Journal of Bioinformatics and Computational Biology

Constructing gene regulatory networks from microarray data using non-Gaussian pair-copula graphical models

O. Chatrabgoun

Department of Statistics, Malayer University, Malayer, Iran.
o.chatrabgoun@malayeru.ac.ir

A. Hosseinian-far

Department of Business Systems & Operations,
University of Northampton, NN2 7AL, UK.
amin.hosseinian-far@northampton.ac.uk

A. Daneshkhah*

Faculty of Engineering, Environment & Computing,
Coventry University, CV1 2JH, UK.
ac5916@coventry.ac.uk

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Many biological and biomedical research areas such as drug design require analysing the Gene Regulatory Networks (GRNs) to provide clear insight and understanding of the cellular processes in live cells. Under normality assumption for the genes, GRNs can be constructed by assessing the nonzero elements of the inverse covariance matrix. Nevertheless, such techniques are unable to deal with non-normality, multi-modality and heavy tailedness that are commonly seen in current massive genetic data. To relax this limitative constraint, one can apply copula function which is a multivariate cumulative distribution function with uniform marginal distribution. However, since the dependency structures of different pairs of genes in a multivariate problem are very different, the regular multivariate copula will not allow for the construction of an appropriate model. The solution to this problem is using Pair-Copula Constructions (PCCs) which is decomposition of a multivariate density into a cascade of bivariate copula, and therefore, assign different bivariate copula function for each local term. In fact, in this paper, we have constructed inverse covariance matrix based on the use of Pair-Copula Constructions when the normality assumption can be moderately or severely violated for capturing a wide range of distributional features and complex dependency structure. To learn the non-Gaussian model for the considered GRN with non-Gaussian genomic data, we apply modified version of copula-based PC algorithm in which normality assumption of marginal densities is dropped. This paper also considers the Dynamic Time Warping (DTW) algorithm to determine the existence of a time delay relation between two genes. Breast cancer is one of the most common diseases in the world where GRN analysis of

*Corresponding author

its subtypes is considerably important; Since by revealing the differences in the GRNs of these subtypes, new therapeutic and drugs can be found. The findings of our research are used to construct GRNs with high performance, for various subtypes of breast cancer rather than simply using previous models.

Keywords: Gene regulatory networks; Gaussian graphical models; Dynamic time warping algorithm; Modified PC algorithm; Pair-copula constructions.

1. Introduction

The past few decades have witnessed numerous developments of GRNs using time series gene expression data which is measured by microarray technology [1,2]. The mechanism of genetic regulation and the interactions of genes are crucial in a wide range of biological and biomedical research. If there is a problem in this mechanism, it can lead to a disease as discussed in [5]. One of the benefits of genetic network modelling is to obtain hypotheses for possible relationships between different genes and the role of each gene followed by the practical confirmation of these assumptions [3]. In addition, these networks provide the possibility of comparing the expression patterns of different genes, and so the unknown genes can be guessed [3].

Since GRNs play an important role in cell processes, numerous methods have been proposed for simulating/modelling GRNs. Among the available techniques linear models, neural networks, (stochastic) differential equations, and the Bayesian networks are more well-known [6,7]. Generally, the existing methods can be classified into three categories. The first category is focused on using association rules to find the regulatory relations between genes [8-10]. Methods belonging to the second category use soft computing approaches to predict GRNs [11,12]. The last category focused on using the probabilistic models, particularly Bayesian network (BN) to predict GRNs. Friedman et al. [13] were the first to use BNs to study GRNs by analysing using bread gene expression data. Moreover, dynamic Bayesian networks (DBNs) are used to construct GRNs (See [10,14,15]).

Under normality assumption for the genes, GRNs can be constructed by assessing the nonzero elements of the inverse covariance matrix (precision matrix) by considering the conditional independencies between genes. In such a model gene i and gene j are conditionally independent if and only if the (i, j) entry in the inverse covariance matrix equals zero. Therefore, much effort has been made to achieve inverse covariance matrix through the use of L_1 (Lasso) regularization such that it makes sense to impose an L_1 penalty to increase its sparsity. However, constructing inverse covariance matrix using Lasso are often confined to Gaussian models. This Lasso technique may not reflect the true connectivity between genes when the normality assumption is not accurate. Nevertheless, such techniques are unable to deal with non-normality, multi-modality and heavy tailedness that are commonly seen in current massive genetic data. In fact, the resulting estimates can be inaccurate when the normality assumption is moderately or severely violated, making the techniques unsuitable for dealing with genetic data [16]. To relax this limitative constraint, one can apply copula function which is a multivariate cumulative distribution function

and marginal probability distribution of each variable is uniform. However, since the dependency structures of different pairs of genes in a multivariate problem are very different, the regular multivariate copula will not allow for the construction of an appropriate model. The solution to this problem is using Pair-Copula Constructions (PCCs) which is decomposition of a multivariate density into a cascade of bivariate copula, and therefore, assign different bivariate copula function for each local term which here we have applied the technique of Bauer and Czado (2016). In fact, in this paper, we have constructed precision matrix based on the use of Pair-Copula Constructions (Lasso technique based on the copula function) when the normality assumption can be moderately or severely violated.

In this regard, Elidan [18] was the first that introduced an innovative by copula function, a marriage between copula functions and graphical models. The combination of copula with graphical model constructs multivariate distribution with univariate marginals and a copula function C that links the marginals. However, the regular copula functions such as Gaussian copula may not be able to accurately depict multi-modal joint distributions in the genomic data [16]. In addition, the non-Gaussian probabilistic graphical model is subject to selection of a copula function for each local term.

In this paper, we propose a novel methodology to construct GRNs based on the so-called Pair-Copula Constructions (PCCs) which are merged with graphical models. In fact, PCC is a more flexible multivariate copula which has recently developed for modelling multivariate dependency [19,20,21]. This modelling structure is based on decomposition of a multivariate density into a cascade of bivariate copula. The only restriction of PCC model is the challenge of selecting the best model structure [22,23,24]. This can be fixed by capturing conditional independence in the graphical model. PCCs create a novel class of multivariate statistical models, that combine the distributional flexibility of PCCs with the parsimony of conditional independence models associated with graphical models, which could be used to construct GRNs. The flexibility of these PCCs allows for capturing a wide range of distributional features such as heavy-tailedness, tail dependence, and non-linear asymmetric dependence in genomic data to construct GRNs. Further details about the combination of PCC with graphical models are provided in [22,23,24]. In fact, in this paper, we have constructed precision matrix based on the use of Pair-Copula Constructions (Lasso technique based on the copula function) when the normality assumption can be moderately or severely violated.

The majority of the existing methods for GRNs construction ignore the time delay regulatory relation [25] between two genes. Hence, in addition to the PCCBNs as a novel method for GRNs construction, we apply Dynamic Time Warping (DTW) that takes the time delay regulatory relation into consideration for GRNs prediction. Therefore, we use DTW, to construct GRNs from microarray datasets. In the other words, the proposed method uses the DTW algorithm to determine the existence of a time delay relation between two genes.

Breast cancer is one of the most common diseases for which many researchers

have been tirelessly working to discover new drugs and treatments. This cancer has several subtypes and each subtype has different therapeutic approaches based on the GRNs. Therefore, analysis of a GRN subtype can be beneficial. Given the mentioned properties of genetic data, we use the method outlined in this paper to analyse different subtypes of breast cancer's GRN.

The paper is organised as follows. In Section 2, we present the PCCs associated with the non-Gaussian BNs of multivariate genomic data. In Section 3, we study some concepts about the breast cancer GRNs and its different subtypes. Moreover, we study the dynamic DTW algorithm to determine the existence of a time delay relation between two genes. The rest of Section 3 include learning Bayesian networks for the breast cancer GRNs using modified version of PC algorithm for non-Gaussian data. In the sequel, we demonstrate breast cancer GRNs construction using described PCCBNs based on the model selection techniques and parameter estimation. A simulation study from constructed GRNs is illustrated in Section 4, and finally we conclude the paper in Section 5.

2. Pair-copula construction for non-Gaussian Bayesian networks

The genes within GRN can be shown as nodes. An interaction between two considered genes can be shown as a directional edge (activation, inhibition). This directional edge actually represents the causal relationship between the two genes. When a specific gene is expressed within the GRNs, due to the causal relationship to the other gene, it can lead to a range activities within the organisms [28]. Therefore, one can use a BN, which is a probabilistic Directed Acyclic Graph (DAG) with Markov property, to model and predict a GRN.

A BN is certainly the most common and applicable probabilistic graphical model to construct and predict GRNs. It represents a set of genes as our random variables and their conditional dependencies via a DAG with Markov property. The preliminary notations of the BNs and their detailed theory with some applications are presented in [29].

The decomposition of a multivariate distribution can be efficiently implemented benefiting from the conditional independencies offered by a DAG. In a GRN, suppose that gene i is expressed by time series $\{X_i : i = 1, 2, \dots d, \quad d \in \mathbb{N}\}$. Then, the density function $f(.)$ of d genes expression, $(X_1, \dots, X_d)$, can be decomposed as a product of d conditional density functions as:

$$f(x_1, \dots, x_d) = \prod_{i=1}^{d} f(x_i | pa(x_i)), \tag{1}$$

where $pa(x_i)$ represents the parent set of gene i that is expressed by gene expression x_i. The density decomposition given in (1) illustrates that once the value of $pa(x_i)$ is learned, knowing the value of the other preceding variables is redundant.

Using copula function, a marriage between copula functions and BNs, conventional BNs models can be extended to a more flexible non-Gaussian BNs. In order

to define copula function, we suppose that gene i is expressed by time series X_i, $i = 1, 2, \ldots, d, d \in \mathbb{N}$ with cumulative distribution function (cdf) F_i. A d-variate copula for d considered genes is an cdf on $[0,1]^d$ such that all univariate marginals are uniform on the interval $[0,1]$. In fact, every cdf F for d considered genes on $\mathbb{R}^d$ with marginals F_i can be written as:

$$F(\mathbf{x}) = C(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)),$$

for $\mathbf{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$, and some suitable copula C. If F is absolutely continuous and $F_1, \ldots, F_d$ are strictly increasing, a similar relationship exists for the probability density function (pdf) f of F, namely

$$f(\mathbf{x}) = c(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)) \prod_{i=1}^{d} f(x_i),$$

for $\mathbf{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$ where the copula pdf $\mathbf{c}$ is uniquely determined (See [30] for more details about the copula function).

While in recent years numerous bivariate copula families (also known as pair-copula families) have been developed, many of these bivariate families have no straightforward multivariate extension. A rich and flexible class of multivariate copulas that uses bivariate (conditional) copulas as building blocks only has recently investigated and applied on the numerous applications in [19,21]. The corresponding decomposition of a multivariate copula into bivariate copulas is called a pair-copula construction (PCC) [19]. The most widely researched copulas arising from PCCs are the vine copulas. These vine copulas admit a graphical representation called a regular vine (R-vine), which consists of a sequence of trees, each edge of which is associated with a certain pair copula in the PCC [30,19].

There have been several attempts to develop a method through using the nice properties of both graphical model and vine model, simultaneously. The main purpose is to benefit from the conditional independence in the graphs and the vine structure. This gap was filled in [22] and [23] by introducing non-Gaussian graphical model by combining useful properties of both pair-copula and DAG which was then called non-Gaussian PCCBNs.

Bauer et al. [22] and Bauer and Czado [23] by merging PCCs with (1), illustrate how the multivariate density given in (1) can be represented in terms of the PCC model. They suppose $D = (V, E)$ to be a DAG for a considered GRN with vertex set V and edge set E, and let f be a multivariate density function on d genes with marginal density f_i and the corresponding cumulative distribution function (CDF) F_i, $i = 1, 2, \ldots, d$. Then f is uniquely determined by its univariate margins f_i, $i = 1, 2, \ldots, d$; its conditional pair-copula $c_{vw|pa(v,w)}, v \epsilon V, w \epsilon pa(v)$ and f can be decomposed as:

$$\begin{aligned} f(x_1, \ldots, x_d) &= \textstyle\prod_{v=1}^{d} f(x_v) \\ &\times \textstyle\prod_{w \in pa(v)} c_{vw|pa(x,w)}(F_{v|pa(v,w)}, F_{w|pa(v,w)}). \end{aligned} \tag{2}$$

Such that $pa(x, w) \in E$ stands for an arrow $x \longrightarrow w$. By making suitable choices of marginal densities and pair-copula functions, the above presentation given in (2) provides us with an approach for the multivariate density. However, in practice, we have to use copula from a convenient class, and this class can be selected by criteria such as AIC and maximum likelihood (ML) and model selection techniques (For more explanations see [22,23]). In the following sections, we address this issue in more details to construct breast cancer GRNs.

3. Data Analysis: Breast Cancer Gene Regulatory Network

Breast cancer first develops from breast tissue and usually there are signs such as a mass in a breast, deformity in a breast, nasal discharge in a breast, secretion of fluid from a nipple, and so on, while many breast cancers have no obvious symptoms at all. Therefore, analysis of the GRNs for the breast cancer tumors is critical. On the other hand, breast cancer as a heterogeneous disease has multiple distinct molecular subtypes which depend on different therapies. There are 4 subtypes of breast cancer based on gene expression profiles: Basal-like (Basal), HER2-enriched (Her2), Luminal A (LumA), and Luminal B (LumB) [31]. LumA cancers are hormone-receptor positive (estrogen-receptor and/or progesterone-receptor positive) and Her2 negative. Her2-enriched cancers are hormone-receptor negative (estrogen-receptor and progesterone-receptor negative) and Her2 positive. LumA cancers tend to grow slowly and have the good prognosis, while basal-like cancers tend to grow fast and have poor prognosis. Therefore, in this section we intend to construct and compare the GRN of these breast cancer subtypes using the method outlined in this paper. It is hoped that this research could be a guide for various researchers to treat various subtypes of breast cancer.

Consider a real data set containing gene expression measurements from breast tumors, obtained from the Cancer Genome Atlas (TCGA) project [32]. The primary goal is to understand the underlying patterns of GRN variation among tumors in different subtypes of breast cancer: Basal, Her2, LumA, and LumB. Therefore, we will have the gene expression information of disease subtype for each tumor. We may regard cancer subtypes as a partial driver of the underlying structure of the gene expression data [2]. Samples from the same subtype will share common genetic variations. The raw data set contains 17814 genes and 343 samples. Out of the 343 samples, there are 4 subtypes of breast cancer with different number of 450 samples in each subtype: Her2 (42), Basal (66), LumA (154) and LumB (81). We preprocessed the data in the same way as in [33]. We first imputed missing values with the k-nearest neighbours algorithm ($k = 10$), then removed genes with low variations across samples (standard deviation smaller than 1.8), and finally mean centred each gene. The result is a 455 column-centred data matrix X with 343 samples and 645 genes.

Another widely cited characteristic of GRN is their abundance of certain repetitive sub-networks known as network motifs. In fact in many cases, due to the large

size of a GRN, the researchers examine an important sub-network that has been frequently repeated within the main network. Network motifs can be regarded as repetitive topological patterns when dividing a large network into smaller blocks [34]. Therefore, given that the GRN associated with breast cancer tumors is typically very large, we shall examine the considered motif for different subtypes of the breast cancer tumor using the method proposed within this paper, i.e. the non-Gaussian PCCBNs. The raw data set contains 17814 genes and 348 samples. We preprocessed the data in the same way as in Lock and Dunson (2013). We first imputed missing values with the k-nearest neighbors algorithm (k = 10), then removed genes with low variations across samples (standard deviation smaller than 1.8), and finally mean centered each gene. The result is a column-centered data matrix X with 348 samples and 10 genes. Note that by reducing the standard deviation of the studied genes, the less effective genes are incorporated into the model and the model components become significantly larger such that eventually leading to an inefficient model. The motif has been selected through an approach explained in [34], and it consists of 10 genes (*EYA2, PRC1, TACC3, GSTP1, AQP5, CSDA, EFS2, TPM2, BCL2* and *HEC*). Hereafter, we denote these genes by *G1* to *G10*, for readers' convenience. Figure 1 shows gene expression for different subtypes of breast cancer tumor Her2, Basal, LumA and LumB, individually. We first examine whether or not the gene expression data for the Her2 subtype follow a Gaussian distribution. The result of descriptive statistics and Kolmogorov-Smirnov normality test are reported in Table 1. As it is evident, the null hypothesis of normality for some genes are rejected (p-value is less than 5% significance level). Thus, a flexible and effective modelling is needed to model both non-Gaussian distributions and complex dependency structure between genes. Similarly, patterns can be used for other subtypes.

Table 1. Descriptive statistics and Kolmogorov Smirnov test for genes of Her2 breast cancer subtype

Gene	Mean	Std	Skewness	Kurtosis	p-value
G1	1.36	2.97	1.16	0.47	0.011*
G2	-3.26	2.17	0.44	-0.71	0.163
G3	3.27	3.25	0.27	-1.29	0.139
G4	0.087	2.26	1.45	0.85	0.000****
G5	0.755	2.81	-1.006	-0.44	0.000****
G6	1.56	2.77	-0.73	-0.42	0.001***
G7	-0.93	2.91	-0.064	-0.97	0.086
G8	2.55	2.15	-1.05	0.43	0.000****
G9	-2.94	2.19	1.12	0.92	0.046*
G10	-3.75	3.27	-0.25	-1.39	0.006**

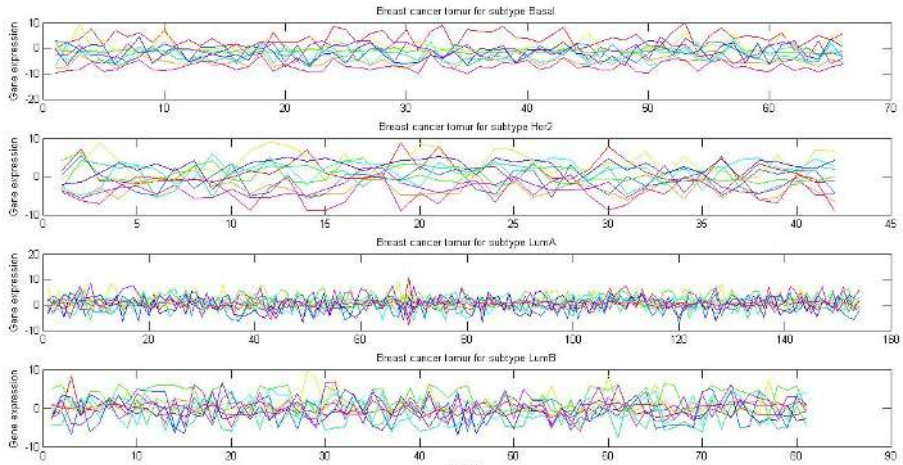


Fig. 1. gene expression for different subtypes of breast cancer tumor: Basal, Her2, LumA and LumB

3.1. *Dynamic Time Warping*

To construct GRNs, when the gene A is expressed, it takes a certain amount of time to express the gene B; therefore there will be a time delay. In many previous studies [26,17,27] during GRN construction, the time delay regulatory relation [25] between two genes has been ignored. We therefore consider the time delay regulatory relation for constructing GRNs using PCCBNs. This can be done using Dynamic Time Warping (DTW) which is a popular technique for comparing time series, providing both a distance measure that is insensitive to local compression and stretches, and the warping which optimally deforms one of the two input series onto the other.

DTW was first introduced in [35]. It has been employed for clustering gene expressions [36,37]. After stretching, the distance between the two genes is computed, by summing the distances of individual aligned elements. Suppose that $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_m)$ are two gene expression with lengths equal to n and m, respectively. Also suppose that a non-negative, local dissimilarity function γ is defined between any pair of elements x_i and y_j by $D(i,j) = \gamma(x_i, y_j) \geq 0$. The most common choice for $\gamma(.,.)$ is to assume the Euclidean distance. DTW finds a vector of ordered pair denoted ϕ_k, warping curve, for $k = 1, 2, \ldots, T$ such that $\phi(k) = (\phi_x(k), \phi_y(k))$, with $\phi_x(k) \in \{1, 2, \ldots, n\}$ and $\phi_x(k) \in \{1, 2, \ldots, m\}$. The warping functions ϕ_x and ϕ_y remap the time indices of X and Y, respectively. Given ϕ, we compute the sum accumulated distortion

between the warped gene expression X and Y by

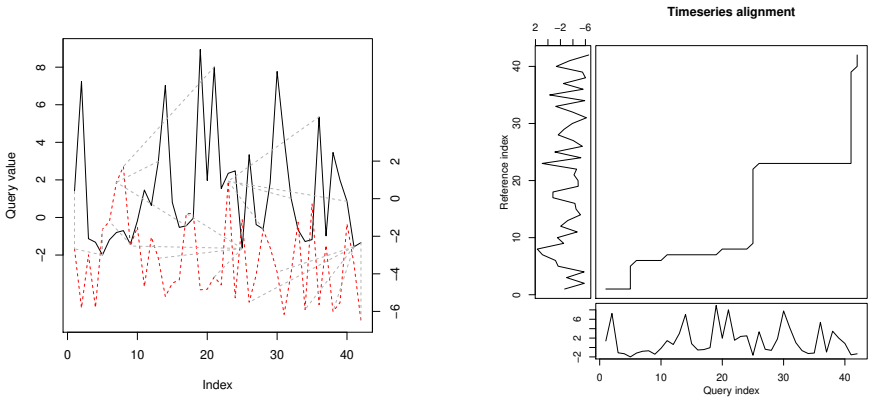
$$d_\phi(X,Y) = \sum_{k=1}^{T} d(\phi_x(k), \phi_y(k)),$$

To ensure reasonable warps, constraints are usually imposed on ϕ. For instance, monotonicity is imposed to preserve their time ordering and avoid meaningless loops such that $\phi_x(k+1) \geq \phi_x(k)$, & $\phi_y(k+1) \geq \phi_y(k)$. The idea underlying DTW is to find the optimal alignment ϕ such that

$$D(X,Y) = \min_{\phi} d_\phi(X,Y).$$

This minimization rule is fully described in literature [38]. Package "dtw" in *R software* provides a unification to compute and visualize DTW alignments [39]. Usually, two- and three-way plots are used to inspect the gene expression data along with their alignment. One intuitive alignment visualization style places both time series in the same plane, and connects the matching point pairs with segments. Another effective layout to display alignments places a gene expression data (G1) horizontally in a small lower panel, the other gene expression data vertically on the left; a larger inner panel holds the warping curve. Hence, the matching points can be recovered by tracing indices on gene G1, moving upwards until the warping curve is met, and then moving leftwards to discover the index of matched gene G2. One of the advantages of this method is the clear visualization.

The above pattern can now be used to determine time delay in breast cancer gene expression data. For instance, two and three way plots can be used to show the alignment for gene G1 and G2 in breast cancer Her2 subtype as presented in Figure 2 (a) and (b), respectively.



(a) (b)

Fig. 2. Two (a) and three (b) way plot for gene expression alignment using DTW for Her2.

3.2. *Learning Bayesian networks for the breast cancer GRNs*

Understanding causal relationships between genes is the primary goal in GRNs. A convenient approach for learning BNs in genetics is the use of expert knowledge. This method is ineffective since the elicitation process for such complex problems is very lengthy and almost impractical. Therefore, alternatively, the computational algorithms which can be implemented based on the existing gene expression data are used. The high dimensionality of the data sets-common in GRNs have led to the development of several learning algorithms focused on reducing computational complexity yet discovering the correct network. These algorithms provide us with invaluable information/statistics about the genes that could or could not be a cause of other genes of interest.

In our method, we initially apply a modified version of PC-algorithm [41] which Gaussian assumption of the marginal distribution is dropped. In other words, we provide an algorithm to approximate the network structure. In particular, this algorithm is well suited to determine a DAG structure of the non-Gaussian continuous variables constituting different subtypes of breast cancer GRNs. In the conventional version of the PC algorithm, a series of tests for conditional independence is performed given the normality assumption of marginals. Alternatively, we shall introduce a novel class of conditional independence tests that are particularly tailored for the algorithm and are applicable to non-Gaussian continuous data. In a Gaussian framework, the test of choice was usually a test for zero partial correlation (see [42]). We therefore need to introduce an equivalent of this test for the non-Gaussian case.

Suppose P is an absolutely continuous probability measure with Markov property for a DAG (V, E) on $[0,1]^d$ with uniform univariate marginal distributions. Moreover, let $\mathbf{u} = \{\mathbf{u}^1, \ldots, \mathbf{u}^n,\ n \in \mathbb{N}\}$ be a random sample realisation of $\{\mathbf{U}^1, \ldots, \mathbf{U}^n\}$ from a random variable $\mathbf{U}$ distributed as P. Considering the PC algorithm, for all distinct vertices $i, j \in V$ and chosen vertex sets $K \in V \backslash i, j$, the null hypothesis $H_0 : U_i \perp U_j | \mathbf{U_K}$ is tested against the general alternative $H_1 : U_i \perp U_j | \mathbf{U_K}$ of conditional dependence. Given a suitable independence test of choice, we should therefore determine H_0 and H_1 at significance level α. By confirming H_0, we remove the edge $i - j$ from considered DAG. In a Gaussian framework, the null hypothesis is then translated into $H_0 : \rho_{ij-K}(X_i, X_j; X_K) = 0$, where $X_k := \Phi^1(U_k)$ for all $k \in V$, and Φ denotes the univariate standard normal cdf. In fact, conditional independence test is translated to the test for zero partial correlation under the assumption of joint normality. We will now introduce a copula-based alternative test for conditional independence that is also applicable to non-Gaussian continuous data. Let $F_{i,j|K}(.,.|\mathbf{v_K})$ denote the conditional cdf of U_i and U_j given $\mathbf{U_K} = \mathbf{v_K}$, and let $C_{i,j|K}(.,.|\mathbf{v_K})$ be the corresponding conditional copula. Moreover, let $C_\perp$ denote the independence copula on $[0,1]^2$. The conditional independence $U_i \perp U_j | \mathbf{U_K}$ holds if and only if

$$F_{i,j|K}(v_i, v_j|\mathbf{v_K}) = C_{i,j|K}(F_{i|K}(v_i|\mathbf{v_K}), F_{j|K}(v_j|\mathbf{v_K})|\mathbf{v_K})$$

$$= F_{i|K}(v_i|\mathbf{v_K})F_{j|K}(v_j|\mathbf{v_K}),$$

for all $v_i, v_j \in [0,1]$ and $\mathbf{v_K} \in [0,1]^{|K|}$. Therefore, the null hypothesis of the conditional independence test can be stated as $H_0^* : C_{i,j|K}(.,.|\mathbf{v_K}) = C_\perp(.,.)$. The new null hypothesis H_0^* can be tested [43] for the transformed observations $W^1_{i|K}, \ldots, W^1_{i|K}$ and $W^1_{j|K}, \ldots, W^1_{j|K}$, where

$$W^k_{i|K} := F_{i|K}(U^k_i|\mathbf{U}^k_K), \quad \textit{and} \quad W^k_{j|K} := F_{j|K}(U^k_j|\mathbf{U}^k_K),$$

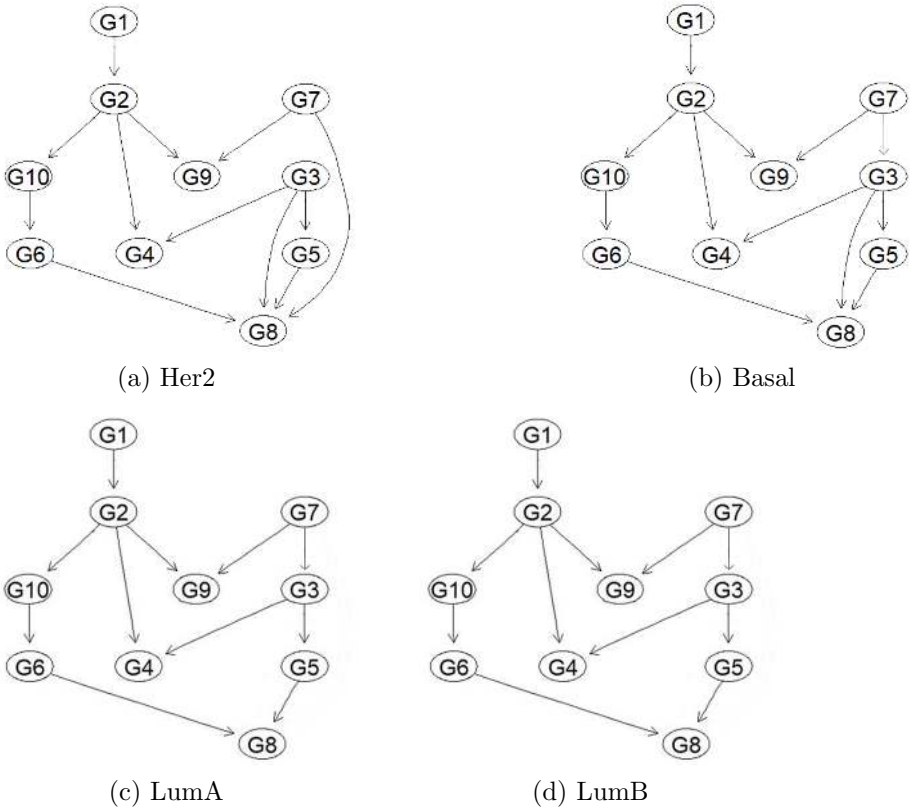
for all $k \in 1, \ldots, n$. In practice, with regards to DAG selection for non-Gaussian data, we can substitute the function *gaussCItest* in the function *pc* from *pcalg* with the independence test of Genest and Favre [44]. This is the modified version of *PC*-algorithm for the non-Gaussian data by the use of copula function property; such that we have used the functions *CDVineCopSelect* and *BiCopHfunc*, implemented in the CDVine package, to compute the Rosenblatt transforms [43], and a significance level of α in the independence tests.

We are now in a position to create a GRN as a non-Gaussian BN for different subtypes of the breast cancer in order to reveal their differences. Therefore, by following the discussed approach for gene expression data of each breast cancer subtype, we can create the GRN. The related GRN for the Her2 breast cancer subtype is illustrated in Figure 3 (a). In addition, this GRN for the other subtypes Basal, LumA, and LumB is presented in Figure 3 (b), (c) and (d), respectively. As it is evident, the number of edges in the Basal and Her2 subtypes is 12, while the LumA and LumB subtypes have 11 edges. In Basal subtype, in comparison to Her2, there exists edge G7$\longrightarrow$G3, while in Her2 this edge is removed, and instead, unlike Basal, G7$\longrightarrow$G8 is active. LumA and LumB are quite similar to each other. However, as shown in Figure 3, there are obvious differences with other subtypes. For example, G3$\longrightarrow$G8 is active in Basal while there is no such edge in LumA and LumB. Furthermore, instead of G7$\longrightarrow$G3, in Basal there are two edges G3$\longrightarrow$G8 and G7$\longrightarrow$G8. As discussed earlier, one of the most important principles in constructing GRNs is to distinguish active and inhibit genes; this can be done by adding and merging PCCs to the constructed DAGs for different breast cancer subtype. We explain this within the next subsection.

3.3. *Model selection*

Given the derived non-Gaussian DAG for different subtypes of breast cancer, we can decompose the multivariate density of our data by merging BNs with PCCs in order to derive PCCBNs model. We fully explain this matter for subtype Her2 and only present the final model for the rest of breast cancer subtypes.

Note that gene G8 has 4 parents (G3, G5, G6, G7) as the order of the parents is based on the heuristic rule of modelling strong bivariate dependences prior to weak dependences. Our decision was based on $\hat{\tau}$ of Kendall's estimates between the variables: $\hat{\tau}_{G8,G3} = 0.248$, $\hat{\tau}_{G8,G5} = 0.269$, $\hat{\tau}_{G8,G6} = 0.155$, and $\hat{\tau}_{G8,G7} = -0.145$. A similar rule can be applied for gene G4 and its parents G2 and G3. The estimated



(a) Her2 (b) Basal

(c) LumA (d) LumB

Fig. 3. The GRN for different breast cancer subtypes.

τ's Kendall between G4 and its parents G2 and G3 are: $\hat{\tau}_{G4,G2} = 0.399$ and $\hat{\tau}_{G4,G3} = -0.129$. The estimated τ's Kendall between gene G9 and its parents G2 and G7 are reported as $\hat{\tau}_{G9,G2} = 0.252$, and $\hat{\tau}_{G9,G7} = 0.142$. Based on these ordering, the resulting multivariate density decomposition using (2) is given by:

$$
\begin{aligned}
f_{1,\ldots,10}(x_1,\ldots,x_{10}) = \prod_{i=1}^{10} f_i(x_i) \times c_{21}(F_1(x_1), F_2(x_2)) \times \\
c_{6,10}(F_6(x_6), F_{10}(x_{10})) \times c_{42}(F_2(x_2), F_4(x_4)) \times \\
c_{92}(F_2(x_2), F_9(x_9)) \times c_{10,2}(F_2(x_2), F_{10}(x_{10})) \times \\
c_{53}(F_5(x_5), F_3(x_3)) \times c_{85}(F_5(x_5), F_8(x_8)) \times \\
c_{97|2}(F_{7|2}(x_7|x_2), F_{9|2}(x_9|x_2)) \times \\
c_{83|5}(F_{3|5}(x_3|x_5), F_{8|5}(x_8|x_5)) \times \\
c_{43|2}(F_{3|2}(x_3|x_2), F_{4|2}(x_4|x_2)) \times
\end{aligned}
$$

$$c_{86|35}(F_{6|35}(x_6|x_3,x_5),F_{8|35}(x_8|x_3,x_5)) \times$$
$$c_{87|356}(F_{7|356}(x_7|x_3,x_5,x_6),F_{8|356}(x_8|x_3,x_5,x_6)). \quad (3)$$

Considering the model selection techniques for the non-Gaussian PCCBNs ([22,23,20]), we can determine the best fitted copulas to any bivariate distribution in (3). For this purpose, we use the package *CDVine* and apply the function *BicopSelect* to the selected best fitted bivariate copulas. This also provides us with the estimates of the parameters and Kendall's τ for each copula. The results of the fitted

Table 2. Maximised log-likelihoods and parameters value for the Gaussian and non-Gaussian PCCBNs corresponding to the Her2 Brest cancer subtypes

		Non-Gaussian PCCBN				Gaussian BN	
	copula	τ	TD(L,U)	Parameters	LL	Parameters	LL
$c_{1,2}$	$BB8_{270}$	-0.35	(0,0)	-4.09(-0.65)	54.58	-0.44	42.29
$c_{2,4}$	Frank	0.39	(0,0)	4.07	65.51	0.49	53.58
$c_{2,9}$	Frank	0.25	(0,0)	2.41	25.16	0.33	23.08
$c_{2,10}$	SBB8	0.53	(0,0)	4.05(0.89)	142.89	0.61	88.6
$c_{7,9\|2}$	t-student	0.37	(0.27,0.27)	0.56(4.3)	87.55	0.43	44.41
$c_{5,3}$	Frank	0.25	(0,0)	2.46	25.94	0.31	19.8
$c_{8,5}$	Gaussian	0.24	(0,0)	0.36	28.8	0.19	28.8
$c_{10,6}$	SG	0.24	(0.31,0)	1.32	31.03	0.33	22.9
$c_{3,4\|2}$	Gaussian	0.45	(0,0)	0.66	68.19	0.48	68.19
$c_{3,8\|5}$	t-Student	-0.45	(0.06,0.06)	-0.66(3.58)	105.11	-0.51	66.96
$c_{6,8\|3,5}$	Clayton	-0.37	(0.56,0)	1.2	121.04	-0.36	83.49
$c_{8,7\|3,6,5}$	Gumbel	0.09	(0,0.03)	0.2	88.09	0.11	69.78

non-Gaussian PCCBN model and Gaussian BN are presented in Table 2. Some of the selected copula functions have two parameters, hence the second parameter is added in brackets. Note that each conditional distribution (or copula) in the Gaussian BN model must be normally distributed. Results, based on Log-Likelihood (LL) and AIC criteria, suggest that the non-Gaussian PCCBN model is much better fit to Her2 breast cancer subtype. As discussed, the flexibility of PCCBNs allow for capturing a wide range of distributional features and complex dependency structure such as heavy-tailedness, tail dependence, and non-linear, asymmetric dependence in the constructed GRN for Her2 subtype. One approach to present these features is contour plot. The contour plots for the estimated copula are presented in Figure 4. As it can be seen from these contour plots, the occurrence of extreme events in some pair of genomic data is evident. Also, in order to study the occurrence of extreme events in genomic data, the pair-wise analysis of upper (U) and lower (L) tail dependence (TD) of the gene expression variables can be implemented using the fitted copula models [30,45]. The computed coefficients of upper and lower tail dependence of any two gene expression variables are presented in Table 2. In genetic data, it is crucial to consider the tail-dependence coefficient in modelling of

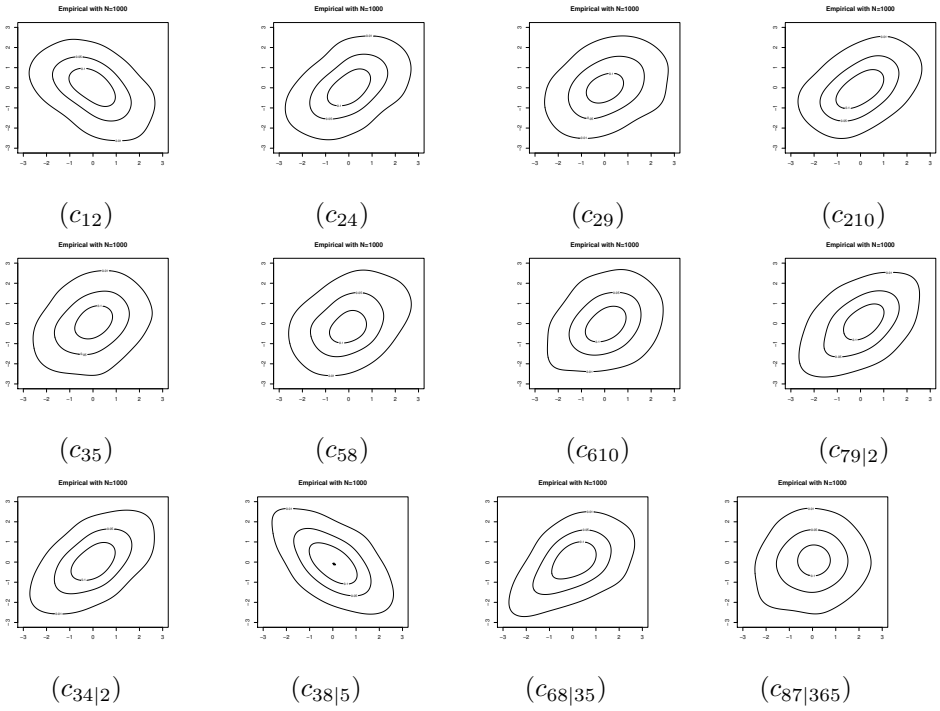
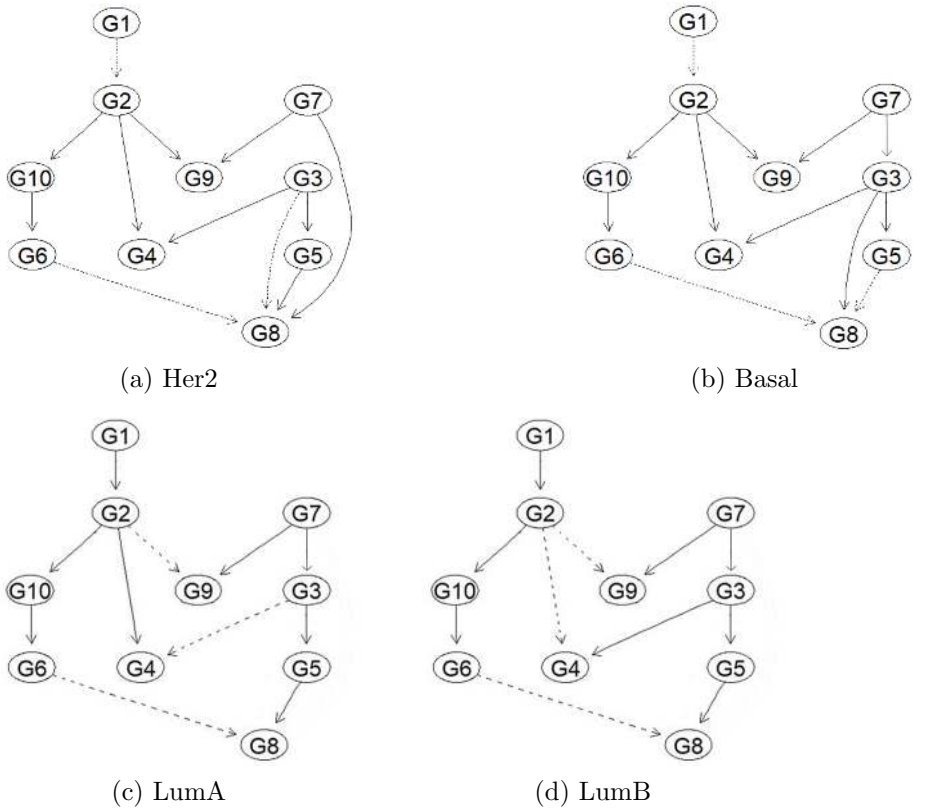


Fig. 4. Contur plot for the estimated copula function of Her2 subtype.

joint GRN construction [45]. Otherwise, it can lead to a serious misspecification of GRNs' structures. Therefore, computing the tail dependence coefficients as precise as possible can provide an accurate interpretation of the constructed GRN accurately. The resulted contour plots for the estimated copula confirms the obtained calculations in Table 2. As an example, the contour plot, demonstrated in Figure 4, between $G3$ and $G8$ given $G5$ confirms the calculated TD in this table. The final

Table 3. Maximised log-likelihoods, numbers of parameters, and AIC values for the Gaussian and non-Gaussian PCCBNs corresponding to the different Breast cancer subtypes

Subtype	Model	Log-Liklihood	Parameters	AIC
Her2	Gaussian	611.88	30	-1163.76
	non-Gaussian	843.99	16	-1655.98
Basal	Gaussian	758.6	26	-1465.2
	non-Gaussian	989.1	15	-1948.2
LumA	Gaussian	586.4	27	-1118.8
	non-Gaussian	721.2	17	-1408.4
LunB	Gaussian	678.9	25	-1307.8
	non-Gaussian	776.3	16	-1520.6



(a) Her2 (b) Basal

(c) LumA (d) LumB

Fig. 5. The final GRN for different breast cancer subtypes.

results for the other subtypes of breast cancer based on the LL and AIC criteria in Table 3 confirms our claims about the superiority of non-Gaussian PCCBN versus Gaussian BN.

Using Kendall's τ correlation criterion obtained from fitted copula function between two interested genes, it can be shown which genes deactivate others. A simple rule in this is that if Kendall's τ correlation coefficient of the fitted copula between two interested genes is negative, then the two genes inhibit each other. Final GRN's for each subtype of breast cancer, following coefficient calculations using a copula function,are shown in Figure 5. For instance, in Figure 5 (a) for Her2 subtype, G1 deactivate G2 (dashed line), and G6 deactivate G8. Similar pattern is repeated for Basal. The difference between the two structures is that in Her2, G3 deactivate G8, while in Basal G5 deactivate G8. Moreover, LumA and LumB cn be easily compared visually. In fact, such network construction for various breast cancer subtypes-with activated and inhibited genes-can help to discover new drugs, new treatments, and even new preventive approaches.

4. Simulation

First, we follow the simulation method proposed in [30] which is based on the sampling of the cumulative distributions and is widely known as the probability integral transform (PIT). This simulation method has later been followed by [22,23]. The same simulation method is used in [20] to draw samples from the approximated PCCBNs where data availability is very limited. The sampling strategy, based on the PIT is as follows: 1) sample from two independent variables, denoted by U_1, U_2 and distributed uniformly on $[0, 1]$; 2); then calculate values of the original variables using the following equations:

$$x_1 = u_1, \qquad x_2 = F^{-1}_{2|1}(u_2|x_1),$$

where x_i's and and u_i's are realization values of X_i's and U_i's respectively. Finally, this can be easily continued to all variables within the PCCBN considering a similar argument [20,19,24].

One of the effective measures to evaluate the performance of non-Gaussian PCCBN against Gaussian BN in the construction of GRN of interest, is the F-score which is defined as:

$$F = \frac{2pr}{p + r}, \tag{4}$$

where p and r denote precision and recall, respectively. They can be computed using the following equation:

$$p = \frac{TP}{TP + FP} \quad \text{and} \quad r = \frac{TP}{TP + FN},$$

where TP is true positive, FP is false positive and FN is false negative. TPs denote the number of the edges that are inferred by the GRN algorithm that actually exist within the true network. FPs are the edges inferred by the GRN algorithm however do not actually exist in the true network. FN edges actually exist in the true network but cannot be inferred by the GRN algorithm. The use of truly and falsely inferred edges (TPs and FPs) only is not sufficient to measure the performance of the underlying model fairly. Missing edges (FNs) in the final network should also be considered in the performance evaluation process. Since the F-score given in Eq. (4) takes into account all types of edges, it is a more reliable measure for completely known networks (mostly the synthetic ones). Table 4 presents F-scores for non-Gaussian PCCBN versus Gaussian BN associated with reconstruction of the different breast cancer subtypes. It can be observed that the F-score of the GRNs for different subtype reconstructed based on the non-Gaussian PCCBN is significantly greater that the ones reconstructed using the Gaussian BN.

5. Conclusions

The Gaussian graphical models with Markov property, in particular BNs, have been widely used in modelling gene regulatory networks in many research studies such as

Table 4. The F-score evaluation of the reconstructed GRN for different breast cancer subtypes.

	$Her2$	$Basal$	$LumA$	$LumB$
GBN	0.56	0.47	0.54	0.57
$PCCBN$	0.63	0.59	0.64	0.61

detecting the activator genes of the genetic diseases, determining the functions of the regulating and regulated genes, obtaining the drug targets of the medical cures, etc. The main drawback of the Gaussian BN models is that the normal assumption of data in most of these applications is not realistic, and the inferred networks will be misleading if the data distributions are non-Gaussian. In the present paper, we propose a novel method of constructing GRNs based on the PCCBN models which benefit from the computational power and flexibility of the PCC models and the parsimony of conditional independence concepts of the BNs. We demonstrated that the effectiveness of using the PCCBNs in constructing GRN where the marginal and conditional distributions are not normal. In addition, we adopted the PC algorithm to select the best network structure for the GRNs of interest by considering the conditional independence/dependence extracted from the considered BN and PCC models. Furthermore, we demonstrated that the flexibility of PCCBNs allow for capturing a wide range of distributional features with the complex dependency including the tail dependence, non-linear, and asymmetric dependence which are common in the underlying GRNs. We also compared the proposed non-Gaussian PCCBNs with Gaussian BN based on various goodness-of-fit measures and the F-score which is a combination measure of TP, FP, and FN. The results clearly suggest that the non-Gaussian PCCBN outperformed Gaussian BN based on the aforementioned criteria illustrated in Tables 3 and 4.

References

1. Cheung,V. G. and Spielman, R. S. (2002) The genetics of variation in gene expression. Nat. Genet., **32**, 522-525.
2. Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. Nature Genetics, **37(7)**:710-717.
3. Deplancke, B. and Gheldof, N. (2012). Gene Regulatory Networks: Methods and Protocols, Springer New York Dordrecht Heidelberg London.
4. Lee, C. Leu, Y. Yang, W. (2012). Constructing gene regulatory networks from microarray data using GA/PSO with DTW, Applied Soft Computing, **12**, 1115-1124.
5. Gupta, R. R. and Achenie, L. E. K. (2007). A network model for gene regulation, Computers and Chemical Engineering, **31**, 950-961.
6. Das, S. Caragea, D. Welch, S. M. Hsu, W. H. (2010). Handbook of Research on Computational Methodologies in Gene Regulatory Networks, USA, Medical Information Science Reference.
7. Rau, A., Jaffrézic, F., Foulley, J. L., Doerge, R. W. (2012). Reverse engineering gene regulatory networks using approximate Bayesian computation, Stat. Comput., 22:1257-1271.

8. Creighton, C. Hanash, S. (2003) Mining gene expression databases for association rules, Bioinformatics, **19**, 79-86.
9. Huang, Z. Li, J. Su, H. Watts, G.S. Chen, H. (2007). Large-scale regulatory network analysis from microarray data: modified Bayesian network learning and association rule mining, Decis. Support Syst. **43**, 1207-1225.
10. Wang, H. C. Lee, Y. S. (2005). Gene network prediction from microarray data by association rule and dynamic Bayesian network, Lect. Notes Comput. Sci. 34-82.
11. Chan, Z. S. H. Havukkala, I. Jain, V. Hu, Y. Kasabov, N. (2008). Soft computing methods to predict gene regulatory networks: an integrative approach on time-series gene expression data, Appl. Soft Comput. **8**, 1189-1199.
12. Tian, T. (2010). Stochastic models for inferring genetic regulation from microarray gene expression data, Biosystems, **99(3)**, 192-200.
13. Friedman, N. Linial, M. Nachman, I. Pe'er, D. (2000). Using Bayesian networks to analyze expression data, J. Comput. Biol. **7**, 601-620.
14. Chan, Z. S. H. Collins, L. Kasabov, N. (2007). Bayesian learning of sparse gene regulatory networks, Biosystems, **87**, 299-306.
15. Zou, M. Conzen, S.D. (2005). A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data, Bioinformatics, **21**, 71-79.
16. Zhang, Q. Shi, X. (2017). A mixture copula Bayesian network model for multimodal genomic data, Cancer Informatics, **16**, 1-11.
17. Gao, B. Cui Y. (2015). Learning directed acyclic graphical structures with genetical genomics data. Bioinformatics. **31(24)**3953-3960.
18. Elidan, G. (2017), Copula Bayesian Networks, Cancer Informatics, **16**, 1-11.
19. Aas, K., Czado, K. C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence, *Insurance, Mathematics and Economics*, **44**, 182-198.
20. Daneshkhah, A., Parham, G., Chatrabgoun, O., and Jokar, M. (2016). Approximation Multivariate Distribution with pair copula Using the Orthonormal Polynomial and Legendre Multiwavelets basis functions. *Comm. in Stat. - Simu. & Comput.*, **45(2)**, 389-419.
21. Bedford, T., Daneshkhah, A., and Wilson, K. (2016). Approximate Uncertainty Modelling with Vine copulas, *Risk Analysis*, **36(4)**, 792-815.
22. Bauer, A., Czad,. C., and Klein, T. (2012). Pair-copula constructions for non-Gaussian DAG models. *The Canadian Journal of Statistics* **40(1)**, 86-109.
23. Bauer, A., Czad,. C. (2016). Pair-copula Bayesian networks. *Journal of Computational and Graphical Statistics,* **25(4)**, 1248-1271.
24. Chatrabgoun, O. Hosseinian-Far, H. Chang, V. Stocks, N. G. Daneshkhah, A.(2018). Approximating Non-Gaussian Bayesian Networks using Minimum Information Vine Model with Applications in Financial Modelling. *Journal of Computational Science,* **24**, 266-276.
25. Dasika, M. S. Gupta, A. Maranas, C. D. Varner, J. D. (2004). A mixed integer linear programming (MILP) framework for inferring time delay in gene regulatory networks, Pac. Symp. Biocomput. **9**, 474-485.
26. Kurt, Z. Aydin, N. Altay, G. (2015). A comprehensive comparision of association estimators for gene network inference algorithms, bioinformatics, **30(15)**, 2142-2149.
27. Zhang, X. F. Yang, L. Q. Yang, S. Hu, X. Yan, H. (2017). DiffGraph: An R package for identifying gene network rewiring using differential graphical models. Bioinformatics. doi.10.1093/bioinformatics/xxxxxx.
28. Schena, M. Shalon, D. Davis, R. W. Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science, **270**,

467-470.
29. Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (2003). Probabilistic Networks and Expert Systems, 2nd ed., Springer, New York.
30. Kurowicka, D., and Cooke. R. (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling*. John Wiley.
31. Schnitt. S. J. (2010). Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy. Mod Pathol, **23**: 60-S64.
32. The Cancer Genome Atlas Network, (2012). Comprehensive molecular portraits of human breast tumours. Nature, **490(7418)**, 61-70.
33. Lock, E. F. and Dunson, D. B. (2013). Bayesian consensus clustering. *Bioinformatics*, **29(20)**, 2610 - 2616.
34. Burda, Z., Krzywicki, A., Martin, O., Zagorski, M. (2011). Motifs emerge from function in model gene regulatory networks. Proc Natl Acad Sci USA, **108**,17263-17268.
35. Bellman, R. Kalaba, R. (1959). On adaptive control processes, IRE Trans. Automat. Control, **4**, 1-9.
36. Aach, J. Church, G. M. (2001). Aligning Gene Expression Time Series with Time Warping Algorithms. Bioinformatics, **17(6)**, 495-508.
37. Hermans, F., Tsiporkova, E. (2007). Merging Microarray Cell Synchronization Experiments Through Curve Alignment. Bioinformatics, **23(2)**, 64-70.
38. Jayadevan, R. Satish, R. K. Pradeep, M. P. (2009). Dynamic time warping based static hand printed signature verification, J. Pattern Recogn. Res. **4**, 52-65.
39. Giorgino, T. (2009), Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package, Journal of Statistical Software, **31(7)**, 6-24.
40. Kalisch, M. Mächler, M. Colombo, D. Maathuis, M. H. Bühlmann, P. (2012). Causal Inference Using Graphical Models with the R Package pcalg, Journal of Statistical Software, **47(11)**, 1-26.
41. Kalisch, M. Buhlmann, P. (2007). Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. Journal of Machine Learning Research, **8**, 613-636.
42. Anderson, T. W. (2003). An Introduction to Multivariate Statistical Analysis. John Wiley & Sons, Chichester, third edition.
43. Rosenblatt. M. (1952). Remarks on a multivariate transformation. The Annals of Mathematical Statistics, **23(3)**:470-472.
44. Genest, C., Favre, A. C., (2007). Everything you always wanted to know about copula modeling but were afraid to ask. J. Hydrol. Eng. **12**, 347-368.
45. Telesca, D., Muller, P., Parmigiani, G., Freedman, R. S. (2012). Modeling Dependent Gene Expression, *The Annals of Applied Statistics*, **6(2)**, 542-560.