



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/cstda

Partial correlation with copula modeling

Jong-Min Kim^{a,*}, Yoon-Sung Jung^b, Taeryon Choi^c, Engin A. Sungur^a^a Statistics Discipline, Division of Science and Mathematics, University of Minnesota at Morris, Morris, MN, 56267, USA^b Office of Research, Alcorn State University, Alcorn State, MS, 39096, USA^c Department of Statistics, Korea University, Seoul, 136-701, South Korea

ARTICLE INFO

Article history:

Received 26 January 2010

Received in revised form 21 September 2010

Accepted 22 September 2010

Available online 16 October 2010

Keywords:

Partial correlation

Gaussian copula

Gene network

ABSTRACT

We propose a new partial correlation approach using Gaussian copula. Our empirical study found that the Gaussian copula partial correlation has the same value as that which is obtained by performing a Pearson's partial correlation. With the proposed method, based on canonical vine and d-vine, we captured direct interactions among eight histone genes.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

The current Pearson partial correlation approach is popular because of the simple computation advantage it confers. But the current approach has many drawbacks: for example, it does not exist if the first or second moments do not exist. Possible values depend on the marginal distributions which are not invariant under non-linear strictly increasing transformations (Kurowicka and Cooke, 2006). This was our motivation to propose a new approach to partial correlation using copula, specifically a Gaussian copula. Since Sklar (1959) proposed the theorem of the copula, numerous copula functions have been introduced in the last five decades. Recently, Nelson (2006) summarized the theories of numerous copula functions and Yan (2007) developed the R package of multivariate dependence with copulas. But most copulas have a limitation which fails to satisfy the copula properties when extended from bivariate to multivariate cases. To overcome this limitation, Aasa et al. (2009) proposed pair-copula constructions of multiple dependence, based on the work of Bedford and Cooke (2002). Since model construction is hierarchical, it is not simple to incorporate more variables in the conditioning sets with pair-copula which uses the inverse of the conditional bivariate distribution function, h -function inverse. But pair-copula constructions by Aasa et al. (2009) are promising way to derive a partial correlation, so we adopted a Gaussian bivariate copula by using the conditional distributions to find a partial correlation. To find a partial correlation, we derive a conditional standard normal distribution by using multivariate normal distribution properties and estimate the partial correlation coefficient by the Gaussian copula. In the general theory of partial correlation, the partial correlation coefficient is a measure of the strength of the linear relationship between two variables after we control for the effects of other variables. If the two variables of interest are Y and X , and the control variables are $Z_1, Z_2, \dots, Z_n$, then we denote the corresponding partial correlation coefficient by $\rho_{YX|Z_1, Z_2, \dots, Z_n}$.

* Corresponding author.

E-mail address: jongmink@morris.umn.edu (J.-M. Kim).

The general formulas to compute a first-order partial correlation and a second-order partial correlation by Pearson (1916) are

$$\rho(YX; Z) = \frac{\rho_{YX} - \rho_{YZ}\rho_{XZ}}{\sqrt{(1 - \rho_{YZ}^2)(1 - \rho_{XZ}^2)}}$$

and

$$\begin{aligned} \rho(YX; Z, W) &= \frac{\rho_{YX;Z} - \rho_{YZ;W}\rho_{XZ;W}}{\sqrt{(1 - \rho_{YZ;W}^2)(1 - \rho_{XZ;W}^2)}} \\ &= \frac{\rho_{YX;Z} - \rho_{YW;Z}\rho_{XW;Z}}{\sqrt{(1 - \rho_{YW;Z}^2)(1 - \rho_{XW;Z}^2)}}. \end{aligned}$$

The general formula for an n th order partial correlation can be computed from correlations with the following recursive formula (Yule and Kendall, 1965):

$$\rho_{YX|Z_1, Z_2, \dots, Z_n} = \frac{\rho_{YX|Z_1, Z_2, \dots, Z_{n-1}} - (\rho_{YZ_n|Z_1, Z_2, \dots, Z_{n-1}})(\rho_{XZ_n|Z_1, Z_2, \dots, Z_{n-1}})}{\sqrt{(1 - \rho_{YZ_n|Z_1, Z_2, \dots, Z_{n-1}}^2)(1 - \rho_{XZ_n|Z_1, Z_2, \dots, Z_{n-1}}^2)}}.$$

Our Gaussian copula method to find a partial correlation is very simple. We derive the conditional distribution of X_1, X_4 given X_2, X_3 as follows:

$$F_{14|23}(X_1, X_4|X_2, X_3) = C^{Ga}(F_{1|23}(X_1|X_2, X_3), F_{4|23}(X_4|X_2, X_3); \rho_{12|34}). \quad (1)$$

Then, using a Gaussian copula, we can estimate a correlation coefficient parameter $\rho_{12|34}$ by the maximum likelihood estimation approach. The estimate of $\rho_{12|34}$ is the partial correlation coefficient of X_1 and X_4 given X_2 and X_3 , $r_{13|2}$. So our proposed method can be applied to many fields such as finance, insurance, and biology.

The properties of copula, the definition of Gaussian copula, and the definition of partial copula and vine copula are introduced in Section 2. The copula parameter estimation methods for the partial correlation by Gaussian copula are presented in Section 3. Its application to gene data is given in Section 4. Section 5 concludes the paper with a discussion of the advantages of the method and future research plans.

2. Method

2.1. Definitions of copula

The dependence structure of a set of random variables is contained within F . The idea of separating F into one part which describes the dependence structure and other parts which describe only the marginal behavior has led to the concept of a copula. A copula is a multivariate uniform distribution representing a way of trying to extract the dependence structure of the random variables from the joint distribution function. It is a useful approach to understanding and modeling dependent random variables. Every joint distribution can be written as

$$F_{XY}(x, y) = C(F_X(x), F_Y(y))$$

where F_X and F_Y are marginal distributions.

Definition 1 (Bivariate Copula). A bivariate copula is a function $C : [0, 1]^2 \rightarrow [0, 1]$, whose domain is the entire unit square with the following three properties:

- (i) $C(u, 0) = C(0, v) = 0$, $\forall u, v \in [0, 1]$
- (ii) $C(u, 1) = C(1, u) = u$, $\forall u \in [0, 1]$
- (iii) $C(u_1, v_1) - C(u_1, v_2) - C(u_2, v_1) + C(u_2, v_2) \geq 0$, $\forall u_1, u_2, v_1, v_2 \in [0, 1]$ such that $u_1 \leq u_2$ and $v_1 \leq v_2$.

Bivariate measures of dependence for continuous variables are as follows:

- Spearman's rho:

$$\rho_C = 12 \int_0^1 \int_0^1 [C(u, v) - uv] dudv.$$

- Kendall's tau:

$$\tau_C = 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1.$$

Sklar (1973) showed that any multivariate distribution function, for example F , can be represented as a function of its marginals, for example G and H , by using a copula C , i.e., $F(x, y) = C(G(x), H(y))$.

We denote the distribution function of standard normal by

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{w^2}{2}\right\} dw.$$

We consider an n -variate normal random vector $z = (z_1, z_2, \dots, z_n)$ with z_k is distributed as $N(0, 1)$ for $k = 1, 2, \dots, n$ and has positive definite, symmetric covariance matrix $\mathbf{V} = (v_{ij})$. With elements

$$v_{ij} = \begin{cases} 1, & \text{if } i = j, \\ \text{corr}(z_i, z_j), & \text{otherwise.} \end{cases}$$

The relation is

$$\frac{\partial \Phi(x, y; \rho)}{\partial \rho} = \phi(x, y; \rho)$$

where

$$\phi(x, y; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp \left\{ -\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)} \right\}$$

and

$$\Phi(z_1, z_2; \rho) = \int_{-\infty}^{z_1} \int_{-\infty}^{z_2} \phi(x, y; \rho) dx dy.$$

The joint density of z is

$$\phi(z_1, \dots, z_n) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{V}|}} \exp \left\{ -\frac{1}{2} z^T \mathbf{V}^{-1} z \right\} dw.$$

The joint distribution is

$$\Phi(z_1, \dots, z_n) = \int_{-\infty}^{z_n} \int_{-\infty}^{z_{n-1}} \cdots \int_{-\infty}^{z_1} \phi(x_1, x_2, \dots, x_n) dx_1 \cdots dx_n.$$

Definition 2 (Gaussian Copula). The copula defined by

$$C^{Ga}(u_1, \dots, u_n) = \Phi(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_n))$$

where $z_1 = \Phi^{-1}(u_1), \dots, z_n = \Phi^{-1}(u_n)$, is called the *Gaussian copula*.

Gaussian copula is by far the most popular copula used in the financial industry in default dependency modeling. There are two reasons for this. First, it is easy to simulate. Second, it requires the right number of parameters equal to the number of correlation coefficients among the underlying names.

2.2. Partial copula

Given an n -dimensional distribution function F with continuous marginal (cumulative) distributions $F_1, \dots, F_n$, there exists a unique n -copula $C : [0, 1]^n \rightarrow [0, 1]$ such that

$$F(x_1, \dots, x_n) = C(F(x_1), \dots, F(x_n)).$$

Suppose Y and Z are real-valued random variables with conditional distribution functions

$$F_{2|1}(y|x) = P(Y \leq y|X = x)$$

and

$$F_{3|1}(z|x) = P(Z \leq z|X = x).$$

Then the basic property of

$$U = F_{2|1}(Y|X) \quad \text{and} \quad V = F_{3|1}(Z|X)$$

is as follows:

Lemma 1. Suppose, for all x , $F_{2|1}(y|x)$ is continuous in y and $F_{3|1}(z|x)$ is continuous in z . Then U and V have uniform marginal distributions.

Proof. By continuity of $F_{2|1}(y|x)$ in y , and with F_1 the marginal distribution function of X ,

$$\begin{aligned} P(U \leq u) &= P(F_{2|1}(Y|X) \leq u) \\ &= \int P(F_{2|1}(Y|x) \leq u) dF_1(x) \\ &= \int_u u dF_1(x) = u. \quad \square \end{aligned}$$

Bergsma (2004) defined a partial copula for testing conditional independence for continuous random variables as follows:

Definition 3. The joint distribution of U and V is called the partial copula of the distribution of Y and Z given X . That is,

$$C(U = F_{2|1}(Y|X), V = F_{3|1}(Z|X)) = F_{23|1}(Y, Z|X).$$

Theorem 1. If $X_1, \dots, X_n$ is a vector of n random variables with absolutely continuous multivariate distribution function F , then the n random variables

$$U_1 = F_1(X_1), \quad U_2 = F_{2|1}(X_2|X_1), \dots, \quad U_n = F_{1|2,\dots,n}(X_n|X_1, \dots, X_{n-1}) \quad (2)$$

are i.i.d. $U(0, 1)$.

To define a copula we begin by considering n standard-uniform random variables $X_1, \dots, X_n$. We do not assume that $X_1, \dots, X_n$ are independent, they may be related. The dependence between the real-valued random variables $X_1, \dots, X_n$ is completely described by their joint distribution function

$$F(X_1, \dots, X_n) = P[X_1 \leq x_1, \dots, X_n \leq x_n]. \quad (3)$$

In the absence of a model for our random variables, correlation (linear or rank) is only of very limited use. On the other hand, if we have a model for $X_1, \dots, X_n$ in the form of a joint distribution F , then we know everything that is to be known about these random variables. We know their marginal behavior and we can evaluate the conditional probabilities that one component takes certain values, given that other components take other values, such as

$$F_{1|2,\dots,n}(x_1, \dots, x_n) = P[X_1 \leq x_1 | X_2 = x_2, \dots, X_n = x_n]. \quad (4)$$

Corollary 1 (Joe, 1997). Assuming $F_{1,\dots,n-1}, F_{2,\dots,n}$ have been defined with a common $(n-2)$ -dimensional margin $F_{2,\dots,n-1}$, the n -variate family is

$$F_{1,\dots,n}(x_1, \dots, x_n) = \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_{n-1}} C_{1n} \left(F_{1|2,\dots,n-1}(x_1|z_2, \dots, z_{n-1}), \right. \\ \left. F_{n|2,\dots,n-1}(x_n|z_2, \dots, z_{n-1}) \right) F_{2,\dots,n-1}(dz_2, \dots, dz_{n-1}), \quad (5)$$

where $F_{1|2,\dots,n-1}, F_{n|2,\dots,n-1}$ are conditional cdfs obtained from $F_{1,\dots,n-1}, F_{2,\dots,n}$.

The Gaussian copula example of Corollary 1 from Joe (1997) can be found.

Corollary 2 (Matteis, 2001). The conditional distribution of U_i given the values of the first $(i-1)$ components of $(U_1, U_2, \dots, U_n)^t$ can be written in terms of derivatives and densities of the i -dimensional marginals

$$C_i(U_i|U_1, U_2, \dots, U_{i-1}) = P[U_i \leq u_i | U_1 = u_1, \dots, U_{i-1} = u_{i-1}] \\ = \frac{\partial^{i-1} C_i(u_1, u_2, \dots, u_i)}{\partial u_1 \cdots \partial u_{i-1}} \bigg/ \frac{\partial^{i-1} C_i(u_1, u_2, \dots, u_{i-1})}{\partial u_1 \cdots \partial u_{i-1}}. \quad (6)$$

2.3. Vine copula

Vines, a fairly new concept within dependence modeling, are tools which employ an intuitive graphical representation to describe what conditional specifications are being made on the joint distribution. Applications of vines can be found where there is a situation of high dimensionality: precipitation data, equity returns, insurance failure market, gene networks, etc. Using the definition by Bedford and Cooke (2002):

Definition 4 (Regular Vine, Vine). V is a vine on n elements if:

- (i) $V = (T_1, \dots, T_m)$.
- (ii) T_1 is a tree with nodes $N_1 = 1, \dots, n$ and a set of edges denoted E_1 .
- (iii) For $i = 2, \dots, m$, T_i is a tree with nodes $N_i \subset N_1 \cup E_1 \cup E_2 \cup \dots \cup E_{i-1}$ and edge set E_i .

A vine V is a regular vine on n elements if:

- (i) $m = n$.
- (ii) T_i is a connected tree with edge set E_i and node set $N_i = E_{i-1}$, with $\#N_i = n - (i - 1)$ for $i = 1, \dots, n$, where $\#N_i$ is the cardinality of the set N_i .
- (iii) The proximity condition holds: for $i = 2, \dots, n-1$ if $a = \{a_1, a_2\}$ and $b = \{b_1, b_2\}$ are two nodes in N_i connected by an edge $(a_1, a_2, b_1, b_2 \in N_{i-1})$, then $\#a \cap b = 1$.

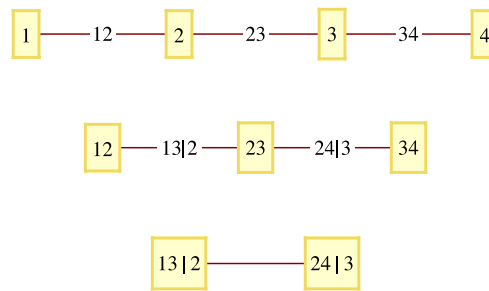


Fig. 1. An example of a four-dimensional *D*-vine.

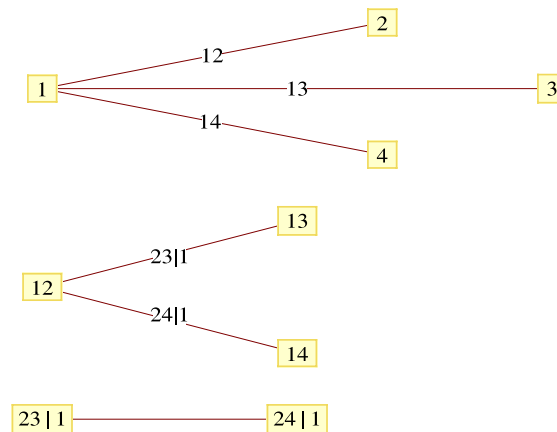


Fig. 2. An example of a four-dimensional canonical vine.

For this paper, we are only interested in regular vines which can be envisioned as layered acyclical trees (i.e. a forest). In the base tree, the nodes represent each particular variable while each edge represents the bivariate probability distributions (bivariate copula in our case) of the two nodes it connects. In the next level, the base edges become the nodes, while the edges represent the conditional bivariate probability distributions for the two variables not contained in both nodes connected by that particular edge. For level n , the edges from $n - 1$ become the nodes, and the edge(s) represent the conditional bivariate probability distribution(s) of the two variables not contained in both nodes connected by that particular edge. Those interested in a more formal introduction and definition of vines can refer to Bedford and Cooke (2002). There are only two types of vines which are relevant to this paper: *D*-vines and canonical vines which are defined as follows:

- If each node in the base tree has a degree of at most two, then the vine is a *D*-vine.
- If each tree has a unique node of degree $n - i$, then the vine is a canonical vine.

It should be noted that canonical vines are generally most effective when there is one obvious variable that tends to dominate the others. *D*-vines, on the other hand, are more effective when all the variables tend to be of equal importance (Aasa et al., 2009). Figs. 1 and 2 provide examples of a *D*-vine and a canonical vine.

3. Copula parameter estimation

3.1. Gaussian copula partial correlation parameter estimation

We use two different parameter estimation methods: the Canonical Maximum Likelihood (CML) method presented in Genest et al. (1995), and the Inference Functions for Margins (IFM) method by Joe (1997). The dependence parameter α by CML is consistent, asymptotically normal and fully efficient at independence. When the consistent estimation of the dependence parameter α is important, the CML method without specifying the marginal distributions is suitable. This approach uses the empirical CDF of each marginal distribution to transform the observations $(X_{i1}, \dots, X_{ip})^T$ into pseudo-observations with uniform margins $(U_{i1}, \dots, U_{ip})^T$ and then estimates α as

$$\hat{\alpha}_{\text{CML}} = \operatorname{argmax}_{\alpha} \sum_{i=1}^n \log c(U_{i1}, \dots, U_{ip}; \alpha).$$

Joe (1997) proposed the estimation of parameters of a copula-based multivariate model, based on the likelihoods of marginal distributions of the model. This method is called Inference Functions for Margins (IFM) method following the

terminology of McLeish and Small (1988) and Xu (1996). The IFM method estimates the marginal parameters β in a first step by

$$\hat{\beta}_{\text{IFM}} = \operatorname{argmax}_{\beta} \sum_{i=1}^n \sum_{j=1}^p \log f_i(X_{ij}; \beta)$$

and then estimates the association parameters α given $\hat{\beta}_{\text{IFM}}$ by

$$\hat{\alpha}_{\text{IFM}} = \operatorname{argmax}_{\alpha} \sum_{i=1}^n \log c(F_1(X_{i1}; \hat{\beta}_{\text{IFM}}), \dots, F_p(X_{ip1}; \hat{\beta}_{\text{IFM}}); \alpha).$$

When each marginal distribution F_i has its own parameters β_i so that $\beta = (\beta_1^T, \dots, \beta_p^T)^T$, the first step consists of an ML estimation for each margin $j = 1, \dots, p$:

$$\hat{\beta}_{j\text{IFM}} = \operatorname{argmax}_{\beta_j} \sum_{i=1}^n \log f_i(X_{ij}; \beta_j).$$

For only Standard normal distribution, we use inference functions for margins method in this paper. An n -dimensional vector of random variables $\mathbf{Z}_{n \times 1} = (Z_1, Z_2, \dots, Z_n)^T$, $-\infty < Z_i < \infty$, $i = 1, \dots, n$, is said to have a multivariate normal distribution if its density function $f(\mathbf{Z})$ is of the form

$$\begin{aligned} f(\mathbf{Z}) &= f(Z_1, Z_2, \dots, Z_n) \\ &= \left(\frac{1}{2\pi} \right)^{n/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{(\mathbf{Z} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{Z} - \boldsymbol{\mu})}{2} \right\} \end{aligned}$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)^T$ is the vector of means and Σ is the variance–covariance matrix of the multivariate normal distribution.

When $n = 2$, $\mathbf{Z}_{2 \times 1} = (Z_1, Z_2)^T$ has the bivariate normal distribution with a two-dimensional vector of means, $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

Assume an n -dimensional random vector $\mathbf{Z}_{n \times 1} = (\mathbf{Z}_1, \mathbf{Z}_2)^T$ has a normal distribution $\mathbf{N}(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)^T$ and

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix}$$

where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are two subvectors of respective dimensions p and q with $p + q = n$.

Theorem 2. The marginal distributions of $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are also normal with mean vector $\boldsymbol{\mu}_i$ and covariance matrix Σ_{ii} where $i = 1, 2$, respectively. That is, the marginal distribution of $\mathbf{Z}_1$ is

$$\begin{aligned} f_1(\mathbf{z}_1) &= \int f(\mathbf{z}_1, \mathbf{z}_2) d\mathbf{z}_2 \\ &= \frac{1}{(2\pi)^{p/2} |\Sigma_{11}|^{1/2}} \exp \left\{ -\frac{(\mathbf{z}_1 - \boldsymbol{\mu}_1)^T \Sigma_{11}^{-1} (\mathbf{z}_1 - \boldsymbol{\mu}_1)}{2} \right\} \end{aligned}$$

and the marginal distribution of $\mathbf{Z}_2$ is

$$\begin{aligned} f_2(\mathbf{z}_2) &= \int f(\mathbf{z}_1, \mathbf{z}_2) d\mathbf{z}_1 \\ &= \frac{1}{(2\pi)^{q/2} |\Sigma_{22}|^{1/2}} \exp \left\{ -\frac{(\mathbf{z}_2 - \boldsymbol{\mu}_2)^T \Sigma_{22}^{-1} (\mathbf{z}_2 - \boldsymbol{\mu}_2)}{2} \right\}. \end{aligned}$$

Theorem 3. The conditional distribution of $\mathbf{Z}_1$ given $\mathbf{Z}_2$ is also normal with mean vector

$$\mathbf{v}_1 = \boldsymbol{\mu}_1 + \Sigma_{12} \Sigma_{22}^{-1} (\mathbf{z}_2 - \boldsymbol{\mu}_2)$$

and covariance matrix

$$\mathbf{Q}_1 = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}.$$

It follows that the conditional density function $f_{1|2}(\cdot|\mathbf{z}_2)$ of $\mathbf{Z}_1$, when $\mathbf{Z}_2 = \mathbf{z}_2$, is specified at the point $\mathbf{z}_1$ by the equation

$$\begin{aligned} f_{1|2}(\mathbf{z}_1|\mathbf{z}_2) &= \frac{f(\mathbf{z}_1, \mathbf{z}_2)}{f(\mathbf{z}_2)} \\ &= \left(\frac{1}{2\pi}\right)^{p/2} \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \exp\left\{-\frac{(\mathbf{z}_1 - \mathbf{v}_1)^T \mathbf{Q}_1^{-1}(\mathbf{z}_1 - \mathbf{v}_1)}{2}\right\}. \end{aligned}$$

The cumulative distribution function is

$$F_{1|2}(\mathbf{z}_1|\mathbf{z}_2) = \int_{-\infty}^{z_p} \cdots \int_{-\infty}^{z_1} f_{1|2}(\mathbf{x}_1|\mathbf{z}_2) dx_1 \cdots dx_p \quad (7)$$

where $\mathbf{z}_1 = (z_1, \dots, z_p)$ and $z_1, \dots, z_p \in \mathbf{R}$.

By using Eq. (7), we can derive the Gaussian conditional distributions, and then by using the CML method by Genest et al. (1995) and the IFM method by Joe (1997), we can estimate the Gaussian copula parameter, an n th order conditional correlation, $\rho_{YX|Z_1, Z_2, \dots, Z_n}$, using the following:

$$F_{Y|Z_1, \dots, Z_n}(Y, X|Z_1, Z_2, \dots, Z_n) = C^{Ga}(F_{X|Z_1, \dots, Z_n}(X|Z_1, Z_2, \dots, Z_n), F_{Y|Z_1, \dots, Z_n}(Y|Z_1, Z_2, \dots, Z_n); \rho_{YX|Z_1, Z_2, \dots, Z_n}). \quad (8)$$

Partial correlation coefficients in normal distributions are correlation coefficients in conditional distributions (Anderson, 2003). The partial correlations of $\mathbf{Z}_1$, when $\mathbf{Z}_2 = \mathbf{z}_2$ are the correlations calculated in the usual way from $\mathbf{Q}_1$. Joe (1997) proved the partial correlation coefficients in normal distributions are correlation coefficients in conditional distributions. The statement holds for higher dimensions (Yule and Kendall, 1965).

Usually, the estimation problems concerning a copula C arise when a conditional joint distribution H expressed as a function of the conditional margins F, G is

$$H(\mathbf{x}, \mathbf{y}|\mathbf{z}; \theta_1, \theta_2, \alpha) = C(F(\mathbf{x}|\mathbf{z}; \theta_1), G(\mathbf{y}|\mathbf{z}; \theta_2); \alpha), \quad (9)$$

and the parametric form of the margins is unknown. In such cases, traditional parametric methods such as maximization of the full likelihood

$$h(\mathbf{x}, \mathbf{y}|\mathbf{z}; \theta_1, \theta_2, \alpha) = c(F(\mathbf{x}|\mathbf{z}; \theta_1), G(\mathbf{y}|\mathbf{z}; \theta_2)|\alpha) f(\mathbf{x}|\mathbf{z}; \theta_1) g(\mathbf{y}|\mathbf{z}; \theta_2)$$

cannot be applied. Likelihood methods involve simultaneous maximization of the model parameters in the conditional joint distribution H . Under the assumption of known parametric families for the conditional margins, representation (9) allows for procedures where the estimation can be carried out separately for F, G and H . In this way the computational burden for high-dimensional problems can be minimized where the full likelihood approach may become very demanding.

The procedures to estimate parameters in this paper can be summarized by the following steps:

- Step 1 . Estimates of the parameters θ_1, θ_2 of the margins F, G are obtained by separately maximizing the corresponding likelihoods $f(\theta_1|\mathbf{x})$ and $g(\theta_2|\mathbf{y})$ of the univariate margins.
- Step 2 . Uses the empirical CDF of each marginal distribution to transform the observations into pseudo-observations with uniform margins and the estimator for the copula parameter α is the value that maximizes the function:

$$\hat{\alpha}_{\text{CML}} = \operatorname{argmax}_{\alpha} \sum_{i=1}^n \log c(\alpha; F'(x_i|z_i), G'(y_i|z_i))$$

where F' and G' are the scaled margins.

3.2. Simulated data for partial correlation comparison

To see how close the values of partial correlation by Gaussian copula are to the original dependence structure, we simulated the data sets in UNICORN, which is a stand-alone uncertainty analysis package (Kurowicka and Cooke, 2006). The main focus of UNICORN is dependence modeling for high-dimensional distributions. A number of dependence structures are available for coupling random variables: dependence trees, vines, and Bayesian belief nets. UNICORN provides a convenient route to model and simulate certain situations: UNICORN also has built-in, easy-to-use tools with which to analyze a simulation. In this paper, we will be using UNICORN to simulate data using D -vine dependence structure. The first thing that we have to do is enter the four variables which have a uniform distribution in $[0, 1]$ into UNICORN's Random Variables Views. UNICORN also has a dependence tab with which you can design your dependence structure; be it a D -vine, canonical vine, dependency tree, BBN, or just a simple correlation matrix. So we will be using a D -vine as the dependency structure. This is done by simply dragging all the variables onto the screen to create the base tree, and then clicking "transcend" which will create the rest of the tree; there will be small boxes to enter the (partial) correlations. At this point, all that needs to be done is to run the simulation and analyze the results. UNICORN provides a report that gives the basic statistics of the four variables (see Table 1). Also, UNICORN provides an interesting graphical tools; in particular, a cobweb plot. Using the simulated data set, we computed the Pearson's (partial) correlations by SAS 9.2, Gaussian copula (partial) correlations by IFM and CML(see Table 2). In an empirical study, we verified that Gaussian copula (partial) correlations are almost the same as Pearson's (partial) correlations.

Table 1

Summarized descriptive statistics of simulated data.

Sample	Mean	Median	Minimum	Maximum	St.D	Skewness	Kurtosis
U1	0.4886052	0.4910	0.00025	0.99975	0.2896216	3.220861e−44	1.024777e−57
U2	0.4828320	0.4545	0.00100	0.99900	0.2838753	2.626588e−43	1.682116e−56
U3	0.4831993	0.4780	0.00025	0.99900	0.2900595	−6.032139e−43	5.096771e−56
U4	0.4864660	0.4855	0.00025	0.99975	0.2877351	−2.726064e−44	8.204424e−58

Table 2

Partial correlation for *D*-vine using simulated data.

Type	r_{12}	r_{23}	r_{34}	$r_{13 2}$	$r_{24 3}$	$r_{14 23}$
Original	0.7	0.8	0.9	0.5	0.6	0.3
Pearson's	0.72	0.8	0.9	0.53	0.62	0.33
Gaussian IFM	0.72	0.8	0.9	0.52	0.6	0.33
Gaussian CML	0.74	0.81	0.91	0.53	0.62	0.33

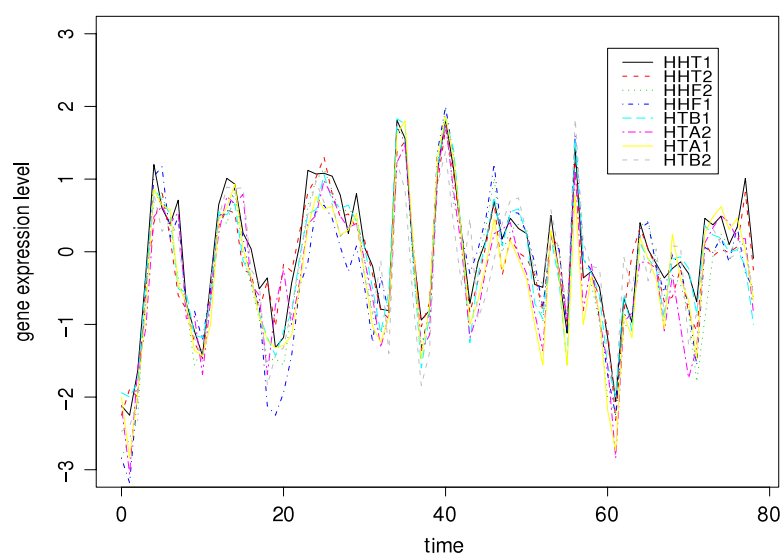


Fig. 3. Time-series plot of gene expressions in histone group.

4. Application to histone genes

In this subsection, we analyze expression data for genes involved in the cell cycle of yeast cells. The data set is composed of measurements on 6221 genes observed at 80 time points. 800 genes regulated by cell cycle were identified. Note that known interactions are still incomplete at present. Using this data, Kim et al. (2008) proposed a directional dependence of genes using asymmetric FGM copula. This study demonstrated pair gene interaction, but did not demonstrate a partial correlation dependence structure. Our present research is an extension of this work. The data include eight histone genes: HHT1 (1), HHT2 (2), HHT1 (3), HHT2 (4), HTA1 (5), HTA2 (6), HTB1 (7) and HTB2 (8). These eight genes encode the four histones (H2A, H2B, H3 and H4). Histones are proteins which “package” DNA into chromosomes. Chromosomes need to be replicated before cell division. Expression of the histone genes should be tightly regulated for the proper functioning of the replication process. Fig. 3 shows a time-series plot of genes in the histone group. It can be easily seen that expression of the eight genes in the histone group is highly correlated. We have used UNICORN (Kurowicka and Cooke, 2006) for plotting data on eight histone genes. Fig. 4 shows a cobweb graph illustrating how the eight genes—HHT1, HHT2, HHT1, HHT2, HTA1, HTA2, HTB1 and HTB2 are linked to each other like eight spiders spinning cobwebs. Canonical vine is more effective than *D*-vine when there is a particular variable that has more importance than the other variables. *D*-vine is more effective than canonical vine when no one variable appears to have greater importance in determining the other variables. Table 3 provides partial correlations using the Gaussian copula in canonical vine and Table 4 provides partial correlations using the Gaussian copula in *D*-vine. We investigated the partial correlations produced by SAS with partial correlations by the Gaussian copula for an empirical study. In this paper, we have used eight standard normal random variables, $X_1, X_2, X_3, X_4, X_5, X_6, X_7$, and X_8 transforming eight histone genes—HHT1 (1), HHT2 (2), HHT1 (3), HHT2 (4), HTA1 (5), HTA2 (6), HTB1 (7) and HTB2 (8). We can derive the conditional cumulative distribution of X_1, X_8 given $X_2, X_3, X_4, X_5, X_6, X_7$ as follows;

$$F_{18|234567}(X_1, X_8|X_2, X_3, X_4, X_5, X_6, X_7) = C^{Ga}(U_1, U_2; \rho_{18|234567})$$

Samples selected: 79

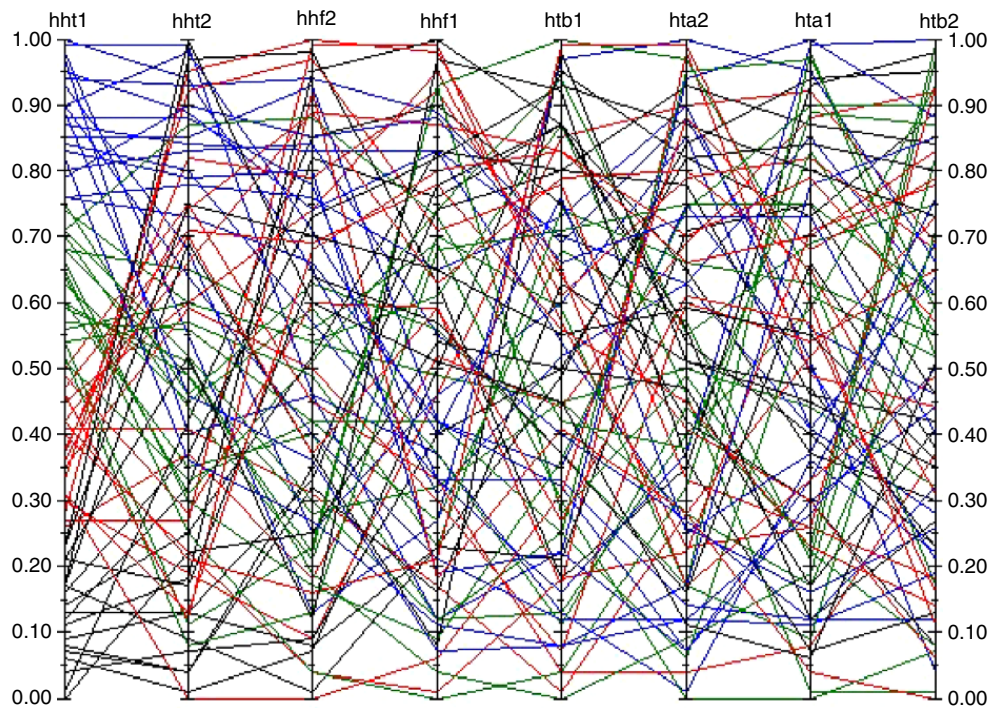


Fig. 4. 8 histone gene plot by Unicorn.

Table 3

Partial correlation table for canonical vine.

Type	$r_{23 1}$	$r_{24 1}$	$r_{25 1}$	$r_{26 1}$	$r_{27 1}$	$r_{28 1}$	$r_{34 12}$	$r_{35 12}$	$r_{36 12}$	$r_{37 12}$	$r_{38 12}$
Pearson's	0.44	0.01	0.31	0.30	0.23	0.02	0.73	0.51	0.19	0.39	0.39
Gaussian IFM	0.46	0.06	0.30	0.25	0.22	−0.02	0.61	0.43	0.11	0.31	0.26
Gaussian CML	0.49	0.09	0.36	0.30	0.24	0.01	0.66	0.50	0.12	0.35	0.31
Type	$r_{45 123}$	$r_{46 123}$	$r_{47 123}$	$r_{48 123}$	$r_{56 1234}$	$r_{57 1234}$	$r_{58 1234}$	$r_{67 12345}$	$r_{68 12345}$	$r_{78 123456}$	
Pearson's	0.09	0.21	0.18	0.28	0.14	0.44	−0.01	0.48	0.25	−0.26	
Gaussian IFM	0.13	0.29	0.17	0.34	0.26	0.40	0.08	0.34	0.22	−0.17	
Gaussian CML	0.13	0.29	0.19	0.29	0.19	0.36	0.03	0.37	0.27	−0.14	

Table 4

Partial correlation table for D-vine.

	$r_{13 2}$	$r_{24 3}$	$r_{35 4}$	$r_{46 5}$	$r_{57 6}$	$r_{68 7}$	$r_{14 23}$	$r_{25 34}$	$r_{36 45}$	$r_{47 56}$	$r_{58 67}$
Pearson's	0.57	−0.34	0.69	0.31	0.71	0.40	0.27	0.27	0.22	0.19	0.33
Gaussian IFM	0.48	−0.27	0.63	0.25	0.57	0.40	0.25	0.26	0.13	0.13	0.28
Gaussian CML	0.48	−0.25	0.66	0.27	0.61	0.41	0.27	0.31	0.14	0.15	0.29
	$r_{15 234}$	$r_{26 345}$	$r_{37 456}$	$r_{48 567}$	$r_{16 2345}$	$r_{27 3456}$	$r_{38 4567}$	$r_{17 23456}$	$r_{28 34567}$	$r_{18 234567}$	
Pearson's	0.29	0.36	0.08	0.46	0.23	0.02	0.15	0.18	−0.03	0.24	
Gaussian IFM	0.30	0.32	0.13	0.37	0.18	0.08	0.06	0.21	−0.05	0.28	
Gaussian CML	0.29	0.38	0.16	0.36	0.17	0.06	0.11	0.21	−0.04	0.26	

where $U_1 = F_{1|234567}(X_1|X_2, X_3, X_4, X_5, X_6, X_7)$ and $U_2 = F_{8|234567}(X_8|X_2, X_3, X_4, X_5, X_6, X_7)$. Using a Gaussian copula with two normal marginals, we can estimate a partial correlation coefficient parameter $\rho_{18|234567}$ by the IFM method and the CML method. The estimate of $\rho_{18|234567}$ is the partial correlation coefficient of X_1 and X_8 given X_2, X_3, X_4, X_5, X_6 , and X_7 , $r_{18|234567}$.

In Table 3, we chose HHT1 as an important variable among the eight histone genes because HHT1 had high correlation with seven other histone genes compared with other cases. Based on the gene network shown in Fig. 2, built for the eight histone genes by Chen et al. (2005), we investigate the partial correlations as follows: HHT2 and HHF1 given HHT1 is 0.46 (Gaussian copula IFM) and 0.49 (Gaussian copula CML), and HHF1 and HHF2 given HHT1 and HHT2 is 0.61 (Gaussian copula IFM) and 0.66 (Gaussian copula CML), and HHA1 and HTB1 given HHT1, HHT2, HHF1 and HHF2 is 0.40 (Gaussian copula IFM) and 0.36 (Gaussian copula CML), and HTA2 and HTB1 given HHT1, HHT2, HHF1, HHF2 and HHA1 is 0.34 (Gaussian copula IFM) and 0.37 (Gaussian copula CML). The gene network of Chen et al. (2005) captured 86% of all the currently reported direct

interactions among these genes. Therefore, these values of the Gaussian copula partial correlation based on canonical vine can be helpful to uncover the gene network.

In addition, the values of partial correlation based on *D*-vine are shown in Table 4. HHF1 and HTA1 given HHF2 is 0.63 (Gaussian copula IFM) and 0.66 (Gaussian copula CML), HTA1 and HTB1 given HTA2 is 0.57 (Gaussian copula IFM) and 0.61 (Gaussian copula CML), HHT2 and HTA2 given HHF1, HHF2 and HTA1 is 0.32 (Gaussian copula IFM) and 0.38 (Gaussian copula CML), HHF2 and HTB2 given HTA1, HTA2 and HTB1 is 0.37 (Gaussian copula IFM) and 0.36 (Gaussian copula CML), HHT1 and HTB2 given HHT2, HHF1, HHF2, HTA1, HTA2 and HTB1 is 0.32 (Gaussian copula IFM) and 0.38 (Gaussian copula CML). With these values of the Gaussian copula partial correlation based on canonical vine and *D*-vine, we may capture all the currently reported direct interactions among these genes.

5. Conclusion

Partial correlation by Gaussian copula has an advantage in terms of simple computation procedures compared to recursive Pearson's partial correlation procedure while the values of partial correlations by Gaussian copula have the same values as those resulting for Pearson's partial correlation. In particular, when we have multivariate data such as that in a gene network, it is not easy to compute the partial correlation by Pearson's partial correlation. But the simple computation procedure built by our Gaussian copula partial correlation is useful in understanding and modeling dependent structures for random variables and possibly in measuring the partial correlation of genes. In our future work, we will investigate dependence properties and measures of association between two or more variables in terms of various copulas. In addition, we will study a model-based mixture pair-copula by using archimedean copula function for various gene data sets.

Acknowledgements

The authors are thankful to the Editor, Associate Editor and the two referees. We will also like to express our gratitude to Andrew Wey for his work on the project, funded by the Morris Academic Partnership (MAP) program of the University of Minnesota at Morris.

References

- Aasa, K., Czado, C., Frigessic, A., Bakken, H., 2009. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics* 44 (2), 182–198.
- Anderson, T.W., 2003. An Introduction to Multivariate Statistical Analysis, third ed. In: Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, NJ.
- Bedford, T., Cooke, R.M., 2002. Vines—a new graphical model for dependent random variables. *Annals of Statistics* 30 (4), 1031–1068.
- Bergsma, W.P., 2004. Testing conditional independence for continuous random variables. PASCAL Project. EURANDOM, Eindhoven.
- Chen, K.C., Wang, T.Y., Tseng, H.H., Huang, C.Y.F., Kao, C.Y., 2005. A stochastic differential equation model for quantifying transcriptional regulatory network in *Saccharomyces cerevisiae*. *Bioinformatics* 21, 2883–2890.
- Genest, C., Ghoudi, K., Rivest, L.P., 1995. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* 82 (3), 543–552.
- Joe, H., 1997. Multivariate Models and Dependence Concepts. Chapman & Hall, London.
- Kim, J.-M., Jung, Y., Sungur, E.A., Han, K., Park, C., Sohn, I., 2008. A copula method for modeling directional dependence of genes. *BMC Bioinformatics* 9, 225.
- Kurowicka, D., Cooke, R., 2006. Uncertainty Analysis with High Dimensional Dependence Modelling. In: Wiley Series in Probability and Statistics, John Wiley and Sons, Ltd.
- Matteis, R.D., 2001. Fitting copulas to data. Ph.D. Thesis. Institute of Mathematics of the University of Zurich.
- McLeish, D.L., Small, C.G., 1988. The Theory and Applications of Statistical Inference Functions. In: Lecture Notes in Statistics, vol. 44. Springer-Verlag, New York.
- Nelson, R.B., 2006. An Introduction to Copulas. Springer-Verlag, New York.
- Pearson, K., 1916. On some novel properties of partial and multiple correlation coefficients in a universe of manifold characteristics. *Biometrika* 11 (3), 231–238.
- Sklar, A., 1959. Fonctions de repartition a *n* dimensions et leurs marges. Publications de l'Institut de Statistique de L'Université de Paris 8, 229–231 (in French).
- Sklar, A., 1973. Random variables, joint distribution functions, and copulas. *Kybernetika (Prague)* 9, 449–460.
- Xu, J.J., 1996. Statistical modeling and inference for multivariate and longitudinal discrete response data. Ph.D. Thesis. Department of Statistics, University of British Columbia.
- Yan, J., 2007. Enjoy the joy of copulas: with a package copula. *Journal of Statistical Software* 21 (4).
- Yule, G.U., Kendall, M.G., 1965. An Introduction to the Theory of Statistics, 14th ed. Charles Griffin & Co., Belmont, California.