Research Article

Francesco Bonacina, Olivier Lopez*, and Maud Thomas

Tree-based conditional copula estimation

https://doi.org/10.1515/demo-2024-0010
received April 12, 2024; accepted December 13, 2024

Abstract: This article proposes a regression tree procedure to estimate conditional copulas. The associated algorithm determines classes of observations based on covariate values and fits a simple parametric copula model on each class. The association parameter changes from one class to another, allowing for non-linearity in the dependence structure modeling. It also allows the definition of classes of observations on which the so-called "simplifying assumption" holds reasonably well. When considering observations belonging to a given class separately, the association parameter no longer depends on the covariates according to our model. In this article, we derive asymptotic consistency results for the regression tree procedure and show that the proposed pruning methodology, i.e., the model selection techniques selecting the appropriate number of classes, is optimal in some sense. Simulations provide finite sample results, and an analysis of data of cases of human influenza presents the practical behavior of the procedure.

Keywords: conditional copula, regression trees, asymptotic theory

MSC 2020: 62G08, 62G20, 62H30, 62P10

1 Introduction

Since Sklar's seminal result, copula theory has emerged as a practical means of describing the dependence between random variables. Allowing one to distinguish between the marginal behavior of each component of a random vector and the dependence structure (represented by a copula function), Sklar's theorem opens the way to flexible modeling of various forms of dependence [38]. In this article, we propose a new method to perform conditional copula analysis based on regression trees and to derive consistency results for this procedure.

Various estimation procedures and analyses of copulas have been studied in the statistical literature (e.g., [3,8,21,40,43]). In the presence of covariates, conditional copula analysis consists in fitting a copula function to the conditional distribution of a random vector. From an application point of view, Dupuis and Jones [12] have shown their importance in modeling certain natural disasters such as hurricanes, or the dependence between different expense lines in actuarial problems. Lopez [36] and Farkas and Lopez [16] have used this type of model for insurance claim management. Another important application, e.g., in finance, can be found in Jaworski et al. [30]. More generally, the study of conditional copulas also appears particularly important in Vine copulas [9]. Abegaz et al. [1] and Gijbels et al. [23,24] have studied both semi-parametric and non-parametric procedures for performing this analysis. Finally, Fermanian and Lopez [18] have examined the case of high-dimensional covariates and relied on a dimension reduction approach to perform the analysis.

*** Corresponding author: Olivier Lopez,** CREST Laboratory, CNRS, Groupe des Écoles Nationales d'Économie et Statistique, Ecole Polytechnique, Institut Polytechnique de Paris, 5 avenue Henry Le Chatelier 91120 Palaiseau, France, e-mail: olivier.lopez@ensae.fr
Francesco Bonacina: Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4 place Jussieu, F-75005 Paris, France; Sorbonne Université, INSERM, Institut Pierre Louis d'Epidémiologie et de Santé Publique, F75012 Paris, France, e-mail: francesco.bonacina@sorbonne-universite.fr
Maud Thomas: Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4 place Jussieu, F-75005 Paris, France, e-mail: maud.thomas@sorbonne-universite.fr

We propose here to use regression trees to perform this conditional copula analysis. Regression trees, along with the *Classification and Regression Tree* (CART) algorithm, were originally introduced by Li et al. [6] algorithm and are now classic tools used for several applications (e.g., see [15,17,26,35]). Apart from the computational efficiency of the CART estimation algorithm used to fit the model, an interesting feature of this approach is the ability to construct classes of individuals (based on their characteristics) with similar behavior. In the context of copula analysis, this corresponds to classes of individuals with the same copula (i.e., dependence) structure. This model can be seen as a means to easily generalize the "simplifying" assumption considered by many authors (see, e.g., [10,11,32]). According to this hypothesis, only the marginal distributions of each component depend on the covariates, while the dependence structure does not vary with them. In contrast, in our model, the copulas are different for each cluster determined by the regression tree, and thus, the simplifying assumption holds separately for each cluster. Our approach is semiparametric, in the sense that the copula fitted to the different leaves of the tree all belong to a parametric copula family (but with a different association parameter depending on the covariates).

The rest of this article is organized as follows. In Section 2, we describe the general framework of regression trees and the algorithm used to fit them to data. Section 3 is devoted to proving theoretical results on the consistency of this procedure. Particular attention is paid to the part of model selection, known as the "pruning step," which consists of selecting an appropriate sub-tree from the maximal tree obtained by iterative partitioning of the data set. In Section 4, the practical behavior is investigated through a simulation study and real data analysis. The proofs of the theoretical results are gathered in the Appendix.

2 Regression trees for conditional copula analysis

2.1 Model and notations

We consider a set of observations $(\mathbf{Y}_i, \mathbf{X}_i)_{1\le i\le n}$ consisting of independent identically distributed copies of the random vector $(\mathbf{Y}, \mathbf{X})$, where $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$ are the covariates, and $\mathbf{Y} = (Y^{(1)}, \ldots, Y^{(k)}) \in \mathbb{R}^k$ is a random vector of response variables $Y^{(j)}$, $j = 1,\ldots, k$. The marginal conditional cumulative distribution functions (c.d.f.) of the random vector $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ are defined as

$$F^{(j)}(t^{(j)}|\mathbf{x}) = \mathbb{P}(Y^{(j)} \le t^{(j)}|\mathbf{X} = \mathbf{x}), \quad t^{(j)} \in \mathbb{R}, \quad j = 1,\ldots,k.$$

From Sklar's theorem [41], the joint conditional c.d.f. $F(\mathbf{t}|\mathbf{x}) = \mathbb{P}(\mathbf{Y} \le \mathbf{t}|\mathbf{X} = \mathbf{x})$ can be expressed as, for all $\mathbf{t} = (t^{(1)}, \ldots, t^{(k)}) \in \mathbb{R}^k$,

$$F(\mathbf{t}|\mathbf{x}) = \mathfrak{C}_{\mathbf{x}}(F^{(1)}(t^{(1)}|\mathbf{x}), \ldots, F^{(k)}(t^{(k)}|\mathbf{x})), \tag{1}$$

where, for all $\mathbf{x}$, $\mathfrak{C}_{\mathbf{x}}$ is a copula function, i.e., a c.d.f. on $[0,1]^k$ with margins uniformly distributed over $[0,1]$. The copula function $\mathfrak{C}_{\mathbf{x}}$ in (1) is unique if all of the conditional margins $F^{(j)}(\cdot|\mathbf{x})$ are continuous for $j = 1,\ldots, k$, which is the assumption that we will make throughout this article. In general, the analyses of the marginal distributions and the dependence structure are therefore made separately.

In the following, we will consider a semi-parametric assumption as in [1] or [36] by introducing a parametric family of copula functions $\mathcal{C} = \{C_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ with $\Theta \subset \mathbb{R}^m$. We assume identifiability of the parametrization of family $\mathcal{C}$, in the sense that $C_\theta \ne C_{\theta'}$ if $\theta \ne \theta'$. We denote $c_{\boldsymbol{\theta}}$ the copula density associated with $C_{\boldsymbol{\theta}}$, i.e.,

$$c_{\boldsymbol{\theta}}(\mathbf{u}) = \frac{\partial^k C_{\boldsymbol{\theta}}(\mathbf{u})}{\partial u^{(1)} \ldots \partial u^{(k)}}, \quad \boldsymbol{u} = (u^{(1)}, \ldots, u^{(k)}) \in [0,1]^k.$$

In the sequel, we assume that, for all $\mathbf{x} \in \mathbb{R}^d$, $\mathfrak{C}_{\mathbf{x}} \in \mathcal{C}$, meaning that there exists a unique $\boldsymbol{\theta}^0(\mathbf{x}) \in \Theta$ such that

$$\mathfrak{C}_{\mathbf{x}} = C_{\boldsymbol{\theta}^0(\mathbf{x})}. \tag{2}$$

Our aim is then to retrieve the function $\boldsymbol{\theta}^0(\mathbf{x})$ from the data $(\mathbf{Y}_i, \mathbf{X}_i)_{1\leq i\leq n}$.

Our estimation strategy is based on regression trees. A tree $\mathbb{T}$ of size K is a partition of $\mathcal{X}$, i.e., $\mathbb{T} = (\mathcal{T}_\ell)_{\ell=1,\dots,K}$, where $\mathcal{T}_\ell \cap \mathcal{T}_{\ell'} = \varnothing$ for $\ell \neq \ell'$ and $\cup_{\ell=1}^K \mathcal{T}_\ell = \mathcal{X}$. The sets $\mathcal{T}_\ell$, $\ell = 1,\dots,K$ are called leaves, and each leaf $\mathcal{T}_\ell$ is obtained as the intersection of conditions of the type $x_{-,\ell}^{(j)} \leq x^{(j)} \leq x_{+,\ell}^{(j)}$ if $X^{(j)}$ is a quantitative variable (continuous or not), and of the type $x^{(j)} \in \mathcal{A}_\ell^{(j)}$, where $\mathcal{A}_\ell^{(j)}$ is a set of potential modalities for a qualitative variable. This particular structure of the partition is associated with a binary tree structure, where the nodes of the tree correspond to conditions on a given covariate and the leaves of the tree to the final classification. The CART algorithm described in Section 2.2 will make this tree structure more obvious.

Given a tree $\mathbb{T}$ with K leaves, we thus consider estimators of $\boldsymbol{\theta}^0(\mathbf{x})$ that are constant on each leaf of $\mathbb{T}$, i.e., of the type $\sum_{\ell=1}^K \theta_\ell \mathbf{1}_{\mathcal{T}_\ell}(\mathbf{x})$, with $\theta_\ell \in \mathbb{R}^m$. In other words, individuals are divided into K classes, for each of which the dependence structure is described by a different copula (from the same parametric family, but with a specific parameter θ_ℓ). In the ideal case, the target function $\boldsymbol{\theta}^0(\mathbf{x})$ is constant on each leaf of the tree $\mathbb{T}$, meaning that $\boldsymbol{\theta}^0(\mathbf{x}) = \theta_\ell^0$ for $\mathbf{x} \in \mathcal{T}_\ell$, where

$$\theta_\ell^0 = \arg\max_{\boldsymbol{\theta}\in\Theta} \mathbb{E}[\log c_{\boldsymbol{\theta}}(\mathbf{U}_i)\mathbf{1}_{\mathbf{X}_i\in\mathcal{T}_\ell}],$$

where $c_{\boldsymbol{\theta}}$ is the copula density associated with the copula function $C_{\boldsymbol{\theta}}$ and $\mathbf{U}_i$ is the random variable defined by

$$\mathbf{U}_i = (F^{(1)}(Y_i^{(1)}|\mathbf{X}_i), \dots, F^{(k)}(Y_i^{(k)}|\mathbf{X}_i)),$$

which has conditional uniform margins over $[0, 1]$ (i.e., uniform conditionally to $\mathbf{X}_i$) and is jointly distributed, conditionally to $\mathbf{X}_i$, according to the c.d.f. $\mathfrak{C}_{\mathbf{X}_i} = C_{\boldsymbol{\theta}^0(\mathbf{X}_i)}$.

However, in practice, a misspecification bias is expected, since the target function $\boldsymbol{\theta}^0(\mathbf{x})$ is not a piecewise constant function, while the estimator function is. For a given tree $\mathbb{T}$, the corresponding estimator $\widehat{\boldsymbol{\theta}}(\mathbf{x}|\mathbb{T})$ is defined as

$$\widehat{\boldsymbol{\theta}}(\mathbf{x}|\mathbb{T}) = \sum_{\ell=1}^K \widehat{\theta}_\ell \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell},$$

where

$$\widehat{\theta}_\ell = \arg\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^n \log c_{\boldsymbol{\theta}}(\widehat{\mathbf{U}}_i)\mathbf{1}_{\mathbf{X}_i\in\mathcal{T}_\ell},$$

and $(\widehat{\mathbf{U}}_i)_{1\leq i\leq n}$ are pseudo-observations, i.e., estimated versions of $(\mathbf{U}_i)_{1\leq i\leq n}$.

Typically, these pseudo-observations are the result of a preliminary estimation of the marginal distribution, namely, $\widehat{U}_i^{(j)} = \widehat{F}^{(j)}(Y_i^{(j)}|\mathbf{X}_i)$, but alternative procedures are possible: for example, a parametric model can be used to handle the margins. In Section 2.3, we also discuss the possibility of relying on tree-based methods to estimate the margins as well, although there is no obligation to use the same type of technique for the dependence structure as for the margins. Therefore, in the following, we will try to keep our results as general as possible, expressing convergence conditions that this step should verify, but without imposing a specific method. However, let us point out that an interesting feature of regression trees is their ability to deal with both quantitative and qualitative covariates, which requires relying on estimation techniques for the margins that satisfy the same requirements.

The rest of the section is devoted to presenting our estimation procedure based on regression trees. We describe the CART procedure consisting of two steps: the construction of the maximal tree (Section 2.2.1), which determines the proper decomposition of the covariate space $\mathcal{X}$ to obtain the regression tree $\mathbb{T}$ and deduce an estimator $\widehat{\boldsymbol{\theta}}(\cdot|\mathbb{T})$, and the pruning step (Section 2.2.2), which corresponds to a selection model step. Fitting the dependence structure requires a preliminary estimation of the margins, which is done once and for all before starting the algorithm. Various methods may be used to deal with this preliminary step, the only requirement being that they satisfy the conditions under which our theoretical results hold. Examples of possible methods to estimate the margins are presented in Section 2.3.

2.2 Regression tree estimation of the dependence structure

Regression trees provide an easy and transparent way to group observations that have a similar behavior in terms of the response variable Y. They constitute a nonparametric regression model capable of reproducing highly nonlinear trends in the data and are thus able to approximate a wide class of functions. In addition, it can include both quantitative and categorical (non-ordinal) covariates.

Originally proposed by Li et al. [6], regression trees are implemented through the CART algorithm, which involves a two-step process. Initially, a maximal tree is constructed, forming a binary structure that assigns observations to numerous classes (leaves), often leading to overfitting. Subsequently, the maximal tree is pruned to identify the subtree that offers the best compromise between complexity and generalization ability. Section 2.2.1 describes how the construction of the optimal tree takes place, making explicit our split criterion based on the maximization of the log-likelihood of the copula mixture model. In Section 2.2.2, we define the penalization criterion and discuss the pruning phase.

2.2.1 Construction of the maximal tree

Recall that, as mentioned before, the computation of the pseudo-observations $\widehat{\mathbf{U}}_i$ is done once and for all before starting the algorithm. The CART procedure is applied to $(\widehat{\mathbf{U}}_i, \mathbf{X}_i)_{1\le i\le n}$with the aim of maximize the log-likelihood of the model described in Section 2.1. Such log-likelihood function can be written as the sum of the log-likelihoods of the parametric copulas estimated for the individual leaves of the tree:

$$\mathcal{L}_n(\theta_1, \dots, \theta_K) = \sum_{\ell=1}^{K}\left(\frac{1}{n}\sum_{i=1}^{n}\log c_{\theta_\ell}(\widehat{\mathbf{U}}_i)\mathbf{1}_{\mathbf{X}_i\in\mathcal{T}_\ell}\right).$$

More precisely, the log-likelihood of the model is maximized conditionally on the covariates as a consequence of the recursive partitioning of the observations. In fact, at each split, the observations are separated by looking at their values for one of the covariates. Formally, if we denote by $D_P = (\widehat{\mathbf{U}}_i, \mathbf{X}_i)_{i\in P}$ the observations that belong to a certain node P (parent) and by $R_P(\mathbf{X})$ the condition of the covariates that identifies those observations – such that $R_P(\mathbf{X}_i)$ is 1 if $(\widehat{\mathbf{U}}_i, \mathbf{X}_i) \in D_P$, 0 otherwise – then the left and right child nodes are determined by conditions of the type $\{X_i^{(j)} \le s, (\widehat{\mathbf{U}}_i, \mathbf{X}_i) \in D_P\}$ and $\{X_i^{(j)} > s, (\widehat{\mathbf{U}}_i, \mathbf{X}_i) \in D_P\}$. Such split is uniquely determined by the pair (j, s), with $j = 1, \dots, p$ and $s \in \mathbb{R}$. (This is true for quantitative covariates, while in the presence of qualitative covariates, the split is performed following Remark 1.)

Initially, all observations are in the root of the tree, implying that the dependence among variables $\widehat{U}_i$ is modeled by a single copula with parameters $\boldsymbol{\theta}^0_{\text{root}}$, which are optimized using the maximum-likelihood estimation (MLE). Subsequently, each split is carried out to maximize the increase in model log-likelihood. This gain simply corresponds to the sum of log-likelihoods estimated for the child nodes – once more, evaluated in correspondence of the parameters optimized via MLE – from which log-likelihood of the parent node is subtracted. In practical terms, the optimal gain, and thus the optimal split, is determined by testing all possible splits. The splitting process ends when further splits fail to enhance the log-likelihood, or more commonly, upon meeting specific stopping criteria. For instance, a common criterion is setting a minimum number of observations per leaf node.

The pseudocode summarizing the construction of the maximal tree is presented in Algorithm 1.

Algorithm 1: Construction of the maximal tree

Data: $\mathsf{D} \leftarrow (\mathbf{X}_i, \widehat{\mathbf{U}}_i)_{i=1,\dots,n}$
function `StoppingCriteria` (D)
  #Defineconditionstostoptreegrowth
  #Forexamplelimittheminimumobservationsperleaf
  return true *if stopping criteria met*, *otherwise* false

Algorithm 1: Construction of the maximal tree

function `FindOptimalSplit` (D_P)

#Initialization
best_gain, best_j, best_s ← (0, −999, −999)
#Grid search over all the possible features and split values
for *each possible* (j, s) **do**
$D_\ell \leftarrow D_P[X^j \leq s]$
$D_R \leftarrow D_P[X^j > s]$
gain ← LogL(D_ℓ) + LogL(D_R) − LogL(D_P)
if gain > best_gain**then**
best_gain, best_j, best_s ← (gain, j, s)**end**
end
return(best_j, best_s)

function `BuildTree`(D):

#Initialization
$R_{root} \leftarrow \{\forall\ \mathbf{X}\}$
ListRulesInternalNodes ← [R_{root}]
ListRulesLeaves ← [ ]
#Tree construction
whilesize(ListRulesInternalNodes) > 0 **do**
#Retrieve the rule and the observations of the parent node to be split
R_P ← ListRulesInternalNodes [0]
D_P ← D[R_PD]
if StoppingCriteria (D_P) **then**
#Move the rule defining this node in the list of the leaves
ListRulesInternalNodes ← RemoveItem(ListRulesInternalNodes, R_P)
ListRulesLeaves ← AddItem(ListRulesLeaves, R_P)
else
#Find the optimal split and compute the rules defining the left/right children
(j*, s*) ← FindOptimalSplit(D_P)
$R_\ell \leftarrow R_P \wedge \{X^{j^*} \leq s^*\}$
$R_R \leftarrow R_P \wedge \{X^{j^*} > s^*\}$
#Replace the rule of the parent node with the ones of its children
ListRulesInternalNodes ← RemoveItem(ListRulesInternalNodes, R_P)
ListRulesInternalNodes ← AddItem(ListRulesInternalNodes, R_ℓ)
ListRulesInternalNodes ← AddItem(ListRulesInternalNodes, R_R)
end
end
return ListRulesLeaves

Remark 1. The implementation of the CART algorithm here proposed as example requires quantitative (or binary) covariates, for which an ordering of values is straightforward. For qualitative variables with $M > 2$ modalities, the algorithm should include an ordering step preliminary to the split research, as suggested in [42]. Specifically, first, the modalities are sorted by increasing values of the $\widehat{\theta}$ parameter estimated by considering the observations associated with each modality. Then, the $M - 1$ possible splits are evaluated and the optimal one identified. An example of this procedure is available in the code we implemented for the application on the human influenza data (4.2).

2.2.2 Pruning step

Obtaining the maximal tree from the CART algorithm is not sufficient to have a proper estimation of the objective function $\boldsymbol{\theta}^0$, since this decomposition leads to overfitting. A subtree must be extracted from this maximal tree. This subtree will achieve a proper compromise between goodness of fit and complexity.

The complexity is here measured in terms of the number of leaves of a given tree $\mathbb{T}$. The selected subtree is thus obtained through the maximization of the following penalized criterion:

$$\bar{\boldsymbol{\theta}}(\mathbf{x}) = \arg\max_{\hat{\boldsymbol{\theta}}(\cdot|\mathbb{T})} \frac{1}{n}\sum_{i=1}^{n} \log c_{\hat{\boldsymbol{\theta}}(\mathbf{X}_i|\mathbb{T})}(\widehat{\mathbf{U}}_i) - \lambda \dim(\mathbb{T}), \tag{3}$$

where the arg max is taken over all subtrees $\hat{\boldsymbol{\theta}}(\cdot|\mathbb{T})$ of $\mathbb{T}$ and $\dim(\mathbb{T})$ is the number of leaves of $\mathbb{T}$. This criterion could give the impression that one needs to consider all the possible subtrees within the maximal tree, and then select the optimal one. Fortunately, the particular shape of the penalty in (3) ensures that the best tree with K leaves (according to this criterion) is a subtree of the best tree with $K + 1$ leaves [6]. This selection is then performed through validation on a test sample or cross-validation.

2.3 Estimation of the margins

Let us consider a given margin $Y^{(j)}$. If the components of $\mathbf{X}$ are all continuous covariates, a simple non-parametric estimator can be obtained via, for example, kernel smoothing. Following Gijbels et al. [23], i.e.,

$$\widehat{F}^{(j)}(t^{(j)}|\mathbf{x}) = \frac{\sum_{i=1}^{n} K\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right) \mathbf{1}_{Y_i^{(j)} \le t^{(j)}}}{\sum_{i=1}^{n} K\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right)}, \tag{4}$$

where $h > 0$ is the bandwidth and the kernel K is, e.g., $K(\mathbf{x}) = \prod_{j=1}^{d} k(x^{(j)})$ with k a positive function and such that $\int k(u)\mathrm{d}u = 1$. As it is classical for kernel estimators, the rate of uniform convergence is $O(h^2 + [\log n]^{1/2} n^{-1/2} h^{-d/2})$ (see, e.g., [14]), where h^2 corresponds to the bias term.

On the other hand, this estimator is not valid if $\mathbf{X}$ contains some qualitative components. In this perspective, consider the case where $\mathbf{X}$ has M modalities, a possible non-parametric estimator is then

$$\widehat{F}^{(j)}(t^{(j)}|\mathbf{x}) = \frac{1}{n_{\mathbf{x}}} \sum_{i=1}^{n} \mathbf{1}_{Y_i^{(j)} \le t^{(j)}} \mathbf{1}_{\mathbf{X}_i = \mathbf{x}},$$

where $n_{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{1}_{\mathbf{X}_i = \mathbf{x}}$. Note that since the covariates are assumed to be random, so is $n_{\mathbf{x}}$. If $n_{\mathbf{x}}$ were not random, the rate of convergence would be typically the same as for an empirical c.d.f., i.e., $n_{\mathbf{x}}^{-1/2}$.

An alternative way to proceed is also to consider a parametric model for the margins, like, for example, generalized linear model (see, e.g., [27,37]). In this case, and under proper assumptions, the convergence rate in the estimation of the margins can become $n^{-1/2}$, up to a strong assumption on the distributions.

Due to the variety of possible approaches in estimating the margins, we will keep the rest of this article as general as possible regarding this point, only requiring generic convergence assumptions on this preliminary step.

3 Consistency results

This section is dedicated to presenting the main theoretical results that validate the asymptotic behavior of the copula tree estimation procedure. We gather and discuss the list of assumptions required to obtain these results in Section 3.1. Moving on to Section 3.2, we explore the consistency of a single tree (with a given number

of leaves K that may tend to infinity). We chose to focus on the stochastic part of the error, while the approximation error is expected to decrease with K (see, for example, the conclusion of [22] p. 378 on the difficulty to analyze such an approximation error). The rate of the decrease depends on the specific shape of the target function $\boldsymbol{\theta}^0(\mathbf{x})$, which remains an open problem in the regression tree literature. In Section 3.3, we investigate the ability of the penalized criterion to achieve a similar performance as if the optimal number of leaves were known.

3.1 Conditions and assumptions

First, Assumption 1 controls the behavior of the pseudo-observations, i.e., the ability to approach the margins.

Assumption 1. Assume that

$$\sup_{\substack{i=1,\dots,n\\ j=1,\dots,k}} \left| \frac{U_i^{(j)}}{\widehat{U}_i^{(j)}} + \frac{1-U_i^{(j)}}{1-\widehat{U}_i^{(j)}} \right| = O_P(1). \tag{5}$$

For some $0 < \alpha < 1/2$,

$$\sup_{\substack{i=1,\dots,n\\ j=1,\dots,k}} \left| \frac{\widehat{U}_i^{(j)} - U_i^{(j)}}{[U_i^{(j)}(1-U_i^{(j)})]^\alpha} \right| = O_P(\varepsilon_n), \tag{6}$$

for some sequence ε_n that tends to 0 as n tends to infinity.

Assumption 1 is here to control the behavior of the pseudo-observations near the border of $[0,1]^k$. If the margins are estimated via empirical distribution functions, this assumption easily holds from (see [46, Remark ii]). Let us note that estimating the marginal by empirical c.d.f. would not be appropriate in the context of conditional copula (where we need to estimate the conditional distributions). In case this assumption would not hold for more complex estimators, it can be avoided through the introduction of trimming, i.e., removing points too close to the boundaries of the unit square. This would introduce a bias that can be controlled through (6) (see remark 2).

As it will appear in the theoretical results of Sections 3.2 and 3.3, the rate ε_n is expected to go faster to zero than the part related to the estimation of the tree itself; otherwise, it will be predominant. It is important to note that (6) is similar to the slightly stronger condition

$$\sup_{\substack{t\in\mathbb{R}\\ j=1,\dots,d}} \left| \frac{\widehat{F}^{(j)}(t|\mathbf{x}) - F^{(j)}(t|\mathbf{x})}{[F^{(j)}(t|\mathbf{x})(1-F^{(j)}(t|\mathbf{x}))]^\alpha} \right| = O_P(\varepsilon_n).$$

If we consider the estimation of the (unconditional) c.d.f. $F^{(j)}(t) = \mathbb{P}(Y^{(j)} \le t)$ by the empirical distribution function, this condition is easily satisfied with $\varepsilon_n = n^{-1/2}$ (see [45], Example 19.12). In the case of the kernel-based estimator, Section A.5 shows that the rate is slower, namely, $\varepsilon_n = (h^2 + [\log n]^{1/2} n^{-1/2} h^{-d/2})$, and, in the case of discrete covariates, Section A.6 shows that $\varepsilon_n = n^{-1/2}$.

Before presenting the rest of the assumptions, we introduce two conditions on classes of functions, which will be necessary in the following.

Condition 2. A class of functions $\mathcal{F} = \{\mathbf{u} \mapsto \varphi_{\boldsymbol{\theta}}(\mathbf{u}) : \boldsymbol{\theta} \in \Theta\} \subset L^2(\mathbb{R}^k)$ (for some $k > 0$) is said to satisfy Condition 2 if

$$|\varphi_{\boldsymbol{\theta}}(\mathbf{u}) - \varphi_{\boldsymbol{\theta}'}(\mathbf{u})| \le B(\mathbf{u}) \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1, \quad \mathbf{u} \in [0,1]^k,$$

where B is a function in $\mathbb{R}^k$ such that $\mathbb{E}[B(\mathbf{U})^2] < \infty$.

For such a class, there exists an envelope function, i.e., a function Φ such that, for all $\boldsymbol{\theta} \in \Theta, |\phi_{\boldsymbol{\theta}}(\mathbf{u})| \leq \Phi(\mathbf{u})$ and $\mathbb{E}[\Phi(\mathbf{U})^2] < \infty$. Taking any point $\widetilde{\boldsymbol{\theta}} \in \Theta$, Φ can be chosen as $\Phi(\mathbf{u}) = \varphi_{\widetilde{\boldsymbol{\theta}}}(\mathbf{u}) + \mathrm{diam}(\Theta)B(\mathbf{u})$, where $\mathrm{diam}(\Theta)$ denotes the diameter of the compact set Θ. The expectations mentioned here are with respect to the (unconditional) distribution of $\mathbf{U}$.

Condition 3. A class of functions $\mathcal{F} = \{\mathbf{u} \mapsto \varphi_{\boldsymbol{\theta}}(\mathbf{u}) : \boldsymbol{\theta} \in \Theta\} \subset L^2(\mathbb{R}^k)$ (for some $k > 0$) is said to satisfy Condition 3 if

- there exist an envelope Φ and a universal constant A_1 such that, for all $\varphi \in \mathcal{F}$,

$$|\varphi(\mathbf{u})| \leq \Phi(\mathbf{u}) \leq A_1 \sum_{r=1}^{k} \frac{1}{\{u^{(r)}[1-u^{(r)}]\}^{\beta_1}}, \quad \mathbf{u} \in [0,1]^k,$$

  with $0 \leq \beta_1 < 1/2$.
- there exists a universal constant A_2 such that for all $\varphi \in \mathcal{F}$, and for all $j = 1, \ldots, k$,

$$|\partial_j \varphi(\mathbf{u})| \leq \frac{A_2}{\{u^{(j)}[1-u^{(j)}]\}^{\beta_2}} \sum_{r=1}^{d} \frac{1}{\{u^{(r)}[1-u^{(r)}]\}^{\beta_3}},$$

  with $0 \leq \beta_2 \leq 1$, $\beta_3 < 1/2$, and where ∂_j denotes the partial derivative with respect to the jth component of $\mathbf{u}$.

These conditions allow controlling the complexity of the class of functions and are related to classical assumptions used for the consistency of classical MLEs (see, e.g., [44]). The second condition is required to control the behavior of the copula log-likelihood and of its derivatives close to the boundaries of $[0,1]^d$. These conditions are similar to the one used, for example, in previous studies [39,43]. They hold for many classical classes of copula functions such as Gaussian, Clayton, Frank, and Gumbel families. We thus consider the two following assumptions.

Assumption 4. Let

$$\mathcal{F}_1 = \{\mathbf{u} \mapsto \log c_{\boldsymbol{\theta}}(\mathbf{u}), \boldsymbol{\theta} \in \Theta\}.$$

Assume that $\mathcal{F}_1$ satisfies Conditions 2 and 3.

Assumption 5. Let

$$\mathcal{F}_2 = \{\mathbf{u} \mapsto \nabla_{\boldsymbol{\theta}} \log c_{\boldsymbol{\theta}}(\mathbf{u}), \boldsymbol{\theta} \in \Theta\}.$$

Assume that $\mathcal{F}_2$ satisfies Conditions 2 and 3.

3.2 Asymptotic theory for a single tree

In this section, we consider a tree $\mathbb{T} = (\mathcal{T}_\ell)_{\ell=1,\ldots,K}$ with K leaves.

Let

$$\boldsymbol{\theta}^0 = (\theta_1^0, \ldots, \theta_K^0) = \arg \max_{(\theta_1,\ldots,\theta_K)} \sum_{\ell=1}^{K} \mathbb{E}[\log c_{\theta_\ell}(\mathbf{U}) \mathbf{1}_{\mathbf{X} \in \mathcal{T}_\ell}],$$

where the maximum is supposed to be achieved at a unique point $(\theta_1^0, \ldots, \theta_K^0)$, and we denote

$$\boldsymbol{\theta}^0(\mathbf{x}|\mathbb{T}) = \sum_{\ell=1}^{K} \theta_\ell^0 \mathbf{1}_{x \in \mathcal{T}_\ell}.$$

Proposition 1 presented below is a consistency result. To that purpose, we consider the L^1-norm to compare our MLE $\widehat{\boldsymbol{\theta}}(\cdot|\mathbb{T})$ and $\boldsymbol{\theta}^0(\cdot|\mathbb{T})$:

$$||\widehat{\boldsymbol{\theta}}(\cdot|\mathbb{T}) - \boldsymbol{\theta}^0(\cdot|\mathbb{T})||_1 = \int ||\widehat{\boldsymbol{\theta}}(\mathbf{x}|\mathbb{T}) - \boldsymbol{\theta}^0(\mathbf{x}|\mathbb{T})||_1 \mathrm{d}\mathbb{P}_{\mathbf{X}}(\mathbf{x}) = \sum_{\ell=1}^{K} |\widehat{\theta}_\ell - \theta_\ell^0| \mathbb{P}(\mathbf{X} \in \mathcal{T}_\ell),$$

where $\mathbb{P}_{\mathbf{X}}$ is the distribution of the covariates $\mathbf{X}$.

Proposition 1. *Under* Assumptions 1 *to* 4, *and if* $n[K \log K]^{-1} \to \infty$,

$$||\widehat{\boldsymbol{\theta}}(\cdot|\mathbb{T}) - \boldsymbol{\theta}^0(\cdot|\mathbb{T})||_1 = o_P(1).$$

By considering an additional assumption on the copula family, namely, Assumption 5 and conditions on the Hessian matrix, we obtain the convergence rate.

Theorem 2. *Under* Assumptions 1 *to* 5, *and assume furthermore that for* $\ell = 1, \dots, K$, *the Hessian matrix* $\nabla^2_\theta \log c_{\boldsymbol{\theta}}(\theta^0_\ell)$ *is invertible. Then*,

$$||\widehat{\boldsymbol{\theta}}(\cdot|\mathbb{T}) - \boldsymbol{\theta}^0(\cdot|\mathbb{T})||_1 = O_P\left(\frac{[K \log K]^{1/2}}{n^{1/2}} + \varepsilon_n\right).$$

It is not surprising to note that the stochastic part of the error deteriorates with K, due to the increase of the complexity of the model. On the other hand, although this part is harder to track, the approximation error is supposed to decrease with K except for highly irregular target functions (since piecewise constant functions can approximate any piecewise continuous function), which means that $\boldsymbol{\theta}^0(\cdot|\mathbb{T})$ is supposed to be closer to the "true" target function $\boldsymbol{\theta}^0(\cdot)$ when the number of leaves of $\mathbb{T}$ increases.

3.3 Oracle property for the pruning step

Let us define the optimal subtree extracted from the maximal tree obtained via Algorithm 1 (which has $K_{\max}$ leaves) as

$$\boldsymbol{\theta}^0(\mathbf{x}) = \arg\max_{\boldsymbol{\theta}^0(\cdot|\mathbb{T})} \mathbb{E}[\log c_{\boldsymbol{\theta}^0(\mathbf{X}|\mathbb{T})}(\mathbf{U})], \tag{7}$$

where the arg max is taken among all functions $x \to \theta^0(x|\mathbb{T}) = \sum_{l=1}^{K} \theta_l \mathbf{1}_{\mathbf{x} \in \mathcal{T}_l}$, where $(\mathcal{T}_l)_{l=1,\dots,K}$ are the leaves of $\mathbb{T}$, and $\mathbb{T}$ is a subtree of the maximal tree obtained via Algorithm 1. Let K^0 denote the number of leaves of $\boldsymbol{\theta}^0$. If K^0 were known, Theorem 2 shows that one may expect a convergence rate of $\sqrt{K^0 \log K^0 / n}$ for the stochastic part. The next result shows that the penalized procedure has the ability to asymptotically achieve this optimal rate even though the number K^0 is unknown. This, of course, requires conditions on the penalizing constant λ.

Theorem 3. *Assume that the assumptions of* Theorem 2 *hold for all of the subtrees of the maximal tree with* $K_{\max}$ *leaves. Then, if* $\lambda \to 0$, *and if* $\lambda n^{1/2}[K_{\max} \log K_{\max}]^{-1/2} \to \infty$,

$$||\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^0||_1 = O_P\left(\frac{[K^0 \log K^0]^{1/2}}{n^{1/2}} + \varepsilon_n\right).$$

All the proofs are postponed to the Appendix section.

4 Empirical evidence

4.1 Simulation study

In this section, we present the functioning of the conditional copula analysis on simulated data.

4.1.1 Regression framework

We consider a bivariate random variable $\mathbf{U} = (U^{(1)}, U^{(2)})$, with uniform margins over $[0, 1]$ and distributed according to an Archimedean copula $C_{\theta(\mathbf{X})}$. Archimedean copulas is a standard family of copulas often used in modeling applications [28,31]. They are determined by a single parameter $\theta \in \mathbb{R}^1$, which is associated with Kendall's τ coefficient through a bijective relationship [21]. In our framework, θ, and thus also τ, depends on two covariates $\mathbf{X} = (X^{(1)}, X^{(2)})$, which are random variables uniformly distributed in $[0, 1]$. Moreover, we assume that $\mathbf{U}_i$ are samples of the true cumulative marginal distributions of some bivariate response variables $\mathbf{Y} = (Y^{(1)}, Y^{(2)})$ conditionally on the covariates $\mathbf{X}$. Specifically, we assume normal distributions for these margins, with the mean parameters being a linear function of $\mathbf{X}$.

The first step of the simulations consists of generating synthetic data for $(\mathbf{X}_i, \theta_i, \tau_i, \mathbf{U}_i, \mathbf{Y}_i)$. Hence, our goal is to estimate the parameters θ_i (or τ_i) from $(\mathbf{X}_i, \mathbf{Y}_i)$, pretending not to know the true observations $\mathbf{U}_i$, as it is usually the case in a real data scenario. Therefore, as a preliminary step to the conditional copula analysis, we first compute the pseudo-observations. We do that by considering two different approaches, a parametric and a non-parametric one, which will result in two vectors of pseudo-observations, namely, $\mathbf{V}$ and $\mathbf{W}$. Eventually, we fit the conditional copula model to both the $\mathbf{V}$ and $\mathbf{W}$ pseudo-observations, other than to the true margins $\mathbf{U}$ for additional comparison. The goodness of the three fits is evaluated against a benchmark model.

4.1.1.1 Definition of different scenarios

To investigate different scenarios, we consider three Archimedean copulas – the Clayton, Frank, and Gumbel copulas. We also consider three types of dependence between τ and the covariates $(X^{(1)}, X^{(2)})$, which we report below:

(i) a step-wise function:

$$\tau_i = \begin{cases} 0.3, & \text{if } X_i^{(1)} < 0.4, X_i^{(2)} < 0.75, \\ 0.5, & \text{if } X_i^{(1)} \geq 0.4, X_i^{(2)} < 0.75, \\ 0.7, & \text{if } X_i^{(1)} < 0.4, X_i^{(2)} \geq 0.75, \\ 0.9, & \text{if } X_i^{(1)} \geq 0.4, X_i^{(2)} \geq 0.75; \end{cases}$$

(ii) a steep sigmoid:

$$\tau_i = 0.3 - \frac{0.2}{1 + \exp(-40(X_i^{(1)} - 0.4))} - \frac{0.4}{1 + \exp(-40(X_i^{(2)} - 0.75))};$$

(ii) a gentle sigmoid:

$$\tau_i = 0.3 - \frac{0.2}{1 + \exp(-15(X_i^{(1)} - 0.4))} - \frac{0.4}{1 + \exp(-15(X_i^{(2)} - 0.75))}.$$

With these constraints, having fixed the covariates $(X^{(1)}, X^{(2)})$, we obtain nine different conditional copulas, from which we sample $\mathbf{U}$ observations. Let us specify that these conditional copulas are defined such that Kendall's τ coefficients always range in the interval $[0.3, 0.9]$, to ensure comparability.

Finally, in all scenarios, the response variables $\mathbf{Y}$ are defined as follows:

$$\begin{cases} Y_i^{(1)} = \Psi^{-1}(U_i^{(1)} - 1 - 0.2X_i^{(1)} - 0.05X_i^{(2)}), \\ Y_i^{(2)} = \Psi^{-1}(U_i^{(2)} - 1 + 0.1X_i^{(1)} - 0.2X_i^{(2)}), \end{cases}$$

where Ψ is the c.d.f. of the distribution $\mathcal{N}(0, 1)$.

4.1.1.2 Pseudo-observation computation

We consider two alternative methods to compute the pseudo-observations.

First, in a parametric approach, we assume that the marginal distributions of $Y^{(j)}$ conditionally on $\mathbf{X}$ can be approximated by normal distributions with variance fixed at 1. Thus, we estimate the mean parameter through a linear model, i.e., $\hat{\mu}_i^{(j)} = LM(\mathbf{X}_i)$, and we compute the pseudo-observations $\mathbf{V}_i^{(j)} = \Psi^{-1}(Y_i^{(j)} - \hat{\mu}_i^{(j)})$.

Second, to avoid assumptions on the form of the margins, we perform a kernel estimation depending on the covariates as defined in (4). We consider a simple Gaussian kernel, with the bandwidth h optimized depending on the scenario; specifically, we used $h = 0.4$ for Clayton and Frank copulas, $h = 0.3$ for the Gumbel copula. This way, the pseudo-observations $\mathbf{W}_i$ are computed as empirical percentiles, where in the calculation of the empirical c.d.f. the different observations are weighted differently according to their distance in terms of covariates.

4.1.1.3 Model's evaluation

As a reference model, we simply fit the Archimedean copula to $\mathbf{U}$ (and to $\mathbf{V}$ and $\mathbf{W}$), ignoring the additional information carried by the covariates. It means that we estimate a unique value for τ, which corresponds to the estimation provided by the root of the regression tree of the conditional copula model. Hence, the prediction errors of the conditional copula model and of the benchmark model are compared. For comparison, we consider the mean squared errors for both the estimates of the τ coefficients and the values of the cumulative copula and the log-likelihood values of the models toward the observations/pseudo-observations.

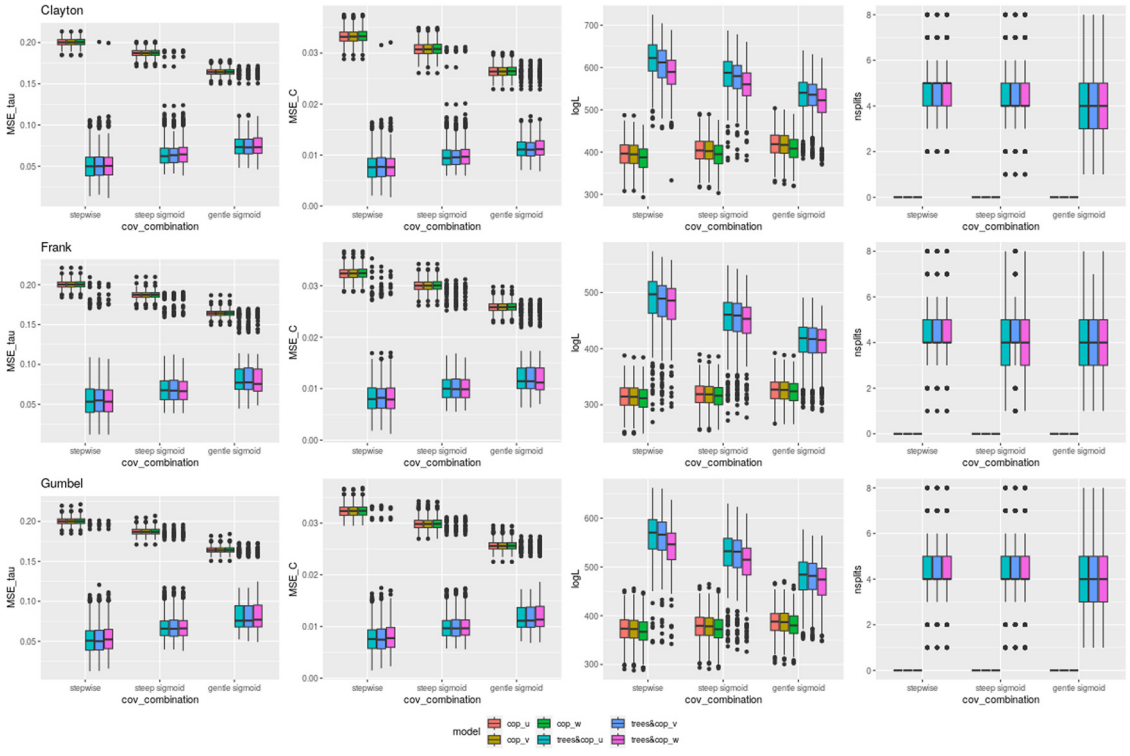

Figure 1: Results of simulations. Results for the Clayton, Frank, and Gumbel copulas are depicted on the different rows. For each copula, the results for the three types of covariate dependence are reported on the *x*-axis. The six colors identify different models: red, orange, and green are for the conditional copula model fitted on the observations $\mathbf{U}$ and on the pseudo-observations $\mathbf{V}$ and $\mathbf{W}$, respectively, while cyan, blue, and magenta are for the benchmark model. In the first two columns, we show results in terms of MSE for the τ estimates and the cumulative copula estimates, in the third column in terms of log-likelihood. In the fourth column, we report the distributions of the number of splits, i.e., the number of leaves minus 1, identified by the regression trees of the conditional copula models. Each boxplot represents the results for 500 datasets of 1,000 points each.

4.1.1.4 Simulation results

For each one of the nine settings presented above, we build 500 triples of datasets, containing 1,000 observations $\mathbf{U}_i$, 1,000 pseudo-observations $\mathbf{V}_i$, and 1,000 pseudo-observations $\mathbf{W}_i$, respectively. Results are presented in Figure 1. In all scenarios, the conditional copula model outperforms the benchmark model, in terms of log-likelihood values and estimates for the τ coefficients and for the cumulative copula values. As expected, the predictions worsen when the dependence on the covariates changes from a step function, which can be perfectly captured by a regression tree, to smoother functions. Finally, no significant changes are noted when models are fitted to observations or pseudo-observations. We note that the conditional copula model most of the time identifies five or six groups of observations, which corresponds to a slightly overfitting of the model, as four groups are expected.

4.2 Real data example

In this section, we present an application of the conditional copula model on epidemiological data of cases of human influenza.

4.2.1 Human influenza: Context and data

Three main influenza strains co-circulate worldwide and infect humans: influenza A\H1N1pdm, influenza A \H3N2, and influenza B. The relative proportions of the three viruses are highly variable in time and space, and the unpredictability of the strains' (co-)dominance patterns poses a major limitation to the mitigation of the upcoming epidemic wave in terms of intervention design and vaccination. Here, we use the conditional copula model to capture some trends of the coupled dynamic of influenza subtypes. In particular, first, we assume that we can use Archimedean copulas to describe the dependence structure of the relative abundance of influenza subtypes across regions and years. Second, we implement the conditional copula model to identify spatio-temporal patterns of such dependence structure.

The World Health Organisation (WHO) provides data on influenza surveillance for several countries, consisting of weekly counts of cases classified by subtype [20,25]. We consider data from 80 countries that reported a minimum of 50 classified cases per year in the period from April 2010 to April 2019. Then, we aggregate counts annually (from April to April) and for each country-year (800 observations in all), we compute the proportion of cases of A\H1N1pdm, A\H3N2, and B. We consider the relative abundance of subtypes as the response variables to be modeled with an Archimedean copula, testing Clayton, Frank, and Gumbel families, and the year and the influenza transmission zone (ITZ) as the relevant covariates. The ITZs are 18 groups of countries with similar influenza transmission patterns identified by the WHO (for the precise definition of the groups, see [47]). Before fitting the conditional copula model, we perform a preprocessing step by applying an additive log-ratio transformation to the relative proportion of subtypes [2]. This is a common procedure when working with percentage data [7,19,29], and it allows us to map bounded vectors (A\H1N1pdm %, A \H3N2 %, B %) $\in S^3$ into unbounded vectors $(Y^{(1)}, Y^{(2)}) \in \mathbb{R}^2$, where S^3 is the so-called 3-part simplex. In particular, we use the isometric log-ratio transformation proposed by Egozcue et al. [13]:

$$\begin{cases} Y^{(1)} = \sqrt{\frac{2}{3}} \ln \frac{B\%}{\sqrt{\text{A\H1N1pdm\%*A\H3N2\%}}} \\ Y^{(2)} = \sqrt{\frac{1}{2}} \ln \frac{\text{A\H1N1pdm\%}}{\text{A\H3N2\%}}. \end{cases}$$

Thus, the actual response variable is $\mathbf{Y} = (Y^{(1)}, Y^{(2)})$, with $Y^{(1)}$ describing the relative abundance between influenza B and the average proportion of influenza A subtypes, while $Y^{(2)}$ denotes the relative amount of A\H1N1pdm and A\H3N2.

4.2.2 Model implementation

4.2.2.1 Estimation of the margins

We consider regression trees to model the relationship between the response variables $Y^{(j)}$ and the covariates year and ITZ. Both the covariates are treated as categorical variables, meaning that *a priori* values have no precise sorting and an ordering step is needed preliminary to the split search, as explained in Remark 1. This allows maximum flexibility to the splitting procedure so that the tree can effectively capture the trends in the data. Once the trees are optimized by cross-validation, the pseudo-observations $\widehat{\mathbf{U}}^{(j)}$ are computed from a mixture of empirical c.d.f. defined over the groups of points identified by the optimal tree, i.e.,

$$\widehat{\mathbf{U}}^{(j)}(t^{(j)}|\mathbf{X}) = \sum_{l=1}^{K}\left(\frac{1}{n_l}\sum_{i=1}^{n_l}\mathbf{1}_{Y_i^{(j)}\leq t^{(j)}}\mathbf{1}_{\mathbf{X}_i\in\mathcal{T}_l^{(j)}}\right),$$

with $\mathcal{T}_{l=1,\dots,K}^{(j)}$ being the terminal nodes of the tree fitted on the $Y^{(j)}$ response variable.

4.2.2.2 Fit of the conditional copula model

We test two-dimensional Clayton, Frank, and Gumbel copulas to model $\widehat{\mathbf{U}}$ conditionally on the covariates year and ITZ, again treated as categorical covariates. We implement a 3-fold cross-validation repeated 50 times to

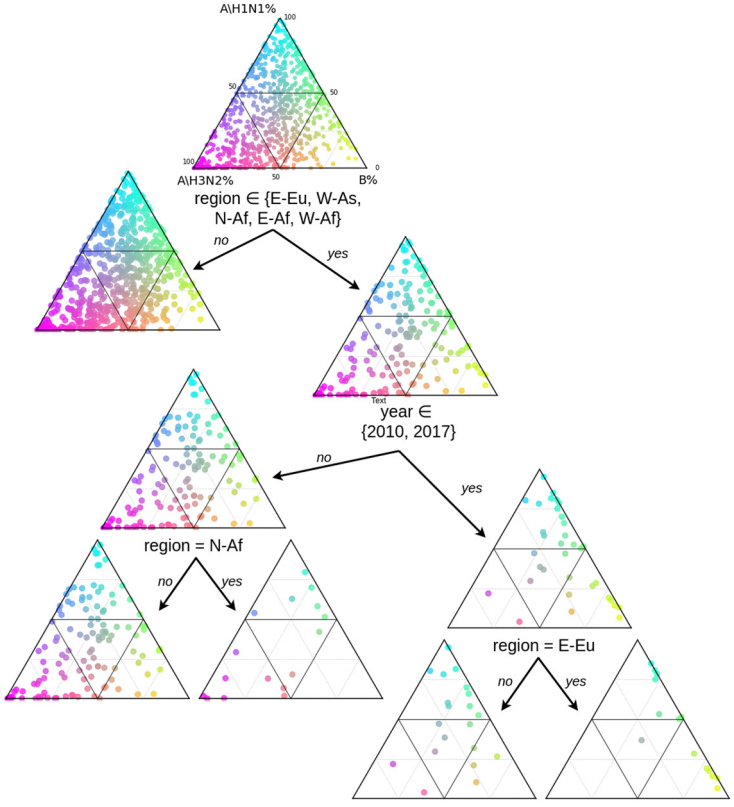

Figure 2: Optimal tree identified by the Frank conditional copula model applied to data of relative abundance of influenza subtypes across countries and regions. Data on the relative abundance of influenza subtypes are considered for 800 countries-years (corresponding to the 800 points in the top ternary plot). Similarly, for each node of the tree, a simplex represents the subtype relative abundance of the countries-years clustered in the node. We use a ternary color code to distinguish countries-years with dominance of A/H1N1pdm (cyan), A/H3N2 (pink), and B (yellow). For each split, the condition used to partition the observations is indicated. From top-left to bottom-right, the number of observations in each leaf is 630, 120, 16, 20, and 14. In the same order, Kendall's τ coefficients equal 0.06, 0.09, 0.3, 0.03, and 0.25.

optimize the pruning of the trees with the penalized approach (3) used to identify the optimal tree. The best conditional model is the one with the Frank copula, which leads to the highest value of log-likelihood. It results in a tree with five leaves (Figure 2), which will be discussed in the next paragraph.

4.2.3 Results and discussions

Thanks to the conditional copula model, we were able to ameliorate the adjustment to the data; the log-likelihood of the Frank simple copula on all the 800 $\widehat{\mathbf{U}}$ pseudo-observations was 1.2, and it increased to 13.8 for the Frank copula mixture model identified by the optimal tree. This improvement is significant, in the sense that the pruning methodology (i.e., the penalized model selection procedure (3)) does not select a tree reduced to its root.

In the estimation of the response variables $Y^{(1)}$ and $Y^{(2)}$, the years are used more often than the regions to perform the splits, indicating that the relative abundance of B vs A and A/H1N1pdm vs A/H3N2, taken independently, varied more in time than in space (see Figure A1 in the Appendix). In other words, to a first approximation, we find consistent temporal dynamics worldwide, going a step further we also identify significant spatial patterns. It is interesting to note that each time the spatial information is used to perform the split, European regions are grouped together, sometimes with other neighboring regions (mainly North Africa and Western and Central Asia), and always separated from countries of the southern hemisphere. These spatial country groupings overall match well the geographical clustering found in other studies with different methods. Previous studies found evidence for an annual reseeding of influenza viruses from tropical and subtropical countries to temperate regions, especially for A/H3N2 viruses [4,5,33,34]. These dynamics could also contribute to determining the patterns in subtype compositions that emerged from our analysis. However, our purely descriptive analysis does not allow us to speculate on any underlying mechanism.

Once $\widehat{U}^{(1)}$ and $\widehat{U}^{(2)}$ are computed, the conditional copula model identifies significant changes in the pseudo-observation dependence across space and time. It results in a tree with five leaves, characterized by different degrees of correlation (Kendall's τ among the $\widehat{U}^{(1)}$ and $\widehat{U}^{(1)}$ pseudo-observation range from -0.09 to 0.3). However, we note that a single leaf contains most of the data points (630 out of 800), meaning that the simplifying assumption would probably provide a reasonable approximation for the majority of the countries-years in our analysis. However, the other leaves allow us to refine the fit of the data and further separate a few country-years that are mainly characterized by a proportion of B infections higher than the average.

5 Conclusion

In this article, we proposed a new methodology to model conditional copulas, based on regression trees. The technique requires the assumption that the conditional copulas all belong to the same family of parametric copulas, with the association parameter changing with the value of the covariates. The procedure presents many advantages. First, the tree structure theoretically allows capturing any form of the conditional association parameter. Second, the simplicity of the final model, if restricted to a single leaf of the tree, allows one to obtain a tractable output. We note that our approach allows a relaxation of the simplifying assumption [10], but this remains valid for each of the subsets of data identified by the tree. Another interesting feature is the ability to deal with discrete and/or continuous covariates.

In addition, let us point out that this method can be easily extended to the case where several families of copulas are tested at each node. This would give a more complex final structure, since not only the association parameter but also the copula family could vary from one leaf to another. However, it increases the complexity of the implementation of the algorithm. Finally, let us note that the potential weakness of the procedure is its instability. Like every regression tree procedure, our method can be very sensitive to new incoming data, as new information may considerably change the structure of the tree and the classes that are made. Careful

attention should be given to this aspect. On the other hand, a direct extension that could reduce this instability would be to consider the corresponding random forest algorithm, i.e., computing many small copula trees on separate bootstrap samples and then aggregating them. The aggregation of these trees would be a way to stabilize the result, but of course, would reduce the interpretability of the model.

R codes

The `R` codes are publicly available at https://github.com/FrancescoBonacina/tree-based-conditional-copula-estimation.

Acknowledgements: The authors are grateful for the reviewer's valuable comments that improved the manuscript.

Funding information: Olivier Lopez received funding from the Excellence chair CARE under the aegis of Fondation du Risque, in partnership with GENES, and with the support of Allianz.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and consented to its submission to the journal, reviewed all the results, and approved the final version of the manuscript.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Appendix

The Appendix section is organized as follows. We first start with preliminary results that are needed to prove our results in Section A.1, including some results on the complexity of the class of functions defined by the model in Sections A.1.1, and A.1.2 provides a general result that will be used repeatedly to handle deviations of the score function. We then prove Proposition 1 in Section A.2,Theorem 2 in Section A.3, and Theorem 3 in Section A.4. Results on the convergence rates of the margins are then shown in Sections A.5 and A.6.

A.1 Preliminary results

In all this section, let us denote

$$\mathfrak{F} = \left\{ (\mathbf{u}, \mathbf{x}) \mapsto \phi(\mathbf{u};\mathbf{x}) = \sum_{\ell=1}^{K} \varphi_\ell(\mathbf{u}) \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell} \text{ with for } \ell = 1, \ldots, K, \varphi_\ell \in \mathcal{F} \text{ satisfying Condition 2} \right\},$$

and, for $\phi \in \mathfrak{F}$,

$$\mathcal{Z}(\phi) = \mathbb{E}[\phi(\mathbf{U};\mathbf{X})], \quad \mathcal{Z}_n^*(\phi) = \frac{1}{n}\sum_{i=1}^{n} \phi(\mathbf{U}_i;\mathbf{X}_i) \quad \text{and} \quad \widehat{\mathcal{Z}}_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} \phi(\widehat{\mathbf{U}}_i;\, \mathbf{X}_i).$$

A.1.1 Bracketing numbers

We first introduce the concept of bracketing numbers to measure the complexity of a class of functions $\mathcal{F}$. For $\varepsilon > 0$, a ε-bracket $[\mathfrak{a}, \mathfrak{b}]$ is the set of functions f such that for all $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{u} \in \mathbb{R}^k$, $\mathfrak{a}(\mathbf{u}, \mathbf{x}) \leq f(\mathbf{u}, \mathbf{x}) \leq \mathfrak{b}(\mathbf{u}, \mathbf{x})$, with the condition that

$$\int (\mathfrak{a}(\mathbf{u}, \mathbf{x}) - \mathfrak{b}(\mathbf{u}, \mathbf{x}))^2 d\mathbb{P}(\mathbf{u}, \mathbf{x}) \leq \varepsilon^2.$$

We then define $\mathcal{N}(\varepsilon, \mathcal{F})$ as the minimal number of ε-brackets required to cover the class of functions $\mathcal{F}$. More details on bracketing numbers can be found in [45, Chapter 19], and [44, Chapter 2.2].

Lemma 4. *For $\varepsilon > 0$,*

$$\mathcal{N}(\varepsilon, \mathfrak{F}) \leq \left(\frac{K^{m/2} C_1 ||\Phi||_2^m}{\varepsilon^m} \right)^K,$$

for some constant C_1 depending only on Θ and m.

Proof. Consider an element $\phi \in \mathfrak{F}$. It can be written as $\phi = \sum_{\ell=1}^K \varphi_\ell \mathbf{1}_{\mathbf{x} \in \mathcal{T}_\ell}$, where each φ_ℓ is in $\mathcal{F}$, and satisfies Condition 2. Then, from [45, Example 19.7], for each $\ell = 1, \ldots, K$, for all $\varepsilon > 0$,

$$\mathcal{N}(\varepsilon, \mathcal{F}) \leq \frac{C_1(m, \Theta) ||\Phi||_2^m}{\varepsilon^m},$$

where C_1 is a constant depending on $\mathrm{diam}(\Theta)$ and m. Let $([a_i, b_i])_{1 \leq i \leq \mathcal{N}(\varepsilon K^{-1/2}, \mathcal{F})}$ denote the brackets of functions that cover $\mathcal{F}$ (with $\int (b_i(\mathbf{u}) - a_i(\mathbf{u}))^2 d\mathbb{P}(u) \leq \varepsilon K^{-1}$).

Therefore, for each φ_ℓ, there exists $i(\ell)$ such that $a_{i(\ell)} \leq \phi_l \leq b_{i(\ell)}$.

Now, let $\mathfrak{i} = (i(1), \ldots, i(K))$, and define

$$\mathfrak{a}_\mathfrak{i}(\mathbf{u}, \mathbf{x}) = \sum_{\ell=1}^K a_{i(\ell)}(\mathbf{u}) \mathbf{1}_{\mathbf{x} \in \mathcal{T}_\ell}; \; \mathfrak{b}_\mathfrak{i}(\mathbf{u}, \mathbf{x}) = \sum_{\ell=1}^K b_{i(\ell)}(\mathbf{u}) \mathbf{1}_{\mathbf{x} \in \mathcal{T}_\ell}. \tag{A1}$$

Clearly, $\mathfrak{a}_\mathfrak{i} \leq \mathfrak{b}_\mathfrak{i}$ and $\phi \in [\mathfrak{a}_i, \mathfrak{b}_i]$. Moreover,

$$\int (\mathfrak{b}_i(\mathbf{u}, \mathbf{x}) - \mathfrak{a}_i(\mathbf{u}, \mathbf{x}))^2 d\mathbb{P}(\mathbf{u}, \mathbf{x}) \leq \sum_{\ell=1}^K \int (b_{i(\ell)}(\mathbf{u}) - a_{i(\ell)}(\mathbf{u}))^2 d\mathbb{P}(\mathbf{u}) \leq \varepsilon^2.$$

Thus, the set of $\varepsilon-$brackets $[\mathfrak{a}_\mathfrak{i}, \mathfrak{b}_\mathfrak{i}]$ covers $\mathfrak{F}$. Their number is less that

$$\left(\frac{K^{m/2} C_1(m, \Theta) ||\Phi||_2^m}{\varepsilon^m} \right)^K,$$

leading to the result. □

A.1.2 General results on sums involving pseudo-observations

The first result of this section shows how to replace pseudo-observations $\mathbf{U}_i$ by their estimated version $\widehat{\mathbf{U}}_i$ in studying the asymptotic behavior of sums involving these quantities. Going back to $\mathbf{U}_i$ then simplifies considerably the study of such quantities, since one goes back to classical i.i.d. quantities.

Lemma 5. *Assume furthermore that there exist $0 \leq \beta_1, \beta_3 < 1/2$, and $0 \leq 1\beta_2 < 1$ and two universal constants A_1 and A_2 such that for all $\varphi \in \mathcal{F}$ satisfies Condition 3.*

Then, under Assumption 1,

$$\sup_{\phi\in\mathfrak{F}}|\widehat{\mathcal{Z}}_n(\phi) - \mathcal{Z}_n^*(\phi)| = O_P(\varepsilon_n),$$

where ε_n *tends to* 0 *when* n *tends to* ∞.

Proof. First, recall that

$$\sup_{\phi\in\mathfrak{F}}|\widehat{\mathcal{Z}}_n(\phi) - \mathcal{Z}_n^*(\phi)| = \sup_{\phi\in\mathfrak{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\{\phi(\widehat{\mathbf{U}}_i;\mathbf{X}_i) - \phi(\mathbf{U}_i;\mathbf{X}_i)\}\right|.$$

Then, from a Taylor expansion,

$$\frac{1}{n}\sum_{i=1}^{n}\{\phi(\widehat{\mathbf{U}}_i;\mathbf{X}_i) - \phi(\mathbf{U}_i;\mathbf{X}_i)\} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\partial_j\phi(\mathbf{U}_i + t[\hat{\mathbf{U}}_i - \mathbf{U}_i];\mathbf{X}_i)[\widehat{U}_i^{(j)} - U_i^{(j)}],$$

for some $t \in (0,1)$.

Let $\tilde{\mathbf{U}}_i = \mathbf{U}_i + t[\hat{\mathbf{U}}_i - \mathbf{U}_i]$. We have, from Condition 3,

$$|\partial_j\phi(\tilde{\mathbf{U}}_i;\mathbf{X}_i)| \le \frac{A_2}{[\tilde{U}_i^{(j)}(1-\tilde{U}_i^{(j)})]^{\beta_2}} \times \left\{\sum_{r=1}^{k}\frac{1}{[\tilde{U}_i^{(r)}(1-\tilde{U}_i^{(r)})]^{\beta_3}}\right\}.$$

Note that

$$\frac{1}{\widetilde{U}_i^{(r)}(1-\widetilde{U}_i^{(r)})} \le \sup_{i=1,\dots,n}\left(\frac{U_i^{(r)}}{\widehat{U}_i^{(r)}} + \frac{1-U_i^{(r)}}{1-\widehat{U}_i^{(r)}}\right)\frac{1}{U_i^{(r)}(1-U_i^{(r)})},$$

leading to

$$|\partial_j\phi(\tilde{\mathbf{U}}_i;\mathbf{X}_i)| \le \frac{A_2\left[\max\left(1, \max_{r=1,\dots,k}\left(\frac{U_i^{(r)}}{\widehat{U}_i^{(r)}} + \frac{1-U_i^{(r)}}{1-\widehat{U}_i^{(r)}}\right)\right)\right]^{\beta_3+\beta_2}}{[U_i^{(j)}(1-U_i^{(j)})]^{\beta_2}}\left\{\sum_{r=1}^{k}\frac{1}{[U_i^{(r)}(1-U_i^{(r)})]^{\beta_3}}\right\}.$$

Let

$$Z_i = \sum_{r=1}^{k}\frac{1}{[U_i^{(r)}(1-U_i^{(r)})]^{\beta_3}}.$$

Since $\beta_3 < 1/2$, $E[Z_i^2] < \infty$. Indeed, for all $r = 1,\dots,k$

$$E\left[\frac{1}{(U_i^{(r)})^{2\beta_3}[1-U_i^{(r)}]^{2\beta_3}}\right] \le 2^{2\beta_3}\left[\int_0^{1/2}\frac{\mathrm{d}u}{u^{2\beta_3}} + \int_{1/2}^{1}\frac{\mathrm{d}u}{(1-u)^{2\beta_3}}\right]$$
$$\le 2^{2\beta_3+1}\int_0^{1/2}\frac{\mathrm{d}u}{u^{2\beta_3}} < \infty.$$

Hence,

$$|\widehat{\mathcal{Z}}_n(\phi) - \mathcal{Z}_n^*(\phi)| \le \frac{A_2}{n}\sum_{j=1}^{k}\sum_{i=1}^{n}\frac{Z_i}{[U_i^{(j)}(1-U_i^{(j)})]^{\beta_2-\beta'}}\sup_{\substack{i=1,\dots,n\\ j=1,\dots,k}}\left|\frac{\widehat{U}_i^{(j)} - U_i^{(j)}}{[U_i^{(j)}(1-U_i^{(j)})]^{\beta'}}\right|$$
$$\times \sup_{\substack{i=1,\dots,n\\ j=1,\dots,k}}\left(\frac{U_i^{(j)}}{\widehat{U}_i^{(j)}} + \frac{1-U_i^{(j)}}{1-\widehat{U}_i^{(j)}}\right)^{\beta_3+\beta_2},$$

with $\beta' = \min(\beta_3, \alpha)$. Then, first, from Cauchy–Schwarz inequality,

$$\mathbb{E}\left[\left\{\frac{Z_i}{[U_i^{(j)}(1-U_i^{(j)})]^{\beta_2-\beta'}}\right\}^2\right] \leq \mathbb{E}[Z_i^2]^{1/2}\mathbb{E}\left[\frac{1}{[U_i^{(j)}(1-U_i^{(j)})]^{2[\beta_2-\beta']}}\right]^{1/2} < \infty.$$

Second, from Assumption 1,

$$\sup_{\substack{i=1,\dots,n\\ j=1,\dots,k}} \left|\frac{\widehat{U}_i^{(j)} - U_i^{(j)}}{[U_i^{(j)}(1-U_i^{(j)})]^{\beta'}}\right| = O_p(\varepsilon_n)$$

and

$$\sup_{\substack{i=1,\dots,n\\ j=1,\dots,k}} \left|\frac{U_i^{(j)}}{\widehat{U}_i^{(j)}} + \frac{1-U_i^{(j)}}{1-\widehat{U}_i^{(j)}}\right|^{\beta_3+\beta_2} = O_P(1). \qquad \square$$

Remark 2. If (5) does not hold, the estimation procedure can be modified by introducing some trimming, i.e., multiplying each term of the log-likelihood by $\mathbf{1}_{\min(1-\hat{U}_i^{(j)},\hat{U}_i^{(j)})\geq\eta_n}$. If η_n tends to zero slower than ε_n, $\hat{U}_i^{(j)} \geq U_i^{(j)}/2$ for n large enough due to (6) for the indexes i where this indicator function is not zero. However, the introduction of trimming induces some bias for the estimator, which can be controlled thanks to Assumption 4.

With at hand Lemma 5 and the complexity bound of Lemma 4, one can derive the main result of this section, which will be used several times in the proof of our main theorems.

Proposition 6. *Assume furthermore that there exist* $0 \leq \beta_1, \beta_3 < 1/2$, $0 \leq /\beta_2 < 1$ *and two universal constants* A_1 *and* A_2 *such that for all* $\varphi \in \mathcal{F}$ *satisfies Condition* 3. *Then,*

$$\sup_{\phi\in\mathfrak{F}}|\widehat{\mathcal{Z}}_n(\phi) - \mathcal{Z}(\phi)| = O_P\left(\sqrt{\frac{K\log K}{n}} + \varepsilon_n\right).$$

Proof. Writing

$$\sup_{\phi\in\mathfrak{F}}|\widehat{\mathcal{Z}}_n(\phi) - \mathcal{Z}(\phi)| \leq \sup_{\phi\in\mathfrak{F}}|\widehat{\mathcal{Z}}_n(\phi) - \mathcal{Z}_n^*(\phi)| + \sup_{\phi\in\mathfrak{F}}|\mathcal{Z}_n^*(\phi) - \mathcal{Z}(\phi)|. \tag{A2}$$

For the first term, from Lemma 5,

$$\sup_{\phi\in\mathfrak{F}}|\widehat{\mathcal{Z}}_n(\phi) - \mathcal{Z}_n^*(\phi)| = O_P(\varepsilon_n). \tag{A3}$$

For the second term, introduce for $\delta > 0$, $J(\delta, \mathfrak{F}) = \int_0^\delta \sqrt{\log \mathcal{N}(\varepsilon, \mathfrak{F})}\,\mathrm{d}\varepsilon$. From [45, Corollary 19.35],

$$\sqrt{n}\mathbb{E}\left[\sup_{\phi\in\mathfrak{F}}|\mathcal{Z}_n^*(\phi) - \mathcal{Z}(\phi)|\right] \leq A_3 J(\|\Phi\|_2, \mathfrak{F}),$$

for some universal constant $A_3 \geq 0$. Then, from Lemma 4,

$$J(\|\Phi\|_2, \mathfrak{F}) \leq \int_0^{\|\Phi\|_2} K^{1/2}\left\{\frac{m}{2}\log K + \log(C_1\|\Phi\|_2^m) + m\log\left(\frac{1}{\varepsilon}\right)\right\}^{1/2}\mathrm{d}\varepsilon.$$

Hence,

$$\frac{\sqrt{n}}{\sqrt{K\log K}}\mathbb{E}\left[\sup_{\phi\in\mathfrak{F}}|\mathcal{Z}_n^*(\phi) - \mathcal{Z}(\phi)|\right] \leq C_2(m, \Theta),$$

which shows that $\sup_{\phi\in\mathfrak{F}}|\mathcal{Z}_n^*(\phi) - \mathcal{Z}(\phi)| = O_P([K\log K]^{1/2}\, n^{-1/2})$. The results then follows from (A2) and (A3). $\square$

A.2 Proof of Proposition 1

We are now ready to prove Proposition 1.

Recall that

$$\|\widehat{\boldsymbol{\theta}}(\cdot|\mathbb{T}) - \boldsymbol{\theta}^0(\cdot|\mathbb{T})\|_1 = \sum_{\ell=1}^{K} |\widehat{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta}^*_\ell| \mathbb{P}(\mathbf{X} \in \mathcal{T}_\ell),$$

so that it suffices to show that

$$\sup_{\ell=1,\dots,K} |\widehat{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta}^0_\ell| = o_P(1). \tag{A4}$$

Let

$$\widehat{\mathcal{L}}_n(\theta_1, \dots, \theta_K) = \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{K} \log c_{\theta_\ell}(\widehat{\mathbf{U}}_i) \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_\ell},$$

$$\mathcal{L}^*_n(\theta_1, \dots, \theta_K) = \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{K} \log c_{\theta_\ell}(\mathbf{U}_i) \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_\ell},$$

$$\mathcal{L}(\theta_1, \dots, \theta_K) = \mathbb{E}\left[\sum_{\ell=1}^{K} \log c_{\theta_\ell}(\mathbf{U}_i) \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_\ell} \right].$$

From [44, Corollary 3.2.3], (A4) holds if

$$\sup_{\theta_1, \dots, \theta_\ell} |\widehat{\mathcal{L}}_n(\theta_1, \dots, \theta_\ell) - \mathcal{L}(\theta_1, \dots, \theta_\ell)| = o_P(1).$$

Let us introduce

$$\mathfrak{F}_1 = \left\{ (\mathbf{u}, \mathbf{x}) \mapsto \sum_{\ell=1}^{K} \log c_{\boldsymbol{\theta}}(\mathbf{u}) \mathbf{1}_{\mathbf{x} \in \mathcal{T}_\ell} \quad \text{with } \boldsymbol{\theta} = (\theta_\ell)_{\ell=1,\dots,K} \in \Theta^K \right\}. \tag{A5}$$

From Assumption 4, Proposition 6 applies to $\mathfrak{F}_1$, leading to

$$\sup_{\theta_1, \dots, \theta_\ell} |\widehat{\mathcal{L}}_n(\theta_1, \dots, \theta_\ell) - \mathcal{L}(\theta_1, \dots, \theta_\ell)| = \sup_{\phi \in \mathfrak{F}_1} |\mathcal{Z}^*_n(\phi) - \mathcal{Z}(\phi)| = O_P\left(\sqrt{\frac{K \log K}{n}} + \varepsilon_n \right),$$

which tends to zero under the condition on K and the result follows.

A.3 Proof of Theorem 2

Introduce

$$\dot{\mathcal{L}}_n(\theta_1, \dots, \theta_K) = \frac{1}{n} \sum_{\ell=1}^{K} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log c_{\theta_\ell}(\widehat{\mathbf{U}}_i) \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_\ell},$$

$$\dot{\mathcal{L}}(\theta_1, \dots, \theta_K) = \sum_{\ell=1}^{K} \mathbb{E}[\nabla_{\boldsymbol{\theta}} \log c_{\theta_\ell}(\mathbf{U}) \mathbf{1}_{\mathbf{X} \in \mathcal{T}_\ell}],$$

and

$$\mathfrak{F}_2 = \left\{ (\mathbf{u}, \mathbf{x}) \to \sum_{\ell=1}^{K} \nabla_{\boldsymbol{\theta}} \log c_{\theta_\ell}(\mathbf{u}) \mathbf{1}_{\mathbf{x} \in \mathcal{T}_\ell} : (\theta_\ell)_{l=1,\dots,K} \in \Theta^K \right\}.$$

From Proposition 6,

$$\sup_{\theta_1,\ldots,\theta_\ell} |\dot{\mathcal{L}}_n(\theta_1,\ldots,\theta_K) - \dot{\mathcal{L}}(\theta_1,\ldots,\theta_K)| = O_P\left(\frac{[K\log K]^{1/2}}{n^{1/2}} + \varepsilon_n\right). \tag{A6}$$

Then, write

$$\begin{aligned}\dot{\mathcal{L}}(\theta_1^0,\ldots,\theta_K^0) - \dot{\mathcal{L}}(\hat{\theta}_1,\ldots,\hat{\theta}_K) = &\{\dot{\mathcal{L}}(\theta_1^0,\ldots,\theta_K^0) - \dot{\mathcal{L}}_n(\theta_1^0,\ldots,\theta_K^0)\}\\ &+ \{\dot{\mathcal{L}}_n(\theta_1^0,\ldots,\theta_K^0) - \dot{\mathcal{L}}_n(\hat{\theta}_1,\ldots,\hat{\theta}_K)\}\\ &+ \{\dot{\mathcal{L}}_n(\hat{\theta}_1,\ldots,\hat{\theta}_K) - \dot{\mathcal{L}}(\hat{\theta}_1,\ldots,\hat{\theta}_K)\}.\end{aligned}$$

The rates of the first and last brackets in this decomposition are given by (A6), while the middle one is

$$\{\dot{\mathcal{L}}_n(\theta_1^0,\ldots,\theta_K^0) - \dot{\mathcal{L}}_n(\hat{\theta}_1,\ldots,\hat{\theta}_K)\} = \dot{\mathcal{L}}_n(\theta_1^0,\ldots,\theta_K^0),$$

has also the same rate. This shows that

$$|\dot{\mathcal{L}}(\theta_1^0,\ldots,\theta_K^0) - \dot{\mathcal{L}}(\hat{\theta}_1,\ldots,\hat{\theta}_K)| = O_P\left(\frac{[K\log K]^{1/2}}{n^{1/2}} + \varepsilon_n\right).$$

From the assumption on the Hessian matrix, and since Proposition 1 applies (which guarantees that each $\hat{\theta}_\ell$ is in an arbitrary small neighborhood of θ_ℓ^* for n large enough), we obtain

$$|\dot{\mathcal{L}}(\theta_1^0,\ldots,\theta_K^0) - \dot{\mathcal{L}}(\hat{\theta}_1,\ldots,\hat{\theta}_K)| \geq \alpha||\hat{\theta} - \theta^0||_1,$$

for some $\alpha > 0$, from a Taylor expansion, and the result follows.

A.4 Proof of Theorem 3

Let $\hat{\boldsymbol{\theta}}^K$ denote the best tree with K leaves, with respect to the log-likelihood (and $\boldsymbol{\theta}^{0,K}$ its corresponding limit), and $\widehat{K}$ denote the number of leaves of $\bar{\boldsymbol{\theta}}$.

Write

$$\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 = [\hat{\boldsymbol{\theta}}^{K^0} - \boldsymbol{\theta}^0]\mathbf{1}_{\widehat{K}=K^0} + \sum_{K\neq K^0}[\hat{\boldsymbol{\theta}}^K - \boldsymbol{\theta}^0]\mathbf{1}_{\widehat{K}=K}.$$

Let $R = \sum_{K\neq K^0}[\hat{\boldsymbol{\theta}}^K - \boldsymbol{\theta}^{0,K}]\mathbf{1}_{\widehat{K}=K}$, and note that

$$\mathbb{P}(R \geq t) \leq \mathbb{P}(\widehat{K} > K^0) + \mathbb{P}(\widehat{K} < K^0).$$

The result is then shown if we prove that $\mathbb{P}(\widehat{K} > K^0)$ and $\mathbb{P}(\widehat{K} < K^0)$ tend to zero when n tends to infinity, which is done below studying each probability separately.

First case: $\mathbb{P}(\widehat{K} > K^0)$.

We will use the notation $\mathcal{L}_n^K$ to denote the log-likelihood associated with $\hat{\boldsymbol{\theta}}^K$. If $\widehat{K} > K^0$, this means that there exists some $K^0 < K < K_{\max}$ such that

$$\mathcal{L}_n^K - \mathcal{L}_n^{K^0} \geq \lambda(K - K^0),$$

i.e.,

$$\mathcal{L}_n^K(\hat{\boldsymbol{\theta}}^K) - \mathcal{L}_n^K(\boldsymbol{\theta}^{0,K}) \geq \lambda(K - K^0),$$

since $\mathcal{L}_n^K(\boldsymbol{\theta}^0) = \mathcal{L}_n^{K^0}(\boldsymbol{\theta}^{0,K})$ for $K \geq K^0$. Whence,

$$\mathbb{P}(\widehat{K} > K^0) \leq \mathbb{P}(\exists\, K > K^0 : \mathcal{L}_n^K(\widehat{\boldsymbol{\theta}}^K) - \mathcal{L}_n^K(\boldsymbol{\theta}^{0,K}) \geq \lambda(K - K^0)).$$

Since $\lambda(K - K^0) \geq \lambda$, and since $\mathcal{L}_n^K(\widehat{\boldsymbol{\theta}}^K) - \mathcal{L}_n^K(\boldsymbol{\theta}^{0,K}) \leq \mathcal{L}_n^{K_{\max}}(\widehat{\boldsymbol{\theta}}^{K_{\max}}) - \mathcal{L}_n^{K_{\max}}(\boldsymbol{\theta}^{0,K_{\max}})$,

$$\mathbb{P}(\widehat{K} > K^0) \leq \mathbb{P}(\mathcal{L}_n^{K_{\max}}(\widehat{\boldsymbol{\theta}}^{K_{\max}}) - \mathcal{L}_n^{K_{\max}}(\boldsymbol{\theta}^{0,K_{\max}}) \geq \lambda). \tag{A7}$$

In the proof of Proposition 1, we showed that

$$\mathcal{L}_n^{K_{\max}}(\widehat{\boldsymbol{\theta}}^{K_{\max}}) - \mathcal{L}_n^{K_{\max}}(\boldsymbol{\theta}^{0,K}) = O_P([K_{\max}\log K_{\max}]^{1/2}\, n^{-1/2} + \varepsilon_n).$$

Hence, the right-hand side of (A7) tends to zero provided that $\lambda n^{1/2}[K_{\max}\log K_{\max}]^{-1/2} \to \infty$.

Second case: $\mathbb{P}(\widehat{K} < K^0)$.

In this case, $\mathcal{L}_n^K - \mathcal{L}_n^{K^0} \leq \mathcal{L}_n^{(K^0-1)} - \mathcal{L}_n^{K^0}$. From the proof of Proposition 1, $\mathcal{L}_n^{(K^0-1)} - \mathcal{L}^{*(K^0-1)} = O_P([K^*\log K^0]^{1/2}\, n^{-1/2})$, and $\mathcal{L}_n^{(K^0)} - \mathcal{L}^{*(K^0)} = O_P([K^*\log K^0]^{1/2}\, n^{-1/2})$. Then, similar to the first case,

$$\mathbb{P}(\widehat{K} < K^0) \leq \mathbb{P}\left(\mathcal{L}_n^{(K^0-1)} - \mathcal{L}^{*(K^0-1)} - \mathcal{L}_n^{K^0} + \mathcal{L}^{*K^0} \geq \frac{\lambda}{2}\right) + \mathbb{P}\left(\mathcal{L}^{*(K^0-1)} - \mathcal{L}^{*K^0} \geq \frac{\lambda}{2}\right).$$

The first probability tends to zero under the same conditions as in the first case, while the second is equal to $\mathbf{1}_{\mathcal{L}^{*(K^0-1)} - \mathcal{L}^{*(K^0)} \geq \lambda/2}$, since the quantity $\mathcal{L}^{*(K^0-1)} - \mathcal{L}^{*K^0}$ is deterministic. This indicator function tends to zero when n tends to infinity if λ tends to zero.

A.5 Convergence rate for the margins for kernel estimators

In this section, we show that Assumption 1 holds for the kernel estimator (4). This is in fact a consequence of Theorem 4 in [14]. We show the result under three additional assumptions on the model:

(1) the density of $\mathbf{X}$ is bounded away from zero on $\mathcal{X}$, i.e., $\inf_{\mathbf{x}\in\mathcal{X}} f_{\mathbf{X}}(\mathbf{x}) > 0$;

(2) we have

$$\sup_{\mathbf{x}\in\mathcal{X},y} \left| \frac{F^{(j)}(y)}{F^{(j)}(y|\mathbf{x})} + \frac{1 - F^{(j)}(y)}{1 - F^{(j)}(y|\mathbf{x})} \right| \leq \mathfrak{a},$$

for some finite constant $\mathfrak{A}$;

(3) the kernel function is a continuous and bounded function, symmetric around 0, such that $\int u^2 K(u)\mathrm{d}u < \infty$, the density $\mathbf{x} \mapsto f_{\mathbf{X}}(\mathbf{x})$ and $\mathbf{x} \mapsto F^{(j)}(t|\mathbf{x})$ are twice continuously differentiable with respect to $\mathbf{x}$, with uniformly bounded derivatives up to order 2.

The first assumption is required to avoid the denominator, in the kernel weights, going too close to zero. The second one is a way to consider that there is some kind of uniform domination of the behavior of the conditional distributions when $\mathbf{x}$ changes. Finally, the third assumption is classical in kernel regression and will help to control the bias term involved in smoothing techniques.

Introducing the kernel estimator of the density of $\mathbf{X}$,

$$\widehat{f}_X(\mathbf{x}) = \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right),$$

we can write, for $t \leq 1/2$,

$$\widehat{f}_X(\mathbf{x})\frac{\widehat{F}^{(j)}(t|\mathbf{x})}{[F^{(j)}(t|\mathbf{x})(1 - F^{(j)}(t|\mathbf{x}))]^a} = \frac{1}{nh^p}\sum_{i=1}^n K\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right) f_t(Y_i^{(j)}),$$

where

$$f_t(y) = \frac{\mathbf{1}_{y\le t}}{[F^{(j)}(t|\mathbf{x})(1-F^{(j)}(t|\mathbf{x}))]^\alpha} \le \frac{1}{[F^{(j)}(y|\mathbf{x})]^\alpha[1-F^{(j)}(1/2|\mathbf{x})]^\alpha} \le \frac{\mathfrak{A}^\alpha}{[F^{(j)}(y)]^\alpha[1-F^{(j)}(1/2|\mathbf{x})]^\alpha}.$$

Since

$$\mathbb{E}\left[\left(\frac{1}{[F^{(j)}(Y_i^{(j)})]^\alpha}\right)^p\right] < \infty,$$

for some $p > 2$ for $\alpha < 1/2$, and since the covering number of the class of functions f_t is controlled (see [45], Example 19.12), then Theorem 4 of [14] applies, showing that

$$\sup_{t\le 1/2,x}\left|\frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{\mathbf{X}_i-\mathbf{x}}{h}\right)f_t(Y_i^{(j)}) - \mathbb{E}\left[f_t(Y_i^{(j)})K\left(\frac{\mathbf{X}_i-\mathbf{x}}{h}\right)\right]\right| = O_P([\log n]^{1/2}\, n^{-1/2}h^{-d/2}).$$

Then, from a Taylor expansion and the third assumption of this section, we obtain

$$\mathbb{E}\left[f_t(Y_i^{(j)})K\left(\frac{\mathbf{X}_i-\mathbf{x}}{h}\right)\right] = \mathbb{E}[f_t(Y_i^{(j)})|\mathbf{X}_i=\mathbf{x}]f_{\mathbf{X}}(\mathbf{x}) + O(h^2).$$

Let us note that this h^2 rate can be improved if one uses a degenerate kernel with a sufficiently high number of moments equal to zero. Then, from the rate of uniform convergence of $\widehat{f}_X(\mathbf{x})$ (from Theorem 1, [14]), we obtain

$$\sup_{t\le 1/2,\mathbf{x}}\left|\frac{\widehat{F}^{(j)}(t|\mathbf{x}) - F^{(j)}(t|\mathbf{x})}{[F^{(j)}(t|\mathbf{x})(1-F^{(j)}(t|\mathbf{x}))]^\alpha}\right| = O_P(h^2 + [\log n]^{1/2}\, n^{-1/2}h^{-d/2}).$$

Studying the supremum for $t > 1/2$ can be done in the same way, by studying $1 - F^{(j)}$ instead of $F^{(j)}$.

A.6 Convergence rate for the margins for discrete covariates

For discrete covariates, recall that

$$\widehat{F}^{(j)}(t|\mathbf{x}) = \frac{\sum_{i=1}^n \mathbf{1}_{Y_i^{(j)}\le t}\mathbf{1}_{\mathbf{X}_i\in C(\mathbf{x})}}{\sum_{i=1}^n \mathbf{1}_{\mathbf{X}_i\in C(\mathbf{x})}}.$$

From the central limit theorem,

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\mathbf{X}_i\in C(\mathbf{x})} = \mathbb{P}(\mathbf{X}\in C(\mathbf{x})) + O_P(n^{-1/2}).$$

The upper part can be studied using similar arguments as [45, Example 19.12], noting that the class of functions $f_t(Y_i^{(j)})\mathbf{1}_{\mathbf{X}_i\in C(\mathbf{x})}$ (where f_t is defined in Section A.5) has a similar covering number as the class of functions f_t. This leads to

$$\sup_{t,\mathbf{x}}\left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{Y_i^{(j)}\le t}\mathbf{1}_{\mathbf{X}_i\in C(\mathbf{x})} - F^{(j)}(t|\mathbf{x})\mathbb{P}(\mathbf{X}\in C(\mathbf{x}))\right| = O_P(n^{-1/2}).$$

Then, we obtain

$$\sup_{t,\mathbf{x}}\left|\frac{\widehat{F}^{(j)}(t|\mathbf{x}) - F^{(j)}(t|\mathbf{x})}{[F^{(j)}(t|\mathbf{x})(1-F^{(j)}(t|\mathbf{x}))]^\alpha}\right| = O_P(n^{-1/2}).$$

A.7 Regression trees for margin estimation in the real data example

We report here the regression trees resulting from fitting the variables $(Y^{(1)}, Y^{(1)})$ as a function of the covariates year and influenza transmission zone (Section 4.2.1).

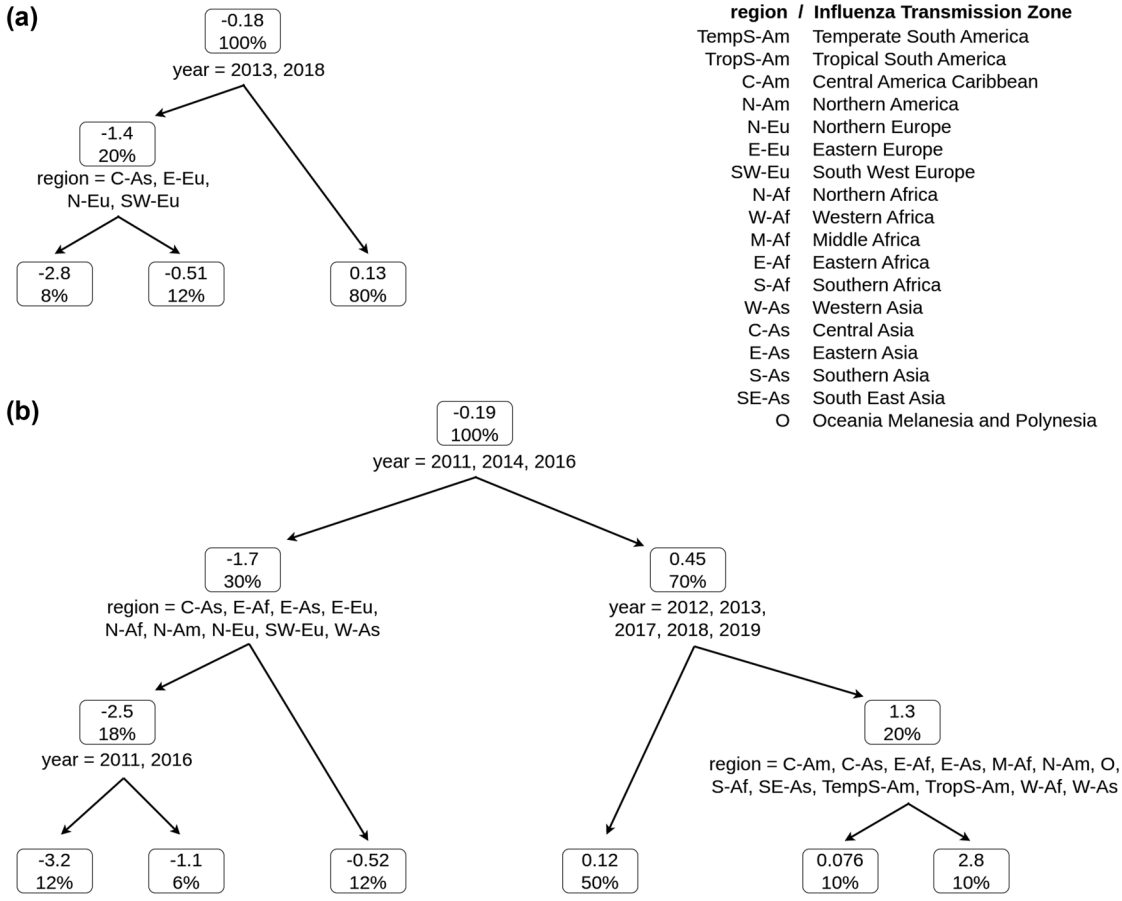

Figure A1: Optimal trees for margin estimation. Country and ITZs are classified by the regression trees to approximate the response variables $Y^{(1)}$ (plot a) and $Y^{(2)}$ (plot b). The coefficients of determination of the two fits are 0.29 and 0.5, respectively. For each node, the average value of the response variable and the percentage of the observations included are indicated. In the top-right corner, a legend illustrates the abbreviations used for the ITZs.

References

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

[2] Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(2), 139–160.

[3] Alquier, P., Chérief-Abdellatif, B.-E., Derumigny, A., & Fermanian, J.-D. (2023). Estimation of copulas via maximum mean discrepancy. *Journal of the American Statistical Association*, *118*(543), 1997–2012.

[4] Bahl, J., Nelson, M. I., Chan, K. H., Chen, R., Vijaykrishna, D., Halpin, R. A., et al. (2011). Temporally structured metapopulation dynamics and persistence of influenza a h3n2 virus in humans. *Proceedings of the National Academy of Sciences*, *108*(48), 19359–19364.

[5] Bedford, T., Riley, S., Barr, I. G., Broor, S., Chadha, M., Cox, N. J., et al. (2015). Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, *523*(7559), 217–220.

[6] Li, B., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees (CART). *Biometrics*, *40*(3), 358–361.

[7] Buccianti, A., Lima, A., Albanese, S., Cannatelli, C., Esposito, R., and De Vivo, B. (2015). Exploring topsoil geochemistry from the coda (compositional data analysis) perspective: The multi-element data archive of the campania region (southern italy). *Journal of Geochemical Exploration*, *159*, 302–316.

[8] Charpentier, A., Fermanian, J.-D., & Scaillet, O. (2007). The estimation of copulas: Theory and practice. *Copulas: From Theory to Application in Finance* (pp. 35–64).

[9] Czado, C., & Nagler, T. (2022). Vine copula based modeling. *Annual Review of Statistics and Its Application*, *9*, 453–477.

[10] Derumigny, A., & Fermanian, J.-D. (2017). About tests of the "simplifying" assumption for conditional copulas. *Dependence Modeling*, *5*(1), 154–197.

[11] Derumigny, A., Fermanian, J.-D., & Min, A. (2023). Testing for equality between conditional copulas given discretized conditioning events. *Canadian Journal of Statistics*, *51*(4), 1084–1110.

[12] Dupuis, D. J., & Jones, B. L. (2006). Multivariate extreme value theory and its usefulness in understanding risk. *North American Actuarial Journal*, *10*(4), 1–27.

[13] Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, *35*(3), 279–300.

[14] Einmahl, U., & Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, *33*(3), 1380–1403.

[15] Farkas, S., Heranval, A., Lopez, O., & Thomas, M. (2021). *Generalized pareto regression trees for extreme events analysis*. arXiv: http://arXiv.org/abs/arXiv:2112.10409.

[16] Farkas, S., & Lopez, O. (2024). Semiparametric copula models applied to the decomposition of claim amounts. *Scandinavian Actuarial Journal*, *2024*(10), 1065–1092.

[17] Farkas, S., Lopez, O., & Thomas, M. (2021). Cyber claim analysis using generalized pareto regression trees with applications to insurance. *Insurance: Mathematics and Economics*, *98*, 92–105.

[18] Fermanian, J.-D., & Lopez, O. (2018). Single-index copulas. *Journal of Multivariate Analysis*, *165*, 27–55.

[19] Filzmoser, P., Hron, K., & Reimann, C. (2009). Univariate statistical analysis of environmental (compositional) data: problems and possibilities. *Science of the Total Environment*, *407*(23), 6100–6108.

[20] Flahault, A., Dias-Ferrao, V., Chaberty, P., Esteves, K., Valleron, A.-J., & Lavanchy, D. (1998). Flunet as a tool for global monitoring of influenza on the web. *Jama*, *280*(15), 1330–1332.

[21] Genest, C., Ghoudi, K., & Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, *82*(3), 543–552.

[22] Gey, S., & Nedelec, E. (2003). Risk bounds for cart regression trees. In: *Nonlinear estimation and classification* (pp. 369–379). Springer.

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

[24] Gijbels, I., Veraverbeke, N., & Omelka, M. (2011). Conditional copulas, association measures and their applications. *Computational Statistics & Data Analysis*, *55*(5), 1919–1932.

[25] GISRS. (2022). *FluNet Database - National Influenza Centres of the Global Influenza Surveillance and Response System and World Health Organisation*. World Health Organization.

[26] Gocheva-Ilieva, S. G., Voynikova, D. S., Stoimenova, M. P., Ivanov, A. V., & Iliev, I. P. (2019). Regression trees modeling of time series for air pollution analysis and forecasting. *Neural Computing and Applications*, *31*, 9023–9039.

[27] Hastie, T. J., & Pregibon, D. (2017). Generalized linear models. *In Statistical models in S* (pp. 195–247). Routledge.

[28] Hennessy, D. A., & Lapan, H. E. (2002). The use of archimedean copulas to model portfolio allocations. *Mathematical Finance*, *12*(2), 143–154.

[29] Jackson, D. A. (1997). Compositional data in community ecology: the paradigm or peril of proportions? *Ecology*, *78*(3), 929–940.

[30] Jaworski, P., Durante, F., Hardle, W. K., & Rychlik, T. (2010). *Copula theory and its applications*, vol. 198. Springer.

[31] Kularatne, T. D., Li, J., & Pitt, D. (2021). On the use of Archimedean copulas for insurance modelling. *Annals of Actuarial Science*, *15*(1), 57–81.

[32] Kurz, M. S., & Spanhel, F. (2022). Testing the simplifying assumption in high-dimensional vine copulas. *Electronic Journal of Statistics*, *16*(2), 5226–5276.

[33] Le, M. Q., Lam, H. M., Cuong, V. D., Lam, T. T.-Y., Halpin, R. A., Wentworth, D. E., et al. (2013). Migration and persistence of human influenza a viruses, Vietnam, 2001–2008. *Emerging Infectious Diseases*, *19*(11), 1756.

[34] Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., et al. (2014). Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza h3n2. *PLoS pathogens*, *10*(2), e1003932.

[35] Loh, W.-Y. (2014). Fifty years of classification and regression trees. *International Statistical Review*, *82*(3), 329–348.

[36] Lopez, O. (2019). A censored copula model for micro-level claim reserving. *Insurance: Mathematics and Economics*, *87*, 1–14.

[37] Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *135*(3), 370–384.

[38] Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

[39] Omelka, M., Hudecová, SSS., & Neumeyer, N. (2021). Maximum pseudo-likelihood estimation based on estimated residuals in copula semiparametric models. *Scandinavian Journal of Statistics*, *48*(4), 1433–1473.

[40] Segers, J. (2012). Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli* (pp. 764–782).

[41] Sklar, M. (1959). Fonctions de répartition àn dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, *8*, 229–231.
[42] Therneau, T., & Atkinson, E. (1997). An introduction to recursive partitioning using the RPART routines. *Mayo Clinic*, *61*, 452.
[43] Tsukahara, H. (2005). Semiparametric estimation in copula models. *Canadian Journal of Statistics*, *33*(3), 357–375.
[44] van der Vaart, A., & Wellner, J. A. (2023). Empirical processes. *In Weak Convergence and Empirical Processes: With Applications to Statistics* (pp. 127–384). Springer.
[45] van der Vaart, A. W. (2000). *Asymptotic Statistics* (vol. 3). Cambridge University Press.
[46] Wellner, J. A. (1978). Limit theorems for the ratio of the empirical distribution function to the true distribution function. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, *45*(1), 73–88.
[47] W.H.O. (2018). Influenza transmission zones. World Health Organization.