DIFFERENT TYPES OF MODELLINGS AND THE INFERENCE OF MODEL PARAMETERS FOR COMPLEX BIOLOGICAL SYSTEMS

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

BY

MELİH AĞRAZ

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

MAY 2017

Approval of the thesis:

DIFFERENT TYPES OF MODELLINGS AND THE INFERENCE OF MODEL PARAMETERS FOR COMPLEX BIOLOGICAL SYSTEMS

submitted by **MELİH AĞRAZ** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Statistics Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver
Dean, Graduate School of **Natural and Applied Sciences**

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

Assoc. Prof. Dr. Vilda Purutçuoğlu
Supervisor, **Department of Statistics, METU**

Examining Committee Members:

Prof. Dr. Gerhard Wilhelm Weber
Institute of Applied Mathematics, METU

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

Prof. Dr. Barış Sürücü
Department of Statistics, METU

Prof. Dr. Birdal Şenoğlu
Department of Statistics, ANKARA University

Assist. Prof. Dr. Neslihan İyit
Department of Statistics, SELÇUK University

Date:

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

Name, Last Name: MELİH AĞRAZ

Signature :

ABSTRACT

DIFFERENT TYPES OF MODELLINGS AND THE INFERENCE OF MODEL PARAMETERS FOR COMPLEX BIOLOGICAL SYSTEMS

AĞRAZ, MELİH
Ph.D., Department of Statistics
Supervisor : Assoc. Prof. Dr. Vilda Purutçuoğlu

May 2017, 105 pages

A reaction set that form a system can be modeled mathematically in different ways such as boolean, ordinary differential equations and stochastic modellings. Among them the random system is merely taken into account by the stochastic approach that is based on the known number of molecules in the reactions and if we consider the behaviour of the system under steady state condition, the modelling can be done via deterministic methods such as the ordinary differential equation. In this thesis, firstly, we aim to estimate the model parameters of a realistically complex biochemical system that is modelled to describe the steady state behaviour of the system. Among alternatives, we implement the Gaussian graphical models (GGM) which is one of the well known probabilistic model in this class. Here initially we develope an alternative approach of GGM in nonparametric distribution. For this purpose, we suggest LMARS (lasso-type multivariate adaptive regression spline) method. Then, we propose a normalization step called Bernstein polynomials for raw data to improve the performance of these models. Finally we suggest another alternative of GGM in parametric class and infer the model parameter via a novel estimation method, called the MMLE (modified maximum likelihood estimator). We evaluate all over findings with simulated and real data compute their accuracies as well as computational time behaviour of the system.

Keywords: Gaussian graphical model, Bernstein polynomials, multivariate adaptive regression spline (MARS), modified maximum likelihood estimator (MMLE)

ÖZ

KARMAŞIK BİYOLOJİK SİSTEMLER İÇİN FARKLI TÜRLERDE MODELLEMELER VE MODEL PARAMETRELERLERİNİN ÇIKARIMI

AĞRAZ, MELİH
Doktora, İstatistik Bölümü
Tez Yöneticisi : Doç. Dr. Vilda Purutçuoğlu

Mayıs 2017 , 105 sayfa

Bir sistem oluşturan boolean, adi diferansiyel denklemler ve stokastik modellemeler gibi bir dizi reaksiyon matematiksel olarak farklı yollarla modellenebilir. Bunların arasında rasgele sistemde, reaksiyonlardaki bilinen molekül sayısına dayanan stokastik yaklaşımla dikkate alınır ve sistemin kararlı durum koşulu altında davranışını göz önüne alırsak, modelleme adi diferansiyel denklem gibi deterministik yöntemlerle yapılabilir. Bu tezde, sistemin kararlı durum davranışını tanımlamak için modellenen gerçekçi karmaşık bir biyokimyasal sistemin model parametrelerini tahmin etmeyi amaçlıyoruz. Alternatifler arasında, bu sınıfta iyi bilinen olasılık modellerinden biri olan Gauss grafiksel modellerini (GGM) uygulamaktayız. Burada başlangıçta, parametrik olmayan dağılımda GGM'e alternatif bir yaklaşım geliştiriyoruz ve bu amaçla LMARS yöntemini öneriyoruz. Ardından, bu modellerin performansını artırmak için ham veriler için Bernstein polinomları adı verilen bir normalizasyon adımını öneriyoruz. Son olarak, parametrik sınıfta başka bir GGM alternatifi öneriyoruz ve model parametresini MMLE olarak adlandırılan yeni bir tahmin yöntemi ile tahminliyoruz. Tüm bulguları simüle edilmiş ve gerçek verilerle değerlendirip, sistemin hesaplama zamanı yanısıra doğruluklarınıda hesaplıyoruz.

Anahtar Kelimeler: Gauss grafiksel modelleme, Bernstein polinomları, çok değişkenli

uyarlamalı regresyon splineları (MARS), modifiye edilmiş ençok olabilirlik yaklaşımı (MMLE) *to my beautiful country*

ACKNOWLEDGMENTS

I would like to thank my supervisor Vilda Purutçuoğlu for her constant support and guidance. Her great experience, guidance, advices, and feedbacks as well as her contributions to my academic career have turned this study to an invaluable learning experience for me.

I would like to thank my examining committee members Prof. Dr. Gerhard-Wilhelm Weber and Prof. Dr. Barış Sürücü for their detailed reviews and their constructive comments.

I owe my special thanks to my friends Onur Çamlı, Gül Inan and Ezgi Ayyıldız for their full support, motivation and nice friendship. Also, I would like to thank all the members of METU Statistics Department.

I give my special thanks to my wonderful mother, Feriha, father, Hüseyin and brother, Fatih Nafi for their endless and unconditional devotion, encourage and supports throughout my life. Finally, I would like to thank specially to my wife Gülhayat, for her professional and personal support.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xiv

LIST OF FIGURES xvii

LIST OF ABBREVIATIONS xx

CHAPTERS

- 1 INTRODUCTION 1
- 2 PROPOSED METHOD AND BACKGROUND 7
  - 2.1 Graphical Models 7
    - 2.1.1 Gaussian Graphical Model 10
    - 2.1.2 Estimation 12
      - 2.1.2.1 Maximum Likelihood Estimation . . . 12
      - 2.1.2.2 Shrinkage Method 14
      - 2.1.2.3 The Least Absolute Shrinkage Approach (LASSO) 15

2.1.2.4 Graphical Lasso (L_1-penalized Method) 16

2.1.2.5 Neighborhood Selection with the Lasso Approach 17

2.2 Multivariate Adaptive Regression Splines and Proposed Method LMARS 18

2.2.0.1 Spline Functions 18

2.2.1 Multivariate Adaptive Regression Splines 19

2.2.2 Development of LMARS 24

2.2.3 LMARS with Interaction Effect 26

2.3 Modified Maximum-Likelihood Estimation 28

2.3.1 Development of Modified Maximum-Likelihood Estimation 32

2.4 Bernstein Operators 33

2.4.1 Bernstein Polynomials 33

2.4.2 Szász-Mirakyan Operator 36

2.4.3 Bleiman-Butzer-Hahn Operator 37

2.4.4 Meyer-König and Zeller Operator 39

2.5 Copula Method and Network Structure 39

2.5.1 Copula Method 39

2.5.1.1 Sklar's Theorem 40

2.5.1.2 Copula Families 42

2.5.1.3 Gaussian Copula 42

2.5.1.4 Students-t Copula 43

2.5.2 Network Structure 43

2.5.2.1 Scalee-Free Network 43

2.5.2.2 Random Network 44

2.5.2.3 Hubs Network 45

2.5.2.4 Cluster Network 45

2.6 Model Selection Criteria 47

2.6.1 Accuracy Measures 47

3 APPLICATION 49

3.1 LMARS Application 49

3.2 LMARS with Interaction 58

3.3 Modified Maximum-Likelihood Estimation Application . . . 68

3.4 Bernstein Operators 70

3.4.1 Application via Bernstein Polynomial Application 70

3.4.2 Application via Bernstein and LMARS Approaches 73

3.4.3 Application via Bernstein-Types of Operators . . . 77

4 CONCLUSION 81

REFERENCES 87

APPENDICES

A APPENDICES 99

CURRICULUM VITAE 103

LIST OF TABLES

TABLES

Table 2.1 Representation of a system with 4 nodes without interactions by LMARS. 24

Table 2.2 Representation of a system with 4 nodes without and with interactions by LMARS. 27

Table 2.3 Confusion matrix of the accuracy measures. 47

Table 3.1 Comparison of the specificity, precision and F-measure computed via LMARS under both AND and OR rules, and GGM and 1000 Monte-Carlo runs based on different dimensional Θ matrices p with normally distributed data with sample size 20 and plans (S_1: scenario 1, S_2: scenario 2). 51

Table 3.2 Comparison of the specificity, precision and F-measure computed via LMARS under AND rule, and GGM and 1000 Monte-Carlo runs based on different dimensional Θ matrices p with student-t distributed data at the degrees of freedom 3 with sample size 20 and plans (S_1: scenario 1, S_2: scenario 2). 52

Table 3.3 Comparison of the specificity, precision and F-measure computed via LMARS under AND rule, and GGM and 1000 Monte-Carlo runs based on different dimensional Θ matrices p with student-t distributed data at the degrees of freedom 4 with sample size 20 and plans (S_1: scenario 1, S_2: scenario 2). 53

Table 3.4 Comparison of the specificity, precision and F-measure computed via LMARS under AND rule, and GGM and 1000 Monte-Carlo runs based on different dimensional Θ matrices p with student-t distributed data at the degrees of freedom 7 with sample size 20 and plans (S_1: scenario 1, S_2: scenario 2). 53

Table 3.5 Comparison of the specificity, precision and F-measure computed via LMARS under AND rule, and GGM and 1000 Monte-Carlo runs based on different dimensional Θ matrices p with student-t distributed data at the degrees of freedom 15 with sample size 20 and plans (S_1: scenario 1, S_2: scenario 2). 54

Table 3.6 Comparison of the real computational time per second calculated via LMARS and GGM under 1000 Monte-Carlo runs based on different dimensional matrices and normally distributed data in scenario 1. 54

Table 3.7 Monte-Carlo comparison of the specificity, precision and F-measure of different dimensional biologically systems computed via GGM and LMARS models under exponential and normal marginals bounded by the Gaussian copula with sample size 20. 60

Table 3.8 Monte-Carlo comparison of the specificity, precision and F-measures of different dimensional biological systems computed via GGM and LMARS models under exponentials marginals bounded by the Gaussian copula with sample size 20. 61

Table 3.9 Monte-Carlo comparison of the specificity, precision and F-measures of different dimensional biological systems computed via GGM and LMARS under multivariate normal distribution with sample size 20. 61

Table 3.10 Comparison of the estimated links via LMARS with/without interaction effects and GGM with approximate glasso. The true links are taken from the study of Bhadra and Mallick [22]. 67

Table 3.11 Comparison of the specificity, precision and F-measure computed via MMLE and GGM under 1000 Monte-Carlo runs based on different dimensional system (Θ) with normally distributed data with sample size 20. 68

Table 3.12 Comparison of the real computational time calculated via MMLE and GGM under 1000 Monte-Carlo runs based on different dimensional systems (p) and normally distributed data. 69

Table 3.13 Comparison of the specificity, precision, F-measure and MCC computed via MMLE and GGM under 1000 Monte-Carlo runs based on different dimensional system (p) with lognormal distributed data with sample size 20. 69

Table 3.14 Comparison of the specificity, precision, F-measure and MCC computed via MMLE and GGM under 1000 Monte-Carlo runs based on different dimensional system (p) with student-t distributed data with sample size 20. 69

Table 3.15 Comparison of the CPU times per second for the simulated datasets generated by 1000 Monte-Carlo runs and scale-free network types. 71

Table 3.16 Comparison of the specificity (Spec), precision (Prec) and F-measure (F) computed via only GGM, Bernstein with GGM and Szasz with GGM under scale-free, hubs, cluster and random networks based on 1000 Monte-Carlo runs for the sample size 20 per gene. 73

Table 3.17 Comparison of the estimated links via transformed and non-transformed human gene expression dataset (Dataset 2). The biologically validated links and the complete name of genes can be found in the study of Bhadra and Mallick [22]. The complete list of genes is given in Appendix. 75

Table 3.18 Comparison of the precision and F-measure value via LMARS under 1000 Monte-Carlo runs based on systems with 500, 900 and 1000 dimensional networks for the sample size 20 per each gene. 76

Table 3.19 Comparison of the precision and F-measure values via GGM estimated by the neighborhood selection (NS) and glasso methods under 1000 Monte-Carlo runs based on 500, 900 and 1000 dimensional Θ with sample size 20. 76

Table 3.20 Comparison of the F-measure and the precision values computed with/without operators in inference of Θ via GGM with 0.9 off-diagonal entries under scale-free networks via sample size 20. 78

Table 3.21 Comparison of the F-measure and the precision values computed with/without operators in inference of Θ via GGM with 0.7 off-diagonal entries under scale-free networks via sample size 20. 78

Table 3.22 Comparison of the F-measure and the precision values computed with/without operators in inference of Θ via GGM with 0.5 off-diagonal entries under scale-free networks via sample size 20. 78

Table 3.23 Comparison of the F-measure and the precision values computed with/without operators in inference of Θ via GGM with 0.9 off-diagonal entries under hubs networks via sample size 20. 79

Table A.1 Protein list of the generated dataset of JAK/STAT pathway [92]. . . 101

LIST OF FIGURES

FIGURES

Figure 2.1 Simple representation of a network with 28 nodes via an undirected graph. 8

Figure 2.2 Representation of (a) an undirected graph and (b) the adjacency matrix of this graph. 9

Figure 2.3 Representation of (a) a directed graph and (b) the adjacency matrix of this graph. 9

Figure 2.4 Representation of the conditional independence between A, B and C nodes for (a) directed and (b) undirected graphs and (c) the independence between A and B once C is known. 9

Figure 2.5 A 0-degree spline of 5 knots. 18

Figure 2.6 A 1-degree spline of 6 knots. 19

Figure 2.7 Simple representation of the smoothing method for the curvature structure via BFs of MARS with t knots. 21

Figure 2.8 Comparison of the linear regression method and MARS on the same data. 22

Figure 2.9 Estimated network for a system with 4 genes represented in Table 2.1 by the LMARS without interaction model. 25

Figure 2.10 Estimated network for a system with 4 genes represented in Table 2.2 by LMARS (a) without interaction and (b) with interaction effects. . . . 28

Figure 2.11 Bernstein basis function for the first order. 35

Figure 2.12 Bernstein basis function for the second order. 36

Figure 2.13 An example of a scale-free network with 50 nodes. 44

Figure 2.14 An example of random network with 50 nodes. 45

Figure 2.15 An example of a hub network with 50 nodes. 46

Figure 2.16 An example of a cluster network with 50 nodes. 46

Figure 3.1 Simple illustration of the elements for the JAK-STAT pathway. . . . 56

Figure 3.2 Structure of the JAK-STAT pathway via (a) true quasi precision, (b) estimated precision via LMARS and (c) estimated precision via GGM under the glasso approach. 57

Figure 3.3 Estimated links of the KSHV pathway via (a) LMARS and (b) GGM under the glasso approach. 58

Figure 3.4 True graphical representation of the cell signalling network (Dataset 1) from Sachs et al. [117]. 63

Figure 3.5 Estimated structure of the real biological network from Dataset 1 via (a) LMARS without interactions, (b) LMARS with interactions and (c) GGM. 64

Figure 3.6 Estimated structure of the human gene interaction network from Dataset 2 by LMARS without interaction effects. 65

Figure 3.7 Estimated structure of the human gene interaction network from Dataset 2 by LMARS with interaction effects. 66

Figure 3.8 Estimated structure of the human gene interaction network from Dataset 2 by GGM. 66

Figure 3.9 (a) Quasi true network of the cell signalling data (Dataset 1), (b) estimated cell signalling data by only the GGM appraoch, (c) estimated network by Bernstein polynomial with the GGM approach and (d)estimated network by Szasz polynomial with the GGM approach. 72

Figure 3.10 (a) Estimated human gene expression data by only GGM appraoch (Dataset 2) and (b) Bernstein polynomial with the GGM approach and under the StARS criterion. 74

Figure A.1 Changes in the number of molecules for proteins P1, P4, P16, P17, P18, P23, P25, P28, P29 and P38. 99

Figure A.2 Changes in the number of molecules for proteins P2, P5, P7, P10, P11, P12, P15, P20, P21 and P35. 99

Figure A.3 Changes in the number of molecules for proteins P3, P6, P13, P19, P31, P32 and P34. 100

Figure A.4 Changes in the number of molecules for proteins P8, P9, P14, P22, P24, P26, P27, P30, P33 and P36. 100

LIST OF ABBREVIATIONS

AIC	Akaike information criteria
BBH	Bleiman-Butzer Hahn operator
BFs	Basis functions
BIC	Bayesian information criteria
CMARS	Canonical Multivariate Adaptive Regression Splines
CME	Chemical master equation
CPU	Central processing unit
CQP	Conic quadratic programming
FN	False negative
FP	False positive
GCV	Generalized cross validation
GELATO	Graph estimation with LASSO thresholding
GGM	Gaussian Graphical Model
GLASSO	Graphical LASSO
HAS	Hybrid adaptive splines
IFN	Type-I interferons
JAK-STAT	Janus kinase/signal transducer and activator of
KSHV	Kaposi's sarcoma associated herpesvirus
LASSO	Least absolute shrinkage and selection operator
LMARS	Lasso Multivariate Adaptive Regression Splines
LR	Logistic regression
LTS	Long-tailed symmetric
MARS	Multivariate Adaptive Regression Splines
MCC	Matthews Correlation Coefficient
MKZ	Meyer-König and Zeller operator
MLE	Maximum Likelihood Estimation
MMLE	Modified Maximum Likelihood Estimation
MPSS	Minimized penalized sum of square

mRNA	Messenger ribonucleic acid
MSE	Mean squared error
NB	Neighborhood selection
ODE	Ordinary differential equations
OLS	Ordinary least squares
PRSS	Penalized residual sum of squares
PSICOV	Prediction using sparse inverse covariance estimation
RCMARS	Robust Conic Multivariate Adaptive Regression Splines
RIC	Rotation information criterion
RMARS	Robust Multivariate Adaptive Regression Splines
SARS	Spatially adaptive regression splines
SCAD	Smoothly clipped absolute deviation
SNPs	Single nucleotide polymorphisms
StARS	Stability approach to regularization selection
STD	Short-tailed distribution
TGDA	Threshold Gradient Descent Approach
TN	True negative
TP	True positive
UTR	Untranslated region

CHAPTER 1

INTRODUCTION

In systems biology, the complex structure of an organism can be understood by means of the interaction of its components. The network is preferred to represent the system of the underlying interactions. In order to construct the network, the mathematical modeling is applied as it enables us to predict the future behavior of the structure and to describe the current relation between their components. The description of the biochemical activations via networks and the mathematical modelling are very powerful approaches to know the actual manner of the biological procedure and to present the structure of the complex systems. There are different levels to present biochemical events. The protein-protein interaction networks and metabolic networks are the two well-known representations. Here, we deal with the former type of networks which aims to explain the functional/physical interactions between proteins. This biological network can be modelled by different techniques under distinct assumptions. As examples of such models, we can list Boolean models [68, 120], Gaussian graphical models (GGM) [77] and stochastic models [59, 149]. In addition, there are various linear and nonlinear modelling approaches to present the biological systems. For instance, the modelling via the kinetic logic implements partial, possibly, linear functions in the representation [25] or the diffusion approximation in stochastic models performs a nonlinear expression to describe the biological events. All these approaches that are based on distinct regression functions can be mainly divided into two parts. These are parametric and nonparametric methods. In this study, we mainly deal with a parametric model GGM and suggest certain parametric pre-processing approaches to the raw data of GGM. Then, we propose another parametric inference method for this model. Later, we develop nonparametric alternatives for GGM too.

In the graph theory, the network structure is modelled by the nodes and edges. The former denotes the elements of the systems which can be proteins, genes or other species of the system in the protein-protein interaction networks and the latter represents the interactions between the system's elements. Hereby, the graphical models explain such structures under the concept of the conditional independency [148].

The GGM, which is related with the graph theory, is one of the well-known methods to construct the structure of the network which is a deterministic and undirected statistical model in explaining complex biolochemical networks [148, 130] under the multivariate normally distributed random variables whose dependency structure is represented by a graph [34]. GGM assumes that the dataset comes from the p-dimensional multivariate normal distribution with a mean vector μ and a covariance matrix Σ, and the interactions in the systems are described by the precision matrix Θ, which is the inverse of the covariance matrix, i.e., $\Theta = \Sigma^{-1}$. Hereby, as long as Σ is nonsingular, a set of nodes and edges that is conditionally independent can be combined to construct a graphical model. This sort of independence is shown via a zero entry in the precision.

In the estimation of such sparse and undirected graphical models, Meinshaussen and Bühlmann [93] propose a lasso idea for every variable in the system by using the others as explanatory variables. In the estimation of the lasso model, an optimization procedure based on a convex function is performed consecutively for each node in the graph. On the other hand, the penalty constant which adjusts the sparsity of the system is chosen via a probability of falsely connecting two or more different connectivity components at very low levels. Accordingly, the zero coefficient indicates the conditional independence between corresponding genes. From the application of this model in different dimensional systems, it has been shown that it overperforms both in terms of the accuracy and the computationally demand regarding the forward selection of the maximum likelihood estimation (MLE) approach in GGM. Whereas, while the dimension of the graph increases, there may be the problem of the non-symmetric estimated precision. Hereby, Friedman et al. [50] suggest a blockwise coordinate descent method in solving the lasso regression. Additionally, Witten et al. [150] propose the block update of the precision elements into the lasso model to get a symmetric precision, and Li and Gui [85] consider the threshold gradient descent approach (TGDA) to find the entries of the precision by decreasing the time of calculation and by getting a symmetric matrix in this model. Among these alternatives, TGDA is also implemented in different dimensional networks by comparing the central processing time (CPU) of the estimation [111]. From the analyses, although certain promising results are obtained regarding the accuracy and the computational cost, the calculation is still demanding, in particular, for large systems.

On the other hand, the multivariate adaptive regression splines (MARS) approach is a well-known statistically nonparametric regression methods that enables us to model the high dimensional data under nonlinearity [46]. Also, it is a particular type of optimization techniques [16, 17, 18] in the sense that MARS aims to transform a non-differentiable problem into a smooth problem [16] by putting binary constrains for the approximation of the optimal value [18]. Hereby, it uses the gradient based schemes to solve the smooth and nonsmooth optimization problems [17]. In order to estimate the model. The forward stage constructs the possibly large model with basis functions (BFs) and the backward elimination reduces the model complexity to get the optimal model.

The approximation theory is concerned with the study of how well given functions can be aproximated by basis functions. In this theory, it is usual to apply the approximating functions in the form of linear positive operators, such as the Bernstein, Szasz-Mirakyan polynomials, the Bleiman-Butzer-Hahn (BBH) operator, Meyer-König and Zeller (MKZ) operator. In biological networks, the complexity implies the large number of genes in a network whose interactions can be described by the scale-free feature [14]. Hence, linear positive operators enable us to transform the data in a new range [19, 89] by using the binomial and poisson distributions. The binomial distribution is the main distribution to describe the biological activation of any gene by means of the chemical master equation [151, 59] and the poisson distribution is just a limiting distribution of the binomial density which can be applicable for large systems [149]. Accordingly, the original data do not loose their biological interpretation when they are transformed by these polynomials since the suggested transformations are dependent on the underlying distributions. Hence, suggested transformations of the data are different from an ordinary standardization of the raw measurements in the range $[0, 1]$. Therefore, we propose to perform these two polynomials as a preprocessing step before any modelling and inference to eliminate batch effects. Because it is known that observations dependent on chemical master equations (CME) are described via the binomial distribution and any variation from these distributions can be caused by the nuisance effect, rather than the biological sources. On the other hand, in order to control whether other Bernstein type of operators such as BBH and MKZ operators [95] show the same advantages, we compare their precisions under GGM and MARS and we find that the Bernstein polynomials have the highest precision.

In the literature, Bernstein polynomials have been already applied in different areas. Voronovskaja [137] proves that the convergence of Bernstein polynomial to $f(x)$ function is bounded on $[0, 1]$ under the large sample size, i.e., $n \to \infty$. Stadtmuller [123] performs them to approximate the unknown regression functions. Vitale [136] uses them to produce the smooth density estimates, Tenbusch [129] developes its application in nonparametric regression functions, Petrone [109] implements it in a fully Bayesian setting, and Babu et al. [10] apply this method to approximate bounded and continuous density functions via its asymptotic properties. These properties are further investigated by Ghosal [58], Petrone and Wasserman [109] and Barrientos et al. [15]. Petrone [108] implements it in a fully Bayesian setting. Besides its implementation in mathematics, this method is also performed in various types of smoothing problems in statistics [10], the numerical analyses and the construction of the Bezier curve in mechanical engineering [20, 21]. Furthermore, Hoshek and Lasser [65] apply them in computer graphics.

On the other hand, the modified maximum likelihood estimation is first introduced by Tiku [131, 132] and Tiku and Suresh [133] to overcome the difficulties in the estimation of the maximum likelihood method when the normal equations derived from the likelihood function have multiple roots and nonlinearity. In this approach, the likelihood equations are obtained by ordering the variates and then linearized them by using the first-order Taylor series expansion. Hence, in this study, we propose the MMLE approach as an alternative of glasso when the states of the systems have the multivariate student-t distribution. Here, the selection of the student-t, in place of normality, also gives us the flexibility for non-normal states while its limiting distribution already covers the normal density.

There are many ways to explain dependency structure of the random variables. Copula is one of the efficient method in this field. Mathematically, the copula is a kind of function which connects the marginals into their multivariate distributions. Statistically, the copula is used to describe the dependency of the random variables. Copula first begins with the question of Frechret [44] whose elements is the set of common distribution functions of two random variables when the marginal distributions are known. Sklar [122] first describes the copula as the relationship between multivariate distribution functions and one dimensional margins of this distribution. According to Fisher (1997), copulas are useful for two reasons. Firstly, they are good functions to study scale-free measures of dependency and secondly, they can be a starting point for constructing bivariate distribution families. There are many copula families for the dependency structure.

In this study, we particularly use the Gaussian copula very often to generate measurements from different joint distribution functions in our analysis. These measurements are applied to evaluate the performance of suggested models based on MARS, processing step dependent on the Bernstein and Bernstein-types of operator and finally inference of the graphical model via the MMLE approach.

Accordingly, in the organization of the thesis, we present the general idea of the graphical model, GGM, MARS, our proposed model LMARS, Bernstein operators and the copulas, which we intensively implement in simulation studies in Section 2. In Section 3, we present our applications under a wide range of Monte Carlo scenarios and real biological datasets. Finally, we summarize our outputs and discuss the future work in Section 4.

CHAPTER 2

PROPOSED METHOD AND BACKGROUND

This chapter mainly presents the methods which we propose in this thesis and the topics associated with our propose methods. We organize this chapter with six sections. In Section 2.1, graphical models, Gaussian graphical models (GGM) and the estimation technique of GGM are defined in detail. The proposed methods, i.e., LMARS method and the development of the modified maximum likelihood estimation, are introduced in Section 2.2 and 2.3, respectively. The bernstein operators, which are used to smooth the data and are introduced as pre-procesing step in this study, is explained in Section 2.4. In Section 2.5, we define the copula method and the network structure, which are used in the data generation of the simulation studies in the application. Finally, the model selection criteria are explained in Section 2.6 since these values are used to check the accuracy of all our analyses.

2.1 Graphical Models

Graphical models are the most important part of our thesis study since all proposed methods are presented as a suggestion to the Gaussian graphical model (GGM) and improvements in its parameter estimation.

In general, the graph is the mathematical structure of the networks whose graphical points connect with the lines. In general, the graph (G) can be expressed in pair V and E, i.e., $G = (V, E)$, in which E represents the elements, known as edge, and V shows the link between pairs of two elements, known as vertices or nodes. When (x, y) shows an edge between two nodes, it means that x is adjacent to y and y is adjacent to x. Therefore, the adjacency matrix becomes important to describe the

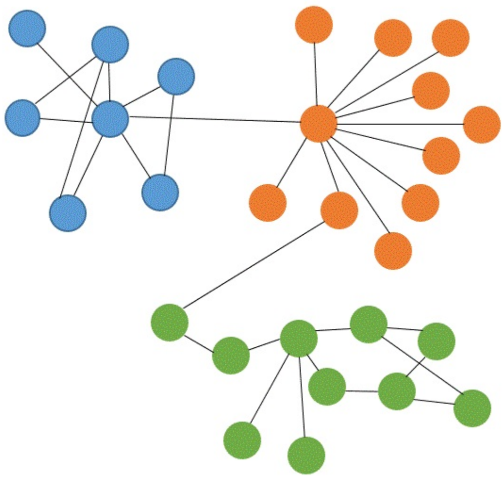


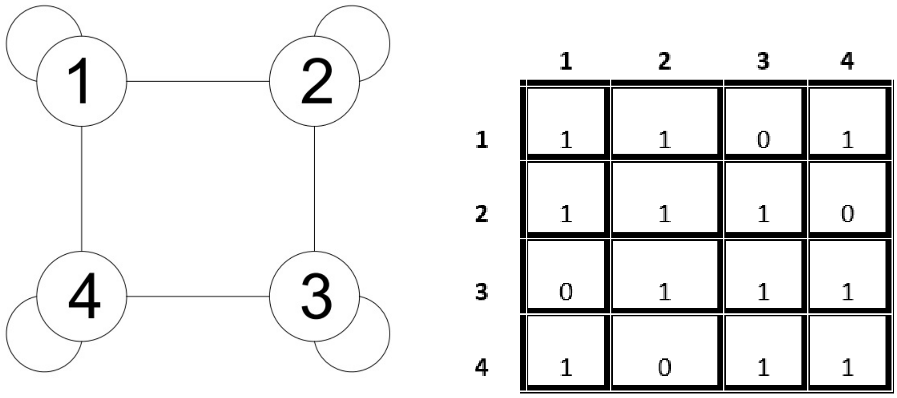
Figure 2.1: Simple representation of a network with 28 nodes via an undirected graph.

graph. The adjacency matrix is and $(n \times n)$ square and binary matrix which includes 0 and 1, and if (i, j) entry is 1, it implies that there is a connection between the ith and the jth nodes, and if there is 0, it refers that there is no connection between two nodes.

The graphical models [82] as shown in Figure 2.1 for illustration is a popular tool for the sparse structure and supplies a graphical representation of random variables under their conditional independencies.

There are two kinds of graph which are mostly used in graphical models. These graphs are directed and undirected graphes. The directed graph is a graph whose nodes are oriented with edges and the undirected graph is a set of objects whose edges are non-directional. These two types of graphes are represented in Figure2.2 and 2.3, respectively.

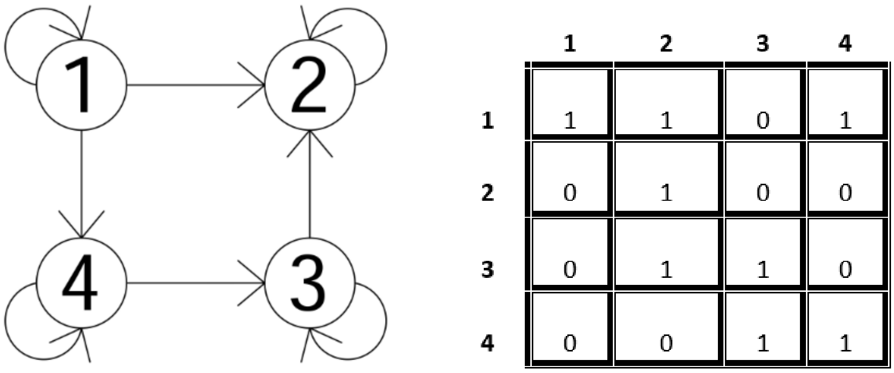
As stated before, lack of links in the undirected graph leads us to apply the rule of the conditional independence. Accordingly, when C is given, A and B are conditionally independent, and that can be represented as $A \perp B \setminus C$ and it means that for each value of C, A and B are conditionally independent on the distribution of C. In an example for three nodes A, B and C, it is obvious that A and B are conditionally independent on C when the structure of A and B is separated by C as seen in Figure 2.4. In gene networks, if two nodes, i.e., genes, are conditionally independent, there is no edge between two nodes and if is represented with a zero entry in the precision matrix.



	1	2	3	4
1	1	1	0	1
2	1	1	1	0
3	0	1	1	1
4	1	0	1	1

(a) (b)

Figure 2.2: Representation of (a) an undirected graph and (b) the adjacency matrix of this graph.



	1	2	3	4
1	1	1	0	1
2	0	1	0	0
3	0	1	1	0
4	0	0	1	1

(a) (b)

Figure 2.3: Representation of (a) a directed graph and (b) the adjacency matrix of this graph.

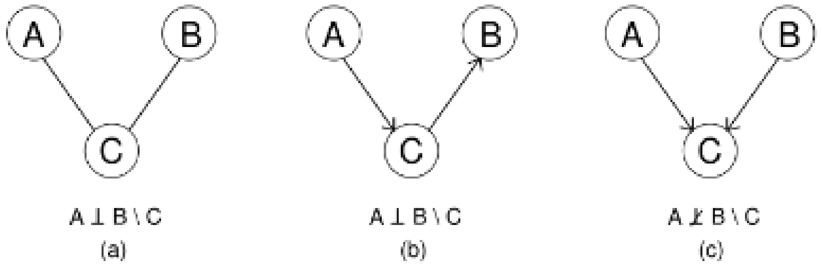


Figure 2.4: Representation of the conditional independence between A, B and C nodes for (a) directed and (b) undirected graphs and (c) the independence between A and B once C is known.

2.1.1 Gaussian Graphical Model

The Gaussian graphical model (GGM) is one of the famous undirected modelling approaches, which shows the graphical interactions over a set of random variables that represents conditional independencies under the multivariate normal distribution. GGM is firstly used in the literature under the name of the covariance selection models by Dempster [34]. But the graphical representation of these models is firstly introduced by Whittaker [148]. Briefly, this modelling is a parametric method in the sense that the states of the system in every time point are thought as the multivariate normal distribution. Moreover, it uses the inverse of the covariance matrix, i.e., precision matrix Θ, to explain the relations between genes. In this method, each node, i.e., protein or species, of the system is regressed by other remaining nodes in such a way that the coefficients of the regression function indicate the conditionally dependent structure between the species [85, 125]. Wermuth [147] also shows that the conditional independence corresponds to nonzero entries in the precision matrix and the zero entries in the inverse of the variance-covariance matrix stand for no interaction between the associated genes. On conclusion from the analyses, it is shown that GGM is successful in modelling the genomic interactions and the estimation of these systems can be applicable by several methods. Yuan and Lin [153] propose a penalized likelihood method to estimate Θ, Banerjee et al. [12] suggest two new algorithms, namely, block coordinate descent algorithm and the Nesterov's first order method. Then, Drton et al. [39] provide a model selection approach, called SINful, to control the overall error rate for incorrect edges in the estimated system, Ravikumar et al. [113] describe a novel method in which the neighborhood of any given node is estimated by a logistic regression based on the L_1-norm, and Augugliaro et al. [7] consider the generalized linear model to increase the accuracy with a low computational cost when the system is sparse. Besides these methods, most familiar approaches can be represented as the methods suggested by Meinshausen and Bühlmann [93] and Friedman et al. [50]. Meinshausen and Bühlmann [93] introduce the NS method within the lasso regression and Friedman et al. [50] suggest the graphical lasso, shortly glasso, approach which is based on the penalized likelihood idea with a penalty term λ to conduct the L_1-norm of the regression coefficients β as shown in Equation 2.14.

The former is simply based on an optimization procedure whose optimality is found via a probability of falsely connecting two or more different connectivity components at very low levels. Whereas, it can produce the non-symmetric precision matrix in the estimation of high dimensional systems. In order to solve the challenge, Witten et al. [150] suggest the block updating of the system, and Li and Gui [85] propose the threshold gradient descending method for the estimates of the precision. On the other hand, the latter applies a block-wise coordinate descent method for solving the lasso regression fitted to GGM.

GGM is used in variety of fields such as modelling the uncertainty in the macroeconomic growth [37], efficiently sampling from the Gaussian-Markov random field [86], and constructing the gene regulatory network [146].

GGM assumes that the dataset has the multivariate normal distribution and the interactions in the systems are described by the precision matrix, Θ, i.e., $\Theta = \Sigma^{-1}$. Therefore in GGM, the multivariate normally distributed nodes can be formulated as $Y = (Y^1, Y^2, ..., Y^p)$ via

$$Y \sim N(\mu, \Sigma), \tag{2.1}$$

where μ is a p-dimensional vector with $\mu = (\mu_1, \mu_2, ..., \mu_p)$ and Σ is a $(p \times p)$-dimensional covariance matrix. So the probability distribution function of Y can be presented by

$$f(Y) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(Y-\mu)'\Sigma^{-1}(Y-\mu)} \tag{2.2}$$

in which Y describes a multivariate normally distributed variable, μ refers to a mean vector and Σ is the variance-covariance matrix as stated beforehand. Finally, $|.|$ denotes the determinant of the given matrix. On the other hand, as stated beforehand, the precision is the inverse of the covariance matrix Σ, denoted by Θ with a $(p \times p)$-dimensional matrix. Thus, the pairwise dependency between two nodes i and j can be shown by $\theta_{i,j}$ as follows.

$$\Theta = \Sigma^{-1} = \theta_{ij}. \tag{2.3}$$

As the precision matrix consists of partial covariances, its diagonal entries are obtained from $\theta_{ii} = 1/var(Y^{(i)}|Y^1, Y^2, \ldots, Y^{i-1}, Y^i, \ldots, Y^p)$ and the partial correlation between j and k, denoted by ρ_{ij} is defined as

$$\rho_{ij} = \frac{-\theta_{ij}}{\sqrt{\theta_{ii}\theta_{ij}}} \quad \forall i \neq j, \tag{2.4}$$

where θ_{ij} shows the partial correlation between Y^i and Y^j and $var(.)$ denotes the variance term in the given random variable. Equation 2.4 means that

$$(i,j) \text{ and } (j,i) \notin E \iff \Sigma_{i,j}^{-1} = 0 \iff \rho_{ij} = 0.$$

In gene networks, the number of nodes p is greater than the number of observations n, i.e., $p >> n$, that leads to the singularity problem. In other words, the estimated sample covariance matrix S is not invertible. There are some approaches to estimate the model parameters of GGM. When the dimension number is less then the number of observation i.e., $p < n$, one can estimate θ by the maximum likelihood estimation (MLE) easily. On the other hand, if $n < p$, then the singularity problem can occur. In order to overcome this challenge, the L_1-penalized method and the NS with the lasso approach are the two well-known approaches. The mathematical details of these techniques are presented as below.

2.1.2 Estimation

2.1.2.1 Maximum Likelihood Estimation

$Y = (Y^1, Y^2, ..., Y^p)$ is the multivaraite Gaussian distributed node vector and the distribuion of $f(Y)$ is denoted in Equation 2.1. So the likelihood function of this distribution can be written as,

$$\begin{aligned}L(\Sigma|Y_1, Y_2, \ldots, Y_n) &= f(Y_1; \Sigma) f(Y_2; \Sigma) \ldots f(Y^p; \Sigma) \\ &= (2\pi)^{-np/2} |\Sigma|^{-\frac{n}{2}} \text{exp}(-\frac{1}{2} \sum_{i=1}^{n} Y_i^{'} \Sigma^{-1} Y_i).\end{aligned}$$

We can take the logarithm for maximizing the likelihood function by

$$L(\Sigma|Y_i) = l(\Sigma|Y) = -\frac{np}{2} \text{log}(2\pi) - \frac{n}{2} \text{log}(|\Sigma|) - \frac{1}{2} \sum_{i=1}^{n} Y_i^{'} \Sigma^{-1} Y_i.$$

In the above expression, $Y_i^{'} \Sigma^{-1} Y_i$ is scalar, so we can write $Y_i^{'} \Sigma^{-1} Y_i = \text{tr}(Y_i^{'} \Sigma^{-1} Y_i)$ and we can use properties of the trace matrix via

$$\begin{aligned}\sum_{i=1}^{n} \text{tr}(Y_i^{'} \Sigma^{-1} Y_i) &= \frac{n}{2} \text{tr}(\sum_{i=1}^{n} \frac{Y_i Y_i^{'}}{n} \Sigma^{-1}) \\ &= \frac{n}{2} \text{tr}(S \Sigma^{-1}),\end{aligned}$$

where $S = \sum_{i=1}^{n} \frac{Y_i Y_i^{'}}{n}$ and tr(.) is a *trace* matrix. This equation is written as the loglikelihood equation by

$$\begin{aligned}l(\Sigma|Y) &= -\frac{np}{2} \text{log}(2\pi) - \frac{n}{2} \text{log}(|\Sigma|) - \frac{n}{2} \text{tr}(S \Sigma^{-1}) \\ l(\Sigma|Y) &= -\frac{np}{2} \text{log}(2\pi) - \frac{n}{2} \big[\text{log}(|\Sigma|) + \text{tr}(S \Sigma^{-1}) \big].\end{aligned}$$

In these expressions, $-\frac{np}{2} \text{log}(2\pi)$ and $\frac{n}{2}$ are constant. Therefore, we maximize the remaining part on the right hand side via

$$\max_{\Sigma} \big[- \text{log}(|\Sigma|) - \text{tr}(S \Sigma^{-1}) \big]. \tag{2.5}$$

If we use $\Theta = \Sigma^{-1}$ and the properties of the logarithm in Equation 2.5, then the function becomes

$$\max_{\Theta} \big[\text{log}(|\Theta|) - \text{tr}(S \Theta) \big]. \tag{2.6}$$

Harville [62] shows the partial derivation of Equation 2.5 as follows.

$$\begin{aligned}\frac{\partial}{\partial \Theta}\big[\log(|\Theta|) - \text{tr}(S\Theta)\big] &= 0,\\ (\Theta^{-1})^{'} - S^{'} &= 0,\\ \Theta^{-1} &= S.\end{aligned}$$

Therefore, the maximum likelihood estimation of the precision matrix is calculated as $\Theta^{-1} = S$.

2.1.2.2 Shrinkage Method

The ordinary least squares (OLS) method is the basic method of estimation in linear regression models. For estimating unknown parameters of the model, it uses the minimization of the residual sum of squares $\sum_{i=1}^{n}(Y - X\beta)^2$. This equation can be derived by

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 = \arg\min_{\beta}(Y - X\beta)^{'}(Y - X\beta) \tag{2.7}$$

and the general β parameter estimation can be derived by $\hat{\beta} = (X^{'}X)^{-1}X^{'}Y$. Finding the parameters by the OLS method is easy. But when there is a singularity problem, OLS can not be applicable. The shrinkage method is suggested to solve this problem.

Hence, the OLS estimates can be replaced with a fairly smaller equality $\tilde{\beta}$ by

$$\tilde{\beta} = \frac{1}{1+\lambda}\hat{\beta} \tag{2.8}$$

and the minimized penalized sum of square (MPSS) can be presented via

$$\tilde{\beta} = \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2. \tag{2.9}$$

The solution of this MPSS equation can be described written as

$$\tilde{\beta} = (X^{'}X + \lambda I_p)^{-1}X^{'}Y, \tag{2.10}$$

where the parameter ($0 \geq \lambda$) controls the shrinkage. If $\lambda = 0$, MPSS turns to OLS and if λ is very big, MPSS can lead to an empty or a null model. The underlying regression, called the ridge regression, does not perform properly when the system is sparse. To handle this problem, some approaches are developed such as the least absolute shrinkage and the selection operator (LASSO) [130], the smoothly clipped absolute deviation (SCAD) [42], elastic net [159] and the NS method [93].

2.1.2.3 The Least Absolute Shrinkage Approach (LASSO)

Let us assume that Y is a vector and all the observed measurements are contained in Y. So a regression model can be constructed between a response variable Y^p and the explanatory variables Y^{-p}. Hereby, the model is described as

$$Y^p = Y^{-p}\beta + \varepsilon. \tag{2.11}$$

In this expression, ε is the error term which has a normal distribution with zero mean and β is the regression coefficient. Thus, the mean vector μ and the variance-covariance matrix Σ of the model in Equation 2.11 can be shown by

$$\mu = \begin{pmatrix} \mu_{-p} \\ \mu_p \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{-p,p} & \sigma_{-p,p} \\ \sigma_{-p,p} & \sigma_{p,p} \end{pmatrix}, \tag{2.12}$$

respectively. Here, μ_{-p} represent the mean vector of all nodes except the pth node, $\Sigma_{-p,p}$ is the $(p-1) \times (p-1)$-dimensional variance-covariance matrix of all nodes except the pth node, $\sigma_{-p,p}$ refers to a $(p-1)$-dimensional vector, and $\sigma_{p,p}$ is the covariance value of the pth node. Interestingly, in Equation 2.11, there is a relation between β and the precision matrix Θ which is formalized by

$$\beta = -\Theta_{-p,p}/\Theta_{p,p}. \tag{2.13}$$

In Equation2.13, the estimated interaction between two nodes in a system is explained by the associated entries of the precision matrix. In this expression, Y^p and Y^j ($j =$

$1, \ldots, p)$ are conditionally independent when $\beta = 0$ and the optimal estimate of β is defined via

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \Big\{ \frac{\|Y - Y^{-p}\beta\|_2^2}{n} + \lambda\|\beta\|_1 \Big\}. \tag{2.14}$$

In Equation 2.14, $||\cdot||_1^1$ and $||\cdot||_2^2$ stand for the L_1-norm and the L_2-norm of the given values, respectively. As seen in the objective function of β, the estimated parameters can be found by different optimization methods under the high dimensional and sparse θ.

2.1.2.4 Graphical Lasso (L_1-penalized Method)

One of the efficient ways to estimate a sparse and symmetric matrix Θ is the graphical lasso approach (glasso) which is introduced by Friedman et al. [50]. With respect to the Lagrangian dual form, the problem is the maximization of the loglikelihood function with respect to the nonnegative matrix as follows.

$$\max_{\Theta} \Bigg(\log(|\Theta|) - \text{tr}(S\Theta) \Bigg), \tag{2.15}$$

where $S = XX'/n$ is an estimated var-covariance matrix. Yuan and Lin [153] show that instead of maximizing Equation 2.15, the penalized loglikelihood function can be maximized via

$$\max_{\Theta} \Bigg\{ \log(|\Theta|) - \text{tr}(S\Theta) - \lambda||\Theta||_1 \Bigg\} \tag{2.16}$$

in which tr$(.)$ denotes the trace matrix as used before. $||\Theta||_1$ is the L_1-norm that is the summation of the absolute values of the entries in the precision matrix. According to the Karush-Kuhn-Tucker condition [150] to maximize Θ, Equation 2.16 must provide the following equation.

$$\Theta^{-1} - S - \lambda\Gamma(\Theta) = 0, \tag{2.17}$$

where $\Gamma(\Theta)$ shows the subgradient of $|\Theta|$ which means if $\Theta_{ij} > 0$, $\Gamma(\Theta_{i,j})$ equivalent to 1. If $\Theta_{ij} < 0$, $\Gamma(\Theta_{i,j})$ sets to -1 and $\Theta_{ij} = 0$.

A sufficient condition for solving the graphical lasso is to block the diagonal matrix with blocks if the inequality $S_{ii'} < \lambda$ is accomplished for all $i \in C_k$, $i' \in C_{k'}$ and $k \neq k'$, in which $C_1, C_2, ..., C_k$ satisfies a partition of p. Furthermore, $\widehat{\Theta}$ is a block diagonal matrix with k blocks by

$$\widehat{\Theta} = \begin{bmatrix} \Theta_1 & & & \\ & \Theta_2 & & \\ & & \ddots & \\ & & & \Theta_k \end{bmatrix}.$$

Here, the kth block of $\widehat{\Theta}$ satisfies Equation 2.15 and $\widetilde{\Theta}$ is estimated. From the findings, it is seen that the blocking idea is computationally efficient in inference.

2.1.2.5 Neighborhood Selection with the Lasso Approach

A popular alternative way to overcome the underlying singularity of the variance-covariance matrix is to apply the neighborhood selection (NS) with the lasso approach [93]. This method is computationally attractive for sparse and high dimensional graphes.

The NS method is a sub-problem of the covariance selection. If Φ is a set of nodes, the neighborhood of ne_p of the node $p \in \Phi$ is the smallest subset of $\Phi \backslash \{p\}$, which denotes the set of nodes except the pth node. So all variables Y^{ne_p} in the neighborhood, Y^p is conditionally independent on all remaining variables. The neighborhoods of the node p consist of the node $b \in \Phi \setminus \{p\}$ so that $(p, b) \in E$ when E defines the set of edges.

By this way, this method is transformed to the standard regression problem and it is efficiently solved by LASSO approach [130]. Hereby, the lasso estimate of Θ, i.e., $\hat{\Theta}$, for the pth node and under the penalty constant λ is given by

$$\widehat{\Theta}^{p,\lambda} = argmin\Big(||Y^p - Y\Theta||_2^2 + \lambda_p||\Theta||_1 \Big), \tag{2.18}$$

where $||\Theta||_1 = \sum_{b\in\Phi(n)} |\Theta_b|$ is the L_1-norm of the coefficient vector and $||.||_2^2$ shows the L_2-norm . But the solution of Equation 2.18 is not unique. Because each selection of the λ penalty determines the neighborhood ne_p of node $p \in \Phi(n)$.

2.2 Multivariate Adaptive Regression Splines and Proposed Method LMARS

Multivariate adaptive regression splines (MARS) is used to estimate the elements of the precision matrix in the proposed method, which we call as LMARS. Before describing the method which we suggest, let us explain the spline functions and MARS.

2.2.0.1 Spline Functions

The function obtained by combining polynomial piecewise linear or non-linear functions that satisfy certain smoothness conditions is called as the spline functions. If the smoothness is based on a constant entry, it is called as the spline of degree “0". But if it is generated by a linear equation, then it is named as the spline of degree “1". A simple example of the spline function for 0 and 1 is seen in Figures 2.5 and 2.6, respectively.

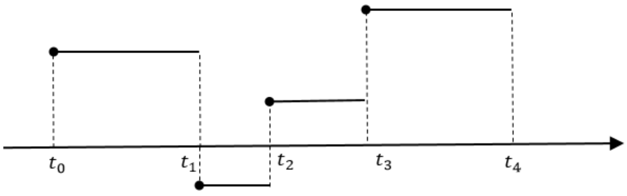


Figure 2.5: A 0-degree spline of 5 knots.

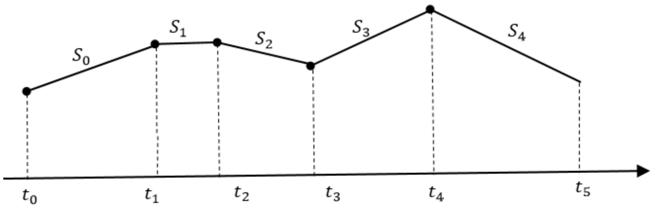


Figure 2.6: A 1-degree spline of 6 knots.

In Figures 2.5-2.6, we represent those two most common splines types for 5 and 6 knots, in order. Hereby, if t_i $(t_1, t_2, \ldots, t_n)$ is the knot and S_i $(S_1, S_2, \ldots, S_n)$ refers to the spline functions, the piecewise functions based on splines can be written as

$$f(n) = \begin{cases} S_0(x) = a_0 x + b, & \text{if } x \in [t_0, t_1) \\ S_1(x) = a_1 x + b, & \text{if } x \in [t_1, t_2) \\ S_2(x) = a_2 x + b, & \text{if } x \in [t_2, t_3) \\ & \vdots \\ S_n(x) = a_n x + b, & \text{if } x \in [t_{n-1}, t_n) \end{cases}$$

in which $S(x)$ is called as the piecewise linear knot.

2.2.1 Multivariate Adaptive Regression Splines

The multivariate adaptive regression splines (MARS) [47], which explains the relation between dependent and independent variables without any assumption have growing applications in many areas of the science over the last few years. As a development of the statistical methodology in this area, the projection pursuit method [46, 48] and the univariate additive version of MARS [49] are intensively studied by Friedman [49]. After the development of MARS, it is studied by many researchers since it creates an adaptable model for the high-dimensional, non-linear and highly-correlated data by introducing piecewise linear regressions. Psichogios et al. [110], Kuhnert and McClure [79] apply MARS as a nonparametric method, Chen [27] suggests quintic function to smooth truncated linear functions, Bakin et al. [11] use second order B-splines except the truncated functions. Munoz and Felicisimo [99] show that MARS combines the linear regression, truncated basis function and binary partitioning to construct the model. Lethwick et al. [84] explain that the predictor variable is divided into piecewise linear functions to describe the relation between dependent and independent variable. Tsai and Chan [135] present a more robust implementation of MARS to develop the stopping rule for the automatic detection of the spline functions in the optimal model. Then, Shih [119] proposes a convex version of MARS to reconstruct the basis functions, Hastie et al. [63] show that MARS is a very convenient modelling technique for high dimensional data. Recently, Weber et al. [144] propose a new method, namely CMARS, which performs a penalized residual sum of squares for the MARS backward algorithm.

The regression method is a statistical modelling technique which defines the relation between dependent (y) and independent (x) variables. Here, the linear regression method is applied to the dataset if the interrelationships between parameters are linear. On the other hand, the nonlinear regression method is implemented to explain nonlinear parametric relations between dependent and independent variables. Thus, the classical nonparametric model can be described as the following structure.

$$y_i = f(\beta, x_i^{'}) + \varepsilon, \tag{2.19}$$

where β is the unknown model parameter and x stands for the independent variable. Moreover, f denotes an unknown functional form, and ε denotes the error terms.

Accordingly, the MARS method affords to proximate the nonlinear functions of f by using piecewise linear basis elements, known as basis functions (BFs) [46]. BFs are used to build the model for each variable (x_{ij}) by the possible t value as a univariate knot and take a constant or a hinge function. Thus, the form of BFs can be shown as $(x-t)_+$ and $(t-x)_+$ in which x is an input variable on the positive side " $+$ ". So

$$(x-t)_+ = \begin{cases} x-t & \text{if } x > t \\ 0 & \text{otherwise} \end{cases}, (t-x)_+ = \begin{cases} t-x & \text{if } x < t \\ 0 & \text{otherwise} \end{cases}. \tag{2.20}$$

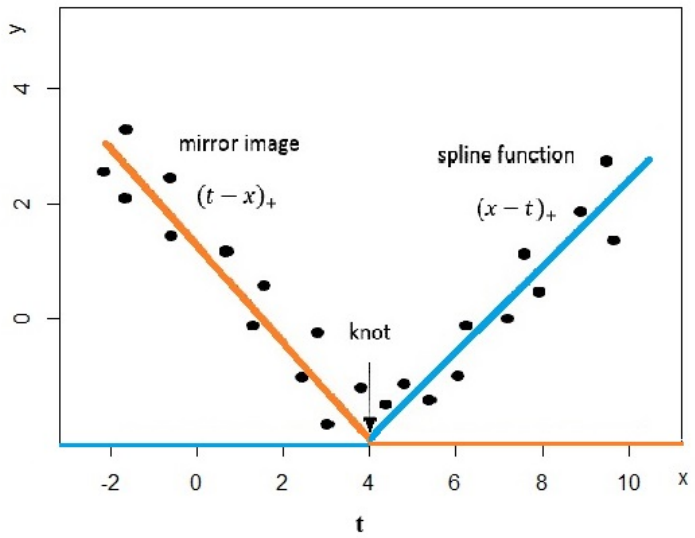


Figure 2.7: Simple representation of the smoothing method for the curvature structure via BFs of MARS with t knots.

In Equation 2.20, t is a single knot taken from the dataset simply shown in Figure 2.7.

MARS can produce two-piece linear models as shown in Figure 2.7 and 2.8 with such BFs. Figure 2.7 shows a mirrored pair, where each function is created piecewisely linear for the knot at the value 4 on the t-axis. These two functions are called the reflected pairs. Then this dataset can be modelled as Equation 2.21 below.

$$Y = c_0 + c_1 \times \max(0, X - 4) + c_2 \times \max(X - 4, 0) \qquad (2.21)$$

In general, MARS can model the nonlinearity better than the linear regression model. The difference can be seen in Figure 2.8. In this figure, the dashed red dots show the linear regression and the black line shows the MARS method fitted on the same data. Here, MARS explains the data via two lines and the linear model describes them by only one linear line. As seen from Figure 2.8, the MARS model fits the nonlinearity better by using piecewise linear BFs than the linear regression model. The purpose of these piecewise linearities is to construct the reflected pairs for every observed value x_{ij}. Therefore, BFs under ($i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, p$) is defined as

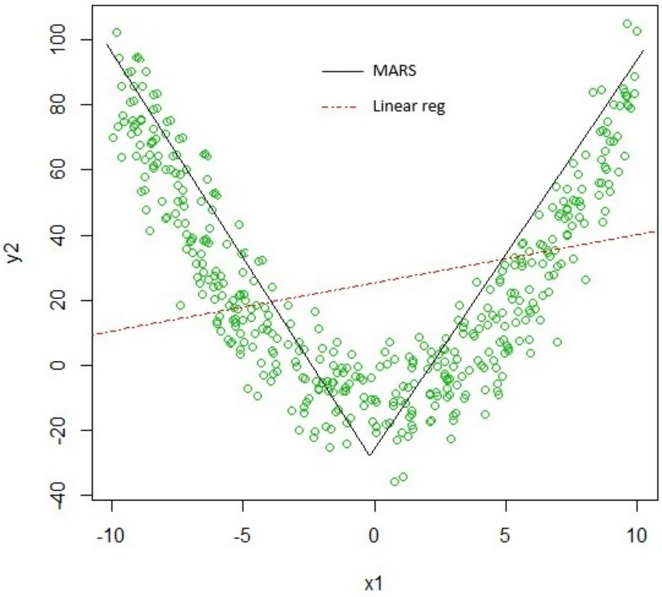


Figure 2.8: Comparison of the linear regression method and MARS on the same data.

$$C = \left\{ (X_j - t)_+, (t - X_j)_+ \} | t \in \{x_{1,j}, x_{2,j}, \dots, x_{N,j}, j \in \{1, 2, \dots, p\} \right\}, \quad (2.22)$$

where N denotes the number of observations and p shows the dimension of the input variables. If all of the input values are well-defined, we can construct $2Np$ basis functions.

The general method to produce the spline fitting in high dimensional systems is to use basis functions as tensor products of univariate spline functions. Hence, the multivariate spline BFs which can be seen in Equation 2.23 are performed as the mth BFs that are tensor products of the univariate spline functions.

$$B_m(x) = \prod_{k=1}^{K_m} (s_{km}(x_{v(km)} - t_{km}))_+^q, \quad (2.23)$$

in which K_m denotes the complete number of truncated linear functions in BF, s_{km} takes the value ∓ 1, $x_{v(km)}$ describes the input variables, and q is the order of splines. Moreover, t_{km} refers to the corresponding knot value and indicates the (right/left) sense of the combined step function. Furthermore, $v(km)$ identifies the predictor variable and t_{km} substitutes for values on the corresponding variable. Finally, $(.)_+$

indicates the partial function as described in Equation 2.20. If we adapt the MARS example in Figure 2.7 to Equation 2.23, t can take the value 4 and q can take 1. Accordingly, the construction of the modelling strategy is same as the forward stepwise linear regression. But different from this model, the functions in the C set are admitted to be used in MARS, instead of the original inputs. Therefore, the MARS model is represented by

$$f(x) = c_0 + \sum_{m=1}^{M} c_m B_m(X) + \varepsilon, \tag{2.24}$$

where $B_m(x)$ is a function of C as shown in Equation 2.22, $X = (X_1, X_2, ..., X_p)^{'}$, c_0 presents the intercepts, c_ms are the regression coefficient for each basis function, and it is estimated by minimizing the residual sum of squares in the linear regression model. Moreover, M denotes the number of basis functions and finally, ε corresponds to the uncorrelated random error term with a zero mean and an unknown constant variance. BFs can be described as the following forms:

1. The constant function, $B_1 = 1$.
2. BFs with the form $B_2 = X$, $B_3 = (x - t_1)_+$, $B_4 = (x - t_2)_+$.
3. BFs with the interaction effect $B_5 = (x - t_3)_+ \times (x - t_4)_+$.

The ultimate goal to construct the model produces a minimum number of BFs. To accomplish this, MARS performs both the forward selection and the backward elimination approach [46]:

1. In the forward selection, BFs are attached to the model and the largest model is obtained. Here, the system starts with a constant BF. Then, all possible BFs are attached to the model. Later, a possible large model that overfits the data is build. The model becomes full when M is maximum (M_{max}). At the end of this procedure, the model is overfitted and the backward elimination procedure is applied.

2. In the backward elimination, the method is reversed and the model is refined by reducing BFs that have no effects in the accuracy of the model. Here, the generalized cross validation (GCV) is actual backward fitting criterion for the

model selection [30] and the modified generalized cross validation (GCV*) is used in the MARS method.

Friedman [47] suggests to perform the underlying modified version of the form of the generalized cross validation criterion (GCV^*) as denoted in Equation 2.25 in order to choose the best model. GCV* produces the best fitted model $\hat{f}_\lambda$ of each size and λ is generated as the final step of the backward method.

$$\mathrm{GCV}^*(\lambda) = \frac{\sum_{i=1}^{N}(y_i - \hat{f}_\lambda(x_i))^2}{(1 - M(\lambda)/N)^2}, \tag{2.25}$$

where N represents the number of observations and $M(\lambda)$ is the effective number of parameters. In this equation, $M(\lambda)$ is found via $M(\lambda) = r + cK$ in which r refers to the number of linear independent basis functions, and K describes the number of selected knots during the forward stage. Additionally, c is the cost in the optimization of BF and the smoothing parameter of the model is generally taken as $c = 3$. However, the model is restricted to be additive on $c = 2$ [63]. Finally, y and $\hat{f}_\lambda$ show the response variable and the estimated f with the data y, respectively.

2.2.2 Development of LMARS

We can explain this system with an example. Assume that we have a system with 4 nodes and each node in this system is estimated by the following sets of lasso equations without interactions in Table 2.1.

Table 2.1: Representation of a system with 4 nodes without interactions by LMARS.

Description of System	Lasso Equations
LMARS without interactions	$y_1 = 2y_2$ $y_2 = y_1 + 5.6y_4$ $y_3 = 4.1 + 5y_2$

In the LMARS model without interactions, as the alternative of GGM, we construct a regression model for each node against all remaining nodes similar to the lasso regression as shown in Equation 2.11 and main effects are selected to show the relation between two genes. In Table 2.1 first row, y_1 is used as a response and the others

(y_2, y_3, y_4) as explanatory variables, and then, the model is constructed and the significant coefficient is taken, which is only y_2 in or toy example in Table 2.1. In this example, it means that y_1 has a relation with y_2. Finally, the binary adjacency matrix is obtained as seen below.

$$\Theta_{woi} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

Then, the graphical representation of this matrix, Θ_{woi}, can be drawn as in Figure 2.9.

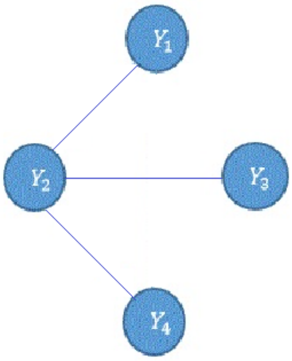


Figure 2.9: Estimated network for a system with 4 genes represented in Table 2.1 by the LMARS without interaction model.

In the development of LMARS, when we compare the performance of LMARS and GGM, we construct the MARS model similar to the lasso regression as shown in Equation 2.11. Therefore, we call this model as the lasso-based MARS, shortly LMARS, model. Accordingly, in order to detect links of a selected gene, we consider that this gene behaves as the predictor and the remainings are accepted as the explanatory variables. On the other side, the mean and the variance of the predicted gene given the rest can be written as the conditional mean and the variance, respectively [52].

Moreover, the mean vector and the covariance matrix of the system can be partitioned as presented in Equation 2.12, resulting in the same expression in Equation 2.13 for the regression coefficients of the MARS model. The calculation steps of the LMARS method used in this study can be also listed as below:

1. Construct the MARS model which contains only the main effects for each node against all remaining nodes as in Equation 2.11.
2. Apply the forward and backward steps of MARS, respectively.
3. Choose the best model which has the highest GCV.
4. Take significant regression coefficients and consider that there is a relation between genes as the predictor and significant covariates.
5. Construct the adjacency matrix where 1 refers to a relationship and 0 describes no interaction between the pairs of genes.
6. Repeat the steps 1 - 5 for all nodes.
7. Construct the adjacency matrix, separately.
8. Compare the estimated adjacency matrix with the true binary precision matrix for every cell.
9. Compute the accuracy measures.

2.2.3 LMARS with Interaction Effect

In this work, we enlarge the underlying LMARS model by adding the interaction effects of the systems' elements. In LMARS, as the alternative of GGM, we construct a regression model for every node against all other nodes similar to the lasso regression as explained in Subsection 2.2.2. But, we also include the second-order interaction terms in this model due to the fact that its physical structure can be thought as the feed-forward loop in the biological networks [4]. For example, assume that we have a system with 4 nodes and each node in this system is estimated by the following sets of lasso equations without interactions and with interactions by LMARS in Table 2.2.

Then, if we describe these sets of equations via the estimated edges in the system, we can report that the first node y_1 has connections with node 2, y_2 has edges with nodes 1 and 4, y_3 is autoregulated and y_4 is bounded with nodes 1 and 2 when the interaction

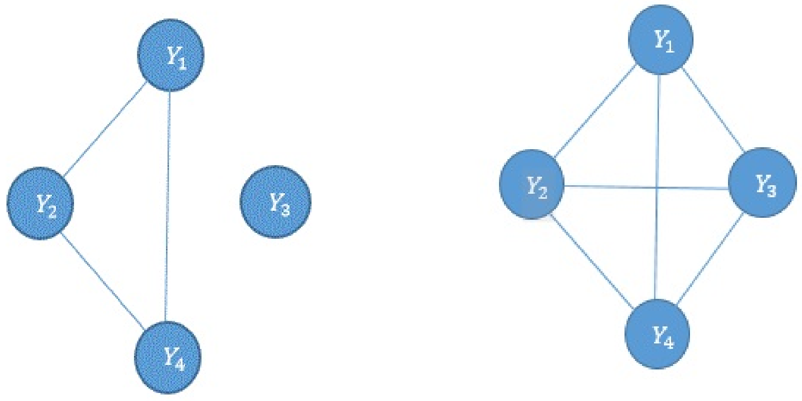
Table 2.2: Representation of a system with 4 nodes without and with interactions by LMARS.

Description of System	Lasso Equations
LMARS without interactions	$y_1 = 2y_2$ $y_2 = y_1 + y_4$ $y_3 = 4.5y_2$
LMARS with interactions	$y_1 = y_2 + 2y_3$ $y_2 = 3y_3 + y_4$ $y_3 = y_2 \times y_4$ $y_4 = y_1 \times 2y_3$

effect is not added into the model. Whereas, if it is included to the model, then, y_1 has connections with y_2, y_3 and y_4, y_2 is connected with y_3 and y_4, y_3 has edges with y_2 and y_4, and finally, y_4 is bounded with y_1 and y_3. Thereby, the associated adjacency matrix of LMARS without interactions (Θ_{woi}) and LMARS with interactions (Θ_{wi}) can be stated as below:

$$\Theta_{woi} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \Theta_{wi} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

where the columns and the rows show the name of the genes from 1 to 4, sequentially. In other words, we consider that the elements of interactions also imply the pairwise relations between each other and between the response, separately. We also represent the graphical view of these matrices in Figure 2.10.



(a) (b)

Figure 2.10: Estimated network for a system with 4 genes represented in Table 2.2 by LMARS (a) without interaction and (b) with interaction effects.

2.3 Modified Maximum-Likelihood Estimation

Maximum-likelihood estimation (MLE) is the most commonly used method in the parameter estimation. The modified maximum-likelihood estimation (MMLE), is another parameter estimation method which is asymptotically equivalent to MLE [131, 132, 133]. In this thesis, after obtaining successful results from the LMARS method, we focus on the MMLE method, where we can estimate the open form of the precision matrix by a likelihood–based approach.

Let consider a likelihood equation to estimate an unknown location parameter θ and think that we already compute its partial derivative with respect to each model parameter. Then we can get the following expression in general.

$$\frac{\text{dyln}L}{d\theta} = \frac{1}{\sigma}\sum_{i=1}^{n} g(z_i) = 0, \text{ for } z_i = \frac{y_i - \theta}{\sigma}, \tag{2.26}$$

where InL denotes the log-likelihood and σ represents the scale parameter for the random sample y_i $(i = 1, 2, \ldots, n)$. Accordingly, z_i implies the standardized form of y_i . Thus, if we assume that σ is known, we can arrange y_i $(1 < i < n)$ as the order statistics. Accepting that $t_{(i)} = E(z_{(i)})$ is the expected value of the ith standardized order variate z_i $(1 < i < n)$, we can expand $g(z_i)$ as a first-order Taylor series around $t_{(i)}$ and realize that g(z) in the interval $a < z < b$ is approximately linear [131]. Then, we can obtain a linear approximation of Equation 2.26 from the Taylor series expansion via

$$\begin{aligned} g(z_i) &= g(t_i) + (z_i - t_i)\frac{dg(z)}{dz} \\ &\simeq \alpha_i + \beta_i z_i \text{ for } 1 < i < n \end{aligned} \tag{2.27}$$

where $\beta_i = \frac{d}{dz_i} g(z)$ and $\alpha_i = g(t_{(i)}) - \beta_i t_{(i)}$.

If $g(z)$ in Equation 2.27 is bounded and $z_{(i)}$ $(1 < i < n)$ tends to its expected value $t_{(i)}$, then

$$g(z_i) - (\alpha_i + \beta_i z_i) \tag{2.28}$$

tends to zero as n approaches to infinity. Hence, the incooperation of Equation 2.26 and 2.27 gives the modified maximum-likelihood equation whose new partial derivative can be defined as belows:

$$\frac{\text{dylnL}}{d\theta} = \frac{1}{\sigma}\sum_{i=1}^{n}(\alpha_i + \beta_i z_i) = 0. \tag{2.29}$$

Since Equation 2.29 is linear in θ, it has an explicit and unique solution, called the modified maximum-likelihood estimator (MML) as presented in Equation 2.30.

$$\hat{\theta} = \frac{\sigma \sum\limits_{i=1}^{n} \alpha_i + \sum\limits_{i=1}^{n} \beta_i y_i}{n} \quad \text{for} \quad m = \sum_{i=1}^{n} \beta_i. \tag{2.30}$$

In Equation 2.30, α_i and β_i stand for the nonlinear functions repeated in the original partial derivation of the log-likelihood function and is described by the order statistics.

Common benefit of this estimation method is seen when the random variable has a non-normal density such as long-tailed symmetric or skewed distributions. In this study, we describe it under the long-tailed symmetric (LTS) distribution since LTS is subset of the heavy-tailed distributions whose tails decrease slower than exponential distributions and the exponential distributional families are generally known as the tails decrease suddenly because of the e^{-x} function. Moreover, these distributions cover the normal or Gaussian distribution, which is the most applicable distribution in systems biology as in GGM, apart from Cauchy and student-t distributions. On the other side, short-tailed distributions (STD) are other sorts of distributions whose tails decrease fast, but they are not cover the normal density. Accordingly, the LTS density can be presented as a scaled form via the student-t distribution at the degrees of freedom $(2p-1)$as shown in Equation2.32 . Hereby, considering that our parametric model for the biological systems has the following structure,

$$Y_i = \beta_0 + \sum_{i=1}^{q} \beta_i Y_{(-i)} + \varepsilon_i, \tag{2.31}$$

where β_0 is the intercept, β_i shows the slope of the ith state, Y_i denotes the state of the ith node which represents the response in the system, and $Y_{(-i)}$ describes the remaining nodes demonstrating the explanatory variable in the equation. Finally, ε_i stands for the error terms that come from the independent and identically distributed long-tailed symmetric distribution having the following expression for the univariate dimension:

$$f(x) = \frac{1}{\sigma\sqrt{k}\beta(\frac{1}{2}, p-\frac{1}{2})}(1+\frac{x^2}{k\sigma^2})^{-p} \tag{2.32}$$

in which $k = 2p-3$ $(p \geq 2)$ and for $p = \infty$, $f(x)$ reduces to the standard normal distribution, i.e., $N(0,1)$ while p represents the shape parameter which adjusts the distribution from Cauchy to normal. Moreover, $\beta(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ when $E(x) = 0$ and $V(x) = \sigma^2$ as the zero mean and constant variance. But, in the application of MMLE we construct the model as, $Y_i = \beta Y_l + \varepsilon_i$ $(l = i+1, i+2, \ldots, n$ and $i = 1, 2, \ldots, n$ also $i < l)$. Thus, for a random sample y_i, the likelihood function can be proportionally written as

$$L = \propto \left(\frac{1}{\sigma}\right)^N \prod_{i=1}^{a}\prod_{j=1}^{n}\left\{1+\frac{z_{i,j}}{k}\right\}^{-p}, \tag{2.33}$$

where N is the total sample size, i.e., $N = a.n$ in our equation, $z_{ij} = \varepsilon_i/\sigma$ $(i = 1, 2, \ldots, a)$ as the indicator of the number of random variable in which e_i is the error term, and denotes $\varepsilon = Y_i - \beta Y_l$ and finally, $j = 1, 2, \ldots, n$ as the indicator of the number of the observation for each random variable. Then, if we take the logarithm of this likelihood function, we can obtain

$$\text{lnL} =\propto -N\text{log}(-\sigma) - p\sum_{i=1}^{N}\text{ln}(1 + \frac{z^2}{k}) \tag{2.34}$$

as the equivalent of Equation 2.33. Later, if the derivative is taken for the unknown parameter, let's say σ, then the partial derivative of Equation 2.34 with respect to σ can be presented by

$$\frac{\partial lnL}{\partial\sigma} = \frac{-n}{\sigma} + \frac{2p}{k\sigma}\sum_{i=1}^{N} z_i g(z_i) = 0 \tag{2.35}$$

in which $g(z_i) = z/(1+z^2/k)$. Here, Equation 2.35 can be computationally challenging as it has multiple roots due to its nonlinearity. In order to unravel this problem, MMLE proposes to use the order statistics and the first order Taylor series expansion. Accordingly, in the calculation, firstly, the random variables are ordered as $y_{i(1)} \leq y_{i(2)\leq} \cdots \leq y_{i(n)}$ $(1 \leq i \leq a)$. Then, z_i is replaced with $z_{i(j)}$ by applying the following linear approximation:

$$\begin{aligned} g(z_{i(j)}) &\simeq g(t_{(i)}) + [z_{i(j)-t_{(i)}}]\left\{\frac{\partial g(z)}{\partial z}\right\} \\ &= \alpha + \beta z_{i(j)}, \end{aligned} \tag{2.36}$$

where $E(z_i) = t_i$. In Equation 2.36, the first two terms of the Taylor series expansion are performed to linearize the nonlinearity in Equation 2.35 and to obtain the following expressions of MMLE.

$$\begin{aligned} \frac{\partial \text{ln}L}{\partial\beta_j} &\simeq \frac{2p}{n\sigma}\sum y_i(\alpha_i + \beta_i)z_{i(j)}, \\ \frac{1}{n}\frac{\partial\,\text{ln}L}{\partial\sigma} &\simeq -\frac{1}{\sigma} + \frac{2p}{nk\sigma}\sum z_i(\alpha_i + \beta_i z_{i(j)}), \end{aligned} \tag{2.37}$$

in which $\alpha_i = (2/k)t_i^2/(1+(1/k)t_i^2)$ and $\beta_i = 1/(1+(1/k)t_i^2)$. Hereby, the explicit solution of Equation 2.37, as the MMLE estimator of σ is found as

$$\hat{\sigma} = \frac{B + \sqrt{B^2 + 4NC}}{2N}, \tag{2.38}$$

where

$$\begin{aligned}
B &= \sum_{i=1}^{a} B_i, \\
C &= \sum_{i=1}^{a} C_i, B_i = \frac{2p}{k}\alpha_i(y_i - \bar{y}_{[.]} i K(y_{[i]} - \bar{y}_{[.]})), \\
C &= \frac{2p}{k} \sum \beta_i (y_{(i)} - \bar{y}_{[.]} - K y_i - \bar{y}_{[.]})^2.
\end{aligned}$$

2.3.1 Development of Modified Maximum-Likelihood Estimation

In MMLE, similar to the lasso-based idea in the estimation of GGM, regression model is constructed. Y is assumed as the location-scale type multivariate distribution in this model. Also, for any pair of (Y_j, Y_l) ($2 \leq j \leq q, j+1 \leq l \leq n$ and $j < l$), the random variables of Y, have the bivariate student-t distribution with v degrees of freedom [70]. Islam [70] proves that the q-variate location-scale type multivariate distribution with Γ location vector and Ω scale matrix can be pairwisely written as the bivariate t distribution. In that case, we biologically examine the relation of the network pairwisely. This structure is also the most common relation type in biological network. By this way, we divide the matrix Θ as a binary form by selecting as the response variable and Y_l as the explanatory variable according to Equation 2.39 as reduced model of Equation 2.11. Hence, the model is constructed as the following formula for all pairwise interactions.

$$Y_j = \beta Y_l + \varepsilon_j. \tag{2.39}$$

In Equation 2.39, similar to Equation 2.11, β is the regression coefficient and ε_j presents the error term of the jth node. After the model construction, the variance of the model, $\sigma_{(j,k)}$, is calculated. Then, this process is repeated sequentially until all genes are regressed on all the corresponding genes as the lasso regression. Later, we take the symmetry of these estimated entries to get the symmetric covariance-variance matrix Σ. Finally, in order to infer the precision matrix Θ, we compute the inverse of this estimated Σ. On the other hand, in the construction of the adjacency matrix from the estimated Θ, we take different threshold values q which means that if the coefficients are less than q, the elements are set to 0, otherwise, they are taken as 1. In this study, we fix q to a small entry in order to not loose any interaction in such sparse scale-free networks. Therefore, we arbitrarily set q to $q = 0.01$ for both GGM and LTS graphical model obtained by MMLE.

2.4 Bernstein Operators

The biological networks are complex and this complexity causes lack of the effective solution on the estimation of parameters in the network structure. Therefore, we propose the Bernstein operators as a preprocessing step before any modelling and inference to eliminate batch effect. The mathematical details of these operators are presented in the following subsections.

2.4.1 Bernstein Polynomials

The Bernstein polynomials [19] are simply the algebraic expressions which can define a continuous function on the closed interval by performing the Weierstrass approximation theorem [60].

Weierstrass Theorem[145] Let $f(x)$ be a real valued continuous function for $[a, b]$, i.e., $f \in C[a, b]$. Then, there exists a sequence of the nth degree polynomial, $P_n(x)$, which converges to $f(x)$ for every $\delta > 0$ such that the norm between $f(x)$ and $P_n(x)$ is bounded via δ as the following inequality.

$$|f(x) - P_n(x)| < \delta, \tag{2.40}$$

when n goes to infinity.

There are many proofs about the Weierstrass theorem, but one of the most famous proofs is demonstrated by Sergei Natanovich Bernstein[19] by introducing the Bernstein polynomials.

Accordingly, the Bernstein polynomials have three major features, namely, the property of symmetry, positivity and the probability of the kernel. The first characteristic, i.e., the feature of symmetry, presents that $B_{i,n}(x) = B_{n-i,n}$ while $i = 1, \ldots, n$. On the other hand, the second feature, i.e., positivity, means that $B_{i,n}(x) \geq 0$ and finally, the third one, i.e., the probability of kernel, implies $\sum_{i=0}^{n} B_{i,n}(x) = 1$ when $0 \leq x \leq 1$. In all these expressions, $B_{i,n}$ indicates the expression below.

$$B_{i,n}(x) = \binom{n}{i} x^i (1-x)^{n-i} \tag{2.41}$$

and is called as the Bernstein kernel for a random variable x having n observations .

Hence assuming that f is a function over the range $C[a, b]$, f can be uniformly approximated by polynomials. Hereby, the Bernstein polynomials are one of the most well-known polynomials with a real-valued function f bounded on the interval $[0, 1]$ by the following equation.

$$B_{k:n}(f; 0) = \sum_{k=0}^{n} f\binom{n}{k} b_{k,n}(t) \tag{2.42}$$

in which n is the degree of the Bernstein polynomials. $f\binom{n}{k}$ is equivalent to the approximation of the values for the function f at points k $(k = 0, \ldots, n)$ in the domain of f implying that any interval $[a, b]$ can be transformed into the interval $[0, 1]$. Finally, $b_{k,n}(t)$ is the Bernstein basis with the degree n on the parameter $t \in [0, 1]$ via

$$b_{k,n}(t) = \binom{n}{k} (1-t)^{n-k} t^k. \tag{2.43}$$

In Equation 2.43, $\binom{n}{k}$ is a binomial coefficient that can be obtained from the Pascals triangle. For $n \geq 0$, there are $(n + 1)$ amounts of basis polynomials. For example, there are 4 basis functions for $n = 3$ which can be listed as:

$b_{0,3} = (1-x)^3$, $b_{1,3} = 3x(1-x)^3$, $b_{2,3} = 3x^2(1-x)^3$ and $b_{0,3} = x^3$.

Thus, the Bernstein basis polynomials of the first degree can be seen as

$b_{0,0} = 1$, $b_{0,1} = 1 - x$ and $b_{1,1} = 1 - x$.

Accordingly, the Bernstein basis polynomials of the second degree can be presented as below

$b_{0,2} = (1-x)^2$, $b_{1,2} = 2x(1-x)$ and $b_{2,2} = x^2$.

Hence, the first ordered Bernstein polynomial basis function can be mapped by Figure 2.11.

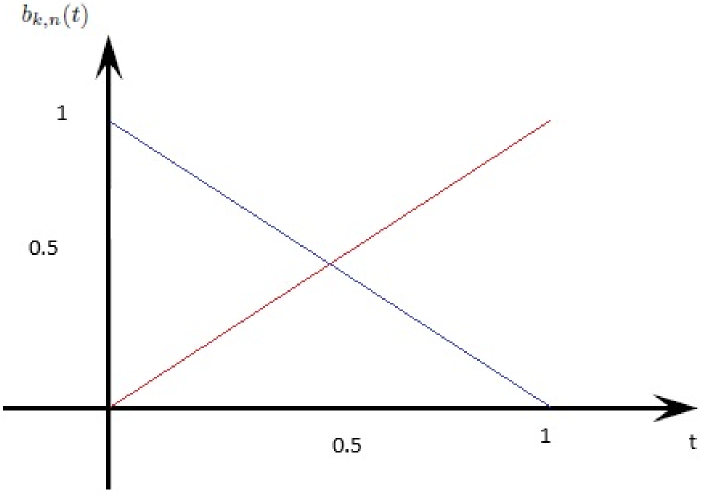


Figure 2.11: Bernstein basis function for the first order.

The second ordered Bernstein polynomial basis function can be mapped by Figure 2.12.

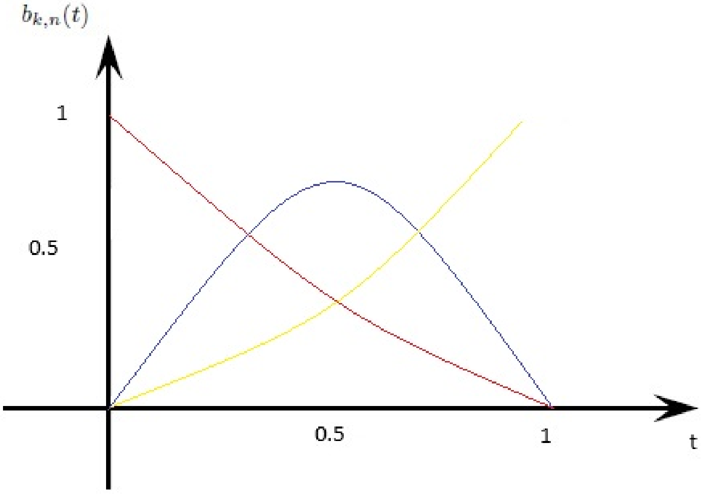


Figure 2.12: Bernstein basis function for the second order.

Furthermore, the Bernstein polynomial approximates to a given function $f(t)$ in such a way that $f(t)$ is always at least as smooth as $f(t)$ is allocated uniformly in $[0, 1]$ for a continuous $f(x)$ on the range $[0, 1]$ as shown in Equation 2.44.

$$\lim_{n \to \infty} B_n(x) = f(x). \tag{2.44}$$

2.4.2 Szász-Mirakyan Operator

The Szász-Mirakyan operator [127], as the generalization of the Bernstein operator for the function on the infinite interval $[0, \infty)$, is studied comprehensively by Cheney and Sharma [29] who show that if f is convex, the sequence of classical Szász-Mirakyan operators decreases as n. Rempulska and Graczyk [114] introduce the generalized Szász-Mirakyan operator on two varibles, Mahmudov [91] proves that the rate of aprroximation by the q-Szász operator, Walczak [139] introduces that specific modification of the Szász-Mirakyan operator, Aral and Gupta [5] show that Szász-Mirakyan operators are convex if the function is convex, and finally, Butzer and Karsli [26] estimates the rate of convergence on the function of Szász-Mirakyan operator.

The Szász-Mirakyan operator is defined by

$$S_n(f;x) = e^{-nx} \sum_{k=0}^{\infty} f\binom{N}{k} \frac{(nx)^k}{k!}, \tag{2.45}$$

where $x \in [0, \infty)$ and the function f is presented in an infinite interval $R^+ = [0, \infty)$. These operators are the generalization of the Bernstein polynomials since the Szász-Mirakyan operators show the properties of the Bernstein Operators. Szász-Mirakyan operators converges to the function f which is continuous in the closed range of $[0, A]$ $(A \in R^+)$ and which is also limited to all positive half-axes via

$$\lim_{n \to \infty} S_n(x) = f(x), \tag{2.46}$$

in which $x \in [0, A]$. Then, the basis of the Szász-Mirakyan operator is defined as in Equation 2.47 by

$$P_{i,n}(x) = e^{-nx} \frac{(nx)^i}{i!}, \tag{2.47}$$

in which $P_{i,n}$ is known as the probability of the Poisson distribution with a mean nx.

2.4.3 Bleiman-Butzer-Hahn Operator

The Bleiman-Butzer-Hahn (BBH) operator [23] which is defined by the Bernstein-type can be represented as below for the nth degree with the k basis polynomials for the value x.

$$L_n(f;x) = \frac{1}{(1+x)^n} \sum_{k=0}^{n} \binom{n}{k} x^k f\left(\frac{k}{n+1-k}\right). \tag{2.48}$$

For Equation 2.48, the following inequality is also satisfied.

$$|L_n(f;x)| \leq ||f||_{C_B} \quad (f \in C_B[0, \infty)). \tag{2.49}$$

Equation 2.49 implies that the BBH operator is linear and bounded for $x \in [0, \infty)$ when $(1+x)^{-n} \sum\limits_{k=0}^{n} \binom{n}{k} x^k = 1$. Here, $C_B[0, \infty)$ is the class of the real-valued function f defined within the interval $[0, \infty)$ and for all functions of f in this interval, $\lim L_n(f;x) = f(x)$ for each $x \in [0, \infty)$ when $n \to \infty$.

Additionally, the property of the uniform approximation of the BBH operator is studied when f belongs to $C[0,\infty]$ for the continuous function on $[0,\infty)$ [134]. Also, Mercer [94] independently derives the Voronovskaya-type theorem which gives an asymptotic error term for the Bernstein polynomials for the functions which are twice differentiable as follows.

$$\lim_{n\to\infty} n((L_n(f;x)-f(x)) = \frac{x(1+x)^2}{2} f^{''}(x) \tag{2.50}$$

for all $f \in C^2[0,\infty)$ with $f(x) = O(x)$ when $x \to \infty$, and $f^{''}(x)$ is the second derivative of the function.

Abel [1] extends this study by giving the complete asymptotic expansion for the BBH operator as the following form.

$$L_n(f;x) = f(x) + \sum_{k=1}^{\infty} c_k(f;x)(n+1)^{-k} \text{ for } n \to \infty. \tag{2.51}$$

Here, c_k represents all the coefficients from $k = 0$ to $k = n$.

The Szasz operator is the limiting operator of BBH [75] and $L_n(f;x) \geq L_{n+1}(f;x) \geq \ldots \geq f(x)$ if f is convex [76] . $L_n(f;x)$ is convex itself if f is a non-increasing convex function.

Jayasri and Sitaraman [71] determine that L_n is a pointwise approximation procedure in the largest subclass of $C[0,\infty)$ for the Bernstein-type of operator. Then, the following functional class is introduced in the study of Hermann [64] via

$$\mathcal{H} = \{f \in C[0,\infty) : \log(|f(x)|+1) = o(x)\}. \tag{2.52}$$

He proves that if f belongs to $\mathcal{H}$, then for each $x > 0$ and $x \to \infty$, the pointwise convergence is $\lim_{n\to\infty} L_n f = f$ on $[0,\infty)$. Moreover, for some $a > 0$, $f(x) = e^{ax}$, then $\lim_{x\to\infty} L_n(f;x) = \infty$. Also, the operator L_n is arisen from the random variable with n observations, X_n, which has the Bernoulli distribution as below:

$$P(\{X_n = k/(n-k+1)\}) = \binom{n}{k} p^k q^{n-k} \tag{2.53}$$

for the parameters $p = x/(1+x)$, $q = 1/(1+x)$ and $k = 0, 1, \ldots, n$.

2.4.4 Meyer-König and Zeller Operator

The Meyer-König and Zeller Operator (MKZ) [95] is given by the equation

$$M_n(f;x) = \sum_{k=0}^{\infty} f\left(\frac{k}{n+k+1}\right) m_{n+1,k}(x) \tag{2.54}$$

in which m is formulated as,

$$m_{n,k}(x) = \binom{n+k}{k} x^k (1-x)^{n+1}, \tag{2.55}$$

for the nth degree and the kth basis polynomial for the value of x as stated previously.

These operators are known as the Bernstein-type of operators. Cheney and Sharma [29] define this operator as a power series of the Bernstein operators.

The MKZ operator can be also obtained from the negative binomial distribution via

$$M_n(f;x) = (1-x)^{n+1} \sum_{k=0}^{\infty} \binom{n+k}{k} f\left(\frac{k}{n+k}\right) x^k. \tag{2.56}$$

2.5 Copula Method and Network Structure

The copulas are perfect tools for modeling and simulating dependent random variables. In this study, the copulas are preferred to generate the datasets for simulation studies so that we can obtain various high dimensional joint distributions under distinct marginals. In our analyses, we apply these data to evaluate the performance of all our suggested approaches under non-normality.

2.5.1 Copula Method

A copula is a multivariate distribution function that separates univariate marginals with a dependency structure. Rényi [115] propose some axioms related with the measure of dependency between two random variables. But the idea of copula was first introduced by Sklar [122] at the same year in the theorem known by his name.

Then, the earliest paper related with Copula was published by Schweizer and Wolff [118]. They showed that the copula of a pair of random variables is invariant when the transformation is increasing. Growing interests in copulas is afforded after 90's. The copula families were generated and extensive surveys by Hutchinson and Lai [69], Joe [73] and Nelsen [100].

Hereby, mathematically, the copula can be defined as

$$H(x,y) = C(F(x),G(y)), \tag{2.57}$$

where X and Y are the continuous random variables, $F(x)$ and $G(y)$ are marginal distributions, and mapping from $C[0,1]^2 \to C[0,1]$ is known as the copula.

Accordingly, by following the description in the study of Nelsen [100], a function

$$C : [0,1]^2 \to [0,1] \quad (u,v) \to C(u,v) \tag{2.58}$$

with the properties

1. $\forall u, v \in [0,1]$
   $C(u,0) = C(0,v) = 0$,
2. $C(u,1) = u$ and $C(1,v) = v$,
3. $\forall u_1, u_2, v_1, u_2 \in [0,1]$ with $u_1 \leq u_2$ and $v_1 \leq v_2$, it holds $C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0$,

is called the (bivariate) copula function for two dimensions.

2.5.1.1 Sklar's Theorem

The definition of the copula is also connected by the Sklar's theorem, which explains the copulas in the statistical modelling. The Sklar's theorem says that [122] any joint distribution function F with the marginal distribution $F_j \, (j = 1, \ldots, n)$ can be written as

$$F(x_1, x_2, \ldots, x_n) = C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)) \; . \tag{2.59}$$

In Equation 2.59, C is called as the copula of F for all random variables x_i when $\forall x \in R$. Here, if the marginals are continuous, then C is unique. The expression of copulas can be also presented as below for marginals F_j and uniform random variables u_i $(i = 1, \ldots, n)$ in $[0, 1]^n$.

$$C(u_1, u_2, \ldots, u_n) = F(F_1^{-1}(u_1), F_2^{-1}(u_2), \ldots, F_n^{-1}(u_n)). \tag{2.60}$$

This allows that the copula density function can be derived from the differentiation of F where F denotes the multivarite distribution function.

We can give the following example to explain its application. Let

$$H(x, y) = \begin{cases} 4xy & \text{if } 0 < x < 1, 0 < y < 1, \\ 0 & \text{if } otherwise, \end{cases}$$

while $H(x)$ is a joint distribution function. To obtain the copula function, it is necessary to first calculate marginals (F and G) because of the Sklar's Theorem. Thus,

$$F(x) = \int_0^1 4xydy = 2x,$$
$$G(y) = \int_0^1 4xydy = 2y,$$

resulting in the inverse of these marginals as below:

$$u = F(x), u = 2x \Rightarrow x = u/2, \; 0 < u < 1,$$
$$v = G(y), v = 2y \Rightarrow y = v/2, \; 0 < v < 1.$$

We can now calculate the copula $C(u, v)$ as $u/2$ and $v/2$ by replacing with x and y, respectively.

$$C(u, v) = 4\frac{u}{2}\frac{v}{2} = uv. \tag{2.61}$$

We can also check Equation 2.61 whether it can provide the properties of the copula function defined in Section 2.5.1.

Accordingly,

$$
\begin{aligned}
&C(u_1, v_1) = u_1.v_1 = x,\\
&C(u_2, v_2) = u_2.v_2 = y,\\
&C(u_1, v_2) = u_1.v_2 = z,\\
&C(u_2, v_1) = u_2.v_1 = t,
\end{aligned}
$$

maintains $x + y - z - t \geq 0$ by the values u_1, v_1, u_2, v_2 in between $[0, 1]$. So, we can say that $C(u, v)$ is a copula function.

There are different types of copula functions in the literature such as the Gaussian copula, student-t copula, Gumbel copula, Clayton copula, Frank copula and the product copula. In this study we only explain the Gaussian and student-t copulas due to the fact that they are close alternatives of each other. Whereas in practical purpose, the student-t is not commonly preferred for its computational complexity. On the other hand, the other copula types cannot be implemented in biological networks as they do not have explicit functional form for the high dimension which is one of the fundamental structures of the biological systems.

2.5.1.2 Copula Families

2.5.1.3 Gaussian Copula

The Gaussian copula can be formulazed as Equation 2.62 as below.

$$C(u, v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)). \tag{2.62}$$

where Φ_ρ denotes the bivariate standard normal distribution with a correlation ρ on the range between -1 and 1, i.e., $\rho \in (-1, 1)$, and Φ^{-1} is the inverse univariate standard Gaussian distribution function.

The Gaussian copula density function is obtained by applying the inverse method to the standard multivariate Gaussian via

$$C(U,V) = \frac{1}{|R|^{1/2}} exp\{-\frac{1}{2}U^{'}(R^{-1} - I)U\}, \tag{2.63}$$

where R is the correlation matrix defined by $R = cov(x_i, x_j)/\sqrt{var(x_i), var(x_j)}$. While $\text{Cov}(.)$ and $Var(.)$ denote the covariance and variance of the given random variables, respectively. Finally, U and V show the transformed form of X_i and X_j, in order.

2.5.1.4 Students-t Copula

A student's-t copula is created by the bivariate student-t distribution. Closely related with the Gaussian copula, the student-t copula can be represented as

$$C(u,v) = t_{v,\rho}(t_v^{-1}(u), t_v^{-1}(v)), \tag{2.64}$$

where v is the degree of freedom parameter, ρ is the correlation coefficient, t_v^{-1} presents the inverse of the univariate standard student-t distribution function

2.5.2 Network Structure

There are different undirected biological network structures to describe the interaction between the nodes, such as the scale-free network, hub network, cluster network and the random network. In the simulation study, different graphical structures are utilized while generating the datasets. These network structures are explained with the graphical representation as described below. In our analyses, we solely implement the scale-free structure as it is the most common topological feature of the biological networks.

2.5.2.1 Scalee-Free Network

Barabasi and Albert propose the scale-free network [13] based on the following two properties. (i) real networks are not constant and they are growing constantly; and probability of connection of two nodes are not uniformly distributed, (ii) in a real

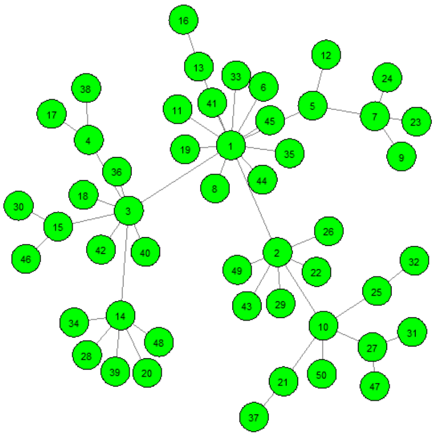


Figure 2.13: An example of a scale-free network with 50 nodes.

network, the new node is connected to a node that has a higher number of connections than the others. In this network model, the network initially starts with a small number of nodes, and a new node is added at every time step via the preferential attachment. They also show that $P(k)$ has an interaction with k as a power law distribution as seen in Equation 2.65 [13]:

$$P(k) \sim k^{-\gamma} \tag{2.65}$$

in which $P(k)$ is the fraction of nodes of the degree k and γ is the exponent of the distribution. A simple example of the scale-free network with 50 nodes can be found in Figure 2.13.

2.5.2.2 Random Network

The random networks were first studied by Erdös and Rényi in 1959 [41]. There are many procedures to obtain a random network and one of the method to construct a random network can be represented by the steps below.

1. In a random network with N nodes, each pair of nodes is connected with a probability p and generated by a random structure with N disjoint nodes. Here, two nodes are randomly choosen and generate a number from a uniform distribution on the interval [0,1]. The choice is greater than p, two nodes with a link

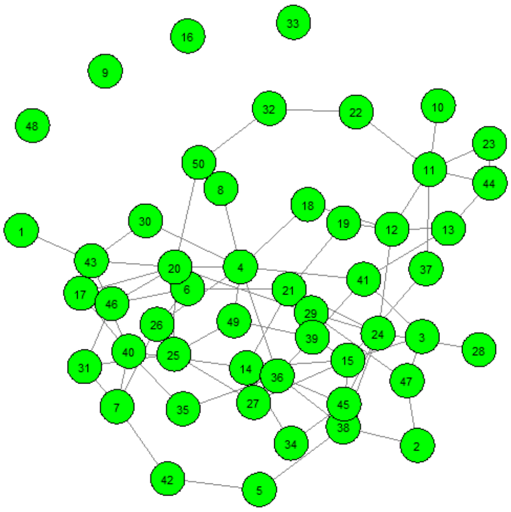


Figure 2.14: An example of random network with 50 nodes.

are connected. Otherwise, the nodes are disconnected.

2. Step is repeated for each $N(N-1)/2$ node pair.

After following these steps, the random network can be obtained as represented in Figure 2.14.

2.5.2.3 Hubs Network

A few nodes with a higher clustering coefficient connected with many links are known as hubs [14]. The hubs are in a central position within the clusters and it can be said that the large number of hubs creates the scale-free network. An example of a network which contain nodes with 3 hubs is shown in Figure 2.15.

2.5.2.4 Cluster Network

The clustering networks as shown in Figure 2.16 for illustration, consist of highly connected subgraphs and if there are separated networks in the system, the clustering method can identify them [14].

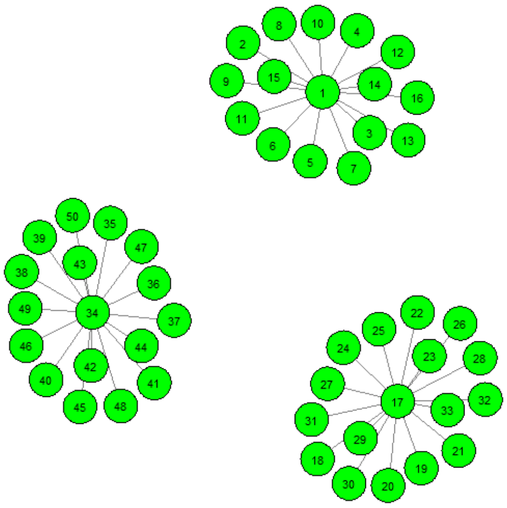


Figure 2.15: An example of a hub network with 50 nodes.

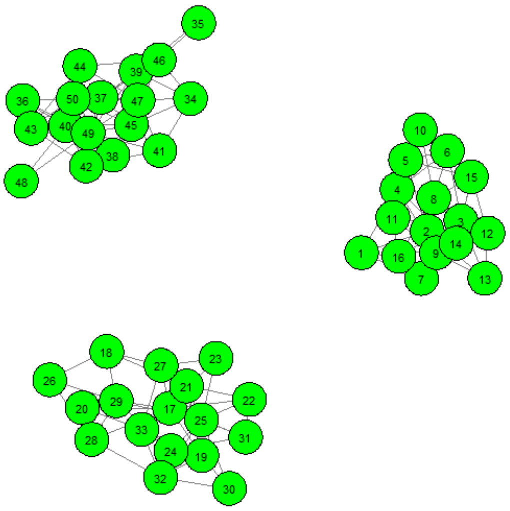


Figure 2.16: An example of a cluster network with 50 nodes.

Table 2.3: Confusion matrix of the accuracy measures.

		Actual Structure		
		Positive	Negative	Row Total
Predicted Structure	Positive	True Positive (TP)	False Positive (FP)	TP+FP
	Negative	False Negative (FN)	True Negative (TN)	FN+TN
	Column Total	TP+FN	FP+TN	

2.6 Model Selection Criteria

In this section, we describe the model selection criteria that are used to check the validity of our proposed model.

The model selection is the process for choosing the best performing model. In the literature, adjusted R^2, Mallow's C_p, Mean-Squared Error (MSE), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are the most commonly used methods. But distinct methods are implemented for the binary classification methods and the accuracy measures are the most common binary classification model selection methods.

2.6.1 Accuracy Measures

In order to compare the performance of actual and predicted classes, specificity, precision, F-measure and Matthews correlation coefficient (MCC) values are used. In the calculation of these values, true positive (TP), true negative (TN), false positve (FP) and false negative (FN) are applied as listed in Table 2.3.

Here, the true positive implies the number of correctly classified objects that have links denoted by 1 and the true negative defines the number of correctly categorized objects that have no link denoted by 0. On the contrary, the false positive shows the number of misclassified objects that have no links, i.e., 0 entries, and false negative shows the incorrectly classified objects that is correct in actual.

Using these classified objects, precision, specifity, F-measure and MCC are computed as in Equations 2.66 - 2.70.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}. \tag{2.66}$$

The precision is a measure of how close the value which we estimate to the real value. It is also called the positive predictive value (PPV).

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}}. \tag{2.67}$$

The specifity measures how much the test can capture the negative cases. It is also known as the true negative rate (TNR).

$$\text{F} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.68}$$

The F-measure is the balance of the recall and precision values since F-measure is the harmonic mean of the precision and the recall where the recall describes the ratio of correctly classified objects with positive labels to the total positive classes in the actual case as shown in Equation 2.69.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}. \tag{2.69}$$

The recall is also named as the sensitivity or the true positive rate (TPR).

Finally, the Matthew's correlation coefficients (MCC) is calculated as in Equation 2.70.

$$\text{MCC} = 2\frac{(\text{TP} + \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP}) \times (\text{FN} + \text{TN}) \times (\text{TP} + \text{FN}) \times (\text{FP} + \text{TN})}}. \tag{2.70}$$

and takes the values between -1 and 1, where -1 indicates a totally wrong classifier and 1 shows the completely true classification.

CHAPTER 3

APPLICATION

This chapter contains the application and the validity of the proposed methods which are described in Chapter 2. In Section 3.1, we check the validation of our proposed method, LMARS, under different dimensional simulation studies and real data applications. In Section 3.2, we evaluate LMARS with interaction effect. Section 3.3 includes the application of the MMLE method which is proposed in the estimation of the precision matrix. The applcations of the Bernstein operators which are proposed for smoothing the biological networks on simulated and real datasets are represented in Subsection 3.4.1, 3.4.2 and 3.4.3.

3.1 LMARS Application

In the application of LMARS, we estimate the precision matrices, Θ, of simulated datasets and calculate the accuracy measures. The true precision matrix of the generated matrices is created under the two different choices which come from normal and nonnormal distributions. For the normal data, we consider two scenarios. In the first scenario, Θ has positive entries which represent positive relations between genes and in the second scenario, the matrix includes negative entries corresponding to inhibitory relations between genes. Under both cases, we perform matrices having different sizes, namely, (100×100), (500×500) and (1000×1000) dimensions. Furthermore, the underlying matrices have distinct sparsity percentages. Because we use a special matrix form similar to the tridiagonal matrix containing a nonzero main diagonal and nonzero parallel diagonals in each side. Our matrix, as given in Equation

3.1, also possesses nonzero entries in the right upper corner and the left lower corner.

$$\mathbf{B} = \begin{bmatrix} \gamma & \gamma & 0 & 0 & \gamma \\ \gamma & \gamma & \gamma & 0 & 0 \\ 0 & \gamma & \gamma & \gamma & 0 \\ 0 & 0 & \gamma & \gamma & \gamma \\ \gamma & 0 & 0 & \gamma & \gamma \end{bmatrix}. \tag{3.1}$$

In Equation 3.1, γ denotes the nonzero entries. Due to this special structure for all matrices, our precisions own different sparsity percentages in each dimension such that (100×100), (500×500) and (1000×1000) matrices have 97%, 99.4% and 99.7% sparsity, respectively. Finally, in our data generation, we consider that each gene is composed of 20 observations.

For the non-normal datasets, we also have two scenarios. In the first case, the matrix generated from the student-t distribution under the degrees of freedom 7 and in the second case, the matrix is generated from the student-t distribution under the degrees of freedom 15. For both conditions, the matrices are simulated under (40×40), (100×100) and (200×200) dimensions. Moreover, as stated beforehand, the inference of all precisions is conducted via LMARS accepting that every gene is assigned as a response and the remaining genes are used as covariates. Furthermore, in the model fitting, we merely take the main effects of all genes and discard all interaction terms, as presented previously. The reason is that we aim to convert the MARS regression into the lasso regression used in GGM so that both models can be comparable. Additionally, similar to the simple linear regression models, it is expected that the main effects explain the major parts of the model [98]. Hence, this process is repeated sequentially until all genes are regressed on all the corresponding genes. Then, according to the estimated regression coefficients, we solely take significant ones without computing the estimated precision matrix and consider that there is a relation between those components. Finally, in the generation of the adjacency matrix, while 1 refers to a relationship and 0 describes no interaction between the pairs of genes, we apply both the AND and OR rules. By using the AND rule, we obtain an entry 1 when both (i, j) and (j, i) entries of the estimated adjacency matrix are one. On the other hand, we apply the OR rule while either (i, j) or (j, i) entry of Θ has value 1, and we accept that it is a sufficient condition to assign the correspond-

Table 3.1: Comparison of the specificity, precision and F-measure computed via LMARS under both AND and OR rules, and GGM and 1000 Monte-Carlo runs based on different dimensional Θ matrices p with normally distributed data with sample size 20 and plans (S_1: scenario 1, S_2: scenario 2).

Plan	p	Specificity			Precision			F-measure		
		LMARS		GGM	LMARS		GGM	LMARS		GGM
		AND	OR		AND	OR		AND	OR	
S_1	100	0.9918	0.9092	0.9907	0.5638	0.1170	0.6867	0.4240	0.1799	0.4273
	500	0.9989	0.9683	0.9871	0.6004	0.0631	0.1395	0.4304	0.1070	0.1978
	1000	0.9989	0.9819	0.9577	0.6125	0.0541	0.0241	0.4327	0.0933	0.0462
S_2	100	0.9918	0.9091	0.9913	0.5630	0.1170	0.7064	0.4241	0.1798	0.4319
	500	0.9980	0.9684	0.9871	0.5995	0.0631	0.1393	0.4302	0.1070	0.1976
	1000	0.9990	0.9820	0.9565	0.6166	0.0541	0.0225	0.4335	0.0933	0.0440

ing entry and its symmetry to one. From both analyses, it is seen that the AND rule generates more sparse and accurate matrices, i.e., networks with less computational demand as the results of one scenario is presented in Table 3.1 and 3.6. Therefore, for the remaining tabulated values apart from Table 3.1 and 3.6, we only represent the findings of the AND rule as the outputs of hte LMARS estimates. In the construction of the adjacency matrix from the estimated regression coefficients, we take significant main effects of the coefficients and assigned it as 1 and others as 0. Finally, in the modelling, we both check backward and forward selection representing that by starting the full model and reducing this model via the one-by-one strategy, we choose the model having the highest GCV. By this way, we get a set of 100, 500 and 1000 regression models for each 100, 500 and 1000-dimensional systems, respectively. On the other hand, in order to compare the performance of both GGM and LMARS, we compute their specificity, precision and F-measure values as shown in Section 2.6.

The findings are presented in Table 3.1. From the results, it is observed that there is no difference in specificity values of LMARS between two scenarios of the true precision matrix. On the other side, when the dimension of the matrix increases, this value is closer to 1 indicating its perfection level. In addition, we observe higher specificity values when we use the AND rule since it generates more sparse matrices. Moreover, F-measure and precision values decrease when the dimension increases under both scenarios [8].

Table 3.2: Comparison of the specificity, precision and F-measure computed via LMARS under AND rule, and GGM and 1000 Monte-Carlo runs based on different dimensional Θ matrices p with student-t distributed data at the degrees of freedom 3 with sample size 20 and plans (S_1: scenario 1, S_2: scenario 2).

Plan	p	Specifity		Precision		F-measure	
		LMARS	GGM	LMARS	GGM	LMARS	GGM
S_1	40	0.9270	0.9399	0.2605	0.3022	0.4131	0.4634
	100	0.9718	0.9655	0.2633	0.2381	0.4166	0.3818
	200	0.9862	0.9777	0.2681	0.2031	0.4227	0.3334
S_2	40	0.9263	0.9405	0.2584	0.3055	0.4105	0.4670
	100	0.9715	0.9653	0.2615	0.2370	0.4144	0.3803
	200	0.9863	0.9783	0.2680	0.2049	0.4228	0.3363

On the other hand, the selected accuracy measures of GGM are computed under the optimal penalty constant based on the StARS (stability approach to egularization selection) criterion and the glasso method in the inference of model parameters. As seen in Table 3.1, GGM has higher specificity, F-measure and precision values for small dimensions. Whereas, when the dimension increases, these values decrease. But LMARS under the AND rule gets higher specificity, precision and F-measure values with respect to the GGM outputs for both scenarios and under all dimensions based on the multivariate normally distributed data.

Additionally, since LMARS is a nonparametric approach, we calculate the performance of this method for nonnormal datasets. For this purpose, we generate data from the student-t distributions with the degrees of freedom 3, 4, 7 and 15 as shown in Tables Table 3.2-3.5, respectively. On the contrary, as found in Table 3.5, GGM becomes better for all measures while the data are closer to normal like the student-t with 15 degrees of freedom in all measures. As a result, we observe that when the dimensions of the systems increase and the measurements are far from the normal density, LMARS outperforms GGM as it is expected. On the other side, when degrees of freedom decrease, i.e., dataset is far from the normal distribution, LMARS and GGM give lower accurate results for all dimension as represented in Table 3.2.

On the other side, as we denote in Table 3.1–3.5, we estimate different size of networks and it is seen that the estimation of the parameters in networks, under distinct

Table 3.3: Comparison of the specificity, precision and F-measure computed via LMARS under AND rule, and GGM and 1000 Monte-Carlo runs based on different dimensional Θ matrices p with student-t distributed data at the degrees of freedom 4 with sample size 20 and plans (S_1: scenario 1, S_2: scenario 2).

Plan	p	Specifity		Precision		F-measure	
		LMARS	GGM	LMARS	GGM	LMARS	GGM
S_1	40	0.9286	0.9386	0.2649	0.2972	0.4186	0.4576
	100	0.9717	0.9686	0.2628	0.2508	0.4161	0.3992
	200	0.9862	0.9802	0.2672	0.2171	0.4216	0.3538
S_2	40	0.9280	0.9376	0.2634	0.2938	0.4167	0.4534
	100	0.9714	0.9690	0.2612	0.2530	0.4140	0.4021
	200	0.9863	0.9781	0.2679	0.2032	0.4225	0.3340

Table 3.4: Comparison of the specificity, precision and F-measure computed via LMARS under AND rule, and GGM and 1000 Monte-Carlo runs based on different dimensional Θ matrices p with student-t distributed data at the degrees of freedom 7 with sample size 20 and plans (S_1: scenario 1, S_2: scenario 2).

Plan	p	Specifity		Precision		F-measure	
		LMARS	GGM	LMARS	GGM	LMARS	GGM
S_1	40	0.9787	0.9979	0.5748	0.9607	0.4340	0.4930
	100	0.9918	0.9907	0.5627	0.6835	0.4242	0.4272
	200	0.9960	0.9838	0.5756	0.2654	0.4248	0.2898
S_2	40	0.9787	0.9976	0.5738	0.9549	0.4328	0.4919
	100	0.9918	0.9909	0.5635	0.6956	0.4241	0.4294
	200	0.9960	0.9840	0.5760	0.2686	0.4249	0.2912

Table 3.5: Comparison of the specificity, precision and F-measure computed via LMARS under AND rule, and GGM and 1000 Monte-Carlo runs based on different dimensional Θ matrices p with student-t distributed data at the degrees of freedom 15 with sample size 20 and plans (S_1: scenario 1, S_2: scenario 2).

Plan	p	Specifity		Precision		F-measure	
		LMARS	GGM	LMARS	GGM	LMARS	GGM
S_1	40	0.9787	0.9979	0.5738	0.9598	0.4328	0.4930
	100	0.9091	0.9911	0.1172	0.6998	0.1803	0.4305
	200	0.9961	0.9839	0.5748	0.2667	0.4247	0.2906
S_2	40	0.9788	0.9979	0.5756	0.9549	0.4343	0.4919
	100	0.9918	0.9909	0.5630	0.6947	0.4242	0.4298
	200	0.9961	0.9840	0.5748	0.2738	0.4246	0.2921

Table 3.6: Comparison of the real computational time per second calculated via LMARS and GGM under 1000 Monte-Carlo runs based on different dimensional matrices and normally distributed data in scenario 1.

Dimensions of Θ	LMARS with AND Rule	LMARS with OR Rule	GGM
(100×100)	18469.3	21039.0	20713.7
(500×500)	36670.5	38118.9	113460.1
(1000×1000)	131857.0	154143.0	1082029.3

dimensions occurs at different times. When the dimension of the system increases, the parameter estimation becomes difficult and the estimation becomes computationally demanding for large networks. For this reason, the timing turns to be essential for researchers, especially for the high-dimensional datasets. So we consider that the computational efficiency can be another performance criterion to choose the best model for the biological systems.

Accordingly, while assessing the computational time as presented in Table 3.6, we observe that LMARS is significantly speedy with respect to GGM. The calculations of LMARS for (100×100), (500×500) and (1000×1000)-dimensional matrices under 1000 Monte-Carlo iterations are completed in 18469.3, 36670.5 and 131857, minutes, respectively, in LMARS with the AND rule. Whereas, GGM does only 1000 iterations for (100×100)-dimensional matrix in 20713.7 minutes. Also as reported in Table 3.6, LMARS is faster than GGM under large systems such that GGM completes the calculations in 1082029.3 seconds for 1000 iterations while LMARS with the OR rule does it in 154143 seconds. Therefore, we conclude that the LMARS method is computationally more efficient than GGM.

We also apply this method in a realistically large system, called the JAK-STAT (Janus kinase/signal transducer and activator of transcription) pathway (Figure 3.1), by using a simulated dataset. This transaction pathway is an important signalling pathway which is activated by Type I interferons (IFN). IFNs control the immune system of living organisms and are used to treat the hepatitis B and C virus infections [92].

In the data generation for the JAK-STAT pathway, we describe the system with 38 proteins and consider that each protein has 10 observations. Here, the list of the proteins, their initial numbers of molecules and their reaction rate constants in the reaction list of the system are used as described in Maiwald et al. [92]. Then, we run the Gillespie algorithm [56] until the total simulation time T sets to 200 and we take the last 10 integer time points from 190 to 199 due to the fact that all the proteins in the system can reach in their steady-state conditions under a long simulation time and the selected time points belong to this phase of the system. The generated time-course data for each protein are shown in the plots presented in Figure A.1, A.2, A.3 and A.4 in Appendix A and the very brief description of the system by using the relations of

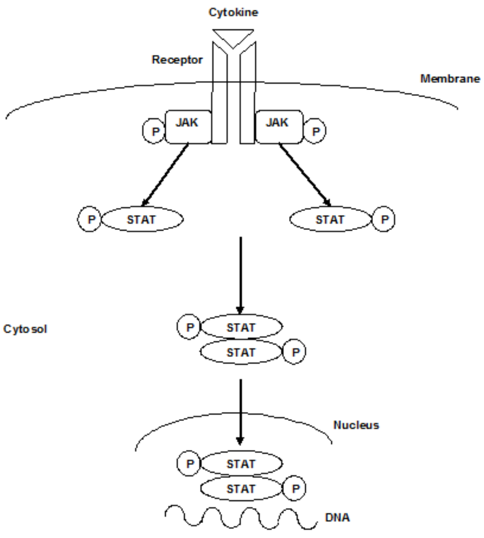
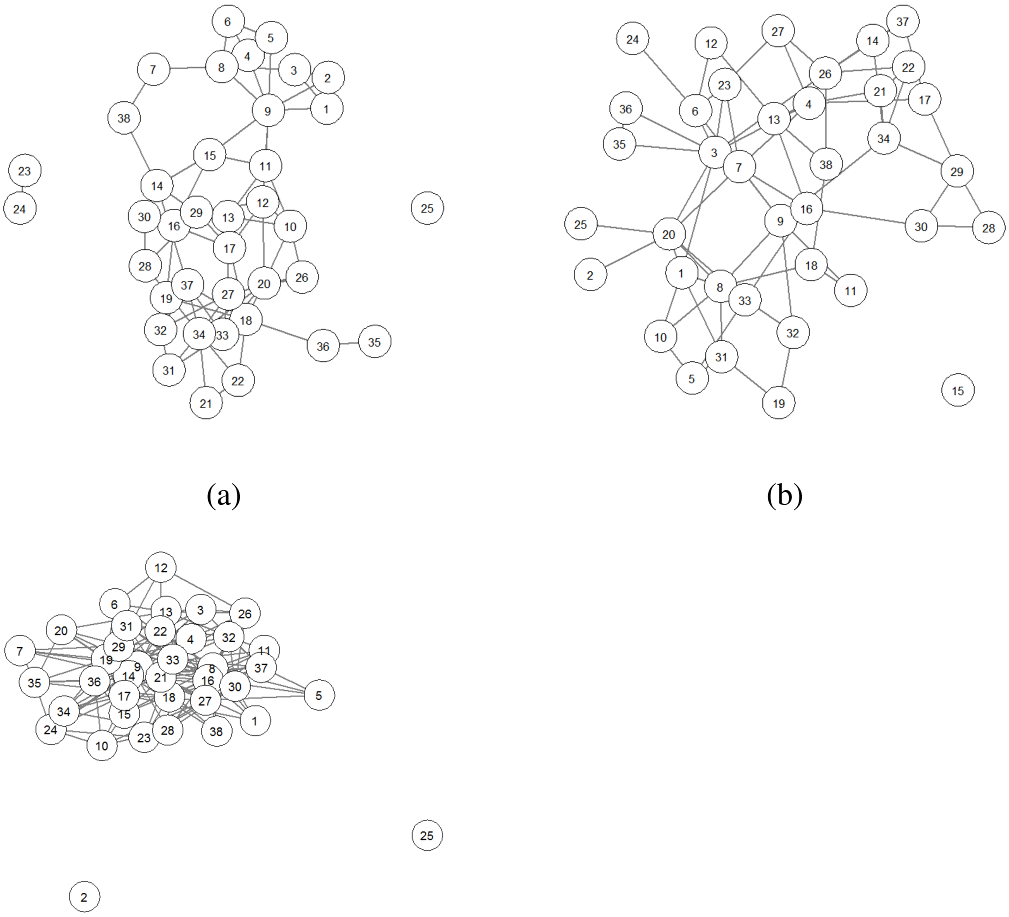


Figure 3.1: Simple illustration of the elements for the JAK-STAT pathway.

the major components can be found in Figure 3.1.

Once the dataset is obtained, we implement LMARS and GGM to estimate the precision matrix Θ, i.e., $\widehat{\Theta}$. Then, we convert both $\widehat{\Theta}$'s into their corresponding adjacency matrices and compare our results with the quasi true adjacency matrix based on the reaction list in Maiwald et al. [92]. Hereby, the specificity measures of LMARS and GGM are found as 0.89 and 0.98, respectively, and the calculations are completed in 3.52 and 328.35 seconds for LMARS and GGM, in order. From these findings, it is seen that similar to the Monte-Carlo studies, although LMARS looses from the accuracy, it is computationally less demanding then GGM. On the other hand, as the system is very sparse, both models are more successful in capturing zero entries in the true Θ, whereas, cannot be effective to estimate the present links. Accordingly, the calculated TN (true negative) values which count the ratio of the zeros are very high and the computed TN (true negative) values which consider the available links are low. Because of this fact, the estimated links of the main components in Figure 3.1 cannot be captured under both GGM and LMARS approaches. In Figure 3.2, we plot the systems based on the quasi true precision and estimated precisions via LMARS and GGM for the visual comparison.

Finally, in order to evaluate the performance of LMARS in the real dataset, we use the data of the protein-protein interactions in the study of Jenner et al. [72]. This

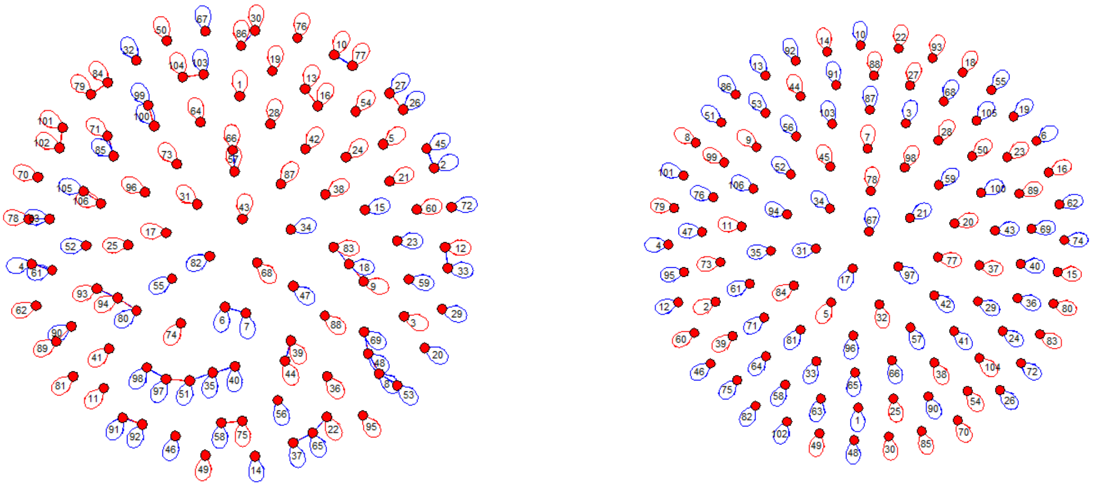


(a) (b)

(c)

Figure 3.2: Structure of the JAK-STAT pathway via (a) true quasi precision, (b) estimated precision via LMARS and (c) estimated precision via GGM under the glasso approach.

dataset consists of 106 genes for 22 time points to describe the pathway of Kaposi's Sarcoma Associated Herpesvirus (KSHV). This pathway is one of the most recently identified human herpesvirus that presents the etiological infectious agent of Kaposi's sarcoma, primary effusion lymphoma and multicentric Castleman's disease [154]. In inference of this pathway via LMARS and GGM, we observe that LMARS gives more biologically validated structure with respect to the GGM results. Because as seen in Figure 3.3, although GGM can merely capture auto-regulated interactions, LMARS can also estimate both auto-regulated links and other interactions. For instance, LMARS estimates the links of Orf45-K8 [154] and K14-K7-K2 [24] which are biologically validated in the related literature besides the auto-regulated links of Orf57-Orf57 [154].



(a) (b)

Figure 3.3: Estimated links of the KSHV pathway via (a) LMARS and (b) GGM under the glasso approach.

3.2 LMARS with Interaction

Comparing the performances of LMARS versus GGM, it is found that LMARS is a fairly well approach as the alternative of GGM. In the application, we perform this model with the interaction effects and compare the results with GGM and LMARS without interaction terms in order to calculate the gain in the extended model. In this part, we implement this model with the interaction effects and make comparative analyses with GGM and LMARS without interaction terms in order to assess the gain in this extended model. In the application of LMARS with/without interaction effects, the same steps are followed as done for LMARS without interaction effect. Thereby, the inferences of precisions are constructed in such a way that every gene is sequentially assigned as a response and the remaining genes are used as covariates. Then, under the estimated regression coefficients, we take significant covariates from each gene-specific model. Here, we accept that these entries imply the significant relations between the response gene and the associated predictor gene, resulting in the entry 1 in the adjacency matrix. Otherwise, we put the zero value in that entry. We compare the underlying calculation for 50, 100 and 500-dimensional systems. The associated datasets are generated by the Gaussian copula function. In the assessment, we particularly choose this copula among alternatives such as Gumbel, Frank and other Archimedean copula types as well as student-t copula. Because as shortly described in Chapter 2, Section 2.5, the Archimedean copulas have explicit forms upto 4 or 5 dimensions [55] and the student-t copula which belongs to the family of elliptical copula like the Gaussian copula [45] is more computationally demanding and less mathematically tractable than the Gaussian copula. Accordingly, in our study, we use three different datasets. In the first dataset, we compare both types of LMARS with GGM under the multivariate normal distribution. Then, in the second dataset, we use the exponential distribution as the marginals and bind them via the Gaussian copula. Finally, in the third part of the simulation, we generate the mixed-distribution dataset where the marginals come from both exponential and normal distributions and the link between them is constructed by the Gaussian copula function. Therefore, we follow the steps that produce our copula bounded marginals whose each random variable has 20 observations and a Σ variance-covariance structure. In order to compare the performance of results from both GGM and LMARS, we compute their specificities, precisions and F-measure values as shown in Section 2.6.

In the calculation of GGM, we infer the model parameters via the glasso method as used in previous analysis. The findings are presented in Tables 3.7 and 3.9. In Table 3.7, marginals come from the exponential distribution with rate 3. From the performance of LMARS without the interaction rule, it is seen that LMARS with

Table 3.7: Monte-Carlo comparison of the specificity, precision and F-measure of different dimensional biologically systems computed via GGM and LMARS models under exponential and normal marginals bounded by the Gaussian copula with sample size 20.

Total number of nodes	Models	Specificity	Precision	F-measure
50	GGM	0.9999	Not Computable	Not Computable
	LMARS without interaction	0.9590	0.3300	0.3400
	LMARS with interaction	1.0000	1.0000	0.6400
100	GGM	0.9799	0	Not Computable
	LMARS without interaction	0.9800	0.3300	0.5000
	LMARS with interaction	1.0000	1.0000	0.5127
500	GGM	0.9980	0.0000	Not Computable
	LMARS without interaction	0.9960	0.3300	0.5000
	LMARS with interaction	1.0000	1.0000	0.4779

interaction is better than LMARS without the interaction model. Furthermore, it is observed that the specifity increases while the dimension increases. But F- measure and precision decrease when we raise the dimension of the system.

In Table 3.7, the marginals are distributed as normal. From the tabulated terms, it is seen that LMARS with interaction effects gives more accurate results than others for all dimensions. But, the accuracy measures of GGM are higher than the LMARS results for low dimensions and GGM cannot calculate the precision and F-measure for 50-dimensional system.

Additionally as seen from Table 3.8 that both LMARS with/without interaction effects have higher accuracies than GGM in all dimensions when the data are extremely far from normality.

On the other hand, in Table 3.8, the marginals are taken as half exponential and half normal distributions, and all the nodes are bounded by the Gaussian copula. By this way, we consider to evaluate the performance of all suggested models under the mixture of two different types of distributions simultaneously. From the results, we detect that LMARS with the interaction effect is more accurate than others for all dimensions.

On the other hand in the GGM analyses, the best penalty constant is chosen by the

Table 3.8: Monte-Carlo comparison of the specificity, precision and F-measures of different dimensional biological systems computed via GGM and LMARS models under exponentials marginals bounded by the Gaussian copula with sample size 20.

Total number of nodes	Models	Specificity	Precision	F-measure
50	GGM	0.9284	0.0000	Not Computable
	LMARS without interaction	0.9590	0.3330	0.5000
	LMARS with interaction	0.9997	0.9956	0.9967
100	GGM	0.9800	0.0000	NaN
	LMARS without interaction	0.9800	0.3330	0.5000
	LMARS with interaction	1.0000	0.9986	0.9986
500	GGM	0.9960	0.0000	Not Computable
	LMARS without interaction	0.9960	0.3330	0.5000
	LMARS with interaction	1.0000	0.9946	0.9946

Table 3.9: Monte-Carlo comparison of the specificity, precision and F-measures of different dimensional biological systems computed via GGM and LMARS under multivariate normal distribution with sample size 20.

Total number of nodes	Models	Specificity	Precision	F-measure
50	GGM	0.9994	Not Computable	Not Computable
	LMARS without interaction	0.9123	0.2637	0.3129
	LMARS with interaction	0.9634	0.6562	0.7923
100	GGM	0.9865	0.4129	0.3972
	LMARS without interaction	0.9092	0.1799	0.1799
	LMARS with interaction	0.9554	0.6043	0.7534
500	GGM	0.9760	0.1210	0.1878
	LMARS without interaction	0.9683	0.0631	0.1070
	LMARS with interaction	0.9870	0.5627	0.7202

rotation information criterion (RIC) [158] as this is the most common measure for the glasso-estimated models. Furthermore, the glasso package is used to estimate the precision matrix [51]. Hereby, in application, two real datasets are used to check the validaty of the new model. The first data [117] is called the cell signalling data and show a small network having 11 phosphoproteins and phospholipids under various experimental conditions in human primary naive CD4+T cells that are measured on 11672 red blood cells. These data are collected after a series of stimulatory cues and then, the inhibitory interventions with cell reactions are stopped at 15 minutes after the stimulation by a fixation in order to profile the effects of each condition on the intracellular signalling networks. In Figure 3.4, the graphical illustration of the accepted signalling molecular interactions is presented and the estimated systems via all alternative models are shown in Figure 3.5. In this figure, it is seen that the LMARS model with interaction effects estimates biologically validated links and already infers the links found via LMARS without interaction effects and GGM. For instance, the link between PKA and PKC proteins (protein numbers 8 and 9 in Figure 3.4), which is biologically validated in the study of Sim and Scott [121], is merely found via LMARS with the interaction effect. Furthermore, the edge between PRAF and PMEK proteins (protein numbers 1 and 2 in Figure 3.4) is correctly inferred [31] in GGM and LMARS with interaction effects. The link between PIP2-PKC proteins (protein numbers 4 and 9 in Figure 3.4) is not estimated via LMARS without interaction model, whereas, it is inferred under the interaction as biologically declared in the study of Kuo [80]. In addition, it is seen that the relation between PRAF and PKC proteins (protein numbers 1 and 9 in Figure 3.5) is validated by LMARS with interaction effects. This interaction can be verified by Figure 3.4 taken from Sachs et al. [117]. In addition, LMARS with interactions can catch the relation between PLCG and PKC proteins (protein numbers 3 and 9 in Figure 3.5). This relation can be also validated by Figure 3.4.

In the second application with a real dataset, we use the large-scale human gene expression data which are gathered by Stranger et al. [124] and are described by Bhadra and Mallick [22] and Chen et al. [28]. This dataset is collected to measure the gene expression in the B-lymphocyte cells from the Northern and Western European ancestry from Utah (CEU). The data are composed of 60 unrelated individuals for

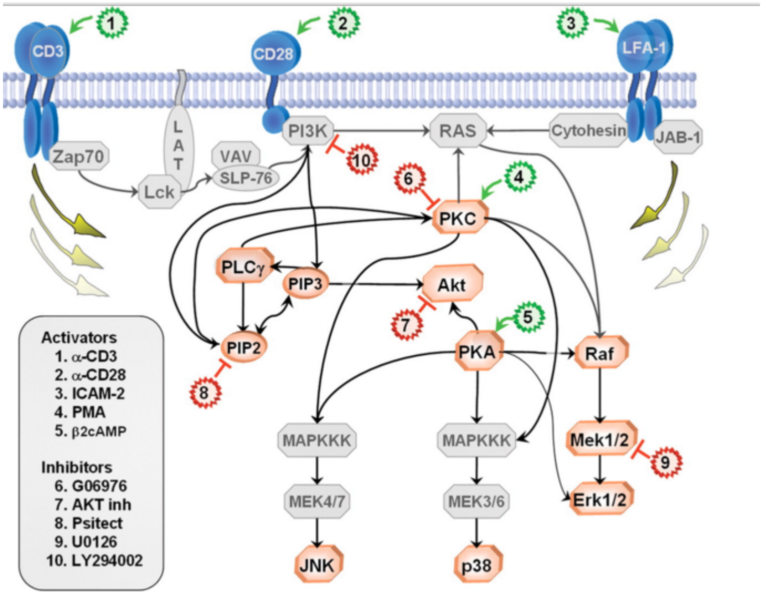


Figure 3.4: True graphical representation of the cell signalling network (Dataset 1) from Sachs et al. [117].

100 probes. 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 0.1. Since UTR of mRNA has an important role in the regulation of the gene expression, the inference of this system has been performed in the previous study [97] via the copula GGM. In this work, we estimate it via both alternatives of MARS and GGM as presented in Figures 3.6-3.8.

From comparative analyses, it is found that both MARS models can capture biologically validated links and GGM cannot detect any interaction. Then, due to the failure of the estimation via GGM, under this high dimension, we also perform an approximate version of the glasso approach, which is particularly designed for such high dimensional systems. Hereby, we infer the parameters of GGM as in the study of Zhao et al. [155] via the huge package in R. This method and its associated package basically estimate the structure of the graph for the multivariate Gaussian distribution by using a multi-step idea. This specific calculation is called as GELATO (graph estimation with LASSO thresholding). GELATO infers the structure of the systems by two stages. In the first stage, an undirected graph is estimated via a threshold among the L_1-norm penalized regression functions, and in the second stage, the variance-covariance matrix and its inverse are estimated via their maximum likelihood estimators. In inference of the interactions, i.e., the edges of the graph, is performed as

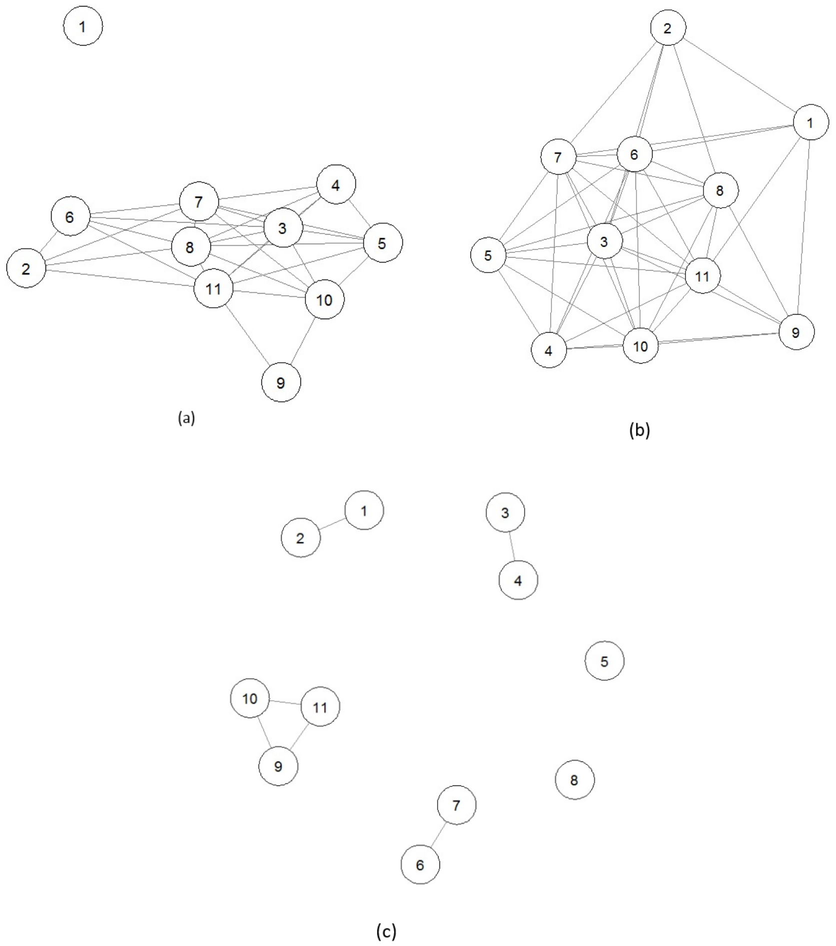


Figure 3.5: Estimated structure of the real biological network from Dataset 1 via (a) LMARS without interactions, (b) LMARS with interactions and (c) GGM.

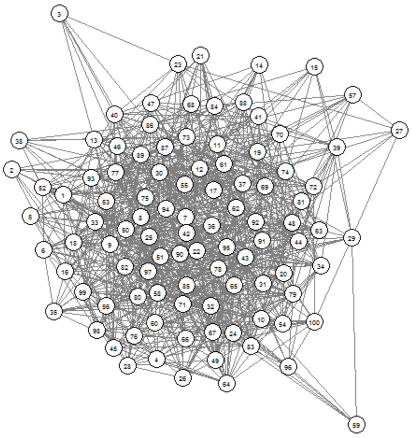


Figure 3.6: Estimated structure of the human gene interaction network from Dataset 2 by LMARS without interaction effects.

in the study of Meinshaussen and Bühlman [93] which implements the NS method within the LASSO regression to estimate the entries of the precision matrix.

Hereby, we list the biologically validated interactions of this dataset via both MARS models and GGM whose inference is conducted by the approximate glasso approach. From this assesment, we see that all three methods can detect 26 interactions which are biologically supported by the study of Bhadra and Mallick [22]. But LMARS with interactions can also find 6 more edges with respect to the findings via LMARS without the interaction model. Finally, GGM under approximate glasso can further observe 2 more links regarding LMARS with the interaction model. Therefore, similar to previous analyses, we see that the application of GGM can be limited for small and moderate dimensional systems since the estimation of the model prameters can be better performed via approximate methods for high dimensional networks.

In the analyses of LMARS with/without interactions and GGM, the R programme language is used. For the exact glasso estimate, we apply the glasso package. Whereas for the approximate glasso estimate, we perform the huge package. In the GGM analyses, the best penalty constant is choosen by the RIC criterion and glasso package is used to estimate the covariance matrix [51]. This matrix is transformed to the binary adjacency matrix and 0.5 is accepted as the cut-off value to convert the estimates into the binary 0-1 entries. Here, since GGM is used based on the pearson correlation coefficients ρ, the estimated ρ's which are greater than 0.5 are accepted as significant

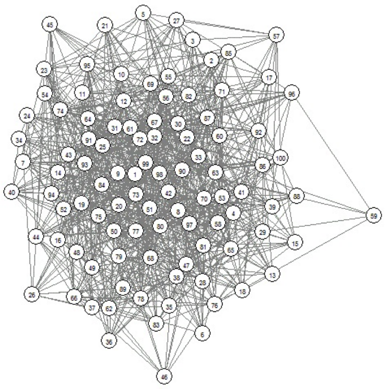


Figure 3.7: Estimated structure of the human gene interaction network from Dataset 2 by LMARS with interaction effects.

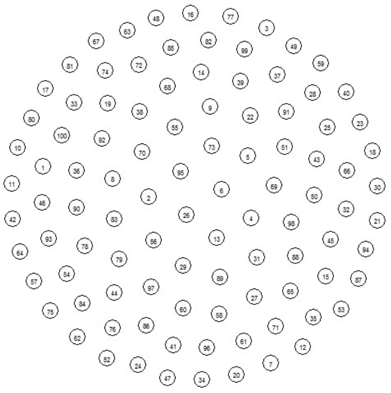


Figure 3.8: Estimated structure of the human gene interaction network from Dataset 2 by GGM.

edge, resulting in the entry 1 in the adjacency matrix. From comparative analyses, apart from the common 26 estimated links via LMARS without interactions, it is observed that the new links in the estimated graphs via LMARS with interactions have the same biologically validated link as in the study of Bhadra and Mallick [22]. In Table 3.10, we list the controlled interactions of this dataset via three methods and also give distinct outputs from each method, which are also validated biologically. From this assessment, we see that GGM failes to catch the links. On the other hand, LMARS with interactions is more successful in the estimation of the true links than LMARS without the interaction model.

Table 3.10: Comparison of the estimated links via LMARS with/without interaction effects and GGM with approximate glasso. The true links are taken from the study of Bhadra and Mallick [22].

Interactions	**LMARS with interactions**	**LMARS without interactions**	**GGM**
Common interactions between the given genes	$GI.7019408.S - GI.4504436.S$	$GI.7019408.S - GI.4504436.S$	$GI.7019408.S - GI.4504436.S$
	$GI.28610153.S - GI.4504436.S$	$GI.28610153.S - GI.4504436.S$	$GI.28610153.S - GI.4504436.S$
	$GI.20070269.S - GI.28610153.S$	$GI.20070269.S - GI.28610153.S$	$GI.20070269.S - GI.28610153.S$
	$GI.18379361.A - GI.20070269.S$	$GI.18379361.A - GI.20070269.S$	$GI.18379361.A - GI.20070269.S$
	$GI.17981706.S - GI.13514808.S$	$GI.17981706.S - GI.13514808.S$	$GI.17981706.S - GI.13514808.S$
	$GI.20302136.S - GI.7661757.S$	$GI.20302136.S - GI.7661757.S$	$GI.20302136.S - GI.7661757.S$
	$GI.4505888.A - GI.41350202.S$	$GI.4505888.A - GI.41350202.S$	$GI.4505888.A - GI.41350202.S$
	$GI.27754767.I - GI.16554578.S$	$GI.27754767.I - GI.16554578.S$	$GI.27754767.I - GI.16554578.S$
	$GI.9961355.S - GI.27754767.I$	$GI.9961355.S - GI.27754767.I$	$GI.9961355.S - GI.27754767.I$
	$GI.27754767.I - GI.27754767.A$	$GI.27754767.I - GI.27754767.A$	$GI.27754767.I - GI.27754767.A$
	$GI.22027487.S - GI.27754767.I$	$GI.22027487.S - GI.27754767.I$	$GI.22027487.S - GI.27754767.I$
	$GI.38569448.S - GI.22027487.S$	$GI.38569448.S - GI.22027487.S$	$GI.38569448.S - GI.22027487.S$
	$GI.34222299.S - GI.22027487.S$	$GI.34222299.S - GI.22027487.S$	$GI.34222299.S - GI.22027487.S$
	$GI.21614524.S - GI.34222299.S$	$GI.21614524.S - GI.34222299.S$	$GI.21614524.S - GI.34222299.S$
	$GI.37537705.I - GI.31652245.I$	$GI.37537705.I - GI.31652245.I$	$GI.37537705.I - GI.31652245.I$
	$GI.18641371.S - GI.41197088.S$	$GI.18641371.S - GI.41197088.S$	$GI.18641371.S - GI.41197088.S$
	$GI.16159362.S - GI.31652245.I$	$GI.16159362.S - GI.31652245.I$	$GI.16159362.S - GI.31652245.I$
	$GI.21389558.S - GI.16159362.S$	$GI.21389558.S - GI.16159362.S$	$GI.21389558.S - GI.16159362.S$
	$GI.28557780.S - GI.16159362.S$	$GI.28557780.S - GI.16159362.S$	$GI.28557780.S - GI.16159362.S$
	$GI.27477086.S - GI.16159362.S$	$GI.27477086.S - GI.16159362.S$	$GI.27477086.S - GI.16159362.S$
	$GI.23510363.A - GI.28557780.S$	$GI.23510363.A - GI.28557780.S$	$GI.23510363.A - GI.28557780.S$
	$GI.27482629.S - GI.23510363.A$	$GI.27482629.S - GI.23510363.A$	$GI.27482629.S - GI.23510363.A$
	$GI.28416938.S - GI.27482629.S$	$GI.28416938.S - GI.27482629.S$	$GI.28416938.S - GI.27482629.S$
	$GI.30795192.A - GI.27482629.S$	$GI.30795192.A - GI.27482629.S$	$GI.30795192.A - GI.27482629.S$
	$GI.24308084.S - GI.27477086.S$	$GI.24308084.S - GI.27477086.S$	$GI.24308084.S - GI.27477086.S$
	$GI.4504700.S - GI.19224662.S$	$GI.4504700.S - GI.19224662.S$	$GI.4504700.S - GI.19224662.S$
Different interactions between the given genes	$GI.33356162.S - GI.17981706.S$		$GI.33356162.S - GI.17981706.S$
	$GI.20373176.S - GI.14211892.S$		$GI.20373176.S - GI.14211892.S$
	$GI.17981706.S - GI.14211892.S$		$GI.17981706.S - GI.14211892.S$
	$GI.14211892.S - GI.20373176.S$		$GI.14211892.S - GI.20373176.S$
	$GI.5454143.S - GI.4504410.S$		
	$GI.27894333.A - GI.27477086.S$		
			$GI.19224662.S - GI.27477086.S$
			$GI.13027804.S - GI.34222299.S$
			$GI.22027487.S - -hmm9615.S$
			$GI.37537697.S - -GI.22027487.S$
Total number of interactions:	32	26	34

Table 3.11: Comparison of the specificity, precision and F-measure computed via MMLE and GGM under 1000 Monte-Carlo runs based on different dimensional system (Θ) with normally distributed data with sample size 20.

	precision		specifity		F		MCC	
Θ	GGM	MMLE	GGM	MMLE	GGM	MMLE	GGM	MMLE
50	0.7510	1	0.9910	1	0.5110	0.5040	0.5210	0.5690
100	0.6480	1	0.9930	1	0.4630	0.5030	0.4720	0.5730
500	0.1120	1	0.9980	1	0.1630	0.5010	0.1880	0.5770

3.3 Modified Maximum-Likelihood Estimation Application

In the application of MMLE, we estimate the precision matrix Θ under the simulated dataset and calculate the accuracy measures for this high-dimensional matrix. Here, the true precision matrix is created from the multivariate normal distribution when the total number of nodes in the systems, i.e., the dimension of Θ is taken as 50, 100 and 500, respectively, so that the performance of both approaches under small, moderate and large systems can be comparable. Furthermore, we set the number of observations for each node as 20 throughout the study and then, we construct the model as explained in Subsection 2.3.1 entitled as "Development of Modified Maximum-Likelihood Estimation". Accordingly, to compare the results of both MMLE and GGM, we calculate the same accuracy measures, namely precision, specifity, F-measure and also add the Matthew's Correlation Coefficient (MCC).

In the comparison of these measures, the mean values of the accuracy measures are calculated from 1000 Monte-Carlo run and the sample size 20 per each gene as performed before and finally, the shape parameter in Equation 2.32 under the high-dimensional density is taken as 50 to get MMLE under the multivariate normal distribution. Finally, the originally developed codes are written and run in the R programme language for both MMLE and GGM as used previously.

On the other side, in the application of GGM, the selected accuracy measures of GGM are computed under the optimal penalty constant based on the RIC criterion and the precision matrix is estimated based on the glasso method as applied before. Then, the estimated precision matrix is compared with the true adjacency matrix and we calculate the accuracy measures to calculate the power of the method.

Table 3.12: Comparison of the real computational time calculated via MMLE and GGM under 1000 Monte-Carlo runs based on different dimensional systems (p) and normally distributed data.

	User Time		System Time		Real Time	
p	GGM	MMLE	GGM	MMLE	GGM	MMLE
50	1055.53	1064.28	32.35	32.43	2 hours	13 minutes
100	2197.24	2216.05	45.92	46.02	3 hours	58 minutes
500	90512.88	209620.2	379.86	470.9	89.88hours	24.4 hours

Table 3.13: Comparison of the specificity, precision, F-measure and MCC computed via MMLE and GGM under 1000 Monte-Carlo runs based on different dimensional system (p) with lognormal distributed data with sample size 20.

	precision		specifity		F		MCC	
p	GGM	MMLE	GGM	MMLE	GGM	MMLE	GGM	MMLE
50	0.0060	1	0.9600	1	0.0200	0.5050	-0.0440	0.5690
100	0.0060	1	0.9620	1	0.0100	0.5030	-0.0270	0.5730
500	0.0020	1	0.9610	1	0.0030	0.5010	-0.0110	0.5770

From the Monte-Carlo runs seen in Table 3.11, we find that when the dimension of the matrix increases, the F-measure, MCC and the precision decrease for the GGM method. On the other hand, MCC increases under MMLE. But MMLE gives the perfect results based on the precision and the specificity which means that the proposed MMLE method can catch the links truely for all dimensions. Moreover, it is observed that MMLE overperforms under all dimesions too. On the other side, when assessing the computational time as presented in Tables 3.12, we observe that GGM is slighly faster than MMLE even though GGM spends more time in the real time which represents the calender date-time.

Table 3.14: Comparison of the specificity, precision, F-measure and MCC computed via MMLE and GGM under 1000 Monte-Carlo runs based on different dimensional system (p) with student-t distributed data with sample size 20.

	precision		specifity		F		MCC	
p	GGM	MMLE	GGM	MMLE	GGM	MMLE	GGM	MMLE
50	0.0400	1	0.9630	1	0.0360	0.5050	-0.0160	0.5690
100	0.1330	0.9970	0.9920	0.9980	0.053	0.4960	0.0510	0.5640
500	0.0110	0.9450	0.9790	0.9890	0.0180	0.4760	0.0100	0.5490

In Tables 3.13 and 3.14, it is seen that MMLE still gives better results for all dimensions and different distributions. In other words, MMLE under the long-tailed distribution fit the data better than GGM.

3.4 Bernstein Operators

After proposing new methods for the prediction of the precision matrix, we also suggest preliminary methods, such as Bernstein polynomials, Szász-Mirakyan Operator, BBH operator and MKZ operator based on the transformation to eliminate the batch effects in the original raw data so that the precision matrix can be estimated more accurately.

3.4.1 Application via Bernstein Polynomial Application

In generating simulated data we still consider 50, 100, and 500-dimensional systems where each gene has again 20 observations. We then generate scale-free, hubs, cluster, and random networks from the multivariate normal distribution by using the huge package under the R programme language [155]. In the calculation, we initially simulate a dataset for the true network and keep its true path for the best model selection in further steps. We later transform this actual dataset via the Bernstein and Szasz polynomials with respect to Equations 2.41 and 2.45, respectively. In these equations, each observation i $(i = 1, \ldots, n)$ is reallocated on the range $[0, 1]$ by using the $f(i/n)$ formula. Finally, these transformed datasets are applied in the glasso approach for the inference of Θ in GGM. In the estimation of Θ, we select the optimal penalty constant, resulting in the optimal model,via the RIC criterion as before. For the assessment, we repeat the underlying process 1000 Monte Carlo runs as before and compute precision, specificity and F-measure values for each run as stated previously and finally, present their means which are defined in Section 2.6 [112].

Moreover, as seen in Tables 3.16, we find that when the dimension of the matrix increases, the F-measure and the precision decrease, whereas, the results of both polynomials, particularly, the Szasz operator, give higher accuracies than the findings of the solely GGM approach under every dimension and most network types. Indeed the

Table 3.15: Comparison of the CPU times per second for the simulated datasets generated by 1000 Monte-Carlo runs and scale-free network types.

Dimension of Θ	Only GGM	Bernstein with GGM	Szasz with GGM
(50×50)	68.53	68.57	68.60
(100×100)	91.55	91.5800	91.61
(500×500)	222.84	222.88	222.92

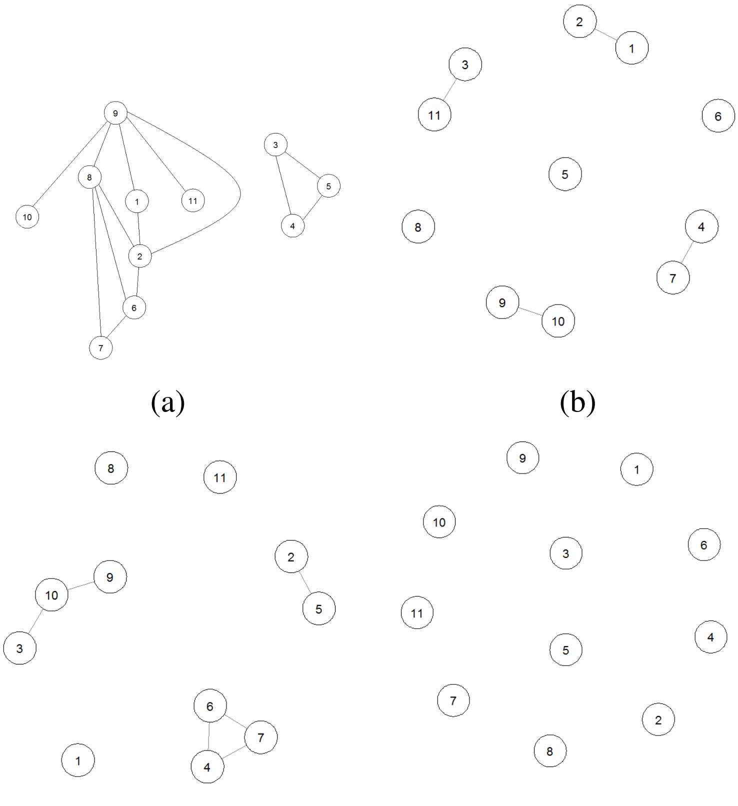
advantage of the Szasz operator over Bernstein is an expected result as the Poisson distribution applied in Szasz can be seen as the limiting case of the binomial distribution when the dimension of the systems p becomes much higher than the number of observations per gene n, that is n, i.e., $p >> n$.

On the other side, from the comparison of the computational burden based on the central processing unit (CPU) as shown in Table 3.15, we observe that all three methods have the same computational demand. This result indicates that the suggested transformation does not cause any additional computational demand even for large systems.

Later, we apply our method to the real biologically networks and use namely the cell signalling in Figure 3.4 and the large-scale human gene expression data, as used in Section 3.2 entitled as the application by LMARS with the interaction effect. For the first dataset, the graphical illustration of the accepted signalling molecular interactions is presented and the estimated systems via all alternative models are shown in Figure 3.9 with the quasi true structure of the network.

From the analyses of the first dataset, as seen in Figure 3.9, glasso approach can catch only 4 links and only the edge between PRAF-PMEK proteins (protein numbers 1 and 2 in Figure 3.9) and PKC-P38 (protein number 9 and 10 in Figure 3.9) are inferred [22] and the rest of the links cannot be proved by literature. On the other side, transformed data with Bernstein polynomial can catch 6 links and the edge between PKC-P38 (protein number 9 and 10 in Figure 3.9) and P44.42-Pakts473 (protein number 6 and 7 in Figure 3.9).

On the other hand, from comparative analyses via the human gene expression data (Dataset 2), as no links detected under the RIC criterion, we change it via StARS.



(a) (b)

(c) (d)

Figure 3.9: (a) Quasi true network of the cell signalling data (Dataset 1), (b) estimated cell signalling data by only the GGM appraoch, (c) estimated network by Bernstein polynomial with the GGM approach and (d)estimated network by Szasz polynomial with the GGM approach.

Table 3.16: Comparison of the specificity (Spec), precision (Prec) and F-measure (F) computed via only GGM, Bernstein with GGM and Szasz with GGM under scale-free, hubs, cluster and random networks based on 1000 Monte-Carlo runs for the sample size 20 per gene.

Scale-free

Dimension of Θ	GGM			Bernstein			Szasz		
	Spec	Prec	F	Spec	Prec	F	Spec	Prec	F
(50×50)	0.9991	0.5942	0.0378	0.7384	0.0403	0.0701	0.7004	0.0398	0.0705
(100×100)	1.0000	0.4790	0.0192	0.8205	0.0203	0.0365	0.7866	0.0200	0.0366
(500×500)	0.9793	0.0040	0.0066	0.9925	0.0038	0.0070	0.9997	0.0040	0.0080

Hubs

Dimension of Θ	GGM			Bernstein			Szasz		
	Spec	Prec	F	Spec	Prec	F	Spec	Prec	F
(50×50)	0.9991	0.4626	0.0411	0.7455	0.0387	0.0673	0.7026	0.0388	0.0689
(100×100)	1.0000	0.8333	0.0220	0.8391	0.0198	0.0353	0.7894	0.0202	0.0370
(500×500)	0.9784	0.0038	0.0063	0.9910	0.0038	0.0070	0.9980	0.0040	0.0080

Cluster

Dimension of Θ	GGM			Bernstein			Szasz		
	Spec	Prec	F	Spec	Prec	F	Spec	Prec	F
(50×50)	0.9991	0.1157	0.1681	0.7334	0.1155	0.1620	0.6994	0.1157	0.1681
(100×100)	1.0000	1.0000	0.0069	0.8164	0.0585	0.0892	0.7864	0.0586	0.0925
(500×500)	0.9791	0.0112	0.0141	0.9913	0.0118	0.0187	0.9997	0.011	0.0190

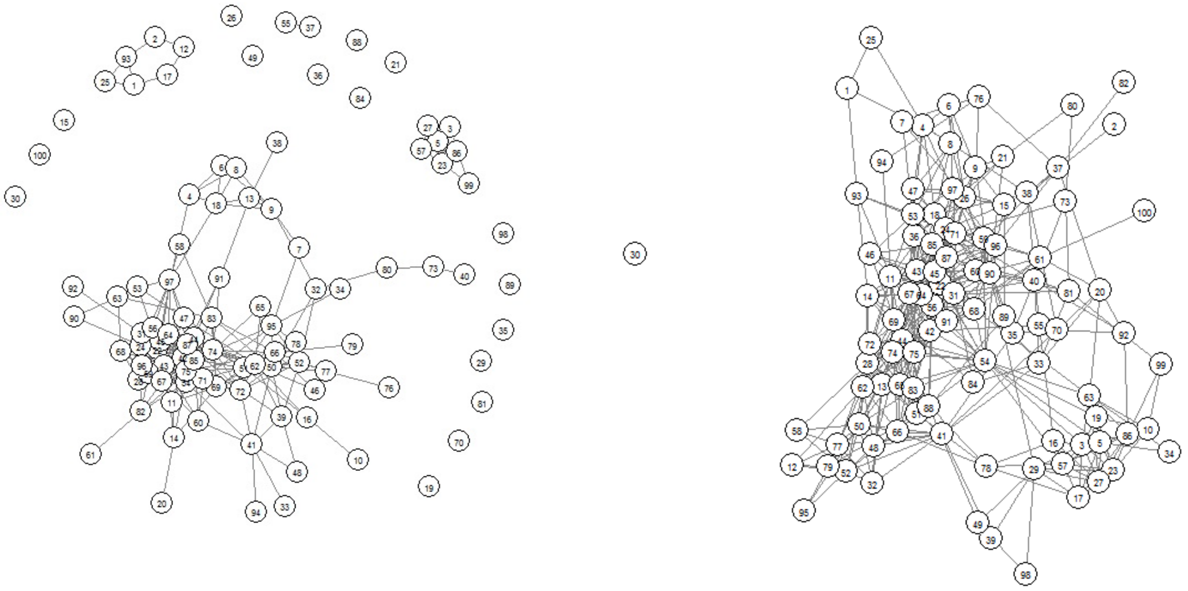
Random

Dimension of Θ	GGM			Bernstein			Szasz		
	Spec	Prec	F	Spec	Prec	F	Spec	Prec	F
(50×50)	0.9991	0.4184	0.0273	0.7342	0.0602	0.0983	0.7001	0.0602	0.1040
(100×100)	1.0000	0.333	0.012	0.8181	0.0300	0.0514	0.7858	0.0302	0.0529
(500×500)	0.9796	0.0061	0.0091	0.9912	0.0060	0.0106	0.9977	0.0057	0.0104

As seen in Table 3.17, we detect 174 links commonly from all the three approaches. But Bernstein with GGM approach further estimates 19 different links which are biologically validated by the study of Bhadra and Mallick [22] and Szasz with GGM cannot estimate any link. Thereby, the findings present that the transformed date are useful to infer the true structure of the large networks.

3.4.2 Application via Bernstein and LMARS Approaches

In the application, we show the comparison of the LMARS and GGM approaches via different estimation techniques together with the Bernstein and Szasz polynomials. For the analyses of both models, we generate 500, 900 and 1000 dimensional datasets in which each gene has 20 observations as usual. In the data generation, we arbitrarily set the off-diagonal of the precision matrix Θ to 0.9 so that the interactions between genes can be clearly observed and we generate scale-free networks [14] under the given Θ by running the huge package in the R programming language. Accordingly,



(a) (b)

Figure 3.10: (a) Estimated human gene expression data by only GGM appraoch (Dataset 2) and (b) Bernstein polynomial with the GGM approach and under the StARS criterion.

in the calculation based on the 1000 Monte-Carlo simulations, we initially produce a network structure for the true network and generate sample datasets from this true network. Then, we transform these data by the Bernstein and Szasz operators and finally, use them for modelling and inferring Θ [3].

In modelling via LMARS, every single node is implemented as a response and the remaining nodes are taken as co-variates as explained in Subsection 2.2.2. Hereby, we consider only main effects and eliminate all interaction terms. Then, we take into account the significant β parameters in Equation 2.11 to estimate Θ. These steps are repeated until every gene i is explained by the remaining other genes as the lasso regression applies. Furthermore, the forward and backward steps are performed for constructing the optimal model and the GCV criterion is calculated to eliminate overfitted coefficients. Finally we convert the estimated Θ to the binary form. To obtain a symmetric Θ, the AND rule is performed. Hereby, if the co-variate j for the lasso model with the response i is significant as well as the co-variate j for the lasso model with the response i is significant ($i, j = 1, \ldots, p$), the entries of (i, j) and (j, i) pairs in the estimated Θ can be assigned as 1 in the binary form. Otherwise, both entries,

Table 3.17: Comparison of the estimated links via transformed and non-transformed human gene expression dataset (Dataset 2). The biologically validated links and the complete name of genes can be found in the study of Bhadra and Mallick [22]. The complete list of genes is given in Appendix.

Interactions	**GGM**	**GGM with Bernstein**	**GGM with Szasz**
Common interactions between the given genes	5 -3, 6 -4, 8 -4, 8 -6, 9 -6, 14 -11, 18 -4 18 -8, 18 -13, 23 -3, 23 -5, 24 -22, 25 -1, 27 -3 27 -5, 27 -23, 31 -22, 41 -33, 41 -39, 42 -22, 42 -28 42 -31, 43 -11, 43 -22, 43 -24, 43 -31, 43 -42, 44 -11 44 -22, 44 -31, 44 -42, 44 -43, 45 -22, 45 -24, 45 -31 45 -42, 45 -43, 45 -44, 47 -24, 48 -41, 50 -32, 50 -41 52 -32, 52 -50, 53 -47, 54 -22, 54 -44, 54 -45, 56 -22 56 -28, 56 -42, 56 -43, 56 -44, 56 -45, 57 -3, 57 -5 57 -23, 57 -27, 59 -22, 59 -24, 59 -45, 59 -54, 60 -43 62 -44, 62 -50, 62 -52, 64 -22, 64 -31, 64 -42, 64 -43 64 -44, 64 -45, 64 -47, 64 -54, 64 -56, 66 -50, 66 -62 67 -11, 67 -22, 67 -24, 67 -28, 67 -43, 67 -45, 67 -56 67 -60, 67 -64, 68 -22, 68 -45, 68 -56, 69 -51, 69 -67 68 -59, 68 -63, 68 -67, 69 -11, 69 -14, 71 -11, 71 -43 71 -47, 71 -60, 71 -67, 72 -42, 72 -50, 72 -62, 73 -40 74 -42, 74 -44, 74 -50, 74 -51, 74 -54, 74 -62, 74 -65 74 -66, 74 -69, 74 -72, 75 -44, 75 -51, 75 -64, 75 -66 75 -67, 75 -68, 75 -74, 77 -50, 77 -52, 77 -62, 78 -51 78 -66, 80 -73, 83 -50, 83 -66, 83 -69, 83 -74, 83 -75 85 -22, 85 -45, 85 -56, 85 -59, 85 -60, 85 -67, 85 -71 86 -3, 86 -5, 86 -23, 86 -27, 86 -57, 87 -22, 87 -31 87 -43, 87 -45, 87 -56, 87 -59, 87 -64, 87 -67, 87 -85 90 -24, 91 -44, 91 -87, 93 -1, 95 -50, 95 -52, 96 -22 96 -24, 96 -45, 96 -54, 96 -59, 96 -68, 96 -85, 96 -87 97 -4, 97 -24, 97 -47, 97 -53, 97 -56, 99 -86	5 -3, 6 -4, 8 -4, 8 -6, 9 -6, 14 -11, 18 -4 18 -8, 18 -13, 23 -3, 23 -5, 24 -22, 25 -1, 27 -3 27 -5, 27 -23, 31 -22, 41 -33, 41 -39, 42 -22, 42 -28 42 -31, 43 -11, 43 -22, 43 -24, 43 -31, 43 -42, 44 -11 44 -22, 44 -31, 44 -42, 44 -43, 45 -22, 45 -24, 45 -31 45 -42, 45 -43, 45 -44, 47 -24, 48 -41, 50 -32, 50 -41 52 -32, 52 -50, 53 -47, 54 -22, 54 -44, 54 -45, 56 -22 56 -28, 56 -42, 56 -43, 56 -44, 56 -45, 57 -3, 57 -5 57 -23, 57 -27 59 -22, 59 -24, 59 -45, 59 -54, 60 -43 62 -44, 62 -50, 62 -52, 64 -22, 64 -31, 64 -42, 64 -43 64 -44, 64 -45, 64 -47, 64 -54, 64 -56, 66 -50, 66 -62 67 -11, 67 -22, 67 -24, 67 -28, 67 -43, 67 -45, 67 -56 67 -60, 67 -64, 68 -22, 68 -45, 68 -56, 69 -51, 69 -67 68 -59, 68 -63, 68 -67, 69 -11, 69 -14, 71 -11, 71 -43 71 -47, 71 -60, 71 -67, 72 -42, 72 -50, 72 -62, 73 -40 74 -42, 74 -44, 74 -50, 74 -51, 74 -54, 74 -62, 74 -65 74 -66, 74 -69, 74 -72, 75 -44, 75 -51, 75 -64, 75 -66 75 -67, 75 -68, 75 -74, 77 -50, 77 -52, 77 -62, 78 -51 78 -66, 80 -73, 83 -50, 83 -66, 83 -69, 83 -74, 83 -75 85 -22, 85 -45, 85 -56, 85 -59, 85 -60, 85 -67, 85 -71 86 -3, 86 -5, 86 -23, 86 -27, 86 -57, 87 -22, 87 -31 87 -43, 87 -45, 87 -56, 87 -59, 87 -64, 87 -67, 87 -85 90 -24, 91 -44, 91 -87, 93 -1, 95 -50, 95 -52, 96 -22 96 -24, 96 -45, 96 -54, 96 -59, 96 -68, 96 -85, 96 -87 97 -4, 97 -24, 97 -47, 97 -53, 97 -56, 99 -86	
Biologically validated different interactions between the given genes	9-63 , 63-53, 96-67, 12-2,72-39,62-51, 78-62	90-63,63-53,97-47,86-57,86-2,85-60,96-85,96-67,91-44,74-44 74-69,74-72,72-39,74-62,78-62,62-51, 95-50,77-50,93-25	
Total number of estimated interactions :	290	483	0

i.e., (i, j) and (j, i), are set to 0. In biological speaking, it means that there is a relation between genes when the associated entry of Θ is 1, and there is no relation between genes when this entry equals to 0.

On the other side, we apply GGM and estimate its model parameters via the NS [93] and glasso methods to extend the performance evaluation of the GGM approach. Hereby, in GGM with the NS method, the inference is performed by fitting the lasso regression. Whereas in modelling via GGM with the glasso method, we implement the lasso regression under the penalized likelihood function. In the application of GGM, firstly, the true precision matrix Θ is estimated and then the estimated Θ under the transformed data via the Bernstein operators' results are compared with the findings under the non-transformed datasets.

In the evaluation of the outcomes based on the underlying dimensional systems, we calculate the F-measure and the precision values for the measures of accuracy by using the formulas as expressed in Section 2.6.

From the outcomes in Tables 3.18 and 3.19, it is observed that F-measure via LMARS

Table 3.18: Comparison of the precision and F-measure value via LMARS under 1000 Monte-Carlo runs based on systems with 500, 900 and 1000 dimensional networks for the sample size 20 per each gene.

Accuracy measure	Dimension of Θ	Only LMARS	LMARS with Bernstein	LMARS with Szasz
Precision	500	0.0017	0.0012	0.0013
	900	0.0000	0.0005	0.0007
	1000	0.0000	0.0004	0.0005
F-measure	500	0.0029	0.0025	0.0026
	900	Not Computable	0.0013	0.0013
	1000	Not Computable	0.0011	0.0012

Table 3.19: Comparison of the precision and F-measure values via GGM estimated by the neighborhood selection (NS) and glasso methods under 1000 Monte-Carlo runs based on 500, 900 and 1000 dimensional Θ with sample size 20.

Inference Method	Accuracy measure	Dimension of Θ	Only GGM	GGM with Bernstein	GGM with Szasz
NS	Precision	500	Not Computable	0.4768	0.4731
		900	Not Computable	0.4679	0.4702
		1000	Not Computable	0.4729	0.4700
NS	F-measure	500	0.0000	0.0179	0.0173
		900	0.0000	0.0092	0.0091
		1000	0.0000	0.0227	0.0289
glasso	Precision	500	Not Computable	0.5002	0.4995
		900	Not Computable	0.4968	0.4972
		1000	Not Computable	0.4951	0.5342
glasso	F-measure	500	0.0000	0.1201	0.1531
		900	0.0000	0.1126	0.0830
		1000	0.0000	0.0775	0.1062

is not computable since the recalls are indefinite, resulting in indefinite F-measure. Whereas GGM with the NS and the glasso methods can calculate F-measure successfully. Moreover, it is seen that GGM overperforms LMARS under the transformed datasets. If we compare the findings of both Bernstein operators, it is seen that the Szasz polynomials are more accurate for all cases. Furthermore, F-measure and precision values decrease when the dimension increases under all conditions. Additionally, we find that the accuracy of the estimates under LMARS is higher when the data are not transformed via the Bernstein operators under relatively low dimensions. But when the dimension of the system raises, the transformed data have higher F-measure for both LMARS and GGM models. On the contrary, when the dimension increases, the precision value decreases too.

3.4.3 Application via Bernstein-Types of Operators

In the assessment of the polynomials' results, we consider four different scenarios. In the first scenario, we estimate the precision matrix Θ from the simulated datasets under different kinds of the Bernstein polynomials and the Bernstein-type of operators, which are the MKZ operator and the BBH operator. For the analyses of the model as used in other analyses, we generate 50, 100 and 500 dimensional datasets in which each gene has 20 observations. Then, we set all the off-diagonal entries of Θ to 0.9 arbitrarily and generate scale-free networks by using the huge package in the R programming language as performed previously. Then, in order to evaluate whether the entry of the off-diagonal terms has any effect in inference, we change it via moderately small and small values too. Hereby, in the second and the third scenarios, the off-diagonal elements of the precision matrix set to 0.7 and 0.5, respectively, when the networks are scale-free. Finally, as the fourth plan, we change the sparsity of the system from the scale-freeness to the hubs property since the former typically indicates a high sparsity level around 90% or above and the latter implies relatively a lower sparsity level at around $80-90\%$ [2].

Accordingly, in all scenarios, we initially generate a dataset for the true network and keep its true path for the best model selection in further steps. Later, we transform this actual dataset via the Berstein polynomial, the Szasz polynomial, the MKZ operator and the BBH operator. Finally, all these non-transformed and transformed data are used to estimate the precision matrices by using the NS method. In the application, we choose NS among alternatives due to its computational gain. For the assessment, we calculate the F-measure and the precision for each run as shown in Section 2.6.

Then, we repeat the calculation of the underlying statistics for 1000 Monte-Carlo runs and their means are computed. The results are presented in Table 3.20.

Table 3.20: Comparison of the F-measure and the precision values computed with-/without operators in inference of Θ via GGM with 0.9 off-diagonal entries under scale-free networks via sample size 20.

Accuracy Measure	Dimension of Θ	Only GGM	GGM with Bernstein	GGM with Szasz	GGM with BBH	GGM with MKZ
F-measure	50	0.0014	0.1714	0.1590	0.1132	0.1492
	100	0.0001	0.0904	0.0863	0.0594	0.0812
	500	0.0000	0.0171	0.0171	0.0094	0.0163
Precision	50	Not Computable	0.4852	0.4789	0.3794	0.4729
	100	Not Computable	0.4769	0.4749	0.4601	0.4699
	500	Not Computable	0.4663	0.4682	0.4466	0.4624

Table 3.21: Comparison of the F-measure and the precision values computed with-/without operators in inference of Θ via GGM with 0.7 off-diagonal entries under scale-free networks via sample size 20.

Accuracy Measure	Dimension of Θ	GGM	GGM with Bernstein	GGM with Szasz	GGM with BBH	GGM with MKZ
F-measure	50	Not Computable	0.0699	0.0711	0.0704	0.0712
	100	Not Computable	0.0361	0.0366	0.0362	0.0368
	500	Not Computable	0.0077	0.0083	0.0078	0.0080
Precision	50	Not Computable	0.0403	0.0403	0.0404	0.0401
	100	Not Computable	0.0201	0.0200	0.0200	0.0200
	500	Not Computable	0.0040	0.0040	0.0040	0.0040

Table 3.22: Comparison of the F-measure and the precision values computed with-/without operators in inference of Θ via GGM with 0.5 off-diagonal entries under scale-free networks via sample size 20.

Accuracy Measure	Dimension of Θ	GGM	GGM with Bernstein	GGM with Szasz	GGM with BBH	GGM with MKZ
F-measure	50	Not Computable	0.0703	0.0710	0.0702	0.0716
	100	Not Computable	0.0356	0.0365	0.0362	0.0367
	500	Not Computable	0.0077	0.0080	0.0077	0.0080
Precision	50	Not Computable	0.0404	0.0402	0.0402	0.0403
	100	Not Computable	0.0198	0.0199	0.0200	0.0199
	500	Not Computable	0.0040	0.0040	0.0039	0.0040

Table 3.23: Comparison of the F-measure and the precision values computed with-/without operators in inference of Θ via GGM with 0.9 off-diagonal entries under hubs networks via sample size 20.

Accuracy Measure	Dimension of Θ	GGM	GGM with Bernstein	GGM with Szasz	GGM with BBH	GGM with MKZ
F-measure	50	Not Computable	0.0668	0.0687	0.0677	0.0691
	100	Not Computable	0.0352	0.0365	0.0358	0.0366
	500	Not Computable	0.0074	0.0076	0.0075	0.0076
Precision	50	Not Computable	0.0383	0.0387	0.0386	0.0387
	100	Not Computable	0.0196	0.0198	0.0198	0.0198
	500	Not Computable	0.0038	0.0038	0.0038	0.0038

From Table 3.20, it is seen that for low dimensions, the Bernstein polynomials give better results than others, in particular, the Szasz polynomials have the highest F-measure. We obtain the same results for the precision values as well. Moreover, we detect that when the dimension of the matrix increases, the F-measure and the precision value decrease. Furthermore, as shown in Tables 3.21-3.22, we observe similar findings in the sense that the Szasz polynomials typically produce better F-measure and precision even though we decrease the correlation between genes (by off-diagonal entries 0.7 and 0.5). Whereas under these scenarios, it is found that the MKZ operator is as good as the Szasz operator in terms of the accuracy of the estimates under certain conditions. Finally, when we evaluate the outputs of Tables 3.21-3.23, we see that the results of both the Szasz polynomial and the MKZ operator are very close to each other and overperfom with respect to the remaining operators.

On conclusion, all these outputs imply that when the sparsity of networks decreases, all Bernstein-type of operators compute similar results and the Szasz polynomial as well as the MKZ operator are slightly better than others. On the contrary, if the sparsity level raises as mostly observed in biological networks, the operators have different accuracy values and the performance of the Bernstein polynomials, especially, the Szasz polynomial, becomes better.

CHAPTER 4

CONCLUSION

In this thesis, we have proposed the LMARS nonparametric regression method, as an alternative of the GGM modelling, to construct the networks and to describe the nonlinearity in the systems. In the comparison of both approaches, we have implemented simulation studies under different scenarios, systems dimensions and sparsities. From the Monte-Carlo results, we have evaluated the specificity, F-measure and the precision values and have concluded that when the dimension and the sparsity of the systems increase, the underlying criteria become higher. Moreover, the computational time of LMARS is less demanding than GGM for all dimensions. We have found similar outputs in the application of both methods in real systems' analyses too. On the other hand, from the comparison of both approaches in nonnormal data, we have observed that LMARS mostly performs better than GGM. Thereby, we consider that LMARS can be seen as a promising alternative approach regarding GGM both in terms of accuracy and computational cost, especially, for the construction of large networks.

On the other side, as the future studies, we think to apply other alternative approaches of GGM within their parametric alternatives in order to relax the strick normality assumption. We consider to apply copula models for this purpose as these approaches can deal with the challenge of the normality by suggesting different distributions to describe the measurement of genes [57, 38]. In order to eliminate the normality assumption, Liu et al. [88] also suggest the nonparanormal SKEPTIC algorithm which is based on the nonparametric optimization method in which the joint function of states is defined as the univariate normal by using the transformed data. But it has been shown that for the continuous distributions, the nonparanormal family is taken as the equivalent to the Gaussian copula family [87]. On the other hand, Voorman et al. [138] propose to fit a generalized additive model by using a penalty value that estimates the optimal basis function in an additive model. The estimation is implemented via the block coordinate descent algorithm. Regarding other suggested approaches, it has been found that this method is successful in modelling the nonlinearity and competitive in the description of the linear interactions in the systems.

As an another extension of this study, we consider to apply extended modelling approaches of MARS. One of these alternatives is the Conic MARS (CMARS) approach [144] which is enhanced as an alternative of the MARS backward step. This method simply suggests to implement a penalized residual sum of squares (PRSS) shown in Equation 4.1 for MARS as a ridge regression, also known as the Tikhonov regularization [6, 61], by eliminating the backward stepwise algorithm of MARS. Moreover, CMARS chooses knots t more far from the input variables x_{ij} for all i and j. These calculations improve the data fitting and produce a better R^2 measure than MARS obtains.

$$PRSS(\alpha, f_1, f_2, \dots, f_p) = \sum_{i=1}^{N}(y_i - \alpha - \sum_{j=1}^{p} f_j x_{ij})^2 + \sum_{j=1}^{p} \lambda_j \int f_j^{''}(t_j)^2 dt_j, \quad (4.1)$$

in which $\lambda_j > 0$ is the tuning parameter [63]. Furthermore, α refers to the coefficient of the mth basis function and $f^{''}$ indicates the second derivative of the basis function. Hence, PRSS can control the complexity and the accuracy of the model and can transform MARS to the Tikhonov regularization problem, resulting in the conic quadratic programming (CQP) in the parameter estimation. The boundaries of this programming is determined by different types of the multi-objective optimization approaches. On the other hand, robust conic MARS (RCMARS) [102] extends CMARS by adding the first and the second-order partial derivatives of the multivariate basis functions, which come from the discretization of the integral while computing the optimal model. Such penalized curvature structure, which can be considered as the generalization of CMARS [103], is called the robustification of MARS and the parameter estimation of this model can be performed by the CQP [128]. From sensitivity analyses in different datasets, it has been found that RCMARS improves the accuracy of MARS with a loss in the computational demand. But this demand can be decreased via the Robust MARS(RMARS) approach [106] which also finds less variance in the estimates of parameters although it has slightly lower accuracy than MARS.

On the other side, it is possible to partition the function used in the knot selection via linear and nonlinear components. The conic generalized partial linear (CGPL) model [32, 105] is a semiparametric approach which uses CMARS for nonlinear variables and the logistic regression (LR) in linear variables [143]. We can also perform the robust conic generalized partial linear (RCGPL) model which implements RCMARS and LR approaches to present nonlinear and linear variables, respectively, in a partially linear model [104, 107]. Finally, the dynamic structure of the networks can be also inferred via the numerical solutions whose model is based on the ordinary differential equations (ODE) [54, 140, 141]. In the parameter estimation of this model, various techniques such as the Euler approximation [53], higher ordered-Heun and Runge-Kutta approximations [40, 152, 142] as well as optimization methods based on the ellipsoidal calculus [78] can be performed. Because from the study of Defterli et al. [33], it has been shown that the results of GGM can be comparable with the ODE approach in order to estimate the gene-environment networks which describe the relations of environmental conditions on the individual genotypes. Whereas, in that study, the results are assessed based on a small system. Hence, their applications can be extended by performing them in large systems as used in this work.

On conclusion, we see that there are lots of probabilistic and deterministic techniques that can be performed to get optimal results in high-dimensional analyses and these studies can be adapted to the problems in computational biology, bioinformatics and medicine, in particular, when the question of interest is based on the penalized type of models. Because it is known that this sort of problems can be solved via both numerical solutions and iterative approaches where various types of statistical inference methods can be also applicable such as the iterative maximum likelihood approach or Bayesian algorithms.

On the other hand, in the extension of LMARS method, we have implemented the LMARS method with interaction terms as an alternative of GGM. The performance of this new model has been evaluated via simulated datasets and two real datasets with biological interpretations of the results. From the outcomes, we have concluded that if the interaction terms are included into the model, it can detect new links which have not been mostly estimated by other methods even though their findings are biologically declared. Thereby, we believe that this extended version of LMARS can be a promising alternative of GGM and estimates the true structure of the system better than the LMARS without interaction effects. Furthermore, we have seen that the application of GGM can be limited for high dimensional systems due to its calculation of the inference. We can perform some approximate methods under this condition. Whereas, both LMARS models do not have such a limitation and can be applicable for realistically complex biological networks.

As a future study of LMARS with interaction effect, precise structural contact prediction using sparse inverse covariance estimation (PSICOV) which use sparse inverse covariance estimation to the problem of protein contact prediction, can used to estimate the precision matrix[74]. We would hope to apply BIG and QUIC method which include a block-coordinate descent method with the blocks via clustering [66]. We consider to use other alternative methods of MARS, known as POLYMARS [126], hybrid adaptive splines (HAS) [90], Bayesian MARS [35, 36], and SARS [158], besides conic-multivariate adaptive regression splines (CMARS) [144] and RCMARS as stated before [102] by converting them as a lasso regression in order to accelerate the speed of calculations and to increase the accuracies of the estimated systems since their advantages over the full description of MARS have been indicated. Moreover, we aim to compare their results with the findings of QUIC [67] and BIGQUIC [66] approaches since these methods are specifically designed for very high dimensional data under high dependency and sparsity.

On the other side, in the Bernstein polynomial application and Bernstein-LMARS, we have proposed an approximation method, called the Bernstein polynomials, in advance of the inference via the GGM and LMARS by transforming the measurements into the $[0, 1]$ interval. In our calculation, we have implemented the Bernstein and Szasz operators, and compared their performances. In the comparative assessment, we have generated different dimensional networks under distinct sparsities. Then, we have computed their precisions, specificities and F-measure values based on 1000 Monte-Carlo runs. From the results, we have observed that the polynomial approaches, specifically, the Szasz method, increase the accuracies of the estimates under most of these cases and it can be used to eliminate batch effect in the data if the modelling is based on the description of the steady-state behaviour of the biochemical systems. Furthermore, from the application with real datasets, we have seen that the polynomial approach is promising to support the known biological findings about the selected systems.

Hereby as the future study, we think to compare these results with other types of deterministic modelling approaches such as the ordinary differential equations approach and check whether these polynomials are still successful in improving the accuracy of the estimates. Moreover, we also consider to perform the extended version of the Bernstein polynomial, called the q-Bernstein approximation [101] which has a higher convergence rate and is asymptotically more efficient. Additionally, the application of the multivariate Bernstein Polynomials [81] can be implemented in order to evaluate the outcomes under univariate and multivariate dimensional observations. Finally, we consider to compare the performance of Bernstein polynomials with spline methods that are performed under penalized regression [116]. Because from the previous studies, it has been reported that these methods are successful in data which indicate distributional feature and our analyses of the MARS (multivariate adaptive regression) approach [46] that is based on the spline functions, support this finding by computing promisingly accurate results when it is compared with the glasso approach in GGM [8].

Finally, in the study of the Bernstein-types of operators' application, we have compared all well-known Bernstein-type of operators to detect which alternative produces the highest accuracy if it is applied with GGM. For this purpose, we have analyzed the Monte-Carlo results of the Bernstein polynomials, BBH and MKZ operators. The results have indicated that the Bernstein polynomials have the highest accuracies if they are performed in advance of the inference of the model parameters of GGM and the MKZ operator can be another good alternative if the sparsity level of the system decreases. But for all choices of operators, we have found that these operators can improve the accuracies of estimates for different dimensional biochemical systems if they are implemented as the pre-processing step before the inference of the precision matrix. As the future work of this study, we consider other approximation methods such as the empirical copula or Fourier transformations in such a way that the distributional feature of the observations can be embedded to the new transformed data in order to smooth the original dataset smartly and get estimates with higher accuracy. Indeed from our preliminary studies via the empirical copula, which is based on the transformation of the original data to the normal distribution, we have already found that it improves the accuracy of the estimates under both normal and nonnormal datasets and can be a good pre-processing step for the biological systems particularly if the dimension of the network is very large such as least higher than 500 genes.

REFERENCES

[1] Abel, U. (1995). The Moments for the Meyer-König and Zeller Operators. *Journal of Approximation Theory*, 82, 352–361.

[2] Ağraz, M. and Purutçuoğlu, V. (2016). Different Types of Bernstein Operators in Inference of Gaussian Graphical Model. *Cogent Mathematics*, 3:1154706.

[3] Ağraz, M. and Purutçuoğlu, V. (2016). Transformations of Data in Deterministic Modeling of Biological Networks. In *Intelligent Mathematics II: Applied Mathematics and Approximation Theory*, Eds. Anastassiou, G., Duman, O., 343–356, Springer.

[4] Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Boca Raton: Chapman and Hall/CRC.

[5] Aral, A. and Gupta, V. (2006). The q-Derivative and Applications to q-Szasz Mirakyan Operators. *Calcolo*, 43 (3), 151–170.

[6] Aster, B. and Teboulle, M. (2004). *Parameter Estimation and Inverse Problems*. New York: Academic Press.

[7] Augugliaro, L., Mineo, A.M., and Wit, E.C. (2013). Differential Geometric Least Angle Regression: A Differential Geometric Approach to Sparse Generalized Linear Models. *Statistical Methodology*, 75(3), 471–498.

[8] Ayyıldız, E., Ağraz, M., and Purutçuoğlu, V. (2016). MARS as the Alternative Approach of GGM in Modelling of Biochemical Systems, *Journal of Applied Statisics, 1-19.* URL: http://www.tandfonline.com/doi/full/10.1080/02664763.2016.1266465 (visited on 28/06/2017).

[9] Ayyıldız, E. (2013). *Gaussian Graphical Models in Estimation of Biological Systems*. MS Thesis, Middle East Technical University.

[10] Babu, G., Canty, A., and Chaubey, Y. (2002). Application of Bernstein Polynomials for Smooth Estimation of a Distribution and Density Function. *Journal of Statistical Planning and Inference*, 105, 377–392.

[11] Bakin, S., Hegland, M., and Osborne, M.R. (2000). Parallel MARS Algorithm Based on B-Splines. *Computational Statistics*, 15, 463–484.

[12] Banerjee, O., Ghaoui, L., and D'Aspremont, A. (2008). Model Selection Through Sparse Maximum Likelihood Estimation. *Journal of Machine Learning Research*, 9, 485–516.

[13] Barabasi, A.L. and Albert, R. (1999). Emergence of Scaling in Random Networks, *Science*, 286(5429), 509–512.

[14] Barabasi, A.L. and Oltvai, Z.N. (2004). Network Biology: Understanding the Cell's Functional Organization. *Nature Reviews Genetics*, 5, 101–113.

[15] Barrientos, A.F., Jara, A., and Quintana, F.A. (2017). Fully Nonparametric Regression for Bounded Data Using Dependent Bernstein Polynomials. *Journal of the American Statistical Association*, 1–20.

[16] Beck, A. and Teboulle, M. (2000). Global Optimality Conditions for Quadratic Optimization Problems with Binary Constraint. *SIAM Journal on optimization*, 11, 179–188.

[17] Beck, A. and Teboulle, M. (2009). Gradient Based Algorithms with Applications in Signal Recovery Problems in Convex Optimization in Signal Processing and Communications. In *Convex optimization in signal processing and communications*, Eds. Palomar, D.P., Eldar, Y.C., 33–88, Cambridge University Press.

[18] Beck, A. and Teboulle, M. (2012). Smoothing and First Order Methods: A Unified Framework. *SIAM Journal on Optimization*, 22, 557–580.

[19] Bernstein, S. (1912). Démonstration du théorème de Weierstrass Fondée Sur le Calcul des Probabilities. *Comminication of the Kharkov Mathematical Society*, 2(13), 1–2.

[20] Bezier, P. (1966). Definition Numerique des Courbes et Surfaces I. *Automatisme*, XI, 625–632.

[21] Bezier, P. (1967). Definition Numerique des Courbes et Surfaces II. *Automatisme*, XII, 17–21.

[22] 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.

[23] Bleimann, G., Butzer, P.L., and Hahn, L. (1980). A Bernstein-Type Operator Approximating Continuous Functions on the Semi-Axis. *Indagationes Mathematicae*, 42 , 255–262.

[24] Bottero, V., Sharma-Waliaa, N., Kerura, N., Paula, A.G., Sadagopana, S., Cannonb, M., and Chandran, B. (2009). Kaposi Sarcoma-Associated Herpes Virus (KSHV) G Protein-Coupled Receptor (vGPCR) Activates the ORF50 Lytic

Switch Promoter: a Potential Positive Feedback Loop for Sustained ORF50 Gene Expression. *Virology*, 392, 34–51.

[25] Bower, J.M. and Bolouri, H. (2001). *Computational Modeling Genetic and Biochemical Networks*. Cambridge: Massachusetts Institute of Technology Press.

[26] Butzer, P.L. and Karsli, H. (2009). Voronovskaya-Type Theorems for Derivatives of the Bernstein-Chlodovsky Polynomials and the Szasz-Mirakyan Operator. *Commentationes Mathematicae*, 49, 33–58.

[27] Chen, V.C.P. (1993). *Applying Experimental Design and Regression Splines to High-Dimensional Continuous-State Stochastic Dynamic Programming*. PhD Thesis, Cornell University.

[28] Chen, L., Emmert-Streib, F., and Storey, J. (2007). Harnessing Naturally Randomized Transcription to Infer Regulatory Relationships Among Genes. *Genome Biology*, 8, R219.

[29] Cheney, E.W. and Sharma, A. (1964). Bernstein Power Series. *Canadian Journal of Mathematics*, 16, 241–252.

[30] Craven, P. and Wahba, G. (1979). Smoothing Noisy Data With Spline Functions. Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation. *Numerische Mathematik*, 31, 317–403.

[31] Cunningham, J., Estrella, V., Lloyd, M., Gillies, R., Frieden, B.R., and Gatenby, R. (2012). Intracellular Electric Field and pH Optimize Protein Localization and Movement. *PLoS One*, 7, 1–12.

[32] Çelik, G. (2010). *Parameter Estimation in Generalized Linear Models with Conic Quadratic Programming*. MS Thesis, Middle East Technical University.

[33] Defterli, Ö., Purutçuoğlu, V., and Weber, G.W. (2014). Advanced Mathematical and Statistical Tools in the Dynamic Modeling and Simulation of Gene-Environment Regulatory Networks . In *Modeling, Optimization, Dynamics and Bioeconomy*, Eds. Zilberman, D., Pinto, A., 235–257, Springer-Verlag.

[34] Dempster., A.P. (1972). Covariance Selection. *Biometrics*, 28(1), 157–175.

[35] Denison, D.G.T, Mallick, B.K., and Smith, A.F.M. (1998). Automatic Bayesian Curve Fitting. *Journal of Royal Statistical Society*, 60, 333–350.

[36] DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001). Bayesian Curve Fitting with Free-Knot Splines. *Biometrika*, 88, 1055–1071.

[37] Dobra, A., Eicher, T., and Lenkoski, A. (2010). Modeling Uncertainty in Macroeconomic Growth Determinants Using Gaussian Graphical Models. *Statistical Methodology*, 7, 292–306.

[38] Dobra, A. and Lenkoski, A. (2011). Copula Gaussian Graphical Models and Their Application to Modeling Functional Disability Data. *Annals of Applied Statistics*, 5 969–993.

[39] Drton, M. and Perlman, M.D. (2008). A SINful Approach to Gaussian Graphical Model Selection. *Journal of Statistical Planning and Inference*, 138(4) 1179–1200.

[40] Dubois, D.M. and Kalisz, E. (2004). Precision and Stability of Euler, Runge-Kutta and Incursive Algorithm for the Harmonic Oscillator. *International Journal of Computing Anticipatory Syststems*, 14, 21–36.

[41] Erdös, P. and Rényi, A. (1959). On Random Graphs. *Publicationes mathematicae (Debrecen)*, 6, 290–297.

[42] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, 96(456), 1348–1360.

[43] Fisher, N.I. (1997). Copulas. In *Encyclopedia of Statistical Sciences*, Eds. Kotz, S., Read, C. B., Banks, D. L., 1, 159-163, NewYork: John Wiley and Sons.

[44] Fréchet, M. (1951). Sur les Tableaux de Corrélation dont les Marges son Données. Annales de l'Universite de Lyon, 14, 53–77.

[45] Frey, R., McNeil, A.J., and Embrechts, P. (2005). *Quantitative Risk Management*. Princeton: Princeton University Press.

[46] Friedman, J.H., Stuetzle, W. (1981). Projection Pursuit Regression. *Journal of the America Statistical Association*, 76(376), 817–823.

[47] Friedman, J.H. (1991). Multivariate Adaptive Regression Splines. *The Annual of Statistics*, 19 (1), 1–67.

[48] Friedman, J.H., Grosse, E., and Stuetzle, W. (1991). Flexible Parsimonious Smoothing and Additive Modelling. *American Statistical Association and the American Society for Quality Control*, 31(1), 3–21.

[49] Friedman, J.H. and Silverman, B. (1991). Multidimensional Additive Spline Approximation. *SIAM Journal on Scientific and Statistical Computing*, 4(2), 291–301.

[50] Friedman, J.H., Hastie, T., and Tibshirani, R. (2008). Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, 9(3), 432–441.

[51] Friedman, J., Hastie, T., and Tibshirani, R. (2014). *glasso: Graphical Lasso-Estimation of Gaussian Graphical Models*. R package version 3.23. URL: http://CRAN.R-project.org/package=glasso (visited on 28/06/2017).

[52] Gamerman, D. and Lopes, H.F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. New York: Chapman and Hall/CRC.

[53] Gebert, J., Lätsch, M., Pickl, S.W., Weber, G.W., and Wünschiers, R. (2004). Genetic Networks and Anticipation of Gene Expression Patterns. *AIP Conference Proceeding*, 718, 474–485.

[54] 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.

[55] Genest, C. and Mackay, J. (1986). The joy of copulas: Bivariate Distributions with Uniform Marginals. *The American Statistician*, 40, 280–283.

[56] Gillespie, D.T. (1977). Exact Stochastic Simulation of Coupled Chemical Reactions. *The Journal of Physical Chemistry*, 81, 2340–2361.

[57] Genest, C. and Favre, A.C. (2007). Everything You Always Wanted to Know About Copula Modeling but Were Afraid to Ask. *Journal of Hydrologic Engineering*, 12, 347–368.

[58] Ghosal, S. (2001). Convergence Rates for Density Estimation with Bernstein Polynomials. *The Annals of Statistics*, 29, 1264–1280.

[59] Golightly, A. and Wilkinson, D.J. (1998). Bayesian Inference for Stochastic Kinetic Models Using a Diffusion Approximation. *Biometrics*, 61, 781–788.

[60] Gonska, H. and Pitul, P. (2005). Remarks on an article of J.P. King. *Commentationes Mathematicae Universitatis Carolinae*, 46, 645–652.

[61] Hansen, P.C. (1998). *Rank-Deficient and Discrete ill-Posed Problems: Numerical Aspects of Linear Inversion*. Philadelphia: Society for Industrial and Applied Mathematics.

[62] Harville, D.A. (1997). *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

[63] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning, Second Edition*. New York: Springer.

[64] Hermann, T. (1991). On the Operator of Bleimann, Butzer and Hahn. In *Approximation Theory*, Eds. Szbados, J. et al., 355–360, North-Holland Publishing Company.

[65] Hoschek, J. and Lasser, D. (1993). *Fundamentals of Computer Aided Geometric Design*. Massachusetts: Taylor and Francis.

[66] Hsieh, C.J, Sustik, M., Dhillon, I.S., Ravikumar, P., and Poldrak, R.A. (2013). Advances in Neural Information Processing Systems. *26 (NIPS)*, 3165-3173.

[67] Hsieh, C.J, Sustik, M., Dhillon, I.S., and Ravikumar, P. (2014). QUIC: Quadratic Approximation for Sparse Inverse Covariance Estimation. *Journal of Machine Learning Research*, 15, 2911–2947.

[68] Huang, S. (2012). Gene Expression Profiling, Genetic Networks, Cellular States: An integrating Concept for Tumorigenesis and Drug Discovery. *Journal of Molecular Medicine*, 77(6), 469–480.

[69] Hutchinson, T.P. and Lai, C.D. (1990). *Continuous Bivariate Distributions Emphasising Applications*. Adelaide: Rumsby Scientific Publishing.

[70] Islam, M.Q. (2014). Estimation in Multivariate Normal Distributions with Stochastic Variance Function. *Journal of Computational and Applied Mathematics*, 255, 698-714.

[71] Jayasri, C. and Sitaraman, Y. (1993). On a Bernstein-Type Operator of Bleimann-Butzer and Hahn. *Journal of Computational and Applied Mathematics*, 47, 267–272.

[72] Jenner, G.R., Alba, M.M., Boshoff, C., and Kellam, P. (2001). Kaposi's Sarcoma-Associated Herpesvirus Latent and Lytic Gene Expression as Revealed by DNA Arrays. *Journal of Virology*, 891–902.

[73] Joe, H. (1997). *Multivariate Models and Dependence Concepts*. London: Chapman and Hall.

[74] Jones, D., Buchan, D.W., Cozzetto, D., and Pontil, M. (2012). PSICOV: Precise Structural Contact Prediction Using Sparse Inverse Covariance Estimation on Large Multiple Sequence Alignments. *Bioinformatics*, 28(2), 184-190.

[75] Khan, R.A. (1988). A Note on a Bernstein-Type Operator of Bleimann, Butzer and Hahn. *Journal of Approximation Theory*, 53, 295–303.

[76] Khan, K.A. (1991). Some Properties of a Bernstein-Type Operator of Bleimann, Butzer and Hahn. In *Progress in Approximation Theory*, Eds. Nevai, P., Pinkus, A., 497–504, Academic Press.

[77] Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models Principles and Techniques*. Massachusetts: MIT Press.

[78] Kropat, E., Weber, G.W., and Rückmann, J. (2010). Regression Analysis for Clusters in Gene-Environment Networks Based on Ellipsoidal Calculus and Optimization. *Dynamics of Continuous, Discrete and Impulsive Systems Series*, 17(5), 639–657.

[79] Kuhnert, P.M., Do, K.A., and McClure, R. (2000). Combining Non-Parametric Models with Logistic Regression: An Application to Motor Vehicle Injury Data. *Computational Statistical and Data Analyses*, 34, 371–386.

[80] Kuo, J.F. (1994). *Protein Kinase C*. New York: Oxford University Press.

[81] Lai, M. (1992). Asymptotic Formulae of Multivariate Bernstein. *Journal of Approximation Theory*, 70, 229–242.

[82] Lauritzen, S.L. (1996). *Graphical Models*. New York: Oxford University Press.

[83] Leathwick, J.R., Rowe, D., Richardson, J., Elith, J., and Hastie, T. (2005). Using Multivariate Adaptive Regression Splines to Predict the Distributios of New Zealand's Freshwater Diadromous Fish. *Freshwater Biology*, 50 , 2034–2051.

[84] Lethwick, J.R., Elith, J., and Hastie, T. (2006). Comparative Performance of Generalized Additive Models and Multivariate Adaptive Regression Splines for Statistic Modelling of Species Distributions. *Ecological Modeling*, 199, 188-196.

[85] Li, H. and Gui, J. (2006). Regularized Estimation for Gaussian Graphical Models, with Applications to Inference of Genetic Networks. *Nonparametric Statistics*, 7(2), 302–317.

[86] Liu, Y., Kosut, O., and Wilsky, A. (2013). Sampling from Gaussian Graphical Models Using Subgraph Perturbations. *Proceedings of the 2013 IEEE International Symposium on Information Theory*, 2498–2502.

[87] Liu, H., Lafferty, J., and Wasserman, L. (2012). The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs. *Journal of Machine Learning Research*, 7, 2295–2328.

[88] Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High Dimensional Semiparametric Gaussian Copula Graphical Models. *Annals of Statistics*, 40, 2293–2326.

[89] Lorentz, G.G. (1953). *Mathematical Exposition No.8: Bernstein Polynomials*. Toronto: University of Toronto Press.

[90] Lou, Z. and Wahba, G. (1997). Hybrid Adaptive Splines. *Journal of the American Statistical Association*, 92, 107—116.

[91] Mahmudov, N.I. (2011). q-Szasz Operators which Preserve x^2. *Journal of Computational Applied Mathematics*, 235(16), 4621-4628.

[92] 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, 4741–54.

[93] Meinshaussen, N. and Buhlmann, P. (2006). High Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics*, 34(3), 1436–1462.

[94] Mercer, A. (1989). A Bernstein-Type Operator Approximating Continuous Functions on the Half-line, *Bulletin of Calcutta Mathematical Societiy*, 31, 133–137.

[95] Meyer-König, W. and Zeller, Z. (1960). Bernsteinche Potenzreihen. *Studia Math*, 19, 89–74.

[96] Stephen Milborrow, S. (2017). *earth: Multivariate Adaptive Regression Spline*. R package verson 3.23. URL:https://cran.r-project.org/web/packages/earth/index.html (visited on 28/06/2017).

[97] Mohammadi, A. and Wit, E. (2015). BDgraph: An R Package for Bayesian Structure Learning in Graphical Models. *Bayesian Analysis*, 10 (1), 109–138.

[98] Montgomery, D.C., Peck, E.A., and Vining, G.G. (2001). *Introduction to Linear Regression Analysis*, New York: Wiley.

[99] Munoz, J. and Felicisimo, A.M. (2005). Comparison of Statistical Methods Commonly Used in Predictive Modelling. *Journal of Veterinary Science*, 15, 285–292.

[100] Nelsen, R.B. (1999). *An Introduction to Copulas*. New York: Springer.

[101] Ostrovska, S. (2003). q-Bernstein Polynomials and Their Iterates. *Journal of Approximation Theory*, 123, 232–255.

[102] Özmen, A., Weber, W.G., and Batmaz, İ. (2010). The New Robust CMARS (RCMARS). *ISI Proceeding of 24th MEC-EurOPT*, 362–368.

[103] Özmen, A., Weber, W.G., and Karimov, A. (2011). Robust Conic Generalized Partial Linear Models Using RCMARS Method-A Robustification of CGPLM. *AIP Conference Proceedings*, 1499, 337.

[104] Özmen, A., Weber, W.G., Batmaz, İ., and Kropat, R. (2011). RCMARS: Robustification of CMARS with Different Scenarios Under Polyhedral Uncertainty Set. *Communications in Nonlinear Science and Numerical Simulation*, 16(12), 4780–4787.

[105] Özmen, A., Weber, W.G., 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.

[106] Özmen, A., Weber, G.W., and Karimov, A. (2014). RMARS: Robustification of Multivariate Adaptive Regression Spline Under Polyhedral Uncertainty. *Journal of Computation and Applied Mathematics*, 259(B), 914–924.

[107] Özmen, A., Kropat, E., and Weber, G.W. (2016). Robust Optimization in Spline Regression Models for Multi-Model Regulatory Networks Under Polyhedral Uncertainty. *Optimization*, 1–21 (preprint).

[108] Petrone, S. (1999). Bayesian density estimation using Bernstein polynomial posteriors. *Canadian Journal of Statistics*, 27, 105–126.

[109] Petrone, S. and Wasserman, L. (2002). Consistency of Bernstein polynomial posteriors. *Journal of the Royal Statistical Society B*, 64, 79–100.

[110] Psichogios, D.C., De Vaux, R.D., and Ungar, L.H. (1992). Non Parametric System Identification: A Emprical Comparison of MARS and Neurol Networks. *American Control Conference*, 1436–1441.

[111] Purutçuoğlu, V., Ayyıldız, E., and Wit, E. (2016). Comparison of Two Inference Approaches in Gaussian Graphical Models. *Turkish Journal of Biochemistry*, 1–18.

[112] Purutçuoğlu, V., Ağraz, M., and Wit, E. (2017). Bernstein Approximations in Glasso-Based Estimation of Biological Networks. *The Canadian Journal of Statistics*, 45 (1), 62–76.

[113] Ravikumar, P., Wainwright, M.J., and Lafferty, J.D. (2010). High-Dimensional Ising Model Selection Using L_1-Regularized Logistic Regression. *The Annals of Statistics*, 38, 1287–1319.

[114] Rempulska, L. and Graczyk, S. (1998). On Generalized Szasz–Mirakyan Opeartors of Functions of Two Variables. *Mathematica Slovaka*, 62, 87–98.

[115] Rényi, A. (1959). On Measures of Dependence. *Acta Mathematica Academiae Scientiarum Hungar*, 10, 441–451.

[116] Ruppert, D., Wand M.P., and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

[117] 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.

[118] Schweizer, B. and Wolff, E.F. (1981). On Nonparametric Measures of Dependence for Random Variables. *The Annals of Statistics*, 9(4), 879–885.

[119] Shih, D.T. (2006). *Convex Version of Multivariate Adaptive Regression Spline and Implementations for Complex Optimization Problem*. PhD Thesis, University of Texas.

[120] Shmulevich, I., Dougherty, E.R., Seungchan, K., and Zhang, W. (2002). Probabilistic Boolean networks: a Rule-Based Uncertainty Model for Gene Regulatory Networks. *Bioinformatics*, 18, 261-274.

[121] Sim, A.T.R. and Scott, J.D. (1999). Targeting of PKA, PKC and Protein Phosphatases to Cellular Microdomains. *Cell Calcium*, 26(5), 209–217.

[122] Sklar, A. (1959). Fonctions de Répartition a n Dimensions et Leurs Marges. *Publications de l Institut de Statistique de L Universite de Paris*, 8, 229-231.

[123] Stadtmuller, U. (1986). Asymptotic Properties of Nonparametric Curve Estimates. *Periodica Mathematica Hungarica*, 17, 83-10.

[124] Stranger, B., Nica, A., Forrest, M., Dimas, A., Bird, C., Beazley, C., Ingle, C., Dunning, M., Flicek, P., Montgomery, S., Tavare, S., Deloukas, P., and Dermitzakis, E. (2007). Population Genomics of Human Gene Expression. *Nature Genetics*, 39, 1217-1224.

[125] Streib, F.E. and Dehmer, M. (2008). *Analysis of Microarray Data: A Network-Based Approach*. Chichester: Wiley-Vch-Verlag.

[126] Stone, C., Hansen, M., Kooperberg, C., and Troung, Y.(1997). Polynomial Splines and Their Tensor Products in Extended Linear Modeling. *Annals of Statistics*, 25, 1371-1470.

[127] Szasz, O. (1950). Generalization of S. Bernsteins Polynomials to the Infinite Interval. *Journal of Research of the National Bureau of Standards*, 45, 239–245.

[128] 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, 675-698.

[129] Tenbusch, A. (1997). Nonparametric Curve Estimation with Bernstein Estimates. *Metrika*, 45, 1-30.

[130] Tibshirani, R. (1996). Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society Series B (Methodological)*, 267-288.

[131] Tiku, M.L. (1967). Estimating the Mean and Standard Deviation from a Censored Normal Sample. *Biometrika*, 54, 155-165.

[132] Tiku, M.L. (1989). *Modified Maximum Likelihood Estimation*. New York: Encyclopedia of Statistical Science, Wiley.

[133] Tiku, M.L. and Suresh, R.P. (1992). A New Method of Estimation for Location and Scale Parameters. *Journal of Statistical Inference and Planning*, 30, 281-292.

[134] Totik, V. (1984). Uniform Approximation by Bernstein-Type Operators. *Nederlandse Akademie van Wetenschappen (Indagationes Mathematicae)*, 50, 87–93.

[135] Tsai, J.C.C. and Chen, V.C.P. (2005). Flexible and Robust Implementations of Multivariate Adaptive Regression Splines within a Wastewater Treatment Stochastic Dynamic Program. *Quality and Reliability Engineering International*, 21, 689–699.

[136] Vitale, R. (1975). A Bernstein Polynomial Approach to Density Function Estimation. *Statistical Inference and Related Topics*, 2, 87–89.

[137] Voronovskaja, E. (1932). Détermination de la Forme Asyptotique de L'Approximation des Fonctions par les Polynomes de M. Bernstein. *Comptes Rendus de l'Académie des Sciences*, 86-92.

[138] Voorman, W.L., Shojaie, A., and Witten, D.M. (2014). Graph Estimation with Joint Additive Models. *Biometrika*,101, 85–101.

[139] Walczak, Z. (2004). On the Convergence of the Modified Szasz-Mirakyan Operators. *Yokohama Math*, 51, 10-18.

[140] Weber, G.W., Tezel, A., Taylan, P., Soyler, A., and Çetin, M. (2008). Mathematical Contributions to Dynamics and Optimization of gene-environment networks. *Optimization*, 57(2), 353–377.

[141] Weber, G.W., Uğur, Ö., Taylan, P., and Tezel, A. (2009). On Optimization, Dynamics and Uncertainty: A Tutorial for Gene-Environment Networks. *Discrete Applied Mathematics*, 157, 2494–2513.

[142] Weber, G.W., Defterli, Ö., Kropat, E., and Alparslan-Gök, S.Z. (2011). Modeling, Inference and Optimization of Regulatory Networks Based on Time Series Data. *European Journal of Operational Research*, 211, 1-14.

[143] 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, 443-457.

[144] Weber, G.W., Batmaz, I., Koksal, G., Taylan, P., and Yerlikaya, F. (2012). CMARS: A New Contribution to Nonparametric Regression with Multivariate Adaptive Regression Splines Supported by Continuous Optimization. *Inverse Problems in Science and Engineering*, 203, 371-400.

[145] Weierstrass, K. (1885). Über die Analytische Darstellbarkeit Sogenannter willkürlicher Functionen Einer Reellen Veränderlichen. *Sitzungsberichte der Kniglich Preuischen Akademie der Wissenschaften zu Berlin*, 789-805.

[146] Werhli, A., Grzegorczyk, M., and Husmeier, D. (2006). Comparative Evaluation of Reverse Engineering Gene Regulatory Networks with Relevance Networks, Graphical Gaussian Models and Bayesian Networks. *Bioinformatics*, 22, (20), 2523-2523.

[147] Wermuth, N. (1976). Analogies Between Multiplicative Models in Contingency Tables and Covariance Selection. *Biometrics*, 32, 95-108.

[148] Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. New York: JohnWiley and Sons.

[149] Wilkinson, D.J. (2006). *Stochastic Modelling for Systems Biology*. Boca Raton: Taylor and Francis.

[150] 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).

[151] Van-Kampen, N.G. (1992). *Stochastic Processes in Physics and Chemistry*. Amsterdam: Elsevier Science.

[152] Yılmaz, F.B. (2004). *A mathematical Modeling and Approximation of Gene Expression Patterns by Linear and Quadratic Regulatory Relations and Analysis of Gene Networks*. MS Thesis, Middle East Technical University.

[153] Yuan, M. and Lin, Y. (2007). Model Selection and Estimation in the Gaussian Graphical Model. *Biometrika*, 94, 19-35.

[154] Zeretzke, C. (2006). *A Genome-Wide Analysis of Protein-Protein Interactions in Kaposi's Sarcoma-Associated Herpesvirus (KSHV)*. PhD Thesis, Munich Ludwig Maximilian University.

[155] Zhao, T., Liu, T., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The Huge Package for High-Dimensional Undirected Graph Estimation in R. *Journal of Machine Learning research,* 13, 1059–1062.

[156] Zhao, T., Li, X., Liu, H., Roeder, K., Lafferty, J. and Wasserman, L. (2015). *huge: High-Dimensional Undirected Graph Estimation*. R package version 3.23. URL: https://cran.r-project.org/web/packages/huge/index.html (visited on 28/06/2017).

[157] Zhou, S., Rütimann, P., Xu, M., and Bühlmann, P. (2012). High Dimemsional Covariance Estimation Based on Gaussian Graphical Models. *Journal of Machine Learning Research*, 12, 2975–3026.

[158] Zhou, S. and Shen, X. (2012). Spatially Adaptive Regression Splines and Accurate Knot Selection Schemes. *Journal of American Statistical Association*, 96, 247–259.

[159] Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

APPENDIX A

APPENDICES

This Appendix section includes the graphes of the generated time-course data which are used for the JAK-STAT application in the LMARS method.

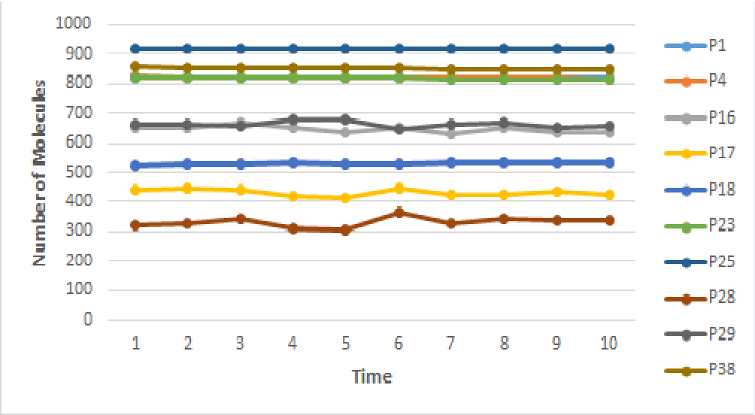


Figure A.1: Changes in the number of molecules for proteins P1, P4, P16, P17, P18, P23, P25, P28, P29 and P38.

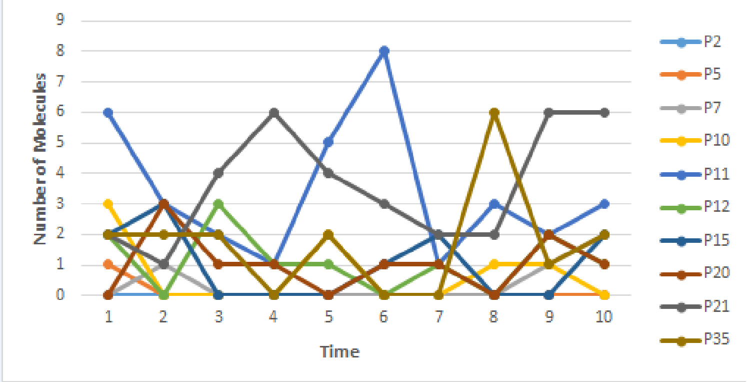


Figure A.2: Changes in the number of molecules for proteins P2, P5, P7, P10, P11, P12, P15, P20, P21 and P35.

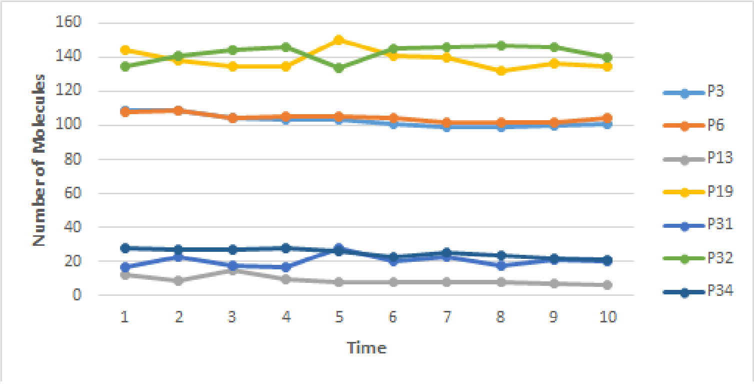


Figure A.3: Changes in the number of molecules for proteins P3, P6, P13, P19, P31, P32 and P34.

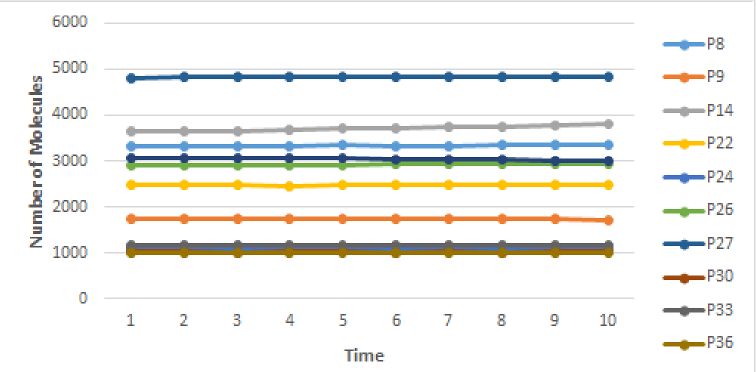


Figure A.4: Changes in the number of molecules for proteins P8, P9, P14, P22, P24, P26, P27, P30, P33 and P36.

Table A.1: Protein list of the generated dataset of JAK/STAT pathway [92].

Protein	**Name of the Proteins**	**Protein**	**Name of the Proteins**
P1	Receptor IFNAR1	P20	IRF9n
P2	TYK	P21	Free TFBS
P3	Receptor Tyk Complex	P22	Occupied TFBS
P4	Receptor IFNAR2	P23	mRNAn
P5	JAK	P24	mRNAc
P6	Receptor Jak Complex	P25	SOCS
P7	IFN_free	P26	Stat2n_IRF9
P8	IFNAR dimer	P27	STAT2n
P9	Active Receptor Complex_Stat2c	P28	CP
P10	STAT2c_IRF9	P29	ISGF-3c_CP
P11	Active Receptor Complex_STAT2c	P30	Stat1c*_Stat2c*_CP
P12	IRF9c	P31	NP
P13	STAT2c	P32	Stat1n*_Stat2n*_NP
P14	STAT1c	P33	ISGF-3n_NP
P15	Active Receptor Complex_STAT2c_STAT1C	P34	Occupied TFBS_NP
P16	STAT1c*_STAT2c*	P35	PIAS
P17	ISGF-3c	P36	PIAS_ISGF-3n
P18	ISGF-3n	P37	STAT1n
P19	STAT1n*_STAT2n*	P38	IFN_influx

CURRICULUM VITAE

PERSONAL INFORMATION

Surnam, Name: Ağraz, Melih
Nationality: Turkish (TC)
Date and Place of Birth: 17.05.1983, Burdur
Marital Status: Marriage
Phone: 0 505 218 88 04

EDUCATION

Degree	Institution	Year of Graduation
M.S.	Dokuz Eylül University, Department of Statistics	2011
B.S.	Dokuz Eylül University, Programme of Mathematics Teaching	2006
B.S.	Gazi University Faculty of Law	2017

PUBLICATIONS

- Ayyıldız, E., **Ağraz, M.**, and Purutçuoğlu, V. (2017). MARS as the Alternative Approach of GGM in Modelling of Biochemical Systems, *Journal of Applied Stat.* URL: http://www.tandfonline.com/doi/full/10.1080/02664763.2016.1266465 (visited on 28/06/2017).

- Purutçuoğlu, V., **Ağraz, M.**, and Wit, E. (2017). Bernstein Approximations in Glasso-Based Estimation of Biological Networks. *The Canadian Journal of Statistics*, 45 (1), 62–76.

- **Ağraz, M.**, M. and Purutçuoğlu, V. (2016).Different Types of Bernstein Operators in Inference of Gaussian Graphical Model. *Cogent Mathematics*, 3:1154706.

CONFERENCE PROCEEDINGS

- **Ağraz M.**, Kılıç B., and Purutçuoğlu, V. (2013). Deterministic of gene networks via parametric and nonparametric approaches, Proceeding of the 29th Meeting of Statisticians, Budapest, Hungary.

- **Ağraz M.** and Purutçuoğlu V. (2016). Deterministic modelling of linear and nonlinear interactions in biological systems , Proceeding of the International Conference on Information Complexity and Statistical Modeling in High Dimensions with Applications (IC-SMHD), Kapadokya, Turkey.

- **Ağraz M.** and Purutçuoğlu V. (2016). A non-parametric model in the construction of biological networks , Proceeding of the 2nd Researchers-Statisticians and Young Statisticians Congress (IRSYSC), Ankara, Turkey.

- **Ağraz M.** and Purutçuoğlu V. (2017). Inference of the Gaussian graphical model via the modified maximum likelihood approach, International Conference on Progress in Applied Science, İstanbul, Turkey.

- **Ağraz M.** and Purutçuoğlu V. (2017). Empirical copula in detection of batch effects, International Workshop on Mathematical Methods in Engineering, Ankara, Turkey.

- **Ağraz M.** and Purutçuoğlu V. (2017). Modeling of biochemical networks via a new Graphical approach, 9th EMR-Italian Region of IBS Conference, Thelesssa, Greece.

BOOK CHAPTER

Ağraz, M. and Purutçuoğlu V. (2016). Transformations of Data in Deterministic Modelling of Biological Networks. In *Intelligent Mathematics II: Applied Mathematics and Approximation Theory*, Eds. Anastassiou, G., Duman, O., 343–356, Springer.

PROJECT

- Scientific and Technological Research Council (TUBITAK). Project title: Stochastic inference of the model parameters for the biochemical systems via the particle filtering method. Project no: TBAG-112T772, April 2013– September 2013.
- Scientific Research Project (BAP). Project title: Estimation of the model parameter of the graphical model based on the Gaussian graphical model and lasso regression using the modified maximum likelihood estimator. Project no: BAP-01-09-2017-002, January 2017- Present.

SUBMITTED AND ONGOING WORKS

- **Ağraz M.** and Purutçuoğlu, V. (2017). Extended Lasso-type MARS model in the description of biological network.
- **Ağraz M.** and Purutçuoğlu, V. (2017). Deterministic modelling of linear and nonlinear interactions in biological systems.
- **Ağraz M.** and Purutçuoğlu, V. (2017). Inference of the Gaussian graphical models via the modified maximum likelihood estimation.
- **Ağraz M.** and Purutçuoğlu, V. (2017). R packaging of modelling the biological networks via LMARS