Assessing Data Support for the Simplifying Assumption in Bivariate Conditional Copulas

Evgeny Levi and Radu V. Craiu

Abstract The paper considers the problem of establishing data support for the simplifying assumption (SA) in a bivariate conditional copula model. It is known that SA greatly simplifies the inference for a conditional copula model, but standard tools and methods for testing SA in a Bayesian setting tend to not provide reliable results. After splitting the observed data into training and test sets, the method proposed will use a flexible Bayesian model fit to the training data to define tests based on randomization and standard asymptotic theory. Its performance is studied using simulated data. The paper's supplementary material also discusses theoretical justification for the method and implementations in alternative models of interest, e.g. Gaussian, Logistic and Quantile regressions.

Keywords Calibration function · Conditional copula · Permutation · Simplifying assumption

1 Introduction

A copula is a mathematical concept often used to model the joint distribution of several random variables. The applications of copula models permeate a number of fields where of interest is the simultaneous study of dependent variables, e.g. [8, 10, 13, 17]. The propagation of copula-related ideas in probability and statistics started with [19] which proved that for a random vector $(Y_1, \ldots, Y_k)$ with cumulative distribution function (CDF) $H(y_1, \ldots, y_k)$ and marginal continuous CDFs $F_i(y_i)$, $i = 1, \ldots, k$, there exists a unique *copula* $C : [0, 1]^k \to [0, 1]$ such that $H(y_1, \ldots, y_k) = C(F_1(y_1), \ldots, F_k(y_k))$. For statistical modelling, it is also useful to note that a k-dimensional copula C and marginal continuous CDFs $F_i(y_i), i = 1, \ldots, k$ are building blocks for a valid k-dimensional CDF, $C(F_1(y_1), \ldots, F_k(y_k))$

E. Levi · R. V. Craiu (✉)
Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada
e-mail: craiu@utstat.toronto.edu

E. Levi
e-mail: evgeny@utstat.utoronto.ca

M. La Rocca et al. (eds.), *Nonparametric Statistics*, Springer Proceedings in Mathematics & Statistics 339, https://doi.org/10.1007/978-3-030-57306-5_29

with ith marginal CDF equal to $F_i(y_i)$, thus providing much-needed flexibility in modelling multivariate distributions. The previous construction can be extended when conditioning on a covariate vector $X \in \mathbf{R}^q$ [14, 17] so that

$$H(y_1, \ldots, y_k|X) = C_X(F_1(y_1|X), \ldots, F_k(y_k|X)), \tag{1}$$

where all CDFs and the copula are conditional on X. For the rest of this paper, we follow [16] and assume: (1) that the copula in (1) belongs to a parametric family that remains the same across the whole range of X, and (2) its one-dimensional parameter depends on X through some unknown function $\theta(X) : \mathbf{R}^q \to \Theta \subset \mathbf{R}$. The range of $\theta(X)$ is usually restricted, so we introduce a known one-to-one link function $g : \Theta \to \mathbf{R}$ such that the *calibration function*, $\eta : \mathbf{R}^q \to \mathbf{R}$, defined as $\eta(X) = g(\theta(X))$ has unrestricted range. Sometimes, it is convenient to parametrize a copula family in terms of Kendall's tau, $\tau(X) : \mathbf{R}^q \to [-1, 1]$, which, for any given value of X, is in one-to-one correspondence with $\theta(X)$ when the copula parameter is one-dimensional. Thus, there is also a known one-to-one function $g'(\cdot)$ such that $\eta(X) = g'(\tau(X))$.

The simplifying assumption (SA) [5] states that copula C_X in (1) is independent of X, or that $\eta(X)$ is constant. Clearly, SA greatly simplifies the estimation in conditional copula models, including their use in vines (see, for instance, [1]). Acar et al. [2] showed that assuming SA when the data generative process has non-constant calibration may bias the inference, while Levi and Craiu [16] showed that SA is violated when important covariates are not included in the model (1). In light of these results, there is a genuine demand for strategies that effectively test whether the SA is appropriate or not. A number of research contributions address this issue for frequentist analyses, e.g. [3, 6, 9, 11].

This contribution belongs within the Bayesian paradigm, following the general philosophy expounded also in Klein and Kneib [12]. In this setting, it was observed in Craiu and Sabeti [4] that generic model selection criteria tend to choose a more complex model even when SA holds. In the next section, we present the problem in mathematical terms and review some of the Bayesian model selection procedures used for SA. A new approach, based on permutations, is described in Sect. 3. The Appendix contains a theoretical justification of the proposed algorithm and a discussion of extensions to other regression problems. A merit of the proposal is that it is quite general in its applicability, but this comes, unsurprisingly, at the expense of power. In order to investigate whether the trade-off is reasonable, we design a simulation study and present its conclusions in Sect. 4. The paper ends with a summary and discussion of future work.

2 Problem Setup

We consider observed data that consist of n independent triplets $\mathscr{D} = \{(x_i, y_{1i}, y_{2i}), i = 1, \ldots, n\}$ where $y_{ji} \in \mathbf{R}$, $j = 1, 2$, and $x_i \in \mathbf{R}^q$. Denote $\mathbf{y}_1 = (y_{11}, \ldots, y_{1n})$, $\mathbf{y}_2 = (y_{21}, \ldots, y_{2n})$ and $\mathbf{X} \in \mathbf{R}^{n \times q}$ the matrix with i^{th} row equal to x_i^T. We rely on (1) to express the *full conditional model* density for Y_1 and Y_2 given $\mathbf{X}$

$$p(\mathbf{y}_1, \mathbf{y}_2 | \mathbf{X}, \omega) = \prod_{i=1}^{n} f_1(y_{1i}|\omega, x_i) f_2(y_{2i}|\omega, x_i) c_{\theta(x_i)} \left(F_1(y_{1i}|\omega, x_i), F_2(y_{2i}|\omega, x_i)\right), \tag{2}$$

where f_j, F_j are the density and, respectively, the CDF for Y_j, and ω denotes all the parameters and latent variables in the joint and marginals models. The copula density function is denoted by c, and it depends on X through unknown function $\theta(X) = g^{-1}(\eta(X))$. The copula family can be selected using several model selection criteria (e.g. [16, 18]). Once the copula family is selected, the objective is to check whether the SA is valid, in other words whether (2) becomes the *reduced model*

$$P(\mathbf{y}_1, \mathbf{y}_2 | \mathbf{X}, \omega) = \prod_{i=1}^{n} f_1(y_{1i}|\omega, x_i) f_2(y_{2i}|\omega, x_i) c_{\theta} \left(F_1(y_{1i}|\omega, x_i), F_2(y_{2i}|\omega, x_i)\right), \tag{3}$$

in which the copula depends only on one parameter, θ. Flexible Bayesian models usually yield posteriors that are analytically intractable, so their characteristics will be estimated using draws $\{\omega^{(t)}\}_{t=1}^{M}$ obtained via a Markov chain Monte Carlo (MCMC) algorithm (e.g. [16, 18]). Data support for the full and reduced models, (2) and (3), may be established using predictive power as a criterion.

2.1 The Cross-Validated Pseudo Marginal Likelihood and Its Conditional Variant

The cross-validated pseudo marginal likelihood (CVML) [7] calculates the average (over parameter values) prediction power for model $\mathscr{M}$ via

$$\text{CVML}(\mathscr{M}) = \sum_{i=1}^{n} \log \left(P(y_{1i}, y_{2i} | \mathscr{D}_{-i}, \mathscr{M})\right), \tag{4}$$

where $\mathscr{D}_{-i}$ is the data set from which the ith observation has been removed. An estimate of (4) for a given model is estimated using posterior draws $\omega^{(t)}$ given the whole data set $\mathscr{D}$, (detailed derivations can be found in Levi and Craiu [16]) via

$$\text{CVML}_{est}(\mathscr{M}) = -\sum_{i=1}^{n} \log\left(\frac{1}{M}\sum_{t=1}^{M} P(y_{1i}, y_{2i}|\omega^{(t)}, \mathscr{M})^{-1}\right). \tag{5}$$

The model with the largest CVML is preferred.

The conditional CVML (CCVML), introduced by Levi and Craiu [16] specifically for copula models, exploits conditional rather than joint predictions

$$\text{CCVML}(\mathscr{M}) = \frac{1}{2}\left\{\sum_{i=1}^{n} \log\left[P(y_{1i}|y_{2i}, \mathscr{D}_{-i}, \mathscr{M})\right] + \sum_{i=1}^{n} \log\left[P(y_{2i}|y_{1i}, \mathscr{D}_{-i}, \mathscr{M})\right]\right\}.$$

Again this criterion can be estimated from posterior samples using

$$\begin{aligned}\text{CCVML}_{est}(\mathscr{M}) = -\tfrac{1}{2}\sum_{i=1}^{n}\Bigg\{&\log\left[\tfrac{1}{M}\sum_{t=1}^{M}\tfrac{P(y_{2i}|\omega^{(t)},\mathscr{M})}{P(y_{1i},y_{2i}|\omega^{(t)},\mathscr{M})}\right] \\ &+ \log\left[\tfrac{1}{M}\sum_{t=1}^{M}\tfrac{P(y_{1i}|\omega^{(t)},\mathscr{M})}{P(y_{1i},y_{2i}|\omega^{(t)},\mathscr{M})}\right]\Bigg\}.\end{aligned} \tag{6}$$

Similar to CVML, the model with the largest CCVML is selected.

2.2 *Watanabe–Akaike Information Criterion*

The Watanabe–Akaike Information Criterion [21] is an information-based criterion that is closely related to CVML, as discussed in [20]. The WAIC is defined as

$$\text{WAIC}(\mathscr{M}) = -2\text{fit}(\mathscr{M}) + 2\text{p}(\mathscr{M}), \tag{7}$$

where the model fitness is

$$\text{fit}(\mathscr{M}) = \sum_{i=1}^{n} \log E\left[P(y_{1i}, y_{2i}|\omega, \mathscr{M})\right], \tag{8}$$

and the penalty

$$\text{p}(\mathscr{M}) = \sum_{i=1}^{n} \text{Var}[\log P(y_{1i}, y_{2i}|\omega, \mathscr{M})]. \tag{9}$$

The expectation in (8) and the variance in (9) are with respect to the conditional distribution of ω given the data and can easily be estimated using the $\omega^{(t)}$ draws. The model with the smallest WAIC measure is preferred.

3 Detecting Data Support for SA

As will be shown in Sect. 4, the criteria described above exhibit unsatisfactory performances when the reduced model is the generative one. While it is expected that the flexibility of the full model will yield good predictions even when SA holds, it was surprising to see that the penalty term in (9) is not large enough to downgrade the full model under the SA null. Therefore, we base our diagnostics on some of the properties that are invariant to the group of permutations when SA holds.

In the first stage, we randomly divide the data $\mathscr{D}$ into training and test sets, $\mathscr{D}_1$ and $\mathscr{D}_2$, with n_1 and n_2 sample sizes, respectively. The full model defined by (2) is fitted on $\mathscr{D}_1$, and we denote $\omega^{(t)}$ the tth draw sampled from the posterior. For the ith item in $\mathscr{D}_2$, compute point estimates $\hat{\eta}_i$ and $\hat{U}_i = (\hat{U}_{1i}, \hat{U}_{2i})$, where $\hat{U}_{ji} = F_j(y_{ji}|\hat{\omega}_j, x_i)$, $j = 1, 2$, $i = 1, \ldots, n_2$, and ω_j denotes the vector of all the parameters and latent variables related to the jth marginal distribution. The marginal parameter estimates, $\hat{\omega}_j$, are obtained from the training data posterior draws. For instance, if the marginal models are $Y_{1i} \sim \mathscr{N}(f_1(x_i), \sigma_1^2)$ and $Y_{2i} \sim \mathscr{N}(f_2(x_i), \sigma_2^2)$, then each of the MCMC sample $\omega^{(t)}$ leads to an estimate $\hat{f}_1^t(x_i)$, $\hat{f}_2^t(x_i)$, $\hat{\sigma}_1^t$, $\hat{\sigma}_2^t$, $\hat{\eta}^t(x_i)$. Then $\hat{U}_i = (\hat{U}_{1i}, \hat{U}_{2i})$ are obtained using

$$(\hat{U}_{1i}, \hat{U}_{2i}) = (\Phi((y_{1i} - \overline{\hat{f}_1(x_i)})/\overline{\hat{\sigma}_1}), \Phi((y_{2i} - \overline{\hat{f}_2(x_i)})/\overline{\hat{\sigma}_2})),$$

where the overline $\overline{a}$ signifies the averages of Monte Carlo draws a^t.

Given the vector of calibration function evaluations at the test points, $\hat{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_{n_2})$, and a partition $\min(\hat{\eta}) = a_1 < \ldots < a_{K+1} = \max(\hat{\eta})$ of the range of η into K disjoint intervals, define the set of observations in $\mathscr{D}_2$ that yield calibration function values between a_k and a_{k+1}, $B_k = \{i : a_k \leq \hat{\eta}_i < a_{k+1}\}$ $k = 1, \ldots, K$. We choose the partition such that each "bin" B_k has approximately the same number of elements, n_2/K.

Under SA, the bin-specific estimates for various measures of dependence, e.g. Kendall's τ or Spearman's ρ, computed from the samples $\hat{U}_i$, are invariant to permutations, or swaps across bins. Based on this observation, we consider the procedure described in Table 1 for identifying data support for SA. The distribution of the resulting test statistics obtained in Method 1 is determined empirically, via permutations. Alternatively, one can rely on the asymptotic properties of the bin-specific dependence parameter estimates and construct a Chi-square test. Specifically, suppose the bin-specific Pearson correlations $\hat{\rho}_k$ are computed from samples $\{\hat{U}_i : i \in B_k\}$), for all $k = 1, \ldots, K$, and let $\hat{\rho} = (\hat{\rho}_1, \ldots, \hat{\rho}_K)^T$ and $\tilde{n} = n_2/K$ be the number of points in each bin. It is known that $\hat{\rho}_k$ is asymptotically normal distributed for each k so that

$$\sqrt{\tilde{n}}(\hat{\rho}_k - \rho_k) \xrightarrow{d} \mathscr{N}(0, (1 - \rho_k^2)^2),$$

where ρ_k is the true correlation in bin k. If we assume that $\{\hat{\rho}_k\colon\ k = 1, \ldots, K\}$ are independent, and set $\rho = (\rho_1, \ldots, \rho_K)^T$ and $\Sigma = diag((1 - \rho_1^2)^2, \ldots, (1 - \rho_K^2)^2)$, then we have

Table 1 Method 1: A permutation-based procedure for assessing data support in favour of SA

A1 Compute the kth bin-specific Kendall's tau $\hat{\tau}_k$ from $\{\hat{U}_i : i \in B_k\})$ $k = 1, \ldots, K$.

A2 Compute the observed statistic $T^{obs} = SD_k(\hat{\tau}_k)$ (where $SD_k(a_k)$ is a standard deviation of a_k over index k). Note that if SA holds, we expect the observed statistic to be close to zero.

A3 Consider J permutations $\lambda_j : \{1, \ldots, n_2\} \to \{1, \ldots, n_2\}$. For each permutation λ_j:

A3.1 Compute $\hat{\tau}_{jk} = \tau(\{\hat{U}_i : \lambda_j(i) \in B_k\})$ $k = 1, \ldots, K$.

A3.2 Compute test statistic $T_j = SD_k(\hat{\tau}_{jk})$. Note if SA holds, then we expect T_j to be close to T^{obs}.

A4 We consider that there is support in favour of SA at significance level α if T^{obs} is smaller than the $(1-\alpha)$-th empirical quantile calculated from the sample $\{T_j : 1 \le j \le J\}$.

$$\sqrt{\tilde{n}}(\hat{\rho} - \rho) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

In order to combine evidence across bins, we define the matrix $A \in \mathbf{R}^{(K-1)\times K}$ as

$$A = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{bmatrix}. \tag{10}$$

Since under the null hypothesis SA holds, one gets $\rho_1 = \ldots = \rho_K$, implying

$$\tilde{n}(A\hat{\rho})^T (A\Sigma A^t)^{-1} (A\hat{\rho}) \xrightarrow{d} \chi^2_{K-1}.$$

Method 2, with its steps detailed in Table 2, relies on the ideas above to test SA.

Method 1 evaluates the p-value using a randomization procedure [15], while the second is based on the asymptotic normal theory of Pearson correlations. To get reliable results, it is essential to assign test observations to "correct" bins which is true when calibration predictions are as close as possible to the true unknown values, i.e. $\hat{\eta}(x_i) \approx \eta(x_i)$. The latter heavily depends on the estimation procedure and sample size of the training set. Therefore, it is advisable to apply very flexible models for the calibration function estimation and have enough data points in the training set. The trade-off we notice is that as more observations are assigned to $\mathscr{D}$, the calibration test predictions improve, even as power decreases due to a smaller sample size in $\mathscr{D}_2$. For our simulations, we have used $n_1 \approx 0.5n$ and $n_2 \approx 0.5n$, and $K \in \{2, 3\}$.

Remarks: The equivalence between SA and equality of $\eta(x)$ across bins is central to both methods and requires some discussion. Below, we assume that only two bins are used and that the estimation is based on very large data so that finite-sample variability is ignored.

1. *Necessity* If SA is true then indeed η must be constant across bins *as long as the copula family is the same* across the whole range of X.

Table 2 Method 2: A Chi-square test for assessing data support in favour of SA

B1 Compute the bin-specific Pearson correlation $\hat{\rho}_k$ from samples $\{\hat{U}_i : i \in B_k\}$), for all $k = 1, \ldots, K$. Let $\hat{\rho} = (\hat{\rho}_1, \ldots, \hat{\rho}_K)^T$, and $\tilde{n} = n_2/K$, the number of points in each bin.

B2 Define $\rho = (\rho_1, \ldots, \rho_K)^T$, $\Sigma = diag((1-\rho_1^2)^2, \ldots, (1-\rho_K^2)^2)$ and $A \in \mathbf{R}^{(K-1)\times K}$ as in (10) then under SA we have that $\rho_1 = \ldots = \rho_K$ and

$$\tilde{n}(A\hat{\rho})^T(A\Sigma A^t)^{-1}(A\hat{\rho}) \xrightarrow{d} \chi^2_{K-1}.$$

Compute $T^{obs} = \tilde{n}(A\hat{\rho})^T(A\hat{\Sigma}A^t)^{-1}(A\hat{\rho})$.

B3 Compute p-value = $P(\chi^2_{K-1} > T^{obs})$ and reject SA if p-value$< \alpha$.

2. *Sufficiency* If SA does not hold, assume that the calibration function takes two values. Assuming consistency of the calibration's estimator, it is expected that bin 1 and bin 2 will contain pairs (u_1, u_2) following distributions $\pi_1(u_1, u_2)$ and $\pi_2(u_1, u_2)$ with corresponding correlations $\rho_1 < \rho_2$, respectively. After a random permutation, pairs in each bin will follow a mixture distribution $\lambda\pi_1(u_1, u_2) + (1-\lambda)\pi_2(u_1, u_2)$ and $(1-\lambda)\pi_1(u_1, u_2) + \lambda\pi_2(u_1, u_2)$ in bins 1 and 2, respectively, with $\lambda \in (0, 1)$. Thus, the post-permutation correlations in bins 1 and 2 are $\lambda\rho_1 + (1-\lambda)\rho_2$ and $(1-\lambda)\rho_1 + \lambda\rho_2$. Observe that each correlation is between ρ_1 and ρ_2 which implies that the absolute difference between them will be less than $\rho_2 - \rho_1$, so we expect to reject the null. This argument offers heuristic support for the method, but obviously cannot be extended to cases where η is non-constant in each bin and finite sample variability must be accounted for.

A theoretical justification for Method 2 and extensions of this idea to other models are available in Appendix.

4 Simulations

In this section, we present the performance of the proposed methods and comparisons with generic CVML and WAIC criteria on simulated data sets. Different functional forms of calibration function, sample sizes and magnitude of deviation from SA will be explored.

4.1 Simulation Details

We generate samples of sizes $n = 500$ and $n = 1000$ from 3 scenarios described below. For all scenarios, the Clayton copula will be used to model dependence between responses, while covariates are independently sampled from $\mathscr{U}[0, 1]$. For all scenarios, the covariate dimension $q = 2$. Marginal conditional distributions $Y_1|X$

and $Y_2|X$ are modelled as Gaussian with constant variances σ_1^2, σ_2^2 and conditional means $f_1(X)$, $f_2(X)$, respectively. The model parameters must be estimated jointly with the calibration function $\eta(X)$. For convenience, we parametrize calibration on Kendall's tau $\tau(X)$ scale.

$$
\begin{aligned}
\text{Sc1}\quad & f_1(X) = 0.6\sin(5x_1) - 0.9\sin(2x_2),\\
& f_2(X) = 0.6\sin(3x_1 + 5x_2),\\
& \tau(X) = 0.5, \sigma_1 = \sigma_2 = 0.2.\\
\text{Sc2}\quad & f_1(X) = 0.6\sin(5x_1) - 0.9\sin(2x_2),\\
& f_2(X) = 0.6\sin(3x_1 + 5x_2),\\
& \tau(X) = \delta + \gamma \times \sin(10X^T\beta)\\
& \beta = (1,3)^T/\sqrt{10}, \sigma_1 = \sigma_2 = 0.2.\\
\text{Sc3}\quad & f_1(X) = 0.6\sin(5x_1) - 0.9\sin(2x_2),\\
& f_2(X) = 0.6\sin(3x_1 + 5x_2),\\
& \tau(X) = \delta + \gamma \times 2(x_1 + \cos(6x_2) - 0.45)/3\\
& \sigma_1 = \sigma_2 = 0.2.
\end{aligned}
$$

Sc1 corresponds to SA since Kendall's τ is independent of covariate level. The calibration function in **Sc2** has single index form for the calibration function, while in **Sc3** it has an additive structure on τ scale (generally not additive on η scale); these simulations are useful to evaluate performance under model misspecification. We note that τ in **Sc2** and **Sc3** depends on parameters δ (average correlation strength) and γ (deviation from SA), which in this study take values $\delta \in \{0.25, 0.75\}$ and $\gamma \in \{0.1, 0.2\}$, respectively.

4.2 Simulation Results

For each sample size and scenario, we have repeated the analysis using 250 independently replicated data sets. For each data, the GP-SIM model suggested by Levi and Craiu [16] is fitted. This method implements sparse Gaussian Process (GP) priors for marginal conditional means and sparse GP-Single Index for calibration function. These non-parametric models are more flexible than parametric ones and can effectively capture various patterns. The inference is based on 5000 MCMC samples for all scenarios, as the chains were run for 10,000 iterations with 5000 samples discarded as burn-in. The number of inducing inputs was set to 30 for all GP. For generic SA testing, GP-SIM fitting is done for the whole data sets, and posterior draws are used to estimate CVML and WAIC. Since the proposed methods requires data splitting, we first randomly divide the data equally into training and testing sets. We fit GP-

Table 3 Simulation results: generic, proportion of rejection of SA for each scenario, sample size and model selection criterion

Scenario	$n = 500$			$n = 1000$		
	CVML (%)	CCVML (%)	WAIC (%)	CVML (%)	CCVML (%)	WAIC (%)
Sc1	33.3	31.1	34.7	38.2	37.3	37.8
Sc2 ($\delta = 0.75, \gamma = 0.1$)	99.1	98.7	99.1	100	100	100
Sc2 ($\delta = 0.75, \gamma = 0.2$)	100	100	100	100	100	100
Sc2 ($\delta = 0.25, \gamma = 0.1$)	80.1	84.4	80.1	99.1	100	99.1
Sc2 ($\delta = 0.25, \gamma = 0.2$)	100	100	100	100	100	100
Sc3 ($\delta = 0.75, \gamma = 0.1$)	76.9	73.3	77.8	85.7	82.2	85.8
Sc3 ($\delta = 0.75, \gamma = 0.2$)	99.1	97.3	99.1	99.1	97.8	99.1
Sc3 ($\delta = 0.25, \gamma = 0.1$)	54.7	56.4	55.6	65.3	68.4	64.9
Sc3 ($\delta = 0.25, \gamma = 0.2$)	89.8	92.0	91.1	99.6	100	99.6

SIM on the training set and then use the obtained posterior draws to construct point estimates of $F_1(y_{1i}|x_i)$, $F_2(y_{2i}|x_i)$ and $\eta(x_i)$ for every observation in the test set. In Method 1, we used 500 permutations. Table 3 shows the percentage of SA rejections for generic Bayesian selection criteria. The presented results clearly illustrate that generic methods have difficulties identifying SA. This leads to a loss of statistical efficiency since a complex model is selected over a much simpler one. Moreover, CVML or CCVML fails to identify SA as both measures do not penalize directly for the complexity of the model. The simulations summarized in Table 4 show that the proposed methods (setting $\alpha = 0.05$) have much smaller probability of Type I error which vary around the threshold of 0.05. It must be pointed, however, that under SA the performance of χ^2 test worsens with the number of bins K, which is not surprising since as K increases, the number of observations in each bin goes down, and normal approximation for the distribution of Pearson correlation becomes tenuous, while the permutation-based test is more robust to small samples. The performance of both methods improves with sample size. We also notice a loss of power between Scenarios 2 and 3, which is due to model misspecification, since in the latter case the generative model is different from the postulated one. All methods break down when the departure from SA is not large, e.g. $\gamma = 0.1$. Although not desirable, this has limited impact in practice since, in our experience, in this case the predictions produced by either model are very similar.

Table 4 Simulation results: proposed method, proportion of rejection of SA for each scenario, sample size, number of bins (K) and method

Scenario	Permutation test				χ^2 test			
	$n = 500$		$n = 1000$		$n = 500$		$n = 1000$	
	$K = 2$ (%)	$K = 3$ (%)	$K = 2$ (%)	$K = 3$ (%)	$K = 2$ (%)	$K = 3$ (%)	$K = 2$ (%)	$K = 3$ (%)
Sc1	4.9	6.2	3.5	5.3	9.7	11.1	10.7	13.7
Sc2($\delta = 0.75, \gamma = 0.1$)	90.2	80.4	99.6	99.1	94.7	94.2	99.6	99.1
Sc2($\delta = 0.75, \gamma = 0.2$)	100	100	100	100	100	100	100	100
Sc2($\delta = 0.25, \gamma = 0.1$)	25.8	18.7	55.1	47.1	30.2	21.8	58.7	53.8
Sc2($\delta = 0.25, \gamma = 0.2$)	91.6	84.9	99.6	99.6	92.4	91.1	99.6	99.6
Sc3($\delta = 0.75, \gamma = 0.1$)	28.0	24.0	57.3	52.9	41.3	45.8	72.4	72.9
Sc3($\delta = 0.75, \gamma = 0.2$)	88.4	85.8	98.7	98.7	94.2	92.0	100	99.1
Sc3($\delta = 0.25, \gamma = 0.1$)	8.0	7.5	11.1	10.7	9.8	10.7	15.1	12.9
Sc3($\delta = 0.25, \gamma = 0.2$)	19.6	18.2	63.6	60.9	24.9	23.6	70.2	69.3

5 Conclusion

We propose two methods to check data support for the simplifying assumption in conditional bivariate copula problems. Both are based on data splitting into training and test sets, partitioning the test set into bins using calibration values obtained in training and using randomization or χ^2 tests to determine if the dependence is constant across bins. Empirically, it was shown that the probability of Type I error is controlled when SA holds. When the generative process does not satisfy SA, these two methods also perform well, showing larger power than generic model selection criteria. Future work will address questions related to the proportion of data that should be assigned to training and test sets as well as bin sizes.

Acknowledgments We thank Stanislav Volgushev and an anonymous referee for suggestions that have greatly improved the paper. Financial support for this work was provided by the Canadian Statistical Sciences Institute (CANSSI) and by NSERC of Canada.

Appendix

Theoretical Discussion

In this section, we prove that under canonical assumptions, the probability of Type I error for Method 2 in Sect. 3 converges to α when SA is true.

Suppose we have independent samples from K populations (groups), $(u_{1i}^1, u_{2i}^1)_{i=1}^{n_1} \sim (U_1^1, U_2^1)$,…,$(u_{1i}^K, u_{2i}^K)_{i=1}^{n_K} \sim (U_1^K, U_2^K)$, the goal is to test $\rho_1 = \cdots = \rho_K$ (here ρ is Pearson correlation).

To simplify notation, we assume $n_1 = \ldots, n_K = n$. Let $\hat{\rho} = (\hat{\rho}_1, \ldots, \hat{\rho}_K)$ be the vector of sample correlations , $\Sigma = diag((1-\rho_1^2)^2, \ldots, (1-\rho_K^2)^2)$ and $(K-1) \times K$ matrix A as defined in Sect. 3, then canonical asymptotic results imply that if $\rho_1 = \cdots = \rho_K$ and as $n \to \infty$,

$$T = n(A\hat{\rho})^T (A\Sigma A^T)^{-1} (A\hat{\rho}) \xrightarrow{d} \chi^2_{K-1}. \tag{11}$$

Based on the model fitted on $\mathscr{D}_1$, we define estimates of $F_1(y_{1i}|x_i)$ and $F_2(y_{2i}|x_i)$ by $\hat{U} = \{\hat{U}_i = (\hat{F}_1(y_{1i}|x_i), \hat{F}_2(y_{2i}|x_i))\}_{i=1}^{n_2}$. Note that $\hat{U}$ depends on $\mathscr{D}_1$ and X. Given a fixed number of bins K and assuming, without loss of generality, equal sample sizes in each bin $\tilde{n} = n_2/K$, we define a test statistic $T(\hat{U})$ as in (11) with $\hat{\rho}_j$ estimated from $\{\hat{U}_{(j-1)\tilde{n}+1}, \ldots, \hat{U}_{j\tilde{n}}\}$, for $1 \le j \le K$.

Note that in Method 2, test cases are assigned to "bins" based on the value of predicted calibration function $\hat{\eta}(x_i)$ which is not taken into account in the generic definition of test statistic $T(\hat{U})$ above. To close this gap, we introduce a permutation $\lambda^* : \{1, \ldots, n_2\} \to \{1, \ldots, n_2\}$ that "sorts" $\hat{U}$ from smallest $\hat{\eta}(x)$ value to largest, i.e. $\hat{U}_{\lambda^*} = \{\hat{U}_{\lambda^*(i)}\}_{i=1}^{n_2}$ with $\hat{\eta}(x_{\lambda^*(1)}) < \hat{\eta}(x_{\lambda^*(2)}) < \cdots < \hat{\eta}(x_{\lambda^*(n_2)})$. Hence, the test statistic in Method 2 has the form $T(\hat{U}_{\lambda^*})$ as in (11) but in this case test cases with smallest predicted calibrations are assigned to the first group, or bin, and with largest calibrations to the Kth group/bin. Finally, define a test function ϕ with specified significance level α to test SA:

$$\phi(\hat{U}|\lambda^*) = \begin{cases} 1 & \text{if} \quad T(\hat{U}_{\lambda^*}) > \chi^2_{K-1}(1-\alpha), \\ 0 & \text{if} \quad T(\hat{U}_{\lambda^*}) \le \chi^2_{K-1}(1-\alpha). \end{cases} \tag{12}$$

Intuitively, if SA is false then we would expect $T(\hat{U}_{\lambda^*})$ to be larger then the critical value $\chi^2_{K-1}(1-\alpha)$.

The goal is to show that this procedure have probability of type I error equal to α, which is equivalent to the expectation of the test function:

$$\text{P(Type I error)} = \int \phi(\hat{U}|\lambda^*) P(\lambda^*|\mathscr{D}_1, X) P(\hat{U}|\mathscr{D}_1, X) P(\mathscr{D}_1) P(X) d\hat{U} d\mathscr{D}_1 dX d\lambda^*. \tag{13}$$

Note that λ^* does not depend on $\hat{U}$ because of the data splitting to training and test sets. Also usually $P(\lambda^*|\mathscr{D}_1, X)$ is just a point mass at some particular permutation. In general the above integral cannot be evaluated, however if we assume that for all test cases:

$$\begin{aligned} \hat{F}_1(y_{1i}|x_i) &\xrightarrow{p} F_1(y_{1i}|x_i) \quad \text{as} \quad n \to \infty \quad \forall i, \\ \hat{F}_2(y_{2i}|x_i) &\xrightarrow{p} F_2(y_{2i}|x_i) \quad \text{as} \quad n \to \infty \quad \forall i, \end{aligned} \tag{14}$$

then under SA and as $n \to \infty$, $P(\hat{U}|\mathscr{D}_1, X) = P(\hat{U}) \approx \prod_{i=1}^{n_2} c(\hat{u}_{1i}, \hat{u}_{2i})$ where c is a copula density and the expectation becomes

$$\begin{aligned} \text{P(Type I error)} &= \int \phi(\hat{U}|\lambda^*) P(\lambda^*|\mathscr{D}_1, X) P(\hat{U}) P(\mathscr{D}_1) P(X) d\hat{U} d\mathscr{D}_1 dX d\lambda^* = \\ &= \int \left(\int \phi(\hat{U}|\lambda^*) P(\hat{U}) d\hat{U} \right) P(\lambda^*|\mathscr{D}_1, X) P(\mathscr{D}_1) P(X) d\mathscr{D}_1 dX d\lambda^* = \alpha. \end{aligned} \tag{15}$$

Since if SA is true, $\int \phi(\hat{U}|\lambda^*) P(\hat{U}) d\hat{U} = \alpha$ for any λ^*. Therefore, if marginal CDF predictions for test cases are consistent then this procedure has the required probability of Type I error for sufficiently large sample size.

Extensions to Other Models

The proposed idea of dividing the data into training and test subsets, splitting the observations on the test set to bins defined in first stage and then using a test to check distributional difference between bins can be extended to other models. For example, one can use a similar construction in a regression problem with conditional mean $f(X)$ and constant variance. First assign test cases to bins by the values of $\hat{f}(X)$, and then compare means in each bin either by randomization procedures or using χ^2 test. This approach can be especially useful when $f(X)$ is assumed to have a complex form, such as generalized additive models, including additive tree structures, splines or based on Bayesian non-parametric methods. Simulations we conducted (not reported here) suggest that for large covariate dimensions, standard F-tests for Gaussian error regression yield large Type I error probabilities, a problem that is attenuated using the permutation-based ideas described in the paper.

References

1. Aas, K., Czado, C., Frigessi, A., Bakken, H.: Pair-copula constructions of multiple dependence. Insur. Math. Econ. **44**, 182–198 (2009)
2. Acar, E., Genest, C., Nešlehová, J.: Beyond simplified pair-copula constructions. J. Multivar. Anal. **110**, 74–90 (2012)
3. Acar, E.F., Craiu, R.V., Yao, F., et al.: Statistical testing of covariate effects in conditional copula models. Electron. J. Stat. **7**, 2822–2850 (2013)
4. Craiu, R.V., Sabeti, A.: In mixed company: Bayesian inference for bivariate conditional copula models with discrete and continuous outcomes. J. Multivar. Anal. **110**, 106–120 (2012)
5. Czado, C.: Pair-copula constructions of multivariate copulas. Copula Theory and its Applications, pp. 93–109. Springer, Berlin (2010)
6. Derumigny, A., Fermanian, J.-D.: About tests of the "simplifying" assumption for conditional copulas (2016). arXiv:1612.07349
7. Geisser, S., Eddy, W.F.: A predictive approach to model selection. J. Am. Stat. Assoc. **74**, 153–160 (1979)

8. Genest, C., Favre, A.-C.: Everything you always wanted to know about copula modeling but were afraid to ask. J. Hydrol. Eng. **12**, 347–368 (2007)
9. Gijbels, I., Omelka, M., Veraverbeke, N.: Estimation of a copula when a covariate affects only marginal distributions. Scand. J. Stat. **42**, 1109–1126 (2015)
10. Hougaard, P.: Analysis of Multivariate Survival Data. Statistics for Biology and Health. Springer, New York (2000)
11. Killiches, M., Kraus, D., Czado, C.: Examination and visualisation of the simplifying assumption for vine copulas in three dimensions. Aust. N. Z. J. Stat. **59**, 95–117 (2017)
12. Klein, N., Kneib, T.: Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. Stat. Comput. 1–20 (2015)
13. Lakhal, L., Rivest, L.-P., Abdous, B.: Estimating survival and association in a semicompeting risks model. Biometrics **64**, 180–188 (2008)
14. Lambert, P., Vandenhende, F.: A copula-based model for multivariate non-normal longitudinal data: analysis of a dose titration safety study on a new antidepressant. Stat. Med. **21**, 3197–3217 (2002)
15. Lehmann, E.L., Romano., J.P.: Testing Statistical Hypotheses. Springer Science & Business Media (2006)
16. Levi, E., Craiu, R.V.: Bayesian inference for conditional copulas using gaussian process single index models. Comput. Stat. Data Anal. **122**, 115–134 (2018)
17. Patton, A.J.: Modelling asymmetric exchange rate dependence*. Int. Econ. Rev. **47**, 527–556 (2006)
18. Sabeti, A., Wei, M., Craiu, R.V.: Additive models for conditional copulas. Statistics **3**, 300–312 (2014)
19. Sklar, A.: Fonctions de répartition à n dimensions et leurs marges. Publications de l'Institut de Statistique de l'Université de Paris **8**, 229–231 (1959)
20. Vehtari, A., Gelman, A., Gabry, J.: Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. **27**, 1413–1432 (2017)
21. Watanabe, S.: Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. **11**, 3571–3594 (2010)