Luciana Dalla Valle, Fabrizio Leisen, and Luca Rossini

Bayesian Nonparametric Conditional Copula Estimation of Twin Data
Bayesian Nonparametric Conditional Copula Estimation of Twin Data

Luciana Dalla Valle  
*University of Plymouth, UK*

Fabrizio Leisen  
*University of Kent, UK*

Luca Rossini  
*Ca’ Foscari University of Venice*

**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, Wang, and Walker (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

**JEL Codes**
C11, C13, C14, C51

**Address for correspondence:**
Luca Rossini  
Department of Economics  
Ca’ Foscari University of Venice  
Cannaregio 873, Fondamenta S.Giobbe  
30121 Venezia - Italy  
Phone: (+39) 041 2349230  
Fax: (+39) 041 2349176  
e-mail: luca.rossini@unive.it

This Working Paper is published under the auspices of the Department of Economics of the Ca’ Foscari University of Venice. Opinions expressed herein are those of the authors and not those of the Department. The Working Paper series is designed to divulge preliminary or incomplete work, circulated to favour discussion and comments. Citation of this paper should consider its provisional character.
Bayesian Nonparametric Conditional Copula Estimation of Twin Data

Luciana Dalla Valle 1  Fabrizio Leisen2
Luca Rossini3

1 University of Plymouth, U.K.  2 University of Kent, U.K.
3 Ca’ Foscari University of Venice, Italy

March 9, 2016

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, Wang, and Walker (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
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)). Bioecological theory states that environmental factors modify the heritability of certain characteristics, such as cognitive ability, that is the readiness for future intellectual or educational pursuits. Several studies have found that cognitive ability is more heritable among those raised in higher socioeconomic status families and therefore the environment can moderate the effect of genes. Bates, Lewis, and Weiss (2013) studied the interactions between environmental and genetic effects to intelligence in twins, showing that higher socioeconomic status is associated with higher intelligence scores.

The aim of this paper is to correctly analyse the effect of environmental 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 red denotes low values, while light yellow denotes high values). From Figure 1 it is clear that the twins’ school performances are strongly correlated and positively influenced by all socioeconomic variables.

In order to model the dependence structure between the twins’ school performances, we used copulas, which are popular modeling approaches in multivariate statistics allowing the separation of the marginal components 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 $C : [0,1]^d \to [0,1]$, 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, dos Anjos, and Vaz de Mendes (2006) and Cherubini, Luciano, and Vecchiato (2004)). In particular, applications of copula models involved lifetime data analysis (Andersen (2005)), survival analysis of Atlantic halibut (Braekers and Veraverbeke (2005)) and transfusion-related AIDS and cancer analysis (Emura and Wang (2012), Huang and Zhang (2008) and Owzar, Jung, and Sen (2007)).

The introduction of covariate adjustments to copulas has attracted an increased interest in recent years, since it allows the dependence structure to be explained by a specific covariate. Craiu and Sabeti (2012) propose a conditional copula approach in regression settings where the bivariate outcome can be mixed or continuous. Patton (2006) introduce time-variation in the dependence structure of ARMA models (see also Jondeau and Rockinger (2006) and Bartram, Taylor, and Wang (2007) for other applications of time-series analysis to dependence modelling). The paper of Acar, Craiu, and Yao (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, Gijbels, and Veraverbeke (2012) and Gijbels, Omelka, and Veraverbeke (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 expectancies of males and females at birth. Following this literature, we adopt a conditional copula approach to model the effect of a covariate, such as the parents’ education or the family income, on the strength of dependence between twins’ school performances.

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. Nonetheless, the choice of the copula family may be controversial and it is still an open problem (see Joe (2014)). To overcome this issues, Wu et al. (2015) 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. 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 included a conditional covariate component to explain the variables 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

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

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

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

(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

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

(2)

### 2.2 Bayesian nonparametric copula density estimation

Wu et al. (2015) introduced a Bayesian nonparametric copula density estimation by using an infinite mixture of Gaussian copulas. Let $\Phi_{\rho}(y_1, y_2)$ denote the standard bivariate normal distribution function with correlation coefficient $\rho$. Then, $C_{\rho}$ is the copula corresponding to $\Phi_{\rho}$, taking the form:

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

(3)

where $\Phi$ is the univariate standard normal distribution function. The Gaussian copula density is:

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

(4)

where the covariance 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

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

(5)

where the weights $w_j$’s sum up to 1 and the $\rho_j$’s vary in $(-1, 1)$. 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

$$f(y_1, y_2) = \sum_{j=1}^{\infty} w_j \mathcal{N}((y_1, y_2)|((\mu_{y_1j}, \mu_{y_2j}), \Sigma_j))$$

(6)
where \( N((y_1, y_2)|(\mu_{1j}, \mu_{2j}), \Sigma_j) \) is the joint bivariate normal density with mean vector \((\mu_{1j}, \mu_{2j})\) and covariance matrix \(\Sigma_j\) (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, Griffin, and Walker (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

Suppose that the dependence parameter of the conditional copula depends on the covariate as follows

\[
C_x(u, v) = C_x(u, v; \theta(x)),
\]

where \( \theta(x) \) is a function of the covariate, called *calibration function*. Following Abegaz et al. (2012), we model the function \( \theta(x) \) as a polynomial of degree \( p \). It is important to highlight that in many copula families the parameter space is restricted. In contrast, polynomial modeling inherently assumes that any value on the real line can be taken. For this reason, we introduce the transformation \( \psi(\theta(x)) = \eta(x) \) such that \( \theta(x) = \psi^{-1}(x) \), assuming that the inverse transformation \( \psi^{-1} \) exists. We suppose that the \((p + 1)\)-th continuous derivative of the \( \eta \) function exists at the \( x \) point. We approximate our calibration function \( \eta(X_i) \) for a data point \( X_i \) by a Taylor expansion of degree \( p \):

\[
\eta(X_i) \approx \eta(x) + \eta'(x)(X_i - x) + \cdots + \frac{\eta^{(p)}(x)}{p!}(X_i - x)^p = x_{i,x}^t \beta \quad (6)
\]

where \( x_{i,x} = (1, X_i - x, \ldots, (X_i - x)^p)^t \) and \( \beta = (\beta_0, \ldots, \beta_p)^t \), with \( \beta_i = \eta^{(i)}(x)/i! \). In our implementation, as in Fan and Gijbels (1996) and Acar et al. (2010), we use the commonly adopted local linear fit, assuming \( p = 1 \).

Our model is based on an infinite mixture of Gaussian copulas, adopting a similar approach to Wu et al. (2015). However, while the authors restrict their attention to unconditional Gaussian copulas, we extend their work by
considering a conditional Gaussian copulas, assuming that the correlation coefficient $\rho$ depends on the covariate $X$. Hence, our Bayesian nonparametric conditional copula model is defined as

$$c(u, v|x) = \sum_{j=1}^{\infty} w_j c_{\rho_j(\theta_j(x))}(u, v|x)$$  \hspace{2cm} \text{(7)}

where the weights $w_j$’s sum up to one and the $\rho_j$’s vary in $(-1, 1)$. The parameter $\rho$ is linked to $\theta(x)$ via the following transformation

$$\rho(x) = \frac{2}{|\theta(x)| + 1} - 1$$  \hspace{2cm} \text{(8)}

where $\theta \in [0, \infty)$ is modelled by the polynomial approximation described in equation (6).

4 Posterior sampling algorithm

Following equation (7), given the observations $(u_i, v_i)$ 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:

$$c(u_i, v_i|x_i) = \sum_{j=1}^{\infty} w_j c_{\rho_j(\theta_j(x_i))}(u_i, v_i|x_i)$$  \hspace{2cm} \text{(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 $\pi_j$ are distributed as a $\text{Be}(1, b)$, $b > 0$. 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:

$$c(u_i, v_i, z_i|x_i) = \sum_{j=1}^{\infty} \mathbb{I}(z_i < w_j) c_{\rho_j(\theta_j(x_i))}(u_i, v_i|x_i).$$  \hspace{2cm} \text{(10)}$$
The introduction of the slice variable \( z_i \) reduces the sampling complexity to the analogous of a finite mixture model. In particular, letting

\[
A_w = \{ j : z_i < w_j \},
\]

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:

\[
c(u_i, v_i, z_i, d_i|x_i) = I(z_i < w_{d_i})c_{\rho_{d_i}(\theta_{d_i}(x_i))}(u_i, v_i|x_i)
\]

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

\[
\prod_{i=1}^{n} c(u_i, v_i, z_i, d_i|x_i) = \prod_{i=1}^{n} I(z_i < w_{d_i})c_{\rho_{d_i}(\theta_{d_i}(x_i))}(u_i, v_i|x_i).
\]

We use the notation \((U, V) = \{ i = 1, \ldots, n : (u_i, v_i) \}, X = \{ x_1, \ldots, x_n \}\) to describe the pseudo-observations and the covariate values, respectively, \( \theta = \{ \theta_1, \theta_2, \ldots, \theta_n \} \) and \( \rho = \{ \rho_1, \rho_2, \ldots, \rho_n \} \) to denote the calibration and the parameter values, and \( D = \{ d_1, \ldots, d_n \}, Z = \{ z_1, \ldots, z_n \} \) and \( \pi = \{ \pi_1, \pi_2, \ldots \} \) to denote the new variables that we have introduced in this Section. Therefore, we used a Gibbs sampler to simulate iteratively from the posterior distribution function, according to the following steps:

1. The stick-breaking components \( \pi \) are updated given \([Z, D, \theta, \rho, (U, V), X] \);
2. The latent slice variables \( Z \) are updated given \([\pi, D, \theta, \rho, (U, V), X] \);
3. The allocation variables \( D \) are updated given \([\pi, Z, \theta, \rho, (U, V), X] \);
4. The calibration and parameter vectors \((\theta, \rho)\) are updated given \([\pi, Z, D, (U, V), X] \).

The Gibbs sampling details are explained in Appendix.
5 Simulation experiments

This section illustrates the performance of our Bayesian nonparametric conditional copula model with simulated data. We generate datasets \((U, V)\) of sizes \(n = 250, 500\) and 1000 from elliptical and archimedean copula families, such as the Gaussian, Frank and Double Clayton copula, which combines the regular Clayton copula with its 90° rotation, allowing positive and negative dependence modelling. 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 simulated data and for the Normal copula mixture, we assume the following calibration functions for \(\theta\):

1. The correlation coefficient is a function of the \(X\) variable as follows:
   \[
   \rho(x) = \frac{2}{(\theta(x) + 1)} - 1,
   \]
   where \(\theta(x)\) is a quadratic calibration function:
   \[
   \theta(x) = \alpha + \beta x^2
   \]
   adopting uniform priors \(U(0, 0.2)\) for \(\alpha\) and \(\beta\).

2. The correlation coefficient is a function of the \(X\) variable as follows:
   \[
   \rho(x) = \frac{2}{(|\theta(x)| + 1)} - 1
   \]
   where \(\theta(x)\) is a quadratic calibration function:
   \[
   \theta(x) = \alpha + \beta x + \gamma \exp(-\delta x^2)
   \]
   adopting uniform priors \(U(0, 0.2)\) for \(\alpha, \beta, \gamma\) and \(\delta\).

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 focussed on 3500 burn-in iterations, since 500 burn-in iterations gave very similar results.

Figures 2, 3 and 4 illustrate the results of the application of the Bayesian nonparametric conditional copula model to data simulated from a gaussian copula, with sample sizes \(n = 250, 500\) and 1000, respectively. Figures
5, 6 and 7 illustrate similar results for the Frank copula; while Figures 8, 9 and 10 illustrate analogous results for the Double Clayton copula. In Figures 2-10, panels (a), (b), (c) and (d) show the scatter plots and histograms of the simulated data and the predictive samples, respectively, 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, respectively, obtained using the second calibration function. The comparison between the simulated and predictive outputs highlights the excellent fit of the Bayesian nonparametric conditional copula model using either calibration function and with different sample sizes. The model performance appears to be consistent across all three copula families, demonstrating that the approach is suitable to model different dependence patterns and tail structures.

6 Real Data applications

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 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.

As discussed in Section 1, the scatterplots in Figure 1 clearly show that
there is a strong correlation between the twins’ school performance and the strength of dependence varies according to the values of a covariate, which is the mother’s (top panel) or father’s level of education (middle panel) or the family income (bottom panel). In Figure 1 the effect of the covariates is illustrated by dots of different colours, where we notice that most of the light yellow dots are grouped in the upper right corner, while the dark red 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.

Adopting the same priors of the simulation studies, we run the Gibbs sampling algorithm described in Section 4 for 4000 iterations. Figures 11, 12 and 13 show, with respect to 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)). From the comparison between the scatterplots and histograms of the real and 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. We note that the good performance of this approach in tail modelling makes it suitable to various applications focussing on extremes. Figure 14 shows the 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). 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 0.43, while for the parents’ education levels the Kendall’s tau decreases from approximately 0.8 to 0.57. 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 than those from high income, highly educated families. This might suggest, as in Loehlin, Harden, and Turkheimer (2009), that environmental factors are more influential in families with lower socio-economic position. On the contrary, other factors, such as
genetic causes, may be more dominant in families with higher socio-economic position.

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 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 belonging to low socioeconomic classes and suggest that genetic factors are more influential in families with higher socioeconomic status.

Although this paper focusses 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

Fabrizio Leisen was supported by the European Community’s Seventh Framework Programme [FP7/2007-2013] under grant agreement no: 630677.

References


A.1 Gibbs sampling details

Let $D_j = \{i = 1, \ldots, n : d_i = j\}$ be the set of indexes of the observations allocated to the $j$-th component of the mixture, while $D = \{j : D_j \neq \emptyset\}$ is the set of indexes of non-empty mixtures components. Let $D^* = \sup \{D\}$ be the number of stick-breaking components used in the mixture. As in Kalli et al. (2011), the sampling of infinite elements of $\pi$ and $(\rho, \theta)$ 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.2 Update of $\pi$

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 $\pi_j$ is distributed
as a Beta \((\mathcal{B}e(1,b))\), the full conditional distribution of \(\pi_j\) is:

\[
\pi_j|\cdots \sim \mathcal{B}e(1 + \#\{d_i = j\}, b + \#\{d_i > j\}) \tag{18}
\]

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

\[
\pi_j|\cdots \sim \mathcal{B}e(1,b).
\]

### A.3 Update of \(Z\)

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

\[
z_i|\cdots \sim \mathcal{U}(0, w_{d_i}) \tag{19}
\]

and it is sampled accordingly.

### A.4 Update of \(D\)

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

\[
P(d_i = j|\cdots) \propto \mathbb{I}(z_i < w_{d_i})c_{\rho_j(\theta_{d}(x_i))}(u_i, v_i|x_i). \tag{20}
\]

### A.5 Update of \((\theta, \rho)\)

The full conditional of the parameters of the calibration function \(\theta(x)\) is:

\[
f(\theta(x)|\cdots) \propto f(\theta(x)) \prod_{d_i=j} c_{\rho_d(\theta_{d}(x_i))}(u_i, v_i|x_i), \tag{21}
\]

where \(f(\theta(x))\) is the prior on \(\theta(x)\). Since the (21) is not a standard distribution, we used a Random Walk Metropolis Hastings with proposal distributed as a truncated Normal in the interval \((0, \infty)\). Consequently \(\rho\) is derived using equation (8).

16
Figure 1: Scatterplots of the twins overall scores with respect to the mother’s (top panel) and father’s level of education (middle panel) and family income (bottom panel).
Figure 2: Gaussian copula with sample size $n = 250$. 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.
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.
Figure 4: Gaussian copula with sample size $n = 1000$. 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.
Figure 5: Frank copula with sample size $n = 250$. 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.
Figure 6: 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.
Figure 7: Frank copula with sample size $n = 1000$. 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.
Figure 8: Double Clayton copula with sample size $n = 250$. 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.
Figure 9: Double Clayton 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.
Figure 10: Double Clayton copula with sample size $n = 1000$. 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.
Figure 11: 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 sample; panels (e) and (f): histograms of the real data and the predictive sample.
Figure 12: 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 sample; panels (e) and (f): histograms of the real data and the predictive sample.
Figure 13: 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 sample; panels (e) and (f): histograms of the real data and the predictive sample.
Figure 14: Estimated Kendall’s tau against the mother’s (top panel) and father’s level of education (middle panel) and the family income (bottom panel).