INFERENCE OF LARGE-SCALE NETWORKS VIA STATISTICAL APPROACHES

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

BY

EZGİ AYYILDIZ DEMİRCİ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
STATISTICS

SEPTEMBER 2019

Approval of the thesis:

INFERENCE OF LARGE-SCALE NETWORKS VIA STATISTICAL APPROACHES

submitted by **EZGİ AYYILDIZ DEMİRCİ** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy 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. Yaprak Arzu Özdemir
Statistics, Gazi University _____

Prof. Dr. Vilda Purutçuoğlu
Statistics, METU _____

Prof. Dr. Gerhard Wilhelm Weber
Marketing and Economic Eng., Poznan Uni. of Tech. _____

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

Assoc. Prof. Dr. Ceylan Yozgatlıgil
Statistics, METU _____

Date: 10.09.2019

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

Name, Surname: Ezgi Ayyıldız Demirci

Signature:

ABSTRACT

INFERENCE OF LARGE-SCALE NETWORKS VIA STATISTICAL APPROACHES

Ayyıldız Demirci, Ezgi
Doctor of Philosophy, Statistics
Supervisor: Prof. Dr. Vilda Purutçuoğlu

September 2019, 73 pages

In system biology, the interactions between components such as genes, proteins, can be represented by a network. To understand the molecular mechanism of complex biological systems, construction of their networks plays a crucial role. However, estimation of these networks is a challenging problem because of their high dimensional and sparse structures. The Gaussian graphical model (GGM) is widely used approach to construct the undirected networks. GGM define the interactions between species by using the conditional dependencies of the multivariate normality assumption. However, when the dimension of the systems is high, the performance of the model becomes computationally demanding, and the accuracy of GGM decreases when the observations are far from normality. In this thesis, we suggest a conic multivariate adaptive regression splines (CMARS) as an alternative to GGM to overcome both problems. CMARS is one of the recent nonparametric methods developed for high dimensional and correlated data. We adapted CMARS to describe biological systems and called it "LCMARS" due to its loop-based description. Here, we generate various scenarios based on distinct distributions and dimensions to compare the performance of LCMARS with MARS and GGM in terms of accuracy measures via Monte Carlo runs. Additionally, different real biological datasets are used to observe the performance of underlying methods. Furthermore, in this study,

we perform various outlier detection methods as a pre-processing step before modeling the networks in order to investigate whether the outlier detection can improve the accuracy of the model. In the analysis, several synthetic and real benchmark biological datasets are used.

Keywords: Gaussian graphical model, Conic multivariate adaptive regression splines, Protein-protein interaction networks, Multivariate adaptive regression splines, Outlier detection

ÖZ

GENİŞ ÖLÇEKLİ AĞLARIN İSTATİSTİKSEL YAKLAŞIMLARLA TAHMİNİ

Ayyıldız Demirci, Ezgi
Doktora, İstatistik
Tez Danışmanı: Prof. Dr. Vilda Purutçuoğlu

Eylül 2019, 73 sayfa

Sistem biyolojisinde gen, protein gibi yapıların etkileşimleri bir ağ yapısı ile temsil edilebilir. Kompleks biyolojik sistemlerin moleküler mekanizmalarının anlaşılabilmesi için ağ yapılarının oluşturulması önemli bir rol oynar. Fakat bu ağların tahmini, yüksek boyutları ve seyrek yapıları sebebiyle zordur. Gaussian grafiksel modeli (GGM) yönsüz ağların oluşturulmasında sıklıkla kullanılan bir yöntemdir. Bu yöntem, bileşenler arasındaki etkileşimi çok değişkenli normallik varsayımı altında koşullu bağımlılık ile tanımlar. Fakat sistemin boyutları büyüdüğünde hesaplama zorluğu ile karşılaşılır ve GGM yönteminin güvenilirliği gözlemler normallikten uzaklaştıkça azalır. Bu çalışmada, bahsedilen problemlerin çözümü için GGM yöntemine alternatif olarak konik çok değişkenli uyarlamalı regresyon eğrilerinin (CMARS) kullanılmasını öneriyoruz. Bu yöntem, yüksek boyutlu ve ilişkili veriler için geliştirilen parametrik olmayan yöntemlerden biridir. CMARS'ı biyolojik sistemleri tanımlamak için uyarladık ve döngü temelli tanımı nedeniyle LCMARS olarak adlandırdık. Çalışmada farklı dağılımlar ve boyutlar altında çeşitli senaryolar üretilmiş ve LCMARS, MARS ve GGM yöntemlerinin performansları farklı doğruluk ölçüleri kullanılarak Monte Carlo simülasyonları ile karşılaştırılmıştır. Ek olarak, farklı gerçek biyolojik veri setleri bu yöntemlerin performanslarını gözlemlemek için kullanılmıştır. Ayrıca bu çalışmada, ağların modellenmesinden önce bir ön işlem

adımı olarak çeşitli aykırı değer saptama yöntemlerinin kullanılmasının modelin doğruluğunu arttırıp arttırmadığı araştırılmıştır. Analizlerde çeşitli sentetik ve gerçek biyolojik veri setleri kullanılmaktadır.

Anahtar Kelimeler: Gaussian grafiksel modeli, Konik çok değişkenli uyarlamalı regresyon eğrileri, Protein etkileşim ağları, Çok değişkenli uyarlamalı regresyon eğrileri, Aykırı değer saptama

To my mom and grandma

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my thesis supervisor Prof. Dr. Vilda Purutçuoğlu for her everlasting support during my research. Her continuous guidance, and feedbacks have turned this study to an immeasurable learning experience for me.

I would like to present my grateful thanks to Prof. Dr. Gerhard Wilhelm Weber and Prof. Dr. Ömür Uğur for valuable ideas and contributions that they made to this study. I would also like to thank the other members of my examining committee, Prof. Dr. Yaprak Arzu Özdemir and Assoc. Prof. Dr. Ceylan Yozgatlıgil for their detailed reviews and constructive comments.

I would like to thank The Scientific and Technological Research Council of Turkey (TÜBİTAK) for their financial support.

I owe my special thanks to my best friend Duygu Varol for her full support, motivation and great friendship. I am grateful to my friends Deniz - Ada Çelikel, Özge Gürer, Seher Gök, Tülay Akal and Gamze Musluoğlu for their nice friendship, everlasting patience, and loving kindness. Also, I would like to thank all the members of the Department of Statistics, METU.

I am grateful to my grandma who supported me very much in my whole life. I miss her so much. I want to thank to Muazzez Gürgan for being such a good cousin.

My special thanks go to my dear husband, Samet Demirci whose love, trust and support make everything possible. I am very thankful to my dog, Jack for making my life wonderful.

Finally, my deepest gratitude to my beloved family. I give my special thanks to my wonderful mother, Nuran Ayyıldız for her endless love and unconditional support throughout my life, to my father Haydar Ayyıldız, for teaching mathematics to me and my brother Mert Ayyıldız, for supporting me through every step of the way.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

BACON	Blocked Adaptive Computationally Efficient Outlier Nominator
BF	Basis Function
CMARS	Conic Multivariate Adaptive Regression Splines
CQP	Conic Quadratic Programming
CV	Cross Validation
FN	False Negative
FP	False Positive
GCGM	Gaussian Copula Graphical Model
GCV	Generalized Cross Validation
GGM	Gaussian Graphical Model
Glasso	Graphical Lasso
GR	Glucocorticoid Receptor
IQR	Inter Quantile Range
LARS	Least Angle Regression
Lasso	Least Absolute Shrinkage and Selection Operator
LCMARS	Loop-based Conic Multivariate Adaptive Regression Splines
LMARS	Loop-based Multivariate Adaptive Regression Splines
MAD	Median Absolute Deviation
MARS	Multivariate Adaptive Regression Splines
MCD	Minimum Covariate Determinant
MCMC	Markov Chain Monte Carlo
MD	Mahalonobis Distance
MLE	Maximum Likelihood Estimation

MSE	Mean Square Error
MSOM	Mean Shift Outlier Model
ODE	Ordinary Differential Equations
OLS	Ordinary Least Squares
PC	Principal Component
PCA	Principal Component Analysis
PLM	Partial Linear Models
PRSS	Penalized Residual Sum of Squares
RCGPLM	Robust Conic Generalized Partial Linear Model
RCMARS	Robust Conic Multivariate Adaptive Regression Splines
THR	Thyroid Hormone Receptor
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TR	Tikhonov Regularization
QP	Quadratic Programming

CHAPTER 1

INTRODUCTION AND MOTIVATION

Rapid developments in biotechnology and medicine over the last few decades have resulted in the production of large amounts of biological, biochemical, and genomic data. Shifting through massive data has created certain computational challenges. One of the issues encountered is the prediction of complex biological networks and the estimation of the unknown system's parameters. Biological systems can be represented in the form of networks such as gene regulatory networks, signal transduction networks, protein-protein interactions and metabolic networks. Biological networks are related to molecules such as DNA, RNA, proteins and metabolites, and the networks describe the interactions between these systems' elements. Gene regulatory and signal transduction networks include interactions among genes and present how genes can be activated or repressed. On the other hand, protein-protein interaction networks represent the interaction between proteins such as the building of protein complexes and the activation of one protein by another protein. By this way, we can understand the molecular mechanism of complex biological systems in a living organism. In a biological network, genes or proteins are denoted by nodes and their interactions are indicated with edges. A simple representation of a network with 500 nodes can be seen in Figure 1.1 for illustration. There are various statistical methods to construct the structure of biological networks. Some of these approaches are based on the implementation of piecewise linear differential equations [106], ordinary differential equations [15, 33], connections with other modeling approaches [16, 21], or the application of stochastic methods [37, 96]. In the calculation of all these modeling approaches, various methods are used, such as particle swarm optimization [20], simulating annealing, nonlinear programming [32, 64, 76], Metropolis Hasting [18], Euler and Runge Kutta methods [24, 47, 91]. All of these approaches are obtained by

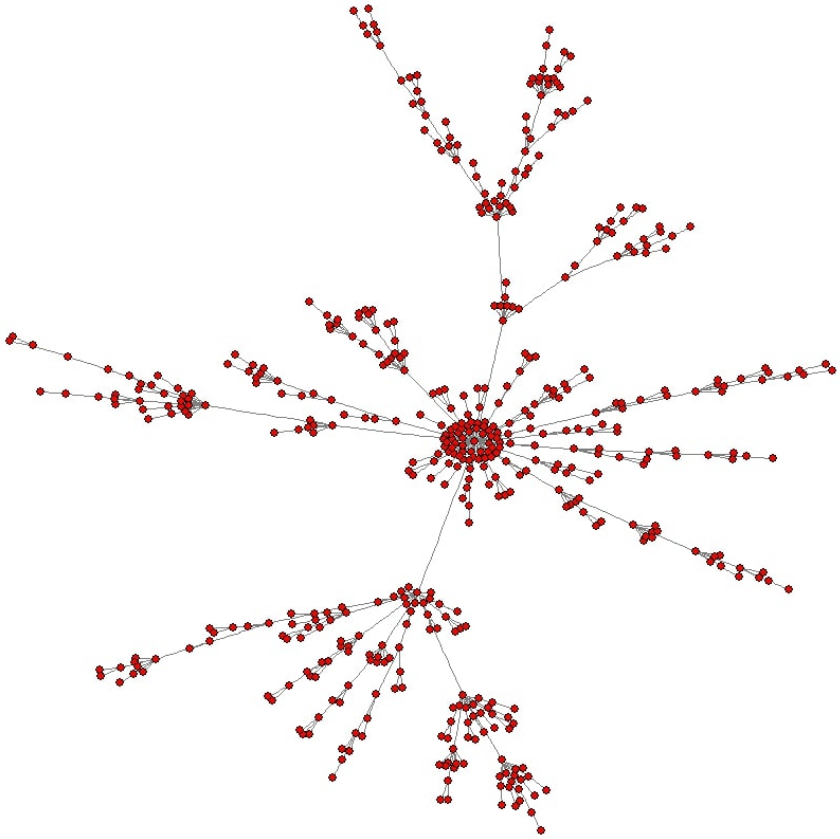


Figure 1.1: Simple representation of a network with 500 nodes.

different types of optimization techniques.

Graphical methods are another type of modelings that are very similar to the description of the continuous-time ordinary differential equations (ODE). Generally, in a biological system, the interactions between species, which refer to the structure of the system, can be represented by graphs. One of the most widely used graphical models is the Gaussian graphical model (GGM) [95]. GGM describes the interactions in the system by using the conditional independency of the species under the assumption that the states of the systems, i.e., the measurements, have the multivariate normal distribution. In GGM, conditional independency can be presented by the zero entries in a precision matrix θ, which is the inverse of the variance-covariance matrix, i.e., $\Sigma^{-1} = \theta$. However, particularly in high-dimensional biological systems, the estimation of the precision matrix can be challenging due to the sparsity in θ [31], and the inference of the model parameter becomes computationally demanding. The Gaussian copula graphical model (GCGM) can be thought as the Bayesian alterna-

tive of GGM by using the Gaussian copula in the separation of the high dimensional multivariate normal distribution. Hereby, the method decomposes θ via the Cholesky method and updates the states of the systems as well as the structure of the networks via various Markov Chain Monte Carlo (MCMC) algorithms such as the reversible jumps MCMC [66] or birth-and-death approach [65].

Linear regression problems in high-dimensional settings occur in several fields. In high-dimensional data, the number of predictor variables can be much larger than the number of observations. In such situations, the variable selection has a crucial role in statistical modeling. In system biology, the biological systems from metabolic networks to protein-protein interaction networks are described via the sparse networks in which there are a few important variables being highly correlated and crucial for the activation of the system and many variables having less connections. But as the number of observations is much more smaller than the number of variables, the detection of the best model which explains these connections is challenging. There are many variable selection methods that can be used under this condition such as all-subsets, stepwise regression, least-angle regression, ridge regression and Lasso-type methods in the most applicable one. The purpose of all these model (variable) selection algorithms are to choose the accurate predictors which give the biologically interpretable and stable findings. In this context, stability means that small changes in the data do not cause large changes in the selected subset of predictors and associated coefficients.

Among many alternatives, each model has its own advantages and drawbacks. For example, all-subsets and stepwise regression cannot handle high-dimensional data, such as gene expression dataset. Under these conditions, the penalized-based methods can be more appropriate. On the other hand, in the ridge regression even the variance reduction is achieved to compare the least-squares approach, it does not produce a sparse model. The lasso model generates sparse models. However when the number of predictors p is much more larger than the number of observations n, it chooses at most n variable. Additionally, the adaptive Lasso model performs different weights for penalized coefficients. But, since it has an oracle property, it obtains better variable selection accuracy compared with Lasso. On the other side, the Gaussian graphical model which is based on the Lasso regression and penalized likelihood

inference has normality assumption. When this assumption is not satisfied, the power of the method decreases substantially. Moreover, the fused Lasso model takes into account the order of features as typically observed in protein mass spectroscopy and microarray types of data. However, the application of this method for high-dimensional problems is computationally challenging with respect to the time and the space. The elastic net method can deal with the grouping effect different from lasso, whereas, the number of estimated parameters increases significantly. But among many alternatives the biologically more relevant and common applications are seen under the GGM. In GGM, even though the main drawback is the strict normality assumption, the interpretation of the optimal model is valid. On the other side, the second challenge of this model is its computational demand in inference of complex networks. In order to overcome this problem, all the previously underlied models and their inference strategies are adapted to GGM. However, for large systems under high sparsity, the inference of GGM is still computationally demanding. Because of this problem and strict normality assumption, a nonparametric model, so called Multivariate Adaptive Regression Splines (MARS), has been proposed as an alternative approach to GGM [4].

MARS is one of the well-known statistically nonparametric regression methods that enables us to model the high dimensional data under nonlinearity [29]. Also, it is a particular type of optimization techniques [7, 8, 9] in the sense that MARS aims to transform a non-differentiable problem into a smooth problem [7] by putting binary constraints for the approximation of the optimal value [9]. Hereby, it uses the gradient based schemes to solve the smooth and nonsmooth optimization problems [8]. Due to these advantages, it has been implemented in different fields from engineering [3] to time-series analyses [59]. In the study of Ayyıldız et al. [4], MARS was applied with only the main terms and the performance of both MARS and GGM was compared in terms of the model accuracy and the computational cost. From the analyses, it has been shown that MARS results with higher accuracy and gets huge gain in computational time, specifically, under large systems.

In this study, first of all, we propose a nonparametric regression model, called the Conic Multivariate Adaptive Regression Splines (CMARS) [86], as an alternate to GGM. The CMARS method is a modified version of MARS. Basically, the MARS

model is a special type of generalized additive model which represents the causes of error propagation via a regularized nonlinear regression model. Hereby, in CMARS by keeping these listed advantages of MARS, a penalized residual sum of squares (PRSS) is implemented by eliminating the backward stepwise algorithm [29, 44], and can be solved by conic quadratic optimization [86]. This model was originally designed for highly correlated datasets without distributional limitations. Due to its flexibility in modeling, it has wide applications from the fleet assignment problem [72] and financial analyses [85] to data mining [101] and eco-finance network analyses [92]. In addition, previous analyses have shown that CMARS produces more accurate results than its close alternates MARS and the so-called partial linear models (PLM) [44, 86]. In this study, we convert the original description of CMARS to a loop form, "LCMARS", by including the main effects of the model by discarding all interaction effects apart from second-order interactions iteratively. In fact, we use the second-order interactions as the representatives of the feed-forward loop [5, 58] in the biological networks' motifs. With this method, we aim to better describe the structure of biological systems and therefore gain a better understanding of complex diseases such as cancer and hepatitis. In our analyses, we compare LCMARS with GGM and LMARS in several real benchmark and simulated datasets under different dimensions based on the accuracy measures: recall, F-measure, Jaccard index, and accuracy values. Our results show promising benefits of LCMARS over GGM and LMARS.

In the second part of this study, we investigate whether the detection of outliers can improve the model accuracy under high dimensional and sparse biological data. For this purpose, we use outlier detection methods as a pre-process in advance of modeling with LCMARS. Outlier detection is a crucial problem in many fields. Although there are too many outlier detection methods in the literature, only a few methods suitable for dependent, sparse and high-dimensional data structure. In this study, we perform various univariate and multivariate outlier detection methods as a pre-processing step before modeling the protein-protein interaction networks in order to investigate whether the outlier detection can improve the accuracy of the model. Within the univariate approaches, we implement the z-score [81] and Box-plot [90] methods which are the most well-known outlier detection approaches. Besides them, we also ap-

ply the multivariate outlier detection methods, called PCOut and Sign [27] which are based on the robust principal component analysis and the blocked adaptive computationally efficient outlier nominators (BACON) method which is a distance-based approach [14]. These methods are applicable for the data type such that the number of variables is bigger than the observations. In the analysis, we use several synthetic and real benchmark biological datasets. From the results, it has seen that all the listed outlier detection methods cannot improve the accuracy of the models when we perform them as a pre-processing step. Furthermore, within the outlier detection methods, there is no any method which outperform the others in the construction of biological networks. We propose that the raw data can be directly used for the mathematical modeling of the protein-protein interaction networks.

Accordingly, in the organization of the study, the most well-known penalized regression methods are introduced in Chapter 2. Gaussian graphical model, nonparametric regression methods and the proposed method, i.e., loop-based CMARS, are explained in details. Then, outlier detection methods are presented in this chapter. In Chapter 3, we compare the proposed method with alternates. In this chapter, to compare the methods we use different dimensional simulated, synthetic and real biological datasets. Here, both the description of real data and their results are represented. Additionally, we investigate the validity of outlier detection methods under both synthetic and real biological datasets. Finally, in Chapter 4, we conclude the findings and discuss possible future works.

CHAPTER 2

PROPOSED METHOD AND BACKGROUND

In this chapter, we present the proposed method which is used to estimate the complex biological systems and its alternates. This chapter consists of five sections. In Section 2.1, the most well-known penalized regression methods are introduced. Gaussian graphical model and the estimations methods of this model are defined in Section 2.2. In Section 2.3, nonparametric regression methods and the proposed method, i.e., loop-based CMARS, are explained in details. Outlier detection methods are presented in Section 2.4. Finally, the model selection criteria which are used throughout the study are defined.

2.1 Regression Methods

In a biological system, the interactions between components can be represented by using graphical models. The graphical models consist of a set of nodes which is totally p, and a set of edges. The nodes represent the components, such as proteins and genes, in the system and the edges display the interaction among these components. The nodes can be formalized as a vector $Y = (Y^{(1)}, ..., Y^{(p)})$.

The sparse graphical models can be obtained by implementing the sparse regression model which searches the sparsity of each node one by one. When we divide the vector Y into two parts such as $Y = (Y^{(-p)}, Y^{(p)}))$ in which $Y^{(-p)} = (Y^{(1)}, ..., Y^{(p-1)})$ denotes the vector of all nodes except for the last one, a regression model for the last node $Y^{(p)}$ is found via

$$Y^{(p)} = Y^{(-p)}\beta + \varepsilon, \tag{2.1}$$

where ε denotes the independent and identically distributed random error.

In biological networks, the Lasso model and its extension such as the bridge regression and the ridge regression are the most common approaches in order to describe the steady-state, i.e., deterministic, behaviour of the system. These models can explain the activation of the species in the system via the conditional probability in the sense that the state of a species is presented linearly by the states of all remaining species, except that species, with a random error. In the following parts, we describe each Lasso-type penalized regression, GGM and its alternatives used for the deterministic modeling of biological networks.

2.1.1 Bridge Regression

The bridge regression [28] is the most comprehensive Lasso type of the penalized regression methods which is also generalized form of the Lasso and ridge regressions. In this model, the vector of regression coefficients β is estimated by minimizing the penalized sum of squares via the L_q-norm. The associated expression is represented as below:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^q. \tag{2.2}$$

Here, λ is a non-negative penalty constant, also called the tuning parameter, and q is a shrinkage parameter and β_0, β_j, x_{ij} and y_i stand for the intercept term, the regression coefficients, associated covariates and response, respectively. In bridge regression denoted in Equation (2.2), the optimal penalty term and shrinkage parameters are detected by the generalized cross validation (GCV) method. In this expression, when the shrinkage parameter sets to 1, the bridge regression corresponds to the Lasso regression. Furthermore, if $q = 2$, it turns to the ridge regression. If q lies in the open interval $0 < q < 1$, the bridge regression produces sparse models. However, if $q > 1$, it results with a nonsparse model meaning that the model contains all predictors. Since in high-dimensional data, the major aim in biological networks is to obtain a sparse model, this method is not an efficient technique for modeling protein-protein interaction networks. On the other hand, on comparison of the bridge regression with Lasso and ridge regression that are presented in the following parts, its computational cost does not differ significantly with respect to its alternates [94].

2.1.2 Ridge Regression

In the ridge regression [49] by using the same model given in Equation (2.1), the coefficients β are estimated by minimizing a penalized sum of squares via the L_2-norm as below:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2, \tag{2.3}$$

in which x_{ij} denotes the jth covariates of the response variable y while i is the index of the sample size ($i = 1, ..., n$) and β_j ($j = 1, ..., p$) stands for the regression coefficient of the jth covariate based on n samples. Moreover, β_0 shows the intercept term and λ denotes the tuning parameter. In Equation (2.3), the first summation term is the residual sum of squares and the second summation term is called a shrinkage penalty. This penalty is not applied to the intercept β_0 since it is a measure of the mean value of the response when x_i's are equal to zero. In ridge regression, the variables are standardized in order to obtain an invariant penalty to the scale of the original data.

When shrinkage penalty sets to 0, i.e., $\lambda = 0$, the penalty term is disappeared and the estimated regression coefficients β_j corresponds to the least-squares estimator. On the other hand, as λ increases, these coefficients approach zero without setting to zero. Hence, this regression includes all predictors with smaller coefficients than the least-squares estimates. But, it cannot exclude any variable from the model.

Finally in this model, for each value of λ, different set of estimated coefficient is produced. So the selection of a good value for the tuning parameter is a crucial issue. On the other side, the ridge regression has an advantage over the least-squares approach in the sense that the latter has no bias but its variance is high. In contrast, the former has lower variance of predictions when λ increases. But this also causes slight increase in bias. Hence, the optimal solution is found via the reduction of the mean square error (MSE). Moreover, when the number of variables p is larger than the number of observations n, i.e., $p > n$, there is no unique solution in the least-squares estimates, on contrary, the ridge regression can be applicable under this situation.

2.1.3 Lasso

The least absolute shrinkage and selection operator [87], in short: Lasso, is a constrained version of the least-squares methods. The Lasso-based approach is mainly used to estimate a sparse network, i.e., the network has small number of edges, resulting in many zero entries in the adjacency matrix. This matrix shows which nodes are adjacent to which other nodes. It uses binary representation and the entry 1 indicates that corresponding two nodes are adjacent. Actually, the sparsity is one of the general features in biological networks.

In this approach, the precision matrix is estimated by using a regression-based method. The networks that are inferred with this method is called the dependency network. Although the regression-based approach has important advantages, such as computational efficiency and good approximations to the joint distribution of the variables, it does not guarantee a symmetric variance-covariance matrix.

In standard regression models, a least-squares criterion is applied to estimate the coefficients β. On the other hand, the Lasso model chooses the regression coefficients β by minimizing the residual sum of squares subject to a constraint on the sum of absolute values of the coefficients. The Lasso model computes the L_1-penalty for β such that $\|\beta\|_1 = \Sigma_i|\beta_{ip}| < \lambda$. Through this difference, the sparsity of the precision matrix can be obtained and the solution is computed by

$$\text{minimize}_\beta \, [\|Y^{(p)} - Y^{(-p)}\beta\|_2^2 + \lambda_p\|\beta\|_1] \tag{2.4}$$

with a tuning parameter λ_p that enables us to estimate of the parameters β. Here, the second term is also named as the L_1-penalty. In Equation (2.4), $\|\cdot\|_1$ refers to L_1-norm which means the sum of the absolute values of the columns [95].

In Lasso, if the penalty parameter λ is sufficiently large, the L_1-penalty forces some of the estimated coefficients to be exactly equal to zero. By this way, some predictors can be excluded from the model. Because of this property, Lasso treats as a variable selection method, resulting in more simpler and more interpretable sparse models that include only a subset of variables. On the other hand, typically this model also has a small increase in bias, by gain in variance. Indeed this enables us to generate more accurate predictions.

Conversely, non-symmetric results are the main problems of this approach. It means that although $Y^{(j)}$ from the rest can be obtained for zero β_{ij}, β_{ji} is not zero when we predict $Y^{(i)}$ from the rest. One solution to unravel this challenge is to apply AND or OR rules. If we use AND rule, we obtain the zero coefficients when both β_{ij} and β_{ji} are zero. Furthermore, if we use OR rule for one of the parameters β_{ij} zero is enough to obtain zero coefficients. So, it is seen that the OR rule generates more sparse networks. However, we cannot calculate the actual strength of the edge since these are two different values because of asymmetry.

In the Lasso-regression approach, another important issue is the selection of the penalty parameters λ_i. To get a false positive less than α, the following penalty equation can be calculated.

$$\lambda_i = 2\sqrt{\frac{s_{ii}}{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2p^2}\right), \tag{2.5}$$

where Φ denotes the cumulative distribution function of the standard normal.

The penalized likelihood idea arises because of the fact that ordinary least-squares (OLS) estimates have problems in high-dimensional data. The prediction accuracy and interpretation are two main problems of OLS. Since, in general, the OLS estimates have a low bias but a large variance, by setting regression coefficients exactly zero, the precision accuracy increases [87]. Moreover, if the number of regression coefficients is greater than the observations, it is difficult to obtain an interpretable model. L_1 and L_2-penalties are used to overcome these challenges. L_1 absolute value penalty and L_2 quadratic (ridge) penalty shrink the coefficients towards zero. However, there is a difference between these two penalized estimation method. The regression coefficients are small but non-zero in the L_2-penalized method. On the other side, the L_1-penalized regression sets the coefficients exactly to 0. This feature causes to obtain an interpretable results in the L_1-penalized regression method [36].

2.1.4 Adaptive Lasso

The adaptive Lasso method is an alternative version of the ordinary Lasso regression which implements the L_1-penalty to estimate β. But additionally, different weights

are assigned for each regression coefficient β_j to penalize [105] as defined below:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}w_j|\beta_j|, \tag{2.6}$$

in which x_{ij} denotes the jth covariates of the response variable y_i and β_j $(j=1,...,p)$ stands for the regression coefficient of the jth covariate based on n samples as stated beforehand. Finally, β_0 shows the intercept term indicating the baseline effect without any covariates. Furthermore, λ is the penalty and w represents the weight vector which is dependent on data. The weights are generated by using the following equation.

$$w_j=\frac{1}{|\hat{\beta}_j|^{\gamma}}, \text{for } \gamma>0, \tag{2.7}$$

where $\hat{\beta}_j$ is the ordinary least-squares estimate. By applying weights, the higher penalty is given to small coefficients and the lower penalty is assigned to large coefficients. In general, the adaptive Lasso reduces the estimation bias and also obtains better variable selection accuracy compared with Lasso in Equation (2.1).

On the other side, this model also satisfies the oracle property. This feature means that the interested model can correctly identify the nonzero coefficients by converging probability to one. Moreover, the estimators of these nonzero coefficients are asymptotically normal with the same mean and covariance [50]. This is a main theoretical advantage of the adaptive Lasso model with respect to the ordinary Lasso model.

In order to estimate the model parameters (i.e., β_0, β_j, γ and λ), this model suggests the least-angle regression (LARS) algorithm which infers the parameters via the least-squares approach with the same computational cost of the Lasso regression if there is no collinearity problem. But if this problem occurs, ridge regression can be an alternative method of adaptive Lasso to estimate coefficients and compute weights. Under this condition, the optimal values of γ and λ are found by two-dimensional cross validation method [105], where the optimal pair of γ and λ can be selected from a grid of values.

On the other side, if the performance of the adaptive Lasso and Lasso are compared

regarding their accuracies in variable selection and prediction, Huang et al. (2008) [50] show that the adaptive Lasso produces smaller (more sparse) models with better predictive performance based on the mean squared error. Fan et al. (2009) [25] implement it in a network modeling to estimate the sparse precision matrix and from their Monte Carlo simulation and real data analyses, it is observed that the adaptive Lasso generates more simpler models than Lasso and outperforms Lasso in terms of the entropy and quadratic loss functions. Moreover, Lasso is applicable for large size problems and is computationally fast. Thus, it is often used to obtain graphical models. However the Lasso penalty generates bias. On the other hand, the adaptive Lasso can overcome this challenge.

2.1.5 Fused Lasso

The fused Lasso model [88] is an extended version of Lasso which takes into account the order of the variables used in the regressions. Such cases can be observed in protein mass spectroscopy and microarray types of data.

In this model, the L_1-penalty is performed on both the coefficients and their successive differences. Hereby, it chooses coefficients by minimizing the residual sum of squares subject to the two different constraints. One of them is on the sum of the absolute values of the coefficients and the other constraint is on the sum of the absolute successive differences of the coefficients as represented below:

$$\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda_1\sum_{j=1}^{p}|\beta_j|+\lambda_2\sum_{j=2}^{p}|\beta_j-\beta_{j-1}|, \tag{2.8}$$

where λ_1 and λ_2 are non-negative regularization, also called tuning or penalty parameters and β_j denotes the regression coefficients for the jth covariate based on n samples, i.e., x_{ij}, and response y_i. In Equation (2.8), the first constraint causes the sparsity in the coefficients and the second one regulates the sparsity in their differences, and encourages adjacent coefficients to have the same values meaning that it creates a fusion between coefficients. In the fused Lasso regression, different from Lasso, the sparsity is applied to the number of sequences of identical non-zero coefficients, rather than the number of non-zero coefficients.

On the other side, in inference of all model parameters, the expression in Equation (2.8) is though as an optimization problem and it is converted as a quadratic function subject to several linear constraints to solve it via the quadratic programming (QP). Hence, the tuning parameters in Equation (2.8) are inferred by two constraints as a quadratic programming problem. Actually, since this equation is strictly convex, a global optimal solution exists. But due to the nondifferentiability of this function, the detection of a solution is computationally demanding. Tibshirani et al. (2005) [88] suggest two approaches to solve this challenge. The first approach is to perform the two-phase active set algorithm SQOPT [34]. This algorithm is an iterative procedure which contains two phases. In order to transform the nondifferentiable function into a smooth function, SQOPT generates additional variables as well as constraints and introduces lower and upper bounds for all variables and constraints. In the calculation, it is assumed that the m components of Ax have lower and upper bounds which are called general constraints of the problem as shown in Equation (2.9). Thus, the SQOPT algorithm produces a set of slack variables s in order to convert these constraints to equalities. So the initial problem can be defined with the following form:

$$\text{minimize}_{x,\,s}\; q(x) \text{ subject to } Ax - s = 0,\; l \leq \begin{pmatrix} x \\ s \end{pmatrix} \leq u, \tag{2.9}$$

where $q(x)$ refers to the objective function. In order to solve the objective function in Equation (2.9), the active set method is used to solve this QP problem. This iterative method includes two phases, namely, feasibility and optimality phases. In the first phase, a feasible point is found by minimizing the sum of infeasibilities. In the second phase, the objection function is minimized by using a sequence of iterations within the feasible region. In general, this algorithm is suitable for small and medium sizes problems as it is computationally challenging with respect to time and space.

On the other hand, the second approach solves the fusion problem by firstly transforming the covariates X to $Z = XL^{-1}$ with $Q = L\beta$, and then applying the LAR procedure and finally transforming it back. The fusion is archived by moving in the direction which is defined by the LAR procedure.

In the end even though, it can unravel one of the limitations of the ordinary Lasso, its inference is fully nonparametric. Hereby, it cannot take into account the probabilistic feature of the observations and the description of the system as Gaussian graphical

model performs.

2.1.6 Elastic Net

The elastic net model [104] is another alternative regularization and variable selection method which is the combination of the Lasso and ridge regressions in the sense that it enables us to capture the highly correlated feature of the variables which generate a group. Because Lasso pays no attention to group variables if the pairwise correlation between variables is high and chooses only one variable from this group.

The elastic net technique is developed from its initial method, namely, the naive elastic net approach. After rescaled naive elastic net coefficients, the elastic net coefficients are estimated. In the naive elastic net, we infer the coefficients by minimizing the following equation after a location and scale transformation to response and predictors, respectively,

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=2}^{p}\beta_j^2, \tag{2.10}$$

where λ_1 and λ_2 are non-negative tuning parameters as used previously. Moreover, β_j, x_{ij} and y_i stand for the regression coefficients, associated covariates and response, in order as described beforehand. Here, when we set α to $\alpha = \lambda_2/(\lambda_1+\lambda_2)$, Equation (2.10) turns to the following optimization problem:

$$\hat{\beta} = \arg\min_{\beta}|y - X\beta|^2 \text{ subject to } (1-\alpha)\sum_{j=1}^{p}|\beta_j| + \alpha\sum_{j=2}^{p}\beta_j^2 \leq t. \tag{2.11}$$

The constraint in Equation (2.11) is called the elastic net penalty which is a convex combination of the Lasso and the ridge penalty. When $\alpha = 1$, the naive elastic net turns to the ridge regression, and when $\alpha = 0$, it turns to Lasso.

The naive elastic net has the two-stage procedure. In the initial stage, the ridge regression coefficients are found for each fixed λ_2. Then in the second stage, the Lasso-type shrinkage is applied. This step can cause double shrinkage resulting in an extra bias with respect to the pure Lasso or ridge regression. Hereby, the elastic net method overcomes this challenge by using the rescaled coefficients of the naive elastic net. In the naive elastic net, the regression coefficients β are estimated by using the following

equation:

$$\hat{\beta}^* = \arg\min_{\beta^*} |y^* - X^*\beta^*|^2 + \frac{\lambda_1}{\sqrt{(1+\lambda_2)}} \sum_{j=1}^{p} |\beta_j^*|. \tag{2.12}$$

On the other hand, the elastic net infers the model parameters by adjusting the estimates of the naive method as defined below:

$$\hat{\beta}(\text{elastic net}) = \sqrt{(1+\lambda_2)}\hat{\beta}^*. \tag{2.13}$$

The relation between the elastic net and the naive method can be also shown by

$$\hat{\beta}(\text{elastic net}) = (1+\lambda_2)\hat{\beta}(\text{naive elastic net}). \tag{2.14}$$

In order to solve the elastic net, Zou and Hastie (2005) [104] propose an algorithm, called LARS-EN. This algorithm is based on the LARS algorithm [23]. To obtain the whole elastic net solution path, LARS is applied for each fixed λ_2, and the fits of the elastic net are sequentially updated.

When the number of predictors p is much more larger than the number of observations n, i.e., $p \gg n$, the LARS algorithm can be computationally slow. On the other side, we know that X^* in Equation (2.12) has a sparse structure under $p \gg n$. The LARS-EN algorithm uses this feature to faster the speed of the calculations. At the kth step of the LARS algorithm, the inverse of the matrix $G_{A_k} = X_{A_k}^{*T} X_{A_k}^*$ is computed when A_k is the active variable set [23]. Hereby, the LARS-EN algorithm computes this inverse by using the sparse Cholesky factorization. This dimension reduction concludes with a computational gain. Furthermore, in each LARS-EN step, just the active variable set and non-zero coefficients are saved. Moreover, when $p \gg n$, the optimal results for the elastic net fit can be obtained at an early stage without completing the total number of iterations in the algorithm. Hence, LARS-EN becomes an efficient algorithm on comparison with LARS under this model.

In the elastic net approach, we need to infer two tuning parameters as presented in Equation (2.10). This increase in the number of parameter can be though as the major drawback of this method. In the calculation, these parameters can be estimated by the cross validation (CV) method as similarly performed in shrinkage approaches. After choosing a grid of values for λ_2, whole solution path can be obtained by the LARS-

EN algorithm for each λ_2. Then, to estimate λ_1, the CV method can be applied. The optimal λ_2 is selected as the entry having the smallest CV error.

On the other side, as the LARS-EN algorithm estimates only the non-zero coefficients, it controls the Type II error. However, it does not deal with false positive rate that controls the Type I error and this one-side control can be another disadvantage of this model.

On the other hand, for the advantage of the elastic net method we can include the grouping effect in our model different than the Lasso approach. In this situation which is typically observed under 'large p, small n' problem, the grouped variables are putted into the model together and the regression coefficients of these variables are disposed to be equal. If the penalty function is strictly convex, these variables have identical coefficients. If $\lambda_2 > 0$, then the elastic net penalty satisfies this property.

Because of the grouping effect property, the elastic net technique can be also performed as a classification method like the principle component analysis or clustering methods. Especially, this method can be applied to get an automatic gene selection with microarray data [104].

2.2 Gaussian Graphical Model

The Gaussian Graphical Model (GGM) [95] is a widely used undirected graphical model whose states are described as the multivariate normal distribution. In GGM, the structure of the graph is constructed by using the conditional independencies of variables under the normality assumption. Therefore, GGM makes the assumption that the state vector Y ($Y = (Y^{(1)}, \ldots, Y^{(p)})^T$ for a system with p nodes, i.e., genes or other environmental items) has a multivariate Gaussian (or normal) distribution [95] via

$$Y \sim N(\mu, \Sigma). \tag{2.15}$$

Here, $\mu = (\mu_1, \mu_2, \ldots, \mu_p)^T$ denotes the mean, Σ shows the $(p \times p)$-dimensional variance-covariance matrix whose entries σ_{ij} present the covariances of $Y^{(i)}$ and $Y^{(j)}$ $(i, j = 1, 2, \ldots, p)$.

In GGM, the conditional independencies of two nodes are indicated with the absence of an edge between these nodes in an undirected graph. Under the normality assumption, the zero covariance and thus zero precision, which is the inverse of variance-covariance matrix, i.e., $\Theta = \Sigma^{-1}$, implies the linear independency between any pair of species, i.e., nodes.

Additionally, the precision matrix is composed of partial covariances. This means that the inverse of the partial variances constitutes the diagonal entries of the precision matrix, such as $\theta_{ii} = 1/var(Y^{(i)} \mid rest)$, and that minus the partial correlation constitutes the scaled off-diagonal entries:

$$\rho_{ij} = \frac{-\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}, \tag{2.16}$$

where ρ_{ij} represents the partial correlation of $Y^{(i)}$ and $Y^{(j)}$ when all the other variables are given.

In GGM, a regression model is build for each node against all remaining nodes as described below [87]:

$$Y^{(p)} = \beta Y^{(-p)} + \varepsilon, \tag{2.17}$$

in which $Y^{(p)}$ represents the state of the pth node, $Y^{(-p)}$ denotes the states of all nodes except the pth node, β shows the regression coefficients, and ε is the independent and normally distributed error term. The conditionally dependent structure is identified by β in Equation (2.17) due to the fact that we can present $Y^{(p)}$ and $Y^{(j)}$ $(j = 1, 2, \ldots, p-1)$ as the two conditionally independent terms if the associated β_j equals zero under the normality.

Hence, the conditionally dependent structure between nodes in the system is identified by the regression coefficient via in Equation (2.18). In this expression, $Y^{(p)}$ and $Y^{(j)}$ $(j = 1, \ldots, p-1)$ are conditionally independent when $\beta_j = 0$ [98]:

$$\beta = -\theta_{-p,p}/\theta_{p,p}. \tag{2.18}$$

In Equation (2.18), the estimated strength of the interaction between two nodes in a system is explained by the associated entries of the precision matrix.

The networks that are generated by using partial correlations are called the gene association network. Since the genomic data generally have a large number of variables,

p, and few samples, n, i.e., $n \ll p$, the application of standard covariance and correlation is not appropriate. Because in small n, large p data, the sample covariance estimator S turns a singular matrix, i.e., non-invertible, because of the large number of zero eigenvalues. Thus, the estimation of positive definited and well-conditioned covariance matrix is a crucial problem.

2.2.1 Maximum Likelihood Approach

The major aim of the GGM approach, as most of the statistical questions of interest, is to estimate the model parameters and, herewith, define a structure for the selected biological networks for us. A network can be inferred by maximizing the likelihood of observed data. As we mentioned before, in GGM, the precision matrix $\mathbf{\Theta}$ is estimated by the inverse of the sample covariance matrix which maximizes the log-likelihood function since $\mathbf{\Theta}$ captures the information of both the direction and the strength of interactions between genes in the system. Then, the partial correlations can be obtained from this estimated matrix. Finally, to decide whether the partial correlations are significantly different from zero, statistical tests can be applied. There are different alternates for this purpose. When the true partial correlation is zero, the estimated partial correlation is distributed as follows:

$$f(r,k) = (1-r^2)^{(k-3)/2} \frac{\Gamma(k/2)}{\sqrt{\pi}\Gamma((k-1)/2)}, \tag{2.19}$$

where $k = n - p - 1$ denotes the degrees of freedom, n is the sample size, and p displays the number of variables. Moreover, $\Gamma(\cdot)$ refers to the Gamma function. This distribution can be used to test the significance of partial correlations that can be performed via different approaches such as the likelihood ratio or Wald statistics.

The z-transformation given below is another alternative approach to check the validity of the partial correlation [98]:

$$z(r) = \frac{1}{2}\sqrt{n-p-1}\ln(\frac{1+r}{1-r}). \tag{2.20}$$

Similarly, the following likelihood ratio test can be applicable:

$$LR(r) = -n\log(1-r^2) \tag{2.21}$$

in which $LR(r)$ has an asymptotic χ_1^2 distribution under the null hypothesis of zero partial correlation.

In general, although the maximum likelihood estimation (MLE) method is successful when the full data are available and $\mathbf{\Theta}$ is nonsingular, it cannot be applicable for non-invertible $\mathbf{\Theta}$ [95].

2.2.2 Graphical Lasso

In order to estimate a sparse and symmetric precision matrix Θ, the L_1-penalty can be applied on the entries of the precision matrix, rather than the regression coefficients [31]. According to the Lagrangian dual form, the penalized likelihood optimization is given by

$$\text{maximize}_{\Theta} \left[\log |\Theta| - \text{Trace}(S\Theta) - \lambda \|\Theta\|_1\right] \tag{2.22}$$

in which λ denotes the non-negative Lagrange multiplier and S represents the sample covariance matrix. Hereby, in the detection of the optimal model, the sparsity of the precision matrix increases when λ increases. In addition, the optimal solution satisfies the requirement of symmetry.

When the dimensionality is too high that means the number of species p is extremely larger than the number of observations n per species, the graphical lasso becomes computationally inefficient [103]. Witten et al. (2011) [99] suggest to write the estimated inverse covariance matrix as the block diagonal form in order to speed up the computation. Because they can describe a standard graphical lasso in each block separately. But necessary and sufficient condition is required so that the estimated inverse covariance matrix can be block diagonal. According to Karush-Kuhn-Tucher conditions [99], the condition which maximizes Equation (2.22), is satisfied by the following equality:

$$\Theta^{-1} - S - \lambda\Gamma(\Theta) = 0, \tag{2.23}$$

in which $\Gamma(x)$ denotes the subgradient of $|x|$. That means, if $\Theta_{ij} > 0$, $\Gamma(\Theta_{ij})$ equals 1. If $\Theta_{ij} < 0$, $\Gamma(\Theta_{ij})$ sets to -1, and if $\Theta_{ij} = 0$, then the value of $\Gamma(\Theta_{ij})$ lies from -1 to 1.

If the inequality $|S_{ii'}| \leq \lambda$ is satisfied for all $i \in C_k$, $i' \in C_{k'}$, $k \neq k'$, where $C_1, C_2, ..., C_k$ represent a partition of the p features, the solution of the graphical Lasso problem becomes the block diagonal matrix with k blocks such that

$$\tilde{\Theta} = \begin{pmatrix} \Theta_1 & 0 & \dots & 0 \\ 0 & \Theta_2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \dots & & \Theta_k \end{pmatrix}. \tag{2.24}$$

According to the results of the simulation study [99], the algorithms which use the blocking idea are computationally efficient. Especially, if the value of tuning parameter λ increases, the number of nodes that are fully unconnected from all other nodes increases. So, the speed of the algorithm also raises.

2.2.3 Pathwise Coordinate Descent

Coordinate descent is a simple optimization method which optimizes the target function with respect to each parameter separately while keeping all the remaining parameters fixed [94]. Then iteratively all parameters are optimized until the convergence is reached. In each iteration, the fixed parameters take the values from previous iteration and the function is minimized over the just one coordinate. Through this basic idea, a complex optimization problem turns to a 1-dimensional optimization. When the target function f is convex and differentiable, a global minimum can be found by minimizing each coordinate separately. However, if the function is convex but not differentiable, the global minimum cannot reach by coordinate descent. On the other hand, the coordinate descent algorithm also converges to a solution when the function f has the following form:

$$f(x) = g(x) + \sum_{i=1}^{p} h_i(x_i). \tag{2.25}$$

Here, g is a convex and differentiable function, and the following nonsmooth part is separable where each h_i is also convex. The coordinate descent method is widely used for problems like Lasso regression, fused Lasso, graphical Lasso, regression

with non-convex penalties and so on. These types of penalized likelihood equations consist of differentiable loss function and separable penalty function.

The pathwise coordinate descent for the Lasso estimates the regression coefficients, $\hat{\beta}$, repeatedly on a grid of λ values [30]. On a grid of decreasing λ values, $\lambda_0 > \lambda_1 > ... > \lambda_k$, λ_0 equals λ_{max} ($\lambda_0 = \lambda_{max}$) and λ_k denotes λ_{min} ($\lambda_k = \lambda_{min}$). Here, λ_{max} represents the smallest value for which all coefficients are zero. Under the orthogonal design, λ_{max} can be computed by using the following equation:

$$\lambda_{max} = max_j |(x_j^T x_j)^{-1} x_j^T y|. \tag{2.26}$$

Additionally, if the design matrix is full rank, λ_{min} can be zero. Otherwise, $\lambda_{min} = \epsilon\lambda_{max}$ for small ϵ such as $\epsilon = 10^{-4}$. By using this strategy, the algorithm takes the advantage of sparsity. Starting with large value for λ corresponds to very sparse model. That means, λ_0 leading to $\hat{\beta}$ has the solution 0 or close to 0. Moreover, at each current step, the previous value of λ is used as the initial value. This procedure is called the warm start.

The steps of algorithm for pathwise coordinate descent for Lasso mainly contain two parts, namely, outer loop and inner loop. The outer loop corresponds to pathwise strategy, whereas, the inner loop is connected with the strategy of the active set.

1. Outer loop
    1.1. Compute a sequence of decreasing tuning parameter λ, $\lambda_0 > \lambda_1 > ... > \lambda_k$.
    1.2. Use solution from the previous stage for initialization (warm start).
2. Inner loop
    2.1. Perform a coordinate cycle and record active set of coefficients which are nonzero.
    2.2. Cycle over coefficients in active set until convergence.

For the following Lasso function

$$\min_{\beta} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.27}$$

one coordinate cycle can be listed as below:

(a) Compute the partial residuals, $r_{ij} = y_i - \sum_{k \neq j} x_{ij}\beta_k$.

(b) Compute the least squares coefficients of these residuals on jth predictor, $\beta_j^* = \frac{1}{N}\sum_{i=1}^{N} x_{ij}r_{ij}$.

(c) Update β_j by using soft thresholding, $\hat{\beta}_j = S(\beta_j^*, \lambda)$.

Here, the soft thresholding operator $S(.)$ can be defined as follows:

$$S(\beta_j^*, \lambda) = sign(\beta_j^*)(|\beta_j^*| - \lambda)_+ = \begin{cases} \beta_j^* - \lambda, & \text{if } \beta_j^* > \lambda, \\ 0, & \text{if } |\beta_j^*| \leq \lambda, \\ \beta_j^* + \lambda, & \text{if } \beta_j^* < -\lambda. \end{cases} \tag{2.28}$$

Although there are many approaches for penalized regression in methodological literature, GGM is the most popular method to determine the conditional relationships between variables [46, 11]. So, graphical Lasso is the baseline method for network inference. Most recent studies are extensions of glasso method such as Unified Graphical Lasso [61], Contrasting Graphical Lasso [61], Time-Varying Graphical Lasso [43], Robust Graphical Lasso [11], Robust Sparse GGM [46] and Bayesian estimation of GGM [97]. Since the glasso is used as a fundamental method in comparison studies, we also compare our proposed method with glasso.

2.3 Nonparametric Regression Methods

2.3.1 Multivariate Adaptive Regression Splines

The Multivariate Adaptive Regression Splines (MARS) is a well-know nonparametric regression approach that is suitable for high-dimensional and correlated data under nonlinearity [29]. It is a data-driven model and used for approximating nonlinearity within the data. This method has a special adaptive procedure in the sense that it can reduce the complexity in nonlinear functions by constructing the linear models. It is in essence with a kind of multivariate regression methods that produces a hierarchical model by using a set of basis functions (BF) and stepwise selection [6]. Here,

there is no assumption between dependent and independent variables. Moreover, the model consists of a two-stage procedure, called the forward stage and the backward stage. In the forward stage, the modeling starts with an intercept term which is the mean of the response values. Then, the basis functions are iteratively added to the model. These basis functions are automatically selected from observed data with a stepwise procedure [6], and the end result is the largest model that includes many basis functions. However, this complex model could have an overfitting problem. Thus, the backward stage is applied to reduce the complexity by removing basis functions that cause the smallest increase in the residual squared error. This stage prevents the overfitting problem.

Therefore, the MARS model can be described as:

$$f(y) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(y), \tag{2.29}$$

in which β_0 denotes the intercept term and β_m represents a regression coefficient that is estimated by minimizing the residual sum of squares. M denotes the number of basis functions in the current model and $h_m(y)$ shows the piecewise linear basis functions (linear splines) which are described as follows:

$$(y-\tau)_+ = \begin{cases} y-\tau, & \text{if } y > \tau, \\ 0, & \text{otherwise,} \end{cases} \quad (\tau - y)_+ = \begin{cases} \tau - y, & \text{if } y < \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{2.30}$$

Here, the $(+)$ sign indicates the positive part of the expression. These piecewise linear functions are connected at some values τ, called the knots. The functions $\big(y-\tau\big)_+$ and $\big(\tau - y\big)_+$ are able to smooth a curve by pieces of linear expressions as shown in Figure 2.1. For each independent random variable y_j, the reflected pairs are obtained and this set of basis functions is represented as given below:

$$C = \{(y_j - \tau)_+, (\tau - y_j)_+ \mid \tau \in \{y_{1j}, y_{2j}, \ldots, y_{nj}\}, j = 1, 2, \ldots, p\}, \tag{2.31}$$

where p is the number of independent variables.

At the end of the forward stage, a large model is obtained. Here, the backward stage can be applied to estimate the best model with a number λ of terms. The optimal λ is chosen by using the generalized cross-validation (GCV) value defined as

$$\text{GCV}(\lambda) = \frac{\sum_{i=1}^{N}(y_i - \hat{f}_\lambda(y_i))^2}{(1-(r+cK)/N)^2} \tag{2.32}$$

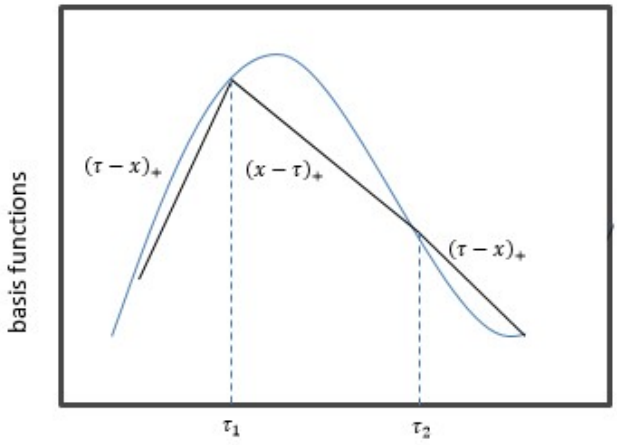


Figure 2.1: Simple representation of the smoothing method for the curvature structure via basis functions of MARS.

in which N denotes the number of observations, r indicates the number of linearly independent basis functions and K denotes the number of knot points used in the forward stage. Additionally, c is the cost for the optimization of the basis function and it is generally set to 3. However, if the model is restricted to obtain an additive model, it is taken as 2. Furthermore, the numerator of GCV equals the residual sum of squares. Finally, $\hat{f}_\lambda(y)$ implies the estimated model with λ number of terms [44]. Hence, the model which minimizes GCV is chosen as a final model in the backward stage.

2.3.2 Conic Multivariate Adaptive Regression Splines

The Conic Multivariate Adaptive Regression Splines (CMARS) is a modified version of MARS. Here, C not only stands for "Conic", but also for, "Convex" and "Continuous". This method proposes to apply a penalized residual sum of squares (PRSS) for MARS as a ridge regression, also known as the Tikhonov regularization (TR), by eliminating the backward stepwise algorithm [101]. After obtaining linear combinations of the special basis functions of MARS in the forward step, PRSS is used for the parameter estimation. Indeed, this is an optimization problem that results from the trade-off between accuracy and complexity. In this context, the accuracy means a (ideally small) sum of error squares, and the complexity refers to the first and

second-order derivatives squared of the basis functions with their parametric multipliers. Thereby, the trade-off is detected by the construction of penalty parameters, and this regularization problem can be solved by the conic quadratic programming. For the MARS model, PRSS has the following form [86]:

$$\text{PRSS} = \sum_{i=1}^{N}(y_i - f(\bar{x}_i))^2 + \sum_{m=1}^{M_{max}} \lambda_m \sum_{|\alpha|=1}^{2} \sum_{\substack{r<s \\ r,s\in V(m)}} \int_{Q^m} \theta_m^2 [D_{r,s}^{\alpha} h_m(t^m)]^2 dt^m. \quad (2.33)$$

Here, $V(m)$ is the variable set associated with the mth basis function h_m, t^m denotes the vector of variables which contribute to the mth basis function and thus shows the integration variable of the complexity term, and finally, Q^m represents a sufficiently large parallelpipe where the integration takes place, i.e., where the input data are located. Additionally, the derivative $D_{r,s}^{\alpha} h_m(t^m)$ is defined by

$$D_{r,s}^{\alpha} h_m(t^m) = \frac{\partial^{\alpha} h_m}{\partial_1^{\alpha} t_r^m \partial_2^{\alpha} t_s^m}(t^m) \quad (2.34)$$

for $\alpha = (\alpha_1, \alpha_2)^T$ where $\alpha_1, \alpha_2 \in \{0, 1\}$. In this way, PRSS can control both the complexity and the accuracy of the model in a balanced way, resulting in conic quadratic programming (CQP) for the parameter estimation. Indeed, the reorganization of Equation (2.33) converts the PRSS calculation into a TR form:

$$\text{PRSS} \approx \|y - h(\bar{d})\theta\|_2^2 + \lambda \|L\theta\|_2^2. \quad (2.35)$$

In Equation (2.35), L is a diagonal $(M_{max}+1)\times(M_{max}+1)$-matrix and θ refers to an $((M_{max}+1)\times 1)$-dimensional vector of parameters. Finally, $\|\cdot\|_2^2$ denotes the L_2-norm of the given term. Here, to construct a TR problem with a single parameter, a uniform penalization is performed by taking the same λ for each derivative term, rather than λ_m [86]. Accordingly, Equation (2.33) has two objectives as a linear combination of $\|y - h(\bar{d})\theta\|_2^2$ and $\|L\theta\|_2^2$. Hence, the solution is the value which minimizes both objective functions by a compromise that can be found via continuous optimization techniques, namely, CPQ, also known as conic quadratic programming as stated before. As a result, with an appropriate choice of a bound M, the optimization problem is stated as follows:

$$\text{minimize } \|h(\bar{d})\theta - y\theta\|_2^2 \text{ subject to } \|L\theta\|_2^2 \leq M. \quad (2.36)$$

To select the penalty parameter λ in PRSS, an efficiency curve, called the L-curve, can be used. The L-curve displays how the regularized solution changes as the parameter

changes to plots the norm of the regularized solution versus the number of parameters, i.e., the "steepness" and "curvatures" in the basis functions. Furthermore, because of the large range of norms, the curve is plotted in the double logarithmic scale. The corner of this curve is interpreted as a good balance between the minimization of the sizes and the corresponding penalty parameters.

2.3.3 Proposed Method - LCMARS

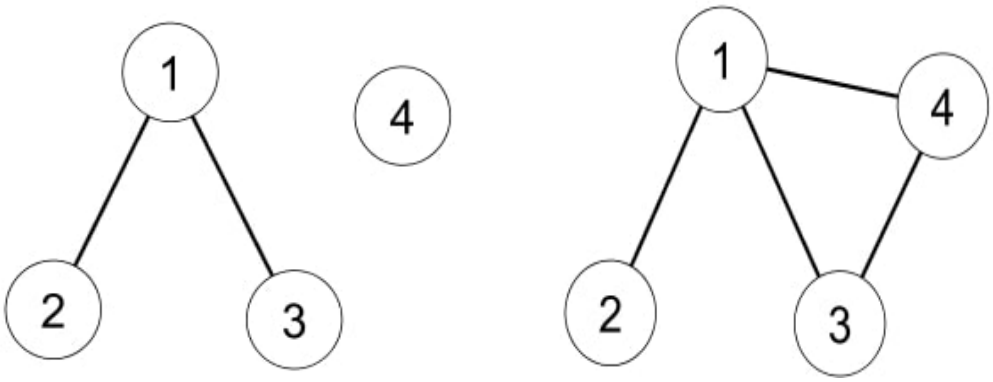
To construct loop-based CMARS (LCMARS), we use the same logic as given in Equation (2.17). Hereby, to define links of a specified gene (or environmental issue), we regress each node in the graph against all the remaining nodes and estimate the regression coefficients to detect the relation between corresponding nodes. Accordingly, in this regression model, since each node represents a gene, the response and the explanatory variables of the model are composed of the expression levels of the genes. For instance, assume that there is a small system with four nodes, namely, y_1, y_2, y_3, and y_4. For the first node, y_1, we can write two separate regression equations for GGM and LCMARS, respectively, as follows:

$$y_1 = 3y_2 + 2y_3, \tag{2.37}$$

$$y_1 = y_2 + 3y_3 + 2y_3y_4. \tag{2.38}$$

For Equation (2.37), node 1, i.e., with expression level y_1, as the response gene, has connections with node 2, and node 3, i.e., y_2 and y_3, for GGM as the explanatory genes. In contrast, Equation (2.38) estimated by LCMARS with the second-order interaction indicates that node 1 as the response gene has interactions with node 2 and node 3 as well as node 4 as the explanatory genes. Furthermore, there is an interaction between node 3 and node 4 due to the significance of y_3y_4 which denotes the feed-forward loop. Indeed, the description of the second-order interaction is not possible in GGM since both models can merely capture the linear relationships between genes. The networks constructed by Equation (2.37) and Equation (2.38) can be also seen in Figure 3.2, for the visual inspection.

In this study, we construct the loop-based CMARS model similar to a regression model as given in Equation (2.17), and call as loop-based CMARS, shortly LCMARS.



(a) (b)

Figure 2.2: (a) The construction of network based on Equation (2.37) and (b) the construction of network based on Equation (2.38)

Thereby, we regress each node in the graph against all the remaining nodes iteratively and estimate the regression coefficients to detect the relation between corresponding nodes. In this calculation, the steps of LCMARS method can be listed as below:

1. Construct p number of models which include both main and the second-order interaction effects by using the forward stage of MARS.

2. Compute PRSS for each model without moving the backward stage of MARS.

3. Determine the best model which has the optimal penalty term.

4. For each selected LCMARS model, assign 1 to associated entry of the adjacency matrix if there is a relation between the response and explanatory variables.

5. Apply the AND rule to build symmetric adjacency matrix. Here, the AND rule means that we assign 1 for each pair of nodes (i, j) in the precision matrix if both (i, j) and (j, i) entries in Θ indicate the edges.

6. Make comparison between the estimated adjacency matrix and the true binary precision matrix to compute accuracy measures, which we choose as recall, F-measure, Jaccard index and accuracy value.

2.4 Outlier Detection Methods

Outlier detection is a crucial problem in many fields. Although there are too many outlier detection methods in the literature, only a few methods suitable for dependent, sparse and high-dimensional data structure. In this study, we aim to investigate whether using an outlier detection method as a pre-processing step before modeling can improve the accuracy of fitted models and which method is the most appropriate for the protein-protein interaction data.

An outlier is an observation that is significantly different from other observations [45]. In the literature, it is also known as abnormalities, discordants, deviants and anomalies [1]. Due to its importance in both the data analyses and modeling the observations, the outlier detection is a crucial problem in many research fields. There are basically two main purposes of the outlier identification. In some cases, the identification of outliers can be the primary goal since it provides new discoveries about the hidden aspects of the data [40]. In other cases, the identification of outliers can be used as a pre-processing step before fitting a statistical model and many methods have been proposed for this purpose. These methods can be divided into two parts, so called, univariate and multivariate methods. Dixon's Q test [22], Grubbs test [38], Z-scores [81] and box plot [90] are the widely known univariate outlier detection methods. The Dixon's Q test and the Grubbs test are very simple methods which enable to find at most 2 outliers. They are applicable under normally distributed data. On the other hand, the box plot is the most well-known nonparametric and robust graphical method which is appropriate to identify outliers. However, the outlier detection with graphical tools is not suitable for more than three dimensions. Therefore, it is necessary to apply a multivariate analysis which can consider the interactions among several variables to the multivariate data. The multivariate outlier detection methods can be basically divided into two parts: (i) Distance-based approaches and (ii) projection based approaches. Mahalanobis distance (MD) and the robust distance (RD) are the well-known distance measures [41]. Moreover, the blocked adaptive computationally efficient outlier nominators (BACON) is an another distance based on a robust multivariate method [14]. Unfortunately, they are not applicable for data structure where the size of the variables are much more than the observations ($p \gg n$)

due to the singularity problem in the estimation of the covariance matrix. In most of genomic datasets, there may be thousands of genes or proteins measured on few samples. Hence, $p \gg n$ situation is commonly observed in biological data. Here, the projection pursuit methods are appeared to be aware. Since they are not restricted with distributional assumptions and are applicable for several data structure, particularly, where the size of the variables is much more than the observations $(p \gg n)$, they are the most appropriate choice for high-dimensional data. A well-known projection pursuit method is the principal component analysis (PCA). Additionally, there are several outlier detection approaches based on PCA such as robust PCA and spherical PCA methods [27, 62].

In the following subsection, we present the brief mathematical details of the well-known outlier detection approaches under univariate and multivariate methods.

2.4.1 Univariate Methods

In univariate data, graphical methods, such as scatter plot and box plot, and statistical tests, such as Dixon's Q test [22], Grubbs test [38] and Z-scores [81] can be performed to detect outlier observations. Accordingly, by using graphs, the observations which show different patterns visually from the rest of data, can be considered as outliers.

In this branch, the box plot [90] is the most well-known graphical method which is appropriate to identify outliers for univariate data. It is a nonparametric and robust approach since it is based on quartiles. In this method, initially the box is constructed with the 1st (Q_1), 3rd (Q_3) quartiles and a line which represents the median. Then, the interquartile range (IQR) is calculated within the range of 1st and 3rd quartiles indicatively the length of the box. In the decision of outliers Tukey (1977) [90] uses measurements which are smaller than $(Q_1 - 1.5 \times \text{IQR})$ or greater than $(Q_3 + 1.5 \times \text{IQR})$. Even though the graphical approaches, in general, are the most preferable methods to investigate outliers due to their friendly usage and simplicity, they cannot be applicable for more than two-dimensional datasets.

On the other hand, among statistical tests, the Dixon's Q approach is a very simple test which enables to find a single outlier. Hence, this test is suitable for data containing

few number of observations and it is solely applicable under normally distributed data. Furthermore, the test cannot be implemented more than once in a dataset since it can identify only one outlier in a given set.

The Grubbs test is an another outlier detection method for the univariate data. Similar to the Dixon's Q test, it applies the normality assumption and finds the value that is the furthest from the sample mean as an outlier. Therefore, the tested data are taken as the minimum and the maximum values. If the test is applied as a one-sided test, a single outlier can be identified by choosing either the minimum or the maximum value of the dataset. On the other side, the two-sided Grubbs test is conducted both the minimum and the maximum values are assigned as outliers.

The Z-scores, also called the standardized value, is the most common and general method among alternatives and represents the distance between the observations and their mean in units of the standard deviation by computing the following expression for each observation x_i:

$$z_i = \frac{x_i - \bar{x}}{s} \tag{2.39}$$

in which $\bar{x}$ is the sample mean and s denotes the standard deviation of sample. In the testing procedure, the Z-scores greater than 3 or less than -3 are considered to be an outlier. However, the major drawback of this method is that it is based on the mean and the standard deviation which are not robust measures against outliers.

On the other hand, there is also an extended version of Z-score called as the modified Z-score method, proposed by Iglewicz and Hoaglin (1993) [52]. This approach uses the median and the median absolute deviation (MAD) instead of the mean and the standard deviation. The median and MAD are robust measures of the central tendency and the dispersion, respectively. Moreover, if the absolute value of this score is greater than 3.5, the corresponding observation is assigned as an outlier.

2.4.2 Multivariate Methods

As mentioned earlier, the outlier detection with graphical tools is not suitable for more than three dimensions. Furthermore, under multivariate dimensional data, analyzing each dimension separately is not an appropriate way since some observations may

be outliers merely in the multivariate space. Therefore, it is necessary to apply a multivariate analysis which can consider the interactions among several variables. Moreover, the datasets with multiple outliers can suffer from masking and swamping effects. In the masking effect, one outlier masks a second outlier in such a way that an observation can be an outlier by itself, however cannot be observed in the presence of the first outlier. Thus, other outlier comes out if and only if the first outlier is deleted. On the other hand, in the swamping effect, one outlier swamps a second observation under this condition, when the first outlier is deleted, the second observation becomes a non-outlying observation [12]. Because of these effects, the identification of outliers in the multivariate data becomes a challenging problem.

The multivariate outlier detection methods can be basically divided into two parts as presented previously. These are (i) Distance-based approaches and (ii) projection-based approaches. The main idea of the distance based methods is to detect outliers by computing a measure of how far each point is from the center of the data. A well-known distance measure in this branch is the Mahalanobis distance (MD). For an observation x, MD is defined by the following way:

$$\text{MD}(x) = \sqrt{(x - \bar{x})^T S^{-1} (x - \bar{x})}, \tag{2.40}$$

where $\bar{x}$ is the sample mean and S refers to the sample covariance matrix. In this equation, $(.)^T$ and $(.)^{-1}$ denote the transpose and the inverse of the given vector or matrix, respectively. Once MDs are computed, the data point which has the biggest MD is labeled as the outlier. Moreover, the MD values under the multivariate normal data are approximately Chi-square distributed. On the other side, a drawback of this measure is that it includes a sample mean and a covariance which are not robust, resulting in affecting from outliers. Thereby, in order to obtain reliable results, MD needs to be estimated with robust measures. One of the most widely used robust estimators for the multivariate location and the scatter is the minimum covariance determinant (MCD) [77]. If the MCD estimators are used in place of MD, then the robust distance (RD) can be computed as [51]:

$$\text{RD}(x) = \sqrt{(x - \hat{\mu}_{MCD})^T \hat{\Sigma}_{MCD}^{-1} (x - \hat{\mu}_{MCD})}, \tag{2.41}$$

in which $\hat{\mu}_{MCD}$ and $\hat{\Sigma}_{MCD}$ denote the MCD estimates of location and covariance parameters, respectively.

The Blocked Adaptive Computationally Efficient Outlier Nominators (BACON) is an another distance based on a robust multivariate method. BACON is an iterative approach which uses the methods of Hadi [41, 42]. But it is computationally more efficient. Therefore, it can be applied for the datasets which include thousands of observations. The Hadi's method initially defines a clean basic subset of the data and this subset is presumed to be outlier-free. Then, the subset iteratively grows by adding one observation in each step with a forward search. This computation can be computationally very demanding for large datasets. On the other hand, in the BACON method, a potentially large number of points, called the block of points, are added to the basic subset in each iteration, rather than adding one by one. As a result, BACON ends up with very few steps without taking into account the sample size [14].

On the other hand, the main idea of the projection pursuit method is to obtain lower dimensional projections which include useful information for discovering the structure of data. Because these methods convert the data via a suitable projection to easily visualize the outliers. Moreover, they are not restricted with distributional assumptions and are applicable for several data structure, particularly, where the size of the variables is much more than the observations ($p \gg n$).

A well-known projection pursuit method is the principal component analysis (PCA). This dimension reduction approach uses a few number of principal components (PC) to represent the data. Here, the direction of orthogonal components is defined by maximizing the variance and a meaningful information can be obtained with a small number of components. In this case the remaining majority of components is accepted as noise and is not counted within the total variance [27]. If the PCA method is applied to detect outliers, it is necessary to use the robust estimator of the covariance matrix since the variance is highly sensitive to outliers [40]. Filzmoser et al. (2008) [27] propose a robust PCA (PCOut) method for this purpose. This method consists of two main parts. The first one is to detect the location outliers (e.g., mean-shift outliers) and the other part is to identify the scatter outliers (e.g., variance inflation outliers). In this computation, initially, the PCOut method robustly scales each vari-

able by using the coordinate wise median and the median absolute deviation (MAD). Then, a semi-robust PCA is obtained. These newly obtained small numbers of PCs explain at least 99% of the total variance. By this way, this method can be easily applied to the data which have larger number of variables than observations ($p \gg n$), as it solves the singularity problem of the covariance matrix. In PCOut, the absolute value of a robust kurtosis measure is calculated for each component to detect the location outliers. On the other hand, the scatter outliers are detected via the semi robust PCs.

Another projection method which is also based on the robust principle components is suggested by Locantore et al. (1999) [62]. In this method, first of all, all the observations are projected onto the boundary of a sphere (or an ellipsoid) with a unit radius. Then, the spatial median and the spatial sign covariance matrix are estimated to obtain robust estimators. Finally, the robust Mahalonobis distance can be computed for each observation. Similar to the other methods, the observations with large distances are the possible outliers. Locantore et al. (1999) [62] call this method as the spherical PCA, whereas, Filzmoser et al. (2008) [27] refer the same method as the Sign approach in their R package, called mvoutlier.

2.5 Model Selection Criteria

In this section, we explain the model selection criteria to compare the performance of suggested model with alternative methods. When we estimate the biological network, we compare this network structure with the true network in order to check the model accuracy. As we mentioned earlier, the estimated undirected network structure can be represented by a binary precision matrix. Hence, to compare the estimated and true precision matrices, we use the accuracy measures which are developed for binary classification. Accuracy measures used in this study are explained in details in the following part.

Table 2.1: Confusion matrix of the accuracy measures.

		Actual Class	
		Positive	**Negative**
Predicted Class	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

2.5.1 Description of Accuracy Measures

In the literature, there are many measures used to evaluate the accuracy of the binary classification. In general, the measurements are the functions of the following four main values which are true positive, true negative, false positive, and false negative. The *true positive* (TP) implies the number of correctly classified objects that have positive labels that means there is an edge, the *true negative* (TN) indicates the number of correctly classified objects that have negative labels which denotes no edge, the *false positive* (FP) shows the number of misclassified objects that have negative labels, and the *false negative* (FN) presents the number of misclassified objects that have positive labels. This information can constitute a confusion matrix as shown in Table 2.1, which represents the actual and the predicted classification.

In this study, we calculate *recall*, *F-measure*, *Jaccard index* and *accuracy* measures whose mathematical expressions are presented in Equations (2.42)-(2.45).

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}, \tag{2.42}$$

$$\text{F-measure} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \tag{2.43}$$

$$\text{Jaccard index} = \frac{\text{TP}}{\text{TP+FP+FN}}, \tag{2.44}$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}. \tag{2.45}$$

Here, recall describes the ratio of correctly classified genes with a positive label to the total of the positive class in the actual case. In other words, it represents the ratio of correctly estimated positive links that means connections between genes over the total number of links in the true network. Recall is also named as the sensitivitiy or

true positive rate (TPR). On the other hand, F-measure indicates the overlap between actual and predicted classes. Jaccard index can be interpreted in the same way as F-measure and it is linearly related to this measure in such a way that

$$\text{Jaccard index} = \frac{(\text{F-measure})}{2 - (\text{F-measure})}. \tag{2.46}$$

Additionally, the accuracy is the ratio of correctly classified genes in both labels to the total of all classified genes. That means it is the ratio of correctly estimated links (both positive and negative) over total links. In all these measures, 1 indicates perfect accuracies, whereas 0, implies no success at all in terms of accuracy.

In our analyses, we choose these four accuracy measures since each of them controls different perspectives of the estimated systems. For example, recall shows the fraction of real edges, i.e., links, that are correctly inferred. So, it only deals with the number of TP and does not pay attention to the number of TN. On the contrary, accuracy and F-measure evaluate the performance of methods from both sides. Therefore, they can be more informative in the assessment of the best model by considering both types of error, i.e., FP and FN simultaneously. So, F-measure and accuracy are widely used measures to evaluate the model performance [43, 61, 97]. Whereas, recall is generally used with false discovery rate to obtain ROC curve [46, 71].

CHAPTER 3

APPLICATION

In this chapter, we compare the proposed method with alternates which are described in Chapter 2. In the application part, to compare the methods we use different dimensional simulated, synthetic and real biological datasets. In Section 3.1, we check the validity of proposed method with simulated data which are generated under different scenarios. The results for different synthetic biological data are presented in Section 3.2. In Section 3.3, both the description of real datasets and their results are represented. Finally, we investigate the validity of outlier detection methods under both synthetic and real biological data in Section 3.4.

3.1 Application of Simulation Study

In the application, we perform LCMARS, LMARS and GGM methods to construct the networks. We implement LCMARS by using the main effects and the second-order interactions of the model. The significant regression coefficients of this model are directly taken from the estimation of the adjacency matrix which is used to indicate the undirected link between nodes in a network. Here, the estimated adjacency matrix can be nonsymmetric since the response variables in each regression model have different significant explanatory variables. So we apply the AND rule which means that we obtain an entry 1 when both (i, j) and (j, i) entries of the estimated adjacency matrix are 1. By using the AND rule, we infer more sparse matrices similar to the real biological networks. In the end, we compare the symmetric adjacency matrix with the true network and evaluate the different accuracy measures such as the recall, F-measure, Jaccard index and accuracy which are explained in details in the previous chapter.

Furthermore, for all schemes, datasets are simulated under various dimensions, p, listed as $40, 100$ and 200 number of genes in the system. Moreover, the number of observations for each dataset equals 20 and all the results are based on the 1000 Monte Carlo runs. Finally in all scenarios, the topology of the networks is chosen as the scale-free networks as this is the most common view of the biological networks [2].

On the other hand, we perform the graphical Lasso, also called glasso, method [31] to infer the model parameters in GGM. This method basically finds the entries of the precision matrix via the penalized likelihood function whose penalty constant is controlled by the L_1-norm. In this study, the optimal penalty constant is selected by the STARS criterion [60].

The simulated datasets are created under mainly two different scenarios such that the generated matrices come from normal and non-normal distributions. The multivariate normally distributed datasets are generated under the scale-free feature by using the huge package in the R programming language. For non-normal datasets, we use different distributions such as Student-t, log-normal with mean=0 and standard deviation=1, Cauchy with location=0 and scale=1 parameters, and mixture of log-normal and normal distributions with equal proportions and same parameters. For Student-t distribution, we use two different degrees of freedom (df), namely, df=7 and df=15, so that we can detect the results of the Student-t and the results of its more normally distributed versions, respectively. The outputs of LCMARS, LMARS and GGM methods under all scenarious are presented in Tables 3.1 - 3.5. In these tables, the best models are shown in boldface and p denotes the total number of genes in the system.

From the analyses, we observe that for normal distribution, recall and accuracy values of both methods are close to each other under all dimensions. On the other side, F-measure and Jaccard index values of LCMARS are greater than GGM. Additionally, as seen in Table 3.1, when the dimension of data increases, the differences between methods get larger.

Additionally, as observed in Table 3.2, under the Student-t distribution for all dimensions and both degrees of freedom (df=7 and df=15), F-measure, Jaccard index and accuracy values of LCMARS are higher than LMARS and GGM. Similar to the pre-

Table 3.1: Comparison of the recall, F-measure, Jaccard index and the accuracy for normally distributed data.

Method	p	Recall	F-measure	Jaccard index	Accuracy
LCMARS	40	0.3483	**0.4608**	**0.2995**	**0.9400**
	100	0.3392	0.4492	0.2896	0.9752
	200	0.3359	**0.4455**	**0.2869**	**0.9876**
GGM	40	**0.3539**	0.4335	0.2773	0.9309
	100	**0.3462**	0.3773	0.2330	0.9655
	200	0.3426	0.3264	0.1958	0.9782
LMARS	40	0.3502	0.4505	0.2910	0.9369
	100	0.3388	**0.4556**	**0.2949**	**0.9759**
	200	**0.3571**	0.1990	0.1105	0.9570

vious result, GGM gets worse when the dimension increases.

In Table 3.3, we observe that the accuracy, F-measure and the Jaccard index values of LCMARS under log-normal datasets are a bit more than GGM when the dimension of the network is small, i.e., $p = 40$. Whereas, the difference between LCMARS and GGM significantly grows for larger dimensions. Besides these, recall values of GGM are higher than LCMARS for high dimensional systems. Moreover, the accuracy, F-measure and the Jaccard index values of LMARS are close to LCMARS under small and moderate systems, i.e., $p = 40$ and 100, respectively. On the other side, the difference between LCMARS and LMARS get larger under bigger systems when we specifically compare F-measure and Jaccard index.

Finally, the results of the Cauchy distribution and mixture of log-normal with normal distributions, are similar to the previous results as they can be seen in Table 3.4 and Table 3.5, respectively. In other words, LCMARS produces higher scores than LMARS and GGM in terms of the accuracy, F-measure and the Jaccard index. Furthermore, although recall values of GGM are higher than LCMARS and LMARS, the differences get smaller when the dimension increases.

Hence, from all findings we detect that apart from small systems, LCMARS performs

Table 3.2: Comparison of the recall, F-measure, Jaccard index and the accuracy for Student-t distributed data with degrees of freedom (df) 7 and 15.

df	Method	p	Recall	F-measure	Jaccard index	Accuracy
7	**LCMARS**	40	0.3379	**0.4468**	**0.2882**	**0.9374**
		100	0.3346	**0.4425**	**0.2844**	**0.9747**
		200	0.3339	**0.4426**	**0.2846**	**0.9873**
	GGM	40	**0.3895**	0.3117	0.1853	0.8676
		100	**0.3780**	0.2064	0.1155	0.9090
		200	**0.3744**	0.1335	0.0718	0.9238
	LMARS	40	0.3422	0.4358	0.2790	0.9335
		100	0.3349	0.4403	0.2826	0.9745
		200	0.3540	0.1996	0.1109	0.9574
15	**LCMARS**	40	0.3382	**0.4468**	**0.2883**	**0.9374**
		100	0.3347	**0.4425**	**0.2844**	**0.9747**
		200	0.3341	**0.4428**	**0.2846**	**0.9873**
	GGM	40	**0.3972**	0.2971	0.1749	0.8564
		100	**0.3833**	0.1909	0.1058	0.8998
		200	**0.3792**	0.1249	0.0668	0.9179
	LMARS	40	0.3415	0.4353	0.2786	0.9335
		100	0.3351	0.4409	0.2830	0.9745
		200	0.3544	0.1997	0.1109	0.9574

better than both LMARS and GGM under all scenarios and dimensions. Additionally, when we compare the results of LCMARS and LMARS, it can be observed than although the results are close to each other under small networks, LCMARS stands out from LMARS for large networks. Therefore, we consider that the suggested approach can be a promising alternative of GGM in the construction of complex biological networks.

Table 3.3: Comparison of the recall, F-measure, Jaccard index and the accuracy for log-normally distributed data.

Method	p	Recall	F-measure	Jaccard index	Accuracy
LCMARS	40	**0.3559**	0.4615	0.3000	0.9388
	100	0.3391	**0.4534**	**0.2931**	**0.9756**
	200	0.3357	**0.4565**	**0.2961**	**0.9880**
GGM	40	0.3390	**0.5063**	**0.3390**	**0.9512**
	100	**0.4664**	0.1198	0.0637	0.7950
	200	**0.4215**	0.0851	0.0444	0.8643
LMARS	40	0.3390	0.4600	0.2990	0.9410
	100	0.3397	0.4424	0.2840	0.9745
	200	0.3586	0.1918	0.1061	0.9548

Table 3.4: Comparison of the recall, F-measure, Jaccard index and the accuracy for Cauchy distributed data.

Method	p	Recall	F-measure	Jaccard index	Accuracy
LCMARS	40	0.3487	**0.4563**	**0.2958**	**0.9388**
	100	0.3385	**0.4588**	**0.2976**	**0.9762**
	200	0.3351	**0.4694**	**0.3070**	**0.9889**
GGM	40	**0.7413**	0.1585	0.0861	0.4191
	100	**0.5532**	0.0906	0.0475	0.6689
	200	**0.4468**	0.0708	0.0367	0.8243
LMARS	40	0.3552	0.4295	0.2737	0.9303
	100	0.3395	0.4422	0.2839	0.9745
	200	0.3570	0.2038	0.1135	0.9583

3.2 Application of Synthetic Biological Data

To evaluate the performance of methods under biological data, we use different synthetic and real benchmark datasets. In this application, the synthetic data are taken from benchmark synthetic data tools such as SynTRen (Synthetic transcriptional reg-

Table 3.5: Comparison of the recall, F-measure, Jaccard index and the accuracy for mixture of log-normal and normal data.

Method	p	Recall	F-measure	Jaccard index	Accuracy
LCMARS	40	0.3471	**0.4621**	**0.3005**	**0.9405**
	100	0.3388	**0.4586**	**0.2974**	**0.9762**
	200	0.3355	**0.4593**	**0.2984**	**0.9881**
GGM	40	**0.4593**	0.2505	0.1435	0.7963
	100	**0.4112**	0.1631	0.0888	0.8739
	200	**0.3816**	0.1264	0.0675	0.9211
LMARS	40	0.3548	0.4322	0.2759	0.9312
	100	0.3403	0.4370	0.2796	0.9739
	200	0.3564	0.2047	0.1140	0.9586

ulatory networks) [17] and GeneNetWeaver (GNW) [80]. The synthetic gene expression datasets from these tools approximate the experimental data and enable us to validate different network construction methods as they present the true network model too.

Accordingly, in this study, the gene expression datasets from DREAM 4 (2009) in-silico size 10 and size 100 are used. DREAM is the International Dialogue for Reverse Engineering Assessments and Methods competition which publishes the challenges in the network inference. Hereby, the DREAM 4 in-silico size 10 challenge consists of five networks involving 10 genes. Moreover, in-silico size 100 challenge includes five networks involving 100 genes. For each network, multifactorial data are available while containing steady-state levels of variations in the network based on gene expression characteristics of two well-studied systems, namely, E.coli and S.cerevisiae. These datasets are accessible from the R package DREAM4. Hence, in our analyses, we use three datasets from in-silico size 10 and two datasets from in-silico size 100. The results of in-silico size 10 and size 100 networks are presented in Table 3.6 and Table 3.7, respectively. From the findings, it is observed that LCMARS has better scores than LMARS and GGM in terms of recall, F-measure and Jaccard index for two in-silico size 10 datasets. Moreover, the accuracy values of LMARS

are higher than other methods for these two datasets. Additionally, for the third data, LCMARS is better under all accuracy measures. On the other hand, in-silico size 100 datasets, accuracy results of LCMARS and LMARS methods are very close to each other. However, the GGM method cannot be applicable since this method cannot work when the system's elements have high correlation.

Table 3.6: Comparison of the recall, F-measure, Jaccard index and the accuracy for DREAM-4 in-silico size 10.

Data	Method	Recall	F-measure	Jaccard index	Accuracy
1	**LCMARS**	**0.647**	**0.647**	**0.478**	0.760
	LMARS	0.471	0.616	0.444	**0.800**
	GGM	0.294	0.454	0.294	0.760
2	**LCMARS**	**0.474**	**0.563**	**0.391**	0.720
	LMARS	0.368	0.538	0.368	**0.760**
	GGM	0.263	0.416	0.263	0.720
3	**LCMARS**	**0.438**	**0.583**	**0.412**	**0.800**
	LMARS	0.375	0.522	0.353	0.780
	GGM	0.312	0.476	0.312	0.780

Table 3.7: Comparison of the recall, F-measure, Jaccard index and the accuracy for DREAM-4 in-silico size 100.

Data	Method	Recall	F-measure	Jaccard index	Accuracy
1	**LCMARS**	**0.275**	0.296	0.174	0.943
	LMARS	0.248	**0.298**	**0.175**	**0.949**
2	**LCMARS**	**0.234**	**0.267**	**0.154**	0.934
	LMARS	0.207	0.258	0.148	**0.939**

Toy Data

We compare the performance of three methods with a simulated toy example of gene expression data generated by the GNW generator by using known biological interaction networks of E.coli. The toy data contain 64 samples and 64 genes and are

available in the R package grndata [10]. We present the outcomes of this dataset in Table 3.8. From the outcomes, we observe that recall, F-measure and Jaccard index of LCMARS are better than others. On the other side, the accuracy values of GGM are slightly higher than LCMARS and it is very close to the LMARS result.

Table 3.8: Comparison of the recall, F-measure, Jaccard index and the accuracy for toy data and Jak-Stat pathway.

Data	Method	Recall	F-measure	Jaccard index	Accuracy
Toy	**LCMARS**	**0.360**	**0.347**	**0.210**	0.842
	LMARS	0.234	0.342	0.207	0.895
	GGM	0.134	0.236	0.134	**0.899**
Jak-Stat	**LCMARS** (BF=10)	0.458	0.343	0.207	0.801
	LCMARS (BF=3)	0.316	0.389	0.241	0.888
	LMARS	0.278	**0.418**	**0.265**	**0.916**
	GGM	**0.626**	0.266	0.153	0.608

JAK - STAT Pathway

The final synthetic data belong to the Janus kinase (JAK) signal transducer and activator of transcription (STAT) pathway. This is an important signal transaction pathway that is activated by Type I interferons (IFNs) [100]. IFN regulates intracellular antimicrobial programmes and influences the development of innate and adaptive immune responses [53]. In this way, IFNs can control the immune system of living organisms and are used to treat the hepatitis B and C virus infections [82, 63].

Since the current real data cannot completely and realistically describe this pathway, we use a simulated dataset taken from Purutçuoğlu et al. (2017) [74]. In these data, the system is described with 38 proteins and each protein has 10 observations. Accordingly, we initially take the list of proteins, the initial numbers of their molecules and their reaction rate constants in the reaction list of the system as presented in Maiwald et al. (2010) [63]. Then, we use these initials in the long-run stochastic simulation of the system via the Gillespie algorithm [35] until all states of the system reach their steady-state activations.

Once the dataset is generated, we apply LCMARS, LMARS and GGM to estimate the network. Then, we compute the recall, F-measure, Jaccard index and the accuracy values to compare model performances. In all calculations of LCMARS, when the true networks of the systems are known, we compute the average clustering coefficient of the system [5] and take this number as the number of maximum basis functions allowed in the estimated regression model. This allows us to control the sparsity of the inferred networks. In addition, for all analyses we also set this number to 10 as the default number since the user may not always know the true structure of the network in advance. In Table 3.8, we list the accuracy scores of LCMARS when we know the true structure of the network and when we do not know this structure, respectively. From the findings under all measurements apart from the recall criterion, we see that LMARS produces more accurate outputs than LCMARS and GGM.

3.3 Application of Real Biological Data

3.3.1 Description of Real Datasets

To evaluate the performance of LCMARS with respect to LMARS and GGM, we use four real biochemical datasets. The names of these data, their numbers of genes and samples are listed in Table 3.9 and the biological details as well as the results of the analyses are presented in the following part.

Table 3.9: List of real pathways (datasets) used in this study with their numbers of genes (p) and samples (n).

Dataset	**Number of genes**	**Number of samples**
1 - Cell signal data	11	11672
2 - Human gene expression (B-lymphocyte cells)	100	60
3 - NCI-60 cell lines (p53) (Biocarta-PGC1A-pathway)	20	33
4 - Human ovarian tumour ((E-GEOD-9891)	11	285

Table 3.10: Comparison of recall, F-measure, Jaccard index, and accuracy measures for datasets which are listed in Table 3.9. BF refers to the number of basis functions.

Pathway	Method	Recall	F-measure	Jaccard index	Accuracy
1	**LCMARS (BF=10)**	**0.822**	**0.725**	**0.569**	0.769
	LCMARS (BF=3)	0.556	0.715	0.556	**0.835**
	LMARS	0.644	0.690	0.527	0.785
	GGM	0.289	0.448	0.289	0.736
2	**LCMARS (BF=10)**	0.885	0.386	0.240	0.838
	LCMARS (BF=4)	0.770	**0.805**	**0.673**	0.978
	LMARS	0.705	0.612	0.441	**0.983**
	GGM	**0.968**	0.178	0.098	0.831
3	**LCMARS (BF=10)**	0.511	0.475	0.312	0.735
	LCMARS (BF=3)	0.225	0.387	0.240	**0.810**
	LMARS	0.340	0.432	0.276	0.790
	GGM	**0.787**	**0.532**	**0.363**	0.675
4	**LCMARS**	**0.703**	**0.825**	**0.703**	**0.703**
	LMARS	0.306	0.469	0.306	0.306
	GGM	0.091	0.167	0.091	0.091

1. Cell Signaling Pathway

As our first dataset we use cell signaling data from Sachs et al. (2005) [78]. This dataset contains the flow cytometry results of 11 phosphorylated proteins and phospholipids measured on 11672 red blood cells. These components belong to the cellular protein-signaling network of human immune system cells. Therefore, the aim of the construction of this network is to understand the native-state tissue signaling biology, complex drug actions and the dysfunctional signals in diseased cells [78]. The graphical representation of the conventionally accepted signaling molecular interaction is presented in Figure 3.1.

In our analyses, the LCMARS method detects 13 links within these 17 biologically

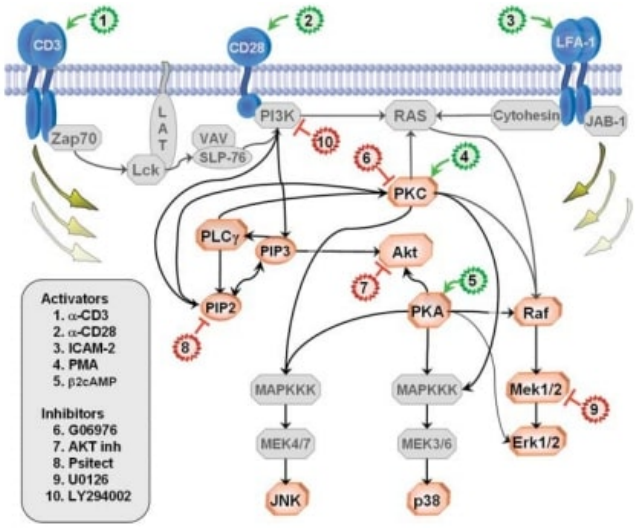


Figure 3.1: True graphical representation of the cell signaling network from Sachs et al. (2005) [78].

validated interactions, while the GGM approach catches only 1 of them. On the other hand, the LMARS method detects 9 links. The true and estimated networks with 3 methods are shown in Figure 3.2. We represent the accuracy measures of all methods in Table 3.10. From the tabulated terms, it is observed that LCMARS has higher recall, F-measure, Jaccard index and accuracy values than both LMARS and GGM.

Table 3.11: List of proteins used in the description of the cell signaling data as in the study of Sachs et al. (2005) [78].

Symbols	**Name of proteins**	**Symbols**	**Name of proteins**
P1	Raf	P7	Akt
P2	Mek	P8	PKA
P3	PLC-γ	P9	PKC
P4	PIP2	P10	p38
P5	PIP3	P11	Jnk
P6	Erk		

2. Human Gene Expression Data

For the second dataset, we implement large-scale human gene expression data originally described in the works of Bhadra and Mallick (2013) [13], Chen and Chen,

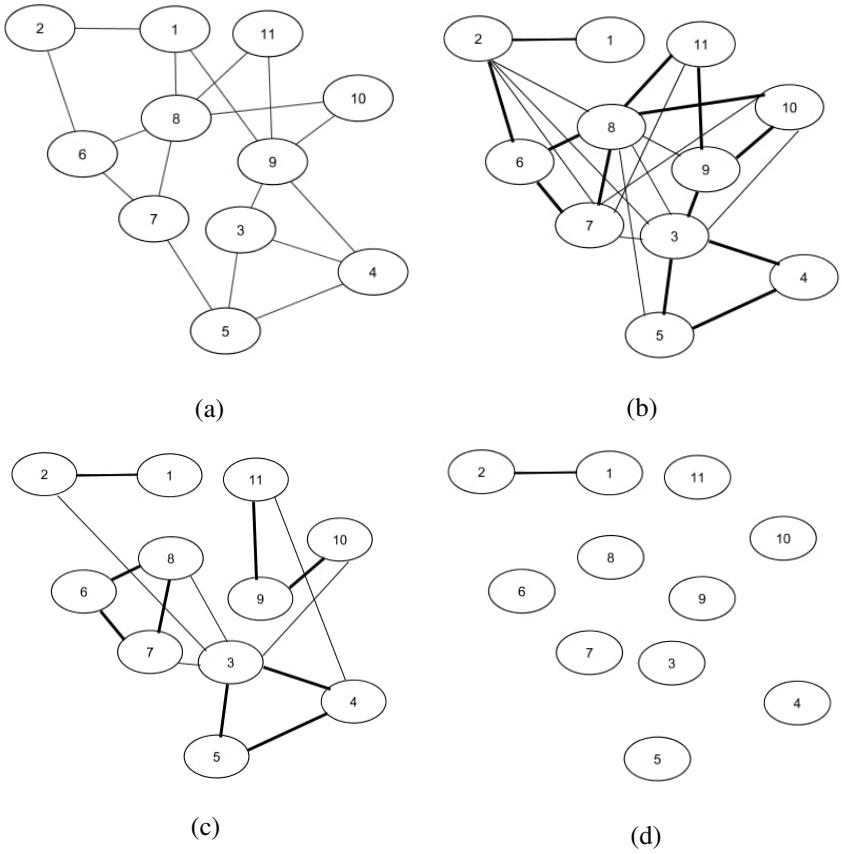


Figure 3.2: (a) True network of the cell signaling data, (b) estimated network via LCMARS, (c) estimated network via LMARS, and (d) estimated network via GGM. The true estimated links are shown in boldface and the complete list of proteins is given in Table 3.

(2008) [19] and Stranger et al. (2007) [83]. This dataset includes the gene expression of B-lymphocyte cells from the Utah residents with Northern and Western European ancestry sample. The genes of 60 unrelated individuals are probed for 100 different transcripts. Here, the focus is on the 3125 Single Nucleotide Polymorphisms (SNPs) that are found in the 5 UTR (untranslated region) of mRNA (messenger RNA) with a minor allele frequency greater than 0.1. The UTR has an important role in the regulation of gene expression. From the 55 biologically validated links, 45 have the names of the transcription factor and target genes in the network of gene expression data [13]. Therefore, for the inference of both models, we use these 45 links for the calculation of the accuracy measures. Hence, from the findings in Table 3.10, similar to the previous datasets, the performance of LCMARS is better than LMARS and GGM based on recall, F-measure and Jaccard Index. Moreover, accuracy value of LMARS is a bit higher than LCMARS. Whereas GGM has the highest score in terms of recall.

3. PGC-1A Pathway

As our third dataset, we apply a part of the NCI-60 cell lines (p53) data which include "Biocarta-PGC1A-Pathway" described in the study of Rahmatallah et al. (2014) [75]. The p53 dataset consists of gene expression profiling of 33 p53 mutated (MUT) cancer cell lines and is taken from the R package "GANPAdata" [26]. Here, we deal with 20 genes which constitute the PGC1A pathway. The peroxisome proliferator-activated receptor gamma coactivator-1 alpha (PGC-1A) is a tissue-specific coactivator that coordinates transcriptional programs important for energy metabolism and energy homeostasis. Thereby, inappropriate increases in the PGC-1A activity are linked to a number of pathological conditions including heart failure and diabetes. On the other side, PGC-1A is a coactivator for many factors including, CBP, Scr-1, PPARa, GR (glucocorticoid receptor), THR (thyroid hormone receptor), several orphan receptors and MEF2. Hereby, when we compare the performance of all three models, we observe that GGM is more successful than LCMARS and LMARS based on recall, F-measure and Jaccard index. Whereas, LCMARS is better in terms of accuracy measures. We consider that this special result may be caused by the particular structure of the pathway. The reason is that this system has very dense links. Thus, the method which is more used to assign links in the adjacency matrix can be more successful. Therefore, GGM is more advantageous in this system. Additionally, apart

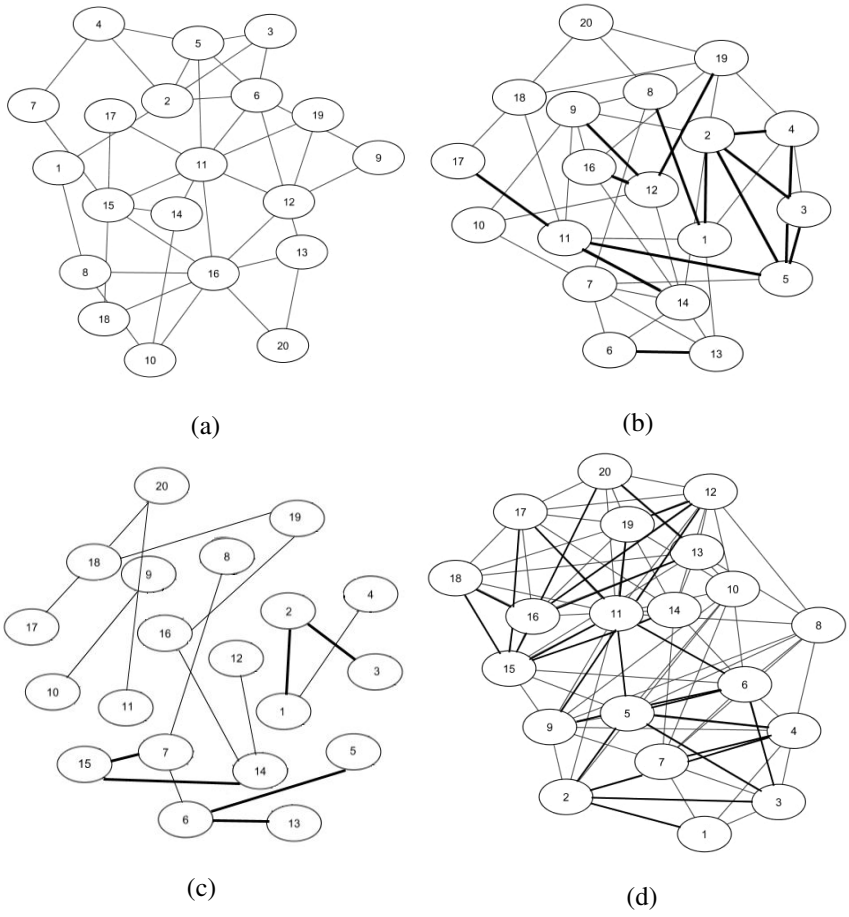


Figure 3.3: (a) True network of the PGC-1A Pathway, (b) estimated network via LCMARS, (c) estimated network via LMARS, and (d) estimated network via GGM. The true estimated links are shown in boldface.

from the accuracy measure, there is no any other measure which controls FN in the estimated system. Hence, the measures which control the presence of link when the system is dense can have higher score than other measures, which check the sparsity when there is no link. As a result, GGM under these measures is better than other methods except the "accuracy" measure. In Figure 3.3, we draw all the estimated networks with the true model for the visual representation of the findings.

4. Human Ovarian Tumour

Finally, as a real biological data, we apply the transcription profiling of 285 human ovarian tumour samples named as E-GEOD-9891. The data are cohorts of 285 patients with epithelial ovarian, primary peritoneal, or fallopian tube cancer, diagnosed

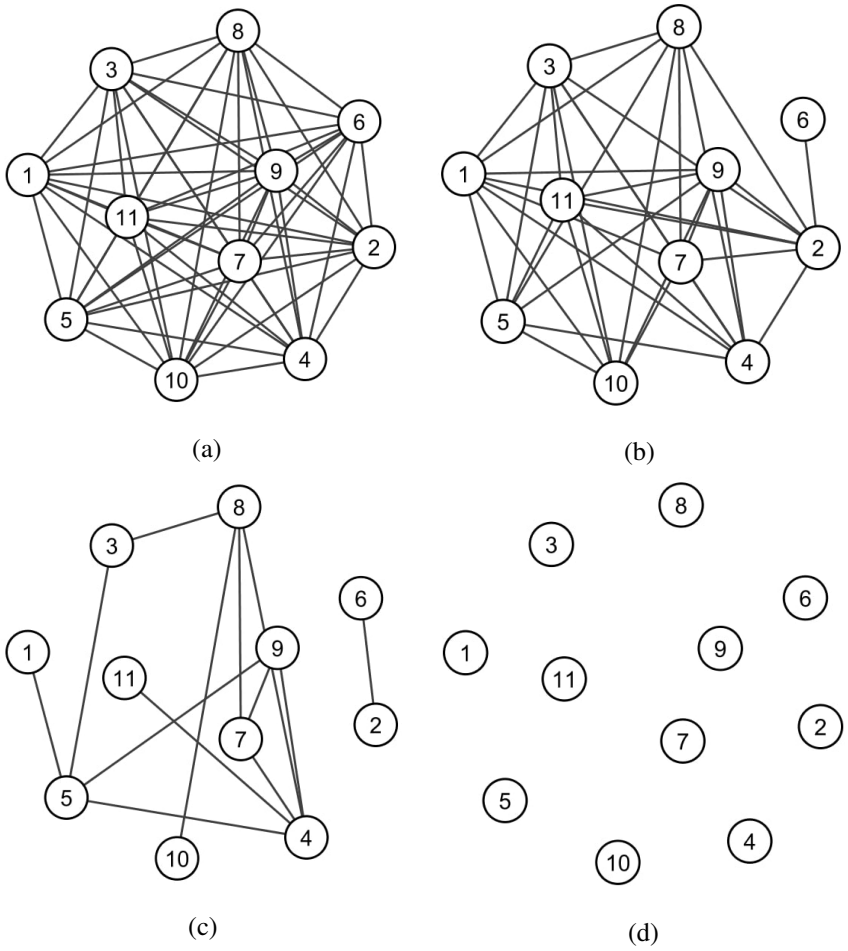


Figure 3.4: (a) True network of the E-GEOD9891 data, (b) estimated network via LCMARS, (c) estimated network via LMARS, and (d) estimated network via GGM.

between 1992 and 2006. The samples are collected through Australian Ovarian Cancer Study with a sample size $(n = 206)$, Royal Brisbane Hospital $(n = 22)$, Westmead Hospital $(n = 54)$ and Netherlands Cancer Institute $(n = 3)$ [89]. In this dataset, there are 11 target genes which are chosen to identify the novel molecular subtypes of the ovarian cancer by the gene expression profiling with a linkage to clinical and pathological features [89]. According to the results in Table 3.10, it is seen that LCMARS performs significantly better than other methods in terms of all measures. Thereby, as shown in Figure 3.4, LCMARS can detect 37 links out of 55 links of the true network. Whereas, LMARS can only capture 13 links in the same 55 links. On the other side, GGM cannot detect any link, i.e., edges between genes.

Table 3.12: List of the datasets.

Dataset ID	Dataset	Number of samples	Number of genes
1	Insilico-size 10-1	10	10
2	Insilico-size 10-3	10	10
3	Insilico-size 10-5	10	10
4	Insilico-size 100-1	100	100
5	Insilico-size 100-4	100	100
6	Gene expression (Toy)	64	64
7	NCI-60 cell lines (p53)	33	20
8	Human gene expression	60	100
9	Egeod-9891	285	11
10	Cell signal data	11672	11

3.4 Application of Outlier Detection Methods

In this part, we aim to investigate whether using an outlier detection method before modeling can improve the accuracy of fitted models and which method is the most appropriate for the protein-protein interaction data. For this purpose, we compare the accuracy results of three different modeling approaches under several outlier detection methods. To evaluate the performance of outlier detection methods under biological data, we use different synthetic and real benchmark biological datasets. The names of all datasets used in this study and their numbers of genes and samples are listed in Table 3.12. Additionally, the biological details of all data are explained in the previous parts.

In this application, we evaluate the performance of outlier detection methods under several biological datasets. Our main aim is to investigate whether an outlier detection approach is necessary as a pre-processing step before modeling. For this purpose, as mentioned earlier, we perform different outlier detection methods designed for univariate and multivariate cases. We apply Z-score and box plot methods within the univariate approaches. Additionally, we use PCOut, Sign method and BACON as multivariate methods. First of all, we detect the outliers in each dataset by using all these methods and remove the outliers. Then, we check the accuracy of estimated

systems with these outlier-free datasets under 3 network models, namely, LCMARS, LMARS and GGM. As accuracy measures, we compute the accuracy and F-measure by comparing the true and the estimated network structures. We compare the accuracy results of models under the full data which include potential outliers with the results of outlier-free datasets. All the accuracy results for each dataset and each method are given in Table 3.13 and Table 3.14. As presented in Table 3.13, applying the Z-score or box plot methods cannot increase the performance of models. Generally, the F-measure and the accuracy values either remain the same or small decreases occur. Additionally, the Z-score method does not identify an outlier in some data in Table 3.13. On the other hand, the results of the multivariate outlier detection methods, namely, PCOut, Sign and BACON, are presented by Table 3.14. From the results, we can see that for the first five data which are synthetic, the PCOut method is a bit better than Sign and BACON, especially, for the higher dimensional Data 4 and Data 5. Additionally, for the remaining real datasets, regression methods with the raw datasets give better results. Rarely, the Sign method gets the upper hand, but the increases in the F-measure and accuracy are very low. In other words, overall, there is no any outlier detection method which outperforms the others. Finally, when we compare the regression models, we can observe that the performance of LCMARS is the best and GGM is the worst as declared in the previous analysis. From our results, we conclude that the outlier detection as a pre-processing step cannot improve the accuracy of the models. Furthermore, there is no unique best method for the outlier detection for protein-protein interaction data. Additionally, there are only a few methods applicable for sparse and high-dimensional systems.

Table 3.13: Comparison of F-measure and accuracy values of Z-score and box plot methods under LCMARS, GGM and LMARS for datasets listed in Table 3.12.

Data	Methods	Full data		Z-score		Box-plot	
		F-measure	Accuracy	F-measure	Accuracy	F-measure	Accuracy
1	**LCMARS**	0.647	0.760	*	*	0.563	0.720
	GGM	0.454	0.760	*	*	0.454	0.760
	LMARS	0.616	0.800	*	*	0.560	0.780
2	**LCMARS**	0.563	0.720	*	*	0.533	0.720
	GGM	0.416	0.720	*	*	0.416	0.720
	LMARS	0.538	0.760	*	*	0.384	0.680
3	**LCMARS**	0.583	0.800	*	*	-	-
	GGM	0.476	0.780	*	*	0.476	0.780
	LMARS	0.522	0.780	*	*	0.476	0.780
4	**LCMARS**	0.296	0.943	0.268	0.942	-	-
	GGM	-	-	0.373	0.966	0.373	0.966
	LMARS	0.298	0.949	0.295	0.951	0.362	0.963
5	**LCMARS**	0.267	0.934	0.280	0.939	0.277	0.946
	GGM	-	-	0.326	0.959	0.326	0.959
	LMARS	0.258	0.939	0.280	0.946	0.321	0.956
6	**LCMARS**	0.347	0.842	0.318	0.837	-	-
	GGM	0.236	0.899	0.780	0.991	0.236	0.899
	LMARS	0.342	0.895	0.326	0.891	0.283	0.898
7	**LCMARS**	0.387	0.810	*	*	0.387	0.810
	GGM	0.532	0.675	*	*	0.500	0.620
	LMARS	0.432	0.790	*	*	0.377	0.785
8	**LCMARS**	0.805	0.978	0.733	0.991	0.636	0.987
	GGM	0.178	0.831	0.168	0.817	0.224	0.911
	LMARS	0.612	0.983	0.594	0.983	0.638	0.988
9	**LCMARS**	0.825	0.703	0.837	0.719	0.790	0.653
	GGM	0.167	0.091	0.167	0.091	0.167	0.091
	LMARS	0.469	0.306	0.487	0.322	0.506	0.339
10	**LCMARS**	0.715	0.835	0.594	0.785	0.594	0.785
	GGM	0.448	0.736	0.594	0.785	0.625	0.752
	LMARS	0.690	0.785	0.649	0.686	0.559	0.570

[1] (*) represents that outliers are not detected and (-) denotes that the models are not computable.

Table 3.14: Comparison of F-measure (F.) and accuracy (Acc.) values of PCOut, Sign and BACON methods under LCMARS, GGM and LMARS for datasets listed in Table 3.12.

Data	Methods	Full data		PCout		Sign		BACON	
		F.	Acc.	F.	Acc.	F.	Acc.	F.	Acc.
1	**LCMARS**	0.647	0.760	-	-	0.647	0.760	0.563	0.720
	GGM	0.454	0.760	0.454	0.760	0.454	0.760	0.454	0.760
	LMARS	0.616	0.800	0.616	0.800	0.616	0.800	0.435	0.740
2	**LCMARS**	0.563	0.720	0.533	0.720	0.500	0.680	0.514	0.660
	GGM	0.416	0.720	0.416	0.720	0.416	0.720	0.416	0.720
	LMARS	0.538	0.760	0.384	0.680	0.462	0.720	0.462	0.720
3	**LCMARS**	0.583	0.800	0.560	0.780	0.560	0.780	0.435	0.740
	GGM	0.476	0.780	0.476	0.780	0.470	0.780	0.476	0.780
	LMARS	0.522	0.780	0.584	0.800	0.545	0.800	0.500	0.760
4	**LCMARS**	0.296	0.943	0.258	0.941	0.281	0.942	0.263	0.938
	GGM	-	-	0.373	0.966	0.373	0.966	0.373	0.966
	LMARS	0.298	0.949	0.373	0.966	0.288	0.950	0.328	0.948
5	**LCMARS**	0.267	0.934	0.281	0.938	0.253	0.934	0.259	0.934
	GGM	-	-	0.326	0.959	0.326	0.959	0.326	0.959
	LMARS	0.258	0.939	0.332	0.959	0.285	0.943	0.259	0.938
6	**LCMARS**	0.347	0.842	-	-	0.295	0.832	0.205	0.807
	GGM	0.236	0.899	0.236	0.899	0.236	0.899	0.236	0.899
	LMARS	0.342	0.895	0.325	0.897	0.325	0.895	0.328	0.888
7	**LCMARS**	0.387	0.810	0.361	0.805	0.413	0.815	0.344	0.790
	GGM	0.532	0.675	0.400	0.565	0.397	0.560	0.370	0.405
	LMARS	0.432	0.790	0.367	0.775	0.438	0.795	0.351	0.815
8	**LCMARS**	0.805	0.978	0.709	0.991	0.704	0.991	-	-
	GGM	0.178	0.831	0.199	0.852	0.207	0.859	-	-
	LMARS	0.612	0.983	0.656	0.986	0.645	0.985	-	-
9	**LCMARS**	0.825	0.703	0.778	0.636	0.825	0.703	-	-
	GGM	0.167	0.091	0.167	0.091	0.167	0.091	-	-
	LMARS	0.469	0.306	0.429	0.273	0.448	0.289	-	-
10	**LCMARS**	0.715	0.835	0.594	0785	0.594	0.785	0.367	0.686
	GGM	0.448	0.736	0.658	0.785	0.690	0.785	0.392	0.719
	LMARS	0.690	0.785	0.554	0.587	0.559	0.570	0.560	0.521

[1] (-) denotes the models that are not computable.

CHAPTER 4

CONCLUSION

In this thesis, we have proposed nonparametric regression model, called LCMARS, as an alternative modeling to GGM on the description of the biological networks. The prediction of complex biological networks and the estimation of the unknown system's parameters have some challenges because of the large amount of data. So there are various statistical methods to construct the structure of biological networks. GGM is one of the most widely used graphical models for this purpose. However, GGM has two major limitations which are the restriction of the normality assumption in the explanation of the states, and the low accuracy in the construction of the actual biological pathways. Particularly in high-dimensional biological systems, the estimation of the structure can be challenging due to the sparsity in the precision matrix and the inference of the model parameter becomes computationally demanding. On the other hand, MARS is one of the well-known nonparametric regression methods that enables us to model the high-dimensional data under nonlinearity. Additionally, the CMARS method is a modified version of MARS. Basically, the MARS model is a special type of generalized additive models which represents the causes of the error propagation via a regularized nonlinear regression model. Hereby, in CMARS by keeping these listed advantages of MARS, a penalized residual sum of squares is implemented by eliminating the backward stepwise algorithm and can be solved by the conic quadratic optimization. This model was originally designed for highly correlated datasets without distributional limitations.

Hereby in this thesis, initially, we have adapted the original CMARS as a loop-based regression with the interaction effects and called it LCMARS due to its loop-based description. Here, the main effects imply the direct relations between genes and the second-order interactions are the representative of the feed-forward-loop motifs'

structures in biochemical systems. We have used the second-order interactions in order to better describe the structure of biological systems. In this thesis, to compare the performance of LCMARS with LMARS and GGM, we have conducted a simulation study under distinct distributions and dimensions. Additionally, several synthetic and real benchmark datasets have used. The accuracy measures such as recall, F-measure, Jaccard index, and accuracy values have used to evaluate the performance of methods.

As a result of our analyses, we have seen that, LCMARS has higher F-measure, Jaccard index and accuracy values than both LMARS and GGM under both normal and non-normal distributions. Especially, for non-normal and high dimensional systems, the difference between methods become larger. Therefore, we have concluded that LCMARS can be a strong alternative to GGM since LCMARS overcomes the limitations of GGM such as the strict normality assumption and a low accuracy under high-dimensional complex systems. Furthermore, as a result of our analyses based on the several biological datasets, we have shown that our proposed model has a higher accuracy than LMARS and GGM based on various accuracy measures. Hence, we believe that LCMARS can be a promising alternate of GGM and LMARS in the description of biological networks.

On the other side, in the second part of the study, we have investigated whether an outlier detection approach is necessary as a pre-processing step before modeling the protein-protein interaction data. For this purpose, we have searched several outlier detection methods designed for univariate and multivariate data. Among these methods, we have chosen the Z-score and box plot under the univariate methods and robust PCA (PCOut), Sign method and BACON within multivariate methods. Once the outliers are detected via these listed approaches, we have implemented GGM, LMARS and LCMARS methods. Then, we have checked the accuracy of the estimated models by using accuracy measures such as, F-measure and accuracy values. For this purpose, we have used several synthetic and real benchmark biological datasets which have different dimensions and different structures. From the results, it has seen that the outlier detection as a pre-processing step cannot improve the accuracy of the models. In some cases, PCOut or Sign methods can give a bit better results. However, the increases in the accuracy values are very low. Hence, within the outlier detection methods, we have not found any method which can significantly outperform the

others in the construction of biological networks.

As the future work of the study, we consider to adapt the proposal modeling approach to its robust version, so-called RCMARS [68]. Shortly, the RCMARS model is suggested to detect the model which has the lowest false positive rate [69] and resulting in finding the minimum number of links in the system. We think that this new model can be converted as a loop-based model similar to LCMARS in order to construct the major core subnetworks. Additionally, robust conic generalized partial linear model (RCGPLM) can be also used to estimate the biological networks. This method consists of a combination of two regression models, namely, logistic regression and RCMARS [70]. RCGPLM can reduce the complexity of CMARS and increase the rate of accuracy [70, 93]. Furthermore, we consider that the description of the spline functions in MARS, CMARS and RCMARS models can be extended by covering not only the linear relations, but also nonlinear relationships between genes. For this purpose, we think to perform the p-splines approach [57] which has unspecified smoothing functions in modeling. Moreover, a threshold value can be applied to define the links in networks by LCMARS, rather than direct usage of significant regression coefficients. In this calculation, we consider to perform the kappa maximized threshold and the maximized sum threshold [39, 55], so that we can obtain more sparse networks. Lastly, a hybrid approach that combines the robust outlier algorithm, called Mean Shift Outlier Model (MSOM) [56], with CMARS can be implemented to construct networks in the presence of outliers [102]. The MSOM-CMARS method aims to minimize the impact of the outliers.

REFERENCES

[1] Aggarwal, C. C. (2013). *Outlier analysis*. New York: Springer-Verlag.

[2] Alon, U. (2007). *An introduction to systems biology: design principles of biological circuits*. Chapman and Hall/CRC.

[3] Attoh-Okine, N. O., Cooger, K., and Mensah, S. (2009). Multivariate adaptive regression (MARS) and hinged hyperplanes (HHP) for doweled pavement performance modeling. *Construction and Building Materials*, 23(9), 3020–3023.

[4] Ayyıldız, E., Ağraz, M. and Purutçuoğlu, V. (2017). MARS as an alternative approach of Gaussian graphical model for biochemical networks. *Journal of Applied Statistics*, 44(16), 2858-2876.

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

[6] Barron, A. R., and Xiao, X. (1991). Discussion: Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1), 67–82.

[7] Beck, A., and Teboulle, M. (2000). Global optimality conditions for quadratic optimization problems with binary constraints. *SIAM Journal of Optimization*, 11(1), 179–188.

[8] Beck, A., and Teboulle, M. (2009). *Gradient based algorithms with applications in signal recovery problems in convex optimization in signal processing and communications*, D.P. Palomar, and Y.C. Eldar, eds, Cambridge University Press.

[9] Beck, A., and Teboulle, M. (2012). Smoothing and first order methods: a unified framework. *SIAM Journal of Optimization*, 22(2), 557–580.

[10] Bellot, P., Olsen, C., and Meyer, P. E. (2014). Grndata: synthetic expression data for gene regulatory network inference. R package version 1.12.0.

[11] Benfenati, A., Chouzenoux, E., and Pesquet, J. C. (2018). A nonconvex variational approach for robust graphical lasso. IEEE International Conference on Acoustics, Speech, and Signal Processing, Calgary, Canada.

[12] Ben-Gal, I. (2005). *Outlier detection*. In Data mining and knowledge discovery handbook. Springer, Boston, MA.

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

[14] Billor, N., Hadi, A. S., and Velleman, P. F. (2000). BACON: blocked adaptive computationally efficient outlier nominators. *Computational Statistics and Data Analysis*, 34(3), 279-298.

[15] Bower, J. M., and Bolouri, H. (2001). *Computational modelling of genetic and biochemical networks*. MIT Press.

[16] Brown, C. (2007). *Differential equations: a modeling approach*. SAGE Publications.

[17] Van den Bulcke T., Van Leemput K., Naudts, B., Van Remortel, P., Ma, H., Verschoren, A., Moor, B., and Marchal, K. (2006). SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, 7(1), 43.

[18] Carlin, B. P., and Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis*. Chapman and Hall/CRC.

[19] Chen, J., and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3), 759-771.

[20] Cheung, N. J., Xu, Z. K., Ding, X. M., and Shen, H. B. (2015). Modeling nonlinear dynamic biological systems with human-readable fuzzy rules optimized by convergent heterogeneous particle swarm. *European Journal of Operational Research*, 247(2), 349–358.

[21] Defterli, Ö., Purutçuoğlu, V., and Weber, G. W. (2014). *Advanced mathematical and statistical tools in the dynamic modeling and simulation of gene-*

environment networks. In Modeling, Dynamics, Optimization and Bioeconomics. Springer, Cham.

[22] Dixon, W. J. (1950). Analysis of extreme values. *The Annals of Mathematical Statistics*, 21(4), 488–506.

[23] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2), 407–499.

[24] Ergenc, T., and Weber, G. W. (2004). Modeling and prediction of gene-expression patterns reconsidered with Runge-Kutta discretization. *Journals of Computational Technologies*, 9(6), 40–48.

[25] Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics*, 3(2), 521–541.

[26] Fang, Z., Tian, W., and Ji, H. (2015). The GANPA datasets package. R package version 1.0.

[27] Filzmoser, P., Maronna, R., and Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics and Data Analysis*, 52(3), 1694-1711.

[28] Frank, L. E., and Friedman, J. H. (1993). A statistical view of some chemometrics regression tool. *Technometrics*, 35(2), 109–135.

[29] Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annuals of Statistics*, 19(1), 1–67.

[30] Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2), 302–332.

[31] Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432–441.

[32] Gebert, J., Laetsch, M., Quek, E. M. P., and Weber, G. W. (2004). Analyzing and optimizing genetic network structure via path-finding. *Journals of Computational Technologies*, 9, 3–12.

[33] Gebert, J., Radde, N., and Weber, G. W. (2007). Modeling gene regulatory networks with piecewise linear differential equations. *European Journal of Operational Research*, 181(3), 1148–1165.

[34] Gill, P. E., Murray, W., and Saunders, M. A. (1997). User's guide for SQOPT 5.3: A Fortran package for large-scale linear and quadratic programming. *Technical Report NA 97-4*, Department of Mathematics, University of California, San Diego.

[35] Gillespie, D. T. (1977). Exact stochastic simulations of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25), 2340–2361.

[36] Goeman, J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, 52(1), 70–84.

[37] Golightly, A., and Wilkinson, D. J. (2005). Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(3), 781–788.

[38] Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, 21(1), 27–58.

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

[40] Hadi, A. S., Imon, A. R., and Werner, M. (2009). Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1), 57–70.

[41] Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3), 761-771.

[42] Hadi, A. S. (1994). A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(2), 393-396.

[43] Hallac, D., Park, Y., Boyd, S., and Leskovec, J. (2017). Network inference via the time-varying graphical lasso. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[44] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. New York: Springer Series in Statistics.

[45] Hawkins, D. M. (1980). *Identification of outliers*. London: Chapman and Hall.

[46] Hirose, K., Fujisawa, H., and Sese, J. (2017). Robust sparse Gaussian graphical modeling. *Journal of Multivariate Analysis*, 161, 172–190.

[47] Helms, V. (2018). *Principles of computational cell biology: from protein complexes to cellular networks*. John Wiley and Sons.

[48] Hesterberg, T., Choi, N. H., Meier, L., and Fraley, C. (2008). Least angle and l_1 penalized regression: A review. *Statistics Surveys*, 2, 61-93.

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

[50] Huang, J., Ma, S., and Zhang, C. H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18, 1603-1618.

[51] Hubert, M., and Debruyne, M. (2010). Minimum covariance determinant. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1), 36–43.

[52] Iglewicz, B., and Hoaglin, D. (1993). *How to detect and handle outliers*. The ASQC Quality Press.

[53] Ivashkiv, L. B., and Donlin, L. T. (2014). Regulation of type I interferon responses. *Nature Reviews Immunology*, 14(1), 36–49.

[54] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York: Springer.

[55] Jiménez-Valverde, A., and Lobo, J. M. (2007). Threshold criteria for conversion of probability of species presence to eitheror presenceabsence. *Acta Oceologica*, 31(3), 361-369.

[56] Kim, S. S., Park, S. H., and Krzanowski, W. J. (2008). Simultaneous variable selection and outlier identification in linear regression using the mean-shift outlier model. *Journal of Applied Statistics*, 35(3), 283-291.

[57] Lang, S., and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13(1), 183-212.

[58] Junker, B. H., and Schreiber, F. (2008). *Analysis of biological networks*. John Wiley and Sons.

[59] Lewis, P. A., and Stevens, J. G. (1991). Nonlinear modeling of time series using multivariate adaptive regression splines (MARS). *Journal of the American Statistical Association*, 86(416), 864–877.

[60] Liu, H., Roeder, K., and Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. *In Advances in Neural Information Processing Systems*, 23, 1432–1440.

[61] Liu, X., Kong, X., and Ragin, A. B. (2017). Unified and contrasting graphical lasso for brain network discovery. In Proceedings of the 2017 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics.

[62] Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., et al. Robust principal component analysis for functional data. *Test*, 8(1), 1-73.

[63] Maiwald, T., Schneider, A., Busch, H., Sahle, S., Gretz, N., Weiss, T. S., Kummer, U., and Klingmüller, U. (2010). Combining theoretical analysis and experimental data generation reveals IRF9 as a crucial factor for accelerating interferon α-induced early antiviral signalling. *The FEBS Journal*, 277(22), 4741–4754.

[64] Moles, C. G., Mendes, P., and Banga, J. R. (2003). Parameter estimation in biochemical pathways: A comparison of global optimization methods. *Genome Research*, 13, 2467–2474.

[65] Mohammadi, A., and Wit, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical model. *Bayesian Analysis*, 10(1), 109–138.

[66] Mohammadi, A., Abegaz, F., Heuvel, E., and Wit, E. C. (2017). Bayesian modelling of Dupuytren disease by using Gaussian copula graphical models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3), 629–645.

[67] Mohammadi, A., and Wit, E. C. (2017). BDgraph: Bayesian structure learning in graphical models using birth-death MCMC, R package version 3.40.

[68] Özmen, A., Weber, G. W., and Batmaz, İ. (2010). The new robust CMARS (RCMARS) method. In: ISI Proceedings of 24th MEC-EurOPT Continuous optimization and information-based technologies in the financial sector, Turkey, 362-368.

[69] Özmen, A., Weber, G. W., Batmaz, İ., and Kropat, E. (2011). RCMARS: Robustification of CMARS with different scenarios under polyhedral uncertainty set. *Communications in Nonlinear Science and Numerical Simulation*, 16(12), 4780-4787.

[70] Özmen, A., Weber, G. W., Çavuşoğlu, Z., and Defterli, Ö. (2013). The new robust conic GPLM method with an application to finance: prediction of credit default. *Journal of Global Optimization*, 56, 233–249.

[71] Peterson, C. B., Stingot, F. C., and Vannucci, M. (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association*, 110(509), 159-174.

[72] Pilla, V. L., Rosenberger, J. M., Chen, V., Engsuwan, N., and Siddappa, S. (2012). A multivariate adaptive regression splines cutting plane approach for solving a two-stage stochastic programming fleet assignment model. *European Journal of Operational Research*, 216, 162–171.

[73] Pearson, E. S. (1963). Some problems arising in approximating to probability distributions using moments. *Biometrika*, 50, 95-112.

[74] Purutçuoğlu, V. , Ayyıldız, E. , Wit, E. (2017). Comparison of two inference approaches in Gaussian graphical models. *Turkish Journal of Biochemistry*, 42(2), 203-212 .

[75] Rahmatallah, Y., Emmert-Streib, F., and Glazko, G. (2014). Gene sets net correlations analysis (GSNCA): a multivariate differential coexpression test for gene sets. *Bioinformatics*, 30(3), 360–368.

[76] Reinker, S., Altman, R. M., and Timmer, J. (2006). Parameter estimation in stochastic biochemical reactions. *IEEE Proceedings Systems Biology*, 153(4), 168–178.

[77] Rousseeuw, P. J., and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212-223.

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

[79] 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*, 4(1),32.

[80] Schaffter, T., Marbach, D., and Floreano, D. (2011). GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16), 2263–2270.

[81] Schiffler, R .E. (1998). Maximum Z scores and outliers. *The American Statistician*, 42(1), 79-80.

[82] Shuai, K., and Liu, B. (2003). Regulation of JAK-STAT signalling in the immune system. *Nature Reviews Immunology*, 3, 900–911.

[83] Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., Montgomery, S., Tavare, S., Deloukas, P., and Dermitzakis, E. T. (2007). Population genomics of human gene expression. *Nature Genetics*, 39(10), 1217–1224.

[84] Tan, W. Y., and Tiku, M .L. (1999). *Sampling distributions in terms of Laguerre polynomials with applications*. New Age International.

[85] Taylan, P., Weber, G. W., and Beck, A. (2007). New approaches to regression by generalized additive models and continuous optimization for modern applications in finance, science and technology. *Optimization*, 56(5-6), 675–698.

[86] Taylan, P., Weber, G. W., and Yerlikaya-Özkurt, F. (2010). A new approach to

multivariate adaptive regression splines by using Tikhonov regularization and continuous optimization. TOP, 18(2), 377–395.

[87] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, 58(1), 267–288.

[88] Tibshirani, R., Saunder, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via fused lasso. *Journal of the Royal Statistical Society*, 67(1), 91–108.

[89] Tothill, R., Tinker, A., George, J., Brown, R., Fox, S., Johnson, D., et al. (2008). Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical cancer research*, 14(16), 5198–5208.

[90] Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA : Addison-Wesley.

[91] Uğur, O., Pickl, S. W., Weber, G. W., and Wünschiers, R. (2006). An algorithm approach to analyse genetic networks and biological energy production: An introduction and contribution where OR meets biology. *Optimization*, 58(1), 1–22.

[92] Weber, G. W., Defterli, Ö., Gök, S. Z. A., and Kropat, E. (2011). Modeling, inference and optimization of regulatory networks based on time series data. *European Journal of Operational Research*, 211, 1–14.

[93] Weber, G. W., Çavuşoğlu, Z., and Özmen, A. (2012). Predicting default probabilities in emerging markets by new conic generalized partial linear models and their optimization. *Optimization*, 61(4), 443–457..

[94] Wenjiang, J. F. (1998). Penalized regression: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3), 397–416.

[95] Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. John Wiley and Sons, New York.

[96] Wilkinson, D. J. (2006). *Stochastic modelling for systems biology*. Boca Raton, FL, Taylor and Francis.

[97] Williams, D. R., Piironen, J., Vehtari, A., and Rast, P. (2018). Bayesian estimation of Gaussian graphical models with predictive covariance selection. arXiv: 1801.05725.

[98] Wit, E., Vinciotti, V., and Purutçuoğlu, V. (2019). *Statistics for biological networks: how to infer networks from data*. Chapman and Hall/ CRC.

[99] 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.

[100] Yamada, S., Shiono, S., Joo, A., and Yoshimura, A. (2003). Control mechanism of JAK-STAT signal transduction pathway. *FEBS Letters*, 534, 190–196.

[101] Yerlikaya-Özkurt, F., Batmaz, İ., and Weber, G. W. (2014). A review and new contribution on conic multivariate adaptive regression splines (CMARS): a powerful tool for predictive data mining. In: Modeling, Dynamics, Optimization and Bioeconomics, Zilberman, I.D., and Pinto, A.A. (Eds.). Springer International Publishing Switzerland, 37, 695–722.

[102] Yerlikaya-Özkurt, F., Askan, A., and Weber, G. W. (2016). A hybrid computational method based on convex optimization for outlier problems: application to earthquake ground motion prediction. *Informatica*, 27(4), 893–910.

[103] Zhou, S., Rütimann, P., Xu, M., and Bühlmann, P. (2011). High dimemsional covariance estimation based on Gaussian graphical models. *Journal of Machine Learning Research*, 12(4), 2975–3026.

[104] Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, 67(2), 301–320.

[105] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.

[106] Zomorrodi, A. R., and Maranas, C .D. (2014). Coarse-grained optimization-driven design and piecewise linear modeling of synthetic genetic circuits. *European Journal of Operational Research*, 237, 665–676.

CURRICULUM VITAE

PERSONAL INFORMATION

Surnam, Name: Ayyıldız Demirci, Ezgi
Nationality: Turkish (TC)
Date and Place of Birth: 17.09.1988, Ankara
Marital Status: Marriage
Phone: 0 536 815 20 97
Email: ezgiayyildizdemirci@gmail.com

EDUCATION

Degree	Institution	Year of Graduation
MS	METU Department of Statistics	2013
BS, Minor	METU Computer Engineering	2012
BS	METU Department of Statistics	2011

PUBLICATIONS

1. **Ayyıldız, E.**, and Purutçuoğlu, V. Modeling of various biological networks via LCMARS, 2018, Journal of Computational Science, 28, 148-154.

2. **Ayyıldız, E.**, and Purutçuoğlu, V. Is it necessary to apply the outlier detection for protein-protein interaction data?, 2018, Journal of Biostatistics-Turkish Clinics, 10(3), 173-186.

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

4. **Ayyıldız, E.**, Ağraz, M. and Purutçuoğlu, V. MARS as an alternative approach

of Gaussian graphical model for biochemical networks, 2017, Journal of Applied Statistics, 44(16), 2858-2876.

5. Purutçuoğlu, V., and **Ayyıldız, E.** Comparison of two inference approaches in Gaussian graphical models, 2017, Turkish Journal of Biochemistry, 42(2), 203-211.

6. Purutçuoğlu, V., and **Ayyıldız, E.** Mathematical modeling of gene networks (in Turkish), 2017, Journal of Biostatistics-Turkish Clinics, 9(2), 143-155.

7. **Ayyıldız, E.**, Purutçuoğlu, V., and Wit, E. A short note on resolving singularity problems in covariance matrices, 2012, International Journal of Statistics and Probability, 1(2), 113-118.

BOOK

- Purutçuoğlu, V. and **Ayyıldız, E.** (December, 2014) Statistics in the Field of Bioinformatics (Biyoinformatik Alanında İstatistik-in Turkish), Nobel Publisher. ISBN: 978-605-320-008-6.

CHAPTER IN BOOK

- **Ayyıldız, E.**, and Purutçuoğlu, V. (2018) Generating various types of graphical models via MARS. Chapter in: Information Complexity and Statistical Modeling in High Dimensions with Applications. Editors: O. Arslan. Springer (Accepted)

PROJECT

- Researcher, January 2016-June 2017, Scientific Research Project (BAP1), METU. Project title: Alternative approaches, inference and copulas in deterministic modellings of complex biological systems. Project no: BAP-01-09-2016-002.

CONFERENCE PROCEEDINGS

1. **Ayyıldız, E.** and Purutçuoğlu, V. "A new steady-state modeling approach for protein-protein interaction networks", Proceeding of the 32nd European Meeting of Statisticians (EMS), Palermo, Italy, 2019.

2. **Ayyıldız, E.** and Purutçuoğlu, V. "Is the outlier detection appropriate for protein-protein interaction data?", Proceeding of the 4th International Researchers, Statisticians and Young Statisticians Congress (IRSYSC), İzmir, Turkey, 2018.

3. Purutçuoğlu, V. and **Ayyıldız, E.** "Construction of ovarian cancer pathway via different mathematical models", Proceeding of the 3rd International Researchers, Statisticians and Young Statisticians Congress (IRSYSC), Konya, Turkey, 2017.

4. **Ayyıldız, E.** and Purutçuoğlu, V. "Modeling of various biological networks via LCMARS", Proceeding of the International Workshop on Mathematical Methods in Engineering (MME), Ankara, Turkey, 2017.

5. **Ayyıldız, E.** and Purutçuoğlu, V. "Generating various types of graphical models via MARS", Proceeding of the International Conference on Information Complexity and Statistical Modeling in High Dimensions with Applications (IC-SMHD), Nevşehir, Turkey, 2016.

6. Purutçuoğlu, V. and **Ayyıldız, E.** "Deterministic inference of networks by capturing their dynamic behaviours", Proceeding of the 6th International Workshop on Differential Equations and Applications, İzmir, Turkey, 2013.

7. **Ayyıldız, E.** and Purutçuoğlu, V. "Inference of the biological systems via L1-penalized lasso regression", Proceeding of the 29th European Meeting of Statisticians (EMS), Budapest, Hungary, 2013.