Computational Statistics and Data Analysis 122 (2018) 115–134

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Bayesian inference for conditional copulas using Gaussian Process single index models

Evgeny Levi, Radu V. Craiu *

Department of Statistical Sciences, University of Toronto, Toronto, Ontario M5S 3G3, Canada

ARTICLE INFO

Article history:
Received 13 May 2017
Received in revised form 14 January 2018
Accepted 21 January 2018
Available online 2 February 2018

Keywords:
Conditional copula
Cross validated marginal likelihood
Gaussian Process
Simplifying assumption
Single index model

ABSTRACT

Parametric conditional copula models allow the copula parameters to vary with a set of covariates according to an unknown calibration function. Flexible Bayesian inference for the calibration function of a bivariate conditional copula is introduced. The prior distribution over the set of smooth calibration functions is built using a sparse Gaussian process (GP) prior for the single index model (SIM). The estimation of parameters from the marginal distributions and the calibration function is done jointly via Markov Chain Monte Carlo sampling from the full posterior distribution. A new Conditional Cross Validated Pseudo-Marginal (CCVML) criterion is used to perform copula selection and is modified using a permutation-based procedure to assess data support for the simplifying assumption. The performance of the estimation method and model selection criteria is studied via a series of simulations using correct and misspecified models with Clayton, Frank and Gaussian copulas and a numerical application involving red wine features.

1. Introduction and motivation

Copulas are useful in modeling the dependent structure in the data when there is interest in separating it from the marginal models or when none of the existent multivariate distributions are suitable. For continuous multivariate distributions, the elegant result of Sklar (1959) guarantees the existence and uniqueness of the copula $C : [0,1]^p \to [0,1]$ that links the marginal cumulative distribution functions (cdf) and the joint cdf. Specifically,

$$H(Y_1, \ldots, Y_p) = C(F_1(Y_1), \ldots, F_p(Y_p)),$$

where H is the joint cdf, and F_i is the marginal cdf for variable Y_i, for $1 \le i \le p$, respectively. This paper's focus is on copula models used in a regression setting in which covariate values are expected to influence the responses $Y_1, \ldots, Y_p$ through the marginal models and the interdependence between them through the copula. The extension to conditional distributions via the *conditional copula* was used by Lambert and Vandenhende (2002) and subsequently formalized by Patton (2006) so that

$$H(Y_1, \ldots, Y_p|\mathbf{X}) = C_{\mathbf{X}}(F_{1|\mathbf{X}}(Y_1|\mathbf{X}), \ldots, F_{p|\mathbf{X}}(Y_p|\mathbf{X})), \tag{1}$$

where $\mathbf{X} \in \mathbf{R}^q$ is a vector of conditioning variables, $C_{\mathbf{X}}$ is the conditional copula that may change with $\mathbf{X}$ and $F_{i|\mathbf{X}}$ is the conditional cdf of Y_i given $\mathbf{X}$ for $1 \le i \le p$. A parametric model for the conditional copula assumes $C_{\mathbf{X}} = C_{\theta(\mathbf{X})}$ belongs to a parametric family of copulas and only the parameter $\theta \in \Theta$ varies as a function of $\mathbf{X}$. Throughout the paper uppercase letters identify random variables, while their realizations are denoted using lowercase. In the remaining of this paper we

* Correspondence to: University of Toronto, 100 St. George Street, Toronto, ON M5S 3G3, Canada.
E-mail address: craiu@utstat.toronto.edu (R.V. Craiu).

https://doi.org/10.1016/j.csda.2018.01.013
0167-9473/© 2018 Elsevier B.V. All rights reserved.

assume that there exists a known one-to-one function $g : \Theta \to \mathbf{R}$ such that $\theta(\mathbf{X}) = g^{-1}(\eta(\mathbf{X}))$ with the *calibration function* $\eta : \mathbf{R} \to \mathbf{R}$ in the inferential focus.

There are a number of reasons one is interested in estimating the conditional copula. First, in regression models with multivariate responses, which is the main focus of this paper, one may want to determine how the dependence structure among the components of the response varies with the covariates. Second, the copula model will ultimately impact the performance of model-based prediction. For instance, for a bivariate response, (Y_1, Y_2), in which one component is predicted given the other, the conditional density of Y_1, given $\mathbf{X} = \mathbf{x}$ and $Y_2 = y_2$, takes the form

$$h(y_1|y_2, \mathbf{x}) = f(y_1|\mathbf{x})c_{\theta(\mathbf{x})}(F_{1|\mathbf{x}}(y_1|\mathbf{x}), F_{2|\mathbf{x}}(y_2|\mathbf{x})), \tag{2}$$

where $c_{\theta(\mathbf{x})}$ is the density of the conditional copula $C_{\theta(\mathbf{x})}$ and $f(y_1|\mathbf{x})$ is the marginal conditional density of y_1 given $\mathbf{X} = \mathbf{x}$. Hence, in addition to the information contained in the marginal model, in Eq. (2) we use for prediction also the information in the other responses.

Third, when specifying a general multivariate distribution, the conditional copula is an essential ingredient. For instance, if U_1, U_2, U_3 are three Uniform(0, 1) variables, when applying a vine decomposition using bivariate copulas (e.g., Czado, 2010) their joint density is

$$c(u_1, u_2, u_3) = c_{12}(u_1, u_2)c_{23}(u_2, u_3)c_{\theta(u_2)}\left(P(U_1 \le u_1|u_2), P(U_3 \le u_3|u_2)\right),$$

where c_{ij} is the density of the copula between variables U_i and U_j and $c_{\theta(u_2)}$ is the density of the conditional copula of $U_1, U_3|U_2 = u_2$. Finally, a conditional copula with predictor values $\mathbf{X} \in \mathbf{R}^q$ in which $\eta(\mathbf{X})$ is constant, may exhibit non-constant patterns when some of the components of $\mathbf{X}$ are not included in the model. This point will be revisited in Section 6.1.

When estimation for the conditional copula model is contemplated, one must consider that there are multiple sources of error and each will have an impact on the model. Even in the simple case in which the estimation of the marginals and copula suffers from errors that depend only on $\mathbf{x}$ one obtains via Taylor expansion:

$$c_{\theta(\mathbf{x})+\delta_3(\mathbf{x})}(F_{1|\mathbf{x}}(y_1|\mathbf{x}) + \delta_1(\mathbf{x}), F_{2|\mathbf{x}}(y_2|\mathbf{x}) + \delta_2(\mathbf{x})) = c_{\theta(\mathbf{x})}(F_{1|\mathbf{x}}(y_1|\mathbf{x}), F_{2|\mathbf{x}}(y_2|\mathbf{x})) \tag{3}$$
$$+ c^{(1,0,0)}_{\theta(\mathbf{x})}(F_{1|\mathbf{x}}(y_1|\mathbf{x}), F_{2|\mathbf{x}}(y_2|\mathbf{x}))\delta_1(\mathbf{x}) \tag{4}$$
$$+ c^{(0,1,0)}_{\theta(\mathbf{x})}(F_{1|\mathbf{x}}(y_1|\mathbf{x}), F_{2|\mathbf{x}}(y_2|\mathbf{x}))\delta_2(\mathbf{x}) \tag{5}$$
$$+ c^{(0,0,1)}_{\theta(\mathbf{x})}(F_{\mathbf{x}}(y_1), F_{\mathbf{x}}(y_2))\delta_3(\mathbf{x}) + \mathcal{O}(\|\delta(\mathbf{x})\|^2), \tag{6}$$

where $c^{(1,0,0)}$, $c^{(0,1,0)}$ and $c^{(0,0,1)}$ are the partial derivatives of $c_z(x, y)$ w.r.t. x, y and z, respectively, and $\delta_i(\mathbf{x})$, $1 \le i \le 3$, denote various estimation error terms due to model misspecification, e.g. $\delta_3(\mathbf{x})$ is the error in estimation of the copula parameter at a given covariate value $\mathbf{x}$. The right hand term in Eq. (3) marks the correct joint likelihood while (4)–(6) show the biases incurred due to errors in estimating the first and second marginal conditional cdfs and the copula calibration function, respectively. It becomes apparent that in order to keep the estimation error low, one must consider flexible models for the marginals and the copula.

Depending on the strength of assumptions we are willing to make about $\eta(\mathbf{x})$, a number of possible approaches are available. The most direct is to assume a known parametric form for the calibration function, e.g. constant or linear, and estimate the corresponding parameters by maximum likelihood estimation (Genest et al., 1995). This approach relies on knowledge about the shape of the calibration function which, in practice, can be unrealistic. A more flexible approach uses non-parametric methods (Acar et al., 2011; Veraverbeke et al., 2011) and estimates the calibration function using smoothing methods. Recently, we have seen a number of developments using nonparametric Bayesian techniques for estimating a multivariate copula using an infinite mixture of Gaussian copulas (Wu et al., 2014), or via flexible Dirichlet process priors (Wu et al., 2015; Ning and Shephard, 2017). The infinite mixture approach in Wu et al. (2014) was extended to estimate any conditional copula with a univariate covariate by Dalla Valle et al. (2017), while an alternative Bayesian approach based on a flexible cubic spline model for the calibration functions was built by Craiu and Sabeti (2012). For multivariate covariates, Sabeti et al. (2014), Chavez-Demoulin and Vatter (2015) and Klein and Kneiß (2015) avoid the curse of dimensionality that appears even for moderate values of q, say $q \ge 5$, by specifying an additive model structure for the calibration function. Few alternatives to the additive structure exist. One exception is Hernández-Lobato et al. (2013) who used a sparse Gaussian Process (GP) prior for estimating the calibration function and subsequently used the same construction for vine copulas estimation in Lopez-Paz et al. (2013). However, when the dimension of the predictor space is even moderately large the curse of dimensionality prevails and it is expected that the q-dimensional GP used for calibration estimation will not capture important patterns for sample sizes that are not very large. Moreover, the full efficiency of the method proposed in Hernández-Lobato et al. (2013) is difficult to assess since their model is build with uniform marginals, which in a general setup is equivalent to assuming exact knowledge about the marginal distributions. In fact, when the marginal distributions are estimated it is of paramount importance to account for the resulting variance inflation due to error propagation in the copula estimation as reflected by Eqs. (3)–(6). The Bayesian model in which joint and marginal components are simultaneously considered will appropriately handle error propagation as long as it is possible to study the full posterior distribution of all the parameters in the model, be they involved in the marginals or copula specification.

Great dimension reduction of the parameter space is achieved under the so-called *simplifying assumption (SA)* that assumes $C_{\theta(\mathbf{X})} = C$, i.e. the conditional copula is constant (Gijbels et al., 2015). The SA condition can significantly simplify the vine copula estimation (for example, see Aas et al., 2009), but it is known to lead to bias when it is wrongly assumed (Acar et al., 2012). Therefore, for conditional copula models it is of practical interest to assess whether the data supports SA or not. A first step towards a formal test for SA can be found in Acar et al. (2013). The reader is referred to Derumigny and Fermanian (2017) for an excellent review of work on SA, and ideas for future developments.

This paper's contribution is two-fold: on one hand we consider Bayesian joint analysis of the marginal and copula models using flexible GP models. Our emphasis is placed on the estimation of the calibration function $\eta(\mathbf{X})$ which is assumed to have a GP prior that is evaluated at $\beta^T\mathbf{X}$ for some normalized β, thus coupling the GP-prior construct with the *single index model (SIM)* of Choi et al. (2011) and Gramacy and Lian (2012). The GP-SIM is more flexible than a canonical linear model and computationally more manageable than a full GP with q variables. The proposed model can be used for large covariate dimension q and for large samples. Both marginal means will be fitted using sparse GP approaches so that large data sets can be computationally manageable. The dimension reduction of the SIM approach has been noted also by Fermanian and Lopez (2015) who used two-stage semiparametric methods to estimate the calibration function. In contrast to Fermanian and Lopez (2015), we use a Bayesian approach and estimate marginals and copula parameters jointly. So far, GP-SIMs have been used mostly in regression settings where the algorithm of Gramacy and Lian (2012) can be used to efficiently sample the posterior distribution. However, the GP-SIM model for conditional copulas involves a non-Gaussian likelihood which requires important modifications of their algorithm.

A second contribution of the paper deals with model selection issues that are particularly relevant for the conditional copula construction. We consider of importance the choice of copula family and identifying whether the simplifying assumption (SA) is supported by the data. For the former task we develop a conditional cross-validated marginal likelihood (CCVML) criterion and also examine its relation with the Watanabe Information Criterion (Watanabe, 2010), while for determining the data support for SA we construct a permutation-based variant of the CVML that shows good performance in our numerical experiments. Finally, we identify an important link between SA and missing covariates in the conditional copula model. To our knowledge, this connection has not been reported elsewhere.

In the next section we review the GP-SIM formulation and introduce the notation. The construction of the conditional copula model, the computational algorithm and the model selection procedures are covered in Sections 3 and 4, respectively. In Section 5 we illustrate the efficiency of the method via simulation and a numerical analysis of wine data. All the contributions relevant to the important issue of SA are included in Section 6. The paper ends with conclusions and directions for future work.

2. Brief review of Bayesian inference for sparse GP

Assume we observe n independent realizations, $y_1, \dots, y_n$, of a random variable $Y \in \mathbf{R}$ and that each observation y_i corresponds to a covariate measurement $x_i \in \mathbf{R}^q$. Henceforth, we assume that $x_1, \dots, x_n$ are fixed by design. The distribution of Y_i has a known form and depends on x_i through some unknown function f and parameter σ so that the joint distribution of the data is

$$P(\mathbf{Y}|x_1, \dots, x_n, \sigma) = P(\mathbf{Y}|f(x_1), \dots, f(x_n), \sigma) = \prod_{i=1}^{n} P(Y_i|f(x_i), \sigma). \tag{7}$$

Usually, the main inferential goal is to estimate the unknown smooth function $f : \mathbf{R}^q \to \mathbf{R}$, while σ is a nuisance parameter. If we let $\mathbf{x} = (x_1, \dots, x_n)^T$ denote the n covariate values, then a Gaussian Process (GP) prior on the function f implies

$$\mathbf{f} = (f(x_1), f(x_2), \dots, f(x_n))^T \sim \mathcal{N}(0, K(\mathbf{x}, \mathbf{x}; \mathbf{w})), \tag{8}$$

where $\mathcal{N}(\mu, \Sigma)$ denotes a normal distribution with mean μ and variance matrix Σ and K is a variance matrix which depends on $\mathbf{x}$ and additional parameters $\mathbf{w}$. In this paper we use the squared exponential kernel to model the matrix $K(\mathbf{x}, \mathbf{x}; \mathbf{w})$, i.e. its (i, j) element is

$$k(x_i, x_j; \mathbf{w}) = e^{w_0} \exp\left[-\sum_{s=1}^{q} \frac{(x_{is} - x_{js})^2}{e^{w_s}}\right], \tag{9}$$

where x_{is} is the sth coordinate value for ith covariate measurement x_i. The unknown parameters $\mathbf{w} = (w_0, \dots, w_q)$ that determine the strength of dependence in (9) are inferred from the data. Of interest is predicting the values of the nonlinear predictor at new observations $\mathbf{x}^* = (x_1^*, \dots, x_m^*)^T$, which we denote as $\mathbf{f}^* = (f(x_1^*), \dots, f(x_m^*))^T$. In the case in which the covariate dimension, q, is moderately large, an accurate estimation of $\mathbf{f}^*$ will require a large sample size, n. Unfortunately, this desideratum is hindered by the computational cost of fitting a GP model when n is large. For example, if $Y_i \sim \mathcal{N}(f(x_i), \sigma^2)$ then Eqs. (8) and (7) yield a joint Gaussian distribution of $\mathbf{Y} = (Y_1, \dots, Y_n)$ and $\mathbf{f}^*$. If $\mathbf{y} = (y_1, \dots, y_n)$ denotes the observed response, then the conditional distribution of $\mathbf{f}^*|\mathbf{Y}$ is $N(\mu^*, \Sigma^*)$ where

$$\mu^* = K(\mathbf{x}^*, \mathbf{x}; \mathbf{w})[K(\mathbf{x}^*, \mathbf{x}; \mathbf{w}) + \sigma^2\mathbf{I}_n]^{-1}\mathbf{y}, \tag{10}$$

$$\Sigma^* = K(\mathbf{x}^*, \mathbf{x}^*; \mathbf{w}) - K(\mathbf{x}^*, \mathbf{x}; \mathbf{w})[K(\mathbf{x}, \mathbf{x}; \mathbf{w}) + \sigma^2\mathbf{I}_n]^{-1}K(\mathbf{x}, \mathbf{x}^*; \mathbf{w}), \tag{11}$$

and $K(\mathbf{x}^*, \mathbf{x}^*; \mathbf{w})$, $K(\mathbf{x}^*, \mathbf{x}; \mathbf{w})$ and $K(\mathbf{x}, \mathbf{x}^*; \mathbf{w})$ have their elements defined using (9).

With the Gaussian sampling model it is clear from (10) and (11) that the MCMC sampling of the posterior requires at each iteration the calculation and inversion of the matrix $K(\mathbf{x}, \mathbf{x}; \mathbf{w}) + \sigma^2 \mathbf{I}_n \in \mathbf{R}^{n\times n}$ which becomes prohibitive when n is large. To make GP models applicable for larger data we refer to the literature on *sparse GP* (Quiñonero Candela and Rasmussen, 2005; Snelson and Ghahramani, 2005; Naish-Guzman and Holden, 2007) in which it is assumed that learning about f can be achieved using a smaller sample of m latent variables, called *inducing variables*, which may be a subsample of the original data or can be built using other considerations as further discussed. The intuitive idea is to use the inducing variables to channel the information contained in the covariate values $\mathbf{x} = \{x_1, \ldots, x_n\}$. We denote the inducing values as $\tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_m)^T \in \mathbf{R}^{m\times q}$ and $K(\mathbf{x}, \tilde{\mathbf{x}}; \mathbf{w}) \in \mathbf{R}^{n\times m}$ the matrix

$$K(\mathbf{x}, \tilde{\mathbf{x}}; \mathbf{w}) = \begin{bmatrix} k(x_1, \tilde{x}_1; \mathbf{w}) & \cdots & k(x_1, \tilde{x}_m; \mathbf{w}) \\ \vdots & \ddots & \vdots \\ k(x_n, \tilde{x}_1; \mathbf{w}) & \cdots & k(x_n, \tilde{x}_m; \mathbf{w}) \end{bmatrix}, \tag{12}$$

where $k(x_i, \tilde{x}_j; \mathbf{w})$ is defined as in (9). The ratio m/n influences the trade-off between computational efficiency and statistical efficiency, as a smaller m will favor the former and a larger m will ensure no significant loss of the latter. If the function values for the inducing points are defined as $\tilde{\mathbf{f}} = (f(\tilde{x}_1), \ldots, f(\tilde{x}_m))^T$ then the joint density of the response vector $\mathbf{Y}$, the latent variable $\tilde{\mathbf{f}}$ and the parameter $\mathbf{w}$ can be expressed only in terms of the m-dimensional vector $\tilde{\mathbf{f}}$ since

$$P(\mathbf{y}, \tilde{\mathbf{f}}, \mathbf{w}|\mathbf{x}, \tilde{\mathbf{x}}) = P(\mathbf{y}|A(\mathbf{x}, \tilde{\mathbf{x}}; \mathbf{w})\tilde{\mathbf{f}}) p_N(\tilde{\mathbf{f}}; 0, K(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}; \mathbf{w})) p(\mathbf{w}), \tag{13}$$

where $p_N(x; \mu, \Sigma)$ is the normal density with mean μ and covariance Σ, $p(\mathbf{w})$ is the prior probability for the parameters $\mathbf{w}$ and

$$A(\mathbf{x}, \tilde{\mathbf{x}}; \mathbf{w}) = K(\mathbf{x}, \tilde{\mathbf{x}}; \mathbf{w}) K(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}; \mathbf{w})^{-1}. \tag{14}$$

The form of $P(\mathbf{y}|A(\mathbf{x}, \tilde{\mathbf{x}}; \mathbf{w})\tilde{\mathbf{f}})$ is derived under the assumption that $\mathbf{f} = A(\mathbf{x}, \tilde{\mathbf{x}}; \mathbf{w})\tilde{\mathbf{f}}$ and depends on form of the sampling model $P(\mathbf{Y}|\mathbf{f}, \sigma)$, e.g., when the latter is $N(\mathbf{f}, \sigma \mathbf{I}_n)$ we obtain $P(\mathbf{y}|A(\mathbf{x}, \tilde{\mathbf{x}}; \mathbf{w})\tilde{\mathbf{f}}) = N(A(\mathbf{x}, \tilde{\mathbf{x}}; \mathbf{w})\tilde{\mathbf{f}}, \sigma \mathbf{I}_n)$.

The posterior distribution $\pi(\tilde{\mathbf{f}}, \mathbf{w}|\mathbf{y}, \mathbf{x})$ is not tractable, but sampling from it will be much less expensive since $K(\mathbf{x}, \tilde{\mathbf{x}}; \mathbf{w}) \in \mathbf{R}^{n\times m}$ and $K(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}; \mathbf{w}) \in \mathbf{R}^{m\times m}$. While the inducing inputs $\tilde{\mathbf{x}}$ can be selected from the samples collected, we will use an alternative approach where we group the covariate values observed, $\mathbf{x}$, into m clusters, and choose the cluster-specific covariate averages as $\tilde{x}_1, \ldots, \tilde{x}_m$. For instance, given a specific value k, one can use a simple k-means algorithm (Bishop, 2006) to classify $\mathbf{x}$ into k clusters and estimate clusters' means using an iterative method. Intuitively, it makes sense to have more inducing points in regions that exhibit more variation in covariate values.

Finally, in order to reduce the dimensionality of the parameter space, we assume that

$$f(x_i) = f(x_i^T \beta), \tag{15}$$

and we set $\tilde{\mathbf{f}} = (f(\tilde{z}_1), \ldots, f(\tilde{z}_m))^T$, where $(\tilde{z}_1, \ldots, \tilde{z}_m)$ are inducing variables in $\mathbf{R}$, $f : \mathbf{R} \to \mathbf{R}$ is an unknown function that is of interest and $\beta \in \mathbf{R}^q$ is normalized, i.e. $\|\beta\| = 1$. Note that without normalization the parameter β is not identifiable. Here $\{\tilde{z}_1, \ldots, \tilde{z}_m\}$ play the same role as $\{\tilde{x}_1, \ldots, \tilde{x}_m\}$ for general sparse GP. They help sample the posterior latent variables much faster and should be spread in the range of $\{x_1^T\beta, \ldots, x_n^T\beta\}$. In the next section we show how to choose the positions of these inducing inputs. The *single index model (SIM)* defined by (15) coupled with the sparse GP approach (henceforth denoted as GP-SIM) has the advantage that it casts the original problem of estimating a general function f in q dimensions based on n observations into the estimation of q-dimensional parameter vector β and of the one-dimensional map f based on $m \ll n$ inducing points. The GP-SIM approach was successfully implemented for mean regression problems (Choi et al., 2011; Gramacy and Lian, 2012) and quantile regression (Hu et al., 2013). It can be used for large covariate dimension and is much more flexible than the simple linear model.

3. GP-SIM for conditional copula

We consider a bivariate response variable $(Y_1, Y_2) \in \mathbf{R}^2$ together with covariate measurement $\mathbf{x} \in \mathbf{R}^q$. Hence, the data $\mathcal{D} = \{(y_{1i}, y_{2i}, x_i),\ i = 1 \ldots n\}$ consist of triplets (y_{1i}, y_{2i}, x_i) where $y_{1i}, y_{2i} \in \mathbf{R}$ and $x_i \in \mathbf{R}^q$. For notational convenience, let $\mathbf{y}_1 = (y_{11}, \ldots, y_{1n})^T$, $\mathbf{y}_2 = (y_{21}, \ldots, y_{2n})^T$ and $\mathbf{x} = (x_1, \ldots, x_n)^T$. We assume that the marginal distribution of Y_j $(j = 1, 2)$ is Gaussian with mean $f_j(x)$ and constant variance σ_j^2. If we let $\mathbf{Y}_j = (Y_{j1}, \ldots, Y_{jn})^T,\ j = 1, 2$, and $\mathbf{f}_j = (f_j(x_1), \ldots, f_j(x_n))^T$ we can compactly write:

$$\mathbf{Y}_j \sim \mathcal{N}(\mathbf{f}_j, \sigma_j^2 \mathbf{I}_n) \quad j = 1, 2. \tag{16}$$

Generally, it is difficult to discern whether the copula structure varies with covariates or not, so we consider a conditional copula to account for the more general situation. Therefore, the likelihood function is

$$\begin{aligned} L(\omega) = \prod_{i=1}^{n} \frac{1}{\sigma_1} \phi\left(\frac{y_{1i} - \mathbf{f}_{1i}}{\sigma_1}\right) \frac{1}{\sigma_2} \phi\left(\frac{y_{2i} - \mathbf{f}_{2i}}{\sigma_2}\right) \times \\ \times c_{\theta(x_i)}\left(\Phi\left(\frac{y_{1i} - \mathbf{f}_{1i}}{\sigma_1}\right), \Phi\left(\frac{y_{2i} - \mathbf{f}_{2i}}{\sigma_2}\right)\right), \end{aligned} \tag{17}$$

where c denotes a parametric copula density function, ω denotes all the parameters in the model, while Φ and ϕ are the cumulative probability function and density function of a standard normal distribution, respectively. The parameter of a copula depends on the unknown function $\theta(x_i) = g^{-1}(f(x_i))$, where f is assumed to take the form given in (15) and g is a known invertible link function that allows an unrestricted parameter space for $\mathbf{f}$. Note that the form of the GP-SIM model used for estimating the copula parameter is invariant to non-linear transformations. This implies that the formulation of the model is the same whether we directly estimate the copula parameter, $\theta(x)$, Kendall's $\tau(x)$, or other measures of dependence. However, this is not true if we use an additive model for $\theta(x)$, since additivity is not preserved by non-linear transformations.

The GP-SIM is fully specified once we assign the GP priors to f_1, f_2, f and the parametric priors for the remaining parameters, as follows:

$$\begin{aligned} f_1 &\sim \mathcal{GP}(\mathbf{w}_1), \quad f_2 \sim \mathcal{GP}(\mathbf{w}_2), \quad f \sim \mathcal{GP}(\mathbf{w}), \\ \mathbf{w}_1 &\sim \mathcal{N}(0, 5\mathbf{I}_{q+1}), \quad \mathbf{w}_2 \sim \mathcal{N}(0, 5\mathbf{I}_{q+1}), \quad \mathbf{w} \sim \mathcal{N}(0, 5\mathbf{I}_2), \\ \beta &\sim \mathrm{U}(S^{q-1}), \quad \sigma_1^2 \sim \mathcal{IG}(0.1, 0.1), \quad \sigma_2^2 \sim \mathcal{IG}(0.1, 0.1). \end{aligned} \tag{18}$$

The $\mathcal{GP}(\mathbf{w})$ is a Gaussian Process prior with mean zero, squared exponential kernel with parameters $\mathbf{w}$, $\mathrm{U}(S^{q-1})$ is a uniform distribution on the surface of the q-dimensional unit sphere and $\mathcal{IG}(\alpha, \beta)$ denotes the inverse gamma distribution. The above prior for $\mathbf{w}$ captures very wiggly functions for small values of $\mathbf{w}$ and almost constant functions for large values of $\mathbf{w}$. The priors for marginal variances are vague and would be conjugate in the absence of the copula term. In our experience, the results are not sensitive to the choice of hyperparameter values. Because the focus of the paper is on inference for the copula, we allow f_1 and f_2 to be evaluated on $\mathbf{R}^q$ while f is on $\mathbf{R}$. In order to avoid computational problems that affect the GP-based inference when the sample size is large, the inference will rely on the Sparse GP method that was described in the previous section. Suppose $\tilde{\mathbf{x}}_1$ are m_1 inducing inputs for function f_1, $\tilde{\mathbf{x}}_2$ are m_2 inducing inputs for function f_2 and $\tilde{\mathbf{z}}$ are m inducing inputs for function f. The number of inducing inputs m_1, m_2 and m can all be different, but in our applications we will choose their values equal and significantly smaller than the sample size, n. The choice is motivated by imperative computational time restrictions, given the large number of numerical simulations we perform to investigate the performance of the approach in terms of estimation and model selection. In practice, the analyst should ideally use the largest number of inducing points supported by the computing environment. As suggested earlier, we define $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ as centers of m_1 and m_2 clusters of $\mathbf{x}$. If $m_1 = m_2$ then the inducing inputs are the same. We cannot use the same strategy for $\tilde{\mathbf{z}}$, since then we would need the centers for the clusters of the variable $\mathbf{x}^T\beta$ which are unknown. If we assume that each covariate x_{is} is between 0 and 1 (this can be achieved easily if we subtract the minimum value and divide by range) then following the Cauchy–Schwarz inequality we obtain

$$\|x_i^T\beta\| \le \sqrt{\|x_i\|^2\|\beta\|^2} \le \sqrt{q} \quad \forall x_i, \beta.$$

Hence we can choose $\tilde{\mathbf{z}}$ to be m equally spaced points in the interval $[-\sqrt{q}, \sqrt{q}]$.

Let $\tilde{\mathbf{f}}_1$ be f_1 evaluated at $\tilde{\mathbf{x}}_1$, $\tilde{\mathbf{f}}_2$ be f_2 evaluated at $\tilde{\mathbf{x}}_2$ and $\tilde{\mathbf{f}}$ be f evaluated at $\tilde{\mathbf{z}}$. Then the joint density of the observed data and parameters is proportional to:

$$\begin{aligned} P(\mathbf{y}_1, \mathbf{y}_2, \tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2, \tilde{\mathbf{f}}, \mathbf{w}_1, \mathbf{w}_2, \mathbf{w}, \sigma_1^2, \sigma_2^2, \beta | \mathbf{x}, \tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \tilde{\mathbf{z}}) &\propto p_N(\mathbf{y}_1; \mathbf{f}_1, \sigma_1^2\mathbf{I}_n) p_N(\mathbf{y}_2; \mathbf{f}_2, \sigma_2^2\mathbf{I}_n) \times \\ &\times \prod_{i=1}^{n} c_{g^{-1}(\mathbf{f}_i)}\left(\Phi\left(\frac{y_{1i} - \mathbf{f}_{1i}}{\sigma_1}\right), \Phi\left(\frac{y_{2i} - \mathbf{f}_{2i}}{\sigma_2}\right)\right) p_N(\tilde{\mathbf{f}}_1; 0, K(\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_1; \mathbf{w}_1)) \times \\ &\times p_N(\tilde{\mathbf{f}}_2; 0, K(\tilde{\mathbf{x}}_2, \tilde{\mathbf{x}}_2; \mathbf{w}_2)) p_N(\tilde{\mathbf{f}}; 0, K(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}; \mathbf{w})) p_N(\mathbf{w}_1; 0, 5\mathbf{I}_{q+1}) \times \\ &\times p_N(\mathbf{w}_2; 0, 5\mathbf{I}_{q+1}) p_N(\mathbf{w}; 0, 5\mathbf{I}_2) p_{IG}(\sigma_1^2; 0.1, 0.1) p_{IG}(\sigma_2^2; 0.1, 0.1), \end{aligned} \tag{19}$$

where $\mathbf{f}_1 = A(\mathbf{x}, \tilde{\mathbf{x}}_1; \mathbf{w}_1)\tilde{\mathbf{f}}_1$, $\mathbf{f}_2 = A(\mathbf{x}, \tilde{\mathbf{x}}_2; \mathbf{w}_2)\tilde{\mathbf{f}}_2$, $\mathbf{f} = A(\mathbf{x}^T\beta, \tilde{\mathbf{z}}; \mathbf{w})\tilde{\mathbf{f}}$ and p_N and p_{IG} are multivariate normal and inverse gamma densities, respectively. Although here we adopt a full GP prior for the marginal models, the approach can be easily adapted to consider GP-SIM models for the marginals too.

The contribution of the conditional copula model to the joint likelihood breaks the tractability of the posterior conditional densities and complicates the design of an efficient MCMC algorithm that can sample efficiently from the posterior distribution. The conditional joint posterior distribution of the latent variables ($\mathbf{f}$) and parameters ($\mathbf{w}$) given the observed data $\mathcal{D}$ does not have a tractable form and its study will require the use of Markov Chain Monte Carlo (MCMC) sampling methods. Specifically, we use Random Walk Metropolis (RWM) within Gibbs sampling for $\mathbf{w}$ (Craiu and Rosenthal, 2014; Rosenthal, 2009; Andrieu et al., 2003) while for $\mathbf{f}$ we will use the elliptical slice sampling (Murray et al., 2010) that has been designed specifically for GP-based models and does not require tuning of free parameters.

3.1. Computational algorithm

Inference is based on the posterior distribution $\pi(\omega|\mathcal{D}, \tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \tilde{\mathbf{z}})$ where $\omega = (\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2, \tilde{\mathbf{f}}, \mathbf{w}_1, \mathbf{w}_2, \mathbf{w}, \sigma_1^2, \sigma_2^2, \beta) \in \mathbf{R}^k$ represents the vector of parameters and latent variables in the model, with $k = 3m + 3q + 7$. Since the posterior is not mathematically tractable, its properties will be explored via Markov chain Monte Carlo (MCMC) sampling. In this section we provide the detailed steps of the MCMC sampler designed to sample from π. The general form of the algorithm falls within the class of Metropolis-within-Gibbs (MwG) samplers in which we update in turn each component of the chain by sampling from its conditional distribution, given all the other components. The presence of the copula in the likelihood breaks the usual conditional conjugacy of the GP models so none of the components have conditional distributions that can be sampled directly.

Suppose we are interested in sampling a target $\pi(\omega)$. A generic MwG sampler proceeds as follows:

Step I Initialize the chain at $\omega_1^{(1)}, \omega_2^{(1)}, \ldots, \omega_k^{(1)}$.

Step R At iteration $t+1$ run iteratively the following steps for each $j = 1, \ldots, k$:

1. Sample $\omega_j^* \sim q_j(\cdot|\omega_j^{(t)}, \omega_{-j}^{(t+1;t)})$ where $\omega_{-j}^{(t+1;t)} = (\omega_1^{(t+1)}, \ldots, \omega_{j-1}^{(t+1)}, \omega_{j+1}^{(t)}, \ldots, \omega_k^{(t)})$ is the most recent state of the chain with the first $j-1$ components updated already (hence the supraindex $t+1$), the jth component removed and the remaining $n-j$ components having the values determined at iteration t (hence the supraindex t).
2. Compute $r = \min\left\{1, \frac{\pi(\omega_1^{(t+1)},\ldots,\omega_{j-1}^{(t+1)},\omega_j^*,\omega_{j+1}^{(t)},\ldots,\omega_k^{(t)})q_j(\omega_j^{(t)}|\omega_j^*,\omega_{-j}^{(t+1;t)})}{\pi(\omega_1^{(t+1)},\ldots,\omega_{j-1}^{(t+1)},\omega_j^{(t)},\omega_{j+1}^{(t)},\ldots,\omega_k^{(t)})q_j(\omega_j^{(*)}|\omega_j^{(t)},\omega_{-j}^{(t+1;t)})}\right\}$.
3. With probability r accept proposal and set $\omega_j^{(t+1)} = \omega_j^*$ and with $1-r$ reject proposal and let $\omega_j^{(t+1)} = \omega_j^{(t)}$.

The proposal density $q_j(\cdot|\cdot)$ corresponds to the transition kernel used for the jth component. Our algorithm uses a number of proposals corresponding to Random Walk Metropolis-within-Gibbs (RWMwG), Independent Metropolis-within-Gibbs (IMwG) and Elliptical Slice Sampling within Gibbs (SSwG) moves.

At the $t+1$ step we use the following proposals to update the chain:

$\mathbf{w}_j$: Use a RWM transition kernel: $\mathbf{w}^* \sim \mathcal{N}(\mathbf{w}_j^{(t)}, c_{w_j}\mathbf{I}_{q+1})$. The constant c_{w_j} is chosen so that the acceptance rate is about 30%, $j = 1, 2$.

$\mathbf{w}$: Use the RWM: $\mathbf{w}^* \sim \mathcal{N}(\mathbf{w}^{(t)}, c_w\mathbf{I}_2)$. The constant c_w is chosen so that the acceptance rate is about 30%.

σ_j^2: Without the copula, the conditional posterior distribution of σ_j^2 would be $\mathcal{IG}(0.1+n/2, 0.1+(\mathbf{y}_j - A_j\tilde{\mathbf{f}}_j^{(t)})^T(\mathbf{y}_j - A_j\tilde{\mathbf{f}}_j^{(t)}))$, where $A_j = A(\mathbf{x}, \tilde{\mathbf{x}}_j; \mathbf{w}_j^{(t+1)})$ for all $j = 1, 2$. We will use this distribution to build and independent Metropolis (IM) type of transition for $\sigma_j^2, j = 1, 2$. The acceptance rate is usually in the range of $[0.25, 0.60]$ and the chain mixes better than it would under a RWM.

β: Since β is normalized we will use RWM on unit sphere using 'Von-Mises–Fisher' distribution (henceforth denoted as $\mathcal{VMF}$). The $\mathcal{VMF}$ distribution has two parameters, μ (normalized to have norm one) representing the mean direction and κ, the concentration parameter. A larger κ implies that the distribution will be more concentrated around μ. The density is symmetric in μ and the argument and is proportional to $f_{VMF}(x; \mu, \kappa) \propto \exp(\kappa x^T \mu)$. The proposals are generated using $\beta^* \sim \mathcal{VMF}(\beta^{(t)}, \kappa)$, where κ is chosen so that the acceptance rate is around 30%.

$\tilde{\mathbf{f}}$'s: For $\tilde{\mathbf{f}}_j$, $j = 1, 2$ and $\tilde{\mathbf{f}}$ we use the elliptical slice sampling proposed by Murray et al. (2010) which does not require the tuning of simulation parameters. Although not needed in our examples, we note that if the chain's mixing is sluggish, one can improve it using the parallelization strategy proposed by Nishihara et al. (2014).

In our experience the efficiency of the algorithm benefits from initial values that are not too far from the posterior mode. Therefore we propose first to roughly estimate the parameters in the two independent regressions for $\mathbf{y}_1$ and $\mathbf{y}_2$ to get $(\tilde{\mathbf{f}}_1, \mathbf{w}_1, \sigma_1^2)^{(1)}$ and $(\tilde{\mathbf{f}}_2, \mathbf{w}_2, \sigma_2^2)^{(1)}$. Then run another MCMC fixing the marginal parameters and only sampling $(\tilde{\mathbf{f}}, \mathbf{w})$. This procedure estimates $(\tilde{\mathbf{f}}, \mathbf{w})^{(1)}$. These 3 short chains (100–200 iterations each) provide good initial values for the joint MCMC sampler. This simple approach shortens the time it would take for the original chain to find the regions of high mass under the posterior. We have also found that the chain's mixing is accelerated when initial value of the second component of $\mathbf{w}$ is small, thus allowing for more variation in the calibration function.

Remark. In our numerical experiments, we will fit the GP-SIM model to data with constant calibration, i.e., with true values $\beta_i = 0$ for all $1 \le i \le q$. The constraint $\|\beta\| = 1$ forbids sampling null values for all the components of β simultaneously, and instead the MCMC draws for β's components are spread randomly in the support. However, the shape of the calibration function is correctly recovered since the sampled values for the second component of $\mathbf{w}$ were large reflecting the perfect dependence between $f(x_i^T\beta)$ and $f(x_j^T\beta)$ for any $1 \le i \ne j \le n$. This led to difficulties in identifying the SA as discussed below, and compelled us to develop a new SA identification procedure that is described in Section 6.2.

4. Model selection

The conditional copula model involves two types of selection. First one needs to choose the copula family from a set of possible candidates. Second, it is often of interest to determine whether a parametric simple form for the calibration is supported by the data. For instance, a constant calibration function indicates that the dependence structure does not vary with the covariates, a conclusion that may be of scientific interest in some applications. Let $\boldsymbol{\omega}^{(t)}$ denote the vector of parameters and latent variables drawn at step t from the posterior corresponding to model $\mathcal{M}$. We investigate the performance of three measures of fit that can be estimated from the MCMC samples $\boldsymbol{\omega}^{(t)}, t = 1 \ldots M$.

4.0.1. *Cross-validated pseudo marginal likelihood*

The cross-validated pseudo marginal likelihood (CVML) (Geisser and Eddy, 1979; Hanson et al., 2011) calculates the average (over parameter values) prediction power for model $\mathcal{M}$ via

$$\mathrm{CVML}(\mathcal{M}) = \sum_{i=1}^{n} \log\left(P(y_{1i}, y_{2i}|\mathcal{D}_{-i}, \mathcal{M})\right), \tag{20}$$

where $\mathcal{D}_{-i}$ is the data set from which the ith observation has been removed. An estimate of (20) can be obtained using posterior draws for all the parameters and latent variables in the model (see, for example, Gelfand et al., 1992). Specifically, if the latter are denoted by $\boldsymbol{\omega}$, then

$$E\left[P(y_{1i}, y_{2i}|\boldsymbol{\omega}, \mathcal{M})^{-1}\right] = P(y_{1i}, y_{2i}|\mathcal{D}_{-i}, \mathcal{M})^{-1}, \tag{21}$$

where the expectation is with respect to conditional distribution of $\boldsymbol{\omega}$ given full data $\mathcal{D}$ and the model $\mathcal{M}$. Based on the posterior samples we can estimate the CVML as

$$\mathrm{CVML}_{est}(\mathcal{M}) = -\sum_{i=1}^{n} \log\left(\frac{1}{M}\sum_{t=1}^{M} P(y_{1i}, y_{2i}|\boldsymbol{\omega}^{(t)}, \mathcal{M})^{-1}\right). \tag{22}$$

The model with the largest CVML is selected.

4.0.2. *Conditional CVML criterion*

The conditional copula construction is particularly useful in predicting one response given the other ones. We exploit this feature by computing the predictive distribution of one response given the rest of the data. The resulting *conditional CVML (CCVML)* is derived from the $P(y_{1i}|y_{2i}, \mathcal{D}_{-i})$ and $P(y_{2i}|y_{1i}, \mathcal{D}_{-i})$ via

$$\mathrm{CCVML}(\mathcal{M}) = \frac{1}{2}\left\{\sum_{i=1}^{n} \log\left[P(y_{1i}|y_{2i}, \mathcal{D}_{-i}, \mathcal{M})\right] + \sum_{i=1}^{n} \log\left[P(y_{2i}|y_{1i}, \mathcal{D}_{-i}, \mathcal{M})\right]\right\}. \tag{23}$$

Note that when the marginal distributions are uniform, CCVML is the same as CVML. One can easily show that

$$\begin{aligned} E\left[P(y_{1i}|y_{2i}, \boldsymbol{\omega}, \mathcal{M})^{-1}\right] &= E\left[\frac{P(y_{2i}|\boldsymbol{\omega}, \mathcal{M})}{P(y_{1i}, y_{2i}|\boldsymbol{\omega}, \mathcal{M})}\right] = P(y_{1i}|y_{2i}, \mathcal{D}_{-i}, \mathcal{M})^{-1}, \\ E\left[P(y_{2i}|y_{1i}, \boldsymbol{\omega}, \mathcal{M})^{-1}\right] &= E\left[\frac{P(y_{1i}|\boldsymbol{\omega}, \mathcal{M})}{P(y_{1i}, y_{2i}|\boldsymbol{\omega}, \mathcal{M})}\right] = P(y_{2i}|y_{1i}, \mathcal{D}_{-i}, \mathcal{M})^{-1}. \end{aligned} \tag{24}$$

Based on (24) CCVML is estimated from MCMC samples using

$$\mathrm{CCVML}_{est}(\mathcal{M}) = -\frac{1}{2}\sum_{i=1}^{n}\left\{\log\left[\frac{1}{M}\sum_{t=1}^{M}\frac{P(y_{2i}|\boldsymbol{\omega}^{(t)}, \mathcal{M})}{P(y_{1i}, y_{2i}|\boldsymbol{\omega}^{(t)}, \mathcal{M})}\right] + \log\left[\frac{1}{M}\sum_{t=1}^{M}\frac{P(y_{1i}|\boldsymbol{\omega}^{(t)}, \mathcal{M})}{P(y_{1i}, y_{2i}|\boldsymbol{\omega}^{(t)}, \mathcal{M})}\right]\right\}. \tag{25}$$

4.0.3. *Watanabe–Akaike Information Criterion*

The Watanabe–Akaike Information Criterion (WAIC, Watanabe, 2010) is an information-based criterion that is closely related to the CVML, as discussed in Watanabe (2013), Gelman et al. (2014) and Vehtari et al. (2017).

The WAIC is defined as

$$\mathrm{WAIC}(\mathcal{M}) = -2\mathrm{fit}(\mathcal{M}) + 2\mathrm{p}(\mathcal{M}), \tag{26}$$

where the model fitness is

$$\mathrm{fit}(\mathcal{M}) = \sum_{i=1}^{n} \log E\left[P(y_{1i}, y_{2i}|\boldsymbol{\omega}, \mathcal{M})\right] \tag{27}$$

and the penalty

$$\mathrm{p}(\mathcal{M}) = \sum_{i=1}^{n} Var[\log P(y_{1i}, y_{2i}|\boldsymbol{\omega}, \mathcal{M})]. \tag{28}$$

Table 1
Parameter's range, Inverse-link functions and the functional relationship between Kendall's τ and the copula parameter.

Copula	Range of parameter (θ)	Inv-Link function	Kendall's τ formula
Clayton	$(-1, \infty) \setminus \{0\}$	$\theta = \exp(f) - 1$	$\tau = \frac{\theta}{\theta+2}$
Frank	$(-\infty, \infty) \setminus \{0\}$	$\theta = f$	No closed form
Gaussian, T	$(-1, 1)$	$\theta = \frac{\exp(f)-1}{\exp(f)+1}$	$\tau = \frac{2}{\pi} \arcsin\theta$
Gumbel	$(1, \infty)$	$\theta = \exp(f) + 1$	$\tau = 1 - \frac{1}{\theta}$

The expectation in (27) and the variance in (28) are with respect to the conditional distribution of $\boldsymbol{\omega}$ given the data and can be computed using Monte Carlo samples from π. For instance, the Monte Carlo estimate of the fit is

$$\widehat{\text{fit}}(\mathcal{M}) = \sum_{i=1}^{n} \log\left(\frac{\sum_{t=1}^{M} P(y_{1i}, y_{2i}|\boldsymbol{\omega}^{(t)}, \mathcal{M})}{M}\right), \tag{29}$$

and p($\mathcal{M}$) can be estimated similarly using the posterior samples. The model with the smallest WAIC is preferred. In the next section we also investigate via simulations the performance of CVML, CCVML and WAIC criteria when identifying data support for a constant calibration function.

In Watanabe (2013) it was demonstrated that CVML and WAIC are asymptotically equivalent, so that CVML($\mathcal{M}$) $\approx$ WAIC($\mathcal{M}$)/(-2) for a large sample size n. This connection can be extended to CCVML using the following two conditional WAICs:

$$\text{CWAIC}_1(\mathcal{M}) = -2\sum_{i=1}^{n} \log E\left[P(y_{1i}|y_{2i}, \boldsymbol{\omega}, \mathcal{M})\right] + 2\sum_{i=1}^{n} Var[\log P(y_{1i}|y_{2i}, \boldsymbol{\omega}, \mathcal{M})], \tag{30}$$

$$\text{CWAIC}_2(\mathcal{M}) = -2\sum_{i=1}^{n} \log E\left[P(y_{2i}|y_{1i}, \boldsymbol{\omega}, \mathcal{M})\right] + 2\sum_{i=1}^{n} Var[\log P(y_{2i}|y_{1i}, \boldsymbol{\omega}, \mathcal{M})], \tag{31}$$

where expectation and variance are with respect to the conditional distribution of $\boldsymbol{\omega}$ given the observed data. An argument that follows directly the one in Vehtari et al. (2017) shows that CCVML and $\frac{1}{2}$ {CWAIC$_1$ + CWAIC$_2$} are also asymptotically equivalent.

5. Performance of the algorithms

5.1. Simulations

The purpose of the simulation study is to assess empirically: (1) the performance of the estimation method under the correct and misspecified models, as well as (2) the ability of the model selection criteria to identify the correct copula structure, i.e. the copula family and the parametric form of the calibration function. For the former aim we compute the integrated mean square for various quantities of interest, including the Kendall's τ. In order to facilitate the assessment of the estimation performance across different copula families, we estimate the calibration function on the Kendall's τ scale. The latter is given by

$$\tau(\mathbf{x}) = 4\left(\iint C(u_1, u_2|\mathbf{x})c(u_1, u_2|\mathbf{x})du_1du_2\right) - 1.$$

We will compare 3 copulas: Clayton, Frank and Gaussian under the general GP-SIM model and the Clayton with constant calibration function. To fit the model with constant copula, we still use MCMC but instead of $\mathbf{f}$, $\tilde{\mathbf{f}}$, $\mathbf{w}$ and β in calibration we have a constant scalar copula parameter, θ. The RWMwG transition is used to sample θ, as the proposal distributions for marginals' parameters and latent variables remain the same.

Table 1 provides inverse-link functions g^{-1} used for calibration, the functional relationship between Kendall's τ and copula parameters and parameter ranges for every copula family used in the paper.

In addition to Kendall's τ we use also the conditional mean of Y_1 given y_2 and x for assessing the estimation. Such conditional means can be useful in prediction when one of the responses is more expensive to measure than the other. The calculation is mathematically straightforward

$$E(Y_1|Y_2 = y_2, \mathbf{x}) = f_1(\mathbf{x}) + \sigma_1 \int_0^1 \Phi^{-1}(z)c\left(z, \Phi\left(\frac{y_2 - f_2(\mathbf{x})}{\sigma_2}\right); \theta(\mathbf{x})\right) dz. \tag{32}$$

The integral in (32) is usually not tractable, but can be easily estimated via numerical integration since it is one-dimensional and defined on the closed interval [0, 1].

5.2. *Simulation details*

We generate samples of size $n = 400$ from each of the next 6 scenarios using the Clayton copula. The covariates are generated independently from Uniform$(0, 1)$ distribution. The covariate dimension q in Scenario 3 is 10, in all other scenarios it is 2.

Sc1 $f_1(\mathbf{x}) = 0.6\sin(5x_1) - 0.9\sin(2x_2)$,
$f_2(\mathbf{x}) = 0.6\sin(3x_1 + 5x_2)$,
$\tau(\mathbf{x}) = 0.7 + 0.15\sin(15\mathbf{x}^T\beta)$
$\beta = (1, 3)^T/\sqrt{10}, \sigma_1 = \sigma_2 = 0.2$
Sc2 $f_1(\mathbf{x}) = 0.6\sin(5x_1) - 0.9\sin(2x_2)$
$f_2(\mathbf{x}) = 0.6\sin(3x_1 + 5x_2)$
$\tau(\mathbf{x}) = 0.3\sin(5\mathbf{x}^T\beta)$
$\beta = (1, 3)^T/\sqrt{10}, \sigma_1 = \sigma_2 = 0.2$
Sc3 $\beta = (1, 10, -3, 6, 1, -6, 3, 7, -1, -5)^T/\sqrt{267}, \sigma_1 = \sigma_2 = 0.2$
$f_1(\mathbf{x}) = \cos(\mathbf{x}^T\beta)$
$f_2(\mathbf{x}) = \sin(\mathbf{x}^T\beta)$
$\tau(\mathbf{x}) = 0.7 + 0.20\sin(5\mathbf{x}^T\beta)$
Sc4 $f_1(\mathbf{x}) = 0.6\sin(5x_1) - 0.9\sin(2x_2)$
$f_2(\mathbf{x}) = 0.6\sin(3x_1 + 5x_2)$
$\tau(\mathbf{x}) = 0.5$
$\sigma_1 = \sigma_2 = 0.2$
Sc5 $f_1(\mathbf{x}) = 0.6\sin(5x_1) - 0.9\sin(2x_2)$
$f_2(\mathbf{x}) = 0.6\sin(3x_1 + 5x_2)$
$\eta(\mathbf{x}) = 1 + 0.7\sin(3x_1^3) - 0.5\cos(6x_2^2)$
$\sigma_1 = \sigma_2 = 0.2$
Sc6 $f_1(\mathbf{x}) = 0.6\sin(5x_1) - 0.9\sin(2x_2)$
$f_2(\mathbf{x}) = 0.6\sin(3x_1 + 5x_2)$
$\eta(\mathbf{x}) = 1 + 0.7x_1 - 0.5x_2^2$
$\sigma_1 = \sigma_2 = 0.2$

Sc1 and **Sc2** have calibration functions for which the SIM model is true for Kendall's τ and, consequently, also for the copula parameter. **Sc1** corresponds to large dependence (τ greater than 0.5) while **Sc2** has small dependence (τ is between -0.3 and 0.3). **Sc3** also has SIM form for calibration function but the covariate dimension is $q = 10$, so this scenario is important in our effort to evaluate how well the algorithms scale up with dimension. **Sc4** corresponds to the covariate-free dependence ($\tau = 0.5$) and allows us to verify the power to detect simple parametric forms for the calibration. Scenarios **Sc5** and **Sc6** do not have SIM form, but have additive calibration function (as in Sabeti et al., 2014). They are used to evaluate the effect of model misspecification on the inference. Note that **Sc6** has almost SIM calibration when $x_2 \in [0, 1]$. Additional simulation scenarios with larger sample size ($n = 1000$) and different β values are included in the online supplemental material. From our experiments we found that when the number of inducing points is $m = 30$ for marginals and we use calibration sparse GPs, we obtain a reasonable CPU time that allows us to perform the desired number of replications while capturing the general form of the estimated functions. On average one MCMC iteration ($n = 400$) with GP-SIM calibration takes 0.02 s, one iteration with constant calibration (and GP for marginals) takes 0.015 s. The MCMC samplers were run for 20,000 iterations for all scenarios.

The first half of the MCMC sample is discarded as burn-in and the second half is used for inference. As noted earlier, starting values were found by running two GP regressions separately to estimate marginal parameters and one MCMC sampler was run in order to estimate calibration parameters. All three samplers were run for only 100 iterations.

5.2.1. *Proof of concept based on one replicate*

In the absence of computable convergence bounds, we used the Gelman–Rubin (Gelman and Rubin, 1992) diagnostic statistics to decide the length of the chain's run. To illustrate using **Sc1**, we ran 10 independent MCMC chains, each for 20,000 iterations, that were started from different initial values. The trace plots for the *potential scale reduction factor (PSRF)*, computed up to 10,000 iterations for β, σ_1^2 and σ_2^2 are displayed in Fig. 1. The plots show that the multivariate PSRF after 10,000 iterations is 1.1. The subsequent 10,000 samples were used for inference.

Parameter Estimation

The simulation results show that inferences performed under **Sc1** and **Sc2** are similar. Since the calibration function in **Sc1** is more complicated, for the sake of reducing the paper's length we present only results for that scenario. The trace-plots, autocorrelation functions and histograms of posterior samples of β, σ_1^2 and σ_2^2 are shown in Fig. 2 when the fitted copula belongs to the correct Clayton family (the horizontal solid red line is the true value). Next we show predictions for the marginals means with 95% credible intervals. Since these are 2-dimensional we estimate 'slices' of this surface at values 0.2 and 0.8, so that we first fix $x_1 = 0.2$ then $x_1 = 0.8$ and similarly for x_2. The results are in Fig. 3 (black is true, green is estimation, red are credible intervals).

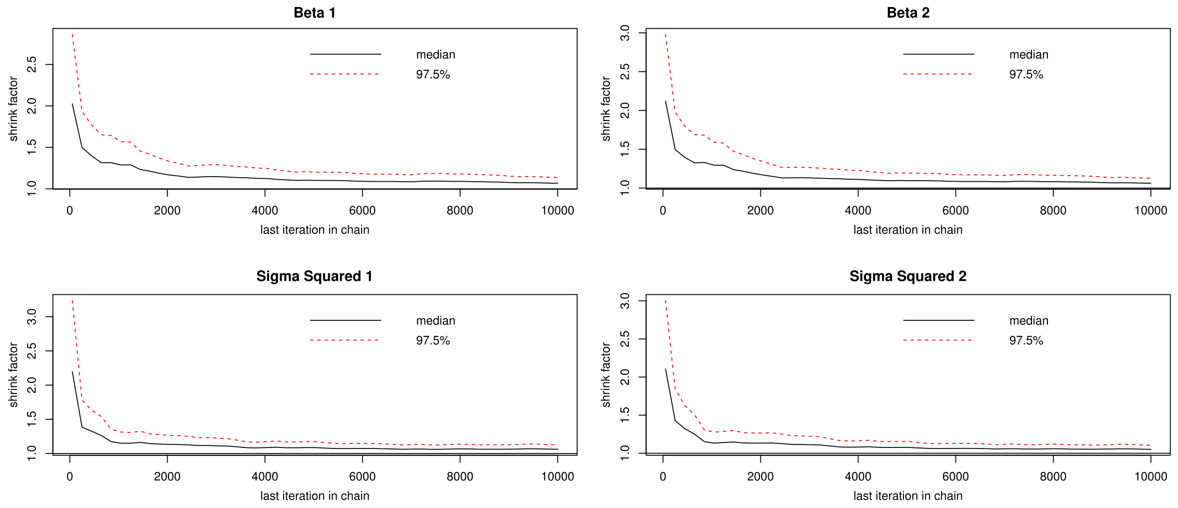

Fig. 1. **Sc1**: Clayton copula, Gelman–Rubin MCMC diagnostic for beta and two variances.

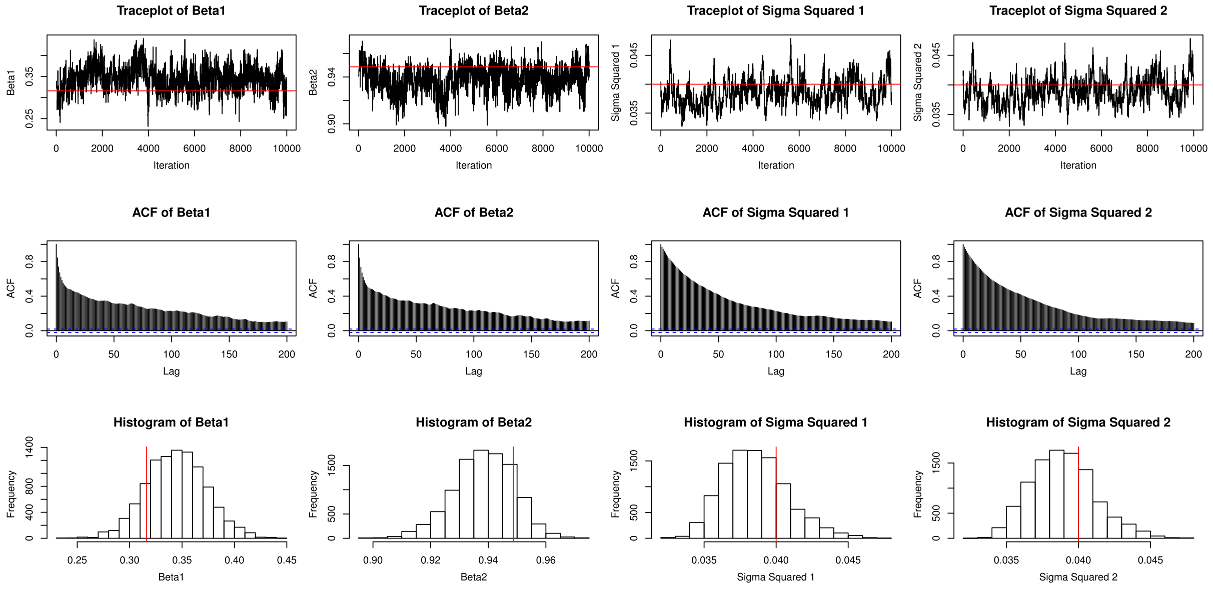

Fig. 2. **Sc1**: Trace-plots, ACFs and histograms of parameters based on MCMC samples generated under the true Clayton family.

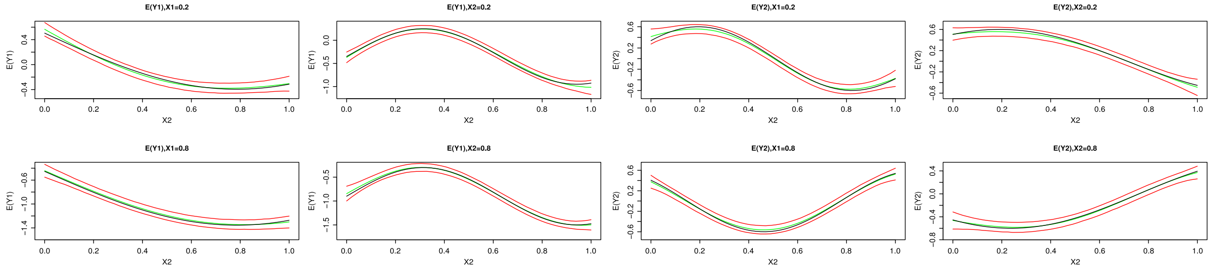

Fig. 3. **Sc1**: Estimation of marginal means. The leftmost 2 columns show the accuracy for predicting $E(Y_1)$ and the rightmost 2 columns show the results for predicting $E(Y_2)$. The black and green lines represent the true and estimated relationships, respectively. The red lines are the limits of the pointwise 95% credible intervals obtained under the true Clayton family. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

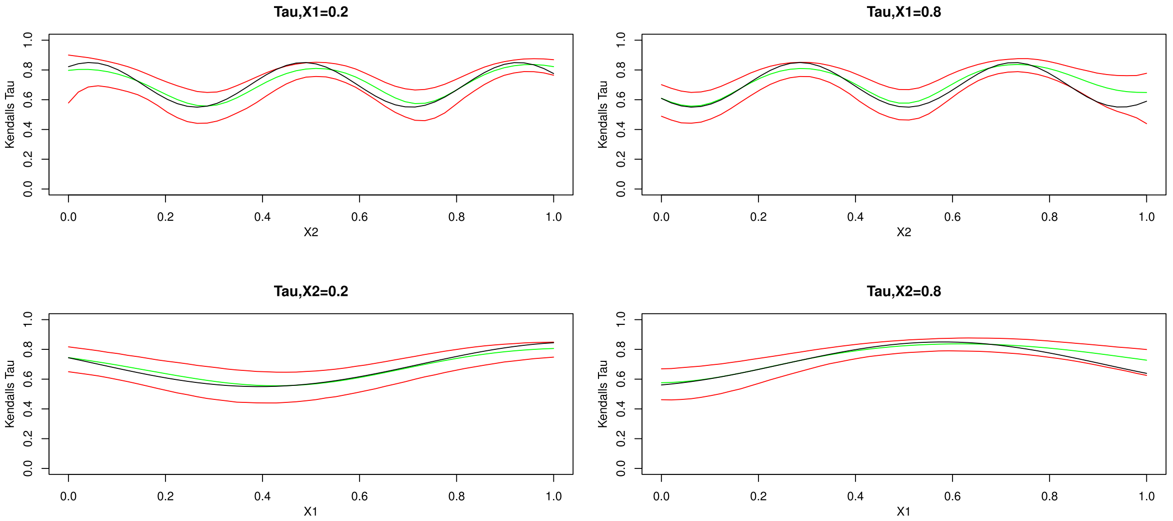

Fig. 4. **Sc1**: Estimation of Kendall's τ one-dimensional projections when $x_1 = 0.2$ or 0.8 (top panels) and when $x_2 = 0.2$ or 0.8 (bottom panels). The black and green lines represent the true and estimated relationships, respectively. The red lines are the limits of the pointwise 95% credible intervals obtained under the true Clayton family. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

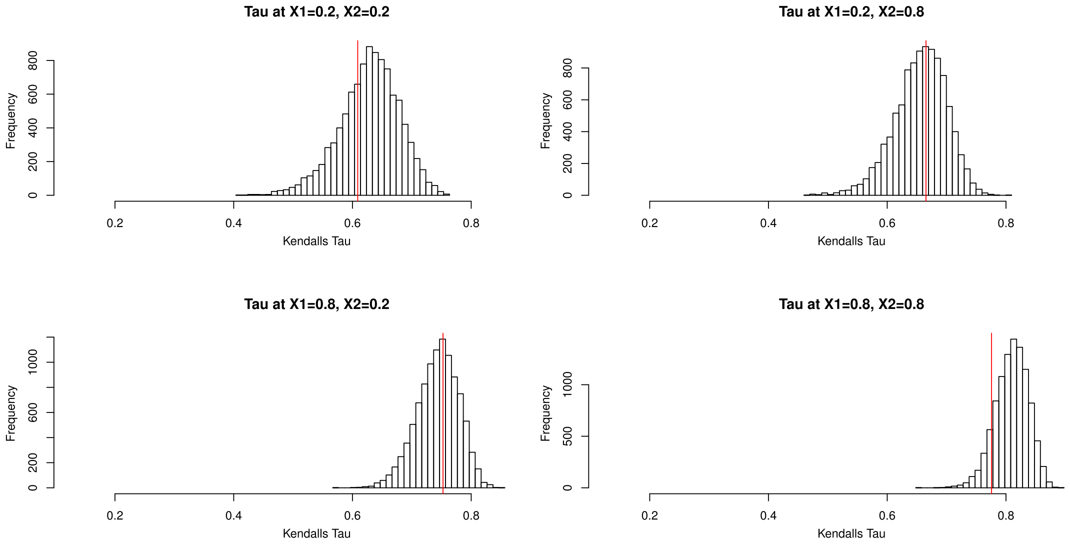

Fig. 5. **Sc1**: Histogram of predicted Kendall's τ values obtained under the true Clayton copula.

One of the inferential goals is the prediction of calibration function or, equivalently, Kendall's τ function.

As with conditional marginal means we estimate one dimensional slices at values 0.2 and 0.8 and the results, shown in Fig. 4, confirm the accuracy of the fit.

The predictive power of the model was assessed by fixing 4 covariate points and estimating the corresponding Kendall's τ values: $\hat{\tau}(0.2, 0.2)$, $\hat{\tau}(0.2, 0.8)$, $\hat{\tau}(0.8, 0.2)$, $\hat{\tau}(0.8, 0.8)$. At each MCMC iteration these predictions are calculated and histograms (Fig. 5) are constructed (red lines are true value of τ). The same estimates are presented in Fig. 6 when the Gaussian copula is used for inference. One can notice that the estimates are biased in this instance, thus emphasizing the importance of identifying the right copula family. Similar patterns have been observed when using the Frank copula.

We also show how well the algorithm estimates calibration function when covariate dimension is large. Fig. 7 shows one dimensional slices of Kendall's τ function for **Sc3** which is estimated by the Clayton GP-SIM model. Each plot is produced by varying one coordinate from 0 to 1 while fixing all other coordinates at $x = 0.5$. We observe that even in this case the estimated curves are very close to true Kendall's τ function.

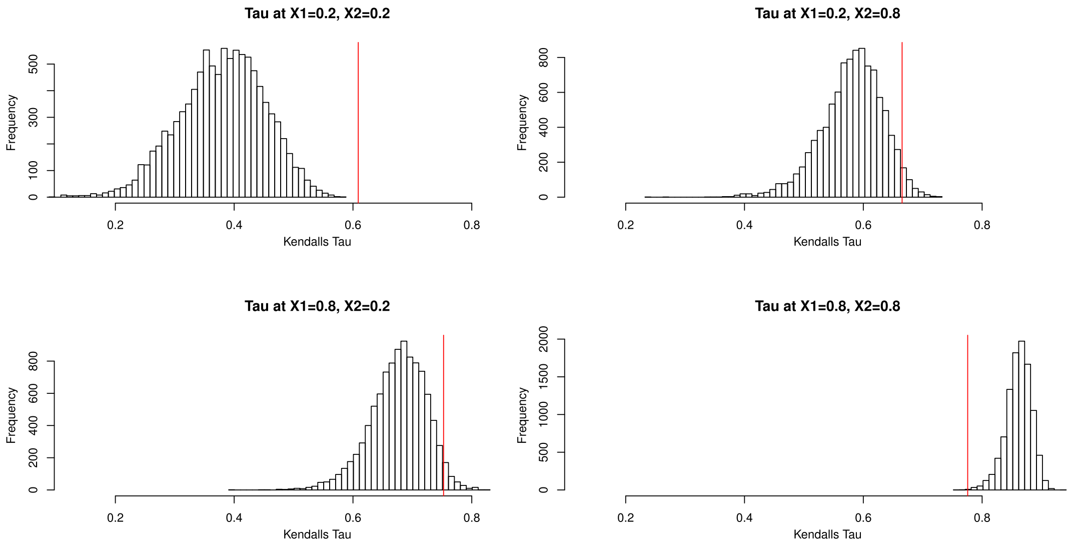

Fig. 6. Sc1: Histogram of predicted τs with Gaussian copula model.

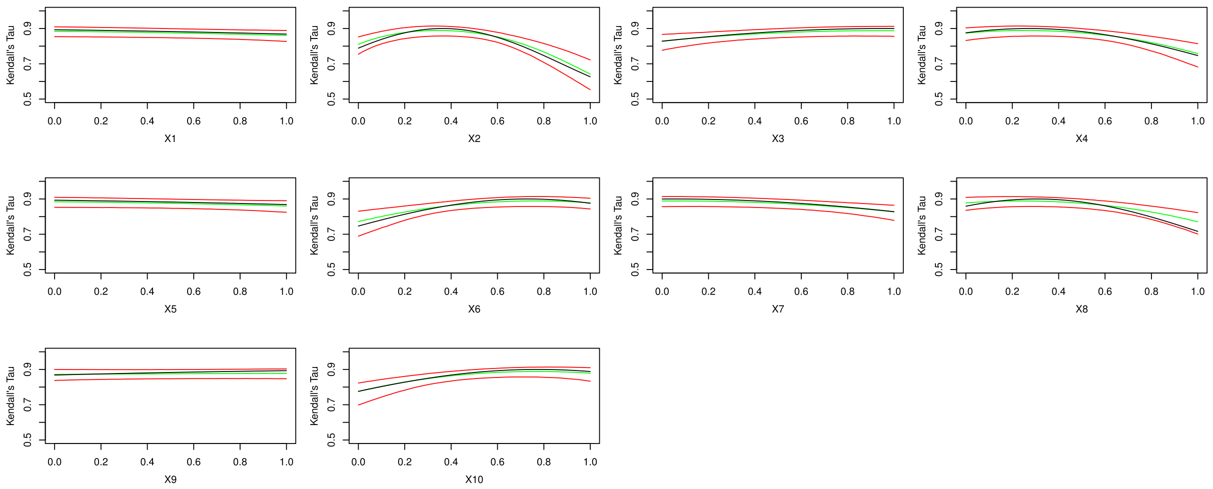

Fig. 7. Sc3: Estimation of Kendall's τ one-dimensional projections for each coordinate fixing all other coordinates at 0.5 levels. The black and green lines represent the true and estimated relationships, respectively. The red lines are the limits of the pointwise 95% credible in intervals obtained under the true Clayton family. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Model Selection

Finally, we focus on the accuracy of CVML, CCVML and WAIC in selecting the correct model. Table 2 shows the values for each scenario and model. Bold values indicate largest CVML/CCVML and smallest WAIC values for each scenario. Observe that all bold values for **Sc1**, **Sc2**, **Sc3**, **Sc5**, **Sc6**, point to the Clayton family, while for **Sc4** they indicate the Clayton family with a constant calibration. We note that the correct copula is selected even when the generative calibration model is additive.

5.2.2. Simulation results based on multiple replicates

So far, the results reported were based on a single implementation of the method. In order to facilitate interpretation, we perform 50 independent replications under each of the six scenarios described previously.

The MCMC sampler was run for 20,000 iterations for all scenarios. As before, the first half of iterations was ignored as a burn-in period. For each data set, 4 estimations were done with Clayton, Frank, Gaussian and constant Clayton copulas. For **Sc5** and **Sc6** we also fitted the Clayton copula with an additive model for the calibration function, as in Sabeti et al. (2014). The marginal distributions models have the general GP form throughout the paper. In order to produce overall measures of fit, we report the integrated squared Bias (IBias^2), Variance (IVar) and mean squared error (IMSE) of Kendall's τ evaluated at covariates $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$. The calculation requires finding points estimates for $\hat{\tau}_r(\mathbf{x}_i)$ for $1 \le r \le N_{rep}$ independently

Table 2
CVML, CCVML and WAIC values for each Scenario and Model.

	CVML	CCVML	WAIC		CVML	CCVML	WAIC
Scenario 1				Scenario 4			
Clayton	**532**	**458**	**−1065**	Clayton	322	254	−644
Frank	422	365	−844	Frank	277	209	−549
Gaussian	397	326	−801	Gaussian	276	207	−547
Clayton-Const	503	433	−1007	Clayton-Const	**323**	**255**	**−647**
Scenario 2				Scenario 5			
Clayton	**166**	**103**	**−333**	Clayton	**324**	**277**	**−650**
Frank	144	82	−289	Frank	256	216	−513
Gaussian	146	84	−293	Gaussian	260	214	−520
Clayton-Const	121	60	−243	Clayton-Const	299	257	−600
Scenario 3				Scenario 6			
Clayton	**613**	**536**	**−1237**	Clayton	**286**	**242**	**−573**
Frank	562	491	−1126	Frank	216	179	−432
Gaussian	494	417	−1002	Gaussian	205	165	−410
Clayton-Const	537	462	−1076	Clayton-Const	283	238	−567

Table 3
Estimated $\sqrt{\text{Bias}^2}$, $\sqrt{\text{IVar}}$ and $\sqrt{\text{IMSE}}$ of Kendall's τ for each Scenario and Model.

Scenario	Clayton			Frank			Gaussian			Clayton constant		
	$\sqrt{\text{IBias}^2}$	$\sqrt{\text{IVar}}$	$\sqrt{\text{IMSE}}$	$\sqrt{\text{IBias}^2}$	$\sqrt{\text{IVar}}$	$\sqrt{\text{IMSE}}$	$\sqrt{\text{IBias}^2}$	$\sqrt{\text{IVar}}$	$\sqrt{\text{IMSE}}$	$\sqrt{\text{IBias}^2}$	$\sqrt{\text{IVar}}$	$\sqrt{\text{IMSE}}$
Sc1	0.0393	0.0575	**0.0697**	0.0357	0.0657	0.0748	0.0679	0.0734	0.1	0.1046	0.0208	0.1066
Sc2	0.0492	0.0665	**0.0827**	0.0695	0.1	0.1218	0.0509	0.0692	0.0859	0.2314	0.0242	0.2327
Sc3	0.0327	0.0744	**0.0813**	0.041	0.0858	0.0951	0.0846	0.1069	0.1363	0.123	0.0134	0.1237
Sc4	0.0061	0.0355	0.036	0.0133	0.0584	0.0599	0.0205	0.0493	0.0534	0.0016	0.0258	**0.0258**
Sc5	0.0723	0.0777	**0.1061**	0.0703	0.0881	0.1127	0.0842	0.0857	0.1202	0.1589	0.024	0.1607
Sc6	0.0147	0.0384	**0.0411**	0.0175	0.05	0.0529	0.0338	0.0559	0.0654	0.0849	0.021	0.0874

replicated analyses and each $i = 1, \dots, n$. The formulas for IBias^2, IVar and IMSE are given by:

$$\begin{aligned}
\text{IBias}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\sum_{r=1}^{N_{rep}} \hat{\tau}_r(\mathbf{x}_i)}{N_{rep}} - \tau(\mathbf{x}_i)\right)^2, \\
\text{IVar} &= \frac{1}{n}\sum_{i=1}^{n} Var_r(\hat{\tau}_r(\mathbf{x}_i)), \\
\text{IMSE} &= \text{IBias}^2 + \text{IVar}.
\end{aligned} \tag{33}$$

We will apply these concepts not only for Kendall's τ but also for $E(y_1|Y_2 = y_2, \mathbf{x})$ for different combinations $(\mathbf{x}, y_2)$.

Estimation

IBias^2, IVar and IMSE for each scenario and each model are shown in Table 3 (bold values show smallest IMSE for each scenario). Note that the smallest IMSE is produced when fitting the correct model and copula family. The Clayton model with GP-SIM calibration has smallest IMSE in all scenarios with the exception of **Sc4**. We note that models with constant calibration have much smaller IVar than models with GP-SIM but have much larger IBias and, consequently, IMSE. Not surprisingly, for **Sc4**, the Clayton copula model with constant calibration yields the smallest IMSE. For each simulated data set and each model, $E(y_1|Y_2 = y_2, \mathbf{x})$ were estimated. For all scenarios except for **Sc3** we let each x_1, x_2 take values in the set $\{0.2, 0.4, 0.6, 0.8\}$ and y_2 in $\{-0.6, -0.2, 0.2, 0.6\}$ making a total of 64 combinations. For **Sc3**, y_2 takes values in $\{-0.5, 0.0, 0.5, 1.0\}$, while $\mathbf{x}$ can take 33 values scattered in $[0, 1]^{10}$, making a total of 132 combinations. The results are presented in Table 4 and largely mimic the patterns found in Table 3, thus showing that the predictive power of the model and the accuracy of dependence estimation are linked.

The results for scenarios **Sc5** and **Sc6** in which the true calibration has an additive form are shown in Table 5. Shown are the global measures of fit for Kendall's τ and $E(y_1|Y_2 = y_2, \mathbf{x})$ when the true Clayton copula is coupled with the GP-SIM and the additive model for representing the calibration function. An astute reader should not be exceedingly surprised to observe that GP-SIM outperforms the additive model under **Sc6** since the calibration function is not far from having a SIM form in this case (due to $0 \le u - u^2 \le 1/4$ for any $u \in [0, 1]$). This is not observed in **Sc5** where GP-SIM performs worse for Kendall's tau estimation than the true additive model.

Model Selection

Finally we show how well CVML, CCVML and WAIC perform in choosing the correct model. For selecting between different copula families or to check whether dependence is covariate-free we just pick the model with largest CVML/CCVML or

Table 4
Estimated $\sqrt{\text{Bias}^2}$, $\sqrt{\text{IVar}}$ and $\sqrt{\text{IMSE}}$ of $E(Y_1|y_2, \mathbf{x})$ for each Scenario and Model.

Scenario	Clayton			Frank			Gaussian			Clayton constant		
	$\sqrt{\text{IBias}^2}$	$\sqrt{\text{IVar}}$	$\sqrt{\text{IMSE}}$	$\sqrt{\text{IBias}^2}$	$\sqrt{\text{IVar}}$	$\sqrt{\text{IMSE}}$	$\sqrt{\text{IBias}^2}$	$\sqrt{\text{IVar}}$	$\sqrt{\text{IMSE}}$	$\sqrt{\text{IBias}^2}$	$\sqrt{\text{IVar}}$	$\sqrt{\text{IMSE}}$
Sc1	0.0231	0.0531	**0.0579**	0.1264	0.0322	0.1304	0.1434	0.0557	0.1539	0.0416	0.0579	0.0713
Sc2	0.0293	0.0464	**0.0549**	0.0802	0.0475	0.0932	0.1098	0.0593	0.1247	0.1213	0.0407	0.128
Sc3	0.0364	0.0707	0.0795	0.214	0.0363	0.217	0.1042	0.0708	0.1259	0.0483	0.0572	**0.0749**
Sc4	0.0174	0.042	0.0454	0.1023	0.0325	0.1074	0.1379	0.0449	0.145	0.0179	0.041	**0.0447**
Sc5	0.0144	0.0413	**0.0437**	0.0909	0.0347	0.0973	0.14	0.051	0.149	0.0355	0.04	0.0534
Sc6	0.0202	0.0456	**0.0498**	0.1046	0.0298	0.1087	0.1367	0.0448	0.1439	0.0237	0.0442	0.0501

Table 5
Estimated $\sqrt{\text{Bias}^2}$, $\sqrt{\text{IVar}}$ and $\sqrt{\text{IMSE}}$ of Kendall's τ and $E(Y_1|y_2, \mathbf{x})$ for GP-SIM and Additive models.

Scenario	Clayton GP-SIM			Clayton Additive		
	$\sqrt{\text{IBias}^2}$	$\sqrt{\text{IVar}}$	$\sqrt{\text{IMSE}}$	$\sqrt{\text{IBias}^2}$	$\sqrt{\text{IVar}}$	$\sqrt{\text{IMSE}}$
	Kendall's Tau					
Sc5	0.0723	0.0777	0.1061	0.0573	0.0516	**0.0771**
Sc6	0.0147	0.0384	**0.0411**	0.0063	0.0458	0.0462
	$E(Y_1|y_2, \mathbf{x})$					
Sc5	0.0144	0.0413	**0.0437**	0.0207	0.0428	0.0475
Sc6	0.0202	0.0456	**0.0498**	0.0236	0.0483	0.0538

Table 6
The percentage of correct decisions for each selection criterion when comparing the correct Clayton model with a non-constant calibration with all the other models: Frank model with non-constant calibration, Gaussian model with non-constant calibration, Clayton model with constant calibration.

Scenario	Frank			Gaussian			Clayton constant		
	CVML	CCVML	WAIC	CVML	CCVML	WAIC	CVML	CCVML	WAIC
Sc1	100%	100%	100%	100%	100%	100%	94%	94%	94%
Sc2	100%	100%	100%	100%	100%	100%	100%	100%	100%
Sc3	98%	96%	98%	100%	100%	100%	100%	98%	100%
Sc5	100%	100%	100%	100%	100%	100%	100%	100%	100%
Sc6	100%	100%	100%	100%	100%	100%	98%	100%	98%

Table 7
The percentage of correct decisions for each selection criterion when comparing the correct Clayton model with a constant calibration with three models: Clayton, Frank and Gaussian, all of them assuming a GP-SIM calibration.

Scenario	Clayton			Frank			Gaussian		
	CVML	CCVML	WAIC	CVML	CCVML	WAIC	CVML	CCVML	WAIC
Sc4	58%	62%	58%	100%	100%	100%	100%	100%	100%

Table 8
The percentage of correct decisions for each selection criterion when comparing the correct additive model with GP-SIM with non-constant calibration.

Scenario	Clayton GP-SIM		
	CVML	CCVML	WAIC
Sc5	92%	94%	90%
Sc6	30%	34%	28%

smallest WAIC. Table 6 shows how often Clayton model is selected over other models using CVML, CCVML and WAIC for **Sc1**, **Sc2**, **Sc3**, **Sc5** and **Sc6**. Similarly, Table 7 shows how often Clayton-constant is selected over other models for **Sc4**.

We can conclude that all selection measures perform quite similarly across scenarios. Also, the numerical study shows that the choice of a copula family is considerably more accurate than correctly determining that the calibration function is constant. The latter difficulty has been reported elsewhere (e.g., Craiu and Sabeti, 2012). In part, this is due to the fact that the models are flexible enough to capture the constant calibration and produce estimates that mislead a cross-validation-based method. In Section 6.2 we return to this problem and develop a new permutation-based procedure that exhibits a drastically improved performance in numerical experiments. Since **Sc5** and **Sc6** were simulated with Clayton additive calibration, we show how often Clayton Additive model is selected over Clayton GP-SIM using different criteria (Table 8). The poor performance for **Sc6** is not that surprising since the additive calibration in this scenario has almost SIM form.

Table 9
Red Wine data: CVML, CCVML and WAIC criteria values different models.

	Clayton	Frank	Gaussian	Gumbel	T-3
CVML	−1858	−1816	**−1788**	−1829	−1810
CCVML	−582	−547	**−522**	−558	−534
WAIC	3713	3634	**3572**	3656	3621

Table 10
Wine data: Posterior means and quantiles of β.

Variable	Posterior mean	95% credible interval
X_{va}	0.274	[0.154, 0.389]
X_{ca}	−0.336	[−0.413, −0.254]
X_{rs}	−0.076	[−0.278, 0.271]
X_{ch}	0.060	[−0.246, 0.259]
X_{fs}	0.276	[0.106, 0.410]
X_{ts}	0.402	[0.248, 0.608]
X_{ph}	0.155	[0.054, 0.286]
X_{su}	0.501	[0.342, 0.601]
X_{al}	0.463	[0.382, 0.517]

Table 11
Wine data: CVML, CCVML and WAIC criteria values for variable selection in conditional copula.

Variables	CVML	CCVML	WAIC
ALL	−1788	−522	3572
$X_{va}, X_{ca}, X_{fs}, X_{ts}, X_{ph}, X_{su}, X_{al}$	−1805	−532	3608
X_{va}	−1823	−552	3646
X_{ca}	−1815	−541	3629
X_{rs}	−1849	−582	3698
X_{ch}	−1842	−578	3688
X_{fs}	−1852	−584	3705
X_{ts}	−1851	−583	3700
X_{ph}	−1816	−557	3633
X_{su}	−1841	−571	3682
X_{al}	−1847	−577	3697
Constant	−1849	−584	3700

5.3. *Red wine data*

We consider the data of Cortez et al. (2009) consisting of various physicochemical tests of 1599 red variants of the Portuguese "Vinho Verde" wine. Acidity and density are properties closely associated with the quality of wine and grape, respectively. Of interest here is to study the dependence pattern between 'fixed acidity' (Y_{fa}) and 'density' (Y_{de}) and how it changes with values of other variables: 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'pH', 'sulphates' and 'alcohol', denoted by $X_{va}, X_{ca}, X_{rs}, X_{ch}, X_{fs}, X_{ts}, X_{ph}, X_{su}, X_{al}$, respectively. Response variables are linearly transformed to have mean 0 and standard deviation of 1, similarly covariates were transformed to be between 0 and 1.

To select the appropriate copula family, we fit GP-SIM with 'Clayton', 'Frank', 'Gaussian', 'Gumbel' and 'T-3' (Student T with 3 degrees of freedom) dependencies. For each model the MCMC was run for 20,000 iterations with 10,000 burn-in period. We used 30 inducing inputs for the marginals and we use calibration function estimation ($m_1 = m_2 = m = 30$). The resulting CVML, CCVML and WAIC values are shown in Table 9.

All model selection measures indicate that among candidate copula families the most suitable one is the Gaussian one. The GP-SIM coefficients (β) fitted under the Gaussian copula family are shown in Table 10.

The credible intervals suggest that not all covariates may be needed to model dependence between responses. For example, 'residual sugars' and 'chlorides' seem to not affect the calibration function so we consider a model in which they are omitted from the conditional copula model. In all models, we include all the covariates in the marginal distributions. For comparison, we have also fitted all Gaussian GP-SIM models with only one covariate, and with no covariates at all (constant). The computational algorithm to fit GP-SIM when the conditional copula depends on only one variable is very similar to the one described above. The main difference is that there is no β variable and the inducing inputs (for calibration function) are evenly spread on [0, 1]. The testing results are shown in Table 11.

Based on the selection criteria results we conclude that all nine covariates are required to explain the dependence structure of two responses. Fig. 8 shows 1-dimensional plots of Kendall's τ calibration curve with 95% credible as a function of covariates. The plots are constructed by varying one predictor while fixing all others at their mid-range values.

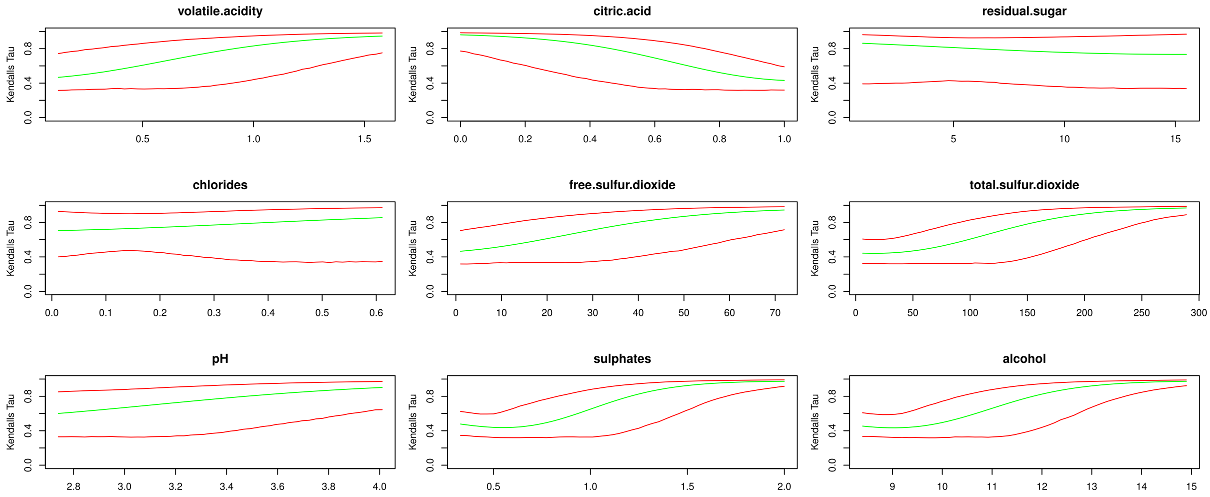

Fig. 8. Wine Data: Slices of predicted Kendall's τ as function of covariates. Red curves represent 95% credible intervals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

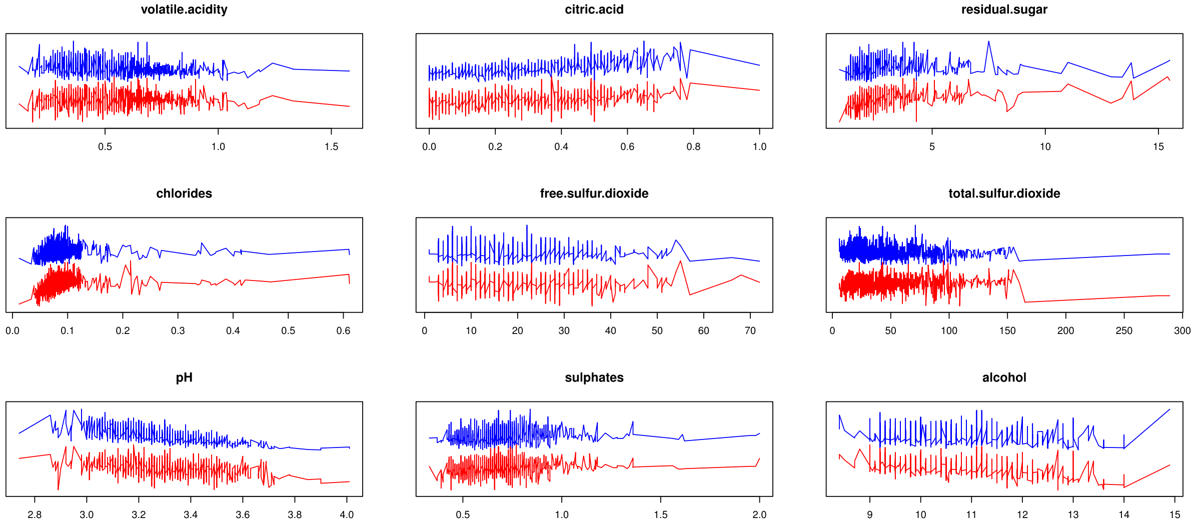

Fig. 9. Wine Data: Plots of 'fixed acidity' (blue) and 'density' (red) (linearly transformed to fit on one plot) against covariates. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The plots clearly demonstrate that when covariates are fixed at their mid-range values, the conditional correlation between 'fixed acidity' and 'density' increases with 'volatile acidity', 'free sulfur dioxide', 'total sulfur dioxide', 'pH', 'sulphates' and 'alcohol', and decreases with levels of 'citric acid'. These relationships can influence the preparation method of the wine.

In order to demonstrate the difficulty one would have in gauging the complex evolution of dependence between two responses as a function of covariates we plot in Fig. 9 the response variables together as they vary with each covariate. It is clear that the model manages to identify a pattern that would be very difficult to distinguish without the help of a flexible mathematical model.

6. Simplifying assumption

6.1. Model misspecification and the simplifying assumption

Understanding whether the data support the SA or not is usually important for the subject matter analysis, since a dependence structure that does not depend on the covariates can be of scientific interest. The SA has also a serious impact on the statistical analysis, because it has the potential to simplify greatly the estimation of the copula. There is however, an interesting connection between model misspecification and SA which, as far as we know, has not been reported elsewhere.

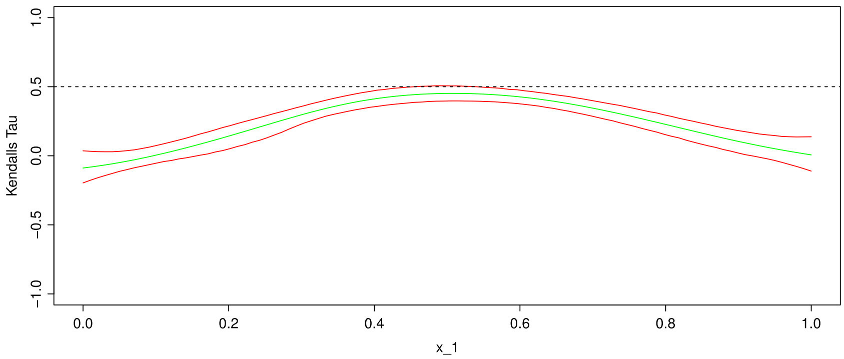

Fig. 10. Estimation of Kendall's τ as a function of x_1 when only first covariate is used in estimation. The dotted black and solid green lines represent the true and estimated relationships, respectively. The red lines are the limits of the pointwise 95% credible in intervals obtained under the true Clayton family. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 12
Missed covariate: CVML, CCVML and WAIC criteria values for model with conditional copula depends on one covariate and when it is constant.

Variables	CVML	CCVML	WAIC
X_1	−508	−174	1017
Constant	−570	−232	1140

To illustrate the point, consider a random sampling design setting with two independent random variables, X_1, X_2 serving as covariates in the Clayton copula model in which SA is satisfied, the sample size $n = 1500$ and

$$\begin{aligned}
&f_1(x) = 0.6\sin(5x_1 + x_2),\\
&f_2(x) = 0.6\sin(x_1 + 5x_2),\\
&\tau(x) = 0.5,\\
&\sigma_1 = \sigma_2 = 0.2.
\end{aligned}$$

When we fit a GP-SIM model with the correct Clayton copula family, but with the X_2 covariate omitted from both marginal and copula models, the estimated Kendall's $\tau(x_1)$ exhibits a clear non-constant shape, as seen in Fig. 10. The CVML, CCVML and WAIC criteria, whose values are shown in Table 12, unanimously vote for a nonconstant calibration function.

While one may expect a non-constant pattern when the two covariates are dependent, this residual effect of X_1 on the copula may be surprising when X_1 and X_2 are independent. We can gain some understanding by considering a simplified example in which $Y_i|X_1, X_2 \sim N(f_i(X_1, X_2), 1)$ for $i = 1, 2$, and $\mathrm{Cov}(Y_1, Y_2|X_1, X_2) = \mathrm{Corr}(Y_1, Y_2|X_1, X_2) = \rho$, hence constant in X_1 and X_2. Hence, for marginal models that include only X_1, yielding residuals $W_i = Y_i - E[Y_i|X_1]$ for $i = 1, 2$, we are interested in explaining the non-constant dependence between $\mathrm{Cov}(W_1, W_2|X_1)$ and X_1. Standard statistical properties of covariance and conditional expectation are used to obtain

$$\mathrm{Cov}(W_1, W_2|X_1) = \mathrm{Cov}(Y_1, Y_2|X_1), \tag{34}$$

and

$$\begin{aligned}
\mathrm{Cov}(Y_1, Y_2|X_1) &= E[\mathrm{Cov}(Y_1, Y_2|X_1, X_2)] + \mathrm{Cov}(E[Y_1|X_1, X_2], E[Y_2|X_1, X_2])\\
&= \rho + \mathrm{Cov}(f_1(X_1, X_2), f_2(X_1, X_2)),
\end{aligned} \tag{35}$$

where the covariance in (35) is with respect to the marginal distribution of X_2. Hence it is apparent that the conditional covariance $\mathrm{Cov}(W_1, W_2|X_1)$ will generally not be constant in X_1. Note that if the true means have additive form, i.e. $f_i(X_1, X_2) = \bar{f}_i(X_1) + \tilde{f}_i(X_2)$, for $i = 1, 2$, then the covariances in (34) are indeed constant in X_1, but the estimated value of $\mathrm{Cov}(Y_1, Y_2|X_1)$ will be biased. Although here we focused on the covariance as a measure of dependence, the argument is extendable to copula parameters or Kendall's tau, but the calculations are more involved.

In conclusion, violation of the SA may be due to the omission of important covariates from the model. This phenomenon along with the knowledge that in general it is difficult to measure all the variables with potential effect on the dependence pattern, suggests that a non-constant copula is a prudent choice.

6.2. A permutation-based criterion to detect data support for the simplified assumption

In this section we modify the CVML and the conditional CCVML method to identify data support for SA after the copula family is selected.

As was shown in previous sections, the selection criteria included in the paper do not perform well in recognizing that the true calibration is constant. This is in line with Craiu and Sabeti (2012) who also noted that the traditional Bayesian model selection criteria, e.g. the Deviance information criterion (DIC) of Spiegelhalter et al. (2002), tend to prefer the more complex calibration model over a simple model with constant calibration even when the latter is actually correct. To illustrate this point with larger sample sizes, we have simulated 50 replicates of sample sizes 1500 using Clayton copula from **Sc1**, **Sc4** and **Sc5**. Each sample is fitted with the general model introduced here and a constant Clayton copula, while marginals are estimated using a general GP. Table 13 shows the proportion of correct decisions for the three scenarios and various selection criteria. Even for a large sample size, the proportion of right decisions for **Sc4**, i.e. when SA holds, is quite low. One of the explanations is that the general model does a good job at capturing the constant trend of the calibration function and yields predictions that are not too far from the ones produced with the simpler (and correct) model. The modified CVML we propose is inspired by two desiderata: (i) to separate the set of observations used for prediction from the set of observations used for fitting the model, and (ii) to amplify the impact of the copula-induced errors in the CCVML calculation. The former will reduce the implicit bias one gets when the same data is used for estimation and testing, while the latter is expected to increase the power to identify SA.

For (i) we randomly partition the data into a training set $\mathcal{D} = \{y_{1i}, y_{2i}, x_i\}_{i=1,\ldots,n_1}$ and a test set $\mathcal{D}^* = \{y^*_{1i}, y^*_{2i}, x^*_i\}_{i=1,\ldots,n_2}$. In our numerical experiments we have kept two thirds of observations in the training set. In order to achieve (ii) we note that permuting the response indexes will not affect the copula term if SA is indeed satisfied and will perturb the prediction when SA is not satisfied. However, one must cautiously implement this idea, since the permutation $\lambda : \{1, \ldots, n_2\} \to \{1, \ldots, n_2\}$ will affect the marginal model fit, regardless of the SA status, as $y_{j\lambda(i)}$ will be paired with x_i, for all $j = 1, 2$. Below we describe the permutation-based CVML criterion that combines (i) and (ii).

Assume that the fitted GP-SIM model yields posterior samples from the conditional distribution of latent variables and parameters $\omega^{(t)} \sim \pi(\omega|\mathcal{D})$, $t = 1 \ldots M$. Then we define the observed data criterion as the predictive log probability of the test cases which can be easily estimated from posterior samples, as follows:

$$
\begin{aligned}
\mathrm{CVML}_{obs} &= \sum_{i=1}^{n_2} \log P(y^*_{1i}, y^*_{2i}|\mathcal{D}, x^*_i) \approx \sum_{i=1}^{n_2} \log \left\{ \frac{1}{M} \sum_{t=1}^{M} P(y^*_{1i}, y^*_{2i}|w^{(t)}, x^*_i) \right\} = \\
&= \sum_{i=1}^{n_2} \log \left\{ \frac{1}{M} \sum_{t=1}^{M} \frac{1}{\sigma_1^{(t)}} \phi\left(\frac{y^*_{1i} - f^{*(t)}_{1i}}{\sigma_1^{(t)}} \right) \frac{1}{\sigma_2^{(t)}} \phi\left(\frac{y^*_{2i} - f^{*(t)}_{2i}}{\sigma_2^{(t)}} \right) \times \right. \\
&\left. \times \; c_{\theta_i^{*(t)}} \left[\Phi\left(\frac{y^*_{1i} - f^{*(t)}_{1i}}{\sigma_1^{(t)}} \right), \Phi\left(\frac{y^*_{2i} - f^{*(t)}_{2i}}{\sigma_2^{(t)}} \right) \right] \right\},
\end{aligned}
$$

where $f^{*(t)}_{1i}, f^{*(t)}_{2i}, \theta^{*(t)}_i$ are the predicted values for the test cases produced by the GP-SIM model. Consider J permutations of $\{1 \ldots n_2\}$ which we denote as $\lambda_1, \ldots, \lambda_J$, and compute J permuted CVMLs as:

$$
\begin{aligned}
\mathrm{CVML}_j &= \sum_{i=1}^{n_2} \log \left\{ \frac{1}{M} \sum_{t=1}^{M} \frac{1}{\sigma_1^{(t)}} \phi\left(\frac{y^*_{1i} - f^{*(t)}_{1i}}{\sigma_1^{(t)}} \right) \frac{1}{\sigma_2^{(t)}} \phi\left(\frac{y^*_{2i} - f^{*(t)}_{2i}}{\sigma_2^{(t)}} \right) \times \right. \\
&\left. \times \; c_{\theta^{*(t)}_{\lambda_j(i)}} \left[\Phi\left(\frac{y^*_{1i} - f^{*(t)}_{1i}}{\sigma_1^{(t)}} \right), \Phi\left(\frac{y^*_{2i} - f^{*(t)}_{2i}}{\sigma_2^{(t)}} \right) \right] \right\}.
\end{aligned} \tag{36}
$$

Note that CVML_{obs} differs from CVML_j only in the values of the copula parameters. While for the former we use $\theta(x^*_i)$, in the latter we use $\theta(x^*_{\lambda_j(i)})$ for the dependence between y^*_{1i} and y^*_{2i}. If calibration is constant then CVML_{obs} and CVML_j should be similar, hence we define the evidence

$$
\mathrm{EV} = 2 \times \min \left\{ \frac{\sum_{j=1}^{J} \mathbb{1}_{\{CVML_{obs} < CVML_j\}}}{J}, \frac{\sum_{j=1}^{J} \mathbb{1}_{\{CVML_{obs} > CVML_j\}}}{J} \right\}. \tag{37}
$$

Under the null model with constant calibration with known marginals and if we assume that CVML_{obs} and $\{\mathrm{CVML}_j : 1 \le j \le J\}$ are iid for each j, then each term inside the min function in (37) has a Uniform(0, 1) limiting distribution when $J \to \infty$. In that case it follows that $P(EV < 0.05) = 0.05$. In practice, the ideal situation just described is merely an approximation since the $\{\mathrm{CVML}_j : 1 \le j \le J\}$ are not independent and we compute EV using a fixed number of permutations. Nevertheless, the ideal setup can be used to build our decision that when $EV > 0.05$ the data support SA, and otherwise they do not.

A similar rule can be build using the CCVML criterion. For instance, its value for test data is

$$
\mathrm{CCVML}_{obs} = \frac{1}{2} \sum_{i=1}^{n_2} \log P(y^*_{1i}|\mathcal{D}, x^*_i, y^*_{2i}) + \frac{1}{2} \sum_{i=1}^{n_2} \log P(y^*_{2i}|\mathcal{D}, x^*_i, y^*_{1i}). \tag{38}
$$

The permutation-based version of (38) can be obtained using the same principle as in (36) thus leading to the counterpart of (37) for CCVML.

Table 13
The percentage of correct decisions for each selection criterion and scenarios. GP-SIM and SA were fitted with Clayton copula, sample size is 1500.

Scenario	CVML	CCVML	WAIC
Sc1	100%	100%	100%
Sc4	74%	78%	74%
Sc5	100%	100%	100%

Table 14
The percentage of correct decisions for each selection criterion and scenario. Predicted CVML and CCVML values based on $n_1 = 1000$ training and $n_2 = 500$ test data, respectively. The calculation of EV is based on a random sample of 500 permutations.

Scenario	CVML	CCVML
Sc1	98%	96%
Sc4	92%	90%
Sc5	100%	100%

Table 14 shows the proportion of correct decisions using proposed methods with 1000 and 500 samples in training and test set respectively, and $J = 500$ permutations. The results, especially those for **Sc4**, clearly show an important improvement in the rate of making the correct selection, with only a slight decrease in the power to detect non-constant calibrations. We can also notice that CVML and CCVML performed similarly.

7. Conclusion and future work

The inclusion of a dynamic copula in the model comes with a significant computational price. The inclusion can be justified by the need for an exploration of dependence, or because it can improve the predictive accuracy of the model.

We have proposed a Bayesian procedure to estimate the calibration function of a conditional copula model jointly with the marginal distributions. In our attempt to move away from an additive model hypothesis we consider a sparse Gaussian process prior used in conjunction with a single index model. The resulting procedure reduces the dimensionality of the parameter space and can be used for moderate number of covariates.

The simplifying assumption is often adopted as a way to bypass the need for estimating a conditional copula model. However, even if the SA is true when conditioning on the true set of covariates, we showed that if one or more covariates are not included in the fitted model, then the SA is violated. We have introduced a couple of selection criteria to help select the copula family from a set of candidates and to gauge data support in favor of the simplifying assumption. While the former task seems to be achieved by all criteria considered, the latter is a particularly difficult problem and we are excited about the good performance exhibited by our permutation-based version of the cross-validated marginal likelihood criterion. Its theoretical properties are the focus of our ongoing work and we plan to extend its use to identifying the set of covariates that do not influence the calibration function.

Acknowledgments

We thank the Co-Editor, Ana Colubi, the Associate Editor and two anonymous referees for their comments and suggestions that have greatly improved the paper. We also gratefully acknowledge Keith Knight's and Stanislav Volgushev's helpful comments on an earlier draft of the paper. Funding support of this work was provided by grant 249547-2012 from the Natural Sciences and Engineering Research Council of Canada and the Collaborative Research Team grant on Copula Dependence Modeling from the Canadian Statistical Sciences Institute.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2018.01.013.

References

Aas, K., Czado, C., Frigessi, A., Bakken, H., 2009. Pair-copula constructions of multiple dependence. Insurance Math. Econom. 44, 182–198. http://dx.doi.org/10.1016/j.insmatheco.2007.02.001.

Acar, E.F., Craiu, R.V., Yao, F., 2011. Dependence calibration in conditional copulas: A nonparametric approach. Biometrics 67, 445–453.

Acar, E.F., Craiu, R.V., Yao, F., et al., 2013. Statistical testing of covariate effects in conditional copula models. Electron. J. Stat. 7, 2822–2850.

Acar, E., Genest, C., Nešlehová, J., 2012. Beyond simplified pair-copula constructions. J. Multivariate Anal. 110, 74–90.

Andrieu, C., De Freitas, N., Doucet, A., Jordan, M.I., 2003. An introduction to MCMC for machine learning. Mach. Learn. 50, 5–43.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer-Verlag New York Inc.

Chavez-Demoulin, V., Vatter, T., 2015. Generalized additive models for conditional copulas. J. Multivariate Anal. 141, 147–167.

Choi, T., Shi, J.Q., Wang, B., 2011. A Gaussian process regression approach to a single-index model. J. Nonparametr. Stat. 23, 21–36.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J., 2009. Modeling wine preferences by data mining from physicochemical properties. Decis. Support Syst. 47, 547–553.

Craiu, R.V., Rosenthal, J.S., 2014. Bayesian computation via Markov chain Monte Carlo. Ann. Rev. Stat. Appl. 1, 179–201.

Craiu, R.V., Sabeti, A., 2012. In mixed company: Bayesian inference for bivariate conditional copula models with discrete and continuous outcomes. J. Multivariate Anal. 110, 106–120.

Czado, C., 2010. Pair-copula constructions of multivariate copulas. In: Copula Theory and its Applications. pp. 93–109.

Dalla Valle, L., Leisen, F., Rossini, L., 2017. Bayesian non-parametric conditional copula estimation of twin data. J. R. Stat. Soc. Ser. C. Appl. Stat.

Derumigny, A., Fermanian, J.D., 2017. About tests of the "simplifying" assumption for conditional copulas. Dependence Modeling 5, 154–197.

Fermanian, J.D., Lopez, O., 2015. Single-index copulae. preprint: arXiv:1512.07621.

Geisser, S., Eddy, W.F., 1979. A predictive approach to model selection. J. Amer. Statist. Assoc. 74, 153–160.

Gelfand, A.E., Dey, D.K., Chang, H., 1992. Model determination using predictive distributions with implementation vias sampling-based methods. In: Bernardo, J.M., Berger, J.O., Smith, A.D. (Eds.), Bayesian statistics, vol. 4. pp. 147–167.

Gelman, A., Hwang, J., Vehtari, A., 2014. Understanding predictive information criteria for Bayesian models. Stat. Comput. 24, 997–1016.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Statist. Sci. 7, 457–472.

Genest, C., Ghoudi, K., Rivest, L.P., 1995. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika 82, 543–552.

Gijbels, I., Omelka, M., Veraverbeke, N., 2015. Estimation of a copula when a covariate affects only marginal distributions. Scand. J. Stat. 42, 1109–1126.

Gramacy, R.B., Lian, H., 2012. Gaussian process single-index models as emulators for computer experiments. Technometrics 54, 30–41.

Hanson, T., Branscum, A., Johnson, W., 2011. Predictive comparison of joint longitudinal-survival modelling: a case study illustrating competing approaches. Lifetime Data Anal. 17, 2–28.

Hernández-Lobato, J.M., Lloyd, J.R., Hernández-Lobato, D., 2013. Gaussian process conditional copulas with applications to financial time series. In: Advances in Neural Information Processing Systems. pp. 1736–1744.

Hu, Y., Gramacy, R.B., Lian, H., 2013. Bayesian quantile regression for single-index models. Stat. Comput. 23, 437–454.

Klein, N., Kneiß, T., 2015. Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. Stat. Comput. 1–20.

Lambert, P., Vandenhende, F., 2002. A copula-based model for multivariate non-normal longitudinal data: analysis of a dose titration safety study on a new antidepressant. Statist. Med. 21, 3197–3217.

Lopez-Paz, D., Hernández-Lobato, Z., Ghahramani, J.M., 2013. Gaussian process vine copulas for multivariate dependence. In: Proceedings of the 30th International Conference on Machine Learning. JMLR: W&CP, Atlanta, Georgia, USA, pp. 10–18.

Murray, I., Adams, R. MacKay, D., 2010. Elliptical slice sampling. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. pp. 541–548.

Naish-Guzman, A., Holden, S., 2007. The generalized FITC approximation. In: Advances in Neural Information Processing Systems. pp. 1057–1064.

Ning, S., Shephard, N., 2017. A nonparametric Bayesian approach to copula estimation. arXiv preprint arXiv:1702.07089.

Nishihara, R., Murray, I., Adams, R.P., 2014. Parallel MCMC with generalized elliptical slice sampling. J. Mach. Learn. Res. 15, 2087–2112.

Patton, A.J., 2006. Modelling asymmetric exchange rate dependence*. Internat. Econom. Rev. 47, 527–556.

Quiñonero Candela, J., Rasmussen, C.E., 2005. A unifying view of sparse approximate Gaussian process regression. J. Mach. Learn. Res. 6, 1939–1959.

Rosenthal, J.S., 2009. Markov chain Monte Carlo algorithms: Theory and practice. In: Monte Carlo and Quasi-Monte Carlo Methods 2008. Springer, pp. 157–169.

Sabeti, A., Wei, M., Craiu, R.V., 2014. Additive models for conditional copulas. Stat 3, 300–312.

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

Snelson, E., Ghahramani, Z., 2005. Sparse Gaussian processes using pseudo-inputs. In: Advances in neural information processing systems. pp. 1257–1264.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of model complexity and fit (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 64, 583–639 (57).

Vehtari, A., Gelman, A., Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. 27, 1413–1432.

Veraverbeke, N., Omelka, M., Gijbels, I., 2011. Estimation of a conditional copula and association measures. Scand. J. Stat. 38, 766–780.

Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. 11, 3571–3594.

Watanabe, S., 2013. A widely applicable bayesian information criterion. J. Mach. Learn. Res. 14, 867–897.

Wu, J., Wang, X., Walker, S.G., 2014. Bayesian nonparametric inference for a multivariate copula function. Methodol. Comput. Appl. Probab. 16, 747–763.

Wu, J., Wang, X., Walker, S.G., 2015. Bayesian nonparametric estimation of a copula. J. Stat. Comput. Simul. 85, 103–116.