log likelihood of the observation computed at θ. Suppose that a restricted model, i.e., $\mathbf{R}$, with its maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{R}}$ can be generated to infer the true paramater $\boldsymbol{\theta}$. For this purpose, $-\sum_i^n \log \boldsymbol{f_i}(\boldsymbol{Y_i}, \boldsymbol{\theta_R})$ is applied as an unbiased observable analogue of the $-\sum_n^i \log \boldsymbol{f_i}(\boldsymbol{Y_i}, \boldsymbol{\theta})$. Also, it provides asymptotic covariance matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}\boldsymbol{R}}$ for MLE $\hat{\boldsymbol{\theta}}_{\boldsymbol{R}}$. This refers to Equation (125).

Focusing on the complexity measure of ICOMP, it is originally inspired by the covariance complexity index of van Emden [119] and proposed to penalize the covariance complexity of the model by using $\boldsymbol{C_0}$,

$$\boldsymbol{C_0}(\boldsymbol{\Sigma}) = \frac{1}{2} \sum_{j=1}^{k} \log \boldsymbol{\sigma}_{jj}^2 - \frac{1}{2} \log |\boldsymbol{\Sigma}|, \tag{132}$$

where *k* denotes the dimension of $\boldsymbol{\Sigma}$ and the σ_{jj}^2 are represented in the diagonal elements of $\boldsymbol{\Sigma}$. When $\boldsymbol{\Sigma}$ is a diagonal matrix, $\boldsymbol{C_0}(\boldsymbol{\Sigma})$ equals to the zero. In case of $\boldsymbol{\theta}$ is a normal random vector with covariance matrix, i.e. $\boldsymbol{\Sigma}$, then $\boldsymbol{C_0}(\boldsymbol{\Sigma})$ can be seen as the K-L distance between multivariate normal density of $\boldsymbol{\theta}$ and the product of the marginal densities of the components of $\boldsymbol{\theta}$. However, $\boldsymbol{C_0}$ whose incapability of in-

variance is restricted under orthonormal transformations. To tackle this, Bozdogan [17, 18, 19] suggested the concept of maximal informational complexity measured by $\boldsymbol{C_1}$.

$$\boldsymbol{C_1}(\boldsymbol{\Sigma}) = \underset{T}{\text{maximize}}\boldsymbol{C_0}(\boldsymbol{\Sigma}) = \frac{k}{2}\log\frac{\text{Trace}(\boldsymbol{\Sigma})}{k} - \frac{1}{2}\log|\boldsymbol{\Sigma}|, \tag{133}$$

where *T* refers the set of orthonormal transformations and *k* stands for the dimension of $\boldsymbol{\Sigma}$.

So, $\boldsymbol{C_1}$ is designed to be invariant in terms of both scalar multiplication and orthonormal transformation. Also, it is known for its monotonicity since $\boldsymbol{C_1}(\boldsymbol{\Sigma})$ is defined as a monotonically increasing function of the dimension k of $\boldsymbol{\Sigma}$.

The complexity of the $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_{\boldsymbol{R}}}$ can be viewed as the K-L distance between the joint density and the product of marginal densities represented by a normal random vector with covariance matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}\boldsymbol{R}}$.

Under orthonormal transformations of that normal random vector, the covariance matrix is guaranteed to be reach its maximum. In other words, under regularity conditions, $\hat{\boldsymbol{\theta}}_{\boldsymbol{R}}$ is approximately normal provided that $\hat{\boldsymbol{\theta}}_{\boldsymbol{R}}$ is certainly normally distributed. Furthermore, the complexity of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_{\boldsymbol{R}}}$ is achieved to occur at its maximum under all orthogonal transforms of the the K-L distance between the joint density and the product of marginal densities for $\boldsymbol{\theta}_{\boldsymbol{R}}$. Thereby, ICOMP is expressed with minus twice sum of log-likelihood $-\sum_i^n \log \boldsymbol{f}_i(\boldsymbol{Y}_i, \boldsymbol{\theta}_{\boldsymbol{R}})$ and its complexity measure, $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_{\boldsymbol{R}}}$.

Complexity of the model $\boldsymbol{R_k}$ is represented by $\boldsymbol{C_1}(\boldsymbol{R_k}) = \boldsymbol{C_1}(\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}, \hat{\boldsymbol{\varepsilon}_{\boldsymbol{k}}})$ with the joint vector of estimated parameters and residuals under model $\boldsymbol{R_k}$, $(\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}, \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{k}})$.

For multiple regression models, the least squares (maximum likelihood) estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$ and the residual vector $\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{k}}$ are assumed to be independent. With the additivity property of $\boldsymbol{C_1}(\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}, \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{k}})$ can be rewritten as $\boldsymbol{C_1}(\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}, \hat{\boldsymbol{\varepsilon}_{\boldsymbol{k}}}) = \boldsymbol{C_1}(\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}) + \boldsymbol{C_1}(\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{k}})$. Also, $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$ is normally distributed by $N(\beta_{(k)}, \sigma^2(X_k'X_k)^{-1})$ in which $\beta_{(k)}$ is the projection of

the true parameter $\boldsymbol{\beta}$ on the space $(\beta_0, \cdots, \beta_k, 0, \cdots, 0)$ and $\boldsymbol{X_k}$ is the matrix of independent variables included in model R_k. So the complexity of $\hat{\boldsymbol{\beta}}_k$ is the complexity of its non-singular covariance matrix [20].

$$\boldsymbol{C_1}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}}) = \boldsymbol{C_1}(\boldsymbol{\sigma^2}(\boldsymbol{X_k'X_k})^{-1}) = \boldsymbol{C_1}((\boldsymbol{X_k'X_k})^{-1}). \tag{134}$$

Equation (134) is achieved via the invariance property of the complexity under scalar multiplication. The problem emerges from n-dimensional residual vector $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{k}}$ of model r_K is normally distributed with $N(0, \sigma^2 P)$ where $\boldsymbol{P} = I - X_k(X_k'X_k)^{-1}X_k'$ and P is singular [111]. This makes the complexity of **P** infinite in such a manner that **P** includes not only *n-q* eigenvalues with value 1, but also *q* eigenvalues equal to 0 where rank$(P) = n - q$ with $q = k + 1$. By this way, the complexity is intended to govern the eigenvalues of the covariance matrix by quantifying the inequalities between them. In the sense of numerical analysis, this plays a pivotal role in order to determine the condition number of the matrix where the lower condition number is though as a good indication for well-conditioned matrix, however, the higher values indicate the ill-conditioned matrix [119]. Here, $\boldsymbol{C_1}$ has ability to manage automatically the condition number by following such a way that the number is taken as an equivalent to the ratio of the maximum to the minimum eigenvalues. Therefore, since it can be identified as a K-L distance, it can be easily combined with the badness-of-fit term whose excitability of K-L distance in the information theory .

This is stated that the complexity of zero for $\hat{\varepsilon}_k$ where the random errors whose covariance matrix is $\boldsymbol{\sigma^2 I_n}$, so it is easy to show that $\boldsymbol{C_1}(\boldsymbol{\sigma^2 I}) = 0$ for $\hat{\varepsilon}_k$ in terms of its projection onto lower-dimensional subspace $\hat{\varepsilon}_k$. We are interesting in minimizing the complexity based information criterion for the multiple regression model, and it is represented by

$$\text{ICOMP}_{a_n}(r_k) = -2\log(\text{maximized likelihood}) + 2a_n\boldsymbol{C_1}((\boldsymbol{X_k'X_k})^{-1}) \tag{135}$$

with a sequence of positive numbers, a_n. For example, the criterion ICOMP_1 is associated with $a_n = 1$ and all *n*.

$$\begin{aligned}\text{ICOMP}_{a_n}(r_k) = n\log(2\pi) + n\log\left[\frac{\text{RSS}_k}{n}\right] + n \\ + a_n q \log\left[\frac{\text{Trace}((\boldsymbol{X_k'X_k})^{-1})}{q}\right] \\ - a_n \log\left[\det(\boldsymbol{X_k'X_k})^{-1}\right].\end{aligned} \tag{136}$$

Also, RSS_k is the residual sum of squares of the model r_k. It is pointed out that, in the case of $q = 1$, $\text{ICOMP}_1(r_k)$ or $\text{ICOMP}_{a_n}(r_k)$ results in $-2\log(\text{maximized likelihood})$. Furthermore, the other measure of a model complexity for a model r_k, the estimated parameters for the complexity is represented by the vector, i.e., $(\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}, \hat{\boldsymbol{\sigma}}^{\boldsymbol{2}}_{\boldsymbol{k}})$ with $\hat{\sigma}^2_k = \text{RSS}_k/n$. It is assumed that $(\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$ and $\hat{\boldsymbol{\sigma}}^{\boldsymbol{2}}_{\boldsymbol{k}})$ are independently estimated parameters and its associated covariance matrix is presented as

$$\mathbf{Q} = \begin{pmatrix} \boldsymbol{\sigma^2}(\boldsymbol{X_k'X_k}) & 0 \\ 0 & 2\boldsymbol{\sigma^4}(\frac{n-q}{n^2}) \end{pmatrix}.$$

As $n \to \infty$, $\boldsymbol{Q}$ is asymptotically equals to the inverse Fisher information matrix IFIM arranged by

$$\mathbf{F}^{-1} = \begin{pmatrix} \boldsymbol{\sigma^2}(\boldsymbol{X_k'X_k}) & 0 \\ 0 & 2\boldsymbol{\sigma^4}/n \end{pmatrix}.$$

The matrix $\mathbf{Q}$ is induced by the distribution of RSS_k/σ^2 [111]. When the estimated parameters whose covariance matrix could not written in closed form, the complexity of $(\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}, \hat{\boldsymbol{\sigma}}^{\boldsymbol{2}}_{\boldsymbol{k}})$ is employed with an estimated inverse-Fisher information matrix. Also, it is represented as $\boldsymbol{C_1}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}\hat{\boldsymbol{\sigma}}^{\boldsymbol{2}}_{\boldsymbol{k}})$=$\boldsymbol{C_1}(\hat{\boldsymbol{F}}^{\boldsymbol{-1}})$ where $\hat{\boldsymbol{F}^{-1}}$ and $\hat{\boldsymbol{\sigma}}_{\boldsymbol{k}}$ is used rather than $\boldsymbol{F^{-1}}$ with $\boldsymbol{\sigma}$. Correspondingly, the covariance matrix of $(\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}, \hat{\boldsymbol{\sigma}}^{\boldsymbol{2}}_{\boldsymbol{k}})$ is observable, it is indicated as $\boldsymbol{C_1}(\hat{\boldsymbol{Q}})$ with $\boldsymbol{C_1}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{k}})$. These two complexities are obtained as so closely with each other when the ratio q/n is small.

So, ICOMPIFIM_{a_n} and ICOMPCOV_{a_n} are represented as below

$$\text{ICOMPIFIM}_{a_n}(m_k) = -2\log(\text{maximized likelihood}) + 2a_n \boldsymbol{C_1}(\hat{\boldsymbol{F}}^{-\mathbf{1}}) \tag{137}$$

$$= n\log(2\pi) + n\log\left[\frac{\text{RSS}_k}{n}\right] + n \tag{138}$$

$$+ a_n(q+1)\log\left[\frac{\text{Trace}(\boldsymbol{X_k'X_k})^{-1} + 2\hat{\boldsymbol{\sigma}}_{\boldsymbol{k}}^{\mathbf{2}}/\boldsymbol{n}}{q+1}\right]$$

$$- a_n \log\det((\boldsymbol{X_k'X_k})^{-\mathbf{1}}) - a_n\log\left(\frac{2\hat{\boldsymbol{\sigma}}_{\boldsymbol{k}}^{\mathbf{2}}}{n}\right),$$

$$\text{ICOMPCOV}_{a_n}(m_k) = -2\log(\text{maximized likelihood}) + 2a_n \boldsymbol{C_1}(\hat{\boldsymbol{Q}}) \tag{139}$$

$$= n\log(2\pi) + n\log\left[\frac{\text{RSS}_k}{n}\right] + n \tag{140}$$

$$+ a_n(q+1)\log\left[\frac{\text{Trace}(\boldsymbol{X_k'X_k})^{-1} + 2\hat{\boldsymbol{\sigma}}_{\boldsymbol{k}}^{\mathbf{2}}(\frac{n-q}{n^2})}{q+1}\right]$$

$$- a_n \log\det((\boldsymbol{X_k'X_k})^{-\mathbf{1}}) - a_n\log\left(2\hat{\boldsymbol{\sigma}}_{\boldsymbol{k}}^{\mathbf{2}}\Big(\frac{n-q}{n^2}\Big)\right).$$

2.2.5 BIC

BIC, motivated from the Bayesian framework, was proposed as a model evaluation tool by Gideon Schwarz [110]. Furthermore, it is aimed to find the smallest model containing the true model whose highly depend on the choice of prior in Bayesian paradigm. In this setup, BIC is not based on the K-L distance which from AIC arises.

In BIC, the regularization parameter, λ, relies on sample size n rather than constant value like in AIC. For $n \geq 8$, it is designed to penalize more strongly than does AIC. Also, it is presented as

$$\text{BIC}(k) = -2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) + k\log(n). \tag{141}$$

Like in AIC, BIC attempts to control both the lack of fit and the lack of parsimony, simultaneously. Furthermore, it is assumed that the observations are independently and identically distributed. Also, it entails asymptotically consistency property with the help of maximum likelihood estimator so that it achieves to choose the model truly with the probability reaching to one [103]. On the other hand, the usual BIC seems more liberal for model selection as the model space increases [30].

In coding theory, BIC is coincided with the Minimum Description Length (MDL) [105].

2.2.6 EBIC

The difficulty in consistency arises in the case of small-*n*-large-*p*. As the number of covariates, *p* increases, BIC procedure leads to inconsistency. The reason for EBIC was designed to produce consistent results in higher dimensional settings [53].

$$\text{EBIC}_{\gamma}(k) = -2\log L(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) + k\log(n) + 2\gamma\log\eta(K_j) \tag{142}$$

$\hat{\boldsymbol{\theta}}(\boldsymbol{k})$ is the maximum likelihood estimator of $\boldsymbol{\theta}(\boldsymbol{k})$ and from the Bayesian framework, $P(k)$ represents the prior probability of models. Note that EBIC differs from BIC such that it prefers to assign probability $\eta^{\xi}(K_j)$ to $p(K_j)$ rather than $\eta(k_j)$ $(j = 1, 2, \cdots, K)$. The prior probability over $\boldsymbol{s} \in S_j$ is selected with probability $\eta^{-\tau}(K_j)\ \tau = 1 - \xi$. In $\text{BIC}_{\eta}(k)$, the first two terms are presented as Laplace approximation to $-2\log(m(\mathbf{Y}|\mathbf{k}))$ and the other term is constant with a constraint.

In this framework, given $\mathbf{k}$ and the prior density of $\boldsymbol{\theta}(\boldsymbol{k})$ in $\pi\theta(k)$, the posterior probability is yielded as

$$P(\mathbf{k}|\mathbf{Y}) = \frac{m(\mathbf{Y}|\mathbf{k})p(k)}{\sum\limits_{k\in K} p(k)m(\mathbf{Y}|\mathbf{k})}, \tag{143}$$

where $m(\mathbf{Y}|\mathbf{k})$ is the likelihood of the model space *K* with

$$m(\mathbf{Y}|\mathbf{k}) = \int f(y; \boldsymbol{\theta}(\boldsymbol{k}))\pi\boldsymbol{\theta}(\boldsymbol{k})d\theta(k). \tag{144}$$

Under Bayesian perspective, $\sum_{k\in K} p(k)m(\mathbf{Y}|\mathbf{k})$ refers a constant whose optimal prior entails $k^* = \underset{k\in K}{\text{maximum}} m(\mathbf{Y}|\mathbf{k})p(k)$. Supposed that the model *K* is divided as equal dimensions in such a way that $\bigcup_{j=1}^{K} K_j$, each subspace K_j over **K** is assigned to an equal probability $p(k|K_j) = 1/\eta(K_j)$ with totally j covariates $\eta(k_j) = \binom{K}{j}$.

2.2.7 modified RIC

RIC, which was motivated by risk inflation criteria, was served as a canonical variable selection procedure under the multiple regression [52].

In the decision making context, RIC was designed to be interested in predictive risk function. On the other hand, in the statistical assessment context, this is coincided with the expected loss function. Since the risk function will be expressed in terms of the expected squared error of prediction, where *X* represents the future prediction values, its risk function, $R(\beta, \hat{\beta})$ will be presented as

$$R(\beta, \hat{\beta}) = E_\beta |X\hat{\beta} - X\beta|^2. \tag{145}$$

Suppose that *X* is taken as fixed values from the generic multiple regression model, $Y = X\beta + \varepsilon$ with $\boldsymbol{X} = [X_1, X_2, X_2, \cdots, X_p]$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$

$$\text{RI}(\hat{\beta}) = \underset{\beta}{\text{maximize}} \frac{E||\mathbf{X}\beta - \mathbf{X}\hat{\beta}||^2}{E||\mathbf{X}\beta - \mathbf{X}\hat{\beta}^*||^2} \tag{146}$$

In the RI function, while nominator is nothing but the usual least square function, the denominator work on the restricted parameter space by taking into the only nonzero $\beta_i's$ account where $\hat{\beta}^*$ is inferred from the $\boldsymbol{\beta}$ vector.

In this case, the smaller RI we have, the better performance we obtain. Correspondingly, in the canonical variable selection, the model achieves to minimize the RI. This is because $\lambda = 2\log(p)$. In the decision making process, this indicates that the model reaches its minimax via RI under the assumption that the predictors are orthogonal [52].

Lysen [90] suggested the use of randomly rotated matrices rather than that of the actual data. In this sense, a permutation matrix is known as a special case of a rotation matrix. The only difference is caused by the stringency of an assumption which we make. For instance, whereas the randomly rotated matrices expect a spherically symmetric error distribution, randomly permuted matrices solely assume that the observations come from an independently distributed arbitrary errors. Therefore, Lysen [90] introduced as a data-dependent variable selection scheme we which call the *permuted inclusion criterion* by preferring to permutations than rotations since it is easy to calculate.

This scheme is motivated from a data generation procedure where the predictor space $\mathbf{X}$, X_j is augmented from the original predictor space, *X*. In this setup, covariance structure is not affected by such procedure, so it remain unchanged. For augmented data, the correlation structure is exploited in the case of variable selection since it is also the same as in the original dataset.

In terms of deciding the penalty parameter α, permuted inclusion criterion is asymptotically coincided with the risk inflation criterion of $2\log(p)$. However, the choice of a tuning parameter, α is based on a specified cutoff value, a quantile of a distribu-

tion. Thereby, this procedure is named as modified RIC with $\alpha = 2\log(p/k)$ and is associated with 0.95 quantile in order to select a variable set.

Unlike the state-of-art variable selection procedures such as AIC or BIC, the modified RIC promotes to eliminate the irrelevant $X_i's$ by reducing theirs' associated $\beta_i's$ to the exactly zero.

Unlike StARS, which uses subsampling or cross-validation, it is interested in directly estimating the optimal regularization parameter by changing the p value with the aim of constructing the true graphical representation with its true connections. However, it suffers from the underselection.

2.2.8 StARS

The undirected graph can be presented in terms of its edges and nodes $\mathbf{G} = (\mathbf{\Gamma}, E), while \mathbf{\Gamma}$ ($\Gamma = 1, \cdots, p$) refers the set of vertices, E denotes the set of edges in the graph. In order to infer the adjacency matrix of the graph **G**, it will be used the **E** notation, also. Suppose that a random vector is presented by $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)$ corresponding with Gaussian p. It is assumed that the absence of an edge between the pair of nodes such as (a, b) implies that X_a and X_b are conditionally independent given the rest $\boldsymbol{X_{\Gamma a,b}}$, provided that $\omega_{ab} = 0$ where $\boldsymbol{\omega}=\boldsymbol{\Sigma^{-1}}$.

$$\hat{\omega} = \underset{\omega \geq 0}{\text{minimize}} \left\{ -\ln \omega + \lambda ||\omega||_1 \right\}. \tag{147}$$

Friedman et al.[56] provides an efficient algorithm to evaluate $\hat{\boldsymbol{\omega}}(\boldsymbol{\lambda})$ with a collection of λ's changing from small to large mentioned in Section 3.4.

StARS is operated in order to decide λ associated with $\hat{\boldsymbol{G}}(\boldsymbol{\lambda})$. As the λ increases, the more sparse network we obtain. Correspondingly, the smaller Λ result in sparse

networks by taking that $\Lambda = 1/\lambda$. In particular, $\Lambda = 0$ denotes to the graph with no edges. Over a grid of regularization parameters $\boldsymbol{G_n} = \Lambda_1, \cdots, \Lambda_K$, it is intended to find the optimal $\hat{\Lambda} \in G_n$ for the parameter selection purposes. The reason behind it attempts to "overselect" rather than "underselect", it is driven from choosing $\hat{\boldsymbol{E}}(\hat{\boldsymbol{\Lambda}})$ encompasses the true graph $\mathbf{E}$ with high probability. In manner the StARS procedure is designed to find Λ pertaining to stability.

It is assumed that N random subsamples $\boldsymbol{S_1}, \cdots, \boldsymbol{S_N}$ are taken from $\boldsymbol{X_1}, \cdots, \boldsymbol{X_n}$, each size of $t = t(n)$ with $1 \leq t(n) \leq n$. There exists totally $\binom{n}{t}$ subsamples, each is randomly drawn without replacement. For each $\Lambda \in G_n$, it is ability to build a graph for each subsample in conjunction with N estimated matrices such that $\hat{E}_1^t(\Lambda) \cdots, \hat{E}_N^t(\Lambda)$.

On closer inspection, $\psi_{ab}^{\Lambda}(.)$ is denoted, along with one edge (a, b) and a particular value of Λ. $\psi_{ab}^{\Lambda}(S_j) = 1$ refers to the existence of edge between (a, b), on the other hand, $\psi_{ab}^{\Lambda}(S_j) = 0$ refers to its absence. This indicates that $\theta_{ab}^t(\Lambda) = P(\psi_{ab}^{\Lambda}(X_1, \cdots, X_n) = 1)$. It is aimed to infer $\hat{\theta}_{ab}^t(\Lambda)$ using the such equality that $\hat{\theta}_{ab}^t(\Lambda) = \frac{1}{N} \sum_{j=1}^{N} \psi^{\Lambda}(S_j)$.

In the StARS procedure, the parameters are denoted by $\xi_{ab}^t(\Lambda) = 2\theta_{ab}^t(\Lambda)(1-\theta_{ab}^t(\Lambda))$. Instead of the parameters, its estimated analogues are used and represented by $\hat{\xi}_{ab}^t(\Lambda) = 2\hat{\theta}_{ab}^t(\Lambda)(1 - \hat{\theta}_{ab}^t(\Lambda))$. Here, it is obvious that the main purpose of the stability procedure lies on calculating the total instability by taking mean of all edges by the way $\hat{D}_t(\Lambda) = \sum_{a \leq b} \hat{\xi}_{ab}^t / \binom{p}{2}$. It is easy to show that $\hat{D}_t(0) = 0$ at zero boundary, while $\hat{D}_t(\Lambda)$ tends to change simultaneously with Λ. Although the increase in Λ results in denser graph, this does not reflect the true nature of the gene networks. Therefore, the choice of $\hat{D}_t(\Lambda)$ is determined in a methodological way that $\bar{D}_t(\Lambda) = \underset{0 \leq t \leq \Lambda}{\text{maximize}} \hat{D}_t(b)$. Specifically, $\hat{\Lambda}_S = \text{maximize} \left\{ \Lambda : \bar{D}_t(\Lambda) \leq \beta \right\}$ with a prespecified boundary β.

Although choosing Λ is closely depend on the choice of β, β takes the default value as 0.05. The reason why the StARS procedure relies on subsampling scheme, it is need to be determined to the effective sample size, t, for each selected graph. For

instance, the subsampling block size t has also greater effect on $\hat{\boldsymbol{E}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\omega}}, \hat{\boldsymbol{D}}$.

In the *huge* package in R [139], StARS [66] was presented as a natural way to determine regularization parameter with the aim of finding the optimal network so that it can be applied with three estimation methods in higher dimensions. Despite bootstrap, it offers a random sampling scheme without replacement for specifying the Λ for high dimensional networks. Although this procedure provides a theoretically efficient way to do so, it suffers from "overselection".

After giving the details of the several model selection criteria for high dimensional context in this section (Section 2.1.4), we will continue with Section 3. In Section 3, we intend to explain the novelty which we introduce under two modelling approaches: GGM and LMARS. Firstly, we apply our data-driven model selection criteria (CAIC, CAICF, ICOMP) on graphical lasso algorithm in order to determine the sparsity for biological network structure. Although RIC[85], EBIC[51] and StARS[66] are preferred as high dimensional model selection criteria and applied with graphical lasso algorithm to select optimal λ parameter when constructing Gaussian graphical models, we prefer to use our suggested K-L information based criteria for sparse network structure, each criterion is combined with graphical lasso procedure and plays an important role for detecting interactions between genes in the network structure. Secondly, it is aimed to use our (K-L) information based criteria at the backward step of the model selection procedure of the LMARS models, although LMARS models were originally introduced with its GCV[30] criterion so as to find an optimal model [6, 7].

CHAPTER 3

APPLICATIONS

3.1 Descriptions of Simulated Data

In this section, several simulation scenarios in high dimensional settings are generated so as to compare data-driven model selection criteria in terms of distinct accuracy measures mentioned in Section in 3.1.3. Our comparisons are separated into two parts according to modeling approaches: parametric and non-parametric.

Firstly, the parametric way of modeling the high dimensional networks, namely, GGMs are constructed under two versatile architectural structure: random and scale-free. Also, the estimation of $\mathbf{\Sigma^{-1}}$ are done by preferring the exact approach, namely, graphical lasso [56] with its blockwise coordinate descent algorithm rather than an approximate option, or MB [95] to do so. The use of log-likelihood in a penalized fashion makes glasso different than MB since the latter could not produce the maximum likelihood estimator. Therefore, glasso offers a way to exploit the data-dependent model selection criteria as a penalty parameter, including EBIC, AIC, BIC, CAIC, CAICF and ICOMP, unlike MB approach. It is concluded that the latter is not applicable for data-adaptive criteria whose based on likelihood [56]. Simulation scenarios are accomplished with 50 observations under three dimensional settings, encompassing 50, 100, 500, for each model selection criterion, including AIC, BIC, RIC EBIC, StARS, CAIC, CAICF, and ICOMP.

Secondly, in the second part of this chapter, we aim to interfere the original model selection procedure of the MARS by the way we replace the our suggestions with GCV. So, the loop-based MARS algorithm is presented with both its original forward

stage and the adjusted backward stage. It is done under two structural form of the network: random and scale-free. Under these settings, the suggested data-adaptive model selection criteria, namely, CAIC, CAICF and ICOMP [17, 20] are compared with state-of-art model selection criteria: AIC [2], BIC [110] and GCV [32]. There are distinct (n $\times$ p)-dimensional settings by including p 50, 100, 500 with 50 observations for each.

All GGM simulations are achieved via the undirected graphical estimation packages: glasso (graphical lasso) [56] and huge [139] packages in the R programming language 3.5. On the other hand, LMARS [6, 9] simulations are done in R by its two specific packages: *earth* package [96] in R, the MARS modelling stage, as well as huge package in the data generation process.

In particular, it is intended to show the novelity how we achieve to insert CAIC, ICOMP, CAICF into both glasso and the adjusted loop-based MARS algorithm. Also, the use of the suggested methods in high dimensional settings.

Our analyses are generally done under two distinct topologies: scale-free and random networks [13] seen from Figure 3.1 and Figure 3.2. Although random networks have been developed previously in the field of graph theory, the scale-free networks are the most representative for the cellular networks among them [13]. Each network setup is based on deep mathematical background. For instance, in the random networks, it is assumed that each pair of nodes is connected with the probability p, leading to most of the genes have the same number of links. On the other hand, scale-free networks are ruled by the power-law degree distribution with probability P^{τ} whose τ refers to the degree component. This is the reason for obtaining not only a relatively small number of nodes whose probability of being connected are higher, but also the most nodes having the less probability than the other counterparts in the random networks. Therefore, in the scale-free networks, interactions are much more concentrated on a relatively small number of nodes known as hubs. Overall, scale-free network structure is coincided with our expectation from the systems biology [13], so our main objective is to fit our novel algorithms into scale-free network with the help of the probability theory behind it.

Figure 3.1: The simulated gene networks show scale-free and random, respectively. They are produced with 500 genes by CAICF.

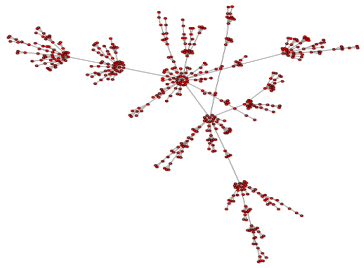
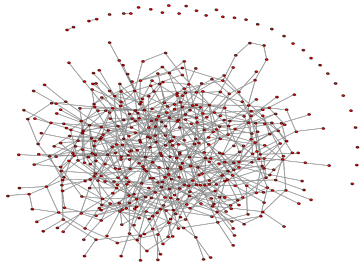
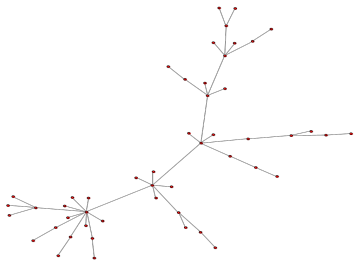
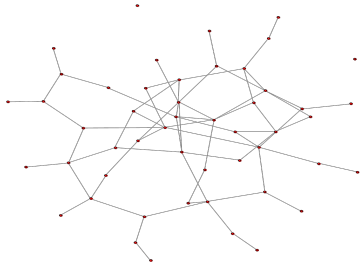
Scale-free	**Random**
	

Figure 3.2: The simulated gene networks show scale-free and random, respectively. They are produced with 50 genes by RIC.

Scale-free	**Random**
	

3.2 Description of Real Data

We apply our likelihood based model selection suggestions to six real datasets where the first two are known as benchmark datasets. While the rest refer to the microarray

data, each contains $p = 11$ genes with different number of observations, except that the Dataset 6 including $p = 9$ genes.

I. We use two benchmark datasets: gene expression and cell-signaling data. The former is based on Affymetrix GeneChip microarrays for gene expression levels and contains $p = 100$ genes and $n = 60$ observations. On the other hand, the second includes $p = 11$ proteins and $n = 11672$ cells. Our dataset resembles a data set known as flow cytometry of Sachs et al. [3]. We use it to produce an undirected network via GGM [56].

   On the other hand, the four microarray datasets whose number of genes are 11 are described in the study of Bahçivancı et al. [10] in details. These are mentioned as Data II, III and IV.

II. **E-GEOD-9891-Transcription Profiling of 285 Human Ovarian Tumour Data:** The data consist a cohort of 285 patients whose epithelial ovarian, primary peritoneal, or fallopian tube cancer has been diagnosed between 1992 and 2006. This data set is mainly extracted from the Australian Ovarian Cancer Study with totally 285 patients. It is aimed to determine the novel subtypes of the overian tumour in relation to both clinical and pathological features.

III. **E-GEOD-63678-Expression Data from Vulvar, Cervical, Endometrial, Carcinoma Tissue:**
   The data include totally 35 samples where 18 cancer samples are from cervical, endometrial, and vulvar, respectively, 5,7 and 6. The others whose 17 normal samples from each type of these cancers are hybridized with the cancer samples on the Affymetrix HG133-A-2.0 platform with 12.000 distinct microarray chips which are represented by different genes. It is aimed to detect the similar features among these cancer types in the embryonic stem cells and the newly discovered cell population of the squamocolumnar junction of the cervix, which it hosts the early cancer's events.

IV. **E-GEOD-81248-Expression Data from HEY Cells:**
   This dataset includes 12 observations for 11 core genes, where each is extracted from mRNA expression on Affymetrix U133 Plus 2 chips. In this dataset, two

samples of distal naive cells whose local cells include both unstimulated (control) and stimulated (LPS or poly(I:C)) exomes so as to indicate whether the exomes with TLR simulated cells have an ability to explain the TLR activation in distal cells in vitro or not.

V. **E-GEOD-48926-Expression data from C33-A cell line (cervix carcinoma cell line)**
In this dataset, there exist 9 genes out of 11 core genes where they are labeled as MAP2K1, MAPK1, CEBPB, CTNNB1, TFAM, PDIA3, IMP3, ERBB2 and CDH4. Also, the levels of gene expression are extracted and hybridized on the Affymetrix platform. Here, C33-A cells are taken from three different cultures where the cells are grown exponentially [3].

3.3 Accuracy Measures

In this section, we identify the performance measures in order to compare performances of the model selection criteria that have mentioned in the Section 3.3. In particular, our focus is on the accuracy assessment whose well-known tools are known as accuracy (acc), precision (pre), recall (rec) and F-measure (F). In a methodological way, they are motivated by the accuracy of the binary classification. In this context, a confusion matrix is used to evaluate such accuracy measures by classifying the objects into different class. There are four distinct classes: true positive (TP), true negative (TN), false positive (FP), false negative (FN). While TP refers the number of correctly specified objects that have positive label, TN shows the number of correctly classified as negative when they indicates the actually negative objects. On the other hand, FP implies the number of misclassified objects that have labeled as wrongly positive and FN demonstrates the number of misclassified objects that have negative label. A confusion matrix is comprised of four classes and it is shown by Table 3.1.

Accuracy: The accuracy refers to the ratio of correctly labelled two objects to all classified objects. It is written mathematically as

Table 3.1: Confusion matrix

		Observed Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \tag{148}$$

Precision: Precision serves as another accuracy measure and is ability to identify the proportion of the true positive objects to the truly labeled two classes by formula

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{149}$$

Recall: Recall is the other accuracy tool and calculates the ratio of objects that are correctly labeled as positive to both the truly detected as positive and falsely labeled as negative. In terms of diagnostic testing, recall is named as *sensitivity* [75].

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{150}$$

F-measure: F-measure is used to evaluate accuracy in terms of the harmonic mean of the precision and the recall. Despite its original formulae [104], in the field of machine learning and data mining, it is aimed to measure balance between precision and recall. It is stated as

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{151}$$

3.4 Applications on GGM

3.4.1 GGM Simulations

In all GGM settings, glasso estimation procedures are applied under the following procedure (*Algorithm 1*).

1. Generate networks whose sample size is 50 under three distinct dimensional settings where the numbers of genes includes 50, 100 and 500 from the *Gaussian graphical model*.

2. Apply an iterative approach to estimate the precision matrix where λ is chosen by RIC as a selection criterion. When $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, *graphical lasso algorithm* offers a solution for this estimation procedure based on penalized log-likelihood and is given by

$$\underset{\Theta}{\text{maximize}}\Big[\log\det\boldsymbol{\Theta} - \text{Trace}(\mathbf{S}\boldsymbol{\Theta}) - \lambda||\boldsymbol{\Theta}||_1\Big], \tag{152}$$

where $\lambda = \lambda_{jk}$ refers to $\lambda_{jk} = \lambda_{kj}$ $\lambda_{jk} = \sqrt{\lambda_j\lambda_k}$ offers different amounts of regularization for each variable. However, we assume that we apply the same amount of regularization for each $\boldsymbol{\Theta}$ in a model. Let $\mathbf{W}$ be the estimate of $\boldsymbol{\Sigma}$ where $\boldsymbol{\Theta}$= $\boldsymbol{\Sigma}^{-1}$ and $\mathbf{S}$ be the empirical covariance matrix. Here, the convex optimization problem will be solved in a cyclical approach. The matrices corresponding with $\boldsymbol{W}$, $\boldsymbol{S}$ and $\boldsymbol{\Theta}$ can be presented by

$$\mathbf{W} = \begin{pmatrix} w_{1,1} & \cdots & w_{1,p} \\ w_{1,p}^T & \cdots & W_{p,p} \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} s_{1,1} & \cdots & s_{1,p} \\ s_{p,1} & \cdots & S_{p,p,} \end{pmatrix}, \quad \boldsymbol{\Theta} = \begin{pmatrix} \theta_{1,1} & \cdots & \theta_{1,p} \\ \Theta_{1,p}^T & \cdots & \theta_{p,p} \end{pmatrix}. \tag{153}$$

The lasso function yields the score function as below.

$$\boldsymbol{\Theta^{-1} - S - \lambda \text{Sign}(\Theta)} = 0. \tag{154}$$

$$\boldsymbol{W - S - \lambda \text{Sign}(\Theta)} = 0. \tag{155}$$

After replacing the derivative of $\log \mathbf{det}(\boldsymbol{\Theta})$, or $\boldsymbol{\Theta^{-1}}$ with $\mathbf{W}$, in this algorithm, we attempts to exemplify the last column by taking $\textbf{Sign}(\beta) = -\textbf{Sign}(\Theta_{-p,p})$

$$\mathbf{W}_{-p,-p}\beta - \mathbf{s}_{-p,p} + \lambda \textbf{Sign}(\beta) = 0. \tag{156}$$

So, the lasso problem can be rewritten as below

$$\underset{\beta}{\text{minimize}}(y - \mathbf{X}\beta)^t(y - \mathbf{X}\beta) + \lambda||\beta||_1. \tag{157}$$

From the linear least squares regression framework, estimates are evaluated with its inner products as $\mathbf{X}^t X$ and $\mathbf{X}^t y$ and are represented as

$$\mathbf{X}^t\mathbf{X}\beta - \mathbf{X}^t y + \lambda \textbf{Sign}(\beta) = 0. \tag{158}$$

The lasso problem is evolved into an optimization problem where it entails three objectives $\textbf{lasso}(W_{-p,-p}, s_{-p,p}, \lambda)$ to solve it. The lasso problem is efficiently solved via the below algorithm, known as *coordinate descent procedure*.

(a) Initialize

$$\boldsymbol{W = S + \lambda \mathbf{I}}. \tag{159}$$

(b) Repeat until convergence

i.

$$W_{-p,-p}\beta - s_{-p,p} + \lambda \text{sign}(\beta) = 0, \tag{160}$$

$$\hat{\beta} \leftarrow \frac{\mathbf{S}\Big(u_j - \sum_{k \neq j} V_{kj}\hat{\beta}_k, \lambda_{ij}\Big)}{V_{jj}} \tag{161}$$

$$\mathbf{S}(x,t) = \mathbf{Sign}(x)(|x| - t)_+, \tag{162}$$

where $V = W_{p,p}$ and $u = s_{-p,p}$. Let $\mathbf{\Sigma^{-1}} = \boldsymbol{W}$. Here, $\mathbf{S}$ is the soft-treshold operator. $\hat{\beta}$ is known for its sparsity. This algorithm works in a fast fashion which requires just rp computations.

ii. Update $w_{-p,p}$ with $W_{-p,-p}\,\hat{\beta}$.

(c) Obtain

$$\hat{\Theta}_{p,p} = \frac{1}{w_{p,p} - w_{-p,p}^t\hat{\beta}}, \text{ and } \quad \Theta_{-p,p} = -W_{-p,-p}^{-1} w_{-p,p}^t \Theta_{p,p}. \tag{163}$$

In this stage, the precision matrices for 20 different penalty parameters which are estimated by RIC. In our case, we provide its corresponding log-likelihood values along the regularization path.

3. Determine the optimal tuning parameter by using a criterion from a collection of them encompassing AIC, BIC, CAIC, CAICF, ICOMP, StARS, EBIC and RIC. In other words, they are employed to pick the best estimate along the whole path.

4. Update $\hat{\mathbf{\Theta}}$ with respect to selected tuning parameter in Step 2(c) in order to obtain the matrix whose elements are 0 and 1's.

For simulation studies the *huge* package in R is implemented to glasso algorithm with scr=FALSE argument, indicating the lossless screening rule [94, 129], rather than lossy [51, 56, 57]. The first rule creates a small block of the inverse covariance matrix where $\boldsymbol{S_{ij}} \geq \lambda$ for $i \neq j$. By this way, the *i*th node or variable is separated from the final estimator. It is motivated from the convex optimization. Therefore, the graphical lasso problem in the previous algorithö is a reduced to a collection of smaller graphical lasso problem, each is aimed to compute block diagonals with corresponding blocks as $C_1, C_2, \cdots, C_K$ is that $|S_{ij}| \geq \lambda$ for all $i \in C_K, j \in C_j, k \neq K$. To achieve such a lossless screening procedure, the related algorithm is demonstrated by the following algorithm (*Algorithm 2*).

The differences between the results arise from the model selection criteria where the same inference method is applied as lasso procedure. In all settings, the lossless [94, 129] screening rule or *graphical lasso with blocks* is employed as a default for which preselecting the neighbourhood is not applied before the graph estimation.

1. Let $\mathbf{A}$ denote a $(p \times p)$-dimensional matrix whose off-diagonal elements are of the form $\mathbf{A}_{ii}^{'} = 1_{|S_{ii}^{'}|>\lambda}$ and whose diagonal elements equal one.
2. Identify the $K \geq 1$ connected components of the graph for which *A* is the adjacency matrix. For each $k = 1, \cdots, K$, let C_k denote the set of indices of the features in the *k*th connected component.
3. Without loss of generality, assume that the features are ordered such that if *i* $\in C_k$, $i^{'} \in C_{k'}$ and $k < k'$ then $i < i'$.
4. The solution of the graphical lasso problem (1.2) takes the following form

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\Theta_1} & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & \boldsymbol{\Theta_2} & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot & 0 & \boldsymbol{\Theta_K} \end{bmatrix},$$

   where $\boldsymbol{\Theta_k}$ solves the graphical lasso problem applied only to the square symmetric submatrix of $\boldsymbol{S}$ consisting of the features whose indices are in C_k. Note that if $C_k = i$, the ith node is completely unconnected from all other nodes then $\boldsymbol{\Theta_k}$ is a scalar equal to $1/(\boldsymbol{S}_{ii} + \lambda)$.

In addition to algorithms we use in our simulation studies, the flowchart representation of the graph estimation from GGM model construction to model selection stage is given in Figure 3.3.

Although StARS is driven to achieve the consistency of the regularization path, RIC works on randomly rotated data to select the minimum regularization without

Table 3.2: Comparisons of model selection criteria under 1000 Monte Carlo runs under different topologies and dimensional settings ($p = 50, 100, 500$) with respect to accuracy measures.

Networks		**Scale-free**			**Random**		
Measures	**Criterion**	**50**	**100**	**500**	**50**	**100**	**500**
F	AIC	0.493	0.461	0.360	0.494	0.427	0.310
pre		0.631	0.613	0.522	0.689	0.619	0.510
rec		0.564	0.503	0.424	0.523	0.498	0.399
acc		0.917	0.954	0.961	0.912	0.936	0.979
F	BIC	0.460	0.402	0.295	0.499	0.414	0.309
pre		0.601	0.548	0.440	**0.697**	0.605	0.492
rec		0.542	0.486	0.393	0.533	0.495	0.403
acc		0.920	0.935	0.973	0.915	0.934	0.976
F	RIC	0.486	0.340	0.297	0.498	0.424	0.304
pre		0.629	0.538	0.441	0.694	**0.619**	0.494
rec		0.560	0.499	0.392	0.528	0.495	0.398
acc		0.915	0.932	0.975	0.915	0.936	0.976
F	EBIC	0.484	0.393	0.296	0.498	0.428	**0.599**
pre		0.631	0.537	0.449	0.689	**0.620**	**0.650**
rec		0.547	0.492	0.388	0.533	0.494	0.691
acc		0.916	0.930	0.975	0.914	0.937	0.978
F	StARS	0.491	0.398	0.291	0.501	0.424	0.310
pre		0.635	0.547	0.434	0.685	0.607	0.507
rec		0.554	0.485	0.388	0.539	0.507	0.402
acc		0.917	0.933	0.974	0.912	0.934	0.976
F	CAIC	0.481	**0.617**	**0.696**	0.485	0.417	0.311
pre		0.616	0.788	**0.906**	0.681	0.612	0.499
rec		0.561	0.615	0.629	0.519	0.486	0.407
acc		0.910	0.953	0.968	0.911	0.934	0.976
F	CAICF	0.292	0.165	0.065	0.383	0.245	0.123
pre		0.179	0.093	0.035	0.251	0.146	0.068
rec		0.796	0.711	0.491	0.814	0.777	0.628
acc		0.770	0.784	0.914	0.793	0.810	0.929
F	ICOMP	**0.507**	0.505	0.525	0.408	0.403	0.398
pre		0.340	0.378	0.578	0.256	0.252	0.248
rec		**1.000**	**1.000**	**0.999**	**1.000**	**1.000**	**1.000**
acc		0.910	0.934	**1.000**	0.941	**0.970**	**0.994**

Figure 3.3: Flowchart simply represents the GGM graph estimation procedure from model construction to model selection.

giving a guarantee for consistency. StARS and RIC suffer from overselection and underselection, respectively [139]. While the dimension increases from 50 to 500, the performance of the both procedures decrease in terms of F-measure, precision and recall. This is also valid for EBIC. The state of art model selection procedures, namely, AIC and BIC also are harmed by increasing the dimension in terms of the recall and precision under both topologies. The results are coincided with our expectation where the both is not designed to work efficiently under high-dimensional settings [25]. On the other hand, CAIC and ICOMP achieve to increase their F-measure, precision and recall as the dimension increases under scale-free network structure, except for ICOMP in the second highest dimension.

Under all dimensional settings, ICOMP provides the best results with respect to the recall under both structures. In terms of accuracy measures, it performs efficiently than its other counterparts in higher dimensions. As the dimension increases, the performance of ICOMP is also raised according to accuracy for all settings, this is valid for also the other types of network structures such as hubs [25]. In particular, we obtain the highest accuracy measure among all counterparts under both network settings. Otherwise, CAIC scale-free networks achieve the best results in terms of the precision measures under the scale-free topology.

3.4.2 GGM Applications on Real Data

The first dataset represents the high dimensional network where $p >> n$. In this sense, ICOMP gives similar results that are obtained with the high-dimensional model selection criteria, namely, StARS, RIC and EBIC, based on all performance measures that we compare. This is valid for also CAIC. However, ICOMP outperforms them in terms of detecting the interactions of the original structure.

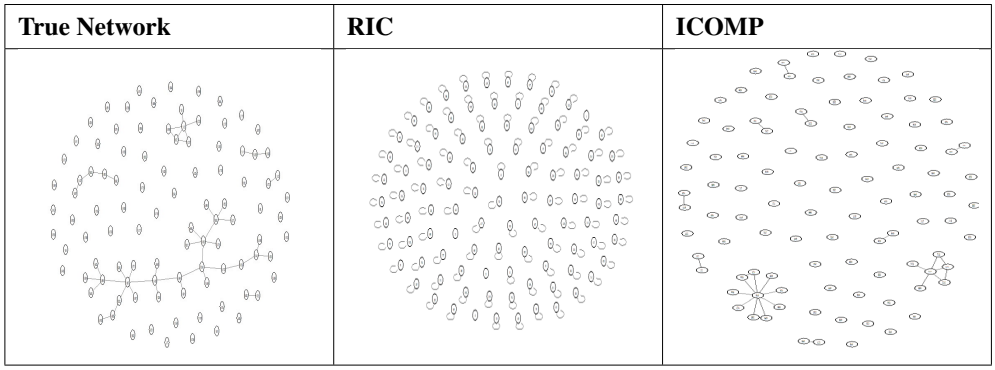


Figure 3.4: From left the right: The true gene expression network, RIC and ICOMP representations. RIC is not able to capture the interactions in the network. ICOMP achieves to detect some interactions in the gene network.

Particularly for precision, it is important to note that all high dimensional criteria obtain one, like ICOMP and CAIC. On the other hand, AIC and BIC could not achieve this.

In terms of accuracy, our three suggestions including CAIC, CAICF and ICOMP, and the criteria produced for high dimensional networks outperform AIC and BIC for two benchmark datasets. Also, they increase their power when we apply them on high dimensional dataset i.e., gene expression, unlike the state-of-art selection techniques.

On the other hand, cell signal dataset do not represent the high dimensional network where $p >> n$. Therefore, in terms of F-measure and precision, AIC and BIC give similar results that lower than the rest obtain.

Our real datasets with 11 genes represent a complete graph where each gene is connected to each other. Also, these datasets are known as dense network and do not reflect the high dimensional network structure. So, we do not expect to obtain optimal results with our suggested criteria as well as RIC, EBIC and StARS. For all complete graphs, there exists no false positive value. So, we get the precision value equals 1 for all criteria which we compare.

Table 3.3: Comparisons of model selection criteria for six real data sets with respect to accuracy measures.

		Datasets					
Measures	**Criterion**	**Data 1**	**Data 2**	**Data 3**	**Data 4**	**Data 5**	**Data 6**
F	AIC	0.355	0.564	0.167	0.754	0.593	**0.880**
pre		0.758	0.758	1.000	1.000	1.000	1.000
rec		0.555	0.555	0.091	0.605	0.422	0.785
acc		0.769	0.769	0.091	0.605	0.422	0.785
F	BIC	0.355	0.564	0.740	0.754	0.593	0.194
pre		0.758	0.758	1.000	1.000	1.000	1.000
rec		0.555	0.555	0.587	0.605	0.422	0.107
acc		0.769	0.769	0.587	0.605	0.422	0.107
F	RIC	0.393	0.690	0.297	0.498	0.424	0.304
pre		**1.000**	**1.000**	1.000	1.000	1.000	1.000
rec		0.526	0.244	0.392	0.528	0.495	0.398
acc		**0.991**	0.717	0.392	0.528	0.495	0.398
F	EBIC	0.393	0.690	0.167	0.198	0.593	0.167
pre		**1.000**	**1.000**	1.000	1.000	1.000	1.000
rec		0.526	0.244	0.091	0.111	0.425	0.091
acc		**0.991**	0.717	0.091	0.111	0.425	0.091
F	StARS	0.393	0.690	0.167	0.198	0.194	0.290
pre		**1.000**	**1.000**	1.000	1.000	1.000	1.000
rec		0.526	0.244	0.091	0.111	0.107	0.174
acc		**0.991**	0.717	0.091	0.111	0.107	0.174
F	CAIC	0.393	0.690	0.625	0.198	0.393	**0.880**
pre		**1.000**	**1.000**	1.000	1.000	1.000	1.000
rec		0.526	0.244	0.455	0.111	0.719	0.785
acc		**0.991**	0.717	0.455	0.111	0.719	0.785
F	CAICF	0.343	0.641	0.740	0.773	0.167	**0.880**
pre		0.208	0.758	1.000	1.000	1.000	1.000
rec		**0.968**	0.555	0.587	0.630	0.09	0.785
acc		0.929	0.769	0.587	0.630	0.09	0.785
F	ICOMP	0.393	0.690	0.593	0.198	0.167	0.296
pre		**1.000**	**1.000**	1.000	1.000	1.000	1.000
rec		0.526	0.244	0.423	0.111	0.091	0.174
acc		**0.991**	0.717	0.423	0.111	0.091	0.174

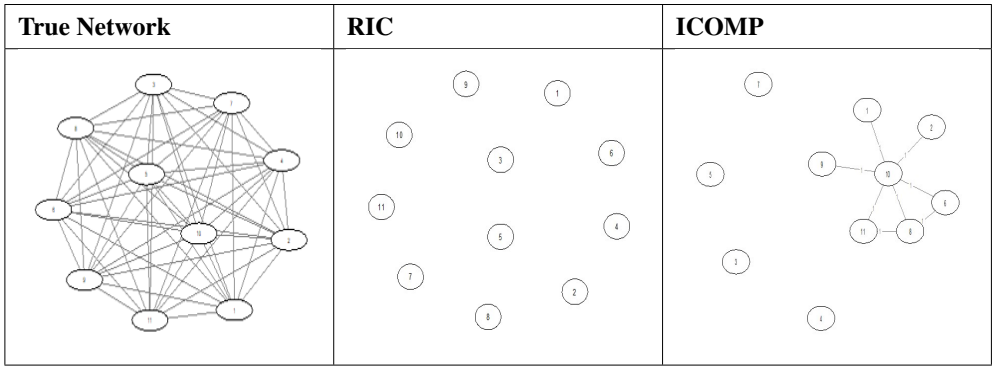


Figure 3.5: From left the right: The true full graph with 11 genes, RIC and ICOMP representations. RIC is not able to capture any interactions in the network. ICOMP achieves to detect some interactions in the gene network.

Also, for all complete graphs, we do not expect to see the true negative value so that the recall and accuracy measures are the same for each criterion.

For datasets including 3, 4 and 6, CAICF achieves to get the best results with respect to recall and accuracy. While it is coincided with BIC for Data 3, CAIC and AIC give the same results in terms of F-measure, recall and accuracy for Data 6 ,like CAICF. For Data 5, AIC and BIC get the same results which refer to the best in terms of recall and accuracy. Similar to RIC, EBIC and StARS, the performance of the ICOMP decreases somehow for the complete data sets where they do not represent high dimensional sparse network structure. However, it can be able to capture some interactions in the true structure.

3.5 Applications on LMARS

Like in the GGM simulations, to compare the accuracy measures of versatile model selection procedures in order to represent the gene network, LMARS scheme is implemented for each gene in a two-stage iterative fashion that including the forward and backward. In our simulation studies, we ameliorate the backward steps of that,

keeping the forward as the same as in the original MARS algorithm. In our LMARS algorithm, the least squares which are suggested in original work is replaced with our proposed model selection procedures including CAIC, CAICF and ICOMP in the backward step. Also, we implement the state-of-art model selection criteria in order to compete the accuracy measures.

Loop based MARS

Under MARS setting, it is required to make adjustments on the original MARS procedure to explain the relations for each gene in a graphical network. In this sense, we modify the original method by LMARS is seen as a nonparametric analogous to the lasso regression [9]. In our LMARS analysis, we aim to construct a MARS regression model by taking only the main effects of the genes for each node against the rest in the network.

To illustrate the LMARS model construction, we design a toy set whose graphical network not only is represented by Figure 2.1, but also its adjancecy matrix and lasso equations are shown in Equation (164), respectively. In our design, $\boldsymbol{p_1}$ is modeled against the others $(\boldsymbol{p_2}, \boldsymbol{p_3})$. Correspondingly, when $\boldsymbol{p_1}$ represents the response in the regression, the others stand for explanatory variables. Here, $\boldsymbol{p_1}$ can be explained by solely taking the effect of $\boldsymbol{p_2}$, which means that $\boldsymbol{p_1}$ is related to $\boldsymbol{p_2}$. Thus, this relation is expressed in the first row of A.

$$\mathbf{A}_{3,3} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \qquad \begin{aligned} \boldsymbol{p_1} &= 1.2 + 3p_2 \\ \boldsymbol{p_2} &= p_1 + 2.7p_3 \\ \boldsymbol{p_3} &= 4.8p_2 \end{aligned} \tag{164}$$

In our LMARS analysis, it is aimed to reveal a gene network representation and the LMARS procedure where we interfere the model selection is shown in Figure 2.1.

3.5.1 LMARS Simulations

In the high dimensional biological network context, we generate our random variables, i.e., genes in the network via the *huge* package with R under the multivariate normality assumption. Under this setting, we generate three and two distinct scenarios in terms of dimensions and topology, respectively. When the dimension of gene networks changes from 50 to 500, there exits two distinct topologies: scale-free and random. Then, we build our Loop-based MARS models so as to modify original MARS strategy to our study. In this sense, we construct separate MARS models for each gene in the network in order to decide the relations between the rest of the genes, only taking its main effects. Below both forward and backward strategies are presented with respect to each MARS model associated with taking one gene as a response in model. Here, the *earth* package in R is used to employ the MARS strategy where all alternative criteria which we use are inserted by hand in its elimination procedure.

The corresponding MARS algorithm can be written in terms of its pseudocodes which are represented in the following procedure (*Algorithm 3*).

In the forward algorithm, we construct a model with respect to LOF. At the step 6, 10 and 11 in the *Algorithm 3*, step functions are used as truncated power basis functions ($q = 1$). In the MARS strategy, Line 6 represents the parent basis function which includes a subset of the complete tensor product basis functions with knots whose first derivatives are continuous at different data values. Here, this algorithm yields $M_{\max} q = 1$.

Under the MARS setting, we aim to interfere the model selection procedure of MARS where we modify its backward algorithm by replacing its original criteria with CAIC, CAICF, ICOMP as well as AIC and BIC in order to determine the basis functions and the corresponding knot values in the final model.

1: **procedure** Construct Full Modelw.r.t. GCV

2: $B_1(\mathbf{x}) \leftarrow 1; M \leftarrow 2 \quad M > M_{max}$;LOF* $\leftarrow \infty$

3: **for** m $\in (1 : M-1)$ **to** : **do**

4: **for** v $\notin \{v(k,m)|1 \leq k \leq K_m\}$ **to** : **do**

5: **for** t $\in \{\mathbf{x}_{vj}|B_m(\mathbf{x}_j) > 0\}$ **to** : **do**

6: $g \leftarrow \sum_{i=1}^{M-1} a_i B_i(\mathbf{x}) + a_M B_m(\mathbf{x})[+(x_v - t)]_+ + a_{M+1} B_m(\mathbf{x})[-(x_v - t)]_+$;

7: $lof \leftarrow$ minimize$_{a_1,a_2,\cdots,a_{M+1}}$LOF(g)

8: **if** lof < lof* **then**,

9: lof* $\leftarrow$ lof; $m^* \leftarrow m$; $v^* \leftarrow v$; $t^* \leftarrow t$;

10: $\mathrm{B_m}(\mathbf{x}) \leftarrow \mathrm{B_{m*}}(\mathbf{x})[+(x_{v*} - t*)]_+$

11: $B_{M+1}(\mathbf{x}) \leftarrow B_{m*}(\mathbf{x})[-(x_{v*} - t*)]$

12: $M \leftarrow M + 2$

Figure 3.6: *Algorithm 3* shows the forward step of MARS.

Therefore, in the backward strategy, we delete our basis functions according to criterion that we prefer rather than using GCV. Also, it is important to mention that removing basis linear functions does not cause discontinuity in the predictor space since the subregions overlap.

In *Algorithm 4*, $V(m) = v(k,m)_1^{K_{max}}$ represents the variable set where the mth basis function is presented by B_m. Totally, in the deletion algorithm, we totally get $M_{\text{max}} - 1$ models by iteratively lowering its number of basis functions one by one in the sequence $\boldsymbol{J}^*$. Then, we determine the best model with its proper basis functions in the prespecified sequence.

In accordance with the MARS procedure, its process can be easily seen from the flowchart representation in Figure 3.8.

By following the steps in the Figure 3.8 for each selection criterion, we conduct our

1: **procedure** Choose BF'sw.r.t. Model Selection Criteria
2: $J^* = 1, 2, \cdots, M_{\text{maximize}}$; $K^* \leftarrow J^*$
3: Criterion* $\leftarrow$ minimize$_{\{a_j | j \in J^*\}}$Criterion$(\sum_{j \in j^*} a_j B_j(\mathbf{x}))$
4: **for** M=M$_{\text{max}}$ **to** 2 **to do** $b \leftarrow \infty$; $L \leftarrow K^*$;
5: **for to** m=2 **to** M **do** $K \leftarrow L - \{m\}$
6: Criterion $\leftarrow$ minimize$_{\{a_k | k \in K\}}$Criterion$(\sum_{k \in K*} a_k B_k(\mathbf{x}))$
7: **if** Criterion < b **then** $b \leftarrow$ Criterion; $K^* \leftarrow K$
8: **if** Criterion < Criterion* **then** Criterion $\leftarrow$ Criterion*; $J^* \leftarrow K$

Figure 3.7: *Algorithm 4* shows the backward step of MARS.

simulation study under the MARS setting. We aim to compare our model selection criteria under two topologies: scale-free and random as the dimension of the gene networks increased from 50 to 500. Thus, we obtain Table 3.4.

For Table 3.4, we obtain the precision values as one for each criterion under three dimensions and two topologies. This indicates that MARS is succeed to not detect falsely positive interaction between the genes under all scenarios.

In generally, when the dimension increases, F-measure decreases or remain the same under each topology. This is valid for also recall values. However, it is important to mention that CAICF achieves to not only increase its F-measure but also to obtain the best F-measure under scale-free structure when the dimension changes from 50 to 100. Like CAICF under scale-free, BIC and ICOMP increases their F-measures in the highest dimension under the random topology, as well as obtaining the best F values in this setup.

In terms of accuracy, each criterion tends to raise their accuracy measures although the dimension increases. This result indicates that the detection of the FN values is lowered when comparing either TP or TN values when the dimension increases.

Under scale-free topology for the both lowest and highest dimensional settings, CAIC get the best values with respect to accuracy, on the other hand, for the moderate dimensional setting, ICOMP is the best. Also, under this topology, CAICF gets the

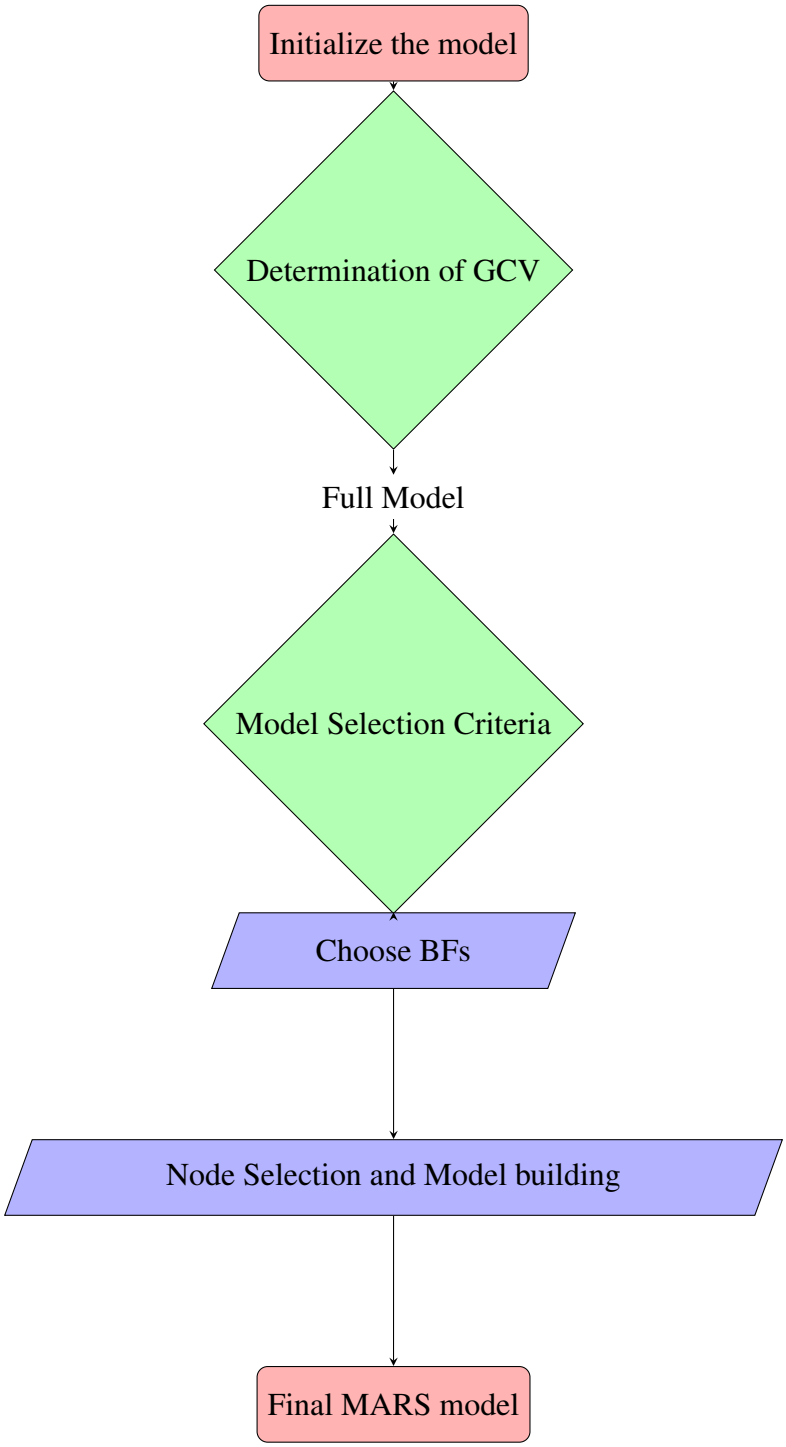


Figure 3.8: Adaptive LMARS model selection procedure: We use the alternative model selection criteria instead of GCV in the backward step in order to determine the final model.

Table 3.4: Comparisons of model selection criteria under 1000 Monte Carlo runs under different topologies and dimensional settings with respect to accuracy measures.

Networks		Scale-Free			Random		
Measures	Criterion	**50**	**100**	**500**	**50**	**100**	**500**
F	AIC	0.399	0.386	0.389	0.402	0.404	0.395
pre		1.000	1.000	1.000	1.000	1.000	1.000
rec		0.253	0.244	0.246	0.256	0.254	0.246
acc		0.931	0.962	**0.993**	0.940	0.970	0.994
F	BIC	0.400	0.390	0.380	0.410	0.390	0.411
pre		1.000	1.000	1.000	1.000	1.000	1.000
rec		0.253	0.247	0.242	0.247	0.253	0.256
acc		0.932	0.964	0.990	0.932	0.964	0.994
F	GCV	0.399	0.393	0.393	0.414	0.397	0.397
pre		1.000	1.000	1.000	1.000	1.000	1.000
rec		0.253	0.248	0.244	0.262	0.248	0.248
acc		0.931	0.964	0.992	0.942	0.969	0.994
F	CAIC	0.402	0.393	0.386	0.416	0.407	0.409
pre		1.000	1.000	1.000	1.000	1.000	1.000
rec		0.255	0.247	0.243	0.264	0.257	0.257
acc		0.932	0.964	0.992	0.943	**0.971**	0.994
F	CAICF	0.399	0.401	0.380	0.416	0.411	0.399
pre		1.000	1.000	1.000	1.000	1.000	1.000
rec		0.253	0.253	0.240	0.264	0.260	0.250
acc		0.930	0.967	**0.993**	0.943	0.941	0.994
F	ICOMP	0.399	0.399	0.388	0.403	0.402	0.404
pre		1.000	1.000	1.000	1.000	1.000	1.000
rec		0.253	0.249	0.245	0.253	0.252	0.253
acc		0.931	**0.970**	0.992	0.940	**0.970**	0.994

compatible results with the CAIC and ICOMP. Under random network setting, CAIC has the best accuracy measures for the first two dimensions, but for the highest dimension, AIC, GCV, ICOMP, CAICF and BIC reach the same highest accuracy value.

3.5.2 LMARS Applications on Real Data

In our application studies, we use the same datasets as we apply in the GGM studies. Therefore, we have three datasets that represent the complete graph of the eleven core genes where each gene is connected to each other with a set of numbers: 12, 35 and 287. Also, there exists another complete graph that contains nine genes with three observations.

From Table 3.5, we reach the similar results with that of the simulation studies in terms of precision, so we get the precision values as one. Although the datasets including 3 to 6 contains only the fully connected genes, the first two datasets contain also the unconnected genes to the some of the nodes in the network. Therefore, obtaining the precision values as one still indicates that under the MARS setting, criteria achieve to not connect the genes falsely (FN=0).

On the other hand, for fully connected gene networks (Data 3, Data 4, Data 5 and Data 6), recall and accuracy values are the same for each criterion since the TN values are also the zero for such dense networks (FP=0).

Particularly for Data 1, each criterion under the LMARS setting succeeds to capture both positive and negative relations correctly the most of the time since they achieve to yield the higher results in terms of F and accuracy measures, simultaneously.

In terms of performance measures that we use, we do not get any significant difference among criteria. Thus, we get the same performance measures in terms of F, precision, recall and accuracy measures for each model selection criterion which we compare.

Table 3.5: Comparisons of model selection criteria under 1000 Monte Carlo runs under different topologies and dimensional settings with respect to accuracy measures.

		Datasets					
Measures	**Criterion**	**Data 1**	**Data 2**	**Data 3**	**Data 4**	**Data 5**	**Data 6**
F	AIC	**0.713**	0.407	0.091	0.091	0.091	0.091
pre		1.000	1.000	1.000	1.000	1.000	1.000
rec		0.557	0.256	0.167	0.167	0.167	0.167
acc		**0.992**	0.734	0.167	0.167	0.167	0.167
F	BIC	**0.713**	0.407	0.091	0.091	0.091	0.091
pre		1.000	1.000	1.000	1.000	1.000	1.000
rec		0.557	0.256	0.167	0.167	0.167	0.167
acc		**0.992**	0.734	0.167	0.167	0.167	0.167
F	GCV	**0.713**	0.407	0.091	0.091	0.091	0.091
pre		1.000	1.000	1.000	1.000	1.000	1.000
rec		0.557	0.256	0.167	0.167	0.167	0.167
acc		**0.992**	0.734	0.167	0.167	0.167	0.167
F	CAIC	**0.713**	0.407	0.091	0.091	0.091	0.091
pre		1.000	1.000	1.000	1.000	1.000	1.000
rec		0.557	0.256	0.167	0.167	0.167	0.167
acc		**0.992**	0.734	0.167	0.167	0.167	0.167
F	CAICF	**0.713**	0.407	0.091	0.091	0.091	0.091
pre		1.000	1.000	1.000	1.000	1.000	1.000
rec		0.557	0.256	0.167	0.167	0.167	0.167
acc		**0.992**	0.734	0.167	0.167	0.167	0.167
F	ICOMP	**0.713**	0.407	0.091	0.091	0.091	0.091
pre		1.000	1.000	1.000	1.000	1.000	1.000
rec		0.557	0.256	0.167	0.167	0.167	0.167
acc		**0.992**	0.734	0.167	0.167	0.167	0.167

From the statistical modeling perspective, all complete graph structures are associated with the overfitted model where it must include all variables or genes. This is the fact that AIC is known for its susceptibility to choose the overfitted model [17, 110]. Therefore, it is expected that AIC tends to select the full model encompassing all variables. Particularly for the high dimensional biological networks, we expect to find the full model or the most complex model among many alternatives by detecting all interactions in the network. However, AIC fails in terms of determining all connections among genes. In our complete gene set applications, all model selection criteria choose the same MARS model with defining same connections in the network structure.

As the sample size increases, all criteria attempt to find more connections among genes in the network. For example, specifically, for nine complete gene structure, each establishes the model by taking only the intercept. However, all criteria detect the relations among three genes in the network where it actually includes eleven connected genes with twelve observations for each gene. On the other hand, for the datasets including the same genes with 35 and 287 observations, they achieve to detect relations among five genes in the network.

Under the MARS setting, we are inspired from the paper that has been published by Kartal-Koc and Bozdogan (2015) [78]. In this study, they interfere the model selection procedure of the original MARS in such a way that ICOMP is replaced with GCV. In this sense, they compare them by constructing both linear and nonlinear models. Also, they concluded that ICOMP outperforms GCV, so it can be applicable on more complex models [78].

Under the LMARS setting, we examine the models that are chosen by each criterion which we compare with respect to complexity for the MARS model [59]. This gives a way to determine the specific choice for the variables. In this sense, we analyze the MARS equations with their basis linear functions by following [78].The corresponding MARS models are presented for each model selection procedure.

Table 3.6: Data 3 is composed of 11 genes and 287 observations.

MODEL:
y=7.92+0.14BF1+0.61BF2-0.07BF3-0.02BF4+0.04BF5+0.22BF6
+0.53BF7+0.01BF8
where BF1=max(0,9.15254-CEBPB), BF2=max(0,CEBPB-9.15254), BF3=max(0,11.2887-CTNNB1), BF4=max(0,CTNNB1-11.2887), BF5=max(0,8.3396-TFAM), BF6=max(0,TFAM-8.3396), BF7=max(0,3.98463-PDIA3), BF8=max(0,PDIA3-3.98463)
Important Variables: PDIA3, TFAM, CEBPB, CTNNB1
GCV: 0.12 **RSS:** 28.8 **RSq:** 0.09.

Here, RSq $= 1 -$ RSS/TSS is used for measuring performance on data as the coefficient of determination where RSS $= \sum_{i=1}^{n}(y - \hat{y})^2$ is the residual sum-of-squares and TSS $= \sum_{i=1}^{n}(y - \bar{y})^2$ is the total sum-of-squares [91].

When we have the fully interconnected gene set with 287 observations, we obtain the same MARS model by using each criteria. In Table 3.6, y represents the "MBD3".

In the sense of modeling fully connected genes, we conclude that an increase in the number of observations help to detect more relations and this is valid for each criterion. When we have Data 5, all model selection criteria achieve to detect the model including "PDIA3" and "TFAM" as basis linear functions. In Table 3.8, the response value is represented by "MBD3".

The other datasets which we use contain one hundred genes and eleven genes, correspondingly with 60 and 16072 observations for each gene in the network. Not all genes are connected to each other in these two sets. Specifically, the first dataset rep-

Table 3.7: Data 4 is composed of 11 genes and 35 observations.

MODEL:
y=7.27+0.56MAP2K1-0.24MAP2K1+0.15TP53+0.26TP53-0.75IMP3+0.03IMP3+0.76CHD4+0.86CHD4
where BF1=max(0,9.1186-MAP2K1, BF2=max(0,MAP2K1-9.1186), BF3=max(0,8.4999-TP53), BF4=max(0,TP53-8.4999), BF5=max(0,5.93102-IMP3), BF6=max(0,IMP3-5.93102), BF7=max(0,9.44477-CHD4), BF8=max(0,CHD4-9.44477)
Important Variables: MAP2K1,TP53,IMP3,CHD4
GCV: 0.18 **RSS:** 1.66 **RSq:** 0.80

Table 3.8: Data 5 is composed of 11 genes and 12 observations.

MODEL:
y=4.78+1.34TFAM+0.02TFAM-0.25PDIA3+1.02PDIA3
where BF1=max(0,10.23-TFAM),BF2=max(0,TFAM-10.23),BF3=max(0,8.009-PDIA3),BF4=max(0,PDIA3-8.009)
Important Variables: PDIA3, TFAM
GCV: 1.04 **RSS:** 0.77 **RSq:** 0.96

resents the sparse graph.

Stone [114] mentions the similarity between AIC and CV that when the sample size increases, minimizing the AIC is equivalent to minimizing CV for any model. So, they yield the same complex model, like in AIC, when the number of observations are increased. Also, this can be described as the overfitting problem, as the model become more flexible, we may face with such a problem. In the MARS setting, as MARS refers a flexible modelling technique, AIC and GCV result in the same overfitted model in the purpose of the model selection [96].

On the other hand, ICOMP has tendency to choose much simpler and interpretable models when comparing GCV in the high dimensional context. Kartal-Koc and Bozdogan [78] discuss the differences between GCV and ICOMP in the sense that when GCV is prone to choose complex models, ICOMP selects the simpler one [116].

In the application of the cell signal dataset, we obtain similar results in terms of the gene network structure. While AIC constructs a MARS model including seven genes and GGV chooses a model with including 5 genes, ICOMP and BIC build the one which uses important genes in the network structure. With respect to the response, each procedure takes the "V4" as a response for its corresponding model.

In Table 3.9 and 3.10, the selected MARS equations are written as with its corresponding true network structure. In the Cell Signal Network, although each gene is connected to at least one gene in the cell, **V3**, **V8** and **V11** play the greater role in the mechanism of this cell.
In Table 3.9, the selected MARS equations are written with respect to its corresponding true network structure in Figure 3.9.

Aiming to determine the important genes in the high-dimensional network structure, ICOMP tends to capture the only crucial ones. On the other hand, AIC obtains the complicated model that makes hard to interpret the cell mechanism. Like AIC, GCV also finds a next complicated model to explain the system. AIC applies the lowest

Table 3.9: Data 2 (i.e., Cell Signal Dataset) is composed of 11 genes and 11672 observations.

MARS Model with AIC

y=35.01-0.02V2-0.62V2+0.401V3+2.095V3+2.13V5+2.15V5-2.12V6+1.085V6-1.18V7+0.275V7-0.04V9+0.33V9+1.05V11-1.34V11

where BF1=max(0,V2-257), BF2=max(0,257-V2), BF3=max(0,V3-963), BF4=max(0,1963-V3), BF5=max(0,V5-75), BF6=max(0,313-V5), BF7=max(0,V6-313), BF8=max(0,313-V6), BF9=max(0,91.4-V7), BF10=max(0,V7-91.4), BF11=max(0,445-V9), BF12=max(0,V9-445), BF13=max(0,V11-256), BF14=max(0,256-V11)

Important Variables: V2, V3, V5, V6, V7, V9, V11

GCV: 10124.69 **RSS:** 117861763 **RSq:** 0.672

MARS Model with GCV and CAIC:

y=58.31+0.43V2-0.38V2+1.93V8+2.074V8-1.12V5+0.95V5-2.12V6+1.085V6-0.86V7+0.638V7

where BF1=max(0,V2-73.0), BF2=max(0,73.0-V2), BF3=max(0,V8-121), BF4=max(0,121-V8), BF5=max(0,V5-189), BF6=max(0,189-V5), BF7=max(0,V6-204), BF8=max(0,204-V6), BF9=max(0,68.3-V7), BF10=max(0,V7-68.3)

Important Variables: V2, V5, V6, V7, V8

GCV: 989.90 **RSS:** 109001050 **RSq:** 0.978

MARS model with BIC, ICOMP and CAICF

y=59.02-3.75V3+2.045V3+1.93V8-2.95V8+2.51V11+0.98V11

where BF1=max(0,20.2-V8), BF2=max(0,V8-20.2), BF3=max(0,87.5-V3), BF4=max(0,V3-87.5), BF5=max(0,V11-313), BF6=max(0,313-V11)

Important Variables: V3, V8, V11

GCV: 936.8001 **RSS:** 109240280 **RSq:** 0.984

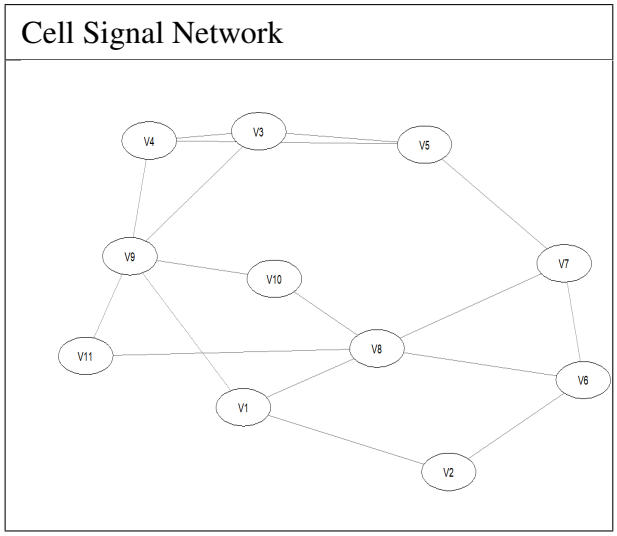


Figure 3.9: Representation of the true network for cell signalling network.

penalty when comparing the other criteria we use, it has a tendency to pick a model with including many terms. This result is clarified by the use of RSS value we mentioned in Section3, since any inclusion reduces to RSS value for a model, even we add the insignificant terms. On the other hand, although GCV is based on RRS in a more penalized version, it suffers from choosing the unnecessary variables to the MARS model, also. Thereby, this result is also verified by the conclusion that the Kartal-Koc and Bozdoğan [78] reach when the numbers are grater than one hundred, ICOMP and BIC choose the simplest with using the true variables.

Also, we use Gene Expression dataset to compare our model selection criteria in terms of the MARS equations. On the other hand, the associated true network is represented in Table 3.8. Here, under the MARS setting, we aim to construct MARS models for this network.

In terms of the MARS equations corresponding with this dataset, GCV and AIC act similarly by the way they choose the overselected model where they include more BF's than that of ICOMP and BIC. Thus, as we mentioned beforehand, we can conclude that this is coincided with our expectation regarding the statistical evaluation theory [71, 116, 141]. Also, it is verified by the systems biology [107].

Table 3.10: (Data 1) Gene Expression Data set is composed of 100 genes and 60 observations.

MODEL with AIC, GCV and CAIC:

y= 9.308-0.510Gene15-1.577Gene15-0.148Gene16+ 0.232Gene16 + 0.199Gene32 + 0.788Gene32 - 0.704Gene43 - 0.601Gene43 - 0.623Gene34 -0.846Gene34 - 0.511Gene37 - 0.234Gene37 - 0.336Gene39 - 0.747Gene39 +0.989Gene44 + 0.159Gene44 + 0.378Gene48+ 0.22Gene48 + 0.229Gene61 + 2.251Gene61

where BF1=max(0,8.578-Gene15), BF2=max(0,Gene15-8.578),BF3=max(0,Gene38.899-Gene16), BF4=max(0,Gene16-8.899), BF5=max(0,9.020-Gene32), BF6=max(0,Gene32-9.020), BF7=max(0,8.873-Gene43), BF8=max(0,Gene43-8.873), BF9=max(0,8.486-Gene34), BF10=max(0,Gene34-8.486), BF11=max(0,7.683-Gene37), BF12=max(0,Gene37-7.683), BF13=max(0,9.225-Gene39), BF14=max(0,Gene39-9.225), BF15=max(0,7.318-Gene44), BF16=max(0,Gene44-7.318), BF17=max(0,8.798-Gene48), BF18=max(0,Gene48-8.798), BF19=max(0,9.034-Gene61), BF20=max(0,Gene61-9.034)

Important Variables: Gene15, Gene16, Gene32, Gene34, Gene37, Gene39, Gene43, Gene44, Gene48, Gene61

GCV: 3.36 **RSS:** 18.45 **RSq:** 0.672

MARS model with ICOMP, BIC and CAICF

y= 11.38- 1.51Gene44+ 1.77Gene44- 0.18Gene74+ 1.433Gene74+ 3.21Gene78+ 2.341Gene78+0.04Gene43-1.23Gene43

where BF1=max(0,9.089-Gene44), BF2=max(0,Gene44-9.089), BF3=max(0, 7.049-Gene74), BF4=max(0,Gene74-7.049), BF5=max(0,14.57-Gene78), BF6=max(0,Gene78-14.57), BF7=max(0,10.09-Gene43), BF8=max(0, Gene43-10.09)

Important Variables: Gene43, Gene44, Gene74, Gene78 **GCV:** 2.96 **RSS:** 10.45 **RSq:** 0.602

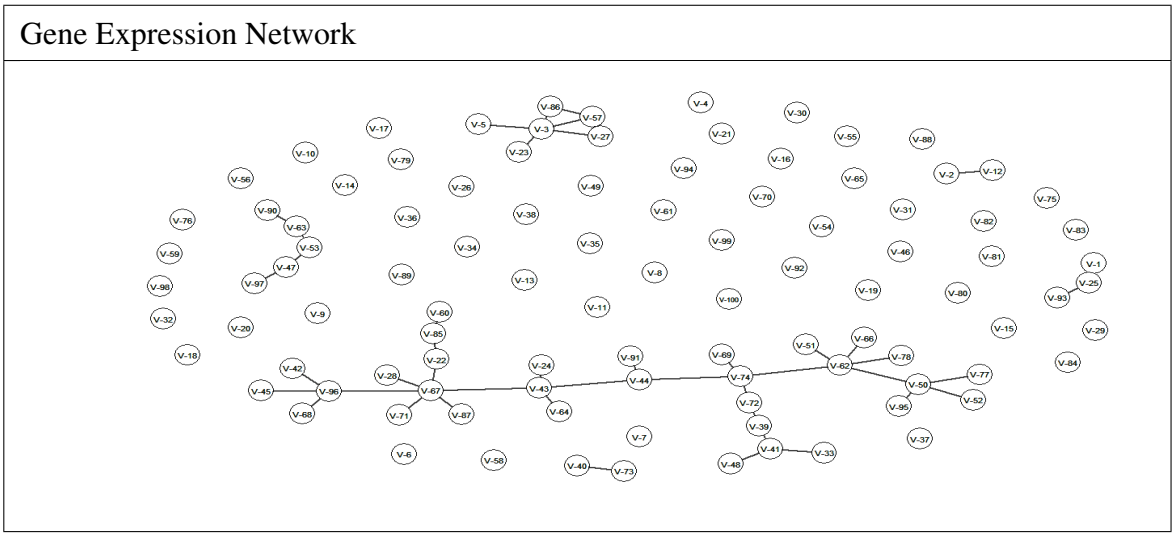


Figure 3.10: Representation of the true network for the gene expression network.

Among our data, the sole Gene Expression Data set yields a sparse network whose relatively small number of genes is connected to each other in the high dimensional setting where the number of genes exceeds the number of number of observations $(p > n)$.

In the representation of the true Gene Expression network, ICOMP detects the important variables in the description network. Here, our two proposed methods, CAICF and ICOMP result in more simpler models than that of AIC and GCV. They are achieved via taking only the important effects, or not including the genes that have moderate effects on the biological mechanism. Therefore, this result is more coincided with our expectation from a sparse biological network structure. In terms of BIC and CAIC, the first tends to select simpler models like ICOMP. On the other hand, the latter acts as AIC in this setup. Also, this result is justified by the statistical theory of the model evaluation. In the statistical model selection context, BIC is attributed as the consistent model selection procedure, but, this is not valid for AIC [97]. Thus, BIC produce more simpler models than that of AIC because of the fact that BIC uses a heavier penalty than AIC applies [112]. Also, Shao [112] mentions that CV does not perform a consistent model selection procedure and employs poorly

when dealing with high dimensional data. [95].

CHAPTER 4

CONCLUSION

In this study, we focus on the model selection procedure in higher dimensions which plays a crucial role when determining the final structure of the gene network. In this thesis, it is expected to capture the true undirected interactions in the network by using two different statistical modeling techniques: GGM and LMARS.

In the first part of the analyses, the background information is reviewed and classified as three main parts: while the first two include modelling approaches, namely, GGM and MARS and one is related to the model selection procedures form the main part of this thesis study.

Hereby, initially, GGM is reviewed as a parametric way to model a high dimensional network where it requires some important assumptions which are conditional independence and normality. Then, its several inference methods are separated into mainly two parts: penalized likelihood methods and regression methods. In this thesis, we focus on graphical lasso (glasso) algorithm which is designed to penalize the covariance matrix, in turn, coefficients in the model by using the assumption of conditional independence. In general, the underlying method belongs to the penalized likelihood method, so it works with a penalty parameter to correctly determine the sparsity in the network. Also, it achieves to define the important genes in the network in such a way that it can be the effects of the some genes to zero in the network structure. In this sense, its other counterparts under penalized likelihood methods are reviewed, so we encompass ridge and lasso approaches as well as its derivatives so-called SCAD, AL, Elastic net as alternative inference methods to determine the sparsity of the network.

In the second part of the background information, we present MARS as a nonparametric flexible way to model a high dimensional network. In this section, the use of MARS as a nonparamteric technique in higher dimensions is justified by explaining its potential capabilities for dealing with nonlinearities and collinearities that arises from higher dimensions. Then, its original model selection procedure, so-called GCV, is introduced. Otherwise, the derivatives of MARS known as conic MARS, robust CMARS, bootstrapping MARS and BCMARS are reviewed. Among many alternatives of the MARS models, LMARS is used since it is designed to construct separate gene models for each by taking the only main effects of the genes.

Finally, the model selection procedures are examined. From the generic statistical point of view, the RSS value is seen as a common way to determine the final model and the philosophy behind it is given. Also, the information theory provides another statistical way to pick up the best approximated model to the true distribution by using the K-L divergence. Since the data-dependent model selection procedures are ruled by this theory, its philosophy behind the K-L divergence is mentioned. Furthermore, the data-dependent model selection procedures are introduced in the methodology part. Here, the selected methods are based on likelihood-based approach, except StARS. To compete them, we have the state-of-art model selection criteria known as AIC and BIC. When both AIC and BIC are powerful under lower dimensions, they lose their power as the dimension increases. To tackle high dimensional problems, as given in the same part, RIC, EBIC and StARS have been proposed. While the first two (RIC and EBIC) stand for information based approaches, StARS conducts a subsampling procedure. So, it does not make use of the good asysmptotical properties of the likelihood theory. Although high dimensional model selection approaches have already been existed in the statistical literature so as to determine the final structure by defining the sparsity of the network, they suffer from either overselection or underselection, respectively, for StARS and both RIC and EBIC. Therefore, we suggest to apply CAIC, CAICF and ICOMP as the model selection procedures in order to represent the true high dimensional network structure under both GGM and MARS settings.

After giving the necessary background information, we continue with the application and the simulation studies under two modelling setups. Therefore, we can separate this section mainly into two parts according to the model which we apply. Here each includes the both real data applications and the simulation studies.

In the sense of real data applications, we use six data sets where the first two refer to the benchmark data, the rest represents dense network structures including fully connected genes. Among all datasets, only the first dataset indicates the sparse gene expression network structure whose number of genes far exceeds the number of observations. However, the number of observations is greater than the number of genes in the network for the others that include eleven core cancer genes, except the sixth one which contains only nine genes.

On the other side, in terms of simulation studies, we generate distinct scenarios under two different topologies: scale-free and random structures. While the important mathematical philosophies support to construct such topologies, scale-free is more coincided with our expectation from the biology since it represents the hubs in the network structure. In other words, under the scale-free topology, some genes, called hubs in the network, have more crucial role than the others, so we expect to see the connections are more concentrated on such hubs in this structure. Otherwise, our scenarios are simulated with respect to different dimensional setups, namely 50, 100, 500 genes, with 50 observations for each gene in the network structure.

Under the GGM setting, the penalized likelihood procedure, namely, the graphical lasso algorithm, requires a penalty parameter to control the sparsity of the network structure. In this section, the graphical lasso algorithm whose coordinate descent procedure and its blockwise version, which we called the lossless screening under the *huge* package in R, are mentioned with its corresponding pseudocodes and its flowchart representation to explain the procedure. This versatile regression procedure simultaneously achieves the variable selection procedure while penalizing the covariance of the model. Thus, the selection of the penalty parameter plays a pivotal role not only to determine sparsity of the network, but also, to decide which genes we include into the final model. So, we suggest to use CAIC, CAICF and ICOMP

as a variable selection procedure in this setup, since they are motivated from the information theory and the particularly ICOMP is designed to take the complexity and interactions into account. Our objective is to compare our suggestions with both the state-of-art model selection criteria, AIC and BIC, and the high dimensional model selection methods: RIC, EBIC and StARS. We compare them in terms of accuracy measures: F-measure, precision, recall and accuracy.

- **Under GGM Simulation Studies**
  - ICOMP outperforms the rest in terms of accuracy measures under both scale-free and random network topologies, even for the highest dimensional setting.
  - ICOMP dominates the others in terms of recall under all dimensional settings for each topology.
- **Under GGM Application Studies**
  - ICOMP achieves to capture some of the interactions in the network while high dimensional model selection procedures (RIC,EBIC and StARS) cannot detect any interactions.
  - Our proposed procedures are more efficient on the sparse networks Data 1 and 2 structure than the dense networks Data 3-6.

Under the MARS setting, we aim to modify two stage iterative algorithms that MARS originally use in order to pick the best final model. This is done by changing the backward algorithm in such a way that we replace GCV with our proposed data-dependent model selection procedures. In this section, the associated LMARS algorithm and the flowchart representation are given. Here, our objective is to compare our suggested model selection procedures with GCV and other classical model selection methods (AIC and BIC) in terms of either accuracy measures or MARS models selected by each criterion.

- **Under LMARS Simulation Studies**

  - When the dimension increases, CAIC obtains the the best results in terms of accuracy under the scale-free topology for both the lowest and the highest settings. Under this topology, ICOMP succeeds to get the best accuracy measure for the moderate dimensional setup. In terms of F measures, CAIC, CAICF and GCV get the best results, respectively, for the dimensions from the lowest to highest. According to recall, ICOMP obtains the best value under the highest scale-free dimension.

  - Under random networks, in terms of accuracy, each criterion achieves to obtain the best for the highest dimensional setting. With respect to recall, CAIC outperforms the others under the highest dimensional random structure.

- **Under MARS Application Studies**

  - While GCV and AIC tend to more complicated MARS models that are hard to interpret, ICOMP achieves to choose the simplest and interpretable model by taking solely the important genes in the network structure.

Overall

- ICOMP outperforms the others in terms recall under all dimensional settings for both topologies under the GGM setting. This also far exceeds the values under the LMARS setting.

- Each criterion achieves to not detect falsely positive value (prec=1) under the LMARS setting.

- In terms of accuracy for the two highest dimensions, ICOMP succeeds in the best values under the GGM setting for each topology. Particularly for the highest setting, ICOMP achieves to obtain an accuracy measure as 1.

- Our suggested model selection methods operate the efficient procedure in terms of accuracy measures for sparse networks than the dense ones.

- Scale-free network structure can be seen as our main objective in order to fit our novel applications under GGM and MARS settings since this structure satisfies our expectation from systems biology along with probability theory.

This thesis broadens our perspectives not only for model selection procedure of LMARS but also for the variable selection procedure where the penalized likelihood approach uses a penalty parameter in order to penalize the regression under GGM setting.

As the future study, since our suggested model selection criteria are compatible with both parametric and nonparametric modelling approaches, we consider to apply our suggested data-dependent model selection criteria on other modelling approaches, i.e., neural networks. Also, they can be combined with several penalized likelihood approaches, i.e., adaptive lasso or group lasso.

Bibliography

[1] Abegaz, F. and Wit, E. (2013). Sparse time series chain graphical models for reconstructing genetic networks. *Biostatistics*, 14(3):586–599.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*.

[3] Alicia Subtil-Rodríguez, Elena Vázquez-Chávez, M. C.-C. M. R.-P. J. I. M.-S. M. E. J. C. R. (2014). The chromatin remodeller chd8 is required for e2f-dependent transcription activation of s-phase genes. 42:2185–2196. data retrieved from NCBI (Nucleic Acids Reserve) PMID: 24265227 `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE48926`.

[4] Alpaydin, E. (2010). *Introduction to Machine Learning*. The MIT Press, 2nd edition.

[5] Aster, R. C., Borchers, B., and Thurber, C. H. (2018). *Parameter estimation and inverse problems*. Elsevier.

[6] Ayyıldız, E., Purutçuoğlu, V., and Weber, G. W. (2018). Loop-based conic multivariate adaptive regression splines is a novel method for advanced construction of complex biological networks. *European Journal of Operational Research*, 270(3):852–861.

[7] Ayyıldız, E., Ağraz, M., and Purutçuoğlu, V. (2017). Mars as an alternative approach of gaussian graphical model for biochemical networks. *Journal of Applied Statistics*.

[8] Ayyıldız, E. and Purutçuoğlu, V. (2013). Gaussian graphical approaches in estimation of biological systems. Master's thesis, Middle East Technical University.

[9] Ağraz, M. and Purutçuoğlu, V. (2017). Different types of modellings and the inference of model parameters for complex biological systems. Master's thesis, Middle East Technical University.

[10] Bahçivancı, B., Purutçuoğlu, V., Purutçuoglu, E., and Ürün, Y. (2018). Estimation of gynecological cancer networks via target proteins. *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*.

[11] Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516.

[12] Banks, D. L., Olszewski, R. T., and Maxion, R. A. (2003). Comparing methods for multivariate nonparametric regression. *Communications in Statistics - Simulation and Computation*, 32(2):541–571.

[13] Barabási, A.-L. A. and Oltvai, Z. N. P. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*.

[14] Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust optimization*, volume 28. Princeton University Press.

[15] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308.

[16] Bower, J. M. and Bolouri, H. (2001). *Computational Modeling of Genetic and Biochemical Networks*. MIT Press, Cambridge MA.

[17] Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*.

[18] Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of mathematical psychology*, 44(1):62–91.

[19] Bozdogan, H. (2010). A new class of information complexity (icomp) criteria with an application to customer profiling and segmentation. *Istanbul University Journal of the School of Business Administration*.

[20] Bozdogan, H. and Haughton, D. M. (1998). Informational complexity criteria for regression models. *Computational Statistics Data Analysis*, pages 51–76.

[21] Breiman, L. (1984). *Classification and Regression Trees*. New York: Routledge.

[22] Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*.

[23] Breiman, L. (2001). Random forests. *Machine Learning*.

[24] Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R., and Kohane, I. S. (2000). Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, 97(22):12182–12186.

[25] Bülbül, G., Purutcuoglu, V., and Purutcuoglulu, E. (2019). Novel model selection criteria on sparse biological networks. *International Journal of Environmental Science and Technology*, pages 1–6.

[26] Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n.

[27] Castelo, R. and Roverato, A. (2006). A robust procedure for Gaussian graphical model search from microarray data with p larger than n. *J. Mach. Learn. Res.*, 7.

[28] Cercignani, C. (1988). *In The Boltzmann equation and its applications*.

[29] Chen, S. S., Donoho, D. L., and Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM*, pages 129–159.

[30] Chen, Z. and Chen, J. (2009). Tournament screening cum ebic for feature selection with high-dimensional feature spaces. *Science in China Series A: Mathematics*, pages 1327–1341.

[31] Cox, D. R. and Hinkley, D. V. (1979). *Theoretical statistics*. Chapman and Hall/CRC.

[32] Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*.

[33] Davis, M. H. A. and Vinter, R. B. (1985). *Stochastic Modelling and Control*. Chapman and Hall, New York.

[34] De La Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574.

[35] Deconinck, E., Zhang, M., Petitet, F., Dubus, E., Ijjaali, I., Coomans, D., and Vander Heyden, Y. (2008). Boosted regression trees, multivariate adaptive regression splines and their two-step combinations with multiple linear regression or partial least squares to predict blood–brain barrier passage: a case study. *Analytica chimica acta*, 609(1):13–23.

[36] Dempster, A. P. (1972). Covariance selection. *Biometrics*, pages 157–175.

[37] Denison, D. G., Mallick Denison, B. K., and Smith, A. F. (1998). Bayesian mars. *Statistics and Computing*, 8(4):337–346.

[38] Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*.

[39] Dobra, A. and Lenkoski, A. (2011). Copula gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*.

[40] Dokuzoğlu, D. and Purutçuoğlu, V. (2016). Application of copulas in graphical models for inference of biological systems. Master's thesis, Middle East Technical University.

[41] Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*.

[42] Donoho, D. L. and Johnstone, I. M. (1994). Threshold selection for wavelet shrinkage of noisy data. volume 1, pages A24–A25. IEEE.

[43] Easley, D. and Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press.

[44] Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American statistical Association*, 81(394):461–470.

[45] Efron, B. and Hastie, T. (2016). *Computer age statistical inference*. Cambridge University Press.

[46] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression.

[47] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868.

[48] El-Nasr, M. S., Yen, J., and Ioerger, T. R. (2000). Flame—fuzzy logic adaptive model of emotions. *Autonomous Agents and Multi-Agent Systems*, pages 219–257.

[49] Emden, H. F. V. and Bashford., M. A. (1971). The performance of brevicoryne brassicae and myzus persicae in relation to plant age and leaf amino acids. *Entomologia Experimentalis et Applicata*.

[50] Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *In Advances in neural information processing systems*.

[51] Fan, J. and Lv, J. (2012). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 849–911.

[52] Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, pages 1947–1975.

[53] Foygel, R. and Drton, M. (2010). Extended bayesian information criteria for gaussian graphical models. *In Advances in neural information processing systems*, pages 604–612.

[54] Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, pages 109–135.

[55] Frank J Fabozziand Kolm, P. N., Pachamanova, D. A., and Focardi, S. M. (2007). *Robust portfolio optimization and management*. John Wiley & Sons.

[56] Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*.

[57] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statical Software*.

[58] Friedman, J. and Popescu, B. E. (2004). Gradient directed regularization. *Unpublished manuscript, http://www-stat. stanford. edu/˜ jhf/ftp/pathlite. pdf*.

[59] Friedman, J. H. (1991). Multivariate adaptive regression splines.

[60] Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics*, 31(1):3–21.

[61] Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*.

[62] Geer, S. A. V. D. and Houwelingen, H. C. V. (1999). High-dimensional data: p» n in mathematical statistics and bio-medical applications. *Bernoulli*, pages 939–943.

[63] Ghaoui, L. E. and Lebret, H. (1997). Robust solutions to least-squares problems with uncertain data. *SIAM Journal on matrix analysis and applications*, 18(4):1035–1064.

[64] Ghasemi, J. B. and Zolfonoun, E. (2013). Application of principal component analysis–multivariate adaptive regression splines for the simultaneous spectrofluorimetric determination of dialkyltins in micellar media. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 115:357–363.

[65] Golightly, A. and Wilkinson, D. J. (2011). Bayesian parameter inference for stochastic biochemical network models using particle markov chain monte carlo. *Interface focus*.

[66] Han Liu, K. R. and Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. *In Advances in neural information processing systems*, 13:1432–1440.

[67] Hansen, P. C. (1999). The l-curve and its use in the numerical treatment of inverse problems.

[68] Hastie, T. and Tibshirani, R. (1987). Generalized additive models: some applications. *Biometrics*, pages 371–386.

[69] Hastie, T., Tibshirani, R., and Wainwright, M. (2005). *Statistical Learning with Sparsity: The lasso and Generalizations*. New York: Chapman and Hall/CRC.

[70] Haughton, D. M. (1988). On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, pages 342–355.

[71] Heinze, G., Wallisch, C., and Dunkler, D. (2018). Variable selection–a review and recommendations for the practicing statistician. *Biometrical Journal*, 3:431–449.

[72] Hoerl, A. E. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

[73] Holmes, C. and Denison, D. (2003). Classification with bayesian mars. *Machine Learning*, 50(1-2):159–173.

[74] Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, pages 1603–1618.

[75] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*.

[76] Johnson, M., Duvenaud, D. K., Wiltschko, A., Adams, R. P., and Datta, S. R. (2016). Composing graphical models with neural networks for structured representations and fast inference. pages 2946–2954.

[77] Jordan, M. I. (2010). Bayesian nonparametric learning: Expressive priors for intelligent systems. *Heuristics, probability and causality: A tribute to Judea Pearl*, 11:167–185.

[78] Kartal-Koc, E. and Bozdogan, H. (2015). Model selection in multivariate adaptive regression splines (mars) using information complexity as the fitness function. *Machine Learning*, 101(1-3):35–58.

[79] Kendall, M. G. and Stuart, A. (1967). *The Advanced Theory of Statistics*.

[80] Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models - Principles and Techniques*. The MIT Press, Cambridge, Massachusetts-London, England.

[81] Korb, K. and Nicholson, A. E. (2003). *Bayesian Artificial Intelligence*. CRC Press, Inc., Boca Raton, FL, USA.

[82] Kriner, M. (2007). *Survival analysis with multivariate adaptive regression splines*. PhD thesis, lmu.

[83] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, pages 79–86.

[84] Küffner, R., Petri, T., Windhager, L., and Zimmer, R. (2004). Petri nets with fuzzy logic (pnfl): Reverse engineering and parametrization. *PLOS ONE*.

[85] Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

[86] Ledoit, O. and Wolf, M. (2004). Honey, i shrunk the sample covariance matrix.

[87] Lee, Y. and Wu, H. (2012). Mars approach for global sensitivity analysis of differential equation models with applications to dynamics of influenza infection. *Bulletin of mathematical biology*, 74(1):73–90.

[88] Li, H. and Gui, J. (2005). Gradient directed regularization for sparse gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*, pages 302–317.

[89] Lindley, D. V. (1968). The choice of variables in multiple regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30:31–53.

[90] Lysen, S. (2009). *Permuted inclusion criterion: a variable selection technique*. PhD thesis, Publicly accessible Penn Dissertations.

[91] Lü, L., Chen, D., Ren, X.-L., Zhang, Q.-M., Zhang, Y.-C., and Zhou, T. (2016). Vital nodes identification in complex networks. *Physics Reports*.

[92] Magwene, P. M. and Kim, J. (2004). Estimating genomic coexpression networks using first-order conditional independence. *Genome biology*, 5(12):R100.

[93] Markowetz, F. and Spang, R. (2007). Inferring cellular networks–a review. *BMC bioinformatics*, 8(6):S5.

[94] Mazumder, R. and Hastie, T. (2012). The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, pages 2125–2149.

[95] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *"The Annals of Statistics"*.

[96] Milborrow, S. (2017). Notes on the earth package retrieved october 31 (2014).

[97] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. pages 758–765.

[98] Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). Asymptotic properties of criteria for selection of variables in multiple regression. *IMA journal of numerical analysis*, pages 389–403.

[99] Ozmen, A., Weber, G. W., İnci Batmaz, and Kropat, E. (2011). Rcmars: Robustification of cmars with different scenarios under polyhedral uncertainty set. *Communications in Nonlinear Science and Numerical Simulation*.

[100] Ozmen, A., Weber, W. G.-W., and Kropat, E. (2015). Robustification of conic generalized partial linear models under polyhedral uncertainty. *Problems of Nonlinear Analysis in Engineering Systems*, 38.

[101] Pearl, J. et al. (1989). *Probabilistic semantics for nonmonotonic reasoning: A survey*. University of California (Los Angeles). Computer Science Department.

[102] Purutçuoğlu, V., Ağraz, M., and Wit, E. (2017). Bernstein approximations in glasso-based estimation of biological networks. *Canadian Journal of Statistics*.

[103] Rao, J. N. and Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the american statistical association*, pages 231–241.

[104] Rijsbergen, D. J. (1979). *Information retrieval*.

[105] Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, pages 465–471.

[106] Sachs, K., Perez, O., Peter, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.

[107] Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). *Akaike information criterion statistics*. Dordrecht, The Netherlands: D. Reidel.

[108] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*.

[109] Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., Vespignani, A., and White, D. R. (2009). Economic networks: The new challenges. *Science*.

[110] Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*.

[111] Scutari, M. and Strimmer, K. (2010). *Introduction to graphical modelling*.

[112] Searle, S. R. and Gruber, M. H. (1971). *Linear models*. New York: Wiley.

[113] Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association*, pages 486–494.

[114] Silvey, J. (1975). *Deciphering data: the analysis of social surveys*. Longman Publishing Group.

[115] Steen, M. V. (2010). Graph theory and complex networks.

[116] Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):44–47.

[117] Tanju, O. and Kalaylıoğlu, Z. (2016). Cluster based model diagnostic for logistic regression. Master's thesis, Middle East Technical University.

[118] Taylan, P., Weber, G.-W., and Yerlikaya-Özkurt, F. (2008). A new approach to multivariate adaptive regression spline by using tikhonov regularization and continuous optimization. *Top*.

[119] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[120] Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282.

[121] Tibshirani, R. and Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 529–546.

[122] Tibshirani, R., Wainwright, M., and Hastie, T. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

[123] Varol, D., Purutçuoğlu, V., and Yılmaz, R. (2014). *Comparative Statistical Microarray Analysis of Yeast Data under Heat Shock Stress*. PhD thesis, Middle East Technical University.

[124] Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, pages 426–482.

[125] Weber, G.-W., Batmaz, I., Köksal, G., Taylan, P., and Yerlikaya-Özkurt, F. (2012). Cmars: A new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization. *Inverse Problems in Science and Engineering*.

[126] West, D. B. (1996). *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, NJ.

[127] Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. Wiley.

[128] Wilkinson, D. (2007). Bayesian methods in bioinformatics and computational systems biology. *Briefings in Bioinformatics*.

[129] Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., and Bühlmann, P. (2004). Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome Biology*.

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

[131] Wu, T. T., Lange, K., et al. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244.

[132] Wynn, T. A., Chawla, A., and W.Pollard, J. (2012). Macrophage biology in development, homeostasis and disease. *Nature*.

[133] Yazici, C. and İnci Batmaz (2011). A computational approach to nonparametric regression: Bootstrapping cmars method. Master's thesis, Middle East Technical University.

[134] Yazıcı, C., Yerlikaya-Özkurt, F., and Batmaz, I. (2015). A computational approach to nonparametric regression: bootstrapping cmars method. *Machine Learning*, 101(1):211–230.

[135] Yerlikaya-Özkurt, F., Batmaz, I., and Weber, G.-W. (2014). Modeling, dynamics, optimization and bioeconomics i. pages 695–722. Springer International Publishing.

[136] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

[137] Zakeri, I. F., Adolph, A. L., Puyau, M. R., Vohra, F. A., and Butte, N. F. (2012). Cross-sectional time series and multivariate adaptive regression splines models using accelerometry and heart rate predict energy expenditure of preschoolers. *The Journal of nutrition*, 143(1):114–122.

[138] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563.

[139] Zhao, P. and Yu, B. (2007). Stagewise lasso. *Journal of Machine Learning Research*, 8:2701–2726.

[140] Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in r. *Journal of Machine Learning Research*, 13:1059–1062.

[141] Zhou, S., Chen, L., and Sun, F. (2007). Optimization of constructal economics for volume-to-point transport. *Applied energy*, 84(5):505–511.

[142] Zhu, Y. and Cribben, I. (2018). Sparse graphical models for functional connectivity networks: best methods and the autocorrelation issue. *Brain connectivity*, 3:139–165.

[143] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.