



A circular-linear dependence measure under Johnson–Wehrly distributions and its application in Bayesian networks

Ignacio Leguey^{a,b,*}, Pedro Larrañaga^b, Concha Bielza^b, Shogo Kato^c

^a *Economía Financiera y Contabilidad e Idioma Moderno department, Facultad de Ciencias Jurídicas y Sociales, Universidad Rey Juan Carlos de Madrid, Spain*

^b *Artificial Intelligence department, Universidad Politécnica de Madrid, Campus de Montegancedo, 28660 Boadilla del Monte, Madrid, Spain*

^c *Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan*

ARTICLE INFO

Article history:

Received 17 August 2017

Revised 17 January 2019

Accepted 21 January 2019

Available online 21 February 2019

2010 MSC:

62H11

94A17

Keywords:

Circular-linear mutual information

Tree-structured Bayesian network

Dependence measures

Directional statistics

ABSTRACT

Circular data jointly observed with linear data are common in various disciplines. Since circular data require different techniques than linear data, it is often misleading to use usual dependence measures for joint data of circular and linear observations. Moreover, although a mutual information measure between circular variables exists, the measure has drawbacks in that it is defined only for a bivariate extension of the wrapped Cauchy distribution and has to be approximated using numerical methods. In this paper, we introduce two measures of dependence, namely, (i) circular-linear mutual information as a measure of dependence between circular and linear variables and (ii) circular-circular mutual information as a measure of dependence between two circular variables. It is shown that the expression for the proposed circular-linear mutual information can be greatly simplified for a subfamily of Johnson–Wehrly distributions. We apply these two dependence measures to learn a circular-linear tree-structured Bayesian network that combines circular and linear variables. To illustrate and evaluate our proposal, we perform experiments with simulated data. We also use a real meteorological data set from different European stations to create a circular-linear tree-structured Bayesian network model.

© 2019 Published by Elsevier Inc.

1. Introduction

Data comprising circular and linear observations are ubiquitous in science. In analysis of such data, the periodic nature of circular data has often been ignored in the literature, and the data have been treated as being linear. However, circular data must be analysed using special circular techniques, owing to their properties [6,12,22,24,26]. Unlike linear data, in the circular domain, the points with module 2π are considered to be the same point (e.g., -360 , 0 , 360 , 720 , etc.). One of the best-known circular distributions is the wrapped Cauchy distribution [21], which was extended to a five-parameter bivariate distribution for toroidal data by Kato and Pewsey [16]. Another popular distribution on the circle is the von Mises distribution [30], which has been extended to the bivariate case [24] and to the multivariate case by Mardia et al. [25] and Mardia and Voss [28].

* Corresponding author at: Economía Financiera y Contabilidad e Idioma Moderno department Facultad de Ciencias Jurídicas y Sociales Universidad Rey Juan Carlos de Madrid, Spain.

E-mail addresses: ignacio.vitoriano@urjc.es (I. Leguey), pedro.larranaga@fi.upm.es (P. Larrañaga), mcbielza@fi.upm.es (C. Bielza), skato@ism.ac.jp (S. Kato).

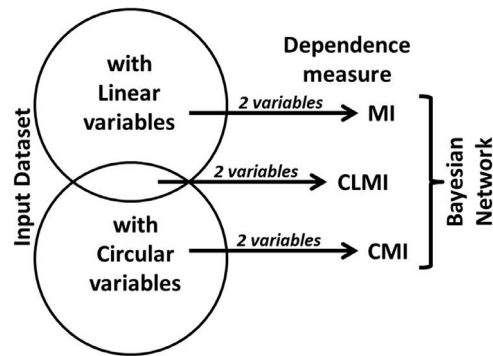


Fig. 1. Flow chart for the choice of dependence measures. MI for two linear variables, CMI for two circular variables, CLMI for a pair of linear and circular variables. These measures allow the modelling of a tree-shaped Bayesian network with circular and/or linear variables.

Several models exist for data consisting of circular and linear observations, many of which focus on a circular-linear regression [7,10,36]. In addition, Mardia and Sutton [27] proposed a bivariate distribution that combines the von Mises distribution with Gaussian distributions on the cylinder. Abe and Ley [1] proposed the WeiSSVM, a cylindrical distribution based on combinations of the sine-skewed von Mises distribution and the Weibull distribution. Furthermore, Johnson and Wehrly [14] presented circular-linear distributions, and proposed a method to obtain a bivariate circular-linear distribution with specified marginal distributions. In this work, we prove that the conditional distributions of a subfamily of Johnson-Wehrly family are well known and mathematically tractable.

Many studies have determined measures for the mutual dependence between linear variables [See 23, 33, 34, among others]. Among them, one of the most well-known is the mutual information measure [4,38]. The latter measure is based on the similarity between the joint density function and the product of its marginal density functions. To extend this measure to the circular domain, Leguey et al. [20] developed a circular-circular mutual-information (CMI) measure. However, this measure is only applicable to circular variables with marginal distributions that follow wrapped Cauchy distributions and has to be approximated using numerical methods. Therefore, for the case when the two variables are in the circular domain, we propose a CMI measure with no constraints on the underlying circular distributions. The CMI measure is shown to have the nice property that it can be expressed in a closed form for a general family of bivariate distributions.

To the best of our knowledge, there are no measures of mutual information for pairs of circular and linear variables. Such a measure is potentially useful for learning Bayesian networks such as the tree-augmented naive Bayes model (TAN) [9], where the dependence between the variables is captured via probabilistic graphical models. Furthermore, the Chow-Liu algorithm [3], which is used to learn a TAN and based on the mutual information of all pairs of variables, guarantees that the graph is a maximum likelihood tree. Therefore, a measure of mutual information for a pair of circular and linear variables is necessary to allow the presence of linear and circular variables in such models. We address this problem and propose a circular-linear mutual-information measure (CLMI) of the dependence between a circular variable and a linear one.

Generally, Bayesian network models that combine linear and circular variables ignore the circular nature and consider all of them as linear-continuous, such as Gaussian Bayesian networks [see 17, 37, among others]. Alternatively, it is also common to discretize [32] the variables (with the related loss of information) to take advantage of a plethora of learning and inference algorithms for discrete Bayesian network models. In both models, the circular variable is not appropriately handled. In this work, we develop a circular-linear Bayesian network model with a tree structure that captures the relationship between circular and linear variables. The model is based on our proposed CLMI and CMI measures (Fig. 1), together with the traditional mutual information of linear variables and the bivariate distribution proposed by Johnson and Wehrly [14].

The paper is organized as follows. Section 2 reviews the wrapped Cauchy distribution and the Gaussian distribution as representatives on the circular and linear domains, respectively. Section 3 reviews the angular-linear Johnson-Wehrly bivariate distribution, and shows that a submodel of their distribution has tractable and well-known conditional distributions. Section 4 discusses the proposed CMI and CLMI measures. Section 5 applies the measures presented in Section 4, and presents the proposed circular-linear tree-structured Bayesian network model, as well as its evaluation in synthetic domains. Section 6 compares the proposed model to a Gaussian Bayesian network model and a discrete Bayesian network model over a real-world meteorological data set recorded from meteorological stations located in Europe. Last, Section 7 concludes the paper and discusses possible avenues for future work.

2. Representative distributions for circular data and for linear data

One of the best-known distributions defined on the circle is the wrapped Cauchy distribution [21]. We use this circular distribution as the underlying marginal model for the circular data in the proposed Bayesian networks. Similarly, we use the Gaussian distribution as the underlying distribution for linear data and apply it to our Bayesian networks.

These two distributions are both well known, and share some nice properties. For example, they are both easy to simulate, computationally very fast and have tractable bivariate extensions. In addition, the marginal and conditional distributions of these extensions belong to the same family and are mathematically tractable.

In the following subsections, we review the wrapped Cauchy distribution, Gaussian distribution, and their bivariate extensions. As known properties related to the bivariate extensions, we discuss their marginal and conditional distributions, as well as other characteristics relevant to our study. These include the complex-form expression for the wrapped Cauchy distribution and the expression for the mutual information between two variables with a bivariate Gaussian density.

2.1. Wrapped cauchy distributions

2.1.1. Definitions

A circular random variable Θ that follows a wrapped Cauchy distribution [21], denoted by $wC(\mu, \varepsilon)$, has the following density function:

$$f(\theta; \mu, \varepsilon) = \frac{1}{2\pi} \frac{1 - \varepsilon^2}{1 + \varepsilon^2 - 2\varepsilon \cos(\theta - \mu)}, \quad \text{with } \theta, \mu \in (0, 2\pi], \varepsilon \in [0, 1), \tag{1}$$

where μ is the location parameter and ε is the concentration parameter. When $\varepsilon = 0$, Eq. (1) is the circular uniform distribution; otherwise, it is unimodal and symmetric about μ . Please refer to [18] and [29] for other properties of wC .

A random vector (Θ_1, Θ_2) that follows the five-parameter circular bivariate wrapped Cauchy distribution [16], denoted by $bwC(\mu_1, \mu_2, \varepsilon_1, \varepsilon_2, \rho)$, has the density function

$$f(\theta_1, \theta_2) = c[c_0 - c_1 \cos(\theta_1 - \mu_1) - c_2 \cos(\theta_2 - \mu_2) - c_3 \cos(\theta_1 - \mu_1) \cdot \cos(\theta_2 - \mu_2) - c_4 \sin(\theta_1 - \mu_1) \sin(\theta_2 - \mu_2)]^{-1}, \quad \text{with } \theta_1, \theta_2 \in (0, 2\pi], \tag{2}$$

where $c = (1 - \rho^2)(1 - \varepsilon_1^2)(1 - \varepsilon_2^2)/4\pi^2$, $c_0 = (1 + \rho^2)(1 + \varepsilon_1^2)(1 + \varepsilon_2^2) - 8|\rho|\varepsilon_1\varepsilon_2$, $c_1 = 2(1 + \rho^2)\varepsilon_1(1 + \varepsilon_2^2) - 4|\rho|(1 + \varepsilon_1^2)\varepsilon_2$, $c_2 = 2(1 + \rho^2)(1 + \varepsilon_1^2)\varepsilon_2 - 4|\rho|\varepsilon_1(1 + \varepsilon_2^2)$, $c_3 = -4(1 + \rho^2)\varepsilon_1\varepsilon_2 + 2|\rho|(1 + \varepsilon_1^2)(1 + \varepsilon_2^2)$, $c_4 = 2\rho(1 - \varepsilon_1^2)(1 - \varepsilon_2^2)$, $\mu_1, \mu_2 \in (0, 2\pi]$, $\varepsilon_1, \varepsilon_2 \in [0, 1)$, and $\rho \in (-1, 1)$. Here, μ_1 and μ_2 are the marginal location parameters, ε_1 and ε_2 are the marginal concentration parameters, and ρ is the parameter controlling the association between Θ_1 and Θ_2 .

2.1.2. Complex form

Using the complex form to represent univariate and bivariate wrapped Cauchy models can simplify their computation issues [16,29].

Let $Z = \exp(i\Theta)$, where Θ is distributed as the univariate wrapped Cauchy given by Eq. (1). Then, Z has a density function given by

$$f(z) = \frac{1}{2\pi} \frac{|1 - |\lambda|^2|}{|z - \lambda|^2}, \quad z \in \Omega, \lambda \in \hat{\mathbb{C}} \setminus \Omega, \tag{3}$$

where $\lambda = \varepsilon \exp(i\mu)$, $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, and $\Omega = \{z \in \mathbb{C} : |z| = 1\}$. Similar to [29], we denote Z distributed as in Eq. (3) as $Z \sim C^*(\lambda)$.

Let $(Z_1, Z_2) = (\exp(i\Theta_1), \exp(i\Theta_2))$, where (Θ_1, Θ_2) is distributed as the bivariate wrapped Cauchy in Eq. (2). Then, the density function of (Z_1, Z_2) is

$$f(z_1, z_2) = \frac{(4\pi^2)^{-1}(1 - \rho^2)(1 - \varepsilon_1^2)(1 - \varepsilon_2^2)}{|a_{11}(\bar{z}_1\eta_1)^q z_2 \bar{\eta}_2 + a_{12}(\bar{z}_1\eta_1)^q + a_{21}z_2 \bar{\eta}_2 + a_{22}|^2}, \quad z_1, z_2 \in \Omega, \tag{4}$$

where $q \in \{-1, 1\}$ is the sign of ρ , $\eta_k = \exp(i\mu_k) \in \Omega$ with $k \in \{1, 2\}$, \bar{z}_k is the complex conjugate of z_k , $\varepsilon_1, \varepsilon_2 \in [0, 1)$, $\rho \in (-1, 1)$, and

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} \varepsilon_1\varepsilon_2 - |\rho| & |\rho|\varepsilon_2 - \varepsilon_1 \\ |\rho|\varepsilon_1 - \varepsilon_2 & 1 - |\rho|\varepsilon_1\varepsilon_2 \end{pmatrix}. \tag{5}$$

We denote (Z_1, Z_2) distributed as in Eq. (4) as $(Z_1, Z_2) \sim bC^*(\eta_1, \eta_2, \varepsilon_1, \varepsilon_2, \rho)$.

2.1.3. Marginals and conditionals

Theorem 1. (Kato and Pewsey [16]) A random vector (Z_1, Z_2) with density as in Eq. (4) has marginals $Z_1 \sim C^*(\varepsilon_1\eta_1)$ and $Z_2 \sim C^*(\varepsilon_2\eta_2)$, and conditionals $Z_1|Z_2 = z_2 \sim C^*(-\eta_1[\mathbf{A} \circ (z_2\bar{\eta}_2)^q])$ and $Z_2|Z_1 = z_1 \sim C^*(-\eta_2[\mathbf{A}^T \circ (z_1\bar{\eta}_1)^q])$, where \mathbf{A} is defined as in Eq. (5), \mathbf{A}^T is the transpose of \mathbf{A} , and

$$\mathbf{A} \circ z = \frac{a_{11}z + a_{12}}{a_{21}z + a_{22}}.$$

2.1.4. Parameter estimation

For the bivariate wrapped Cauchy distribution, the method of moments [2] for the parameter estimation has several advantages over the method of maximum likelihood [16]. The method of moments for the bivariate wrapped Cauchy distribution is computationally very fast and easier to implement. In addition, the formulae for the parameter estimates can all be expressed in a closed form, and all estimates are always within a range.

Let $\{(\theta_{1j}, \theta_{2j}) \mid j = 1, \dots, N\}$ be a random sample from a $bwC(\mu_1, \mu_2, \varepsilon_1, \varepsilon_2, \rho)$ (Eq. (2)). Then, the estimators obtained using the method of moments for $\mu_1, \mu_2, \varepsilon_1, \varepsilon_2$ and ρ are [16]

$$\hat{\mu}_k = \arg(\bar{R}_k), \quad \hat{\varepsilon}_k = |\bar{R}_k|, \quad k = 1, 2,$$

with $\bar{R}_k = \frac{1}{N} \sum_{j=1}^N \exp(i\theta_{kj})$, and

$$\hat{\rho} = \frac{1}{N} \left(\left| \sum_{j=1}^N \exp(i(\Phi_{1j} - \Phi_{2j})) \right| - \left| \sum_{j=1}^N \exp(i(\Phi_{1j} + \Phi_{2j})) \right| \right), \tag{6}$$

where $\Phi_{rj} = 2 \arctan \left(\frac{1+\hat{\varepsilon}_r}{1-\hat{\varepsilon}_r} \tan \left(\frac{\theta_{rj} - \hat{\mu}_r}{2} \right) \right)$, $r = 1, 2$.

2.1.5. Simulation

The simulation of data from a univariate wrapped Cauchy distribution can be performed easily. An efficient method is to use the Möbius transformation over a circular uniform distribution, $U(0, 2\pi)$, as explained in [29]. There is also another method, already implemented in the “Circular” R package [40]. This method consists of wrapping (i.e., $\text{mod } 360$) from the simulation of a (linear) Cauchy distribution with location parameter μ and scale parameter $-\log(\varepsilon)$. If $\varepsilon = 1$, the distribution is a point mass at μ , whereas if $\varepsilon = 0$, the wrapped Cauchy distribution is distributed as the circular uniform, $U(0, 2\pi)$.

The bivariate wrapped Cauchy distribution has (univariate) wrapped Cauchy marginals and conditionals. Therefore, the random variates from the bivariate wrapped Cauchy can be generated from two circular uniform $U(0, 2\pi)$ random variates by using the relationship $f(z_1, z_2) = f(z_1|z_2)f(z_2)$ (see Theorem 1). Here, once $f(z_2)$ is simulated, it is possible to simulate $f(z_1|z_2)$ and, hence, to obtain $f(z_1, z_2)$.

2.2. Gaussian distributions

2.2.1. Definitions

The Gaussian distribution is widely known. Please refer to [13,19,41] for good reviews on this topic.

In this study, we denote a linear random variable X that follows a Gaussian distribution as $N(\iota, \sigma)$. Here, $\iota \in \mathbb{R}$ is the mean parameter and $\sigma \in (0, \infty)$ is the variance parameter.

Similarly, we denote a random vector (X_1, X_2) that follows a bivariate Gaussian distribution as $bN(\iota_1, \iota_2, \sigma_1, \sigma_2, \gamma)$. In this case, $\iota_1, \iota_2 \in (-\infty, \infty)$ are the mean parameters of the marginals, and $\sigma_1, \sigma_2 \in (0, \infty)$ are the variance parameters of the marginal. Then, $\gamma \in [-1, 1]$ is the correlation between X_1 and X_2 .

2.2.2. Marginals and conditionals

A random vector (X_1, X_2) that follows $bN(\iota_1, \iota_2, \sigma_1, \sigma_2, \gamma)$ has marginals $X_1 \sim N(\iota_1, \sigma_1)$ and $X_2 \sim N(\iota_2, \sigma_2)$ and conditionals $X_1|X_2 = x_2 \sim N(\iota_1 + \frac{\sigma_1}{\sigma_2}\gamma(x_2 - \iota_2), (1 - \gamma^2)\sigma_1^2)$ and $X_2|X_1 = x_1 \sim N(\iota_2 + \frac{\sigma_2}{\sigma_1}\gamma(x_1 - \iota_1), (1 - \gamma^2)\sigma_2^2)$.

2.2.3. Parameter estimation

Let $\{(x_{1i}, x_{2i}) \mid i = 1, \dots, N\}$ be a random sample from $bN(\iota_1, \iota_2, \sigma_1, \sigma_2, \gamma)$.

For Gaussian distributions, the estimates from the maximum likelihood method and the method of moments of the parameters $\iota_1, \iota_2, \sigma_1^2, \sigma_2^2$ and γ are the same, and given by

$$\hat{\iota}_k = \frac{1}{N} \sum_{i=1}^N x_{ki}, \quad \hat{\sigma}_k = \sqrt{\hat{\sigma}_k^2}, \quad k = 1, 2,$$

with $\hat{\sigma}_k^2 = \frac{1}{N} \sum_{i=1}^N (x_{ki} - \hat{\iota}_k)^2$ and

$$\hat{\gamma} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1^2 \hat{\sigma}_2^2}, \tag{7}$$

where $\hat{\sigma}_{12} = \frac{1}{N} \sum_{i=1}^n (x_{1i} - \hat{\iota}_1)(x_{2i} - \hat{\iota}_2)$.

2.2.4. Mutual information

Let X_1 and X_2 be Gaussian variables. Then, the mutual information (MI) between X_1 and X_2 is as follows [4]:

$$MI(X_1, X_2) = -\frac{1}{2} \log(1 - \gamma^2),$$

where γ is the correlation coefficient between X_1 and X_2 , as defined in Eq. (7).

3. Circular-linear distribution of Johnson and Wehrly

In this section, we review the method proposed by Johnson and Wehrly [14] to obtain angular-linear bivariate distributions with arbitrary marginal distributions.

After reviewing the general form of the Johnson–Wehrly angular-linear distribution, we propose a subfamily of their model in which the conditionals and marginals are mathematically tractable.

3.1. Definition

Let $f_{\Theta}(\theta)$ and $f_X(x)$ be probability density functions on the circle and on the line, respectively. Suppose that $F_{\Theta}(\theta)$ is the cumulative distribution function (CDF) of $f_{\Theta}(\theta)$ defined with respect to a fixed, arbitrary origin and that $F_X(x)$ is the CDF of $f_X(x)$. Then, the distribution of Johnson and Wehrly [14] is defined by the density

$$f(\theta, x) = 2\pi g(2\pi F_{\Theta}(\theta) - 2\pi qF_X(x))f_{\Theta}(\theta)f_X(x), \tag{8}$$

where $0 \leq \theta < 2\pi$, $x \in \mathbb{R}$, $q \in \{-1, 1\}$ decides the positive or negative association between the two variables, and $g(\cdot)$ is a density on the circle.

Let a random vector (Θ, X) have the density given by Eq. (8). Then, the marginal distribution of Θ has the density $f_{\Theta}(\theta)$ and CDF $F_{\Theta}(\theta)$, while the marginal distribution of X has the density $f_X(x)$ and CDF $F_X(x)$. However, the conditional distributions are not tractable in general.

3.2. Conditionals

Let a random vector (Θ, X) follow the distribution (8). Then, changing the variables $U = 2\pi F_{\Theta}(\Theta)$ and $V = 2\pi F_X(X)$, the density function of (U, V) can be expressed as

$$f(u, v) = \frac{1}{2\pi} g(u - qv),$$

where $g(\cdot)$ is a density on the circle and $q \in \{-1, 1\}$.

We propose the following subfamily of the family given by Eq. (8) which has tractable conditional distributions.

Theorem 2. Let a random vector (Θ, X) follow the distribution given by Eq. (8) with $g(\cdot)$ being the wrapped Cauchy density given by Eq. (1) with location parameter μ_g and concentration parameter ε_g . Assume that $U = 2\pi F_{\Theta}(\Theta)$ and $V = 2\pi F_X(X)$. Then,

$$U|X = x \sim \text{wC}(2\pi qF_X(x) + \mu_g, \varepsilon_g) \quad \text{and} \quad V|\Theta = \theta \sim \text{wC}(q(2\pi F_{\Theta}(\theta) - \mu_g), \varepsilon_g).$$

In particular, if $\Theta \sim \text{wC}(\mu_{\theta}, \varepsilon_{\theta})$, $X \sim N(\mu_x, \sigma_x^2)$ and $F_{\Theta}(\theta) = \int_0^t f_{\Theta}(t) dt$, then, it holds that

$$\Theta|X = x \sim \text{wC}(\mu, \varepsilon),$$

where $\mu = \arg(\hat{\phi}_{\theta|x})$, $\varepsilon = |\hat{\phi}_{\theta|x}|$,

$$\hat{\phi}_{\theta|x} = \frac{\varepsilon_g \exp(i(2\pi qF_X(x) + \mu_g - v)) + \varepsilon_{\theta} \exp(i\mu_{\theta})}{1 + \varepsilon_g \varepsilon_{\theta} \exp(i(2\pi qF_X(x) + \mu_g - \mu_{\theta} - v))},$$

and $v = \arg\{(1 - \varepsilon_{\theta} \exp(-i\mu_{\theta})) / (1 - \varepsilon_{\theta} \exp(i\mu_{\theta}))\}$.

Proof. The conditional density function of Θ given $X = x$ and that of X given $\Theta = \theta$ can be expressed as

$$f(\theta|x) = \frac{f(\theta, x)}{f_X(x)}$$

and

$$f(x|\theta) = \frac{f(\theta, x)}{f_{\Theta}(\theta)},$$

respectively. Changing the variable $U = 2\pi F_{\Theta}(\Theta)$ in $f(\theta|x)$, we obtain

$$f(u|x) = f(\theta|x) \left| \frac{\partial \theta}{\partial u} \right| = g(u - 2\pi qF_X(x)),$$

where $q \in \{-1, 1\}$. Similarly, changing the variable $V = 2\pi F_X(X)$ in $f(x|\theta)$ leads to

$$f(v|\theta) = f(x|\theta) \left| \frac{\partial x}{\partial v} \right| = g(2\pi F_{\Theta}(\theta) - qv).$$

Let $g(\cdot)$ be the wrapped Cauchy density function given by Eq. (1), with location parameter μ_g and concentration parameter ε_g , as defined in Kato [15]. Then, the following hold:

$$U|X = x \sim \text{wC}(2\pi qF_X(x) + \mu_g, \varepsilon_g)$$

and

$$V|\Theta = \theta \sim \text{wC}(q(2\pi F_\Theta(\theta) - \mu_g), \varepsilon_g). \tag{9}$$

Consider the case where $\Theta \sim \text{wC}(\mu_\theta, \varepsilon_\theta)$ and $X \sim N(\iota_x, \sigma_x^2)$. Without loss of generality, the origin of the cumulative distribution function of Θ is assumed to be zero (i.e., $F_\Theta(\theta) = \int_0^\theta f_\Theta(t)dt$).

The conditional of Θ given $X = x$ has a wrapped Cauchy distribution. To see this, we first note that U and Θ have the following relationship:

$$\exp(iU) = \exp(iv) \frac{\exp(i\Theta) - \varepsilon_\theta \exp(i\mu_\theta)}{1 - \varepsilon_\theta \exp(i(\Theta - \mu_\theta))}$$

or

$$\exp(i\Theta) = \frac{\exp(i(U - v)) + \varepsilon_\theta \exp(i\mu_\theta)}{1 + \varepsilon_\theta \exp(i(U - v - \mu_\theta))},$$

where $\exp(iv) = (1 - \varepsilon_\theta \exp(-i\mu_\theta))/(1 - \varepsilon_\theta \exp(i\mu_\theta))$. McCullagh [29] showed that if $\exp(iU) \sim C^*(\alpha_1 \exp(i\beta_1))$, then

$$\frac{\exp(iU) + \alpha_2 \exp(i\beta_2)}{1 + \alpha_2 \exp(i(U - \beta_2))} \sim C^*\left(\frac{\alpha_1 \exp(i\beta_1) + \alpha_2 \exp(i\beta_2)}{1 + \alpha_1 \alpha_2 \exp(i(\beta_1 - \beta_2))}\right),$$

where $0 \leq \alpha_1, \alpha_2 < 1$ and $0 < \beta_1, \beta_2 \leq 2\pi$. Using this result, we have

$$\exp(i\Theta)|X = x \sim C^*(\hat{\phi}_{\theta|x})$$

or, in polar-coordinate form,

$$\Theta|X = x \sim \text{wC}(\arg(\hat{\phi}_{\theta|x}), |\hat{\phi}_{\theta|x}|),$$

where

$$\hat{\phi}_{\theta|x} = \frac{\varepsilon_g \exp(i(2\pi qF_X(x) + \mu_g - v)) + \varepsilon_\theta \exp(i\mu_\theta)}{1 + \varepsilon_g \varepsilon_\theta \exp(i(2\pi qF_X(x) + \mu_g - \mu_\theta - v))}.$$

□

For the rest of this subsection, we consider a subfamily presented in Theorem 2 in which the marginal of the circular variable has the wrapped Cauchy and that of the linear variable has the Gaussian.

Therefore, the conditional of the circular variable given the linear variable follows a known and tractable distribution (i.e., a wrapped Cauchy distribution (Eq. (1))). It follows from Theorem 2 that the conditional $X|\Theta = \theta$ itself does not follow any well-known distribution. However, Theorem 2 implies that

$$2\pi \Phi\left(\frac{X - \iota_x}{\sigma_x}\right)|\Theta = \theta \sim \text{wC}(q(2\pi F_\Theta(\theta) - \mu_g), \varepsilon_g),$$

where Φ denotes the CDF of the standard Gaussian distribution $N(0, 1)$, namely, $\Phi(x) = \int_{-\infty}^x \phi(t)dt$, where ϕ is the Gaussian density with $\iota_x = 0$ and $\sigma_x = 1$. Since it is easy to evaluate the CDF of the standard Gaussian distribution numerically, numerical calculations associated with the conditional distribution of X given $\Theta = \theta$ can be conducted efficiently.

4. General circular-circular and circular-linear mutual information

Mutual dependence measures between two linear variables have been studied at length, including the works of [23,34], among many others. In the case of linear data, one of the best-known measures is mutual information (MI) [4,38], which determines the similarity between the joint density and the product of its marginal densities. For circular data, the first mutual-information measure was developed recently by [20] and is called CMI. However, CMI is defined only for a bivariate wrapped Cauchy variable.

In this section, we redefine the CMI such that the measure can be used for any bivariate circular variables. Then, we present a closed-form expression for the CMI for a general family of bivariate circular distributions. This study is also the first to propose a mutual-information measure for circular and linear variables, which we call CLMI.

4.1. Circular-circular mutual information

In [20], the CMI is approximated using numerical methods. Here, we present a simple expression for CMI for a general family of distributions.

Let Θ, Ψ be a pair of circular variables. Then, the CMI between Θ and Ψ is defined by

$$\text{CMI}(\Theta, \Psi) = \int_0^{2\pi} \int_0^{2\pi} f(\theta, \psi) \log \left\{ \frac{f(\theta, \psi)}{f(\theta)f(\psi)} \right\} d\psi d\theta, \tag{10}$$

where $f_\Theta(\theta)$ is the marginal density of Θ , $f_\Psi(\psi)$ is the marginal density of Ψ , and $f(\theta, \psi)$ is the joint density of (Θ, Ψ) .

Following a similar method to that in Johnson and Wehrly [14], Wehrly and Johnson presented a general family for circular variables in [42]. Here, we consider the following subfamily with the joint density function

$$f(\theta, \psi) = 2\pi \delta(2\pi F_\Theta(\theta) - 2\pi q F_\Psi(\psi)) f_\Theta(\theta) f_\Psi(\psi), \tag{11}$$

where $0 \leq \theta, \psi < 2\pi$, $f_\Theta(\theta)$ and $f_\Psi(\psi)$ are any probability density functions on the circle, $F_\Theta(\theta)$ and $F_\Psi(\psi)$ are the CDFs of $f_\Theta(\theta)$ and $f_\Psi(\psi)$, respectively, $q \in \{1, -1\}$ decides the positive or negative association between the two variables, and $\delta(\nu)$ is the density of the wrapped Cauchy distribution $wC(\mu_\delta, \varepsilon_\delta)$.

Assume that a random vector (Θ, Ψ) has the distribution given by Eq. (11). Then, it holds that the marginal distribution of Θ has the density $f_\Theta(\theta)$ and the CDF $F_\Theta(\theta)$ and that the marginal distribution of Ψ has the density $f_\Psi(\psi)$ and the CDF $F_\Psi(\psi)$. As pointed out in [16], the bivariate wrapped Cauchy distribution from Eq. (2) is a special case of the Wehrly and Johnson family [42].

Theorem 3. Let (Θ, Ψ) have the distribution given by Eq. (11). Then, the CMI between Θ and Ψ defined in Eq. (10) is given by

$$\text{CMI}(\Theta, \Psi) = 2\pi \int_0^{2\pi} \delta(t) \log\{\delta(t)\} dt.$$

In particular, if δ is the density of the wrapped Cauchy distribution $wC(\mu_\delta, \varepsilon_\delta)$, then

$$\text{CMI}(\Theta, \Psi) = -\log(1 - \varepsilon_\delta^2). \tag{12}$$

Proof. Let $f(\theta, \psi)$ be the joint density function given by Eq. (11) and defined in Johnson and Wehrly [42]. Then, the CMI between Θ and Ψ is given by

$$\begin{aligned} \text{CMI}(\Theta, \Psi) = & \int_0^{2\pi} \int_0^{2\pi} 2\pi \delta(2\pi F_\Theta(\theta) - q2\pi F_\Psi(\psi)) f_\Theta(\theta) f_\Psi(\psi) \\ & \times \log\{2\pi \delta(2\pi F_\Theta(\theta) - q2\pi F_\Psi(\psi))\} d\psi d\theta. \end{aligned}$$

Changing the variables $u = 2\pi F_\Theta(\theta)$ and $v = 2\pi F_\Psi(\psi)$, we have

$$\text{CMI}(\Theta, \Psi) = \frac{1}{2\pi} \int_0^{2\pi} \int_0^{2\pi} \delta(u - qv) \log\{2\pi \delta(u - qv)\} dudv.$$

Note that (U, V) has the distribution of Kato [15]. It follows from the change of variables $t_1 = u - qv$ and $t_2 = v$ that

$$\text{CMI}(\Theta, \Psi) = \int_0^{2\pi} \delta(t_1) \log\{2\pi \delta(t_1)\} dt_1.$$

Since $\delta(\cdot)$ is the density of the wrapped Cauchy $wC(\mu_\delta, \varepsilon_\delta)$, it follows that

$$\begin{aligned} \text{CMI}(\Theta, \Psi) = & \int_0^{2\pi} \frac{1}{2\pi} \frac{1 - \varepsilon_\delta^2}{1 + \varepsilon_\delta^2 - 2\varepsilon_\delta \cos(t_1)} \log\left\{ \frac{1 - \varepsilon_\delta^2}{1 + \varepsilon_\delta^2 - 2\varepsilon_\delta \cos(t_1)} \right\} dt_1 \\ = & -\log(1 - \varepsilon_\delta^2), \end{aligned}$$

where the second equality follows from Equation (4.396.16) of Gradshteyn and Ryzhik [11]. \square

This CMI measure is expressed in a closed form. Therefore, it is computationally very fast and can be used for pairs of variables having any case of the family given by Eq. (11) that allows the calculation of the CDFs.

To estimate ε_δ in Eq. (12), which is unknown, we use an estimate from [16] based on a sample of size N : $\{(\theta_1, \psi_1), \dots, (\theta_N, \psi_N)\}$, given by

$$\hat{\varepsilon}_\delta = \frac{1}{N} \left(\left| \sum_{j=1}^N \exp(i(2\pi \hat{F}_\Theta(\theta_j) - (2\pi \hat{F}_\Psi(\psi_j)))) \right| - \left| \sum_{j=1}^N \exp(i((2\pi \hat{F}(\theta_j) + (2\pi \hat{F}(\psi_j)))) \right| \right),$$

where $\hat{F}_\Theta(\theta_j)$ and $\hat{F}_\Psi(\psi_j)$ denote the empirical CDFs of Θ and Ψ in the sample, respectively.

Considering the particular case where bivariate variables follow the bivariate wrapped Cauchy distributions, as described in Eq. (2), the estimate of ε_δ simplifies to the correlation parameter ρ between Θ and Ψ given in Eq. (6).

4.2. Circular-linear mutual information

Following the development of the CMI, we next present the CLMI, which allows a closed-form expression for the mutual-information measure for a general family of distributions.

Let Θ be a circular variable, and let X be a linear variable. Then, the CLMI between Θ and X is defined by

$$\text{CLMI}(\Theta, X) = \int_{-\infty}^{\infty} \int_0^{2\pi} f(\theta, x) \log\left\{ \frac{f(\theta, x)}{f_\Theta(\theta) f_X(x)} \right\} d\theta dx, 0 \leq \theta < 2\pi, x \in \mathbb{R}, \tag{13}$$

where $f_\Theta(\theta)$ is the marginal density of Θ , $f_X(x)$ is the marginal density of X , and $f(\theta, x)$ is the joint density of (Θ, X) .

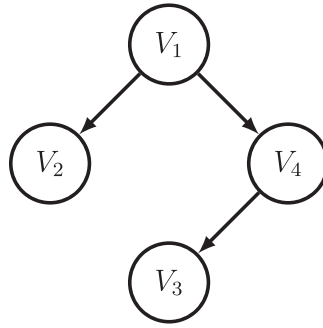


Fig. 2. Structure of a tree-structured Bayesian network with four variables and V_1 as the root node. Each node has a maximum of one parent.

Theorem 4. Let (Θ, X) have the distribution density of Eq. (8). Then, the CLMI between Θ and X , defined in Eq. (13), is given by

$$\text{CLMI}(\Theta, X) = 2\pi \int_0^{2\pi} g(t) \log\{g(t)\} dt.$$

In particular, if g is the density of the wrapped Cauchy distribution $wC(\mu_g, \varepsilon_g)$, then

$$\text{CLMI}(\Theta, X) = -\log(1 - \varepsilon_g^2). \tag{14}$$

Proof. Theorem 4 can be proved in a similar manner to Theorem 3 by changing the variables $u = 2\pi F_\Theta(\theta)$ and $v = 2\pi F_X(x)$. \square

Because ε_g in Eq. (14) is unknown in the usual settings, we need to estimate ε_g in order to estimate the value of CLMI. Similar to the CMI, an estimate of ε_g based on a sample $N : \{(\theta_1, x_1), \dots, (\theta_N, x_N)\}$ is given by

$$\hat{\varepsilon}_g = \frac{1}{N} \left| \sum_{j=1}^N \exp(i(2\pi \hat{F}_\Theta(\theta_j) - 2\pi \hat{F}_X(x_j))) \right| - \left| \sum_{j=1}^N \exp(i(2\pi \hat{F}_\Theta(\theta_j) + 2\pi \hat{F}_X(x_j))) \right|,$$

where $\hat{F}_\Theta(\theta)$ and $\hat{F}_X(x)$ are the estimated distribution functions of Θ and X , respectively. It can be seen that $\hat{\varepsilon}_g$ is a method of moments estimator of ε_g by noting that

$$|E[\exp(i(2\pi F_\Theta(\Theta) - 2\pi F_X(X)))]| = \varepsilon_g$$

and

$$|E[\exp(i(2\pi F_\Theta(\Theta) + 2\pi F_X(X)))]| = 0.$$

Similar to the CMI, the CLMI for the Johnson–Wehrly family of distributions provided by Eq. (8) is computationally very fast because it is expressed in closed form. In addition, it can be used to compute the mutual-information measure between circular and linear variables without needing any assumptions on their marginal distributions.

If we consider the particular case where $\Theta \sim wC(\mu_\theta, \varepsilon_\theta)$ and $X \sim N(\mu_x, \sigma_x)$, then the estimate of ε_g simplifies to

$$\hat{\varepsilon}_g = \frac{1}{N} \left| \sum_{j=1}^N \exp(i(\Phi_j - \tau_j)) \right| - \left| \sum_{j=1}^N \exp(i(\Phi_j + \tau_j)) \right|,$$

where $\Phi_j = 2\pi \hat{F}_\Theta(\theta_j) = 2 \arctan\left(\frac{1+\hat{\varepsilon}_\theta}{1-\hat{\varepsilon}_\theta} \tan\left(\frac{\theta_j - \hat{\mu}_\theta}{2}\right)\right)$ and $\tau_j = 2\pi \hat{F}_X(x_j) = \pi \operatorname{erf}\left(\frac{x_j - \hat{\mu}_x}{\hat{\sigma}_x \sqrt{2}}\right)$, where $\hat{\varepsilon}_\theta$ and $\hat{\mu}_\theta$ are as in Section 2.1, $\hat{\mu}_x$ and $\hat{\sigma}_x$ are as in Section 2.2, and erf is the Gauss error function.

5. Circular-linear tree-structured Bayesian network learning

In this section, we show an application of the CLMI and the CMI measures presented in Section 4.

The study of Bayesian networks involves probability and graph theories. There are two different ways to build the structure of a Bayesian network model. The first is based on supergraphs of the network skeleton, and testing the conditional independence of given subsets of the skeleton. The best-known algorithm for this structure-learning method is the PC [39] algorithm. However, this learning method is not considered here owing to the lack of conditional independence tests for circular domains. The second, and chosen method, is the score-based search. We use a score and search algorithm as an optimization problem for the structure learning in a Bayesian network with a tree-structure, where nodes and arcs represent the variables and dependence relationships between them, respectively (Fig. 2). Therefore, each node is allowed to have only one parent node. The importance of this application is that it allows the use of both linear and circular variables to develop this kind of Bayesian network model.

Algorithm 1 Circular-linear tree-structure algorithm.

- 1: Given $\Theta_1, \Theta_2, \dots, \Theta_n$ circular random variables and X_1, X_2, \dots, X_m linear random variables,
 - (i) Estimate the parameters of all marginals $f(\theta_i)$ and of $f(x_i)$
 - (ii) Estimate the dependence parameters ε_δ , γ , and ε_g of the joint density functions $f(\theta_i, \theta_j)$, $f(x_i, x_j)$, and $f(\theta_i, x_j)$, respectively, for all pairs of variables based on the estimated parameters of the marginals
- 2: Using these distributions, compute all values of $\text{CLMI}(\Theta_i, X_j)$, $\text{CMI}(\Theta_i, \Theta_j)$ and $\text{MI}(X_i, X_j)$ (i.e., the $(n+m)(n+m-1)/2$ edge weights), and order them
- 3: Assign the largest two edges to the undirected tree to be represented
- 4: Examine the next-largest edge and add it to the current tree, unless it forms a loop, in which case discard it and examine the next largest edge
- 5: Repeat step 4 until $n+m-1$ edges have been selected (and the spanning undirected tree is complete).
- 6: Choose a root node and follow the structure to create the maximum weight spanning directed tree structure

Table 1

Mean accuracy \pm standard deviations of the simulation results for the circular-linear tree-structured Bayesian network for different combinations of circular and linear variables.

Circular variables	Linear variables						
	3	5	7	10	15	20	30
3	0.710 \pm 0.191	0.723 \pm 0.185	0.703 \pm 0.185	0.691 \pm 0.191	0.697 \pm 0.194	0.710 \pm 0.192	0.689 \pm 0.208
5	0.744 \pm 0.173	0.741 \pm 0.169	0.774 \pm 0.166	0.749 \pm 0.172	0.736 \pm 0.169	0.749 \pm 0.170	0.766 \pm 0.163
7	0.804 \pm 0.146	0.780 \pm 0.144	0.791 \pm 0.139	0.796 \pm 0.140	0.791 \pm 0.136	0.811 \pm 0.137	0.788 \pm 0.147
10	0.830 \pm 0.119	0.828 \pm 0.123	0.825 \pm 0.120	0.808 \pm 0.125	0.831 \pm 0.115	0.817 \pm 0.123	0.837 \pm 0.122
15	0.858 \pm 0.096	0.854 \pm 0.090	0.866 \pm 0.094	0.867 \pm 0.092	0.863 \pm 0.096	0.860 \pm 0.095	0.863 \pm 0.095
20	0.879 \pm 0.077	0.884 \pm 0.075	0.886 \pm 0.079	0.876 \pm 0.084	0.879 \pm 0.078	0.882 \pm 0.079	0.878 \pm 0.084
30	0.900 \pm 0.061	0.901 \pm 0.060	0.910 \pm 0.067	0.903 \pm 0.062	0.899 \pm 0.060	0.900 \pm 0.065	0.904 \pm 0.060

Our proposed algorithm (Algorithm 1) is based on the algorithm introduced by Chow and Liu [3] to find a maximum weight spanning tree structure. The weight between a pair of variables is measured as the CLMI, CMI, or MI between them, depending on the nature of the two variables involved. Note that the range of values for the mutual-information measures is $[0, \infty)$. The complexity of this algorithm is quadratic in the number of variables.

Given the undirected tree structure from Algorithm 1, there are $n+m$ possible trees, depending on the selected root node. Note that if the estimate of the parameters method in Step 1 is performed via MLE, the generated maximum weight spanning directed tree is also a maximum likelihood tree. Otherwise, while we can confirm that the resulting tree is a maximum spanning tree, we cannot ensure that it is a maximum likelihood tree.

5.1. Experimental results

We used the R software package [40] to report the results of experiments applied over synthetic Bayesian network structures. Owing to the properties explained in Section 2 that make the structure-learning phase easier, the data used in the experiments are simulated from Gaussian distributions and wrapped Cauchy distributions. We generate 12,250 different simulated structures: 250 different structures for every combination of 3, 5, 7, 10, 15, 20, and 30 linear Gaussian variables with 3, 5, 7, 10, 15, 20, and 30 circular wrapped Cauchy variables (i.e., 49 combinations of variables). From each of these structures, we simulate a data set with $N = 500$ instances, where the parameters of the variables are assigned based on the uniform distributions on the following ranges: $-1000 < \iota < 1000$ and $0 < \sigma < 100$ for the Gaussian marginal distributions and $0 < \iota < 2\pi$ and $0 < \varepsilon < 1$ for the wrapped Cauchy marginal distributions. Since σ is uniformly distributed on the wide range, most of the linear variables are likely to have large variance; note that $E(\sigma) = 50$. On the other hand, the concentration of the circular variables is less likely to become too small because $E(\varepsilon) = 0.5$. We have made this assumption in order to ensure that our learning algorithm appropriately estimates the structures so that most of the circular variables are more influential than most of the linear ones.

In addition, to enforce the dependence between the parent and children nodes of the network, we assign a correlation parameter between them as a uniform random value $0.5 < |\rho| < 1$, with $q = 1$. Then, we apply the structure-learning algorithm proposed in Algorithm 1 over the generated data sets. To measure the accuracy of the results, we compare the misplaced arcs of the learned structure to the initial simulated structure. The results of these experiments are displayed in Table 1, where the results of the 49 combinations of variables are shown as the mean accuracy \pm standard deviations. We refer to accuracy as the percentage of non-misplaced arcs (i.e., how accurate a replicated structure is when compared against the original structure).

To better understand the algorithm behaviour when varying the number of circular and linear variables, we apply the Friedman test [8] (significance level $\alpha = 0.05$). This statistical test evaluates equality between sets of results that follow

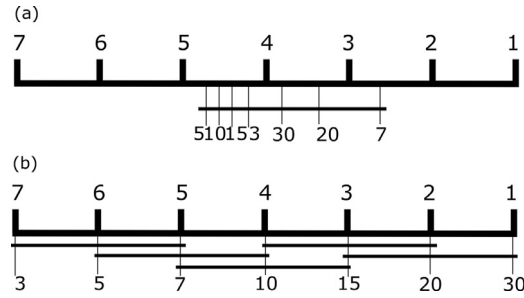


Fig. 3. Demšar diagram to compare the experimental results by varying the number of (a) linear variables or (b) circular variables.

unknown distributions in order to detect significant differences among the different numbers of linear or circular variables as a whole set. Here, the null hypothesis of equality between the sets was rejected (p -value ≤ 0.05). Therefore, we proceeded with the Nemenyi post-hoc test [31] (significance level $\alpha = 0.05$) to compare the sets of results with each other.

Fig. 3 provides the statistical test results using Demšar's diagram [5]. The top axis plots the average Friedman test ranks of the sets. The lowest (best) rank classifiers are to the right and are considered to be better. Then, to show Nemenyi test results, we connect those sets of results that are not significantly different (p -value > 0.05 in the Nemenyi post-hoc test).

Analysing Table 1 and Fig. 3(a), which shows the results of changing the number of linear variables, we find that the Friedman test is not rejected (p -value = 0.61). This implies that there is no statistically significant difference between the experimental results when the number of linear variables varies. Nevertheless, Fig. 3(b), in which the number of circular variables varies, shows that the Friedman test is rejected (p -value = 0.000018). Furthermore, the accuracy increases as the number of circular variables increases. The case of 30 circular variables is the most accurate, and the case with 3 circular variables is the least accurate. In addition, conducting Nemenyi's post-hoc tests for every pair of cases, we find no statistically significant differences between consecutive trios of cases (i.e., 3-5-7, 5-7-10, 7-10-15, 10-15-20, and 15-20-30).

6. Real example

In this section, we apply the proposed structure learning for our circular-linear Bayesian network to a meteorological data set.

Recently, the study of wind characteristics has become an important field owing to the importance of the location and orientation of wind turbines for profitable wind energy utilization. We use our proposed circular-linear tree-structured Bayesian network to model the relationship between wind speed, wind direction, and other meteorological features from various stations around Europe.

The data include circular and linear measurements collected from seven wind stations located in Europe: Baltic Sea in Poland, Black Sea in Romania, Dwejra in Malta, Hegyhatsal in Hungary, Lampedusa in Italy, Pallas-Sammaltunturi in Finland, and the M Ocean station in Norway (Fig. 4). Each station records wind direction and wind speed measurements. In addition, relative humidity, atmospheric temperature, and atmospheric pressure are recorded at the Lampedusa station. These data are available at the World Data Centre for Greenhouse Gases¹ (WDCGG). In our data set, we consider the records for the period 16 July 1992 to 29 December 2005. Since some stations report more than one record per day, we calculate the mean value per day for linear measurements and the mean direction per day for circular measurements. Thus, the data set has $n + m = 7 + 10 = 17$ variables (i.e., wind speed and wind direction from every station, as well as the additional three measures from the Lampedusa station) and $N = 3301$ instances.

Table 2 shows the names of the variables, the numbers of samples of the variables, and the parameter estimates of the marginal distributions. Note that some circular variables are close to circular uniformity because $\hat{\varepsilon} \simeq 0.05$. Owing to the nature of the data, which are recorded over a long period, the circular data show low concentration parameters. The highest concentration parameter is in the Dwejra station in Malta, where $\hat{\varepsilon} = 0.39$, with $\hat{\mu} = -0.77$ (approximately 311 degrees). Note that Lampedusa station, which is close to Dwejra station, shows a similar mean wind direction of $\hat{\mu} = -0.58$ (approximately 327 degrees) and a concentration parameter of $\hat{\varepsilon} = 0.20$. With regard to wind speed linear variables, note that the highest values are shown in the M Ocean station and Baltic Sea station, both of which are located in the sea or on the coast of the Northern Europe territory.

The circular-linear tree-structured Bayesian network model (Fig. 5) reveals the conditional relationship between the variables measured from each meteorological station. Note that most Lampedusa station nodes are connected to Dwejra station nodes. This is because these two stations are geographically close together, as previously mentioned. The figure also shows other geographically close node connections, such as the arc between the Pallas-Sammaltunturi and M Ocean stations, both located at the Northern Europe territory, and the arc between the Black Sea and the Hegyhatsal stations, which are the

¹ WDCGG URL: <http://ds.data.jma.go.jp/gmd/wdcgg/>.

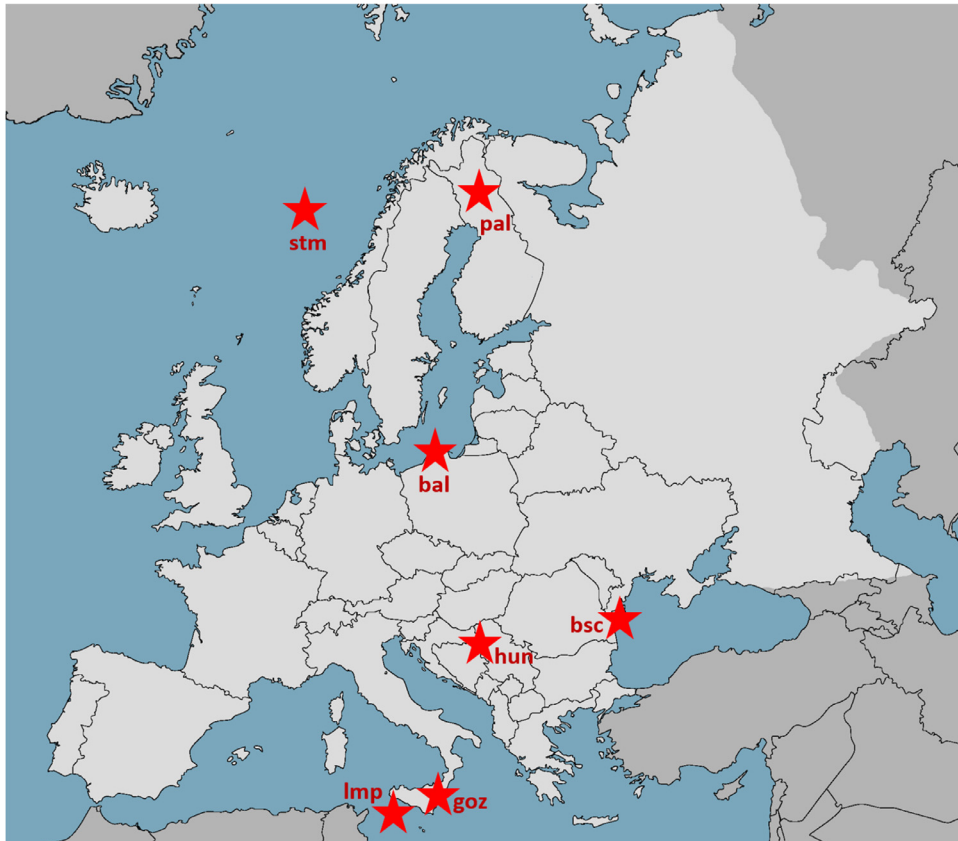


Fig. 4. European locations of the meteorological stations from the WDCGG data set.

Table 2

Station name, variable described, variable used in the model, number of non-missing cases (N_i), circular mean $\hat{\mu}$ or linear mean $\hat{\iota}$ (where applicable), unit of measure, concentration $\hat{\varepsilon}$ or standard deviation $\hat{\sigma}$ (where applicable), and type of variable (C: Circular; L: Linear) for the 17 numeric variables of the WDCGG data set. The circular variables range from 0 to 2π .

Station	Variable	Name	N_i	$\hat{\mu}/\hat{\iota}$	Units	$\hat{\varepsilon}/\hat{\sigma}$	Type
M Ocean	Wind direction	stmWD	1288	-1.86	radians	0.06	C
	Wind speed	stmWS	1288	8.93	m/s	3.91	L
Pallas-Sammaltunturi	Wind direction	palWD	150	-2.76	radians	0.18	C
	Wind speed	palWS	150	6.4	m/s	3.21	L
Lampedusa	Wind direction	impWD	446	-0.58	radians	0.20	C
	Wind speed	impWS	446	6.62	m/s	3.94	L
	Relative humidity	impRH	446	56.4	%	27.8	L
	Atmospheric pressure	impAP	446	1010	hPa	6.29	L
Hegyhatsal	Atmospheric temp.	impAT	446	19.59	Celsius	5.26	L
	Wind direction	hunWD	557	-0.67	radians	0.07	C
Dwejra	Wind speed	hunWS	557	3.82	m/s	3.20	L
	Wind direction	gozWD	157	-0.77	radians	0.39	C
Black Sea	Wind speed	gozWS	157	2.79	m/s	2.17	L
	Wind direction	bscWD	550	0.47	radians	0.22	C
Baltic Sea	Wind speed	bscWS	550	4.73	m/s	2.56	L
	Wind direction	balWD	1169	-1.77	radians	0.21	C
	Wind speed	balWS	1169	9.93	m/s	4.90	L

two Eastern Europe stations. Therefore, it seems that our circular-linear model is capturing the dependence relationships between the variables of the data set properly.

The Schwarz Bayesian information criterion (SBIC) [35] is a model selection criterion, where the lowest value is preferred. It is based on the likelihood function with an overfitting penalty, and is defined as

$$SBIC = -2\ln\hat{L} + \ln(N)w,$$

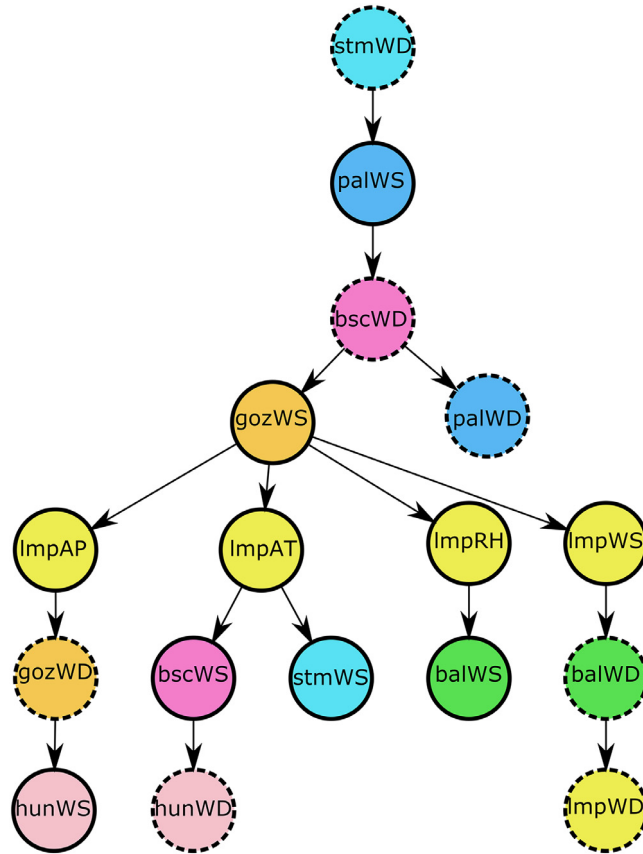


Fig. 5. Circular-linear tree-structured Bayesian network for the WDCGG meteorological data set. The names of the variables are shown in Table 2. The selected root node is the wind direction at the *M* Ocean station. Dashed border node lines indicate circular variables, while solid border node lines indicate linear variables. Nodes with the same colour are recorded at the same station. Nodes with similar colour tones are located close to each other geographically.

Table 3

SBIC comparison between the circular-linear Bayesian network model, Gaussian Bayesian network model, and discrete Bayesian network model for the WDCGG data set.

Bayesian network	
Model	SBIC
Circular-Linear	$-5.9197 * 10^{192}$
Gaussian	$-2.0896 * 10^{169}$
Discrete	$-3.9851 * 10^4$

where \hat{L} is the likelihood function value, N is the sample size, and w is the number of parameters to be estimated in the model.

We use the SBIC to compare our model to a Gaussian Bayesian network model, where we assume that all variables follow Gaussian continuous distributions. We also compare our model to a discrete Bayesian network model using the discretization of variables. This discretization is carried out by considering each variable as linear, and then creating 10 partitions of equal width (these vary for each variable, depending on their corresponding domain). Note that for those cases where the variables are circular, the linear domain is considered to be between 0 and 2π .

Comparing the SBIC values in Table 3, we observe that our circular-linear model clearly outperforms the other two models, which ignore the circular nature of the circular variables and treat them as (linear) continuous or discrete variables.

7. Conclusions

Circular data are often observed together with linear data in the sciences. In this study, we showed that our subfamily of Johnson & Wehrly bivariate distributions has tractable properties, such as well-known marginals and conditionals and a closed-form expression for the estimators of the parameters. We presented a CLMI measure that measures the mutual dependence between a circular variable and a linear variable by determining the similarity between the joint density and the product of their marginal densities. We also extended the definition of the CMI measure. We have shown that the CLMI and CMI can be expressed in a simple and closed form for our distributions for circular-linear data and bivariate circular data, respectively.

In addition, we described experimental results that illustrate how to use these measures (i.e., the CLMI and CMI) with the well-known mutual information between linear variables. To the best of our knowledge, this study is the first to develop a circular-linear tree-structured Bayesian network model that can capture the dependence between any possible pair of linear and circular variables.

Then, we applied our algorithm for a tree-structured Bayesian network model to a real data set in order to model the relationships between circular and linear measurements recorded at seven meteorological stations located in Europe. Here, we observed that the proposed model captures the strong dependence between variables recorded in geographically close stations well and outperforms other models, which assume all variables to be Gaussian or discrete.

There are several potential applications of our tree-structured Bayesian network model, some of them are related to sports (e.g., baskets in basketball or goals in football), social behaviour (e.g., hand gesture recognition, arms movement) and meteorological events (e.g., twister progress, earthquake epicentre location and expansion), among many others.

Working with a combination of circular and linear statistics is a non-trivial task. Applications within Bayesian networks and machine learning research for graphical models open a challenging field. As future work, we intend to adapt the proposed circular-linear graphical model to perform real-time classification. In addition, dropping the dimension constraint (of one parent) in our model would be another interesting path to explore in order to extend this model to a more general Bayesian network case allowing more than one parent per node.

References

- [1] T. Abe, C. Ley, A tractable, parsimonious and flexible model for cylindrical data, with applications, *Econom. Stat.* (2016) Inpress., doi:[10.1016/j.ecosta.2016.04.001](https://doi.org/10.1016/j.ecosta.2016.04.001).
- [2] K. Bowman, L. Shenton, *Methods of moments*, *Encycl. Stat. Sci.* 5 (1985) 467–473.
- [3] C. Chow, C. Liu, Approximating discrete probability distributions with dependence trees, *IEEE Trans. Inf. Theory* 14 (3) (1968) 462–467, doi:[10.1109/TIT.1968.1054142](https://doi.org/10.1109/TIT.1968.1054142).
- [4] T.M. Cover, J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 2012.
- [5] J. Demšar, Statistical comparisons of classifiers over multiple data sets, *J. Mach. Learn. Res.* 7 (2006) 1–30.
- [6] N.I. Fisher, *Statistical Analysis of Circular Data*, Cambridge University Press, 1995.
- [7] N.I. Fisher, A.J. Lee, Regression models for an angular response, *Biometrics* 48 (3) (1992) 665–677, doi:[10.2307/2532334](https://doi.org/10.2307/2532334).
- [8] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *J. Am. Stat. Assoc.* 32 (200) (1937) 675–701, doi:[10.1080/01621459.1937.10503522](https://doi.org/10.1080/01621459.1937.10503522).
- [9] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, *Mach. Learn.* 29 (1997) 131–163, doi:[10.1023/A:100746552](https://doi.org/10.1023/A:100746552).
- [10] A.L. Gould, A regression technique for angular variates, *Biometrics* 25 (4) (1969) 683–700, doi:[10.2307/2528567](https://doi.org/10.2307/2528567).
- [11] I.S. Gradshteyn, I.M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed., Academic Press, 2007.
- [12] S.R. Jammalamadaka, A. Sengupta, *Topics in Circular Statistics*, World Scientific, 2001.
- [13] N.L. Johnson, S. Kotz, N. Balakrishnan, *Distributions in Statistics: Continuous Univariate Distributions*, Houghton Mifflin, 1970.
- [14] R.A. Johnson, T.E. Wehrly, Some angular-linear distributions and related regression models, *J. Am. Stat. Assoc.* 73 (363) (1978) 602–606, doi:[10.1080/01621459.1978.10480062](https://doi.org/10.1080/01621459.1978.10480062).
- [15] S. Kato, A distribution for a pair of unit vectors generated by Brownian motion, *Bernoulli* 15 (3) (2009) 898–921, doi:[10.3150/08-BEJ178](https://doi.org/10.3150/08-BEJ178).
- [16] S. Kato, A. Pewsey, A Möbius transformation-induced distribution on the torus, *Biometrika* 102 (2) (2015) 359–370, doi:[10.1093/biomet/asv003](https://doi.org/10.1093/biomet/asv003).
- [17] C. Kenley, *Influence Diagram Models with Continuous Variables*, Stanford University, 1986 Ph.D. thesis.
- [18] J.T. Kent, D.E. Tyler, Maximum likelihood estimation for the wrapped Cauchy distribution, *J. Appl. Stat.* 15 (2) (1988) 247–254, doi:[10.1080/02664768800000029](https://doi.org/10.1080/02664768800000029).
- [19] S. Kotz, N. Balakrishnan, N.L. Johnson, *Continuous Multivariate Distributions, Models and Applications*, John Wiley & Sons, 2004.
- [20] I. Leguey, C. Bielza, P. Larrañaga, Tree-structured Bayesian networks for wrapped Cauchy directional distributions, in: *Advances in Artificial Intelligence*, vol. 9868, Springer, 2016, pp. 207–216, doi:[10.1007/978-3-319-44636-3_19](https://doi.org/10.1007/978-3-319-44636-3_19).
- [21] P. Lévy, L'addition des variables aléatoires définies sur une circonférence, *Bulletin de la Société Mathématique de France* 67 (1939) 1–41.
- [22] C. Ley, T. Verdebout, *Modern Directional Statistics*, CRC Press, 2017.
- [23] S. Lloyd, On a measure of stochastic dependence, *Theory. Probab. Appl.* 7 (3) (1962) 301–312, doi:[10.1137/1107028](https://doi.org/10.1137/1107028).
- [24] K.V. Mardia, *Statistics of directional data*, *J. R. Stat. Soc. Ser. B (Methodological)* 37 (3) (1975) 349–393.
- [25] K.V. Mardia, G. Hughes, C.C. Taylor, H. Singh, A multivariate von Mises distribution with applications to bioinformatics, *Can. J. Stat.* 36 (1) (2008) 99–109, doi:[10.1002/cjs.5550360110](https://doi.org/10.1002/cjs.5550360110).
- [26] K.V. Mardia, P.E. Jupp, *Directional Statistics*, John Wiley & Sons, 2009.
- [27] K.V. Mardia, T.W. Sutton, A model for cylindrical variables with applications, *J. R. Stat. Soc. Ser. B (Methodological)* 40 (2) (1978) 229–233.
- [28] K.V. Mardia, J. Voss, Some fundamental properties of a multivariate von mises distribution, *Commun. Stat.* 43 (2014) 1132–1144, doi:[10.1080/03610926.2012.670353](https://doi.org/10.1080/03610926.2012.670353).
- [29] P. McCullagh, Möbius transformation and cauchy parameter estimation, *Ann. Stat.* 24 (2) (1996) 787–808, doi:[10.1214/aos/1032894465](https://doi.org/10.1214/aos/1032894465).
- [30] R. von Mises, Über die äganzzahligkeit der atomgewichte und verwandte fragen, *Zeitschrift für Physik* 19 (1918) 490–500.
- [31] P. Nemenyi, Distribution-free multiple comparisons, *Biometrics* 18 (2) (1962) 263.
- [32] F. Nojavan, S. Qian, C. Stow, Comparative analysis of discretization methods in Bayesian networks, *Environ. Modell. Softw.* 87 (2017) 64–71, doi:[10.1016/j.envsoft.2016.10.007](https://doi.org/10.1016/j.envsoft.2016.10.007).
- [33] A. Rényi, On measures of dependence, *Acta Math. Acad. Sci. Hungarica* 10 (3–4) (1959) 441–451, doi:[10.1007/BF02024507](https://doi.org/10.1007/BF02024507).
- [34] A. Rényi, On the dimension and entropy of probability distributions, *Acta Math. Acad. Sci. Hungarica* 10 (1–2) (1959) 193–215, doi:[10.1007/BF02063299](https://doi.org/10.1007/BF02063299).
- [35] G. Schwarz, Estimating the dimension of a model, *Ann. Stat.* 6 (2) (1978) 461–464, doi:[10.1214/aos/1176344136](https://doi.org/10.1214/aos/1176344136).

- [36] A. Sengupta, On the construction of probability distributions for directional data, *Bull. Indian Math. Soc.* 96 (2) (2004) 139–154.
- [37] R. Shachter, C. Kenley, Gaussian influence diagrams, *Manage. Sci.* 35 (5) (1989) 527–550, doi:[10.1287/mnsc.35.5.527](https://doi.org/10.1287/mnsc.35.5.527).
- [38] C.E. Shannon, A mathematical theory of communication, *Mob. Comput. Commun. Rev.* 5 (1) (2001) 3–55, doi:[10.1145/584091.584093](https://doi.org/10.1145/584091.584093).
- [39] P. Spirtes, C.N. Glymour, R. Scheines, *Causation, Prediction, and Search*, MIT Press, 2000.
- [40] R.D.C. Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2008.
- [41] Y. Tong, *The Multivariate Normal Distribution*, Springer, 1990.
- [42] T.E. Wehrly, R.A. Johnson, Bivariate models for dependence of angular observations and a related Markov process, *Biometrika* 67 (1) (1980) 255, doi:[10.1093/biomet/67.1.255](https://doi.org/10.1093/biomet/67.1.255).