Scientia Iranica E (2019) 26(4), 2495–2505

Sharif University of Technology
Scientia Iranica
Transactions E: Industrial Engineering
http://scientiairanica.sharif.edu

Copula Gaussian graphical modeling of biological networks and Bayesian inference of model parameters

H. Farnoudkia and V. Purutçuoğlu*

Department of Statistics, Middle East Technical University, Ankara, Turkey.

Received 25 August 2017; received in revised form 12 April 2018; accepted 20 May 2019

KEYWORDS
Copula Gaussian graphical model;
Reversible jump Markov chain Monte Carlo algorithm;
Biological networks;
F-measure;
Systems biology.

Abstract. A proper understanding of complex biological networks facilitates a better perception of those diseases that plague systems and efficient production of drug targets, which is one of the major research questions under the personalized medicine. However, the description of these complexities is challenging due to the associated continuous, high-dimensional, correlated and very sparse data. The Copula Gaussian Graphical Model (CGGM), which is based on the representation of the multivariate normal distribution via marginal and copula terms, is one of the successful modeling approaches to presenting such types of problematic datasets. This study shows its novelty by using CGGM in modeling the steady-state activation of biological networks and making inference of the model parameters under the Bayesian setting. In this regard, the Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm is suggested in order to estimate the plausible interactions (conditional dependence) between the systems' elements, which are proteins or genes. Furthermore, the open-source R codes of RJMCMC are generated for CGGM in different dimensional networks. In this regard, real datasets are applied, and the accuracy of estimates via F-measure is evaluated. From the results, it is observed that CGGM with RJMCMC is successful in presenting real and complex systems with higher accuracy.

1. Introduction

In recent years, the term *network or system* has become one of the most popular concepts in various sciences, ranging from computer engineering to biology. Although its property varies in all these disciplines, which is why it is implemented by means of distinct assumptions, it should be noted that a common structure constructs a mechanism specialized for one or more functions. In terms of biology, a network represents a set of reactions that describes a particular process of a species by means of genomic particles, denoting genes or proteins and their interactions. Thereby, understanding such complexity and describing it mathematically can open new avenues for researchers for both understanding various diseases and producing appropriate treatment under personalized medicine.

Therefore, graphical modeling is one of the very common tools to represent any dimensional network where each variable is shown by a node, and the relationship between two nodes is presented by an undirected edge. The *Gaussian Graphical Model* (GGM) is the probabilistic version of the graphical approach, where nodes, also called states, 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 [1]. The *precision matrix*, which is the inverse of $\boldsymbol{\Sigma}$, also

*. *Corresponding author.*
E-mail addresses: hajar.farnoudkia@metu.edu.tr (H. Farnoudkia); vpurutcu@metu.edu.tr (V. Purutçuoğlu)

doi: 10.24200/sci.2019.5071.1076

denoted by $\boldsymbol{\theta} = \boldsymbol{\Sigma^{-1}}$, is an expression that represents the conditional dependence between nodes, such that the significantly large values show highly possible dependency between the two related nodes, given the remaining nodes in the network. In this respect, the mathematical description of the model is shown below:

$$Y_p = \boldsymbol{\beta}\mathbf{Y_{-p}} + \varepsilon, \tag{1}$$

where Y_p stands for the state of the pth node, and $\mathbf{Y_{-p}}$ shows the states of all other nodes except the pth node, respectively. $\boldsymbol{\beta}$ is the vector of the regression coefficient associated with $\mathbf{Y_{-p}}$, ε shows p-dimensional vector for the random error. Accordingly, the distribution of $\mathbf{Y}$ is shown as follows:

$$f\left(\mathbf{Y}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right) = (2\pi)^{-\frac{n}{2}} \det\left(\boldsymbol{\Sigma}\right)^{-\frac{1}{2}}$$

$$\exp\{-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{Y}-\boldsymbol{\mu}\right)\}. \tag{2}$$

Herein, $det(.)$ and $(.)^T$ represent the determinant and the transpose of the given matrix, respectively. Thus, in the inference of this model, $\boldsymbol{\beta}$ has a direct relation with $\boldsymbol{\theta}$ via $\boldsymbol{\beta} = \boldsymbol{\theta_{-pp}}/\boldsymbol{\theta_{pp}}$ 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. Thus, the knowledge of $\boldsymbol{\beta}$ implies the knowledge of $\boldsymbol{\theta}$, resulting in the information about the conditional dependency between the related nodes. In the estimation of $\boldsymbol{\theta}$, different methods can be applied. Among many alternatives, Friedman et al. (2008) [2] considered the graphical lasso, also known as glasso, approach by inferring the entries of $\boldsymbol{\theta}$ via the penalized likelihood method whose penalty constant controls $\boldsymbol{\beta}$ via l_1-norm. On the other hand, Meinshausen and Bühlmann (2006) [3] suggested the neighborhood selection method, which is fully nonparametric and is based on the threshold gradient descent algorithm for the estimation of $\boldsymbol{\theta}$. However, the major challenge of all these algorithms is the computational limitation in the inference of realistically high-dimensional systems.

This study aims to implement the Bayesian framework as an alternative to the underlying frequentist and non-parametric approaches. The suggested method is implemented in the form of the combination of GGM and the Gaussian copula [4], resulting in the copula GGM model. The main advantage of this model is that it can overcome the modeling problem of high-dimensional systems by describing the complex GGM model and its multivariate normal density via pieces of marginal and copula terms within the copula GGM representation. CGGM has been already proposed in the study of Dobra and Lenkoski (2011) [5] to describe the functional disability data. In that work, the inference is conducted via the *Reversible Jump Markov Chain Monte Carlo method* (RJMCMC). Here, the novelty of this study lies in adopting this approach to construct the structure of a biological network and infer its model parameters. Furthermore, as the second novelty, this study writes the functional codes via the R programming language so that the codes can be applicable to all biological networks under the steady-state condition and distinct dimensions, i.e., the number of nodes or proteins. Of note, the codes are available upon request. Accordingly, to compare the performance of our algorithm with others, three datasets have been used. Initially, the social survey dataset of Dobra and Lenkoski (2011) [5], which is also applied as a benchmark dataset in comparative analyses, has been implemented [6,7]. Then, an actual biological network, called cell signaling network, is used, and the results are interpreted. Lastly, an ovarian cancer dataset, whose true interactions can be biologically validated from the literature, should be implemented. Finally, our findings are compared with the outputs of Mohammadi (2015) [6], where the inference is performed via the birth-and-death algorithm in place of RJMCMC for the same CGGM.

Hence, in the process of organizing this study, the Gaussian graphical model and copulas are introduced in Section 2. In Section 3, the method of inference is introduced in detail. Then, in Section 4, the suggested methods are applied to different datasets. Lastly, our findings are summarized, and some suggestions for the future works are made in Section 5.

2. Materials and methods

In this part, initially, the general idea of graphical networks, which is one of the common ways to show the relationship between factors in a mathematical model, is explained. Based on the statistical analysis of biological networks, when the number of genes or other kinds of variables is large and their correlation matrix is sparse, the application of a graphical version of the network may boost readers' imagination about the structure of genes or variables. In this representation, the Gaussian copula graphical model is performed as it enables one to partition a high-dimensional joint distribution function as pieces of marginal that are bound by a separate copula term when the data are described by multivariate normal distribution. By means of normality, we can also simplify the correlation between variables due to the property of the conditional independency. Finally, by using the RJMCMC algorithm in the inference of the network, we can benefit from the flexibility of the Bayesian method when the data are limited and the dimension of the network is large.

Hence, in the following parts, the mathematical details of the graphical model, the Gaussian graphical model, the Gaussian copula approach, and the selected Bayesian algorithm are presented in order.

2.1. Graphical model

Let a data matrix $\mathbf{Y}$ with p variables and n samples be presented; herein, an attempt is made to obtain the relationship between Y_i and Y_j for $i \neq j$ given the remaining variables. In this type of networks, which is common in social surveys and biological aspects, each variable is shown by a node in a graph and the conditional dependence between two nodes is presented by an undirected edge. Hereby, if $\mathbf{E}$ denotes the set of available edges under an undirected structure, $(i,j) \in \mathbf{E}$ equals $(j,i) \in \mathbf{E}$, showing that Y_i and Y_j are conditionally dependent ($i,j = 1,2,\ldots,p$ and also $Y_i \perp Y_j | Y_{V \setminus i,j}$ for $V = 1,2,\ldots,p$). This structure is called the pairwise Markov property [1].

2.1.1. Gaussian graphical model

Now, it is assumed here that vector $\mathbf{Y}$ follows a p-dimensional multivariate normal distribution via $N_p\left(0, \boldsymbol{\theta}^{-1}\right)$. Here, $\boldsymbol{\theta}$ is the inverse of the covariance matrix, which is also called the precision matrix. Hence, for n samples, the likelihood function of $\mathbf{Y}$ can be written as follows:

$$\left(\mathbf{Y}^{1:n}|\boldsymbol{\theta}\right) \propto |\boldsymbol{\theta}|^{\frac{n}{2}} \exp\left\{-\frac{1}{2} tr\left(\boldsymbol{\theta}^T \mathbf{U}\right)\right\}, \tag{3}$$

where $|.|$ and $tr(.)$ describe the determinant and the trace of the given expression, respectively, and $\boldsymbol{\theta}^T$ is the transpose of the given matrix as used beforehand. Finally, $\mathbf{U}$ is the trace of the $\mathbf{Y^TY}$ matrix. Thus, a graphical model with $\mathbf{V}$ nodes and $\mathbf{E}$ edges, $(\mathbf{V},\mathbf{E})$, for $N_p\left(0,\boldsymbol{\theta}^{-1}\right)$ is presented where $\mathbf{V} = (1,2,...,p)$ is called the Gaussian Graphical Model (GGM).

2.2. Gaussian copula

If the normality assumption does not hold for the data matrix, the copula can solve the problem by combining data such that their joint distribution is Gaussian with the same covariance matrix [4,1,8]. For binary and ordinal categorical data, a continuous latent variable $\mathbf{Z}$ is introduced [5,6,9,10] by defining some increasing thresholds $\tau_v = (\tau_{v,0}, \tau_{v,1}, \ldots, \tau_{v,\omega_v})$. Therefore:

$$y_v^j = \sum_{l=1}^{\omega_v} l \times 1_{\tau_{v,l-1} < z_v^j \leq \tau_{v,l}}, \tag{4}$$

for $j = 1,2,...,n$. The relationship between Y_{ij} and Z_{ij} satisfies the following constraint.

$$y_{ij} < y_{ik} \rightarrow z_{ij} < z_{ik}, z_{ij} < z_{ik} \rightarrow y_{ij} \leq y_{ik}. \tag{5}$$

Then, by defining the interaction of the correlation matrix in terms of $\boldsymbol{\theta}$ as:

$$\mathbf{Y}_{i,j}\left(\boldsymbol{\theta}\right) = \frac{\left(\theta^{-1}\right)_{i,j}}{\sqrt{\left(\theta^{-1}\right)_{i,i}\left(\theta^{-1}\right)_{j,j}}}, \tag{6}$$

and $\mathbf{Z}_V N_p\left(0,\boldsymbol{\theta}^{-1}\right)$, a one-to-one correspondence with observed data can be obtained as follows:

$$\tilde{Z}_i = Z_i/(\theta_{i,i}^{-1})^{\frac{1}{2}} \text{ and}$$

$$Y_i = F^{-1}\left(\Phi\left(\tilde{Z}_i\right)\right). \tag{7}$$

In Eq. (6), $\theta_{i,i}$ and $\theta_{j,j}$ indicate the diagonal entries of the ith and jth nodes, respectively. Accordingly, $\theta_{i,j}$ is the precision value between the ith and jth nodes. On the other hand, in Eq. (7), F^{-1} and Φ stand for the inverse of the cumulative distribution functions (cdf) and cdf of the normal distribution, respectively. Hence, by standing $\mathbf{C}(u_1,\ldots,u_p|Y)$ as the Gaussian copula with $(p \times p)$-dimensional correlation matrix for the p random sample from the standard uniform distribution, we have:

$$\begin{aligned} p\left(Y_1 < y_1, \ldots, Y_p < y_p\right) \\ = \mathbf{C}(F_1\left(y_1\right), \ldots, F_p\left(y_p\right)|\mathbf{Y}\left(\boldsymbol{\theta}\right)). \end{aligned}$$

This study decomposes the multivariate normal distribution of the states via the Gaussian copula model with the normal marginal distributions. This new probability distribution function is used to calculate the likelihood within the Bayesian framework. In doing so, GGM can be performed under any dimensional systems since the high-dimensional multivariate normal density can be partitioned via the copula term.

2.3. Reversible jump Markov chain Monte Carlo method

The Reversible Jump Markov Chain Monte Carlo method (RJMCMC) is an approach that mostly deals with the Cholesky decomposition to obtain a positive definite precision matrix due to its conjugate advantages in prior distribution for the precision matrix, which is considered as the G-Wishart distribution [11] with a density:

$$p\left(\boldsymbol{\theta}|G\right) = \frac{1}{I_G\left(\delta, \mathbf{D}\right)} \det\left(\boldsymbol{\theta}^{\frac{\delta-2}{2}}\right) \exp\left\{-\frac{1}{2} tr\left(\boldsymbol{\theta}^T \mathbf{D}\right)\right\}. \tag{8}$$

In this expression, G implies the given graphical structure of the data. On the other hand, the G-Wishart prior is a generalized version of the chi-square distribution and the conjugate with the multivariate normal density. The sampling algorithm from the G-Wishart distribution was performed by Lenkoski (2013) [12]. Thus, the posterior distribution, $\boldsymbol{\theta}$, of the given G is presented as the G-Wishart distribution with parameters $\delta + n$ and $\mathbf{D} + \mathbf{U}$. In this expression, $\delta > 2$, $\mathbf{D} = I_p$ is the p-dimensional identity matrix and $\mathbf{U} = \sum_{j=1}^{n} y_j y_j^T$, i.e., the trace of $\mathbf{Y^TY}$, as defined beforehand. In Eq. (8), the normalizing constant $I_G(\delta, \mathbf{D})$ is not always easy to obtain [13]. When G is not a complete graph and is non-decomposable, this constant is calculated by a Monte Carlo method. Accordingly, a double reversible jump algorithm was introduced by Lenkoski (2013) [12] to obtain the normalizing constant of the G-Wishart distribution.

Further, the Cholesky decomposition partitions the matrix into a lower triangle matrix and its transpose in a way that $\boldsymbol{\theta} = \boldsymbol{\varphi}^T\boldsymbol{\varphi}$ denotes a chi-square distribution into the square of the standard normal distribution. Here, $\boldsymbol{\varphi}$ is the upper triangle matrix, in which zero implies no relationship between the two corresponding elements. Finally, under the normality of data, two strictly positive precision parameters $\sigma_p = \sigma_g = 0.1$ and RJMCMC are repeated in the following steps until convergence is achieved [4].

2.3.1. Resampling the latent data

In the first stage of the RJMCMC algorithm, the latent variable $\mathbf{Z}$ is used instead of $\mathbf{Y}$ if $\mathbf{Y}$'s are not normal [5,3]. Here, $\mathbf{Z}$ is an $(n \times p)$-dimensional matrix and, for each column, which is related to each node, we calculate its minimum $\mathbf{L}$ and its maximum $\mathbf{U}$ as the vectors of p elements.

In this step, by using matrix $\boldsymbol{\theta}$ and vectors $\mathbf{L}$ and $\mathbf{U}$, other Z_i's are generated from truncated normal in the L_i and U_i distributions in the following form:

$$Z_i|Z_i\backslash i \sim N\left(\mu_i, \sigma_i^2\right), \tag{9}$$

where:

$$\mu_i = -\sum_{y\in bd(i)} \frac{\theta_{i,y}}{\theta_{i,i}} z_{y,j},$$

for:

$$bd\left(i\right) = \left\{y \in \left(1, \ldots, p\right) : \left(i, j\right) \in \mathbf{E}\right\},$$

when:

$$\mathbf{E} = \{\left(i, j\right)|\theta_{i,y} \neq 0, \quad i \neq y\}, \qquad \sigma_i^2 = \frac{1}{\theta_{i,i}},$$

and:

$$\Sigma_{i,i} = \frac{1}{\theta_{i,i}}.$$

In the second step, these $z_{i,j}$'s will be used.

2.3.2. Resampling the Precision Matrix

In this step, matrix $\boldsymbol{\theta}$ is calculated by using the latent variables from the previous stage, and the Cholesky decomposition of matrix $\boldsymbol{\theta}$ is applied. For non-zero diagonal elements, a Metropolis-Hasting update [2] of $\boldsymbol{\varphi}$ is done by sampling a γ value from a normal distribution truncated below at zero with a mean $\varphi_{i,i}$ and a variance σ_p^2. Then, γ is replaced by the related diagonal elements of $\boldsymbol{\varphi}$ and $\boldsymbol{\varphi}$ and transformed to $\boldsymbol{\varphi}'$ with a probability $\min\{R_p, 1\}$, where:

$$R_p = \frac{\Phi\left(\frac{\varphi_{i,i}}{\sigma_p}\right)}{\frac{\gamma}{\sigma_p}}(\frac{\gamma}{\varphi_{i,i}})^{\delta+n+nb(i)-1}R'_p, \tag{10}$$

denoting that:

$$R'_p = \exp\{-\frac{1}{2}tr(\boldsymbol{\theta}' - \boldsymbol{\theta})^T \left(\mathbf{D} + tr\left(\mathbf{Z}^T\mathbf{Z}\right)\right)\}. \tag{11}$$

In addition, the candidate value $\boldsymbol{\theta}' = \boldsymbol{\varphi}'^{\boldsymbol{T}}\boldsymbol{\varphi}'$ results in $\boldsymbol{\theta} = \boldsymbol{\varphi}^{\boldsymbol{T}}\boldsymbol{\varphi}$.

For non-diagonal elements of $\boldsymbol{\varphi}$, a new γ is sampled from $N(\mu_i, \sigma_g^2)$. In these cases, $\boldsymbol{\varphi}$ is transformed to $\boldsymbol{\varphi}'$ with a probability $\min\{R'_p, 1\}$.

2.3.3. Resampling the graph

In the third step, only one element of the Cholesky matrix $\varphi_{i,j}$, which is obtained in the previous step, is selected randomly. If there is no edge between Y_i and Y_j, it will be changed by a value from $N(\varphi_{i,j}, \sigma_g^2)$ in $\boldsymbol{\varphi}$ with a probability $\min\{R_p, 1\}$, where:

$$\begin{aligned} R_p =& \sigma_g\sqrt{2\pi}\varphi_{i,i}\frac{I_G\left(\delta, \mathbf{D}\right)}{I_{G'}\left(\delta, \mathbf{D}\right)} \\ &\times \exp\{-\frac{1}{2}tr(\left(\boldsymbol{\theta}' - \boldsymbol{\theta}\right)^T\left(\mathbf{D} + tr\left(\mathbf{Z}^T\mathbf{Z}\right)\right)) \\ &+ \frac{\left(\varphi'_{i,j} - \varphi_{i,j}\right)^2}{2\sigma_g^2}\}, \end{aligned} \tag{12}$$

Here, $\boldsymbol{\varphi}'$ stands for the proposal φ and G' is a graph in which all elements coincide with G except $G_{i,j}$, which is supposed to be the edge between related nodes. If there is an edge between Y_i and Y_j, it will be replaced by zero in $\boldsymbol{\varphi}$ with a probability $\min\{R'_p, 1\}$, where:

$$\begin{aligned} R'_p =& (\sigma_g\sqrt{2\pi}\varphi_{i,i})^{-1}\frac{I_G\left(\delta, \mathbf{D}\right)}{I_{G'}\left(\delta, \mathbf{D}\right)} \\ &\times \exp\{-\frac{1}{2}tr(\left(\boldsymbol{\theta}' - \boldsymbol{\theta}\right)^T\left(\mathbf{D} + tr\left(\mathbf{Z}^T\mathbf{Z}\right)\right)) \\ &+ \frac{\left(\varphi'_{i,j} - \varphi_{i,j}\right)^2}{2\sigma_g^2}\}. \end{aligned} \tag{13}$$

In this stage, since the dimensionality of the parameter space changes by a one-unit increase or one-unit decrease, the reversible jump Markov chain methodology is performed. Then, this graph in the first step of the algorithm is used, and the process continues until convergence is reached.

3. Applications

In order to evaluate the RJMCMC method in terms of accuracy and assess its performance for the first time in real biological systems, the code of the R programming language is originally generated with a function for each stage by deriving it from the sample precision matrix as the initial matrix after 200,000 iterations for three datasets. The first case of data is the Rochdale dataset and is used to conduct comparative analysis of different inference methods [14]. The second data are the real data applied to construct the cell signaling pathway, and the third data represent the combination of endometrial and ovarian carcinoma [15-17].

3.1. Rochdale data

The Rochdale data are binary (yes/no) data collected from 665 samples to assess the relationship among eight factors affecting the economic activities specific to women. These eight variables are named as follows: a (wife economically active), b (age of wife > 38), c (husband unemployed), d (child ≤ 4), e (wife's education at the high-school level or beyond), f (husband's education at the high-school level or beyond), g (Asian origin), and h (other household member working). In the case of analyses, it is claimed that there are at least two-way interaction effects whose minimal sufficient statistics are the following pair of variables: {fg, ef, dh, dg, cg, cf, ce, bh, be, bd, ag, ae, ad, ac}. Then, by including variable h, which changes as very fast and very slow as the two new random variables, the data are observed in Table 1. More details about this dataset can be also found in the studies of Whittaker (1990) [1] and Dobra and Lenkoski (2011) [5]. Hereby, from the inference of this dataset via RJMCMC, 15 edges are found: (a; c), (a; d), (a; e), (a; g), (b; d), (b; e), (b; h), (c; e), (c; f), (c; g), (d; g), (d; h), (e; f), (e; g), and (f; g). The 14 edges that represent the validated links from the study of Whittaker (1990) [1] are exactly the same as what have been obtained from the RJMCMC codes except (e; g) edge.

As a result, in these analyses, the data are initially transformed to Gaussian. Then, by applying our RJMCMC codes, the latent variables $\mathbf{Z}$ are resampled based on entries of the initial matrix, which is considered as the sample covariance matrix in our study. Next, the precision matrix is resampled by taking the latent data produced in the first step, and the graph is resampled by only one element, which is selected randomly from the Cholesky decomposition of the precision matrix in the second step. This process continues until convergence is reached. In this example, this process is iterated up to 1,000,000 times, of which the first 200,000 runs are supposed to be in the burn-in period. The adjacency matrix is obtained from the mean of the estimated entries of the precision matrix and, thereby, represents the estimated structure of the links, as shown in Table 2.

In this matrix, the entry 1 indicates the link between the pairs of variables, and the entry 0 implies no link between them. Further, Figure 1 presents some examples from the estimated density of selected pairs in the precision matrix after the burn-in period with 200,000 MCMC runs. In these plots, it is observed that each estimated density is unimodal, and the model parameters reach convergence.

Table 2. The adjacency matrix of the Rochdale data estimated by 1,000,000 RJMCMC iterations, where the first 200,000 runs take place in the burn-in period.

	a	***b***	***c***	***d***	***e***	***f***	***g***	***h***
a	0	0	1	1	1	0	1	0
b	0	0	0	1	1	0	0	1
c	1	0	0	0	1	1	1	0
d	1	1	0	0	0	0	1	1
e	1	1	1	0	0	1	1	0
f	0	0	1	0	1	0	1	0
g	1	0	1	1	1	1	0	0
h	0	1	0	1	0	0	0	0

Table 1. The lexicographical ordered Rochdale data [7].

5	0	2	1	5	1	0	0	4	1	0	0	6	0	2	0
8	0	11	0	13	0	1	0	3	0	1	0	26	0	1	0
5	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0
4	0	8	2	6	0	1	0	1	0	1	0	0	0	1	0
17	10	1	1	16	7	0	0	0	2	0	0	10	6	0	0
1	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0
4	7	3	1	1	1	2	0	1	0	0	0	1	0	0	0
0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
18	3	2	0	23	4	0	0	22	2	0	0	57	3	0	0
5	1	0	0	11	0	1	0	11	0	0	0	29	2	1	1
3	0	0	0	4	0	0	0	1	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	25	0	1	37	26	0	0	15	10	0	0	43	22	0	0
0	0	0	0	2	0	0	0	0	0	0	0	3	0	0	0
2	4	0	0	2	1	0	0	0	1	0	0	2	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

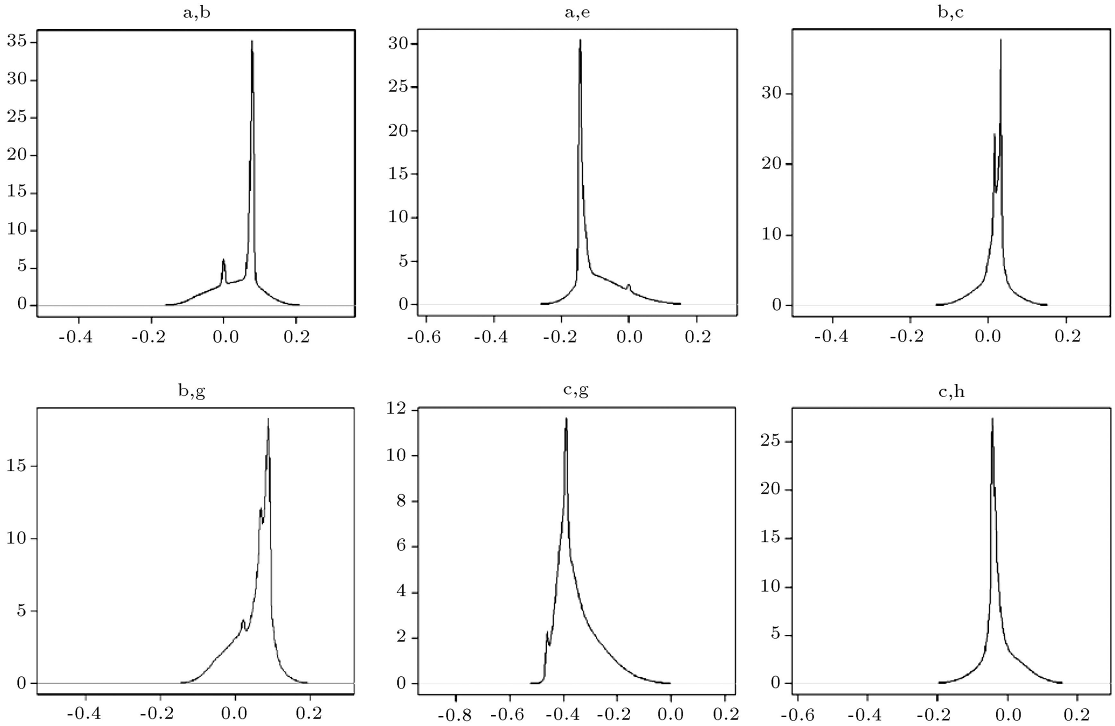

Figure 1. The density of some of the estimated entries in the precision matrix for the Rochdale data after the 1,000,000 MCMC iterations, where the first 200,000 runs take place in the burn-in period.

Table 3. The adjacency matrix of the cell signaling pathway data estimated by the 1,000,000 iterations of the birth-and-death algorithm [23], where the first 200,000 runs take place in the burn-in period.

	a	***b***	***c***	***d***	***e***	***f***	***g***	***h***
a	0	1	1	1	1	0	1	1
b	1	0	0	1	1	1	1	1
c	1	0	0	0	1	1	0	1
d	1	1	0	0	1	0	0	0
e	1	1	1	1	0	1	1	0
f	0	1	1	0	1	0	1	1
g	1	1	0	0	1	1	0	0
h	1	1	1	0	0	1	0	0

Furthermore, in order to check the accuracy of our estimates and codes, the F_1-score, also known as F-measure, is computed as shown below, and the obtained results are compared with the estimated parameters by the birth-and-death method. This method has been suggested as an alternative to RJMCMC in the literature, and its R coding has been developed under the BDgraph package [9]. The estimated adjacency matrix is presented by the birth-and-death method, as shown in Table 3.

$$F_1-\text{ score} = \frac{2TP}{2TP+FP+FN}, \tag{14}$$

where TP, FP, and FN represent the values of the True Positive, False Positive, and the False Negative, respectively. The F_1-score is always between 0 and 1, where 1 is its perfection level. Hence, by taking the same number of the MCMC iterations from both methods, F_1-score = 0.96 is obtained where $TP = 14$, $FP = 1$, and $FN = 0$ in the RJMCMC iterations. Furthermore, the same measures are found as in F_1-score = 0.69, where $TP = 11$, $FP = 7$, and $FN = 3$ by using the birth-and-death algorithm for the same dataset. Therefore, it can be concluded that the RJMCMC method is successful in the inference of the copula GGM, and the estimated links found by our open-source R code validate the true links about the data.

3.2. Cell signaling data

For the second application, a real cell signaling dataset that contains 11 phosphoproteins and phospholipids is used under various experimental conditions in human primary naive CD4+T cells that are measured on 11672 red blood cells [18]. In the inference of this system, our RJMCMC codes and the birth-and-death algorithm are run for 10,000 iterations. Then, the estimated systems from both approaches are compared with respect to the F_1-score based on the true structure of the system in the study of Sachs et al. (2005) [18]. In this assessment, the directed true network is converted into the undirected one since the copula GGM approach is designed for the undirected graphs. Thereby, from the findings of RJMCMC with 10,000 iterations based

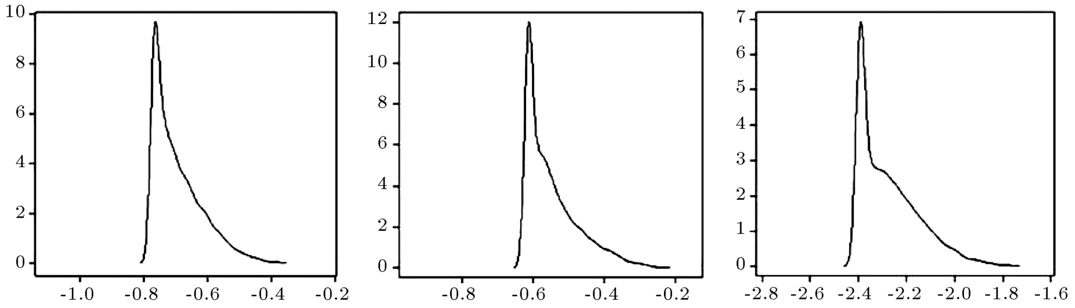

Figure 2. The density of some of the estimated entries in the precision matrix for the cell-signaling data after the 10,000 MCMC iterations, where the first 2,000 runs take place in the burn-in period. PIP2-PIP3, Plcy-PIP2, and Raf-Mek in lexicographical order.

Table 4. The adjacency matrix of the ovarian cancer data estimated by the 10,000 RJMCMC iterations, where the first 2,000 runs take place in the burn-in period.

	MAP2K1	**MK01**	**CEBPB**	**CTNNB1**	**TFAM**	**TP53**	**PDIA3**	**IMP3**	**ERBB2**	**CHD4**	**MBD3**
MAP2K1	1	1	1	1	1	1	1	1	1	1	1
MK01	1	1	1	1	1	1	1	1	1	1	1
CEBPB	1	1	1	1	1	1	1	1	1	1	1
CTNNB1	1	1	1	1	1	1	1	1	1	1	1
TFAM	1	1	1	1	1	1	1	1	1	1	1
TP53	1	1	1	1	1	1	1	1	1	1	1
PDIA3	1	1	1	1	1	1	1	1	1	1	1
IMP3	1	1	1	1	1	1	1	1	1	1	1
ERBB2	1	1	1	1	1	1	1	1	1	1	1
CHD4	1	1	1	1	1	1	1	1	1	1	1
MBD3	1	1	1	1	1	1	1	1	1	1	1

on the 2,000 runs taking place the burn-in period, we obtain F_1-score as F_1-score $= 0.63$, where $TP = 8$, $FP = 10$, and $FN = 11$. Eight nodes, found by our codes compatible with the true network, include Raf-Mek, Erk-Mek, PIP2-PLCY-, PIP3-PIP2, Erk-Akt, Erk-PKA, Raf-PKC, and PKC-Mek. The density plots of the estimated links after the burn-in are also shown in Figure 2. On the other hand, the values of the birth-and-death algorithm are computed as F_1-score $=$ 0.50, while $TP = 7$, $FP = 1$, and $FN = 13$. Based on these outputs, it is seen that RJMCMC enjoys better accuracy than the birth-and-death method for this dataset.

3.3. Ovarian cancer data

In this analysis, we specifically deal with gynecologic cancer including the ovarian, cervix, and endometrial cancers. This type of cancer is the second most prevalent cancer in women in the world after breast cancer. In our study, we initially search the biological literature and detect 11 core genes, which are active in gynecologic cancer [15-17]. These genes are named as MPK2K1, MK01, CEBPB, CTNNB1, TFAM, TP53, PDIA3, IMP3, ERBB2, CHD4, and MBD3. Then, based on the ArrayExpress database, an Affymetix dataset is considered and collected under the ovarian cancer, and the observations belonging to the underlying 11 genes are selected.

In the data, each gene has 14 samples and the true network composed of these genes is complete, i.e., its adjacency matrix has the value of one in all entries. The estimated adjacency matrix from RJMCMC and BDMCMC is presented in Tables 4 and 5, respectively. In the estimation, similar to previous analyses, 10,000 MCMC iterations are conducted, and the first 2,000 runs are discarded in the burn-in period. From the outcomes, we calculate F_1-score $= 1$ for RJMCMC and F_1-score $= 0.79$ for BDMCMC. Thereby, as observed from other analyses, the findings show that RJMCMC outperforms BDMCMC with higher accuracy.

4. Results and discussion

This study extended the idea of the Copula GGM (CGGM) model in the description of biological networks since GGM is one of the successful probabilistic

Table 5. The adjacency matrix of the ovarian cancer data estimated by the 10,000 BDMCMC iterations, where the first 2,000 runs take place in the burn-in period.

	MAP2K1	**MK01**	**CEBPB**	**CTNNB1**	**TFAM**	**TP53**	**PDIA3**	**IMP3**	**ERBB2**	**CHD4**	**MBD3**
MAP2K1	1	1	1	1	1	0	0	1	0	0	1
MK01	1	1	0	0	0	0	1	1	1	1	1
CEBPB	1	0	1	0	0	1	1	1	1	1	1
CTNNB1	1	0	0	1	1	1	1	1	1	1	1
TFAM	1	0	0	1	1	0	1	1	1	1	0
TP53	0	0	1	0	0	1	0	0	0	1	0
PDIA3	0	1	1	1	1	0	1	0	1	1	0
IMP3	1	1	1	1	1	0	0	1	1	1	1
ERBB2	0	1	1	1	1	0	1	1	1	1	0
CHD4	0	1	1	1	1	1	1	1	1	1	1
MBD3	1	1	1	1	0	0	0	1	0	1	1

modeling approaches for explaining the steady-state behavior of the biological systems, and the copulas enable us to separate any high-dimensional joint function as marginals. Hereby, CGGM can be also used instead of sole GGM, especially for high-dimensional systems as it can partition the high-dimensional joint density into small parts, resulting in simplicity of estimating the model parameters. In the inference of the underlying model, the Reversible Jump Markov Chain Monte Carlo (RJMCMC) approach as another alternative to the birth-and-death (BDMCMC) algorithm for CGGM is implemented [9,6]. In the computation, the RJMCMC approach has been adopted to estimate the biological networks by writing all codes as the open-source R codes and making all necessary calibrations in the calculations while converting the implementation of these techniques into the system's biology. In the application, the bench-mark Rochdale dataset is used and applied to compare different modeling and inference approaches in the system's biology to validate the performance of the current calculation. Then, we have also implemented it for the inference of the cell signaling pathway and the ovarian cancer data. According to the comparative analyses via the BDMCMC algorithm [9,6], we have observed that RJMCMC gives more accurate results in all analyses.

As the extension of this study, the split-merge method [19] and the Gibbs sampling [8] are used as the new alternatives to RJMCMC in selecting the dimension for the precision matrix. Because even though these listed methods have been also suggested in place of RJMCMC theoretically, their application to real-life and high-dimensional network problems has not been performed yet. Further, although the accuracy of RJMCMC is significantly high, its calculation via R can be computationally demanding. Hereby, any improvement in the selection procedure of the precision matrix can be deemed useful to deal with the existing challenge during the computational time. The construction of complex systems via CGGM with RJMCMC or its new alternates can help us describe the actual biological activations better and identify any malfunctions in the systems that cause illnesses.

Furthermore, RJMCMC has been performed for Time Series CGGM (TSCGGM) [1] so that the measurement based on different time-course data [20,21] can be applied to estimate the biological networks. In this model, we are interested in estimating two matrices: the precision matrix and the autoregressive coefficient matrix. The last one shows time dependency between variables in the vector autoregression VAR(1) modeling, which can be extended to VAR(p) by our recommended method. In the study of Abegaz and Wit (2013) [22], this calculation is done via the penalized likelihood approach to the state space model. In addition, we think that this model can be extended by using the vine rather than the Gaussian copulas [23,24]. In doing so, the strict normality assumption of the measurements can be relaxed by accepting other non-normal distributions. Because, in some cases, the normality assumption or the normalization of the data can dissemble the structure, particularly in dealing with a sample of lower size, which is a common challenge in biological datasets. This question is one of the major interests in computational and systematic biology whose applications can be seen from various biological sciences, ranging from genetics to pharmacology.

Furthermore, another powerful alternate to RJMCMC in inference of the relationship between variables can be a Multivariate Adaptive Regression Spline (MARS), which is, in brief, a non-parametric regression technique and can be seen as an extension of linear models that automatically describe nonlinearities and interactions between variables. From previous studies, it has been shown that RJMCMC can be used in certain parts of MARS [25], and we think that this idea can be adapted for the construction of biological networks, too. From the recent literature about the MARS model, it has been found that the conic version of MARS, called CMARS [26,27], and its robustification, called robust CMARS or shortly RCMARS [28,29], are two other extended versions of MARS to improve the accuracy of the nonlinear and correlated data. Among these alternatives, the CMARS model has been implemented to construct biological networks; based on the results, it has been observed that the accuracy of the model can increase in comparison to the MARS model [30]. Moreover, the CMARS model is also extended by different bootstrapping regression methods to obtain the empirical distributions of the parameters of CMARS [31]. In addition, Yerlikaya-Özkurt et al. (2016) [32] and Taylan et al. (2014) [33] developed a new scheme to minimize the impact of outliers on regression estimators of CMARS. On the other hand, the RCMARS model was applied to build a precipitation model of the continental central Anatolia region of Turkey [34]. Then, it is also performed for the presentation of the regulatory networks [35]. However, the performance of this model has not been compared yet with CGGM in terms of accuracy and evaluation of the computational demand of different biological systems' models. It is supposed here that such a comparative study can be useful for detecting the most accurate model that, particularly, fit with protein-protein interaction data.

Furthermore, all these models from CGGM, MARS, CMARS, and RCMARS, which can describe the steady-state activation of the biological systems, can be extended by considering the randomness in the nature of the systems. Under this condition, the stochastic models can be beneficial. Among alternatives, the diffusion model [36], the discretized version of the diffusion model [37-40], and the Stochastic Hybrid Systems (SHS) [41] are implemented in modeling biological networks. In these models, SHS is further extended by adding jumps to describe the abrupt changes in the data [42,43], whereas the application of this model to biological networks and the Bayesian inference of this jump model have not been studied yet. Hereby, the application of this approach to signal transaction data has been considered, and an attempt has been made to adapt RJMCMC for this model. In the end, such a novelty in SHS can open new avenues for the representation of the biological systems under stochastic models.

Acknowledgements

The authors would like to thank the TÜBİTAK (The Scientific and Technological Research Council of Turkey) grant (Project no: 114E636) for their support and would like to thank Prof. Dr. Alexander R. de Leon for his helpful discussion and insights. Moreover, the authors would like to thank anonymous referees and the editors of the journal for their constructive comments and recommendations, which improved the readability and quality of the paper.

References

1. Whittaker, J. *Graphical Models in Applied Multivariate Statistics*, John Wiley and Sons, New York (1990).

2. Friedman, J., Hastie, T., and Tibshirani, R. "Sparse inverse covariance estimation with the graphical lasso", *Biostatistics*, **9**, pp. 432-441 (2008).

3. Meinshausen, N. and Bühlmann, P. "High dimensional graphs and variable selection with the lasso", *The Annals of Statistics*, **34**, pp. 1436-1462 (2006).

4. Green, P.J. "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination", *Biometrika*, **82**(4), pp. 711-732 (1995).

5. Dobra, A. and Lenkoski, A. "Copula Gaussian graphical models and their application to modeling functional disability data", *Annals of Applied Statistics*, **5**, pp. 969-993 (2011).

6. Mohammadi, A. "Bayesian model determination in complex systems", PhD Thesis. University of Groningen, Netherland (2015).

7. Richardson, S. and Green, P.J. "Bayesian analysis of mixtures with an unknown number of components", *Journal of Royal Statistical Society B*, **59**, pp. 731-792 (1997).

8. Walker, S. "A Gibbs sampling alternative to reversible jump MCMC", Report no.: IMS-EJS-EJS_2009_383, pp. 1-3 (2009).

9. Mohammadi, A. and Wit, E.C. "Bayesian structure learning in sparse Gaussian graphical models", *Bayesian Analysis*, **10**, pp. 109-138 (2015).

10. Skrondal, A. and Rabe-Hesketch, S. "Structural equation modeling: Categorical variables", *Entry for the Encyclopedia of Statistics in Behavioral Science*, Wiley, pp. 1-8 (2005).

11. Wang, H. and Zhengzi, S. "Efficient Gaussian graphical model determination under G-Wishart prior distributions", *Electronic Journal of Statistics*, **6**, pp. 168-198 (2012).

12. Lenkoski, A. "A direct sampler for G-Wishart variates", *Statistics*, **2**, pp. 119-128 (2013).

13. Atay-Kayis, A. "A Monte Carlo method for computing the marginal likelihood in non-decomposable Gaussian graphical models", *Biometrika*, **92**(2), pp. 317-335 (2005).

14. Ai, J. "Reversible-jump MCMC methods in Bayesian statistics", MSc Thesis, The University of Leeds, United Kingdom (2012).

15. Hu, Z., Zhu, D. etc. "Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomolgy-mediated integration mechanism", *Nature Genetics*, **47**(2), pp. 158-163 (2015).

16. The Cancer Genome Atlas Research Network. "Integrated genomic analyses of ovarian Carcinoma", *Nature*, **474**, pp. 609-615 (2011).

17. Levine, D.A. and The Cancer Genome Atlas Research Network "Integrated genomic characterization of endometrial carcinoma", *Nature*, **497**, pp. 67-73 (2013).

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

19. Trivedi, P.K. and Zimmer, D.M. "Copula modeling: An introduction for practitioners", *Foundations and Trends R in Econometrics*, **1**, pp. 1-111 (2005).

20. Weber, G.W., Defterli, Ö., Alparslan Gök, S.Z., and Kropat, E. "Modeling, inference and optimization of regulatory networks based on time series data", *European Journal of Operational Research*, **211**(1), pp. 1-14 (2011).

21. Sima, C., Hua, J., and Jung, S. "Inference of gene regulatory networks using time-series data", *A Survey, Current Genomics*, **10**(6), pp. 416-429 (2009).

22. Abegaz, F. and Wit, E. "Sparse time series chain graphical models for reconstructing genetic networks", *Biostatistics*, **14**(3), pp. 586-599 (2013).

23. Wawrzyniak, M.M. "Dependence concepts", MSc Thesis, Delft University of Technology, Netherland (2006).

24. Brechmann, E.C. and Schepsmeier, U. "Modeling dependence with C- and D-vine copulas: The R package CDVine", *Journal of Statistical Software*, **52**(3), pp. 1-27 (2013).

25. Holmes, C.C. and Denison, D.G.T. "Classification with Bayesian MARS", *Machine Learning*, **50**, pp. 159-173 (2003).

26. Yerlikaya-Özkurt, F., *CMARS: A New Contribution to Nonparametric Regression with MARS*, Lap Lambert Academic Publishing (2011).

27. Weber, G.W., Batmaz, İ., Köksal, G., Taylan, P., and Yerlikaya-Özkurt, F. "CMARS: a new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization", *Inverse Problems in Science and Engineering*, **20**(3), pp. 371-400 (2012).

28. Özmen, A. *Robust Optimization of Spline Models and Complex Regulatory Networks*, Springer International Publishing, Switzerland (2016).

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

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

31. Yazıcı, C., Yerlikaya-Özkurt, F., and Batmaz, İ. "A computational approach to nonparametric regression: bootstrapping CMARS method", *Machine Learning*, **101**(1-3), pp. 211-230 (2015).

32. Yerlikaya-Özkurt, F., AŞkan, A., and Weber, G.W. "A hybrid computational method based on convex optimization for outlier problems: Application to earthquake ground motion prediction", *Informatica*, **27**(4), pp. 893-910 (2016).

33. Taylan, P., Yerlikaya-Özkurt, F., and Weber, G.W. "An approach to the mean shift outlier model by Tikhonov regularization and conic programming", *Intelligent Data Analysis*, **18**(1), pp. 79-94 (2014).

34. Özmen, A., Kropat, E., and Weber, G.W. "Robust optimization in spline regression models for multi-model regulatory networks under polyhedral uncertainty", *Optimization*, **66**(12), pp. 2135-2155 (2017).

35. Özmen, A., Batmaz, İ., and Weber, G.W. "Precipitation modeling by polyhedral RCMARS and comparison with MARS and CMARS", *Environmental Modeling and Assessment*, **19**(5), pp. 425-435 (2014).

36. Bower, J. and Bolouri, H., *Computational Modeling of Genetic and Biochemical Networks*, MIT Press, London (2001).

37. Golightly, A. and Wilkinson, D.J. "Bayesian inference for stochastic kinetic models using diffusion approximation", *Biometrics*, **61**(3), pp. 781-788 (2005).

38. Golightly, A. and Wilkinson, D.J. "Bayesian sequential inference for stochastic kinetic biochemical network models", *Journal of Computational Biology*, **13**(3), pp. 838-851 (2006).

39. Purutçuoğlu, V. "Inference of stochastic MAPK pathway by modified diffusion bridge method", *Central European Journal of Operational Research*, **21**(2), pp. 415-429 (2013).

40. Purutçuoğlu, V. and Wit, E. "Bayesian inference for the MAPK/ERK pathway by considering the dependency of the kinetic parameters", *Bayesian Analysis*, **3**(4), pp. 851-886 (2008).

41. Li, X., Omotere, O., Qian, L., and Dougherty, E.R. "Review of stochastic hybrid systems with applications in biological systems modeling and analysis", *EURASIP Journal on Bioinformatics and Systems Biology*, **8**, pp. 1-12 (2017).

42. Savku, E. "Advance in optimal control of markov regime-switching models with applications in finance and economics", PhD Thesis. Middle East Technical University, Turkey (2017).

43. Savku, E., Azevedo, N., and Weber, G.W. "Optimal control of stochastic hybrid models in the framework of regime switches", *Modeling, Dynamics, Optimization and Bioeconomics*, II. Editors: Pinto, A. and Zilberman, D., Springer, pp. 371-387 (2017).

Biographies

Hajar Farnoudkia (1986) is a PhD student of Statistics at Middle East Technical University (METU) since 2015. She received her MSc and BSc degrees from the University of Tabriz. Her research interests include Gaussian graphical models and their application to biological datasets. She is currently working on time series chain graphical models and C-vine and D-vine copulas.

Vilda Purutçuoğlu is a Professor at the Department of Statistics at Middle East Technical University (METU) and also affiliated faculty in the Informatics Institute, Institute of Applied Mathematics and Department of Biomedical Engineering at METU. She has completed her BSc and MSc in Statistics and holds minor degree in Economics. She received her doctorate from the Lancaster University. Dr. Purutçuoğlu's current research interests lie in the field of bioinformatics, systems biology, and biostatistics. She has a research group who has been working on deterministic and stochastic modeling of biological networks and their inferences via Bayesian and frequentist theories.