Dalla Valle, Luciana, Leisen, Fabrizio and Rossini, Luca (2018) *Bayesian non-parametric conditional copula estimation of twin data.* Journal of the Royal Statistical Society: Series C (Applied Statistics), 67 (3). pp. 523-548. ISSN 0035-9254.

Citation for published version

Dalla Valle, Luciana and Leisen, Fabrizio and Rossini, Luca (2018) Bayesian non-parametric conditional copula estimation of twin data. Journal of the Royal Statistical Society: Series C (Applied Statistics), 67 . pp. 523-548. ISSN 0035-9254.

DOI

https://doi.org/10.1111/rssc.12237

Link to record in KAR

http://kar.kent.ac.uk/62861/

Document Version

Author's Accepted Manuscript

Bayesian Nonparametric Conditional Copula Estimation of Twin Data*

Luciana Dalla Valle [1]   Fabrizio Leisen[2]

Luca Rossini[3,4]

[1] University of Plymouth, U.K.   [2] University of Kent, U.K.

[3] Ca' Foscari University of Venice, Italy

[4] Free University of Bozen-Bolzano, Italy

June 9, 2017

Abstract

Several studies on heritability in twins aim at understanding the different contribution of environmental and genetic factors to specific traits. Considering the National Merit Twin Study, our purpose is to correctly analyse the influence of the socioeconomic status on the relationship between twins' cognitive abilities. Our methodology is based on conditional copulas, which allow us to model the effect of a covariate driving the strength of dependence between the main variables. We propose a flexible Bayesian nonparametric approach for the estimation of conditional copulas, which can model any conditional copula density. Our methodology extends the work of Wu et al. (2015) by introducing dependence from a covariate in an infinite mixture model. Our results suggest that environmental factors are more influential in families with lower socio-economic position.

Keywords: Bayesian nonparametrics, Conditional Copula models, Slice sampling, National Merit Twin Study, Social Science.

*Email Addresses for correspondences: fabrizio.leisen@gmail.com (Fabrizio Leisen); luca.rossini@unibz.it (Luca Rossini);luciana.dallavalle@plymouth.ac.uk (Luciana Dalla Valle).

1

1 Introduction

The literature on heritability of traits in children often focuses on twins, due to the shared environmental factors and the association of genetical characteristics. Among studies on the heritability of diseases, Wang et al. (2011) applied an efficient estimation method to mixed-effect models to analyze disease inheritance in twins.

One of the main purposes of studies on heritability is to estimate the different contribution of genetic and environmental factors to traits or outcomes (see, for example, the latent class twin method of Baker (2016)). Bates et al. (2013) studied the interactions between environmental and genetic effects to intelligence in twins, showing that higher socioeconomic status is associated with higher intelligence scores. Bioecological theory states that environmental factors may significantly influence the heritability of certain characteristics, such as cognitive ability, which is the readiness for future intellectual or educational pursuits. Several studies have found that cognitive ability is more pronounced and evident among children raised in higher socioeconomic status families. Such families can offer greater opportunities to children, due to their socioeconomic wealth status, and represent stimulating environments where children's inherited capabilities may become more manifest.

The aim of this paper is to correctly analyse the effect of socioeconomic factors on the relationship between twins' cognitive abilities. From a sample of 839 US adolescent twin pairs who completed the National Merit Scholarship Qualifying Test, we consider each twin's overall school performance (measured by a total score including English, Mathematics, Social Science, Natural Science and Word Usage), the mother's and father's education level and the family income. The data are plotted in Figure 1, which shows the scatterplots of the twins' school performances, on each axis, against the socioeconomic variables, whose values are in different colours (dark brown denotes low values, while light brown denotes high values). Figure 1 indicates that the twins' school performances are positively correlated and their dependence is influenced by the values of the socioeconomic variables (the mother's (panel (a)), the father's level of education (panel (b)) and the family income (panel (c))). Indeed, most of the light brown dots (denoting high values of the covariates) are grouped in the upper right corner, while the dark brown dots (denoting low values of the covariates) lie in the bottom left corner of each plot. Hence, the

higher the parents' education or family income, the higher the twins' school performance. This means that the twins' performance scores are functions of each covariate and they vary according to the values of the covariates.

(a) Mother's education

(b) Father's education
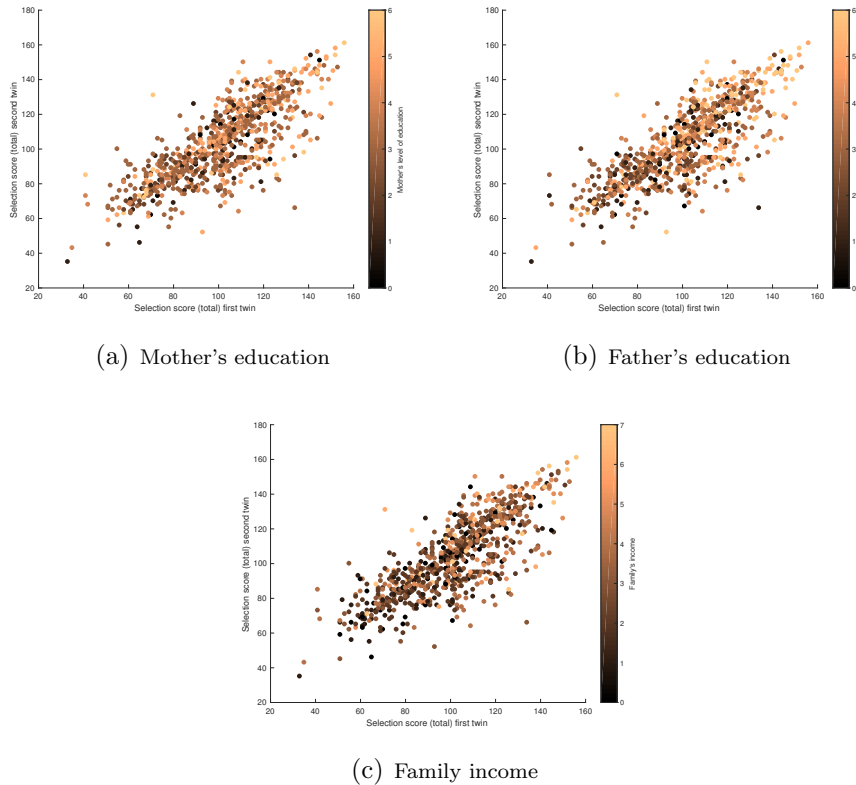
(c) Family income

Figure 1: Scatterplots of the twins overall scores with respect to the mother's (panel (a)) and father's level of education (panel (b)) and family income (panel (c)).

In Figure 2 we produced three-dimensional histograms of the twins' performance scores for different values of the mother's education. Similar plots for the father's level of education and family income are included in Appendix B, Figures B.1 and B.2. The different shapes of the histograms corresponding to different levels of the covariates suggest that the dependence structure between the twins' school outcomes changes according to the values of the mother's and father's education and family income. Therefore, a flexible model, able to capture the effect of a covariate on the dependence

3

between the kids' performance scores is necessary.

(a) $x = 0$, $n_0 = 19$ (b) $x = 1$, $n_1 = 54$ (c) $x = 2$, $n_2 = 105$

(d) $x = 3$, $n_3 = 305$ (e) $x = 4$, $n_4 = 195$ (f) $x = 5$, $n_5 = 124$

(g) $x = 6$, $n_6 = 37$
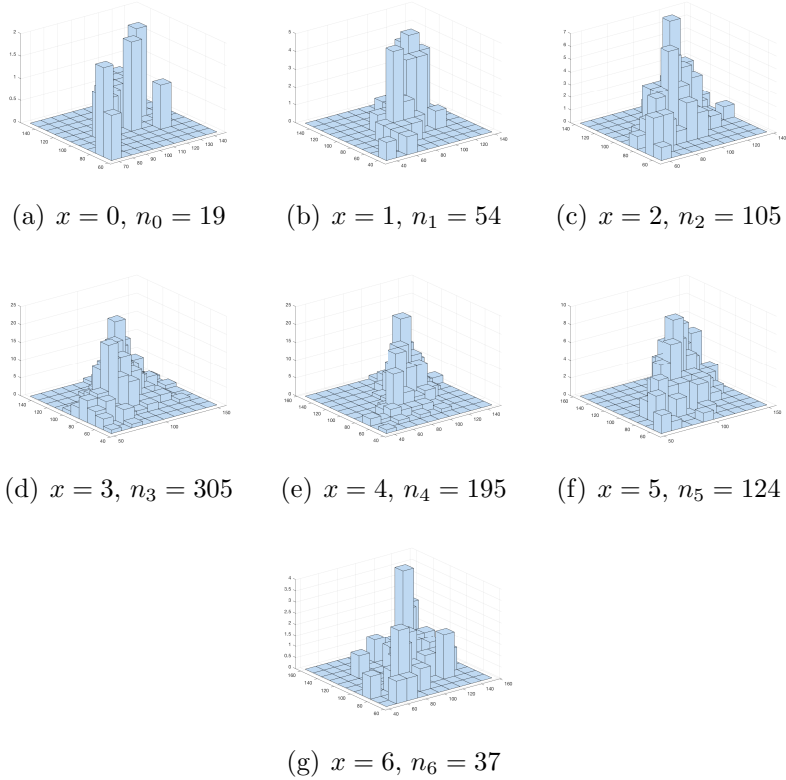
Figure 2: Three-dimensional histograms of the twins' performance scores for different values of the mother's education. In each panel, x denotes the level of the covariate and n_x denotes the relevant sample size.

In order to model the dependence structure between the twins' school performances, we use copulas, which are popular modeling approaches in multivariate statistics allowing the separation of the marginal components of a joint distribution from its dependence structure. More precisely, Sklar (1959) proved that a d-dimensional distribution H of the random variables $Y_1, \ldots, Y_d$ can be fully described by its marginal distributions and a function $C : [0,1]^d \rightarrow [0,1]$, called copula, through the relation $H(y_1, \ldots, y_d) = C(F_1(y_1), \ldots, F_d(y_d))$. In the literature, copulas have been applied to model the dependence between variables in a wide variety of fields (see Kolev et al. (2006) and Cherubini et al. (2004)). In particular, applications of copula models involved lifetime data analysis (Andersen (2005)), survival analysis of

4

Atlantic halibut (Braekers and Veraverbeke (2005)) and transfusion-related AIDS and cancer analysis (Emura and Wang (2012), Huang and Zhang (2008) and Owzar et al. (2007)).

The introduction of covariate adjustments to copulas has attracted an increased interest in recent years. Craiu and Sabeti (2012) propose a conditional copula approach in regression settings where the bivariate outcome can be continuous or mixed. Patton (2006) introduces time-variation in the dependence structure of ARMA models (see also Jondeau and Rockinger (2006) and Bartram et al. (2007) for other applications of time-series analysis to dependence modelling). The paper of Acar et al. (2010) provides a nonparametric procedure to estimate the functional relationship between copula parameters and covariates, showing that the gestational age drives the strength of dependence between the birth weights of twins. Abegaz et al. (2012) and Gijbels et al. (2012) propose semiparametric and nonparametric methodologies for the estimation of conditional copulas, establishing consistency and asymptotic normality results for the estimators. The methodology is then applied to examine the influence of the gross domestic product (GDP), in USD per capita, on the life expectancy of males and females at birth.

In a similar vein, parametric models such as Bayesian regression copulas allow the specification of Bayesian marginal regressions for a set of outcomes, linking the marginals to covariates, and combining them via a copula to form a joint model. The general framework of Bayesian Gaussian regression copulas with discrete, continuous or mixed outcomes is presented by Pitt et al. (2006) and allows to handle a multivariate regression with Gaussian and non-Gaussian marginal distributions.

Yin and Yuan (2009) adopt a Bayesian regression copula model in cancer clinical trials for dose finding to account for the synergistic effect of combinations of multiple drugs. A copula constructed from the skew t distribution is employed by Smith et al. (2012) to capture asymmetric and extreme dependence between variables modelled via Bayesian marginal regressions. While most Bayesian regression copula models focus on covariate adjustments for the marginals, recently Klein and Kneiss (2016) proposed simultaneous Bayesian inference for both the marginal distributions and the copula. Other contributions along the same lines are Taglioni et al. (2016), Stander et al. (2015a) and Stander et al. (2015b). However, the authors selected the copula family by using the deviance information criterion, which may suffer from limitations, as discussed for example by Plummer

(2008). Indeed, the choice of the copula family may be controversial and it is still an open problem (see Joe (2014)). The literature offers a rich range of copula families, such as elliptical copulas (e.g. Gaussian and Student's t) and archimedean copulas (e.g. Frank, Gumbel, Clayton and Joe) to accommodate various dependence structures. In this paper, we adopt a Bayesian nonparametric approach which allows us to overcome the issue of the choice of copula and we adopt a conditional copula approach to model the effect of a covariate on the dependence between variables. Our methodology builds on Wu et al. (2015), who propose a Bayesian nonparametric procedure to estimate any unconditional copula density function. The authors combine the well-known Gaussian copula density with the modeling flexibility of the Bayesian nonparametric approach, proposing to use an infinite mixture of Gaussian copulas. Burda and Prokhorov (2014) propose to use nonparametric univariate Gaussian mixtures for the marginals and a multivariate random Bernstein polynomial copula for the link function under the Dirichlet process prior. Our paper extends the work of Wu et al. (2015) to the conditional copula setting, by proposing a novel methodology which combines the advantages of a conditional copula approach with the modeling flexibility of Bayesian nonparametrics. In particular, we include a conditional covariate component to explain the variable dependence structure, allowing us further flexibility to the copula density modelling. Up to our knowledge, this is the first Bayesian nonparametric proposal in the conditional copulas literature.

The outline of the paper is the following. In Section 2 we briefly review the literature about conditional copulas and Bayesian nonparametric copula estimation. In Section 3 we introduce our novel Bayesian nonparametric conditional copula setting. Section 4 provides an algorithm for estimating the posterior parameters and Section 5 illustrates the performance of the methodology. Section 6 is devoted to the application of our methodology to the analysis of the National Merit Twin Study. Concluding remarks are given in Section 7.

2 Preliminaries

In this Section, we review some preliminary notions about conditional copulas and illustrate the Bayesian nonparametric copula density estimation introduced in Wu et al. (2015). In what follows, we focus on the bivariate

case for simplicity, however the arguments can be easily extended to more than two dimensions.

2.1 The conditional copula

Let Y_1 and Y_2 be continuous variables of interest and X be a covariate that may affect the dependence between Y_1 and Y_2. Following Gijbels et al. (2012), Abegaz et al. (2012) and Acar et al. (2010), we suppose that the conditional distribution of (Y_1, Y_2) given $X = x$ exists and we denote the corresponding conditional joint distribution function by

$$H_x(y_1, y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2 | X = x).$$

If the marginals of H_x, denoted as

$$F_{1x}(y_1) = P(Y_1 \leq y_1 | X = x), \qquad F_{2x}(y_2) = P(Y_2 \leq y_2 | X = x),$$

are continuous, then according to Sklar's theorem there exists a unique copula C_x which equals

$$C_x(u, v) = H_x(F_{1x}^{-1}(u), F_{2x}^{-1}(v)) \tag{1}$$

where $F_{1x}^{-1}(u) = \inf\{y_1 : F_{1x} \geq u\}$ and $F_{2x}^{-1}(v) = \inf\{y_2 : F_{2x} \geq v\}$ are the conditional quantile functions and $u = F_{1x}(y_1)$ and $v = F_{2x}(y_2)$ are called pseudo-observations. The conditional copula C_x fully describes the conditional dependence structure of (Y_1, Y_2) given $X = x$. An alternative expression for (1) is

$$H_x(y_1, y_2) = C_x(F_{1x}(y_1), F_{2x}(y_2)). \tag{2}$$

2.2 Bayesian nonparametric copula density estimation

Let $\Phi_\rho(y_1, y_2)$ denote the standard bivariate normal distribution function with correlation coefficient ρ. Then, C_ρ is the copula corresponding to Φ_ρ, taking the form:

$$C_\rho(u, v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)) \tag{3}$$

where Φ is the univariate standard normal distribution function. The Gaussian copula density is:

$$c_\rho(u, v) = |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\Phi^{-1}(u), \Phi^{-1}(v))(\Sigma^{-1} - \mathbf{I}) \begin{pmatrix} \Phi^{-1}(u) \\ \Phi^{-1}(v) \end{pmatrix}\right\} \tag{4}$$

where the correlation matrix is:

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Wu et al. (2015) proposed to use an infinite mixture of Gaussian copulas for the estimation of a copula density, as follows

$$c(u,v) = \sum_{j=1}^{\infty} w_j c_{\rho_j}(u,v) \tag{5}$$

where the weights w_j's sum up to 1 and the ρ_j's vary in $(-1,1)$. Given a set of n observations $(u_1, v_1), \ldots, (u_n, v_n)$, their model can be described through a hierarchical specification, i.e.

$$
\begin{aligned}
(u_i, v_i) \mid \rho_i &\overset{\text{ind}}{\sim} c_{\rho_i}(u_i, v_i), & i = 1, \ldots, n, \\
\rho_i \mid G &\overset{\text{iid}}{\sim} G, \\
G &\sim DP(\lambda, G_0),
\end{aligned}
\tag{6}
$$

where G is a Dirichlet Process prior with total mass λ and base measure G_0. This proposal is motivated by the fact that bivariate density functions on the real plain can be arbitrarily well approximated by a mixture of a countably infinite number of bivariate normal distributions of the form

$$f(y_1, y_2) = \sum_{j=1}^{\infty} w_j N((y_1, y_2) \mid (\mu_{1j}, \mu_{2j}), \Sigma_j)$$

where $N((y_1, y_2) \mid (\mu_{1j}, \mu_{2j}), \Sigma_j)$ is the joint bivariate normal density with mean vector (μ_{1j}, μ_{2j}) and correlation matrix Σ_j (see Lo (1984) and Ferguson (1983)). Roughly speaking, the authors are mimicking the Dirichlet process mixture model in the copula setting (see Escobar (1994) and Escobar and West (1995)). The sampling strategy follows the slice sampler of Walker (2007) and Kalli et al. (2011). The authors show that the Gaussian mixture is flexible enough to accurately approximate any bivariate copula density.

3 Conditional copula estimation with Dirichlet process priors

The data object of study requires a model which can take into account the effect of a covariate. We build on the model introduced by Wu et al. (2015)

and illustrated in the previous section. The idea is to replace the Gaussian copula with a conditional version where the correlation is a function of the covariate, i.e.
$$c_\rho(u, v|x) = c_{\rho(x)}(u, v).$$
The function $\rho(x)$ can be modelled as preferred, for instance, with a generalized linear model or with a non-linear function. In any case, we have that $\rho(x)$ will depend on a vector of parameters $\boldsymbol{\beta}$, so that

$$c_{\rho(x)}(u, v) = c_{\rho(x|\boldsymbol{\beta})}(u, v).$$

We assume a Dirichlet process prior on the vector of parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$. Following the model description provided in equation (6), we can summarize our model as follows,

$$
\begin{aligned}
(u_i, v_i) \mid \rho(x_i|\boldsymbol{\beta}_i) &\stackrel{\text{ind}}{\sim} c_{\rho(x_i|\boldsymbol{\beta}_i)}(u_i, v_i), \qquad i = 1, \ldots, n, \\
\boldsymbol{\beta}_i \mid G &\stackrel{\text{iid}}{\sim} G, \\
G &\sim DP(\lambda, G_0),
\end{aligned}
\tag{7}
$$

where G is a Dirichlet process prior with total mass λ and base measure G_0. In the numerical experiments, we consider a base measure G_0 which is a multivariate normal distribution with zero mean vector and variance $\sigma^2 I_d$, where I_d is the d-dimensional identity matrix and $\sigma^2 > 0$. As in Wu et al. (2015), our model can be described as an infinite mixture of Normal distributions,

$$c_\rho(u, v|x) = \sum_{j=1}^{\infty} w_j c_{\rho(x|\boldsymbol{\beta}_j)}(u, v), \tag{8}$$

and hence suitable for implementing a slice sampling algorithm, as explained in the next section.

In order to model the function $\rho(x|\boldsymbol{\beta})$, we would like to follow some standard approaches in the literature. Abegaz et al. (2012) model the dependence of the parameter of interest, with respect to the covariate, through a *calibration function* $\theta(x|\boldsymbol{\beta})$. It is important to highlight that in many copula families the parameter space is restricted. In contrast, the calibration function $\theta(x|\boldsymbol{\beta})$ can assume any value on the real line. In our case, the parameter is restricted to the interval $(-1, 1)$ and we need a transformation which can link the

calibration function $\theta(x|\boldsymbol{\beta})$ to $\rho(x|\boldsymbol{\beta})$. In this paper, we adopt the following transformation,

$$\rho(x|\boldsymbol{\beta}) = \frac{2}{|\theta(x|\boldsymbol{\beta})| + 1} - 1.$$

In our simulated and real data examples we focus on two particular calibration functions studied in the literature, which are

$$\theta(x|\boldsymbol{\beta}) = \beta_1 + \beta_2 x^2$$
$$\theta(x|\boldsymbol{\beta}) = \beta_1 + \beta_2 x + \beta_3 \exp\left(-\beta_4 x^2\right)$$

respectively, such that $\theta(x|\boldsymbol{\beta}) \in (-\infty, +\infty)$ and, consequently, $\rho(x|\boldsymbol{\beta}) \in (-1, 1)$.

4   Posterior sampling algorithm

The observations (y_{1i}, y_{2i}), for $i = 1, \ldots, n$, are transformed into the corresponding pseudo-observations (u_i, v_i) using a nonparametric estimation approach, as in Gijbels et al. (2011). The pseudo-observations are then plugged into the copula. Following equation (8), given (u_i, v_i) for $i = 1, \ldots, n$, and the conditional variable x_i, the conditional copula density function for each pair (u_i, v_i) can be written as an infinite mixture of conditional Gaussian copulas, such that:

$$c(u_i, v_i | x_i) = \sum_{j=1}^{\infty} w_j c_{\rho(x_i|\boldsymbol{\beta}_j)}(u_i, v_i) \tag{9}$$

where w_j's are the stick-breaking weights, i.e.

$$w_j = \pi_j \prod_{l=1}^{j-1} (1 - \pi_l)$$

where the π_j are distributed as a $\mathcal{B}e(1, \lambda)$, $\lambda > 0$. In order to sample from the infinite mixture displayed in equation (9), we use the slice sampling algorithm for mixture models proposed by Walker (2007) and Kalli et al. (2011). To reduce the dimensionality of the problem, the authors introduce a latent

variable z_i for each i which allows us to write the infinite mixture model as follows:

$$c(u_i, v_i, z_i | x_i) = \sum_{j=1}^{\infty} \mathbb{I}(z_i < w_j) c_{\rho(x_i | \boldsymbol{\beta}_j)}(u_i, v_i). \tag{10}$$

The introduction of the slice variable z_i reduces the sampling complexity to the analogous of a finite mixture model. In particular, letting

$$A_w = \{j : z_i < w_j\}, \tag{11}$$

then it can be proved that the cardinality of the set A_w is almost surely finite. Consequently, there is a finite number of parameters to be estimated. By iterating the data augmentation principle further, we introduce another latent variable d_i, which is called allocation variable, allowing us to allocate each observation to one component of the mixture model. Then, the conditional copula density $c(u_i, v_i, z_i, d_i | x_i)$ takes the form:

$$c(u_i, v_i, z_i, d_i | x_i) = \mathbb{I}(z_i < w_{d_i}) c_{\rho(x_i | \boldsymbol{\beta}_{d_i})}(u_i, v_i) \tag{12}$$

where $d_i \in \{1, 2, \dots\}$. Hence, the full likelihood function of the conditional copula model is:

$$\prod_{i=1}^{n} c(u_i, v_i, z_i, d_i | x_i) = \prod_{i=1}^{n} \mathbb{I}(z_i < w_{d_i}) c_{\rho(x_i | \boldsymbol{\beta}_{d_i})}(u_i, v_i). \tag{13}$$

We use the notation $(U, V) = \{i = 1, \dots, n : (u_i, v_i)\}$, $X = \{x_1, \dots, x_n\}$ to describe the pseudo-observations and the covariate values, respectively. We denote with $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots\}$ the vector of parameters and $D = \{d_1, \dots, d_n\}$, $Z = \{z_1, \dots, z_n\}$ and $\boldsymbol{\pi} = \{\pi_1, \pi_2, \dots\}$ the new variables introduced so far.

Therefore, we used a Gibbs sampler to simulate iteratively from the posterior distribution function, according to the following steps:

1. The stick-breaking components $\boldsymbol{\pi}$ are updated given $[Z, D, \boldsymbol{\beta}, (U, V), X]$;

2. The latent slice variables Z are updated given $[\boldsymbol{\pi}, D, \boldsymbol{\beta}, (U, V), X]$;

3. The allocation variables D are updated given $[\boldsymbol{\pi}, Z, \boldsymbol{\beta}, (U, V), X]$;

4. The vector of parameters $\boldsymbol{\beta}$ is updated given $[\boldsymbol{\pi}, Z, D, (U, V), X]$.

The Gibbs sampling details are explained in Appendix A.
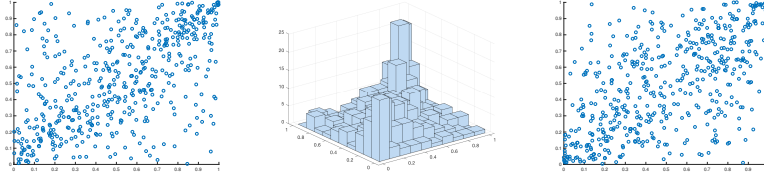
(a) Simulated, 1st cal.fun. (b) Simulated, 1st cal.fun. (c) Predictive, 1st cal.fun.
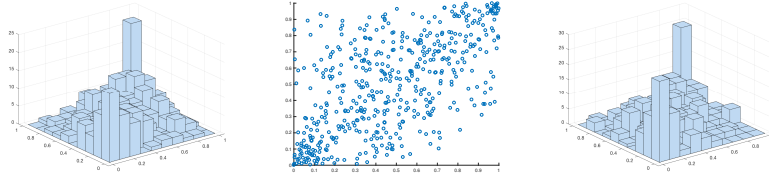
(d) Predictive, 1st cal.fun. (e) Simulated, 2nd cal.fun. (f) Simulated, 2nd cal.fun.
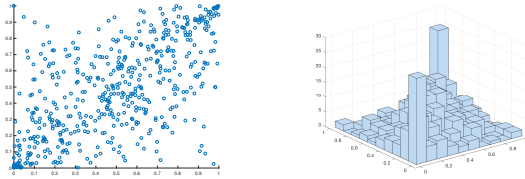
(g) Predictive, 2nd cal. fun. (h) Predictive, 2nd cal. fun.

Figure 3: Gaussian copula with sample size $n = 500$. Panels (a), (b), (c) and (d) depict the scatter plots and histograms, obtained with the first calibration function, of the simulated and predictive samples, respectively; panels (e), (f), (g) and (h) depict the scatter plots and histograms, obtained with the second calibration function, of the simulated and predictive sample, respectively.

5 Simulation experiments

This section illustrates the performance of the Bayesian nonparametric conditional copula model with simulated data. We generate datasets (U, V) of sizes $n = 250, 500$ and 1000 from different copula families, such as the Gaussian and Frank copulas.

The copula dependence parameter is considered as a function of the exogenous variable X, which is simulated from an Uniform distribution in the interval $[-2, 2]$.

For the Dirichlet process prior, we fix the total mass $\lambda = 1$ and, for the
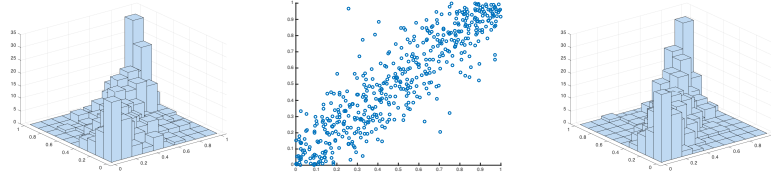
12

(a) Simulated, 1st cal.fun. (b) Simulated, 1st cal.fun. (c) Predictive, 1st cal.fun.

(d) Predictive, 1st cal.fun. (e) Simulated, 2nd cal.fun. (f) Simulated, 2nd cal.fun.

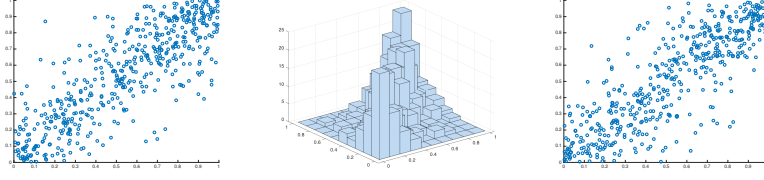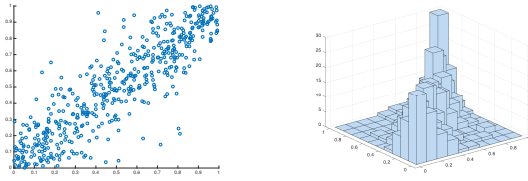(g) Predictive, 2nd cal. fun.　(h) Predictive, 2nd cal. fun.

Figure 4: Frank copula with sample size $n = 500$. Panels (a), (b), (c) and (d) depict the scatter plots and histograms, obtained with the first calibration function, of the simulated and predictive samples, respectively; panels (e), (f), (g) and (h) depict the scatter plots and histograms, obtained with the second calibration function, of the simulated and predictive sample, respectively.

base measure G_0, we adopt a bivariate normal distribution with zero mean vector and covariance matrix $\sigma^2 \boldsymbol{I}$, where $\sigma^2 = 100$. The following calibration functions are selected for $\theta(x|\boldsymbol{\beta})$:

$$\theta(x|\boldsymbol{\beta}) = \beta_1 + \beta_2 x^2$$
$$\theta(x|\boldsymbol{\beta}) = \beta_1 + \beta_2 x + \beta_3 \exp\left(-\beta_4 x^2\right).$$

As highlighted in Section 3, we link the calibration functions $\theta(x|\boldsymbol{\beta})$ with

$\rho(x|\boldsymbol{\beta})$ through the following transformation:

$$\rho(x|\boldsymbol{\beta}) = \frac{2}{|\theta(x|\boldsymbol{\beta})| + 1} - 1.$$

This ensures that $\rho(x|\boldsymbol{\beta})$ assumes values between $(-1, 1)$.

We run the Gibbs sampler algorithm described in Section 4 for 4000 iterations with (i) 500 burn-in iterations and (ii) 3500 burn-in iterations. Aiming at a parsimonious representation of the results, we focused on 3500 burn-in iterations, since 500 burn-in iterations gave very similar results.

(a) Gaussian Copula   (b) Frank Copula

Figure 5: Number of components (y-axis) generated at each MCMC iteration (x-axis) for the Gaussian (left) and for the Frank copula (right) with sample size $n = 500$.

Figure 3 illustrates the results of the application of the Bayesian nonparametric conditional copula model to data simulated from a Gaussian copula, with sample size $n = 500$. Figure 4 illustrates similar results for the Frank copula. Since the performances of the model with sample sizes $n = 250$ and 1000 for both copula families were analogous, here we omit the results.

In Figures 3 and 4, panels (a), (b), (c) and (d) show the scatter plots and histograms of the simulated data and the predictive samples obtained using the first calibration function; while panels (e), (f), (g) and (h) show the scatter plots and histograms of the simulated data and the predictive sample obtained using the second calibration function. The comparison between the simulated and predictive outputs highlights the excellent fit of the Bayesian

14

nonparametric conditional copula model using either calibration function and for different copula families. The model performance appears to be consistent across both copula families, demonstrating that the approach is suitable to model different dependence patterns and tail structures. Figure 5 shows

(a) Gaussian Copula   (b) Frank Copula

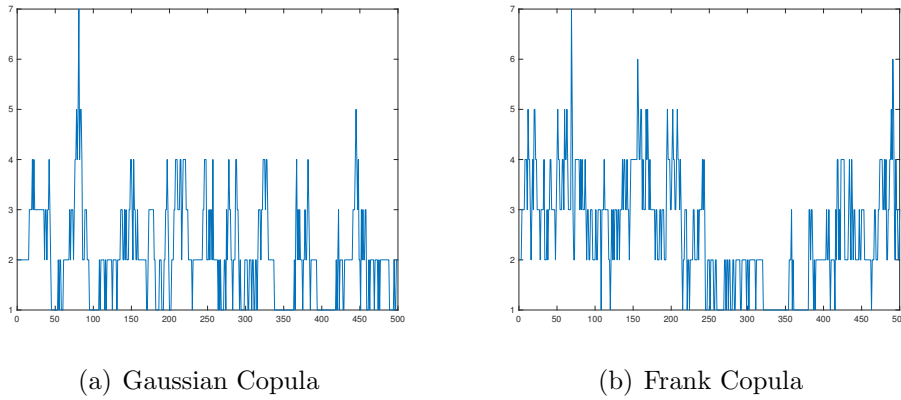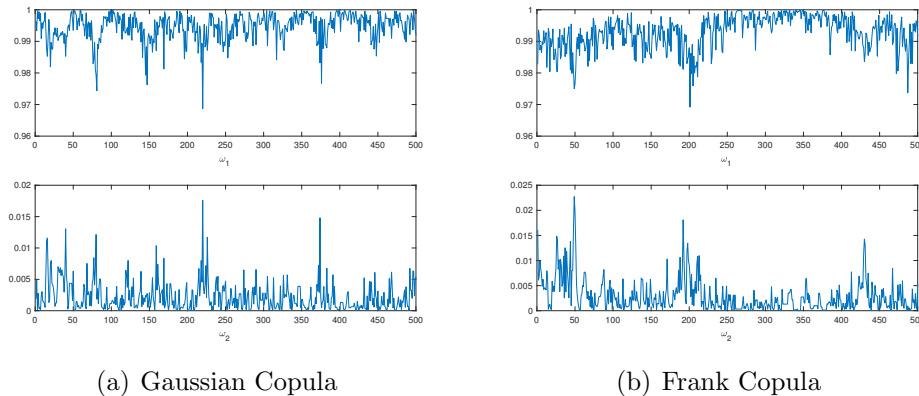Figure 6: Plots of the values of the first two most significant weights (y-axis) generated at each MCMC iteration (x-axis) for the Gaussian (left) and for the Frank copula (right) with sample size $n = 500$. The values of the first weight are plotted in the top panels, while the values of the second weight are plotted in the bottom panels.

the plot of the number of components generated at each MCMC iteration for both the Gaussian and the Frank copula. In Figure 5, we focus on the first calibration function, since the second calibration function gave similar results. Table 1 shows the summary statistics of the number of components generated at each MCMC iteration for both copulas, indicating that the posterior median of the number of components is equal to 2. For the two most significant components, we estimated the weights generated at each MCMC iteration for both copulas. In Figure 6 we show the trace plots of the last 500 iterations of the first two weights, as defined in Section (4). Figure 6 suggests that the first weight is much more important than the second weight, since the first weight tends to take values close to 1, while the second weight takes values close to 0. For each of the two components we also estimated the posterior mean of the copula correlation coefficient ρ_j (defined in eq. 5), obtaining, for the Gaussian copula, a value of 0.6701 for the first component and -0.9860 for the second component. On the other hand, for the Frank

15

copula we obtained a posterior mean of 0.7941 for the first component and -0.9722 for the second component, respectively.

	Min.	1st Quant.	Median	Mean	3rd Quant.	Max.
Gaussian	1	1	2	2.102	3	7
Frank	1	2	2	2.546	3	7

Table 1: Summary statistics of the number of components generated at each MCMC iterations for the first calibration function.

Finally, we consider an alternative Gaussian copula scenario, using the second calibration function. Table 2 shows the estimated weights of the mixture. The previous illustration, depicted in Figure 6, showed that one component was dominating the others. On the contrary, this experiment shows two dominating weights, $w_1 = 0.3237$ and $w_4 = 0.6286$, motivating the use of the proposed approach.

Comp	Posterior Mean	Comp	Posterior Mean	Comp	Posterior Mean
1	0.3237	4	0.6286	7	0.0024
2	0.0083	5	0.0096	8	0.0018
3	0.0193	6	0.0041	9	8.9908×10^{-04}

Table 2: Posterior means of the weights obtained with the second calibration function and sample size $n = 500$ for the Gaussian copula.

6 Real Data application

We now apply the proposed Bayesian nonparametric conditional copula method to a sample of 839 adolescent twin pairs, which is a subset of the National Merit Twin Study (Loehlin and Nichols, 2009, 2014). The dataset contains questionnaire data from 17 years old twins and their parents, where the twins were identified among 600.000 US high school juniors who took part to the National Merit Scholarship Qualifying Test (NMSQT).

The NMSQT was designed to measure cognitive aptitude, that is students' readiness for future intellectual or educational pursuits. The participants to the test include identical twins and same-sex fraternal twins who were asked to fill in a complete questionnaire in order to understand
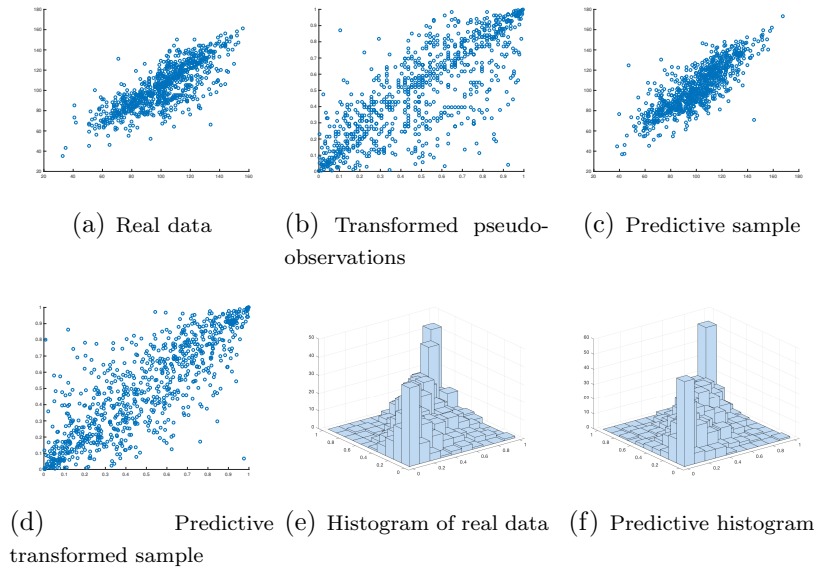
(a) Real data	(b) Transformed pseudo-observations	(c) Predictive sample

(d) Predictive transformed sample	(e) Histogram of real data	(f) Predictive histogram

Figure 7: Panels (a) and (b): scatterplots of the twins' overall scores for the real and pseudo-observations with respect to the mother's level of education; panels (c) and (d): scatterplots of the predictive and transformed predictive samples; panels (e) and (f): histograms of the real data and the predictive sample.

their school performance and attitude. Our purpose is to examine whether the relationship between twins' cognitive ability, measured by the NMSQT, is influenced by their socioeconomic status, measured by parent education and parental income. The variables we considered from this study are the overall measures of each twin's performance at school (obtained as the sum of individual scores in English Usage, Mathematics Usage, Social Science Reading, Natural Science Reading and Word Usage/Vocabulary), the mother's and father's level of education and the family income. The overall scores range from 30 to 160, the education covariates range from 0 to 6, while the family income covariate ranges from 0 to 7. The levels of the education covariates correspond to: less than 8-th grade, 8-th grade, part high school, high school graduate, part college or junior college, college graduate, and graduate or professional degree beyond the bachelor's degree. The levels of the income covariate correspond to values going from less than $5000 per year to over than $25000 per year.

As discussed in Section 1, the scatterplots in Figure 1 clearly show that there is a positive correlation between the twins' school performance and the
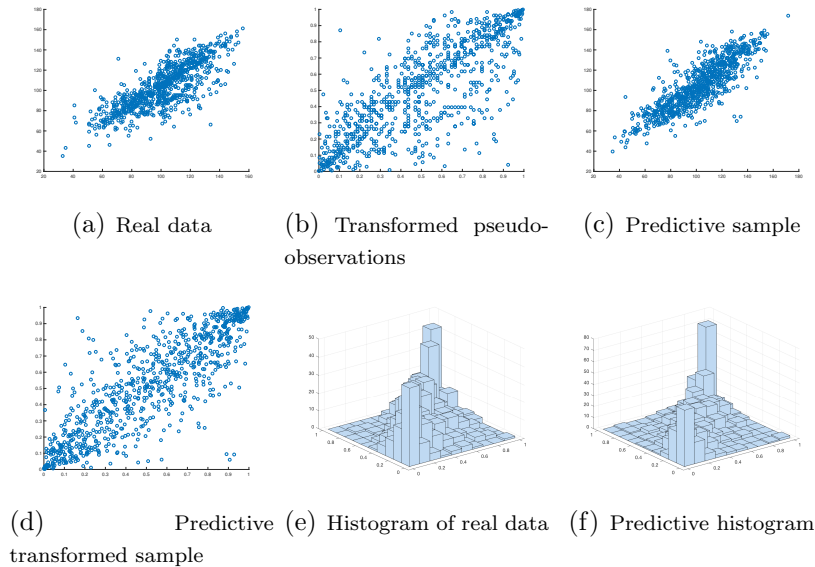
(a) Real data   (b) Transformed pseudo-observations   (c) Predictive sample

(d)   Predictive transformed sample   (e) Histogram of real data   (f) Predictive histogram

Figure 8: Panels (a) and (b): scatterplots of the twins' overall scores for the real and pseudo-observations with respect to the father's level of education; panels (c) and (d): scatterplots of the predictive and transformed predictive samples; panels (e) and (f): histograms of the real data and the predictive sample.

strength of dependence varies according to the values of a covariate, which is the mother's (panel (a)) or father's level of education (panel (b)) or the family income (panel (c)). In Figure 1 the effect of the covariates is illustrated by dots of different colours, where we notice that most of the light brown dots are grouped in the upper right corner, while the dark brown dots lie in the bottom left corner. Therefore, the higher the parents' education or family income, the higher the twins' school performance. In order to model the effect of a covariate, such as the mother's and father's education and family income, on the dependence between the overall scores of the twins, we implement the Bayesian nonparametric conditional copula model.

Figure 9 shows the relationship between the covariates of the twins dataset, where the lower-triangular panels represent pairwise scatterplots, the upper-triangular panels show pairwise Pearson's correlation coefficients, and the diagonal panels represent the histograms of each covariate. The scatterplots and Pearson's correlation coefficients in Figure 9 indicate a rather strong positive correlation between each pair of covariates, especially between the mother's and father's level of education. The high correlations indicate

18

that the data do not contain much information on the independent effects of each covariate, and suggest the inclusion of only one of them in the model. For this reason we decided to include only one of the redundant covariates at a time. Note that, with a different dataset, the methodology may be extended to include more than one covariate. However, model specification issues and increased computational costs must be carefully considered.
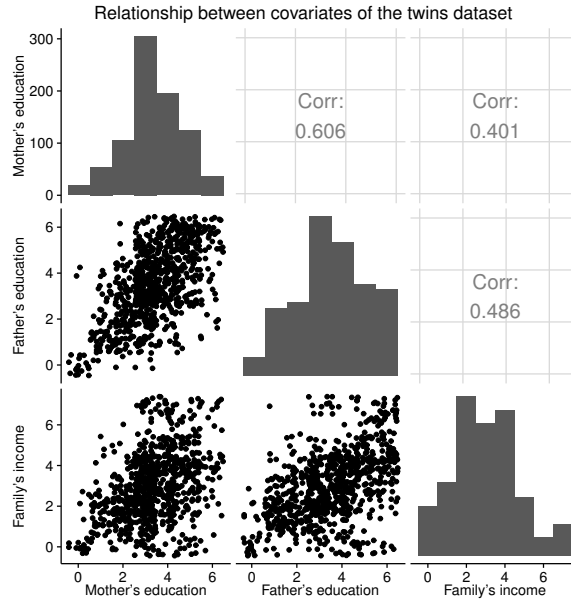
Figure 9: Relationship between the covariates of the twins dataset. The lower-triangular panels represent pairwise scatterplots, the upper-triangular panels show pairwise Pearson's correlation coefficients, and the diagonal panels represent the histograms of each covariate. (Note that jittering was used in the scatterplots to prevent overplotting).

Adopting the same priors of the simulation studies, we run the Gibbs sampling algorithm described in Section 4 for 4000 iterations. Figures 7, 8 and 10 show, for the mother's and father's education and family income, respectively, the scatterplots of the twins' overall scores using the real and transformed pseudo-observations (panels (a) and (b)), the scatterplots of the predictive and transformed predictive samples (panels (c) and (d)) and the histograms of the real and the predictive samples (panels (e) and (f)). Note that the pseudo-observations are obtained using the nonparametric estimation approach described in Section 4. From the comparison between the scatterplots and histograms of the real and

19

| (a) Real data | (b) Transformed pseudo-observations | (c) Predictive sample |

(d) Predictive transformed sample   (e) Histogram of real data   (f) Predictive histogram

Figure 10: Panels (a) and (b): scatterplots of the twins' overall scores for the real and pseudo-observations with respect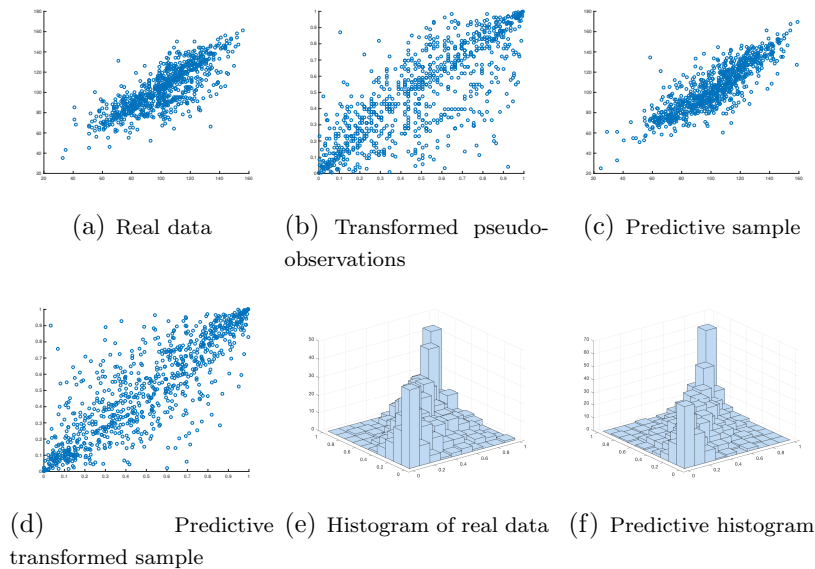 to the family income; panels (c) and (d): scatterplots of the predictive and transformed predictive samples; panels (e) and (f): histograms of the real data and the predictive sample.

predictive samples obtained with the three different covariates, it emerges that the Bayesian nonparametrics conditional copula model accurately captures the tail structures and the dependence patterns between the twins' overall scores. Moreover, the posterior means of the number of mixture components for the conditional copula are 26.82, 24.49 and 27.11, for the mother's and father's level of education and the family income, respectively, supporting the need for non-Gaussian copulas. Tables B.1, B.2 and B.3 in Appendix B list the posterior means of the component weights when the three different covariates are considered. We note that the good performance of this approach in tail modelling makes it suitable to various applications focusing on extremes. To quantify the degree of dependence between the twns' scores, we use the conditional Kendall's tau, which is a nonparametric measures of correlation, known as concordance, between two ranked variables (Y_1, Y_2) with respect to a covariate $X = x$. The conditional Kendall's tau takes the following form:

$$\tau(x) = 4 \int \int C_x(u_1, u_2) dC_x(u_1, u_2) - 1$$

20

where C_x is the appropriate conditional copula. Figure 11 shows Kendall's tau estimated from the model against the mother's (top panel) and father's level of education (middle panel) and the family income (bottom panel), together with 95% credible intervals. The plots clearly illustrate the negative effect of all three covariates on the dependence between the twins' overall scores. The effect is greater for the family income, where the Kendall's tau decreases from approximately 0.83 to 0.45, while for the parents' education levels the Kendall's tau decreases from approximately 0.8 to 0.6. Therefore, the higher the parents' education and family income, the better the socioeconomic status and the higher the differences between the twins' school performances. The cognitive aptitudes of twins from less advantaged families are more similar to each other than those from high income, highly educated families. Families of high socioeconomic status provide supportive and challenging environments, able to offer a wide range of opportunities and choices to their children, and allowing them to express themselves freely. Hence, twins raised in wealthy families are encouraged to develop differences in their traits, and may show rather dissimilar cognitive abilities, albeit high on average. On the contrary, families of low socioeconomic status offer scarce opportunities to their children and may provide limiting and restrictive environments. In less advantaged families, twins cannot develop their full potential and individuality, hence both tend to show low cognitive abilities.

This might suggest, as in Loehlin et al. (2009), an interaction between genetic and environmental factors. Genes multiply environmental inputs that support intellectual growth such that an increased socioeconomic status raises the average cognitive ability but also magnifies individual differences in cognitive ability (see Bates et al. (2013)).

7 Conclusion

In this paper we proposed a Bayesian nonparametric conditional copula approach to model the strength and type of dependence between two variables of interest and we applied the methodology to the National Merit Twin Study. In order to capture the dependence structure between two variables, we introduced two different calibration functions expressing the functional form of a covariate variable. The statistical inference was obtained implementing a slice sampling algorithm, assuming an infinite mixture model for the copula. The methodology combines the advantages of the conditional
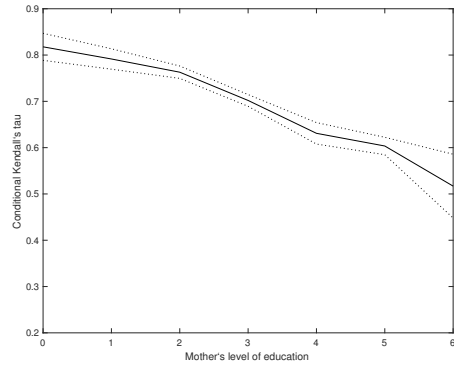
(a) Mother's level of education
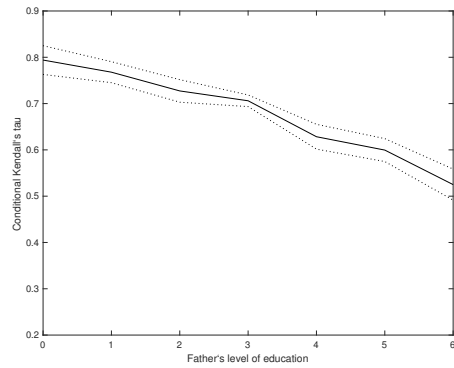
(b) Father's level of education
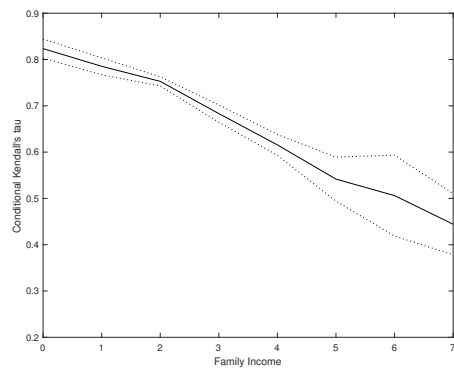
(c) Family income

Figure 11: Estimated Kendall's tau against the mother's (top panel) and father's level of education (middle panel) and the family income (bottom panel) and an approximate 95% credible intervals (dotted lines)

22

copula approach with the modeling flexibility of Bayesian nonparametrics.

The simulation studies illustrated the excellent performance of our model with three distinct copula families and different sample sizes. The application to the twins data revealed the importance of the environment in the development of twins' cognitive abilities and suggests that environmental factors are more influential in families with higher socioeconomic position. On the contrary, other factors, such as genetic causes, may be more dominant in families with lower socioeconomic position.

Although this paper focuses on bivariate copula models, the methodology can be extended to multivariate copulas including more than one covariate. However, the inclusion of multiple covariates needs special attention regarding the choice of variables prior to estimate the calibration functions. Moreover, the increasing computational cost due to the additional covariates should be taken carefully into consideration.

Acknowledgements

References

Abegaz, F., Gijbels, I., and Veraverbeke, N. (2012). Semiparametric estimation of conditional copulas. *Journal of Multivariate Analysis* **110,** 43–73.

Acar, E. F., Craiu, R. V., and Yao, F. (2010). Dependence calibratin in conditional copulas: a nonparametric approach. *Biometrics* **67,** 445–453.

Andersen, E. (2005). Two-stage estimation in copula models used in family studies. *Lifetime Data Analysis* **11,** 333–350.

Baker, S. (2016). The latent class twin method. *Biometrics* **3,** 827–834.

Bartram, S., Taylor, S., and Wang, Y. (2007). The Euro and European financial market dependence. *Journal of Banking and Finance* **31,** 1461–1481.

Bates, T., Lewis, G., and Weiss, A. (2013). Childhood socioeconomic status amplifies genetic effects on adult intelligence. *Psychological Science* **24,** 2111–2116.

Braekers, R. and Veraverbeke, N. (2005). A copula-graphic estimator for the conditional survival function under dependent censoring. *Canadian Journal of Statistics* **33,** 429–447.

Burda, M. and Prokhorov, A. (2014). Copula based factorization in Bayesian multivariate infinite mixture models. *Journal of Multivariate Analysis* **127,** 200–213.

Cherubini, U., Luciano, E., and Vecchiato, W. (2004). Copula methods in finance. *John Wiley and Sons* .

Craiu, R. V. and Sabeti, A. (2012). In mixed company: Bayesian inference for bivariate conditional copula models with discrete and continuous outcomes. *Journal of Multivariate Analysis* **110,** 106–120.

Emura, T. and Wang, W. (2012). Nonparametric maximum likelihood estimation for dependent truncation data based on copulas. *Journal of Multivariate Analysis* **110,** 171–188.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* **89,** 268–277.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90,** 577–588.

Ferguson, T. (1983). *Bayesian density estimation by mixtures of normal distributions*, pages 287–303. New York: Academic Press.

Gijbels, I., Omelka, M., and Veraverbeke, N. (2012). Multivariate and functional covariates and conditional copulas. *Electronic Journal of Statistics* **6,** 1273–1306.

Gijbels, I., Veraverbeke, N., and Omelka, M. (2011). Conditional copulas, association measures and their applications. *Computational Statistics and Data Analysis* **55,** 1919–1932.

Huang, X. and Zhang, N. (2008). Regression survival analysis with an assumed copula for dependent censoring: a sensitivity analysis approach. *Biometrics* **64,** 1090–1099.

Joe, H. (2014). *Dependence Modeling with Copulas.* Chapman & Hall.

Jondeau, E. and Rockinger, M. (2006). The copula-GARCH model of conditional dependencies: An international stock market application. *Journal of International Money and Finance* **25,** 827–853.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing* **21,** 93–105.

Klein, N. and Kneiss, T. (2016). Simultaneous inference in structured additive conditional copula regression models: a unifying bayesian approach. *Statistics and Computing* **26,** 841–860.

Kolev, N., dos Anjos, U., and Vaz de Mendes, B. (2006). Copulas: a review and recent developments. *Stochastic Models* **22,** 617–660.

Lo, A. (1984). On a class of Bayesian nonparametric estimates i: density estimates. *Annals of Statistics* **12,** 351–357.

Loehlin, J., Harden, K., and Turkheimer, E. (2009). The effect of assumptions about parental assortative mating and genotype–income correlation on estimates of genotype–environment interaction in the national merit twin study. *Behavior Genetics* **39,** 165–169.

Loehlin, J. and Nichols, R. (2009). The National Merit twin study. *Harvard Dataverse, V3, http://hdl.handle.net/1902.1/13913* .

Loehlin, J. and Nichols, R. (2014). Heredity, environment and personality: A study of 850 sets of twins. *University of Texas Press* .

Owzar, K., Jung, S.-H., and Sen, P. K. (2007). A copula approach for detecting prognostic genes associated with survival outcome in microarray studies. *Biometrics* **63,** 1089–1098.

Patton, A. J. (2006). Modelling asymmetric exchange rate dependence. *International Economic Review* **47,** 527–556.

Pitt, M., Chan, D., and Kohn, R. (2006). Efficient bayesian inference for gaussian copula regression models. *Biometrika* **93,** 537–554.

Plummer, M. (2008). Penalized loss functions for bayesian model comparison. *Biostatistics* **9,** 523–539.

Sklar, A. (1959). Fonctions de réparation à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris* **8,** 229–231.

Smith, M. S., Gan, Q., and Kohn, R. J. (2012). Modelling dependence using skew t copulas: Bayesian inference and applications. *Journal of Applied Econometrics* **27,** 500–522.

Stander, J., Dalla Valle, L., Taglioni, C., and Cortina Borja, M. (2015a). Bayesian copula modelling in the presence of covariates. In *Book of Abstracts, 8th CSDA International Conference on Computational and Methodological Statistics (ERCIM 2015)*, page 179. Angela Blanco-Fernandez and Gil Gonzalez-Rodriguez.

Stander, J., Dalla Valle, L., Taglioni, C., and Cortina Borja, M. (2015b). Bayesian copula modelling in the presence of covariates. In *Abstracts Booklet, RSS conference 2015*, page 136.

Taglioni, C., Stander, J., Dalla Valle, L., and Cortina-Borja, M. (2016). Bayesian copula modelling in the presence of covariates. In *Book of Abstracts, ISBA 2016 World Meeting on Bayesian Statistics, Cagliari (Italy)*, pages 413–414. S. Cabras and M. Guindani, CUEC Cooperativa Universitaria Editrice Cagliaritana.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation* **36,** 45–54.

Wang, X., Guo, X., He, M., and Zhang, H. (2011). Statistical inference in mixed models and analysis of twin and family data. *Biometrics* **67,** 987–995.

Wu, J., Wang, X., and Walker, S. (2015). Bayesian nonparametric estimation of a copula. *Journal of Statistical Computation and Simulation* **85,** 103–116.

Yin, G. and Yuan, Y. (2009). Bayesian dose finding in oncology for drug combinations by copula regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **58,** 211–224.

A   Gibbs sampling details

Let $\mathcal{D}_j = \{i = 1, \ldots, n : d_i = j\}$ be the set of indexes of the observations allocated to the j-th component of the mixture, while $\mathcal{D} = \{j : \mathcal{D}_j \neq \emptyset\}$ is the set of indexes of non-empty mixtures components. Let $D^* = \sup\{\mathcal{D}\}$ be the number of stick-breaking components used in the mixture. As in Kalli et al. (2011), the sampling of infinite elements of $\boldsymbol{\pi}$ and $\boldsymbol{\beta}$ is not necessary, since only the elements of the full conditional probability density functions of D are need.

The maximum number of stick-breaking components to be sampled is:

$$N^* = \max\{i = 1, \ldots, n | N_i^*\},$$

where N_i^* is the smallest integer such that $\sum_{j=1}^{N_i^*} w_j > 1 - z_i$.

A.1   Update of π

We update the stick-breaking components and consequently the weights w_j based on the equation $w_j = \pi_j \prod_{k<j}(1 - \pi_k)$. Assuming that π_j is distributed as a Beta $(\mathcal{B}e(1, \lambda))$, the full conditional distribution of π_j is:

$$\pi_j | \cdots \sim \mathcal{B}e(1 + \#\{d_i = j\}, \lambda + \#\{d_i > j\}), \tag{14}$$

where $\#\{d_i = j\}$ are the number of d_i equal to j and $\#\{d_i > j\}$ is the number of d_i greater than j for $j < D^*$.

On the other hand, if $j = D^* + 1, \ldots, N^*$ we have that

$$\pi_j | \cdots \sim \mathcal{B}e(1, \lambda).$$

A.2   Update of Z

From the full likelihood function (13), z_i follows a uniform distribution

$$z_i | \cdots \sim \mathcal{U}(0, w_{d_i}) \tag{15}$$

and it is sampled accordingly.

27

A.3 Update of D

The allocation variable d_i values lie between 0 and N_i and the density of d_i satisfies

$$P(d_i = j | \dots) \propto \mathbb{I}(z_i < w_{d_i}) c_{\rho(x_i | \boldsymbol{\beta}_{d_i})}(u_i, v_i). \tag{16}$$

A.4 Update of β

The full conditional of the vector of parameters $\boldsymbol{\beta}_k$, for $k \geq 1$ is:

$$f(\boldsymbol{\beta}_k | \dots) \propto \pi(\boldsymbol{\beta}_k) \prod_{d_i = k} c_{\rho(x_i | \boldsymbol{\beta}_k)}(u_i, v_i), \tag{17}$$

where $\pi(\boldsymbol{\beta}_k)$ is the prior on $\boldsymbol{\beta}$. Since the (17) is not a standard distribution, we used a Random Walk Metropolis Hastings.

B Supplementary Materials

Comp	Posterior Mean	Comp	Posterior Mean	Comp	Posterior Mean
1	0.0376	17	0.0393	33	1.0542×10^{-04}
2	0.0402	18	0.0389	34	6.0567×10^{-05}
3	0.0454	19	0.0233	35	3.1244×10^{-05}
4	0.1029	20	0.0233	36	2.0493×10^{-05}
5	0.0832	21	0.0285	37	1.2760×10^{-05}
6	0.0578	22	0.0221	38	4.8162×10^{-06}
7	0.0504	23	0.0167	39	3.5861×10^{-06}
8	0.0408	24	0.0180	40	1.4768×10^{-06}
9	0.0360	25	0.0091	41	9.4360×10^{-07}
10	0.0467	26	0.0129	42	1.3644×10^{-07}
11	0.0352	27	0.0051	43	6.5268×10^{-08}
12	0.0431	28	0.0034	44	5.4506×10^{-08}
13	0.0402	29	0.0036	45	3.4268×10^{-08}
14	0.0351	30	9.4652×10^{-04}	46	5.4973×10^{-13}
15	0.0353	31	4.6736×10^{-04}	47	6.0754×10^{-12}
16	0.0240	32	2.5834×10^{-04}	48	1.0977×10^{-11}

Table B.1: Posterior means of the mixture component weights for the twins' performance scores with respect to the mother's level of education.

Comp	Posterior Mean	Comp	Posterior Mean	Comp	Posterior Mean
1	0.0514	18	0.0313	35	7.4762×10^{-06}
2	0.0516	19	0.0260	36	4.8028×10^{-06}
3	0.0536	20	0.0257	37	3.5777×10^{-06}
4	0.0379	21	0.0197	38	7.2414×10^{-07}
5	0.0675	22	0.0246	39	5.6380×10^{-07}
6	0.0743	23	0.0063	40	1.5620×10^{-07}
7	0.0414	24	0.0097	41	5.4960×10^{-08}
8	0.0523	25	0.0049	42	1.0994×10^{-07}
9	0.0384	26	0.0037	43	1.5524×10^{-08}
10	0.0497	27	0.0013	44	1.1536×10^{-08}
11	0.0351	28	5.7852×10^{-04}	45	6.4516×10^{-09}
12	0.0495	29	3.5660×10^{-04}	46	6.2531×10^{-09}
13	0.0304	30	1.6825×10^{-04}	47	1.4255×10^{-08}
14	0.1066	31	9.9642×10^{-05}	48	3.4571×10^{-09}
15	0.0270	32	4.6900×10^{-05}	49	7.0123×10^{-10}
16	0.0319	33	2.6103×10^{-05}	50	1.8651×10^{-10}
17	0.0469	34	1.1375×10^{-05}	51	8.0720×10^{-10}

Table B.2: Posterior means of the mixture component weights for the twins' performance scores with respect to the father's level of education

Comp	Posterior Mean	Comp	Posterior Mean	Comp	Posterior Mean
1	0.0836	17	0.0359	33	1.1455×10^{-04}
2	0.0681	18	0.0236	34	6.7836×10^{-05}
3	0.0486	19	0.0277	35	3.8637×10^{-05}
4	0.0750	20	0.0233	36	1.9610×10^{-05}
5	0.0074	21	0.0205	37	1.0765×10^{-05}
6	0.0569	22	0.0209	38	6.4166×10^{-06}
7	0.0492	23	0.0172	39	4.1155×10^{-06}
8	0.0415	24	0.0105	40	2.7336×10^{-06}
9	0.0319	25	0.0097	41	9.7839×10^{-07}
10	0.0340	26	0.0080	42	3.6532×10^{-07}
11	0.0288	27	0.0042	43	3.5966×10^{-07}
12	0.0451	28	0.0023	44	1.7294×10^{-07}
13	0.0395	29	0.0018	45	1.9279×10^{-08}
14	0.0365	30	0.0013	46	1.5505×10^{-08}
15	0.0281	31	4.1135×10^{-04}	47	8.1074×10^{-09}
16	0.0377	32	2.2382×10^{-04}	48	8.5996×10^{-09}

Table B.3: Posterior means of the mixture component weights for the twins' performance scores with respect to the family income.

(a) $x = 0$, $n_0 = 24$ (b) $x = 1$, $n_1 = 90$ (c) $x = 2$, $n_2 = 98$

(d) $x = 3$, $n_3 = 212$ (e) $x = 4$, $n_4 = 178$ (f) $x = 5$, $n_5 = 122$
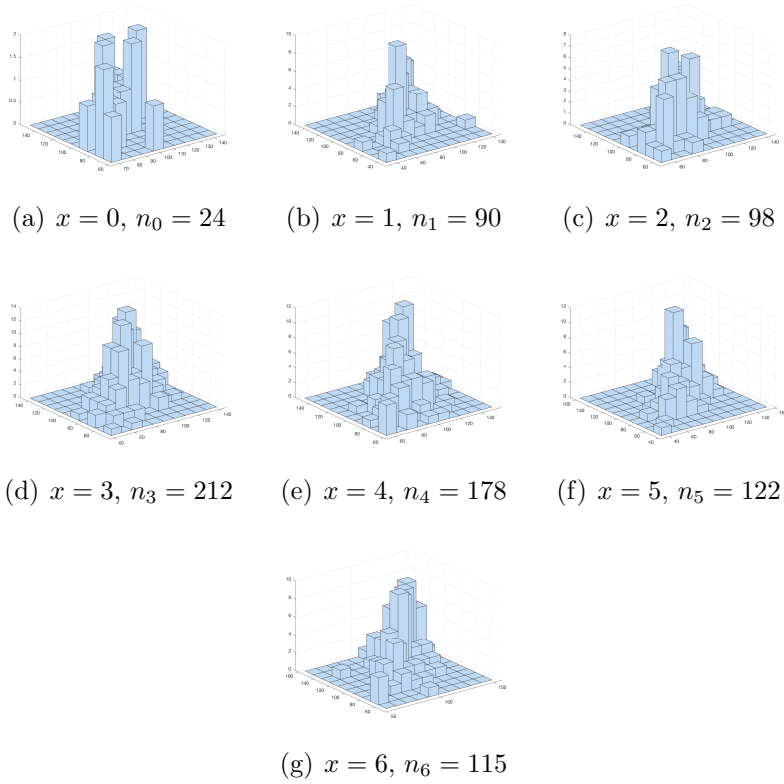
(g) $x = 6$, $n_6 = 115$

Figure B.1: Three-dimensional histograms of the twins' performance scores for different values of the father's education. In each panel x denotes the level of the covariate and n_x denotes the relevant sample size.

(a) $x = 0$, $n_0 = 62$ (b) $x = 1$, $n_1 = 92$ (c) $x = 2$, $n_2 = 200$

(d) $x = 3$, $n_3 = 166$ (e) $x = 4$, $n_4 = 183$ (f) $x = 5$, $n_5 = 73$

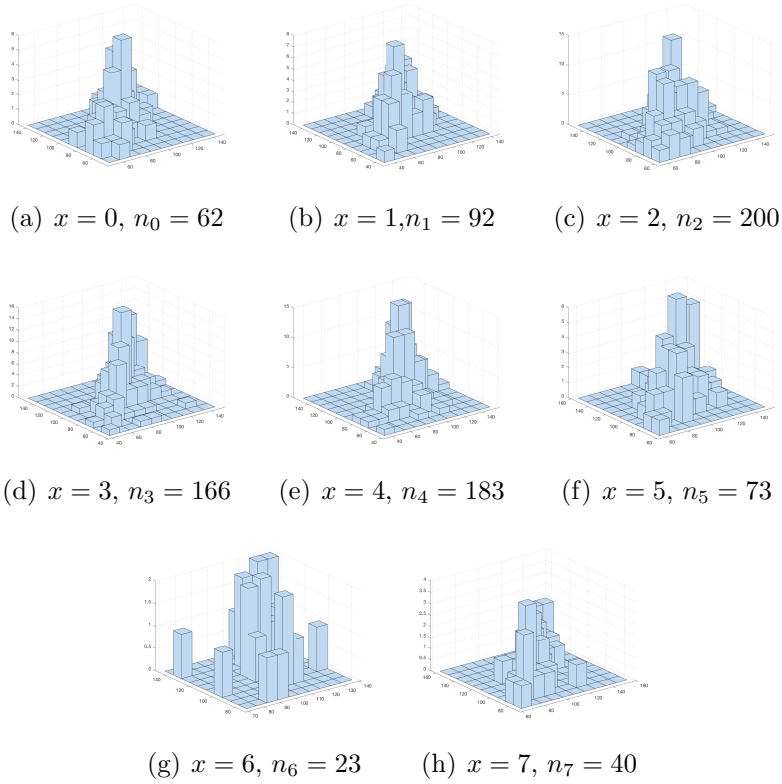(g) $x = 6$, $n_6 = 23$ (h) $x = 7$, $n_7 = 40$

Figure B.2: Three-dimensional histograms of the twins' performance scores for different values of the family income. In each panel x denotes the level of the covariate and n_x denotes the relevant sample size.