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 $1 \leq i \leq p$, 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 $1 \leq i \leq p$. 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/

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