APPLICATION OF COPULAS IN GRAPHICAL MODELS FOR INFERENCE OF BIOLOGICAL SYSTEMS

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

BY

DAMLA DOKUZOĞLU

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

JUNE 2016

Approval of the thesis:

APPLICATION OF COPULAS IN GRAPHICAL MODELS FOR INFERENCE OF BIOLOGICAL SYSTEMS

submitted by **DAMLA DOKUZOĞLU** in partial fulfillment of the requirements for the degree of **Master of Science 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, **Statistics Dept., METU**

Examining Committee Members:

Assoc. Prof. Dr. Ceylan Talu Yozgatligil _____

Statistics Dept., METU

Assoc. Prof. Dr. Vilda Purutçuoğlu _____

Statistics Dept., METU

Assoc. Prof. Dr. Yeliz Yolcu Okur _____

Graduate School of Applied Mathematics, METU

Assist. Prof. Dr. Ceren Vardar Acar _____
Statistics Dept., METU

Assist. Prof. Dr. Bala Gür Dedeoğlu _____

Institute of Biotechnology, Ankara University

Date: June 29, 2016

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: Damla Dokuzoğlu

Signature :

ABSTRACT

APPLICATION OF COPULAS IN GRAPHICAL MODELS FOR INFERENCE OF BIOLOGICAL SYSTEMS

Dokuzoğlu, Damla

MS, Department of Statistics

Supervisor: Assoc. Prof. Dr. Vilda Purutçuoğlu

June 2016, 81 pages

Naturally, genes interact with each other by forming a complicated network and the relationship between groups of genes can be showed by different functions as gene networks. Recently, there has been a growing concern in uncovering these complex structures from gene expression data by modeling them mathematically. The Gaussian graphical model (GGM) is one of the very popular parametric approaches for modelling the underlying types of biochemical systems. In this study, we evaluate the performance of this probabilistic model via different criteria from the change in dimension of the systems to the change in the distribution of the data. Hereby, we generate high dimensional simulated data via copulas and apply them in GGM to compare sensitivity, specificity, F-measure and various other accuracy measures. We also assess its performance under real datasets. We consider that such comprehensive analyses can be

helpful for assessing the limitation of this common model and for developing alternative approaches to overcome its disadvantages.

Keywords: Gaussian graphical model, Monte Carlo simulations, copula, accuracy measures, gene networks

ÖZ

BİYOLOJİK AĞLARIN TAHMİNİNDEKİ GRAFİKSEL MODELLEMELERDE KOPULALARIN KULLANIMI

Dokuzoğlu, Damla

Yüksek Lisans, İstatistik Bölümü

Tez Yöneticisi: Doç. Dr. Vilda Purutçuoğlu

Haziran 2016, 81 sayfa

Doğal ortamda, genler birbiri ile karmaşık bir ağ oluşturarak etkileşime girmektedir ve gen grupları arasındaki ilişki, gen ağları gibi farklı fonksiyonlar aracılığıyla gösterilebilir. Son zamanlarda, gen anlatım verilerinden matematiksel modelleme yaparak genler arasındaki bağlantıyı ortaya çıkarmak, artan bir ilgiye sahip olmaya başladı. Gaussian Grafiksel Modelleme (GGM), bahsedilen türdeki biyokimyasal sistemlerin modellenmesinde oldukça popüler olan, parametrik yaklaşımlardan biridir. Bu çalışmada, yukarıda bahsedilen olasılık modelinin performansı, sistemlerin boyutlarındaki değişimden, verilerin dağılımlarındaki değişime kadar, çeşitli ölçütler kullanılarak değerlendirilecektir. Bu amaçla, yüksek boyutlu veri, kopula fonksiyonları ve simülasyon yöntemi kullanılarak oluşturulacaktır. Oluşturulan bu veri; duyarlılık, özgüllük, F-ölçü ve diğer çeşitli doğruluk ölçülerini karşılaştırmak için Gaussian Grafiksel Modele uygulanacaktır. Ayrıca gerçek veri kümelerinde de performans değerlendirilecektir. Bu denli kapsamlı analizlerin, bu yaygın modelin sınırlamalarının değerlendirilebilmesinde

ve dezavantajlarını aşabilen alternatif yaklaşımlar geliştirilebilmesinde yardımcı olmasını düşünmekteyiz.

Anahtar Kelimeler: Gaussian grafiksel model, Monte Carlo simülasyonları, doğruluk ölçütleri, kopula, gen ağları

To my family

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my supervisor Assoc. Prof. Dr. Vilda Purutçuoğlu for her constant guidance, support, encouragement and empathy. Her advanced level of academic knowledge leads me to learn a lot and I am sincerely thankful for all her effort.

Next, I express my appreciations to the members of my thesis committee; Assoc. Prof. Dr. Ceylan Talu Yozgatligil, Assoc. Prof. Dr. Yeliz Yolcu Okur, Assist. Prof. Dr. Ceren Vardar Acar, Assist. Prof. Dr. Bala Gür Dedeoğlu.

I would like to thank TUBİTAK Research Grant (Project no: 114E636) for their support.

I am very thankful to Prof. Dr. Barış Sürücü and my coordinator Ülkü Önder. They are always understanding throughout the process.

Also, I would like to thank my friend, Ali Kanatlı, for being there for me and being supportive and understanding during stressful times.

I would like to convey my thanks to my sister Duygu Dokuzoğlu Oğuz and my brother in law Engin Berk Oğuz for their intense and priceless support. They always make me stronger and decided.

Last but absolutely not least, I want to express my gratitudes the reason of my existence, my family. My parents, Senem Dokuzoğlu and Tefik Dokuzoğlu. They always encourage me and are tower of my strength. Words are not sufficient to express my feelings for them.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

AC	Accuracy
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GGM	Gaussian Graphical Model
KS	Kolmogorov Smirnov
MLE	Maximum Likelihood Estimator
mRNA	Messenger RNA
RIC	Rotation Information Criterion
SNPs	Single Nucleotide Polymorphisms
STARS	Stability Approach to Regularization Selection
TN	True Negative
TP	True Positive
TPR	True Positive Rate
UTR	Untranslated Region

CHAPTER 1

INTRODUCTION

The inference of gene networks plays an important role in enlightening the connections among genes that may lead to a better understanding of molecular mechanisms in organisms. The biologists routinely use high-throughput technologies, microarrays to measure expression of genes. Moreover, the statisticians are often in charge of exploring interactions among genes through statistical analysis by using large datasets. Accordingly, it is usual to apply multivariate methodologies to analyze these large datasets. Because multivariate methods may disclose various interactions among genes that cannot be established from individual gene-based approaches.

In this study, I focus on a graphical modeling approach that purposes at finding relationships among a group of genes, where a graph is used for encoding relationships among multiple variables. When a graph is used for a gene network, the nodes represent genes and the edges indicate relationships between the linked genes (Figure 1.1). Here, the edges can be explained with various relationships among genes. For instance, the pairwise correlations are used to define edges in a “relevance network”.

Moreover, we can define edges by means of the conditional independence. It implies that if any two genes connect each other with an edge, indirectly, they can be affected from other genes. Therefore, the appearance profiles of two genes are correlated as long as they are both regulated by some other genes.

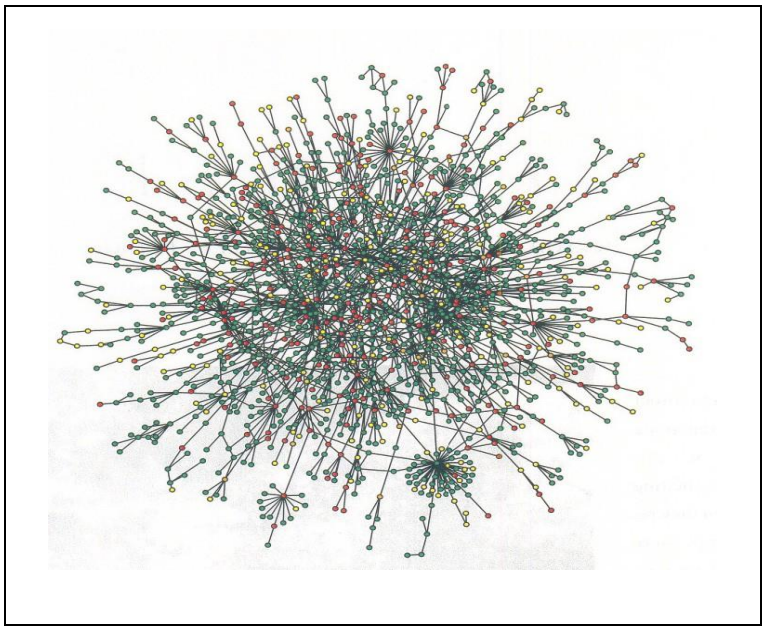


Figure 1.1: An example for a representation of a complex biological network structure [5].

Additionally, it is necessary to assume gene networks together under various conditions, rather than considering them separately. The reason is that the large parts of the gene networks are probable to share common topologies corresponding to similar underlying biological processes [5]. Hereby, the large datasets allow us to infer the relationships among genes and the Gaussian graphical model (GGM) is one of the alternative approaches to get these findings.

Accordingly, in the following chapters, the description of GGM and gene networks is shown. Morover, Gaussian graphical model application and the results of the simulations is represented.

1.1.Aim of the study

The aim of the study is to understand how the Gaussian graphical model deals with the different dimensional systems and data types when inferencing the gene networks.

In order to meet requirements of this aim, below questions are established.

- Under which conditions does GGM work properly?
- How does GGM work when the normality assumption is violated?
- How does GGM behave when the dimension of the networks are getting larger?
- What is the difference between gene structures when modeling with GGM?
- How does GGM work with copula functions?
- Which copula function is the most suitable to model gene networks?
- Which marginals should be used to create multivariate biological data with the copula function?
- How does GGM model give the result under different copula marginals?

1.2. Motivation

Recently, there exists a huge interest in the gene interactions. Most of the scientist try to understand their behaviors and relations with each others. Most of the biological activities and diseases are related with gene networks. In this scope, there exist some useful statistical tools such as graphical models to capture the biological networks. For this reason, I focus on a graphical modeling approach that purposes at finding relationships among a group of genes, where a graph is used for encoding relationships among multiple variables. Here, the Gaussian graphical model (GGM) is quite common and it is used to discover and estimate the biological networks. In this study, my motivation is to discover capabilities and deficiencies of GGM.

1.3. Thesis Overview

This thesis starts with the description of the Gaussian graphical model and includes the theoretical background of the study. After the GGM section, it continues with the description of copulas, the measures of accuracy and the application. Then, the results of related simulations are presented for different scenarios. Finally, I review and discuss the key results derived from the simulations as well as the real dataset.

Hereby, *Chapter 2* shows the methods of the study. In this chapter, firstly, I explain the Gaussian graphical model and gene networks. Then, I discuss the estimation methods of the covariance matrix which are listed as the maximum likelihood approach, shrinkage covariance matrix and the lasso-based graphical model. After that, I explain the theoretical background of copulas. I review the dependence structure of the copula function and give detailed information about the most-known copula functions which are *product copula, Gaussian (Normal) copula, student-t copula, Gumbel copula, Farlie-Gumbel-Morgenstern copula, Clayton copula* and *Frank copula*. Moreover, I turn to specific application areas of copula functions. Lastly, I explain the related measures of accuracy and show their calculations. Moreover, the general confusion matrix is added to define *true positive, false negative, false positive* and *true negative values*.

Chapter 3 describes the application part of the study. This chapter consists of two main sections that are *multivariate normal data and multivariate data via the Gaussian copula*. In the multivariate normal data section, I simulate the normal data under different dimensions and network structures which are *scale-free, random, cluster* and *hubs*. Then, I model data with the Gaussian graphical model and show the results of the simulations for each dimension. Following that, in the multivariate data via the Gaussian copula section, the high dimensional data are created by the help of the Gaussian copula with different margins. Here, the student-t, log-normal, normal and the exponential distributions are used as margins of the Gaussian copula. I put the last touches on this

thesis by considering the semi-exponential and semi-normal marginal application with the Gaussian copula. Modeling with GGM and results of the simulations are shown in related tables.

Chapter 4 presents the findings of the real data applications. Here, I use three different types of biologic real data and evaluate the performance of GGM.

Finally, *in Chapter 5,* I conclude the findings of all simulations and real data and then, discuss possible further research topics.

CHAPTER 2

METHODS

2.1. Gaussian Graphical Models

The Gaussian graphical model (GGM) is simply dependent on the estimated partial correlation matrix, whose interpretation is straightforward under the normality assumption of the data [46]. Here, the zero entry implies no relation between the associated pair of genes due to the feature of independence under the multivariate/univariate normal distribution.

On the other hand, in spite of this underlying advantage, it also means that the linear interaction between genes can be descripted merely under the normality assumption. Unlike the Bayesian networks which are directed graphical model, the Gaussian graphical model produces undirected networks [27]. Statistically, the undirected edges imply the conditional independence. Accordingly, in GGM, if there is no edge between two nodes, it can be accepted as the evidence of the conditionally independent nodes [46].

For instance, there is no edge between the node 1 and the node 3 in Figure 2.1 even though they are related to each other via the node 2. Therefore, we can conclude that the node 1 and the node 3 are conditionally independent.

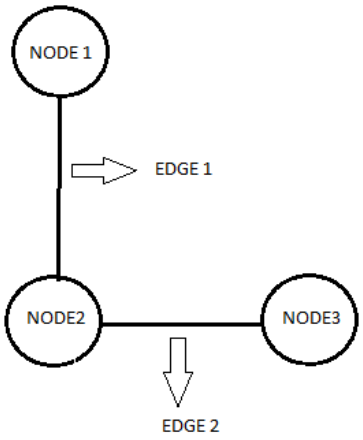


Figure 2.1: The simple the representation of the conditional independence between the node 1 and the node 3 for the given node 2.

Under the GGM assumption, the graph structure can be estimated using the sparsity pattern of the inverse covariance matrix. This independence structure is implicitly related with the variance-covariance matrix Σ [27]. But the basic concept of the independence structure is directly related to the inverse of the variance-covariance matrix, Θ which is the precision, also called the concentration matrix; $\Theta = \Sigma^{-1}$.

The covariance matrix Σ has a distinct role in GGM, such that the zero entry in the covariance matrix implies the marginal independence while the zero entry in the precision matrix Θ means the conditional independence between related nodes, i.e., variables [46]. For this reason, the precision matrix Θ can be written in terms of the partial covariance. In the GGM approach, the derivation of the partial correlation from the precision matrix is quite common and there exists several approaches to find those partial correlations from data.

In high dimensional datasets, as typically seen in gene expression networks, the estimation is always problematic. Below, we list the most common methods for the

covariance matrix selection methods among alternatives, which are designed specifically to unravel this challenge and explain most well-known ones in details in the following subsections. In this study, we use the glasso method to estimate the precision matrix of the system.

Accordingly, the coordinate-wise descent algorithm within the l_1-penalized lasso [16], fused lasso [40], grouped lasso [48], adaptive lasso [50] and elastic net method [51] are well-known estimation techniques of GGM. These methods are highly related with the construction of the penalized maximum likelihood function under both l_1-norm and l_2-norm terms. On the other hand, there are some recent approaches to estimate precision matrix without any use of optimization methods. These methods are fully parametric approaches that are based on the Bayesian techniques such as the birth-death Markov chain Monte Carlo approach [44] and some specific semi-parametric approaches such as the non-paranormal SKEPTIC algorithm [24]. There are also other Monte Carlo approaches as Atay –Kayis and Massam suggest [1]. They apply Monte Carlo methods for the calculation of Θ when there is non-decomposable network graph [1]. Furthermore, Dobra et al. [14] and Wang and Li [42] propose the reversible jump Markov Chain Monte Carlo (MCMC) method by combining the graphical model with Gaussian copula and Dauwels et al. [13] perform the Mont Carlo expectation maximization method in place of the reversible MCMC for the same type of models.

Below, we explain the most common covariance matrix selection methods which are maximum likelihood approach, shrinkage covariance matrix and lasso-based graphical model.

2.1.1. Maximum Likelihood Approach

Let $Y_{(1)}$, …., $Y_{(p)}$ be a random sample of the *p*-dimensional vector *Y* whose likelihood function under the multivariate normal distribution is shown as in Equation (1).

$$L(\mu,\Sigma) = (2\pi)^{-np/2} \prod_{i=1}^{n} det(\Sigma)^{-1/2} exp\{-\frac{1}{2}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu)\}, \quad (1)$$

where μ and Σ denote the mean vector and the variance-covariance matrix, respectively. Furthermore, *n* is the sample size and *p* represents the total number of genes in the system. Finally, *det (.)* stands for the determinant of the given matrix. Thereby, by converting Equation (1) into an expression based on Θ, the log-likelihood of Equation (1) can be written as

$$\mathrm{L}(\Theta) = \frac{n}{2}\log|\Theta| - \frac{n}{2}Trace(S\Theta), \quad (2)$$

in which *S* is the sample covariance matrix and Θ refers to the precision matrix. However, the solution via the maximum likelihood method has some challenges such that it can estimate fully connected Θ. Moreover, there is a possibility to not infer precision matrix Θ if the covariance matrix is non-invertible. These problems generally occur when the number of nodes *p* is higher than the number of observations *n*, i.e., *p*>>*n*. Under this condition, the MLE procedure may produce multiple solutions. Moreover, when the number of nodes is much higher than the number of observations, the computational demand of MLE increases as its estimators are found via further iterative steps [26].

2.1.2. Shrinkage Covariance Matrix

If the sample size *n* is small and the number of variables *p* is large, the emprical estimators of the covariance may not be unique [33]. In order to infer the network under this challenge, the shrinkage estimator can be useful. Basically, the shrinkage estimator

is about taking a weighted average of the sample covariance matrix and a target matrix of the same dimensions as in Equation (3).

$$S' = \lambda T + (1 - \lambda)U, \tag{3}$$

where *T* is the low dimensional target, λ denotes the shrinkage intensity lying in the range [0, 1] and *U* represents the unbiased sample covariance [33]. The unbiased estimator *U* exhibits a large variance because of the number of parameters. Whereas, the low dimensional *T* shows a lower variance with a considerable bias. In order to solve this problem, the shrinkage approach combines both estimators by using a weighted average. In Equation (2), λ shows the level of the shrinkage. For $\lambda = 1$, the shrinkage estimate equals to the shrinkage target *T* and when $\lambda = 0$, it equals to the unbiased sample covariance *U*. Hence, the main challenge in this estimator is the selection of the optimal value for the shrinkage intensity. According to Schafer [33], the most appropriate way to unravel this problem is to minimize the mean squared error as in Equation (4).

$$\mathrm{R}(\lambda) = \mathrm{E}(\mathrm{L}(\lambda)) = \mathrm{E}[\sum_i (\lambda T + (1 - \lambda)U - \sigma)^2]. \tag{4}$$

In this equation, E (.) implies the expectation of the given random variable. Here, the value of λ is correlated with the variance of the sample covariance. That is, when *U* becomes smaller, λ also gets smaller. Therefore, when the sample size increases, the dimension of the target *T* reduces.

2.1.3. Lasso-based Graphical Model

This method is originated from the Lasso regression with the ℓ_1-norm. Normally, in the multivariate regression method, there exists a response variable (dependent variable) $y \in R^n$ and a predictor variable (independent variable) matrix $X \in R^{nxp}$. Assuming that $X_1, \ldots, X_P$ are linearly independent, the linear regression solves the least squares problem and finds the unique solution for $\hat{\beta}$. But if the dimensions of data are much higher than the size of data, i.e., $p>>n$, this method becomes problematic [22]. In order to handle this challenge, the lasso regression can be a solution. Basically, the lasso regularization assumes that the observations have a multivariate Gaussian distribution with the mean vector μ and the covariance matrix Σ such that

$$\sum_{i=1}^{N}(y_i - \sum_j \beta_j X_{ij})^2 + \lambda \sum_j \beta_j. \quad (5)$$

Equation (5), the lasso regression is subjected to the ℓ_1-norm which is the least absolute deviations (LAD) or least absolute errors (LAE) [26]. Statistically, the ℓ_1-norm minimizes the sum of the absolute differences between dependent variable (y_i) and the estimated values of independent value (X_i), where, the ℓ_2-norm basically minimizes the sum of squares of differences between the dependent variable (y_i) and the estimated values of the predictor variable (X_i).

Moreover, the ℓ_1-penalty is important for the estimation of Σ^{-1} to increase its sparsity [39]. By sparsity, we mean the property that all parameters, which are zero, are actually estimated as zero with a probability tending to one. The reason behind this conversion is to avoid over-fitting due to either high-dimensional data or collinearity of the variables [22].

In addition, the estimation of a sparse graphical model is highly important in the high dimensional setting.

In recent years, the researchers have proposed several estimation methods for the sparse undirected graphical models by using the ℓ_1-lasso regularization.

Meinshausen and Bühlmann (2006) [26] apply the lasso regression method for the selection of the covariance matrix. In this method, each node is linearly regressed on the rest of the nodes with an ℓ_1-penalty. Basically, it uses the neighborhood selection method to find the relation between nodes. Later, they combine the neighborhoods to estimate the full graphical structure. In more detail, under the Gaussian neighborhood regression method, they study the lasso regression models as the Gaussian graphical model by proposing the neighborhood selection method, which tries to discover the smallest index. In this approach, the node $Y^{(p)}$, i.e., the last node, depends on the rest of the nodes $Y^{(-p)}$ [26] as shown by the following expression.

$$Y^{(\mathrm{p})} = \beta\, Y^{(\mathrm{-p})} + \varepsilon, \tag{6}$$

where β denotes the p-dimensional regression coefficients. Here, it is assumed that ε has p-dimensions and is a multivariate normally distributed random error with the mean vector zero and the covariance matrix $\sum_{pxp}$. The estimator of β can be found by the least squares criterion in such a way that the lasso regression model provides the sparsity by applying the ℓ_1-penalty on these regression coefficients. Thus, it is shown that the neighborhood selection aims to estimate the individual neighborhood of any given node. Moreover, it is found that the neighborhood selection method is much better than the MLE method in terms on computation time and its consistency for high-dimensional data.

In 2008, by adding to above methods, Friedman [17] applied the coordinate descent algorithm for the lasso regression which is named as the *graphical lasso, "glasso"*. This

approach provides a computationally efficient method for performing the lasso-regularized estimation of the sparse covariance matrix.

In Equation (7), we present the objective functions that is maximized with respect to Θ under the ℓ_1-penalized log-likelihood function.

$$\max_{\Theta} \quad \log\det(\Theta)-\mathrm{tr}(S\,\Theta)-\lambda\|\,\Theta\,\|_1\,, \tag{7}$$

where *n* is the number observation and *p* denotes the number of nodes. Hence, we have the multivariate normal distribution with the mean μ and the covariance Σ. $\Theta = \Sigma^{-1}$ indicates the precision matrix, i.e., inverse of the variance-covariance matrix, as used beforehand and *S* shows the empirical covariance matrix. Furthermore, λ is the non-negative Lagrange multiplier. When λ is getting larger, the biological network becomes sparser [46]. Finally, tr (.) describes the trace and $\|\Theta\|_1$ represents the ℓ_1-norm of the precision matrix.

In equation (7), the optimal selection of λ can be succeed by using different methods such as the Banerjee method [4], the block diagonal update of the matrix [47] and k-cross validation method [17]. Furthermore, we can also use the rotation information criterion (RIC), the stability approach to regularization selection (STARS) and the extended Bayesian Information Criterion (EBIC). Whereas, in the application we select the optimal penalty constant by RIC among these alternatives since it is the most common measure of GGM if the inferences are conducted by the glasso method [49].

2.2. Copula

A copula method for understanding multivariate distributions has a relatively short history in the statistical literature. Most of the statistical applications has arisen in the last twenty years. However, by the Sklar theorem 1959 [37], it is showed that a multivariate distribution can be represented in terms of its underlying margins by gathering them together via a copula function. But the reason behind the popularity of copulas is that it can be applied to many fields from finance to insurance.

In general, the copulas [28] provide the theoretical framework in which the multivariate associations can be modeled separately from the univariate distributions of the observed variables. Therefore, a copula is a function which connects the univariate marginals of variables to their multivariate distributions. Hereby, the copulas are useful approach for generating the joint distribution with a variety of dependence structures between variables by eliminating the influence of marginals.

Typically, the pairwise dependence between two variables is explained with families of bivariate distributions such as bivariate normal, log normal and gamma. But this approach has a limitation. Because, the individual behavior of variables must be explained with the same family of the univariate distribution. But, the copula models relax these assumptions and apply different marginal distributions of the random variables by the following way.

$$H(x,y)= C\{F(x), G(y)\}. \tag{8}$$

Here, if x and y are two continuous random variables, the copula function C is unique. Otherwise, it is not unique. The reason is that the uniqueness is based on the multiplication of the range of x and y. Accordingly, $H(x,y)$ is the joint distribution function of x and y. Moreover, $F(x) = P(X \leq x)$ refers to the marginal cumulative distribution of x,

$G(y)=P(Y \leq y)$ denotes the marginal cumulative distribution of y and C presents the unique copula, which characterizes the joint dependency of x and y. That is, $H(x, y)$ is the multivariate distribution with margins $F(x)$ and $G(y)$.

As a result, the properties of the copula can be listed as below [41].

1. N dimensional copula function C^n has the domain as the n-dimensional identity matrix.

2. C^n consists of margins; $C^n=C(1,....1, u,1,....,1)= u$; for every $u \in$ I=[0,1].

3. $C^n=C(u_1,\ldots\ldots, u_n)=0$ if any $u_m=0$; $m \leq n$.

4. C is the n-increasing. That is every n-copula is non-decreasing in each argument separately.

In this list, the fourth property comes from the properties of the cumulative distribution. Last but not least is the invariance property. Hereby, the copula function is invariant with respect to the increasing and continuous transformation of marginal distributions. In this way, we can implement the same copula function for different transformations of random variables. For example, the copula function C which creates a joint distribution of x and y, can be used for the logarithmic transformation of x and y (lnX, lnY) [28].

In order to see the relationship between cumulative distributions and the copula, the following example can be given: Let $F(x,y)$ be the bivariate distribution function with univariate distributions $F_x(X)$ and $F_y(Y)$. The inverse functions of these cumulative distributions can be denoted as $x=F_x(u_1)^{-1}$ and $y=F_y(u_2)^{-1}$, respectively.

So, let u_1 and u_2 be distributed as uniform due to the fact that the cumulative distribution functions are continuous. Hence,

$$F(x,y)=[\ x=F_x(u_1)^{-1},\ y=F_y(u_2)^{-1}\ ]$$

$$=Prob\ (U_i \leq u_i;\ where\ i=1,\ 2)$$

$$=C\ (u_1,\ u_2). \tag{9}$$

As seen in Equation (9), the main advantage of the copula function is that the joint distribution or the dependence structure between *X* and *Y* can be determined without considering the marginal distributions of *X* and *Y*. As a result, the dependency structure of the copula function can be defined as

$$F(x,y)=C(F_x(X),\ F_y(Y);\ \theta), \tag{10}$$

where θ is the dependency parameter of the copula. This method is called as the inverse method and is based on the Sklar's Theorem [37]. More mathematical details about this dependency parameter will be given in the following part.

In literature, to regarding the dependency and the data structure, there exists many kinds of copulas. So it appears that the copulas form the dependence structure of the model. Thus, the choice of the copula that is going to fit the data is very important.

2.2.1. Dependency Structure

"The dependence function" and "uniform representation" are the two alternative names of the copula [28]. Since the copula function is used to define the dependence structure between random variables, this dependence structure is invariant to increasing and

continuous transformations of marginal distributions thanking to the invariant property of copula function.

Generally, the dependent structure between two variables (X, Y) is explained by the Pearson correlation coefficient. The correlation coefficient of two variables has some useful properties such that it implies the linear dependency, symmetry and the invariance property with respect to the linear transformation and finally, it has a range between -1 and +1.

Here, the zero correlation means that two variables are independent under normality. However, the correlation coefficient has some deficiencies too. For example, X and Y are two random variables and $Y=X^2$ as well as cov [X, Y] =0. It means that the correlation of X and Y is zero and they are independent. But, it is obvious that X and Y are dependent [17]. Thus, the correlation coefficient can only handle the linear dependency structure. Moreover, it is not good at some heavy-tailed distributions and it is not invariant to non-linear transformations. Because of these deficiencies, there exists alternative measures of dependency structures, namely, the rank correlations and the tail dependence.

So, the copulas dependence structure can be represented by the Kendall's Tau and Speaman's rho under the rank correlations and the tail dependence coefficients, but not the Pearson's linear correlation coefficient [41]. To calculate the Pearson's linear correlation, the effects of the marginal distributions must be taken into consideration. This is totally contrary to the logic of the copula function.

Below, we present the mathematical details of correlation measures and their plotting approach.

a) Spearman Rank Correlation:

The Spearman rank correlation is a non-parametric approach, explaining the functional dependency between two variables (*X, Y*) within the range of -1 and +1 [17]. This value does not have any assumption about the distribution of the data and is related with the concordance and discordance terms. The former means ordered variables in a same way, while the latter implies the differently ordered variables. Furthermore, it is invariant to firmly increasing transformations and is symmetric, co-monotonic and counter-monotonic [41].

Hereby, the expression of the Spearman rank correlation for the random variable *X* and *Y* with respect to the copulas *C(.,.)* is presented as below.

$$P_S(X, Y) = 12\int\int_0^1 \{C(u_1,u_2) - u_1u_2\} d_{u1} d_{u2} \ . \quad (11)$$

b) Kendall's Tau Correlation:

The Kendall's Tau is the useful rank correlation for the copula application. It is a non-parametric approach that measures the dependence between two variables within the range of -1 and +1 [17]. It is also subject to the concordance. Like Spearman's rho, Kendall's tau is invariant to firmly increasing transformations. It has some useful properties which are symmetry, normalization, co-monotonic and counter- monotonic. Finally, the expression of the Kendall's tau for two random variable (*X, Y*) can be defined as the probability of the concordance minus the probability of the discordance [41] and its copula formulation can be written as

$$P_\tau(X, Y) = 4\int\int_0^1 \{C(u_1,u_2)\ d\ \mathrm{C}\ (u_1, u_2) - 1. \quad (12)$$

c) Graphical Tools (Chi-Plots, K-Plots)

In order to visualize the dependent structure of the copula function, the 2-dimensional scatter plot can be helpful. In general the scatter plot is a useful tool to show the tail dependence of the copula function [17]. In addition to the scatter plot, Chi-plots and K-plots are also well-known visual tools to show the dependency. The Chi-plots are created by depending on the dependent structure of the chi-square distribution and it shows the distance between pairs (X, Y). On the other side, K-plots visualize the dependency by using the expected value of the i^{th} variable pairs. Hereby, in these plots, there exists two lines; one of them shows perfectly independent pairs. And other describes perfectly positive dependent pairs. Both graphs are based on the rank of the variables [17].

2.2.2. Copula Types

There are seven major types of copulas. Each of them presents distinct ranges for the random variables and denotes different levels of correlations. These functions can be listed as below and shortly presented in Table 2.1.

a. Product Copula
b. Gaussian (Normal) Copula
c. Student-t Copula
d. Gumbel Copula
e. Farlie-Gumbel-Morgenstern Copula
f. Clayton Copula
g. Frank Copula

a. Product Copula

This copula [6] corresponds to the independence between random variables such that

$$C(x, y) = xy, \tag{13}$$

for the random variable x and y.

b. Gaussian (Normal) Copula

The Gaussian copula [6] can be shown by

$$C(x, y; \theta) = \varphi_G\,(\varphi^{-1}(x), \varphi^{-1}(y); \theta), \tag{14}$$

where φ is the standard normal distribution and $\varphi_G(x,y)$ denotes the standard bivariate normal distribution by explaining the correlation parameter $\theta \in (-1,1)$. The normal copula is flexible in negative and positive dependence and shows a symmetric dependence in both tails. Moreover, it is suitable for every dimensional dataset as the copula parameter has a direct relation with the Pearson correlation coefficients. Therefore, it is the most common copula type.

c. Student's-t Copula

The Student's t-copula [6] which is created by the bivariate student-t distribution can be shown by

$$C(x, y; \theta) = T_v\,(T_v^{-1}(x), T_v^{-1}(y); \theta), \tag{15}$$

where Tv is the Student -t distribution with a degree of freedom ν and the shape matrix Σ. The Student–t copula belongs to the elliptical copula family with the correlation matrix Σ. This copula shows heavy tails and the symmetric dependence. For this reason, it is useful for modelling high correlations extreme values which are observed in the tails of the distribution.

d. Gumbel Copula

The Gumbel copula [6] can be represented as the expression below;

$$C(x, y; \theta) = exp\ \{-[(-\log x)^{\theta} + (-\log y)^{\theta}]^{1/\theta}\}. \quad (16)$$

In this expression, the copula parameter θ may take all values in the interval [1, ∞), where it does not allow negative dependence. This copula family is useful for capturing the upper tail dependence. Because it shows asymmetric dependence. In other words, if the results are strongly correlated with high values instead of low values, the Gumbel copula can be a good estimate [41]. The Gumbel copula also covers the Archimedean class and an extreme-value copula. The Archimedean copulas are very famous because of their capability of the large dependency range. Moreover, their mathematical applications and estimations are relatively simple.

e. Farlie-Gumbel-Morgenstern Copula (FGM)

It is remarkable when the dependence between two marginals is modest in terms of their magnitudes. Like the Gaussian copula, the Farlie-Gumbel-Morgenstern copula [41] indicates a symmetric dependence but, it cannot handle a large dependency structure. This copula can be represented by

$$C(x, y; \theta) = xy\ (1+\theta\ (1-x)\ (1-y)), \quad (17)$$

for the random variable *x* and *y*. Here, the copula parameter θ may take all values in the interval [-1, 1].

f. Clayton Copula

The Clayton copula [6] is well-known as it has a straightforward functional application and belongs to the Archimedean copula family. The Clayton copula is defined as

$$C(x, y; \theta) = (x^{-\theta}+y^{-\theta}-1)^{-1/\theta}, \quad (18)$$

when $\theta \in (0,\infty)$. Here, while θ converges to zero, the marginals become independent. Furthermore, it is not preferable for the negative dependence. This copula family is useful for capturing the lower tail dependence. Because it shows an asymmetric dependence.

g. Frank Copula

This copula [41] permits the negative and symmetric dependence for both tails. Moreover, it is also related with the weak tail dependence. The Frank copula can be expressed as the following way.

$$C(x, y; \theta) = -\theta^{-1} \log \{1+ \frac{(\boldsymbol{exp}\{-\boldsymbol{\theta x}\}-\mathbf{1})\ (\boldsymbol{exp}\{-\boldsymbol{\theta y}\}-\mathbf{1})}{(\boldsymbol{exp}\{-\boldsymbol{\theta}\}-\mathbf{1})}\}, \quad (19)$$

where $\theta \in (-\infty, \infty)$. Due to its range, it is good for the description of strong negative and positive correlational structures. Similar to the Gumbel and Clayton copula, the Frank copula belongs to the Archimedean copula family and it creates a wide range of dependence.

Table 2.1: Copula Types, their ranges and functional forms for the two random variable *x* and *y*.

Copula Type	***Range of Copula θ***	***Functional Form***
Gaussian Copula	$\theta \in (-1,1)$	$\varphi_G\,(\varphi^{-1}(x), \varphi^{-1}(y); \theta)$
Student's t-Copula	$\theta \in (-1,1)$	$T_v\,(T_v^{-1}(x), T_v^{-1}(y); \theta)$
Gumbel Copula	$\theta \in [1, \infty)$	$exp\,\{-[(-\log x)^{\theta} + (-\log y)^{\theta}]^{1/\theta}\}$
FGM Copula	$\theta \in [-1,1]$	$xy\,(1+\theta\,(1-x)\,(1-y))$
Clayton Copula	$\theta \in (0,\infty)$	$(x^{-\theta}+y^{-\theta}-1)^{-1/\theta}$
Frank Copula	$\theta \in (-\infty, \infty)$	$-\theta^{-1}\log\{1+\frac{(exp\{-\theta x\}-1)\,(exp\{-\theta y\}-1)}{(exp\{-\theta\}-1)}\}$

2.2.3. Fields Where Copula Can Be Used

The most famous application field for the copulas is the finance. Maximizing the asset return and allocating bank capitals are very tough topics in finance due to the high risk factor. Generally, maximizing expected return is solved with the Gaussian assumption while the capital allocation is the topic of the joint distribution, namely, both related to the logic of copulas. The copula functions are useful in managing risk and also they are

very good at the value-at-risk computation. Moreover, the asset return data have nonlinear dependency structure (asymmetric dependence). This type of structures can be explained by the Archimedean copula family. So, the Student's t-copula and the Gaussian copula are commonly used copula functions in finance [8].

The second most famous application field for the copulas is the economics. For example, the risk management and modelling the non-linear dependence between economic and financial multivariate time series models are created by gathering univariate time series models via the copula function. To explain the correlation structure of the multivariate distribution in economical application, the Pearson's linear correlation is used. But, the Pearson correlation matrix is fully applicable when the economic variables are coming from the multivariate normal distribution. As long as the economic variables are not normal, the copula functions are required to model the correlation structure. Under this condition, as mentioned before, Spearman's Rho and Kendall's Tau approaches could be the best alternatives to the linear correlation coefficient in the copula application, instead of the Pearson's linear correlation. [30]

Accordingly, in this study, we evaluate the performance of the GGM method under different multivariate distributions. In the implementation, we use the Gaussian copula since it is applicable for any dimensional networks. Because for the remaining types, apart from the student's t-copula, they have explicit functional forms for only small/ toy systems. On the other hand, the student's t-copula is more flexible, whereas, it is computationally problematic and cannot produce very different results than the outputs of the Gaussian copula [43].

2.3. Measure of Accuracy

Generally, in the comparative analyses, different measures of accuracy can be used to evaluate the performance of different methods or control the findings. In this research, in order to assess the performance of the Monte Carlo runs, we apply the well-known accuracy measures which are the precision, recall, F_1-score, false positive rate, true positive rate and the accuracy. All these values are the functions of the scores in Table 2.2. This table, also called the confusion matrix [21], gives information about the actual and the predicted classification. The meanings of its entries are presented below.

Table 2.2: The general confusion matrix.

		PREDICTION CLASS	
		Positive (1)	Negative(0)
ACTUAL CLASS	Positive (1)	True Positive	False Negative
	Negative(0)	False Positive	True Negative

True Positive (TP): The number of correct prediction of actually positive entry.

False Positive (FP): The number of incorrect prediction of actually negative entry.

False Negative (FN): The number of incorrect prediction of actually positive entry.

True Negative (TN): The number of correct prediction of actually negative entry.

Hereby, the description of the selected accuracy measures can be defined as follows:

Precision: It is also known as the positive predictive value (PPV). The higher the precision value, the better classification. A low precision shows that there exists a large number of false positives. In the calculation, the number of correctly estimated edges is divided by the total number of edges in the estimated graph. It is calculated by using the below equation:

$$\text{Precision}=\frac{TP}{TP+FP}\ . \tag{20}$$

Recall: It is also named as the true positive rate. A low recall implies that there exist many false negatives. Finally, in the computation, the number of correctly estimated edges is divided by the total number of edges in the true graph, as below.

$$\text{Recall}=\frac{TP}{TP+FN}\ . \tag{21}$$

F_1-score: It is a weighted average of the precision and recall scores. F1-score reaches its best value at 1 and worst score at 0. Hence, it is calculated by performing the below equation:

$$F_1\text{-Score} = 2 \times \frac{PrecisionXRecall}{Precision+Recall}\ . \tag{22}$$

Accuracy: The accuracy (AC) is the proportion of the total number of correct predictions to the total number of all classified object and is determined by the following equation.

$$\text{AC}=\frac{TP+TN}{TP+TN+FP+FN}. \tag{23}$$

False Positive Rate: This values (FPR) is the proportion of negatives which are incorrectly classified as positives among the total number of negatives in the true graph and is calculated by using the expression below.

$$\text{FPR}=\frac{FP}{TN+FP}. \quad (24)$$

In Chapter 2, we have explained the methods in the application. In that scope, GGM, Copula functions and measure of accuracies have been represented in details. Hence, in *Chapter 3,* we show the application of these methods and discuss the results.

CHAPTER 3

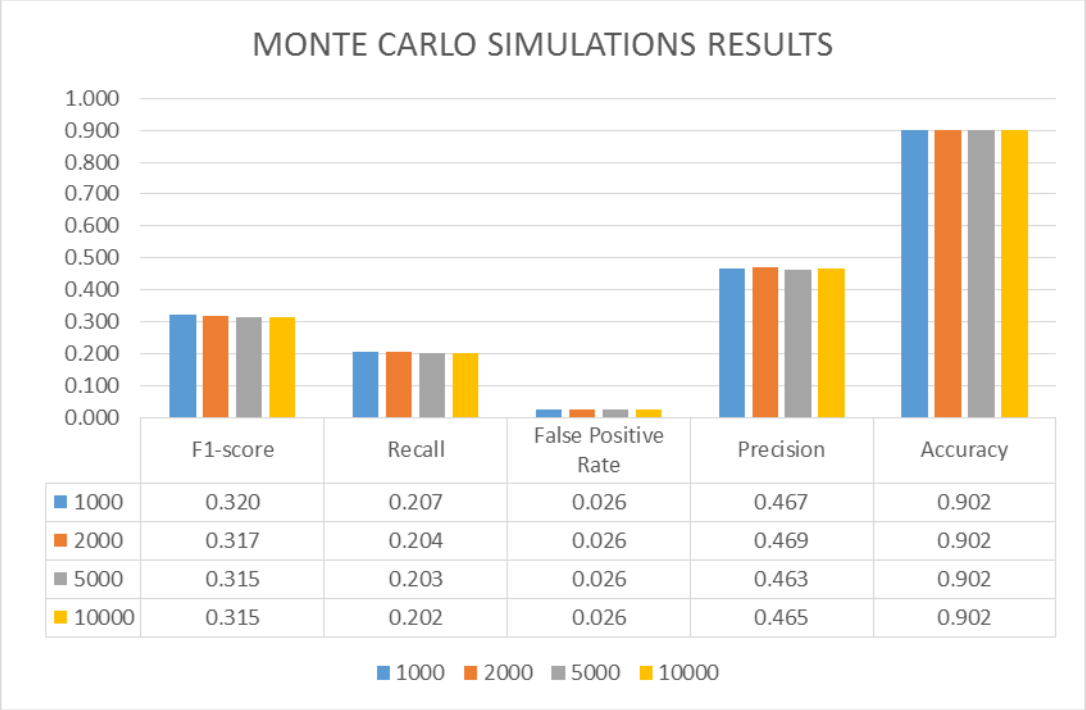
APPLICATION UNDER SIMULATED DATA

In this chapter, more detailed information is given about the application of the Gaussian graphical model under different multivariate data types. In the light of the assessment for the limitation of this common model, different runs are completed under various scenarios by performing the Monte Carlo simulation. In each analysis, we use 1000 iterations.

As we mentioned in Chapter 1, the precision matrix is our main interest to understand the network structure of the biological systems. Thus, the main purpose of this study is to estimate the precision matrix by using GGM. To evaluate the adequacy of GGM, we compare the exact graph path, i.e., the population graph path, with the estimated graph path, i.e., sample graph path, under different dimensions, graph structures and copula functions.

In order to decide on the suitable number of trials for the Monte Carlo simulation, we complete the first simulation under different trials. In that scope, we generate the multivariate normal data under the scale-free network structure with 20 numbers of nodes and a random sample of size twenty *(n=20)* under 1000, 2000, 5000 and 10000 Monte Carlo runs. According to the results of different trials as seen in Figure 3.1, the accuracy measures are identical under each trial.

Hereby, in the remaining of analyses, we set the number of Monte Carlo iterations as 1000 for convenience.



	F1-score	Recall	False Positive Rate	Precision	Accuracy
1000	0.320	0.207	0.026	0.467	0.902
2000	0.317	0.204	0.026	0.469	0.902
5000	0.315	0.203	0.026	0.463	0.902
10000	0.315	0.202	0.026	0.465	0.902

Figure 3.1: The results of the Monte Carlo simulations with different trials, namely, 1000, 2000, 5000 and 10000 runs.

Moreover, in our simulation, we take the total number of nodes, i.e. genes, that is also named as the dimension of the networks as 20, 50 and 100 nodes. For each dimension, a random sample of size twenty *(n=20)* is drawn from the simulated multivariate data to be modeled by GGM. The findings are shown tables below. In these comparative analyses, the simulations are separated into two parts which are the GGM application in multivariate normal data and multivariate data with the Gaussian copula function.

First of all, the multivariate normally distributed data are created under different biologic networks and dimensions. Four types of biologic networks, which are scale-free, cluster, random and hubs are used to assess their differences in the implementation of GGM. Thereby, after generating 20 observations for each node in the system, the graphical lasso (glasso) method is implemented to infer the graph path. To compare the estimated graph

path with the actual one; the values of *the precision, recall, F_1-score, accuracy* and *the false positive rate* are calculated.

Then in the second stage of the study, the multivariate data are simulated via the copula functions and applied in GGM. Due to the computational feasibility, the Gaussian (Normal) copula function is taken to create multivariate data under different dimensional systems. In the Gaussian copula application, distinct marginals are considered. In this assessment, we use the normal, student-t, log-normal and the exponential marginals within the copula function with their suitable parameters. Then, similar to the previous part of analyses, the generated data are modelled via GGM and the inference of the model parameters is conducted by the glasso approach.

Finally, as selected for the first part of the analyses, the accuracies of the results are evaluated under a wide range of measures, namely, the *precision, recall, F_1-score, accuracy* and *the false positive rate*. By this way, we aim to comprehensively analyze the plausible limitation of GGM and its advantages.

3.1. Simulation under Multivariate Normal Distribution

Nowadays, thanks to the high-throughput data-collection techniques and the microarrays, we can visualize the negotiation of cell's components at any time. Also new technologies which are “Protein Chips” or “semi-automated Yeast Two-Hybrid Screens”, allow us to understand how and when these cell's components interact with each other [5]. In order to describe the underlying interactions, there exist different types of undirected biological networks structures. These are scale-free, hubs, cluster and random networks. Among them, the scale-free networks are the most common type [5]. Although, most of the cellular networks are scale-free, we also want to check the capability of GGM on other biological network types which are hubs, cluster and random.

To be more precise, the scale-free network means that most nodes participate in only a few interactions while a few nodes participates in dozens that is in strong contrast to the random networks. Whereas, the random networks assume that a fixed number of nodes is connected randomly to each other. The major reason behind the popularity of the scale-free network is that new nodes are prone to link to the node which has many links. On the other hand, the hubs networks imply that most of the nodes have only a few links when a few nodes have a very large number of links. So we can conclude that the presence of the large hubs create scale-free networks.

Lastly, the cluster networks describe locally distributed various subgraphs of highly linked groups of nodes. These subgraphs capture specific patterns of connections [5].

In this part, we show the application of the Gaussian Graphical Model in the multivariate normally distributed data. In more details, to understand the adequacy of GGM, we simulate multivariate normal data under the mentioned different graph structures and the biological systems under distinct dimensions.

In the application of the Gaussian graphical model, as stated in previous chapters, there exist three different methods, listed as the maximum likelihood approach, the shrinkage covariance matrix and the lasso-based graphical methods in order to estimate the precision matrix. Among them, we use the lasso-based graphical method (glasso) in inference due to its computational efficiency and accuracy in the application of the sparse covariance matrix.

In the application of the glasso method, the estimation of the Lagrange multiplier, λ, has a crucial role. In order to find to the optimal λ, there exist different methods such as the Baneerjee method and k-cross validation method [4]. Among alternatives, we apply the rotation information criterion (RIC) to find the optimal λ due to its gain in accuracy [49].

In our analyses, we initially estimate the precision matrix and compare the population (true) precision matrix with the estimated graph one based on their adjacency (path) matrices. The adjacency matrix is a square matrix and represents the graph linkages. It

contains zero and one entries and its elements show whether pairs of nodes are adjacent or not in the biological graph. The table, in the following part, show the simulation results of quantitative comparisons of different methods and dimensional systems, where we repeat the experiments 1000 times and report the average of the *precision, recall, F_1-score, false positive rate, true positive rate* and *the accuracy values.*

3.1.1. Scale-Free Network

As it's mentioned previously, the scale-free network whose representation for 50 genes is shown in Figure 3.2 as example, means that many of the genes cooperate with only a few genes while a few cooperate with dozens. Moreover, most of the cellular networks have the scale-free graph structure [5]. Here, we show the results of the Gaussian graphical model in the application of the multivariate normal data which have the scale-free graph structure. There are four different tables below. The confusion matrices in Tables 3.1-3.3 show the true positive, false negative, false positive and the true negative values for the different dimensional systems. Additionally, Table 3.4 presents the comparison of the precision, recall, F_1-score, false positive rate, true positive rate and the accuracy values for the selected dimensional scale-free networks.

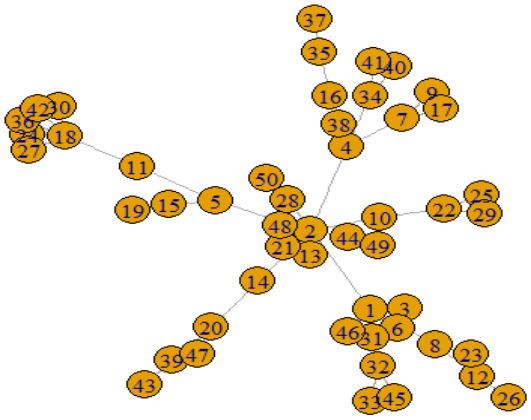


Figure 3.2: An example of Scale-free network with 50 nodes whose data are generated from multivariate normal distribution.

Table 3.1: The confusion table for a scale-free system with 20 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.020	0.075
	Negative(0)	0.022	0.883

Table 3.2: The confusion table for a scale-free system with 50 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.001	0.038
	Negative(0)	0.001	0.960

Table 3.3: The confusion table for a scale-free system with 100 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.000	0.020
	Negative(0)	0.000	0.980

As we see in the above tables, the number of true positive percentage, which is the number of the correct prediction of the actual linkage over all edges in the system (sum of the direct edges and conditional edges), is quite low. When the system has twenty nodes, GGM model only finds 2% of ones truly. However, it can find 88% of zeros correctly. When the dimension of the system is getting larger, the effectiveness of GGM to find the edges between genes decreases. On the other hand, GGM is good at to find the conditional independence and it can capture the sparsity of the graph by finding the true zero values. As stated beforehand, GGM applies the sparsity pattern of the inverse covariance matrix

to estimate the graph structure. This situation becomes better when the data have larger dimensions. We mean that GGM can detect the zero values better than ones.

Table 3.4: The accuracy table for a scale-free system based on 1000 Monte Carlo runs.

	Precision	Accuracy	Recall (TPR)	FPR	F_1-score
Perfection Level / Number of Nodes	1	1	1	0	1
20	0.467	0.902	0.207	0.0262	0.320
50	0.390	0.960	0.016	0.001	0.030
100	0.249	0.980	0.001	0.000	0.001

In Table 3.4, we represent the accuracy measures and their perfection levels which are the best values. From the results, it is seen that the accuracy measures of GGM decrease sharply while the network becomes larger. Accordingly, the accuracy value shows the overall performance of the model to estimate zeros and ones, properly. Moreover, we expect FPR closer to zeros. Because it is the proportion of wrongly estimated zeros in the sample to all zeros in the population.

Here, when the dimensions increase, the precision values get smaller and become far from to 1. It means that the model loses the classification power when the dimensions raise. Similar to the precision, the recall and the F_1-score values decrease when GGM describes larger systems.

To the contrary, the accuracy measure increases when the dimension of the system reaches 100 nodes. Because GGM is more successful in estimating the zero values. Also, as we expected, FPR decreases when the dimension of the scale-free network increases.

Since, the ability of GGM to predict ones decreases, while to predict zeros increases under higher dimensional systems.

3.1.2. Random Network

As it is mentioned beforehand, the random networks whose visual example is represented in Figure 3.3 for system with 50 nodes assume that a fixed number of genes is connected randomly to each other [5]. Here, we show the application of GGM on the multivariate normal data which have the random network feature. There are four different tables which list the true positive, false negative, false positive and the true negative values for the 20, 50 and 100 dimensional systems under this network type. Additionally, Table 3.8 indicates the comparison of the precision, recall, F_1-score, false positive rate, true positive rate and the accuracy values for these random systems.

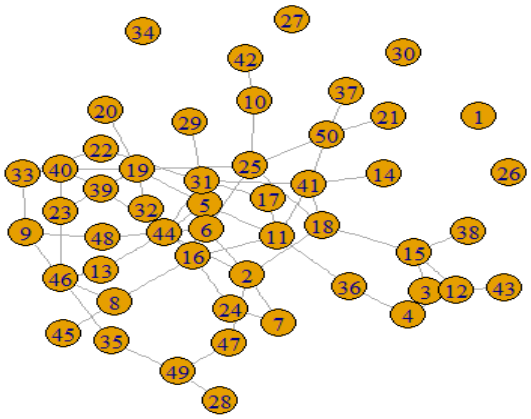


Figure 3.3: An example of Random network with 50 nodes whose data are generated from multivariate normal distribution.

Table 3.5: The confusion table for a random system with 20 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
		Positive (1)	Negative(0)
True Path	Positive (1)	0.027	0.115
	Negative(0)	0.023	0.835

Table 3.6: The confusion table for a random system with 50 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
		Positive (1)	Negative(0)
True Path	Positive (1)	0.001	0.057
	Negative(0)	0.001	0.941

Table 3.7: The confusion table for a random system with 100 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
		Positive (1)	Negative(0)
True Path	Positive (1)	0.000	0.030
	Negative(0)	0.000	0.970

In Table 3.5, GGM only finds 2.7% of the actual linkage of the population path (ones). Also, it can find 83.5% of the conditional dependence of the population path (zeros), correctly. The percentage of the correctly estimated conditional dependence of the

population path is lower than the scale-free network results. For the estimation of the true edge, the situation is vice versa. When the dimension of the system is getting larger, in Table 3.6 and 3.7, the effectiveness of GGM to find the edges between genes decreases. However, GGM is good at to find the conditional independence, i.e., zero values. Because, it can capture the sparsity of the graph much better. When the data have larger dimensions, GGM is getting better to find sparse relation of the graph path. As a conclusion, GGM can detect the zero values better than ones, especially, when the dimension of the system increases.

Table 3.8: The accuracy table for a random system based on 1000 Monte Carlo runs.

	Precision	Accuracy	Recall (TPR)	FPR	F_1-score
Perfection Level / Number of Nodes	1	1	1	0	1
20	0.540	0.862	0.192	0.028	0.303
50	0.497	0.942	0.018	0.001	0.035
100	0.492	0.970	0.001	0.000	0.002

In Table 3.8, there exist the mean values of accuracy measures for 1000 Monte Carlo runs and their perfection levels which are the best scores. Accordingly, the accuracy value shows the overall performance of the model to estimate zeros and ones, properly. For the random network results, GGM has better precision values than the scale-free networks. Because the random network has more ones (1) in its path. With detail, 14% of the random

network has direct edges (1) for 20 dimensional systems when only 9.5% of the scale-free networks has direct edges (1) under this condition. Moreover, FPR, which is the power of the estimated conditional dependence, closer to zero. Once again, it is seen that the accuracy measures of GGM decrease sharply while the network becomes larger. For the precision value, it means that the model loses the classification power. Similar to the precision, the recall and the F_1-score values decrease when GGM works with the larger dimensional systems. To the contrary, the accuracy measure increases when the dimension of the system reaches 100 nodes because of the power of GGM under sparsity.

3.1.3. Cluster Network

As distinct from other network structures, the cluster networks, as simply drawn via 50 genes in Figure 3.4 for an illustration, create the network's subgroups within genes. In detail, the cluster networks mean that there exist locally distributed various subgraphs of highly linked groups of genes. These subgraphs capture the specific patterns of connections [5]. Here, we present the application of GGM on the multivariate normal data which have this network feature. Tables 3.9-3.11 display the associated results. From these tables, we observe the true positive, false negative, false positive and the true negative percentages for the three different dimensions, i.e., network with 20, 50, 100 nodes, and Table 3.12 indicates the comparison of the precision, recall, F_1-score, false positive rate, true positive rate and the accuracy values for these systems.

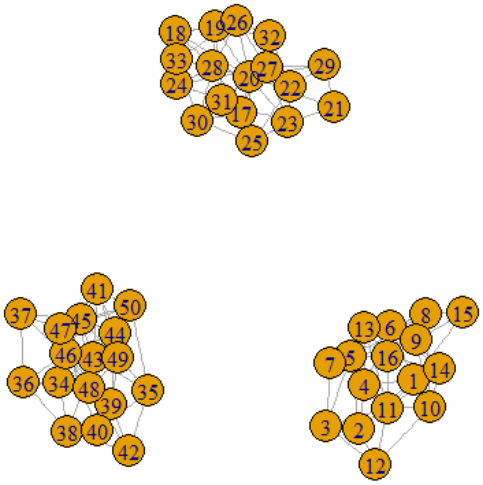


Figure 3.4: An example of Cluster network with 50 nodes whose data are generated from multivariate normal distribution.

Table 3.9: The confusion table for a cluster system with 20 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.031	0.238
	Negative(0)	0.024	0.707

Table 3.10: The confusion table for a cluster system with 50 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.000	0.110
	Negative(0)	0.000	0.890

Table 3.11: The confusion table for a cluster system with 100 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.000	0.057
	Negative(0)	0.000	0.943

In Table 3.9, GGM only finds 3.1% of the actual linkage of the population path, i.e., ones, and can detect 71% of the conditional dependence of the population path, i.e., zeros, correctly. The percentage of the truly estimated conditional dependence of the population path is lower than both the scale-free and the random networks. Furthermore, for the estimation of the true edge, GGM gives the highest percentage among other network types. Moreover, when the dimension of the system gets larger, the effectiveness of GGM to find the edges between genes decreases as similar to previous cases. To conclude, GGM can detect the zero values better than ones; especially, as the dimension of the system

increases. But it has the best performance to truly estimate direct edge between nodes among others.

Table 3.12: The accuracy table for a cluster system based on 1000 Monte Carlo runs.

	Precision	Accuracy	Recall (TPR)	FPR	F_1-score
Perfection Level / Number of Nodes	1	1	1	0	1
20	0.573	0.738	0.118	0.035	0.198
50	0.540	0.888	0.015	0.002	0.030
100	0.485	0.943	0.000	0.000	0.002

In Table 3.12, there exists the mean value of accuracy measures for 1000 Monte Carlo runs and their perfection levels. From the results of the cluster network results, it is observed that GGM has the highest precision value because of the power to estimate present links. According to the precision value, the model loses the classification power moderately when the dimension raises. Also, FPR, which is the power of the estimation of the conditional dependence, is closer to the zero when the dimension of the systems gets larger. Because when the dimension increases, GGM starts to assign more zeros to the estimated path due to its sparse nature. So, the recall, FPR and the F_1-score of GGM reach zero while the network has 100 nodes.

3.1.4. Hubs Network

Shortly, the hubs networks mean that most of the genes have only a few links when a few genes have a very large number of links. A simple visual representation of this network type for n systm with 50 nodes is shown in Figure 3.5. Moreover, the large number of hubs creates scale-free networks by getting together [5]. Here, we present the application of the GGM on the multivariate normal data which have the hubs network feature. Similar to previous findings, we construct Tables 3.13-3.15 to show the true positive, false negative, false positive, true negative values and the comparison of the precision, recall, F_1-score, false positive rate, true positive rate and the accuracy values for 20, 50 and 100 dimensional systems.

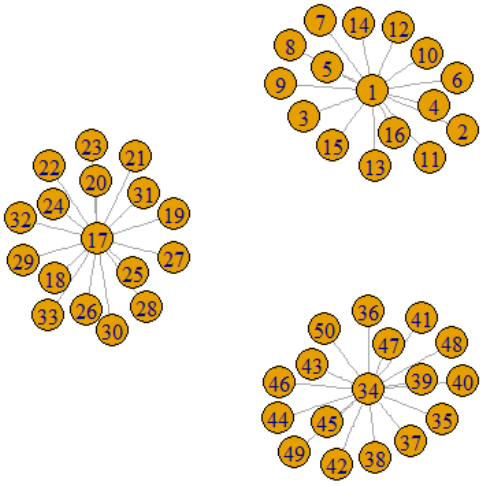


Figure 3.5: An example of Hubs network with 50 nodes whose data are generated from multivariate normal distribution.

Table 3.13: The confusion table for a hubs system with 20 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.027	0.063
	Negative(0)	0.028	0.882

Table 3.14: The confusion table for a hubs system with 50 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.001	0.036
	Negative(0)	0.001	0.961

Table 3.15: The confusion table for a hubs system with 100 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.000	0.073
	Negative(0)	0.000	0.927

From Table 3.13, it is observed that the results of the hubs almost are close to the outputs of the scale-free. But GGM under the hubs network is more effective to estimate the actual

linkage than its performance under the scale-free network. Moreover, the true positive percentage, which is the number of the correct prediction of the actual linkage is divided by all edges in the system (sum of the direct edges and conditional edges), is 2.7%. When the system has twenty nodes, GGM models only find 88% of zeros truly. As expected, when the dimension of the system becomes larger, the effectiveness of GGM to find the edges between genes decreases. Additionally, GGM is successful in finding the conditional independence and it can capture the sparsity of the graph by finding the true zero values. This situation becomes better when the data have larger dimensions.

Table 3.16: The accuracy table for a hubs system based on 1000 Monte Carlo runs.

	Precision	Accuracy	Recall (TPR)	FPR	F_1-score
Perfection Level / Number of Nodes	1	1	1	0	1
20	0.487	0.909	0.296	0.033	0.428
50	0.456	0.962	0.033	0.002	0.061
100	0.466	0.926	0.002	0.000	0.004

On the other hand, the accuracy value shows the overall performance of the model to estimate zeros and ones, simultaneously. From the results of Table 3.16, we represent the accuracy measures and their perfection levels. In this table, the precision values do not change too much when the system becomes larger. The possible reason behind it could be the compact structure of the hubs networks. It is seen that the recall, FPR and the F_1-

score of GGM decrease sharply while the network has high dimensions. Here, when the dimensions increase, the precision values do not change much. On the other side, for the hubs networks, the model has a similar classification power when the dimensions raise. But for the accuracy, TPR and the FPR, GGM loses the power in estimation when the dimensions increase.

3.2. Simulation under Multivariate Data with the Gaussian Copula

In this part, we show the application of GGM in the multivariate data bounded via the Gaussian Copula. Hereby, at this chapter, the main aim is to understand how GGM works with the copula functions when modeling the cellular network. Here, similar to previous applications of multivariate normal data, we compare the structure of the population graph with the structure of the sample graph.

In our assessment, we produce the multivariate data by using the Gaussian copula function. Because, GGM assumes that the linkage between genes can be explained by using the precision matrix which is an inverse of the covariance matrix. Here, to explain the multivariate dependence, we have to put all dependent measurements of variables into a complete and positive defined matrix which implies the covariance structure in the Gaussian copula. On the other hand, the Archimedean copula families, which are composed of the Gumbel, Frank and Clayton, copulas are constructed with only a single dependency parameter θ [28]. Furthermore, according to Whittaker [45], it is not clear which parameters create a reliable model under which values and which dependence structure can be created by the given copula function. Moreover, the Archimedean family, the Gumbel and Clayton copula do not have explicit density expressions if we infer their copula terms [43]. Hereby, as the Gaussian copula does not have these limitations, we perform it in our analyses with a wide variety of marginals.

Accordingly, the multivariate data are simulated with student- t, log-normal and semi-normal, semi-exponential marginals and fully exponential scenarios. At this point, we choose the log-normal distribution as the marginal of the Gaussian copula because of its wide application in biological systems [23].

In the application of GGM, as stated in previous chapters, there exist three different methods, listed as the maximum likelihood approach, the shrinkage covariance matrix and the lasso-based graphical methods in order to estimate the precision matrix. Among them, we use the lasso-based graphical method (glasso) in inference due to its computational efficiency and the accuracy in the sparse covariance matrix.

In the application of the glasso method, the selection of the Lagrange multiplier, λ, is another important point. In order to find the optimal λ, there exist various approaches such as the Baneerjee method and the k-cross validation method [49]. In this study, we apply the rotation information criterion (RIC) to find the optimal λ due to its high accuracies among alternatives [2].

The following sections represent the mean results of the 1000 Monte Carlo simulations for distinct marginals and dimensional systems. To carry out the evaluation of GGM under the Gaussian copula function, we tabulate the average of all selected accuracy measures, namely, *the true positive, false positive, false negative, true negative, precision, recall, F_1-score, false positive rate, true positive rate* and *the accuracy* by comparing the estimated adjacency matrix with the true ones generated under the scale-free feature.

3.2.1. The Student-t Marginals

The student-t distribution has the rising sloping curve at the right-hand side and the descending sloping curve at the left-hand side. These tails tell us that there exist the extreme quantiles for the student-t distributions. The tails of the student-t distributions decrease more slowly than the tails of the normal distribution. Hereby, the distribution of student-t decreases when the degrees of freedom increase. Accordingly, when the degrees of freedom approach to infinity, this distribution approaches to the standard normal distribution. Hence, the student-t distribution with an infinity number of degrees of freedom implies that it is completely same as the standard normal distribution. But in general, the student- t distribution has more widespread shape than the normal distribution [15].

Thus, when simulating the multivariate data with the Gaussian Copula function, we apply the student-t margins with the degrees of freedom 10 to assess its performance far from normality. Because as mentioned above, the higher degrees of freedom bring us the wider data distribution. Then, similarly, we repeat the experiments under 1000 Monte Carlo runs and report the average of these related values. The results of these simulations are given in Tables 3.17-3.19. The findings indicate that the performance of GGM under all accuracy measures decreases with respect to the outputs under the normality.

Table 3.17: The confusion table for student-t margins with 10 degrees of freedom and 20 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.005	0.090
	Negative (0)	0.004	0.901

Table 3.18: The confusion table for student-t margins with 10 degrees of freedom and 50 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.000	0.039
	Negative (0)	0.000	0.961

Table 3.19: The confusion table for student-t margins with 10 degrees of freedom and 100 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.000	0.020
	Negative (0)	0.000	0.980

In Table 3.17, there are 400 edges between 20 nodes, i.e., (20x20)-dimensional matrix, and almost 10% of these edges are direct, i.e., the entries ones in θ. But GGM only finds almost 1% of them under the student-t marginals. Hence, it is seen that the student-t marginals are modelled with suitable parameters (degree of freedom for similar features to the normal distribution), the result are drastically low. Moreover, when the dimension of the system gets larger, the effectiveness of GGM to find the edges between genes decreases. For the (100x100)-dimensional precision matrix, GGM cannot find any relation between nodes, i.e., it can only assign zero values.

Table 3.20: The accuracy table for student-t margins with 10 degrees of freedom based on 1000 Monte Carlo runs.

	Precision	Accuracy	Recall (TPR)	FPR	F_1-score
Perfection Level / Number of Nodes	1	1	1	0	1
20	0.575	0.906	0.053	0.004	0.092
50	0.533	0.961	0.000	0.000	0.000
100	Not Computable	0.980	0.000	0.000	0.000

In Table 3.20, the precision values are quite high even higher than the result of the multivariate normal data. But other measures such as the recall and the F_1-score are quite low. The precision values are high. Because if we check Table 3.17, we realize that the model estimates only 1% of the edges as one (1) and 99% of them as zero (0). Indeed, this is the deficiency of GGM in the sense that when the data are far from the normality, GGM starts to assign zero values (conditional independence) for each node. Moreover, these values reach zero when the dimensions of the systems reach 50 nodes. The accuracy measure, which is the proportion of the total number of correct predictions to the total number of all classified object, reaches 0.98, almost one. So as it is expected, GGM cannot capture the direct edges between nodes successfully, rather, it can assign zero values in the majority of the entry in the precision matrix.

3.1.2.2. The Log-Normal Marginals

The log-normal distribution is an asymmetric and continuous distribution which is a good representative for the data with the extreme positive values. Furthermore, it has a strong alliance with the normal distribution. Because if the log transformation of random variables distributes normally, the random variables can distribute log-normally. Although, the normal distribution is the most well-known density and has many application areas, the log-normal distribution is prevalent in many fields too [23]. For example, the multiplicative rule which is essential in chemistry and physics also valid for the log-normal distribution while the normal distribution has the additive rule [20].

Hence, we choose this density as a marginal of the Gaussian copula as the most biological mechanisms (exponential growth), chemical phenomenon (the velocity of a simple reaction) and biological systematics induce in this distribution [20]. Additionally, according to Kapteyn (1903) [19], if the data from one-dimensional measurements in nature fit the normal distribution, but if two and three dimensional results such as surfaces and volumes cannot be symmetric, the log-normal distribution has quite profitable features.

Moreover, at different standard deviation levels, it has different distributional shapes in such a way that for higher standard deviations, it becomes more skewed and captures more extreme values. When the standard deviation decreases, the shape of the distribution looks like more symmetric similar to the normal distribution.

Thus, in order to evaluate this characteristics, we simulate multivariate data under different parameters of the log-normal distribution. In this regard, we take the value of the standard deviation as 8 to assess the performance of GGM under the skewed log-normal. Also, we generate data with a lower standard deviation value, taken as 0.5, in order to evaluate the outcomes under a more symmetric log-normal density.

The findings of the simulations can be seen in Tables 3.21-3.23. From the results, it is observed that the performance of GGM becomes even worse than the results under the student-t distribution and all accuracy measures decrease sharply regarding the outputs under the normality.

Table 3.21: The confusion table for log-normal margins (mean=10, standard deviation=8) with 20 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.000	0.095
	Negative (0)	0.000	0.905

Table 3.22: The confusion table for log-normal margins (mean=10, standard deviation =8) with 50 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.000	0.039
	Negative (0)	0.000	0.961

Table 3.23: The confusion table for log-normal margins (mean=10, standard deviation =8) with 100 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
		Positive (1)	Negative(0)
True Path	Positive (1)	0.000	0.020
	Negative (0)	0.000	0.980

Table 3.24: The accuracy table for log-normal margins (mean=10, standard deviation=8) based on 1000 Monte Carlo runs.

	Precision	Accuracy	Recall (TPR)	FPR	F_1-score
Perfection Level / Number of Nodes	1	1	1	0	1
20	Not Computable	0.905	0.000	0.000	0.000
50	Not Computable	0.9608	0.000	0.000	0.000
100	Not Computable	0.9802	0.000	0.000	0.000

Table 3.25: The confusion table for log-normal margins (mean=10, standard deviation=0.5) with 20 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.000	0.095
	Negative (0)	0.000	0.905

Table 3.26: The confusion table for log-normal margins (mean=10, standard deviation=0.5) with 50 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.000	0.039
	Negative (0)	0.000	0.961

Table 3.27: The confusion table for log-normal margins (mean=10, standard deviation=0.5) with 100 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.000	0.020
	Negative (0)	0.000	0.980

Table 3.28: The accuracy table for log-normal margins (mean=10, standard deviation=0.5) based on 1000 Monte Carlo runs.

	Precision	Accuracy	Recall (TPR)	FPR	F_1-score
Perfection Level / Number of Nodes	1	1	1	0	1
20	Not Computable	0.905	0.000	0.000	0.000
50	Not Computable	0.9608	0.000	0.000	0.000
100	Not Computable	0.9802	0.000	0.000	0.000

We compare the results of two data with different standard deviations, but results are identical. For each standard deviation, GGM cannot capture any direct edge between nodes and the model assigns only zeros. In Tables 3.24 and 3.27, the recall and F_1-score are zero and the precision value cannot be calculated because of zero results. Hence, we conclude that, because of the violation of the normality assumption (even working with the Gaussian copula and log-normal marginals), GGM cannot give effective results for the estimated path.

3.2.3. Semi-Exponential and Semi-Normal Marginals

Previously, the multivariate data are simulated by the help of the Gaussian copula function with single types of marginals. In these simulations, student's-t, log-normal and

the exponential distributions are used as marginals, separately. In addition to the simulation of the Gaussian copula with these single marginal types, we also model the copula function with mixed marginals. Thereby, our main purpose is to observe how results of GGM change when the Gaussian copula function is described under combined marginals. Hereby, we bind the exponential and normal distributions, in our measurements so that the Gaussian copula function can be used for half exponential margins and half normal margins. Here, we take the exponential marginals' rate parameter λ as 4 and the normal marginals' mean as 0 with the standard deviation 2. The results of the Monte Carlo simulations are represented in Tables 3.29- 3.32.

Table 3.29: The confusion table for semi-exponential (rate=4), semi normal marginals (mean=0, standard deviation=2) with 20 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.000	0.095
	Negative(0)	0.000	0.905

Table 3.30: The confusion table for semi-exponential (rate=4), semi normal (mean=0, standard deviation=2) marginals with 50 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.000	0.039
	Negative(0)	0.000	0.961

Table 3.31: The confusion table for semi-exponential (rate=4), semi normal (mean=0, standard deviation=2) marginals with 100 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.000	0.020
	Negative(0)	0.000	0.980

Table 3.32: The accuracy table for semi-exponential (rate=4), semi normal (mean=0, standard deviation=2) marginals based on 1000 Monte Carlo runs.

	Precision	Accuracy	Recall (TPR)	FPR	F_1-score
Perfection Level / Number of Nodes	1	1	1	0	1
20	Not Computable	0.905	0.000	0.000	0.000
50	Not Computable	1	0.000	0.000	0.000
100	Not Computable	1	0.000	0.000	0.000

Similar to the result of the log-normal marginals, GGM cannot capture any direct edge between nodes and the model can merely assign zeros. Due to the deviation from the

normality assumption (even working with the semi normal data), GGM cannot give effective result for the estimated path. Accordingly, in Table 3.32, the recall and F_1-score are zero for each dimensional size and the precision value cannot be calculated because of these zero results.

3.2.4. Exponential Marginals

The exponential distribution is a well- known continuous distribution with the rate parameter λ ($\lambda>0$) which is the only parameter of the distribution. Furthermore, it is positive defined with an interval $[0, \infty)$. Moreover, it has the positive-skewed (inverse J) shape [3]. In this study, when simulating the multivariate data with the Gaussian copula function, we use the exponential margins with the rate $\lambda=4$.

Tables 3.17-3.19 show the simulation results from the comparison of different dimensional systems under 1000 Monte Carlo runs. Here, we report the average of these related values.

Table 3.33: The confusion table for exponential margins with 20 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
True Path		Positive (1)	Negative(0)
	Positive (1)	0.070	0.026
	Negative(0)	0.428	0.476

Table 3.34: The confusion table for exponential margins with 50 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
		Positive (1)	Negative(0)
True Path	Positive (1)	0.020	0.019
	Negative(0)	0.331	0.630

Table 3.35: The confusion table for exponential margins with 100 nodes based on 1000 Monte Carlo runs.

		Estimated Path	
		Positive (1)	Negative(0)
True Path	Positive (1)	0.007	0.013
	Negative(0)	0.240	0.740

In Tables 3.33, 3.34 and 3.35, surprisingly, GGM can capture some direct edges between nodes. It can find almost 7% of the true links for 20-dimensional system. But it is not good at the detection of zeros truly as much as under the multivariate normal data. For the dimensional size 20, it is approximately equal to 0.5 chance to assign 0 and 1 values. The reason behind this results can be the skewed shape of the exponential distribution which can simplify the visibility of the link in the estimation via the penalized maximum likelihood approach. Finally, similar to previous findings, when the dimension of the system increases, GGM loses the power.

Table 3.36: The accuracy table for exponential margins (rate=4) based on 1000 Monte Carlo runs.

	Precision	Accuracy	Recall (TPR)	FPR	F_1-score
Perfection Level / Number of Nodes	1	1	1	0	1
20	0.138	0.545	0.723	0.501	0.648
50	0.058	0.650	0.515	0.352	0.550
100	0.028	0.748	0.347	0.247	0.434

In Table 3.36, GGM has the highest recall and F_1-score values which indicate even better output under the multivariate normality. On the other hand, the precision values are low since GGM can estimate some of the true links. Whereas, the accuracy values still increase while the number of nodes increases as the model still tends to assign more zero values.

CHAPTER 4

REAL DATA APPLICATION

In this chapter, in order to detect the deficiency of GGM on real systems, we implement GGM on the real data and show their results. Previously, we have found that GGM starts to fail when the data are not normally distributed. In here, we apply GGM on the two different biological dataset whose descriptions are presented below.

4.1. Application via Cell Signalling Protein Data

In the application of GGM in biological data, we use the cell signalling data which contain information about 11 phosphoproteins and some phospholipids [32]. These 11 proteins are called as praf, pmek, plcg, PIPP2, PIP3, p44.42, pakts473, PKA, PKC, P38 and pjnk where each of them has 1000 observations, resulting in 11000 measurement totally. Here, our main purpose is to model the data with GGM and to compare the estimated network with the true system to evaluate the efficiency of GGM. This dataset is gathered to measure the biologic relations of proteins and the true structure of the network is represented in Figure 4.1.

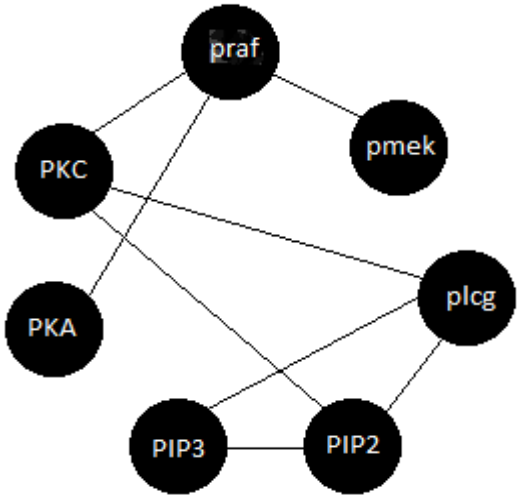


Figure 4.1: The true network of the cell signaling proteins from the study of Sachs' et al., (2005) [32].

According to the study Sachs et al., (2005) [32], we create the above figure (Figure 4.1) to visualize the true network of the proteins. In Figure 4.1, the protein PIP2 has direct edges with the protein PIP3, PKC and the protein plcg. Also, the protein plcg has direct linkages with PIP2, PIP3 and PKC. The protein praf has direct edges with the protein PKC, PKA and pmek. Moreover, according to Sim and Scott (1999) [36], there exists the direct linkage between the PKA and PKC proteins [36]. But, from the modeling of this dataset via GGM, we find none of the underlying true links in Figure 4.1. GGM can merely assign zero entries in the precision matrix for all estimated interactions. As a conclusion, GGM cannot capture any true network between 11 proteins.

In order to critic the plausible reason behind this estimation, we check the QQ-plots of each protein. As seen in Figure 4.2, we find that the distributions of each marginal protein are far from the normal density, although the structure of the system is suitable for the GGM-type of the mathematical modelling

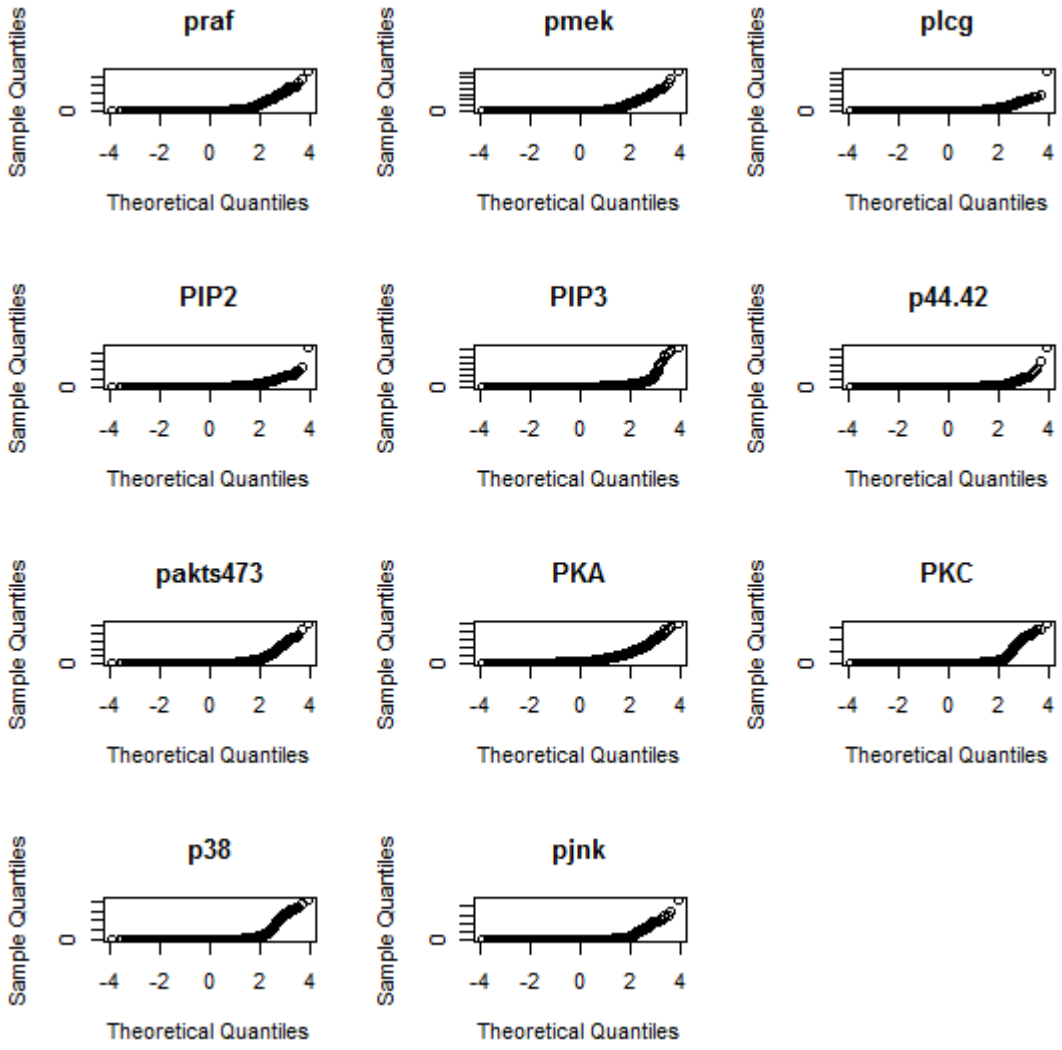


Figure 4.2: The QQ-plots of cell signalling data by comparing the normal density.

Hence, we check whether the data can be multivariate normal after transformations. As seen in Figure 4.2, since the protein have heavy right hand side tails, we initially eliminate outliers and among the remaining 320 observations, we perform log and log (log) transformations as these two types of the transformations are biologically meaningful.

We see that both transformations cannot solve the non-normality problem. On the other hand, we also find that some proteins indicate constant value, i.e. no change during whole activation time of experiment. Such type of fixed value can be observed for the proteins controlling the growth factor or the proteins whose degradation rates, i.e. half- life, are very slow. These type of proteins typically have crucial role in the initiation of the cellular activation of apoptosis. Thus, they may not be converted normal in any transformation. On the other side, if we remove these proteins in the modelling, we cannot use the same true network for comparison. Because, the activation of whole system can change as the proteins are dependent on each other from the nature of their activations.

4.2. Application via Human Gene Expression Data

In the second real data application, we use the human gene expression data which contain 100 transcripts (with unique Illumina TargetID) measured on 60 unrelated individuals. The data are collected by Stranger et al., (2007) [38] and are defined by Bhadra and Mallick (2013) [7] and Chen et al., (2007) [11]. The purpose of the data is to understand the gene expression in the B-lymphocyte cells from the Northern and Western European ancestry from Utah (CEU). The main focus of these studies is the 3125 Single Nucleotide Polymorphisms (SNPs) which are found in the 5 UTR (untranslated region). Because UTR (untranslated region) of mRNA is quite important to control the gene expression.
In order to find the biological links in these data, we use GGM with the glasso approach. But, similar to the previous results GGM cannot discover any of the validated links presented in Table 4.1. On the other hand, according to the study of Bhadra and Mallick (2013) [7], 26 biological interactions of these data are discovered. We create Table 4.1 from the study of Bhadra and Mallick (2013) [7] to show all the validated interactions with the name of the genes. Thus, similar to the cell signaling data, we conclude that the application of GGM can be restricted for small and moderately large systems as its

inference can be better accomplished via approximate methods for high dimensional networks.

Here, in order to check the normality of this dataset as the source of deficiencies of the model, we compute the Shapiro-Wilk test for the multivariate normality in the R programme and we take the significance level 0.05. As a result, we obtain p-value < $2.2e^{-16}$ which is smaller than any significance level and we conclude that at least one of the variable is not coming from the normal distribution. Also, the QQ-plots of the first fifteen genes are drawn, as examples in Figure 4.3, to visualize the distribution of the gene expressions. Both test results and plots show that the human gene expression data do not have normal distribution.

Table 4.1: Biologically validated links in the human gene expression data [7].

LINKS	
GI.7019408.S-GI.4504436.S	GI.21614524.S-GI.34222299.S
GI.28610153.S-GI.4504436.S	GI.37537705.I-GI.31652245.I
GI.20070269.S-GI.28610153.S	GI.18641371.S-GI.41197088.S
GI.18379361.A-GI.20070269.S	GI.16159362.S-GI.31652245.I
GI.17981706.S-GI.13514808.S	GI.21389558.S-GI.16159362.S
GI.20302136.S-GI.7661757.S	GI.28557780.S-GI.16159362.S
GI.4505888.A-GI.41350202.S	GI.27477086.S-GI.16159362.S
GI.27754767.I-GI.16554578.S	GI.23510363.A-GI.28557780.S
GI.9961355.S-GI.27754767.I	GI.27482629.S-GI.23510363.A
GI.27754767.I-GI.27754767.A	GI.28416938.S-GI.27482629.S
GI.22027487.S-GI.27754767.I	GI.30795192.A-GI.27482629.S
GI.38569448.S-GI.22027487.S	GI.24308084.S-GI.27477086.S
GI.34222299.S-GI.22027487.S	GI.4504700.S-GI.19224662.S

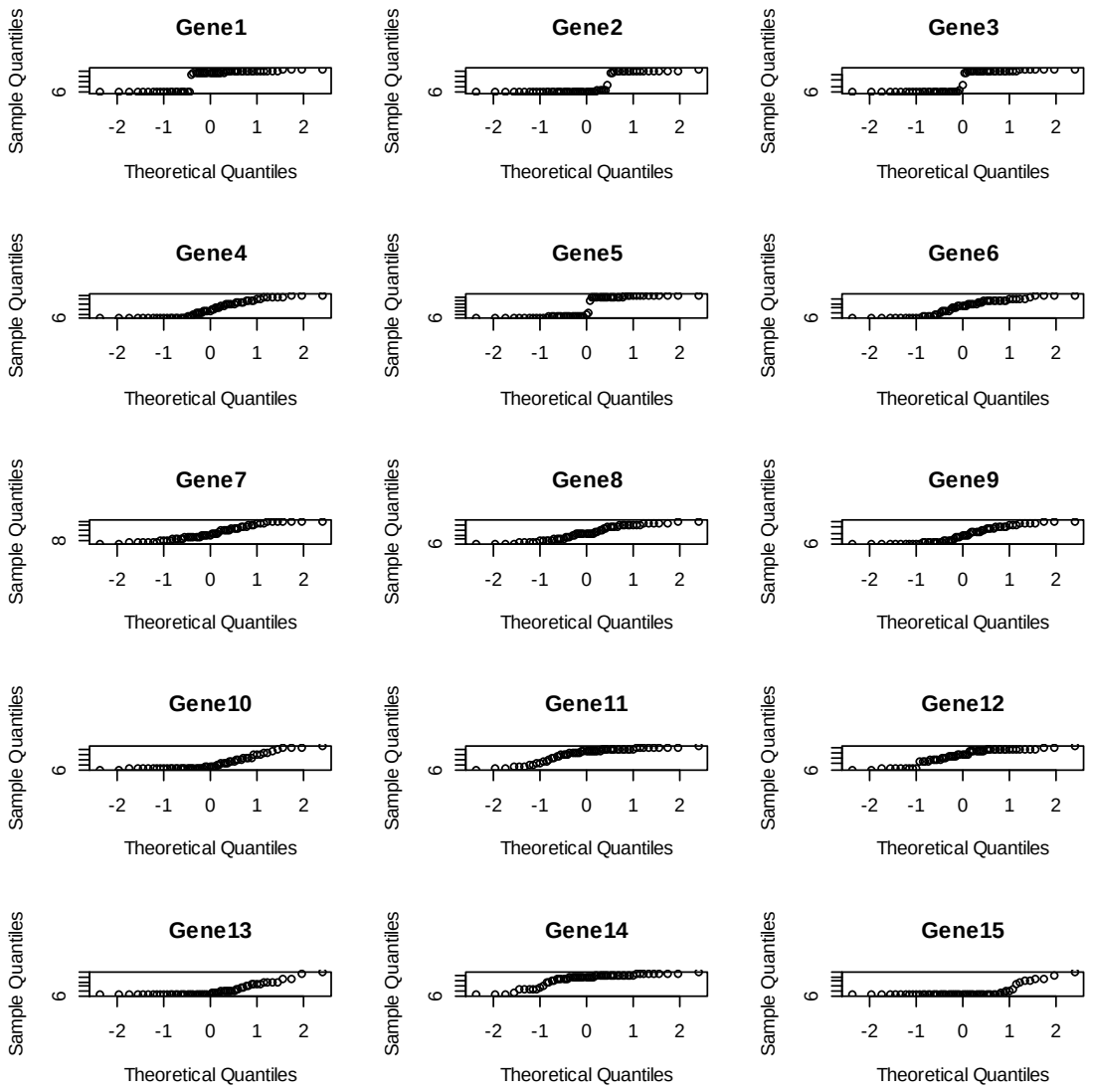


Figure 4.3: The QQ-plots of the human gene expression data [7] by comparing the normal density.

According to tests' results and plots, it is detected that the human gene expression data are not normal. Furthermore, like cell signaling data, we cannot apply the transformation to the measurements to make them normal. Because, we face with the same challenges in

the cell signaling protein data in the sense that they are heavily right tailed and certain genes show no change in their activation throughout the experiment.

4.3. Application via Palm Oil Data

As a part of the real data application we use the palm oil data which contain information about the lipid contents of the developing palm oil related with the major lipid metabolites [29]. The data are gathered by the Oo et al., (1985) [29]. The lipid biosynthesis is an important topic in order to develop the nutritional and technical properties of the crop oil. The lipid biosynthesis process is highly related with the formation of the triacyl-glycerols (TAGs). Because they are the end products of the lipid biosynthesis process. TAGs have commercial interest and they can be modified in their relative quantities to increase the quality of the overall oil due to their feature to store oils [31]. The data show the changes in the lipid content (gram) of triacylglycerols (TAG), fatty acids (FA), diacylglycerols (DAG), monoacylglycerols (MAG) and polar lipids (PL) for five different measurement periods as presented in Table 4.2.

Table 4.2: The palm oil dataset [29].

Weeks	TAG	FA	DAG	MAG	PL
8	0.01	0.01	0.02	0.00	0.05
12	0.02	0.01	0.03	0.00	0.08
16	5.08	0.51	0.39	0.08	0.34
20	36.75	6.17	2.69	1.76	0.41
24	19.32	39.06	3.95	0.21	4.15

In Table 4.2, the entries under the proteins names show the observed concentration during the given weeks.

In order to discover the network of the lipids, we model this dataset by using GGM with the glasso method and compare the findings with the quasi true structure of the network given in Figure 4.4. As a result, we detect that GGM cannot infer any link between the selected variable.

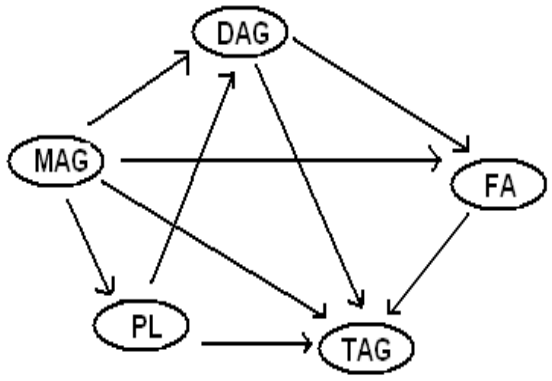


Figure 4.4: The network of the lipid metabolites

In order to investigate the cause of nonmorality behind this result, we apply the one-sided Kolmogorov-Smirnov (KS) test. The tabulated values with respect to the significance level $\alpha=0.05$ are shown in Table 4.3.

Here, the null hypothesis is taken as the normality of the data. According to the findings of the test, we conclude that FA, DAG, MAG and PL proteins are distributed normally, except TAG since all the associated p-values are greater than the significance level, $\alpha=0.05$.

Table 4.3: The results of the Kolmogorov- Smirnov (KS) Test under significance level $\alpha=0.05$

Variables	P-values of KS Test	Conclusion
TAG	0.030	Reject Null Hypothesis
FA	0.158	Fail to Reject Null Hypothesis
DAG	0.102	Fail to Reject Null Hypothesis
MAG	0.164	Fail to Reject Null Hypothesis
PL	0.088	Fail to Reject Null Hypothesis

Then to convert all measurements to normal, we apply log transformation to the TAG Lipid content as it is the only non-normal protein. After the transformation, we compute the KS value again and show that the new measurements are also normal at $\alpha=0.05$. Whereas, after performing GGM in this normal dataset, we observe that GGM cannot still estimate any true link. We consider that this situation may be caused by the limitation of GGM under very mall dataset even under normality.

CHAPTER 5

CONCLUSION and OUTLOOK

In this study, we have considered to comprehensively evaluate the performance of Gaussian graphical model (GGM) which is one of the common modelling approaches for the description of the steady-state behaviors of biological systems. For this purpose, we have assessed the findings of GGM, first of all, under different dimensions and then the topology of the networks and under various distributions.

In all these calculations, we have computed the accuracy of the estimates based on various accuracy measures.

Thus, for the analyses in the first stage we have applied GGM in multivariate normal data under distinct graph structures and dimensional sizes. In the analyses via Monte Carlo simulations, we have detected that the hubs network have the best precision, TPR and F_1-score values. Then, the scale-free networks and the hubs network have relatively better performance in terms of the underlying accuracy measures. Because these network types have quite similar features. Although, the random and the cluster networks have higher precision value, their F_1-score and TPR values are lower. But, in general, for all networks types, the power of the GGM estimation decreases when the size of the system increases. In other words, with moderately small number of genes, GGM can successfully explain the true structure of the networks, whereas, with large number of genes and multivariate normally distributed data, GGM cannot find the true links.

On the other side, in the second stage of the analyses, we have conducted simulations with the Gaussian copula under different marginals. We have prefered the Gaussian copula since the Archimedean copulas which are Gumbel, Frank and Clayton, are considered by only one parameter θ and they do not have explicit solutions [43]. Furthermore, only the Gaussian copula can create multivariate data by using the positive defined covariance matrix.

At first stage, as marginals of the Gaussian copula, we have used the student-t and the log-normal because of their similar features with the normal distribution. After that, we have also used semi marginals (semi-normal and semi–exponential) choices and lastly we use the exponential distribution as margins.

From the analyses, we have observed that the results of the student-t are similar to the results of the multivariate normal data. Also, it has higher precision values. But for other accuracy measures, it has lower values. Here, we have found that this is the deficiency of GGM in the sense that when data are far from the normality, GGM starts to assign zero values (conditional independence) for each node. Moreover, these values reach zero when the dimensions of the systems reach 50 nodes.

But for the log-normal and the semi-normal, semi- exponential data, the results become worse. When the marginals of the joint function contain both normal and exponential distributions, the true networks cannot be modelled well under the Gaussian copula. GGM cannot capture any direct edge between nodes and the model can merely assign zeros. Due to the deviation from the normality assumption (even working with the semi-normal data), GGM cannot give effective results for the estimated paths.

On the other side, the results of the exponential margins are surprisingly good in such a way that they have higher TPRs and F_1-scores than the application of the multivariate normal data. Furthermore, we have seen that with small number of genes and Gaussian Copula under exponential marginals, GGM is succesful in capturing the true links. Here, GGM can give better classification due to the fact that its underlying estimation method,

MLE, is working with the skewed data better. More specifically, the exponential distribution is good for the signaling data and the measurements of our networks indicate a kind of signal data containing extreme values.

On conclusion from the Monte Carlo simulation analyses, GGM is effective in modelling the small and moderately large systems under the multivariate normally distributed data. However, the performance of GGM becomes worse when the data are far from normality.

On the other hand, in real data applications, we model the three different datasets which are cell signaling, human gene expression and palm oil. Although, the true networks of the variable are already known, GGM cannot find any linkage between variables. On the contrary, it assigns only zero values for each network between variables.

Therefore, from this study, we have realized that GGM is not sufficient to estimate network structures under all conditions and it has certain strong drawbacks. Generally, it cannot capture true links when the size of the system increases and the normality assumption is violated. Hence, we consider that the modelling of complex biological systems can be performed by non-parametric models in place of GGM as they are free from any distributional assumptions. Thereby, MARS method can be applied as it is a good substitute for dealing with linear and nonlinear relationship between variables when they are highly correlated [25]. In order to handle the non-normal data, the non-paranormal SKEPTIC algorithm can be also another strong alternative that is based on the non-parametric optimization [25].

Moreover, Conic MARS (CMARS) and Robust Conic MARS (RCMARS) which are the extended version of the MARS method can be performed since they are promising competitive of MARS. Furthermore the algorithms which are designed for particularly high dimensional and correlated measurements such as the random forest algorithm can be implemented for the construction of networks [44].

Additionally, in our study, we only use to the RIC criterion for the estimation of the inverse covariance matrix. Hence, GGM can be applied under different criteria such as ICOMP [9], CAIC [10] and BIC [35] during the selection of the best fitted model which means the selection of the optimal penalty constant.

Finally, in the simulation part of the study, we have faced with the difficulties in the construction of multivariate copulas. Creating a bivariate copula function is not problematic, but when the dimension of copulas increases, obtaining the density of copulas get progressively difficult. To handle that, we can create multivariate distributions via a copula-vine method by identifying marginals and their dependence structures. So, the analyses of the copula can be extended by the application of the Canonical vine (C-vine) and D-vine method [34]. These vines have similar constructions for any number of variables. Here, in C-vine, one variable links to all other variables and determines the dependency structure. In the D-vine method, the linkage is more symmetric and each node has at most two links [43]. Therefore, both methods, C-vine and D-vine can be used for more comparative analyses with GGM under large scales of copulas.

Moreover, in this study, we have only applied GGM on the biologic datasets. As the future work, we intent to evaluate its performance on the financial datasets. Because modelling the non-linear dependence between economic and financial variables have an increasing concern and GGM can be applied for modelling these datasets.

REFERENCES

1. Atay-Kayis, A., & Massam, H. (2005). A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika,* 92, 317-335.

2. Ayyıldız, E. (2013). Gaussian graphical model in estimation of biological systems, Statistics Department, Middle East Technical University, Ankara.

3. Balakrishnan, N., & Basu, A. P. (1995). Exponential Distribution: Theory, Methods and Application. Columbia: Gordon and Breach.

4. Banerjee, O., Ghaoui, L., & D'Aspremont, A. (2008). Model selection through parse maximum likelihood estimation. *Journal of Machine Learning Research*, 9, 485-516.

5. Barabási, A. L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization, *Nature Reviews Genetics,* 5,101-113.

6. Berg, D. (2008). Using Copulas. An introduction to practitioners. Oslo: ASTIN.

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

8. Bouye, E., Durrleman, V., Nikeghbali, A., Riboulet, G., & Roncalli, T. (2000). Copulas for finance: A reading guide and some applications, 2, 7-40.

9. Bozdoğan, H. (1987). Model selection and akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52 (3), 345-370.

10. Bozdoğan, H. (1988), ICOMP: A new model selection criterion, H.H. Bock (Ed.), Classification and Related Methods of Data Analysis, North-Holland, Amsterdam, 599-608.

11. Chen, L., Emmert-Streib, F., & Storey, J. (2007), Harnessing naturally randomized transcription to infer regulatory relationships among genes, Genome Biology, 8, R219.

12. Çelik, G. (2010). Parameter estimation in generalized linear model with conic quadratic programming, Institute of Applied Mathematics, Middle East Technical University, Ankara.

13. Dauwels, J., Yu, H., Xu, S., & Wang, X. (2013). Copula Gaussian graphical model for discrete data. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver.

14. Dobra, A., Lenkoski, A., & Rodriguez, A. (2011). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*, 106, 1418-1433.

15. Finner, H., Dickhaus T., & Roters, M. (2008). Asymptotic tail properties of student's t-distribution. *Communications in Statistics - Theory and Methods*, 37 (2), 175-179.

16. Friedman, J. H., Hastie, T., & Tibshiran, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, 2, 302-332.

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

18. Genest, C., & Favre, A. C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydroelectric Engineering*, 12 (4), 347-368.

19. Kapteyn, J. C. (1903). Skew Frequency Curves in Biology and Statistics. Groningen: Noordhoff.

20. Koch, A. L. (1966). The logarithm in biology. II. Mechanisms generating the lognormal distribution exactly. *Journal of Theoretical Biology*, 23, 276–290.

21. Kohavi, R., & Provost, F. (1998). On Applied Research in Machine Learning. In Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process (Vol. 30). New York, NY: Columbia University Press.

22. Li, X., & Xu, R. (2009). High-Dimensional Data Analysis in Cancer Research. New York, Springer.

23. Limpert, E., Stahel, W. A., & Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience*, 51 (5), 341-351.

24. Liu, H., Han, F., Yuan, M., Lafferty, J., & Wasserman, L. (2012). High dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40, 2293-2326.

25. Liu, H., Lafferty, J., & Wasserman, L. (2012). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs, Journal of Machine Learning Research.

26. Meinshausen, N., & Buhlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34 (3), 1436–1462.

27. Muhammadi, A., & Wit, E. C. (2014). Bayesian structure learning in sparse Gaussian graphical models. Bayesian Analysis, 10, 108-138.

28. Nelsen, R. B. (2006). An Introduction to Copulas. Second Edition. Portland: Springer Science & Business Media.

29. Oo, K. C., The, S. K., Khor, H. T., & Ong, S. H. (1985). Fatty acid synthesis in the oil palm (Elaeis guineensis): incorporation of acetate by tissue slices of the developing fruits, *Lipids*, vol. 20, pp 205-210.

30. Patton, A. J. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis*, 1, 2-10.

31. Quek, E., Purutcuoglu, V., Sambanthamurthi, R., & Weber G.-W. (2013). Modelling lipid biosynthesis pathways of oil palm by Boolean and graphical approaches, HIBIT 2011 conference in Izmir, Turkey, 129- 135.

32. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D., & Nolan, G. (2005), Causal proteinsignaling networks derived from multiparameter single-cell data, *Science*, 308 (5721), 523– 529.

33. Schafer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4 (1), 1-27.

34. Schirmacher, D., & Schirmacher, E. (2008). Multivariate dependence modeling using pair-copula. Technical report, *Society of Actuaries*.

35. Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6 (2), 461-464.

36. Sim, A. T. R., & Scott, J. D. (1999), Targeting of PKA, PKC and protein phosphatases to cellular microdomains, *Cell Calcium*, 26 (5), 209–217.

37. Sklar, A. (1959), Fonctions de répartition à n dimensionset leurs marges. Publ. Institute du Statistics, Université de Paris. Paris, 8, 229-231.

38. Stranger, B., Nica, A., Forrest, M., Dimas, A., Bird, C., Beazley, C., Ingle, C., Dunning, M., Flicek, P., Montgomery, S., Tavare, S., Deloukas, P., & Dermitzakis, E. (2007), Population genomics of human gene expression, *Nature Genetics*, 39, 1217–1224.

39. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1), 267-288.

40. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society*, Series B, 67, 99-108.

41. Trivedi, P. K., & Zimmer, D. M. (2005). Copula Modeling: An Introduction for Practitioners. 1(1), 1-111.

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

43. Wawrzyniak, M. M., & Kurowicka, D. (2006). Dependence concepts. Delft University of Technology, Delft Institute of Applied Mathematics, Delft, Netherlands.

44. Weber, G. W., Batmaz, I., Kösal, G., Taylan, P., & Yerlikaya-Özkurt, F. (2012). CMARS: a new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization, Inverse Problems in Science and Engineering, pp. 371-400.

45. Whittaker, J. (1990). Graphical models in Applied Multivariate Statistics. Chichester: John Wiley & Sons.

46. Wit, E. Vinciotti, V., & Purutçuoğlu, V. (2010). Statistics for biological networks: short course notes. 25th International Biometric Conference (IBC), Florianopolis, Brazil.

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

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

49. Zhao, T., Luin, H., & Simon, N. (2012). The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research*, 13, 1059-1062.

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

51. Zou, H., & Hastie, T. (2005). Regularization and variable selection via elastic net. *Journal of the Royal Statistical Society*, Series B, 67, 91-108.