Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (4) (2021), 1172 – 1184
DOI : 10.15672/hujms.728352

Research Article

Vine copula graphical models in the construction of biological networks

Hajar Farnoudkia[1], Vilda Purutçuoğlu*[2]

[1] *Department of Management, Faculty of Economics and Administrative Sciences, Başkent University, Ankara, Turkey*

[2] *Department of Statistics, Faculty of Arts and Science, Middle East Technical University, Ankara, Turkey*

Abstract

The copula Gaussian graphical model (CGGM) is one of the major mathematical models for high dimensional biological networks which provides a graphical representation, especially, for sparse networks. Basically, this model uses a regression of the Gaussian graphical model (GGM) whose precision matrix describes the conditional dependence between the variables to estimate the coefficients of the linear regression model. The Bayesian inference for the model parameters is used to overcome the dimensional limitation of GGM under sparse networks and small sample sizes. But from the application in bench-mark data sets, it is seen that although CGGM is successful in certain systems, it may not fit well for non-normal multivariate observations. In this study, we propose the vine copulas to relax the strict normality assumption of CGGM and to describe networks from a variety of copulas alternates besides the Gaussian copula. Accordingly, we evaluate the best fitted bivariate copula distribution for every pairwise gene and compute the estimated adjacency matrix which denotes the presence of an edge between the corresponding genes. We assess the performance of our proposed approach in three network data via distinct accuracy measures by comparing the outputs with the results of the CGGM.

Mathematics Subject Classification (2020). 62H05, 92Bxx, 62Fxx

Keywords. Vine copulas, biological networks, reversible jump MCMC, systems biology

1. Introduction

The construction of biological networks is a challenging problem since the amount of available data increases sharply and the interpretation of these data sets becomes important to understand complex systems diseases like cancers. Hereby, many mathematical models are suggested to better describe this complexity. The graphical models are one of the successful modeling groups in this field. Basically, these models present the interactions between systems elements, which are genes or proteins affecting the flow of activation in networks via undirected edges that are typically computed by pairwise correlations between underlying elements. The copula Gaussian graphical model (CGGM),

*Corresponding Author.
Email addresses: hajer.farnoudkia@baskent.edu.tr (H. Farnoudkia), vpurutcu@metu.edu.tr (V. Purutçuoğlu)
Received: 28.04.2020; Accepted: 08.04.2021

which is one of the recent and promising graphical models, explains the functional relationship between genes under the multivariate normal distribution. In this model, the joint function is represented by the Gaussian copula in a lasso regression. The inference of this model is conducted via different Bayesian algorithms. The reversible jump Markov chain Monte Carlo (RJMCMC) algorithm is one of the methods used to construct the conditional dependence between the nodes in the CGGM introduced by [12]. Herein, the captured dependence structure is the undirected edge between the nodes and the used copula is the Gaussian copula because of its exclusive property that the uncorrelateness implies the independence. But in some cases, Gaussian may not be an appropriate model between the marginals of the variables since it requires the symmetry and a zero tail of the dependence. So another copula could be more appropriate for modeling this type of data sets. Thereby, in this study, we aim to use vine copulas which enable us the flexibility to select the non-Gaussian copula for every pair of genes in the construction of protein-protein interaction networks models. As the common properties of all these mentioned models is that they are parametric approaches. But, in the literature of the construction of biological networks, there exists different types of non-parametric models as well. One of the recent methods is called the loop-based multivariate adaptive regression splines (LMARS) [1, 4] and its conic version [5]. These two models adapt the multivariate adaptive regression splines (MARS) and conic MARS (CMARS) model for the protein-protein interaction networks by iteratively performing MARS and CMARS, respectively, for each protein via main effects and second-order interaction terms. Indeed, these two models have wide application in different fields such as in finance [2], supply chain management [29], optimization problems in mathematics [46], analysing the environmental statistics [28] and neuroscience [9]. Besides MARS and its extension, there are some other alternative non-parametric approaches that have been already adapted to the description of the complex biological networks or can be adapted in this field with some modifications. The random forest [14, 37], generalized partial linear model [30, 44], neural networks [18, 23], support vector machines [22] and ordinary differential equation models [15] are some examples for this type of non-parametric methods. On the other side, in terms of handling the uncertainty there are some general methods for complex models. The robust optimization [25, 27] stochastic optimal control [36, 41] and the chance constrained optimization methods [3, 16, 21] can be seen some known examples in this area. In general, these methods are successful in the description of the systems. Whereas, they do not take into account the distributional knowledge of the data for their model descriptions. Thus, the parametric approaches are considered if the distributional features of the data are known. Hereby, in order to include this information, distinct parametric approaches are suggested. The Gaussian graphical model is one of the strong alternatives while presenting the systems' changes via the multivariate normal distribution [20, 38]. Whereas, its inference is limited for the large system. Accordingly, the CGGM whose inference is conducted either the Bayesian algorithm [24, 32] or the vine copula approaches, as mentioned previously, are the alternative solutions for the limitation of the Gaussian graphical model.

Accordingly, in the first section, RJMCMC is introduced briefly. Then, in the Materials and Methods part, the copula is defined as its types and the formats of vine copulas. Finally, in the Application and Conclusion parts, we compare the mentioned methodologies by some accuracy measures such as F_1 score and Matthews correlation coefficient and summarize our results, respectively.

2. Materials and methods

Gaussian copula that is used for CGGM via RJMCMC in inference has some advantages such as using the correlation to get the dependence and is also being conjugate with the G-Wishart distribution. In RJMCMC, the inverse of the correlation matrix which is called the

precision matrix, shows the conditional dependence in such a way that each zero element of the precision matrix implies the conditionally independence between corresponding variables. In the graphical models, each variable is indicated by a node and the conditional dependence between them is described by an undirected edge. But, we do not need to know the strength of the edges in CGGM. On the other hand, the vine copula deals with the joint distribution function which can be written by bivariate (pair) copulas in a way that all of the pair copulas can be from a different copula type with different parameters. Hereby, the advantage of the vine copula is its flexibility and its ability to define the full model without the necessity of any specific assumption. In the following part, initially, we explain briefly the RJMCMC algorithm within CGGM and then, present shortly some methods for the selection of the best pairwise vine copula models.

2.1. Reversible jump markov chain Monte Carlo method

The Gaussian graphical model (GGM) is the probabilistic version of the graphical approach where the nodes are described by a multivariate normal distribution with a p-dimensional mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)$ and a $(p \times p)$-dimensional covariance matrix $\boldsymbol{\Sigma}$ for totally p nodes. The precision matrix, which is the inverse of $\boldsymbol{\Sigma}$ and also denoted by $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, represents the conditional dependence between nodes in a way that the significantly large values point a highly possible dependency between two related nodes given remaining nodes in the network. Thereby, the mathematical description of the model is denoted as below

$$Y_p = \boldsymbol{\beta} Y_{-p} + \boldsymbol{\epsilon}, \tag{2.1}$$

where Y_p stands for the state of the pth node and $\boldsymbol{Y_{-p}}$ shows the states of all other nodes except the pth node, respectively. $\boldsymbol{\beta}$ is a vector of the regression coefficient associated with $\boldsymbol{Y_{-p}}$ and $\boldsymbol{\epsilon}$ refers to the p-dimensional vector of the normally distributed random error. Accordingly, the distribution of Y is indicated as

$$f(Y|\boldsymbol{\mu}, \boldsymbol{\Theta}) = (2\pi)^{-n/2} \det(\boldsymbol{\Theta})^{n/2} \exp\{-\frac{1}{2}(Y - \boldsymbol{\mu})^T \boldsymbol{\Theta} (Y - \boldsymbol{\mu})\}. \tag{2.2}$$

Herein, det(.) and $(.)^T$ describe the determinant and the transpose of the given matrix, in order. Thus, in the inference of this model, $\boldsymbol{\beta}$ has a direct relation with $\boldsymbol{\Theta}$ via $\boldsymbol{\beta} = \frac{\boldsymbol{\Theta}_{-pp}}{\Theta_{pp}}$ in which $\boldsymbol{\Theta}_{-pp}$ is the $((p-1) \times p)$-dimensional submatrix of $\boldsymbol{\Theta}$ when the associated term of the pth node is discarded. So, the knowledge of $\boldsymbol{\Theta}$ implies the knowledge of $\boldsymbol{\beta}$, resulting in the information about the coefficients of the regression expression came from the conditional dependency between the related nodes.

RJMCMC is an approach which deals with mostly the Cholesky decomposition to get a positive definite precision matrix due to its conjugate advantageous in the prior distribution for $\boldsymbol{\Theta}$ which is supposed as the G-Wishart distribution with a density

$$p(\boldsymbol{\Theta}|G) = \frac{1}{I_G(\sigma, D)} \exp\{-\frac{1}{2} tr(\boldsymbol{\Theta}^T \boldsymbol{D})\}. \tag{2.3}$$

In this expression, G implies the given graphical structure of the data. On the other hand, the G-Wishart prior is the generalized version of the chi-square distribution and the conjugate with the multivariate normal density [43]. Thus, the posterior distribution $\boldsymbol{\Theta}$ of the given G is presented as the G-Wishart distribution with parameters $(\sigma + n)$ and $(\boldsymbol{D} + \boldsymbol{U})$. Accordingly, the RJMCMC algorithm performs a three-stage procedure by utilizing the Metropolis-Hasting algorithm to calculate the probability of the update in every step, namely, resampling the latent data, resampling the precision matrix and resampling the graph iteratively, until all parameters are convergent. The mathematical details of each step can be found in [12].

2.2. Vine copula

The base theorem of the general version of the copula method is the Sklar's theorem in which every joint distribution function of two or more variables can be written by their marginal distributions and a copula as in the below

$$F(\boldsymbol{y}) = C(F_1(y_1), F_2(y_2), .., F_d(y_d), \tag{2.4}$$

where $\boldsymbol{y} = (y_1, y_2, .., y_d)^T$ and F is the d-dimensional cumulative distribution function of the random variable Y. It means for every two or more random variable, there is a unique copula that defines the joint distributions of the variables based on their marginal distributions and a copula term for the dependence between them.In order to use the multivariate copula some strict assumptions are needed which are discussed in detail in the study of [19]. As a solution for the complexity problem of the multivariate copula, the vine copula decomposes the joint distribution function $f(\boldsymbol{y})$ into the bivariate cases to reduce the complexity of the multivariate model. Therefore the vine copula is based on the conditional bivariate distributions.

To show how the multivariate distribution function is decomposed to some bivariate functions by applying the Sklar's theorem and the chain rule of probability and the conditional distribution definition, the simplest form of the multivariate joint distribution .i.e, $d = 3$ is used. In the first step the chain rule of probability is used as below

$$f(y_1, y_2, y_3) = f_1(y_1) f(y_2|y_1) f(y_3|y_2, y_1). \tag{2.5}$$

The $f(y_2|y_1)$ term of Equation 2.5 can be written as its definition for the conditional probability function via $\frac{f(y_2,y_1)}{f_1(y_1)}$. Then, the Sklars theorem is applied in a way that $f(y_1, y_2) = c_{1,2}(F_1(y_1), F_2(y_2)) f_1(y_1) f_2(y_2)$. Accordingly the conditional distribution function of y_1 and y_1 is written as

$$f(y_2|y_1) = c_{1,2}(F_1(y_1), F_2(y_2)) f_2(y_2). \tag{2.6}$$

Hence, $f(y_3|y_1, y_2) = c_{(2,3|1)}(F(y_2y_1), F(y_3y_1)) c_{1,3}(F_1(y_1), F_3(y_3)) f_3(y_3)$. Finally, we have

$$f(y_1, y_2, y_3) = c_{1,2}(F_1(y_1), F_2(y_2)) c_{2,3|1}(F(y_2|y_1)F(y_3|y_1)) c_{1,3}(F_1(y_1)F_3(y_3)) f_1(y_1) f_2(y_2) f_3(y_3).$$

Meanwhile, there are more than one way to write the joint probability function as pair copula terms and marginal probability functions. In the above statement, the order is $1, 2, 3$. But, it can have another order of the variables and the structure of the network depends completely on the associated order. Hereby, below we introduce the types of vine copulas and some analytical tools that can select the order and the best pair-copula among the possible pair-copula for the systems network. In the graphical representation of the network, each node states one variable and each edge or connection undirected line shows the dependence structure between the corresponding variables (nodes). This representation is allocated to the pair copula construction while the multivariate copula representation is not possible by nodes and edges where there are more than two variables connected (depend) each other. So, the other advantage of the vine copula is its ability to be represented graphically.

2.2.1. Types of vine copula. The family types of the copula is more than two while there are two main families, namely, the Elliptical and Archimedean copulas. These families will be used to represent the dependence structure between two variable in the vine copula. Some of them have only one parameter and some have two parameters. The elliptical bivariate copula is written as $u_1, u_2 \in [0, 1]$ in the form of $C(u_1, u_2) = F(F_1^{(-1)}(u_1), F_2^{(-1)}(u_2))$, where F is a bivariate distribution function with invertible F_1 and F_2. This copula family includes the Gaussian and student-t copulas which are both symmetric with one and two parameters, respectively. Whereas, the Gaussian copula

has no tail dependence while the student-t has an extra parameter ν showing the tail-dependence, adjusted via the degrees of freedom. That means the bivariate Gaussian (student-t) function is used as the pair copula function and the inverse marginal cumulative function of y_1 and y_2 are used as u_1 and u_2 in the copula function.

On the other side, the formula of the Archimedean copula families are not straightforward unlike the Elliptical family. Therefore, the generator function ψ is used which can give the function by using the following statement

$$C(u_1, u_2) = \psi^{[-1]}(\psi_1(u_1) + \psi_2(u_2)),$$

here ψ is a strictly decreasing and continuous generator function while the pseudo inverse of the generator function is defined as $\psi^{[-1]}(t) = I_{[0,\psi(0))}(\psi^{(-1)}(t))$. Some of the one-parameter Archimedean copulas are Clayton, Gumbel, Frank and Joe. There are some other two-parameter copulas made by the combination of two one-parameter copulas that make them more flexible about the shape and the tail-dependence. The properties of the mentioned one-parameter and two-parameter Archimedean copula family are represented in Table 1.

Table 1. Denotation and properties of the bivariate Archimedean families

Name	Generator func (ψ)	Prm range	Kendall's τ	Tail dependence
Clayton	$\frac{1}{\theta}(t^{-\theta}-1)$	$\theta > 0$	$\frac{\theta}{\theta+2}$	$(2^{-\frac{1}{\theta}}, 0)$
Gumbel	$-(logt)^{\theta}$	$\theta \geq 1$	$1-\frac{1}{\theta}$	$(0, 2-2^{-\frac{1}{\theta}})$
Frank	$-log(\frac{e^{-\theta t}-1}{e^{-\theta}-1})$	$\theta \in R$	$1-\frac{4}{\theta}+4\frac{D(\theta)}{\theta}$	(0,0)
Joe	$-log(1-(1-t)^{\theta})$	$\theta > 1$	$1+\frac{4}{\theta^2}\int tlog(t)(1-t)^{2(1-\theta)/\theta}dt$	$(0, 2-2^{-\frac{1}{\theta}})$
BB1	$(t^{-\theta}-1)^{\sigma}$	$\theta > 0, \sigma \geq 1$	$1-\frac{2}{\sigma(\theta+2)}$	$(2^{-\frac{1}{\theta\sigma}}, 2-2^{\frac{1}{\sigma}})$
BB6	$(-log(1-(1-t)^{\theta}))^{\sigma}$	$\theta > 0, \sigma \geq 1$	$1+\frac{4}{\theta^2}\int(1-log(1-(1-t)^{\theta}))dt$	$(0, 2-2^{-\frac{1}{\theta\sigma}})$
BB7	$(1-(1-t)^{\theta})^{-\sigma}-1$	$\theta \geq 1, \sigma > 0$	$1+\frac{4}{\theta\sigma}\int(-(1-(1-t)^{\theta}))^{\sigma+1}dt$	$(2^{-\frac{1}{\sigma}}, 2-2^{\frac{1}{\theta}})$
BB8	$-log(\frac{1-(1-\sigma\theta)^{\theta}}{1-(1-\sigma)^{\theta}})$	$\theta \geq 1, \sigma \in (0,1)$	$1+\frac{4}{\theta\sigma}\int(-log(\frac{(1-t\sigma)^{\theta}-1}{(1-\sigma)^{\theta}-1}))dt$	$(0,0)$

The BB1, BB6, BB7 and BB8 stand for the Clayton-Gumbel, the Joe-Gumbel, the Joe-Clayton and the Joe-Frank copulas. These families are more flexible as they are the combined version of other Archimedean families which can describe the one-sided tail dependence between two variables apart from their complicated joint distributions. For instance the BB7 copula is the combination of Joe and Clayton families has two-parameter: θ for Joe copula and σ for the Clayton copula and can define the models with two-sided non-symmetric tail dependence.

The vine copulas, as discussed previously, are in the form of the bivariate or pair copula. There are two types of the vine copula which are in good order named by the canonical (shortly denoted C-vine) and the drawable (shortly denoted D-vine) copulas. The structure of the C-vine copula has the shape of the star via all trees and the nodes should be determined in advance. Thus, the order of variables is determined based on the roots of each tree. On the contrary, the structure of the D-vine copula is a path and, in this type, the first tree is the root tree and all other trees are made from the first tree. In both of vine copula types, the number of roots equal to $(d-1)$. Figure 1 shows the structure of both vine copula types for $d = 5$ as an example.

2.2.2. Analytical tools. The general form of the vine copula is called the regular vine copula (shortly denoted R-vine) is the disordered vine which includes the combination of both C and D-vine copulas. That means in some points the connection in in the star shape like the C-vine and in some other points it is in the form of a path like the D-vine. Figure 2 represents a R-vine structure.

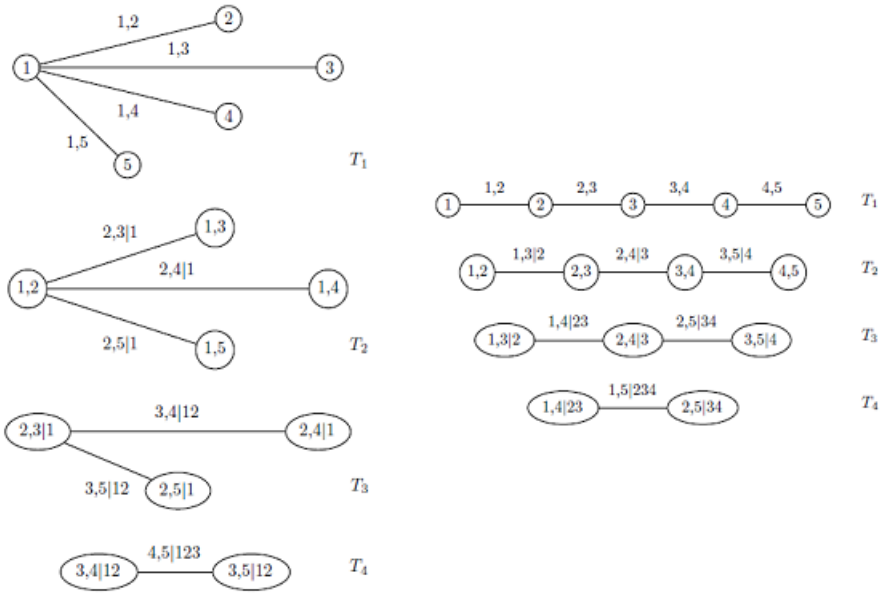

Figure 1. The examples of 5-dimensional C-(left panel) and D-vine trees (right panel) with edge indices [7]

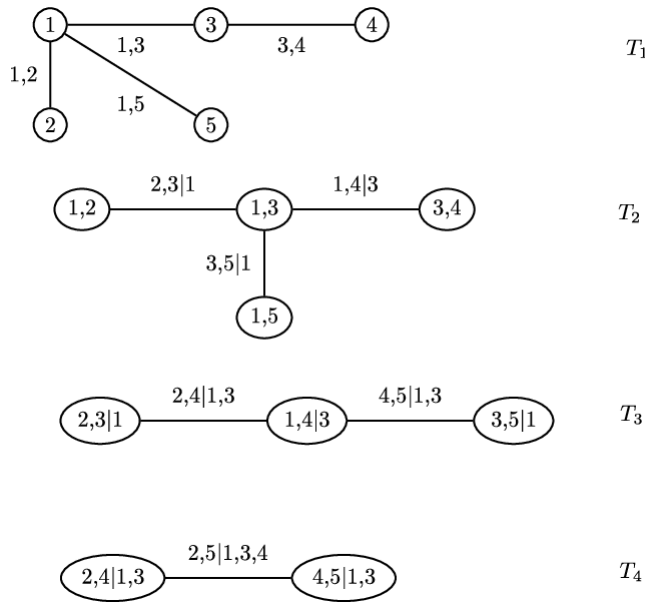

Figure 2. The examples of 5-dimensional R-vine trees (right panel) with edge indices [39]

There are several ways to write a joint density function via pair copulas as there are $\frac{(d(d-1))}{2}$ pair copulas. As it is mentioned previously, the order of variables determines the root of each tree in the C-vine and the path in the first tree in the D-vine. The algorithm of the order selection for the C-vine copula is briefly described as follows [8]:

- Compute the empirical distribution function of the data to transform them into the uniformly distributed data.
- Compute the Kendalls τ correlation coefficient of the new data and select the variable with the largest τ as the first root.

- Select the best copula for each node between the first root and other variables and then, estimate the parameter(s).
- Transform the data by conditioning to the first selected variable via a function by using the parameters estimated from the previous step.
- Select the variable among the new data which have the largest Kendalls τ as the second root.
- Continue the process until the $(d-1)$th root is found.

In order to estimate the model parameters by the maximum likelihood estimation (MLE), the order and the copula families are computed. There are some methods to select the best pair copula between the nodes, such as graphical tools like the contour plot and some other statistical tests like the Vuonge-Clarke test, which are special kinds of the goodness of fit test. Similarly, there are some other tools to compare two models by using the AIC and BIC criteria as well as the Vuonge test [7].

3. Application

In this study, we use three benchmark data sets. The first set is called the CellSignal data [35] which have 11 genes with 11672 samples. We call the second set as Data 2 which has 10 genes with 285 samples [34] and finally, the third data set is a kind of binary data to see the relationship between eight factors. These factors have effects on womens economical activities with 665 numbers of observations. The descriptions of the data sets are given in the following part. In our analysis, we compare the accuracy of RJMCMC and vine-copula methods for the data sets and in the comparison, we use the F_1-score and the Matthews correlation coefficient (MCC) whose expressions are presented in Equation 3.1 and 3.2, respectively. In these expressions, TP and FP denote the true positive and false positive, in order, and similarly, TN and FN represent the true negative and false negative values, respectively.

$$F_1 - \text{score} = \frac{2\text{TP}}{(2\text{TP} + \text{FP} + \text{FN})} \in [0, 1]. \tag{3.1}$$

$$\text{MCC} = \frac{((\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN}))}{\sqrt{((\times TP + \times FP)(\times TP + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN}))}} \in [-1, 1]. \tag{3.2}$$

Apart from $F_1 -$ score and MCC, the following accuracy measures control one type of error rates unlike the F_1-score and MCC that control every element of the confusion matrix.

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP}+\text{FN})}. \tag{3.3}$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN}+\text{FP})}. \tag{3.4}$$

Indeed, these kinds of measures are not applicable all the times as they control one sided error rate as stated previously. Hence, the assumptions and the sensitivity of the network can be helpful to select which measure should be noticed more in the given network.

3.1. The CellSignal data

The CellSignal data [35] are attached to the BDgraph package [24] with 11672 samples in which each independent measurement consists of quantitative amounts of each of the 11 phosphorylated molecules. These molecules are measured from single cells. The corresponding true network is drawn in Figure 3. In the construction of its network, we infer the system via the C-vine copulas by using the strategy given in Section 2.2.2 and obtain the order of genes as Akt, PKC, PIP2, Mek, Jnk, PIP3, Plcy, PKA, Raf, P38 and Erk.

Table 2 presents the estimated matrix of the families which is obtained by the VineCopula package [7] under the R programming language.

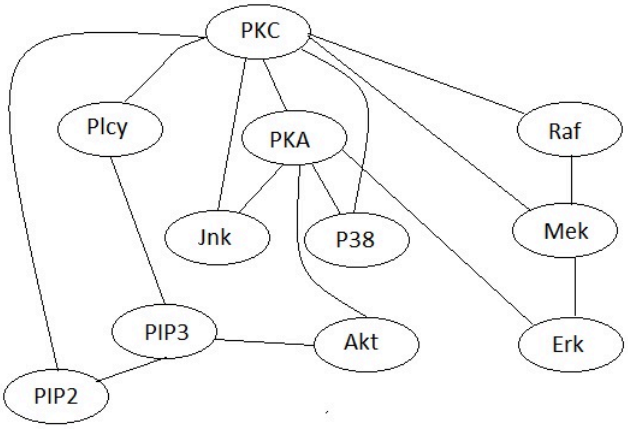

Figure 3. The true network of the CellSignal data.

Table 2. The upper triangular of the estimated adjacency matrix of CellSignal data with copula families via numbers in VineCopula package [7]. The full matrix is symmetric with respect to the diagonal

Name	**PKC**	**PIP2**	**Mek**	**Jnk**	**PIP3**	**Plcy**	**PKA**	**Raf**	**P38**	**Erk**
Akt	10	10	10	10	5	10	2	10	10	17
PKC		2	16	10	13	1	2	40	2	9
PIP2			1	1	9	10	30	1	13	9
Mek				9	13	40	29	40	2	20
Jnk					30	6	30	7	5	20
PIP3						0	26	0	0	0
Plcy							0	0	0	0
PKA								0	0	0
Raf									0	0
P38										0

In this table, the values are related to a pair of copulas and all of the zero in the upper triangle of the adjacency matrix implies the (conditional) independence between associated genes. The result of the comparison Table 2 and its true graph in the study of [35], are shown in Table 3 apart from the results of RJMCMC for this data set.

Hence, it is seen that F_1-score decreases under the vine copula with the MLE method in inference, whereas, MCC improves. As the accuracy measure, MCC can capture all TP, FP, TN and FN values resulting in a more comprehensive measure of the accuracy comparing to F_1-score. When we compare the sensitivity and the specificity measures, it is seen that the sensitivity is lower under RJMCMC with respect to vine copula, whereas, the specificity improves under RJMCMC. But the vine copula has the reverse performance under these two measures. Therefore, as shortly discussed before, regarding the importance of the TP and TN values for the study, either sensitivity or specificity values can be controlled via RJMCMC or vine copula.

Table 3. Results of some accuracy measures for the RJCMCMC and R-vine approach applied for the CellSignal data

Method	TP	FP	FN	TN	F_1	MCC	Sensitivity	Specificity
True graph	16	0	0	38	1	1	1	1
RJMCMC	8	10	11	26	0.43	0.14	0.42	0.72
R-vine	13	28	3	11	0.46	0.01	0.81	0.28

On the other hand, for this data set, other methods were used in the study of [10]. The results are shown in Table 4. In general these results indicate that the non-parametric approaches can be more preferable to describe complex networks as seen LCMARS and LMARS. On the other side, among the parametric approaches, the performance of GGM via the penalized maximum likelihood approach is closer to the performance of GGM via RJMCMC (with F_1-score 0.44 versus 0.43). Whereas, the performance of vine copula model improves this accuracy (with F_1-score 0.46). Additionally, since the inference can be conducted under MLE, it can decrease the computational demand in the estimation regarding RJMCMC. For this data set, the computational demand via the vine copula approach takes less than one minute, whereas, RJMCMC completes the calculation in more than two hours. Furthermore, the accuracy of non-parametric methods are better than parametric methods for this data set as seen from Table 4.

Table 4. The comparison of the accuracy between some non-parametric methods such as Loop-based Conic Multivariate Adaptive Regression Splines(LCMARS), Loop-based Multivariate Adaptive Regression Splines (LMARS) and Gaussian Graphical Model (GGM) under F_1-score

Method	$F_1 - score$
LCMARS	0.72
LMARS	0.69
GGM	0.44

3.2. Data 2

The second data which we use to see the performance of Vine copula, belong to a gynecological cancer network whose observations are assembled from the ArrayExpress database [31]. In these kinds of data sets, we need to have the true graph which is obtained from biological literature. This data set includes ten proteins, named as MP2K, PDA, MPK, IMP, ERB, TFM, MBD, CHD, CTNB and CBPB. These genes are also selected as the core genes in the literature of gynaecological cancer and the quasi true network structure of these genes is represented by a complete graph meaning that all the entries of the adjacency matrix are composed of ones [6]. By using the algorithm described in the study of [8] and explained in Section 2.2.2 under the application of the C-vine approach, the order of the variables is estimated as (4, 6, 5, 7, 8, 2, 10, 9, 1, 3) and the computed C-vine copulas are listed in Table 5. In this analyses, we apply the VineCopula package in R [7].

The accuracy of the R-vine copula is represented in Table 6 by comparing with the true graph measures and RJMCMC performance for this data set. Hereby, as seen in Table 5, the graph related to this network is a full graph. By comparing it with the related true network, we find F_1-score=1 indicating higher accuracy via the C-vine copula for Data 2. Whereas, the F_1-score is computed as 0.94 with RJMCMC. On the other side, MCC and specificity measures cannot be computed for this data set as both TN and FN are observed as zero. This result indicates a better accuracy under the C-vine copula model.

Table 5. The upper triangular of the estimated adjacency matrix of Data 2 with copula families via numbers in VineCopula package [7]

Name	**PDIA**	**MPK1**	**IMP**	**ERB2**	**TFM**	**MBD3**	**CHD4**	**CTNB1**	**CBPB**
MP2K	5	1	5	14	5	5	5	5	14
PDIA	0	3	13	1	1	19	5	2	1
MPK1	0	0	40	5	3	23	10	30	23
IMP	0	0	0	23	10	1	5	13	1
ERB2	0	0	0	0	3	1	5	1	2
TFM	0	0	0	0	0	14	5	1	2
MBD3	0	0	0	0	0	0	33	33	5
CHD4	0	0	0	0	0	0	0	1	5
CTNB1	0	0	0	0	0	0	0	0	5

Table 6. Results of some accuracy measures for the RJCMCMC and R-vine approach applied for Data 2

Method	TP	FP	FN	TN	F_1	Sensitivity
True graph	45	0	0	0	1	1
RJMCMC	41	0	4	0	0.95	0.91
R-vine	45	0	0	0	1	1

3.3. The Rochdale data

The Rochdale data present the eight binary (yes or no) factors that influence women activities named by **a**: wife economically active, **b**: wife age > 38, **c**: husband unemployed, **d**: the number of children 4, **e**: education level of wife, (high-school+), **f**: education level of husband (high-school+), **g**: Asian origin, and **h**: other household member working. So, the data are in eight variables done with 665 cases. The true network based on the study by [45] is in the form of fg, ef, dh, dg, cg, cf, ce, bh, be, bd, ag, ae, ad, ac. Initially, we transformed the data to the Gaussian data through a method suggested by [17] and then, the R-vine method by the algorithm designed by [11] was applied to this latent data set in order to see the relationship of those eight factors which have influences in womens activities. Our proposed method can catch $\{ef, dg, cg, cf, ce, bh, be, bd, ag, ae, ad, cd\}$. This means that 10 of 13 relationships are caught by the method and it has an overestimated relationship between c and d variables. In order to show the performance of the proposed method, it is compared with the true networks and the network found by RJMCMC via some accuracy measures listed in Table 7.

Table 7. Results of some accuracy measures for the RJCMCMC and R-vine approach applied for the Rochdale data

Method	TP	FP	FN	TN	F_1	MCC	Accuracy	Sensitivity	Specificity
True graph	14	0	0	14	1	1	1	1	1
RJMCMC	13	1	1	13	0.93	0.85	0.93	0.93	0.93
R-vine	10	1	3	14	0.83	0.72	0.86	0.77	0.93

The accuracy measures for RJMCMC is taken from [33]. Indeed, regarding this outcome, it is seen that although both accuracy measures decrease slightly under the vine copula approach, the computational demand is decreased significantly by the vine approach. Similar to the results of Data 1, the former is based on the MLE method, which is very fast and spends one minute for the computation and the latter is conducted by the Bayesian algorithm whose estimates are found via 10^6 MCMC (Markov chain Monte Carlo) runs [13].

4. Conclusion

In this study, we have been discussed two kinds of methodologies to estimate undirected biological networks. By comparing their performances and speeds, we have observed that the novel proposal approach based on the vine copula methods with the MLE inference can be a strong alternate of CGGM with RJMCMC due to its competitive performance in accuracy and gain in computational time during inference. Furthermore, with the help of the proposal vine copula approach, we can estimate complex biological systems via frequentist methods without the restriction of the Gaussian copula. As the future works, we consider to evaluate the performance of the proposed approach in different simulated data sets which can be generated under distinct network typologies, the number of observations per genes and the number of genes in the system. Furthermore, machine learning techniques [40,42] and more advanced machine learning methods such as the deep learning algorithms [26] can be applied to catch the relationship between variables in biological data sets under the non-parametric methods. Moreover, modeling based on non-parametric approaches such as the robustification of the CMARS method [28] can be also adapted to explain the large networks. We consider that the performance of these models by comparing their outcomes with the vine copula methods can be applied to detect more accurate model in the construction of the biological networks.

Acknowledgment. The second author thanks the COSTNET Project (No: CA15109) for their support. Both authors thank the editor and the anonymous referees for their valuable comments which improve the quality of the paper.

References

[1] M. Ağraz and V. Purutçuoğlu, *Extended lasso-type MARS (LMARS) model in the description of biological network*, J. Stat. Comput. Simul. **89** (1), 1-14, 2019.

[2] Ö.S. Alp, E. Büyükbebeci, A. İşcanog, F.Y. Özkurt, P. Taylan and G.W. Weber, *CMARS and GAM & CQP-modern optimization methods applied to international credit default prediction*, J. Comput. Appl. Math. **235** (16), 4639-4651, 2011.

[3] S.K. Alparslan-Gök, S. Miquel and S.H. Tijs, *Cooperation under interval uncertainty*, Math. Methods Oper. Res. **69** (1), 99-109, 2009.

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

[5] E. Ayyıldız and V. Purutçuoğlu, *Modeling of various biological networks via LCMARS*, J. Comput. Sci. **28**, 148-154, 2018.

[6] B. Bahçivancı, V. Purutçuoğlu, E. Purutçuoğlu and Y. Ürün, *Estimation of gynecologic cancer networks via target proteins*, J. Multidiscip. Eng. Sci. Technol. **5** (12), 9296-9302, 2018.

[7] E.C. Brechmann and U. Schepmeier, *Modeling dependence with C- and D-vine copulas: The R package CDVine*, J. Stat. Softw. **52** (3), 1-25, 2013.

[8] C. Czado, U. Schepsmeier and A. Min, *Maximum likelihood estimation of mixed C-vines with application to exchange rates*, Stat. Model. **12** (3), 229-255, 2012.

[9] A. Çevik, G.W. Weber, B.M. Eyüboğlu, K.K. Oğuz and Alzheimers Disease Neuroimaging Initiative, *Voxel-MARS: a method for early detection of Alzheimers disease by classification of structural brain MRI*, Ann. Oper. Res. **258** (1), 31-57, 2017.

[10] E.A. Demirci, *Inference of large-scale networks via statistical approaches*, PhD thesis, Middle East Technical University, 2019.

[11] J. Dissmann, E.C. Brechmann, C. Czado and D. Kurowicka, *Selecting and estimating regular vine copulae and application to financial returns*, Comput. Statist. Data Anal. **59**, 52-69, 2013.

[12] A. Dobra and A. Lenkoski, *Copula Gaussian graphical models and their application to modeling functional disability data*, Ann. Appl. Stat. **5** (2A), 969-993, 2011.

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

[14] B. Fellinghauer, P. Bühlmann, M. Ryffel, M. Von Rhein and J.D. Reinhardt, *Stable graphical model estimation with random forests for discrete, continuous, and mixed variables*, Comput. Statist. Data Anal. **64**, 132-152, 2013.

[15] J. Gebert, N. Radde and G.W. Weber, *Modelling gene regulatory networks with piecewise linear differential equations*, Challenges of Continuous Optimization in Theory and Applications of European Journal of Operational Research **181** (3), 1148-1165, 2007.

[16] B. Häussling Löwgren, J. Weigert, E. Esche and J.U. Repke, *Uncertainty analysis for data-driven chance-constrained optimization*, Sustainability **12** (6), 2450, 2020.

[17] P.D. Hoff, *Extending the rank likelihood for semiparametric copula estimation*, Ann. Appl. Stat. **1** (1), 265-283, 2007.

[18] A. Karacayir, *Short term electricity Load forecasting with multiple linear regression and artificial neural network*, MSc. Term Project Report/Thesis, Middle East Technical University, 2012.

[19] I. Kojadinovic and J. Yan, *Modeling multivariate distributions with continuous margins using the copula R package*, J. Stat. Softw. **34** (9), 1-20, 2010.

[20] D. Koller and N. Friedman, *Probabilistic Graphical Models Principles and Techniques*, MIT Press, Massachusetts, 2009.

[21] E. Kropat, G.W. Weber and B. Akteke-Öztürk, *Eco-finance networks under uncertainty*, in: Proceedings of the International Conference on Engineering Optimization, Rio de Janeiro, Brazil, 2008.

[22] S. Kuter, B.B. Ciftci and G.W. Weber, *Snow cover mapping from satellite data by artificial neural networks and support vector machines - An OR contribution to land-use, water management and development*, International Conference on OR for Development ICORD 2017, Quebec, Canada, July 13-14, 2017.

[23] S. Kuter, G.W. Weber and Z. Akyurek, *Artificial neural networks vs. multivariate adaptive regression splines for sub-pixel snow mapping from satellite data*, Workshop on the State of the Art and Future Development, Poznan, Poland, July 3-6, 2016.

[24] A. Mohammadi and E.C. Wit, *BDgraph: Bayesian structure learning of graphs in R*, Bayesian Analysis **10** (1), 109-138, 2015.

[25] J.M. Mulvey, R.J. Vanderbei and S.A. Zenios, *Robust optimization of large-scale systems*, Operations Research **43** (2), 264-281, 1995.

[26] M.A. Nielsen, *Neural Networks and Deep Learning*, Determination Press, San Francisco, CA, 2015.

[27] A. Özmen, *Robust Optimization of Spline Models and Complex Regulatory Networks*, Springer International Publishing, Switzerland, 2016.

[28] A. Özmen, İ. Batmaz and G.W. Weber, *Precipitation modeling by polyhedral RCMARS and comparison with MARS and CMARS*, Environ. Model. Assess. **19** (5), 425-435, 2014.

[29] A. Özmen, G.W. Weber, İ. Batmaz and E. Kropat, *RCMARS: Robustification of CMARS with different scenarios under polyhedral uncertainty set*, Commun. Nonlinear Sci. Numer. Simul. **16** (12), 4780-4787, 2011.

[30] A. Özmen, G.W. Weber and E. Kropat, *Robustification of conic generalized partial linear models under polyhedral uncertainty*, Methods **20** (21), 22, 2012.

[31] H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk and R. Mani, *ArrayExpressa*

public database of microarray experiments and gene expression profiles, Nucleic Acids Res **35** (suppl-1), D747-D750, 2007.

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

[33] V. Purutçuoğlu and H. Farnoudkia, *Gibbs sampling in inference of copula gaussian graphical model adapted to biological networks*, Acta Physica Polonica A **132** (3), 2017.

[34] Y. Rahmatallah, F. Emmert-Streib and G. Glazko, *Gene sets net correlations analysis (GSNCA): A multivariate differential coexpression test for gene sets*, Bioinformatics **30** (3), 360368, 2014.

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

[36] E. Savku and G.W. Weber, *A stochastic maximum principle for a Markov regime-switching jump-diffusion model with delay and an application to finance*, J. Optim. Theory Appl. **179** (2), 696-721, 2018.

[37] D. Seçilmiş and V. Purutçuoğlu, *Modeling of biochemical networks via classification and regression tree methods*, Mathematical Methods in Engineering, 87-102, 2019.

[38] I. Shmulevich, E.R. Dougherty and K. Seungchan, *Sparse inverse covariance estimation with the graphical lasso*, Bioinformatics **18**, 261274, 2002.

[39] J. Stöber, H.G. Hong, C. Czado and P. Ghosh, *Comorbidity of chronic diseases in the elderly: Patterns identified by a copula design for mixed responses*, Comput. Statist. Data Anal. **88**, 28-39, 2015.

[40] V. Strijov, G.W. Weber, R. Weber and S.O. Akyuz, *Editorial of the special issue in data analysis and intelligent optimization with applications*, Machine Learning **101**, 1-4, 2015.

[41] E. Todorov, *Stochastic optimal control and estimation methods adapted to the noise characteristics of the sensorimotor system*, Neural Comput. **17** (5), 1084-1108, 2005.

[42] G. Üstünkar, S.Ö. Akyüz, G.W. Weber and Y.A. Son, *Analysis of SNP-complex disease association by a novel feature selection method*, in: Operations Research Proceedings 2010, Springer, Berlin, Heidelberg, 21-26, 2011.

[43] H. Wang and S. Zhengzi, *Efficient Gaussian graphical model determination under G-Wishart prior distributions*, Electron. J. Stat. **6**, 168-198, 2012.

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

[45] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*, Wiley Publishing, 1990.

[46] F. Yerlikaya-Özkurt, C. Vardar-Acar, Y. Yolcu-Okur and G.W. Weber, *Estimation of the Hurst parameter for fractional Brownian motion using the CMARS method*, J. Comput. Appl. Math. **259**, 843-850, 2014.