IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 5, 2020, accepted August 14, 2020, date of publication August 18, 2020, date of current version August 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3017529

Binary Trees for Dependence Structure

QINGSONG SHAN AND QIANNING LIU
Department of Statistics, Jiangxi University of Finance and Economics, Nanchang 330013, China
Corresponding author: Qianning Liu (qianningliu@outlook.com)

This work was supported in part by the Education Department of Jiangxi Province under Grant GJJ190253 and Grant GJJ190259.

ABSTRACT In a data set with many categorical variables and several continuous valuables, the relationship between continuous random variables may differ from category to category for a given categorical variable. To study how categorical variables may affect the dependent structure of continuous variables, we proposed two splitting criteria constructed based on copula entropy to build decision trees serving for different purposes. One type of tree can be used to identify the attributes or combinations of them under which the continuous variables have a strong relationship. The other type of tree is used to classify regions with different strength of relationship. Applying these methods to the survey data on the status of poor families of Sichuan province, it is found that the method successfully evaluated the effectiveness of the poverty alleviation policies.

INDEX TERMS Copula, dependence structure, decision trees, entropy.

I. INTRODUCTION AND MOTIVATION

In 2018, the Chinese government launched a survey on the status of poor households after getting rid of poverty. On the one hand, the purpose of the survey is to understand the current living conditions of those families who have been lifted out of poverty with the help of poverty alleviation policies. On the other hand, it is also hoped that the survey can lead to future policy formulation. The survey is very extensive, involving more than a dozen provinces in the country. This article only takes Sichuan Province as an example.

The survey was conducted by questionnaire. To get the data, instead of sampling, commissioners were sent to conduct a household-by-household survey, with a total of 11,329 households surveyed. There are about 180 questions in the questionnaire, which not only cover all aspects of poor family life: housing, diet, children's schooling, medical care and work, but also include many questions about whether they are satisfied with various policies. For example, one question is whether attending employment training is helpful to increase income. Another question is whether the industrial poverty alleviation policy is helpful to increase income. This kind of question usually only needs to answer "yes" or "no". The questionnaire investigated the income of each family in detail. Under normal circumstances, family income includes three sources: wage income, operational income and state subsidies, as well as transfer income from relatives, friends. Since the main sources of household income are wage income and operational income, we only consider these two items in the following analysis.

One of the purposes of the paper is to figure out how much effective these policies are in improving family income and how to evaluate the roles of the policies. The difficulty is that without historical data, quantifying the changes of incomes caused by policies will not be possible. In this article, we will try to evaluate the policies through people's answers and rank them in the order of importance from the questions asking whether a policy is helpful.

Let X and Y be the two sources of incomes and questions are called attributes. Then the questionnaire can be abstracted into a data set with two objective continuous random variables and some categorical attributes, which is shown in table 1.

TABLE 1. Sample data.

Attribute A	Attribute B	Attribute C	$\cdots$	X	Y
Y	N	Y	$\cdots$	x_1	y_1
N	Y	N	$\cdots$	x_2	y_2
Y	Y	N	$\cdots$	x_3	y_3

Some of these questions are targeted. For example, when asked whether the industrial poverty alleviation policy is helpful, it is aimed at operational income. Whether the employment poverty alleviation policy is helpful is aimed at wage income. Other questions, such as whether have loan

The associate editor coordinating the review of this manuscript and approving it for publication was Shuihua Wang.

or not, do not target any specific source of income. For the questions, which have specific target income, one typical way of analyzing the effectiveness of these policies is to take one attribute and divide the population into groups with positive answers (who think the policy is helpful in improving income) and negative answers (who think the policy is not helpful), then compare the difference in median of incomes of the two groups. There are two issues we need to consider for this method. First, this method works well for a single attribute. When considering a combination of questions, the number of groups will increase exponentially. For example, a combination of 3 questions will generate $2^3 = 8$ subgroups. As the number of questions increase, the calculation burden will increase rapidly. Second, this method is not suitable for studying the relationship between random variables, which should be taken into account in our case.

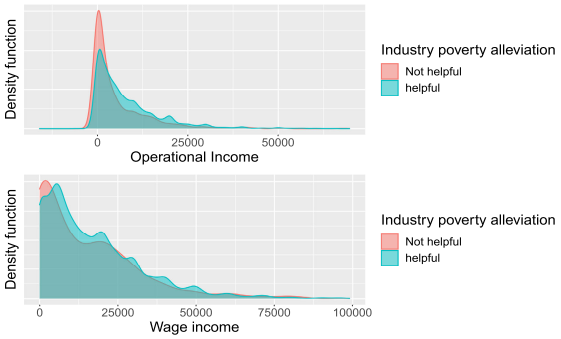


FIGURE 1. The distributions of household income of different respondents to the question "Is employment poverty alleviation policies helpful?".

We will next make a comparison of the effects of the two policies with relatively clear pertinency on wage income and operational income, that are, employment poverty alleviation policy and industry poverty alleviation policy. The first step is to divide all families into two categories based on the criteria of whether employment poverty reduction policy helps or not and compare the difference of wage income and operational income between the two categories of families. Figure 1 shows the distribution curves about the two answers whether employment poverty reduction policy helps or not of all the incomes estimated by nonparametric kernel method. Figure 2 shows the density function curve of each income under the answers of whether industry poverty alleviation is helpful. Both figures showed that the families which answered that industry poverty alleviation is helpless (red) are with relatively low income. But this positive relationship between income and answers does not simply implies the effectiveness of policies. There are two possibilities. First, the policies are effective, so people answered "helpful", in the meanwhile the effective policies raised people's income. Second, people with high income tend to answer "helpful", while people with low income tend to answer "not". Of course, it might be a mixture of the above two reasons. For the second case, if we assume the tendency for different questions are the same, then we can compare the effectiveness of policies by comparing the changes of incomes under different questions. Next, we will elaborate the method.

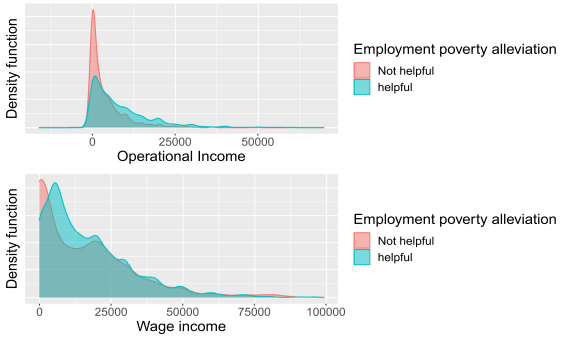


FIGURE 2. The distributions of household income of different respondents to the question "Is industrial poverty alleviation policies helpful?".

Since the distributions are highly skewed, to quantify the difference between two categories, we use Wilcoxon test to test the differences of the medians of two categories. More specifically, let X be the operational income of families which answered employment poverty reduction policy helps, Y be the operational income of families which answered employment poverty reduction policy not helpful. M_X and M_Y are the medians of X and Y respectively. We then construct 95% confidence interval for $M_X - M_Y$. We do the same calculation for operational income of the families that answered whether industry poverty alleviation helps, as well as the wage income, so that a total of four confidence intervals are obtained and listed in table 2.

TABLE 2. Confidence intervals for the increase of incomes under different policies.

	95% Confidence Interval for $M_x - M_y$	
	Employment poverty alleviation	Industry poverty alleviation
Operational income	$(2460, 3000)$	$(1500, 2200)$
Wage income	$(1500, 3200)$	$(1000, 3000)$

Here we use the median number of the groups that hold the industry poverty alleviation is helpful minus the median number of the groups who hold it unhelpful, thus the received positive value can be used to measure the effectiveness of a policy. On the main diagonal of the matrix formed by two sets of confidence intervals, it is not surprisingly to see that the employment poverty alleviation raises the wage income and the industry poverty alleviation raises the operational income. The interesting thing is that the vice diagonal is not only non-zero, whose value is relatively large instead. The confidence interval of the 95% of the income difference between operational incomes under the different answers about employment poverty alleviation in Table 2 is (2460,3000), while the 95% of the income difference between operational incomes under the different answers about industry poverty alleviation (1500,2200). In other words, the industry poverty alleviation policy does not increase more operational income

than the employment poverty alleviation does operational income. This is clearly against common sense. Some plausible explanations to this may include: first, the information is inaccurate. For example, the responses from respondents were affected, resulting in responses not reflecting the truth. Second, there is an integrated effect between policies and incomes. For example, when employment alleviation policies work better, there may be a shift of family members from originally being engaged in business activities to obtaining employment, which can lead to a decrease in household industrial income, or a significant difference in the effects of policies within different groups of people. Here, we make reason two the focus of our study, that is, the interaction among policies.

As can be seen from the above analysis, even when considering the effect of policies generally considered more targeted, such as industry poverty alleviation aimed at operational income and employment poverty alleviation aimed at wage income, it cannot be limited to one kind of income. Therefore, when analyzing the policies involved in the survey, it cannot and there is no way for it to be analyzed separately which must be considered in a comprehensive manner. Below, we consider the impact of each policy on operational income and wage income as a whole. We take the employment poverty alleviation policy as an example, the employment poverty alleviation policy may have influence on wage income or operational income (the effect of which may not be direct or positive, as in the previous case of a career shift leading to reduced operational income). When both effects exist, the relationship between wage income and operational income will change. Below we will focus on the impact of policies on the relationship between wage income and operational income.

To study relationship between two random variables, copula is a powerful tool. Since they are not affected by marginal distribution so that we can focus on the dependence structure between random variables. We will consider how the copulas between X and Y will be affected by different combinations of attributes. The method we proposed is similar to a decision tree, but the difference is that this time the target variable is copulas or relationship between random variables, not a single variable.

The first difficulty of this procedure is about estimating copula functions. Considering the fact that usually people have almost no knowledge about the relationship between two sources of incomes, nonparametric estimation of copulas is a reasonable choice. Generally speaking, there are two types of methods: empirical method and smoothing method. The empirical method was introduced by [1]. Then [2], [3] proved the consistency of empirical copula process for copulas with continuous partial derivatives. This method is robust, but it cannot be used straightforwardly to derive an estimate of the copula density. Various smoothing methods have been provided. [4] suggested to use kernel methods to estimate copula densities, which was further discussed in [2] and [5]. After that, various modern smoothing techniques was applied to the estimation of copulas, e.g. beta kernel method was introduced by [6] to remedy boundary bias, wavelet based estimation was suggested by [7] for copula estimation and [8] for copula density estimation, Bernsein polynomials studied by [9], [10] and [11], B-splines studied by [12], [13] claimed that using penalized hierarchical B-spline together with sparse grids can weaken the curse of dimensionality. In this article, we will adopt the transformation method, which was introduced to kernel copula density estimation by [6], since [14] has shown that this estimator outperforms other estimators.

Due to the simplicity and flexibility in handling both real-valued and categorical features, tree-based methods have been used in a broad range of areas including machine learning, engineering, finance and business. A key step in the process of constructing a decision tree is to choose a split criterion. Some classic split criteria are: ID3 [15], C4.5 [16], and CART [17]. The selection and construction of split criteria is always a hot topic. [18] built a decision tree based on Pearson correlation coefficient. Deng entropy was used as a measure of splitting rules to construct a decision tree for fuzzy data set classification [19]. Both articles used relation-based measures as the splitting criteria. Compared with other measures, copula entropy is a more general measure of dependence [20]–[24], which makes it an ideal tool in studying relationship. Besides the different choice of the measures used in constructing the splitting criteria, the main difference between this article and the other two is the research object. In those two articles, the authors creatively constructed two splitting criteria to improve classification accuracy. The two split criteria constructed in this article are both used to study the relationship between variables.

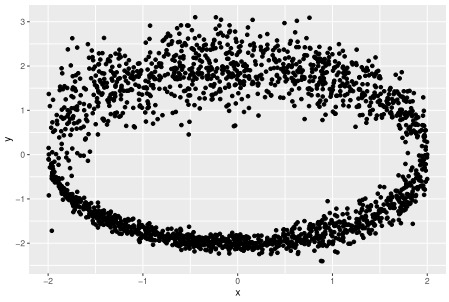


FIGURE 3. A circular data set with changing variance.

II. DECISION TREE FOR RELATIONSHIP

To illustrate the idea, let's consider a sample of 2000 points randomly spread on a circle of radius 2 with changing variance as is shown in Figure 3. Three factors A, B and C split the data with $x = 0$, $y = 0$ and $x^2 + y^2 = 4$, respectively. Denote $A_1 = \{(x, y)|x \leqslant 0\}$, $A_2 = \{(x, y)|x > 0\}$, $B_1 = \{(x, y)|y \leqslant 0\}$, $B_2 = \{(x, y)|y > 0\}$, $C_1 = \{(x, y)|x^2 + y^2 \leqslant 4\}$, $C_2 = \{x^2 + y^2 > 4\}$. Points in the third quadrant has constant variance 0.1, then variance will gradually increase anticlockwise, until it reaches its maximum 0.6 on the second

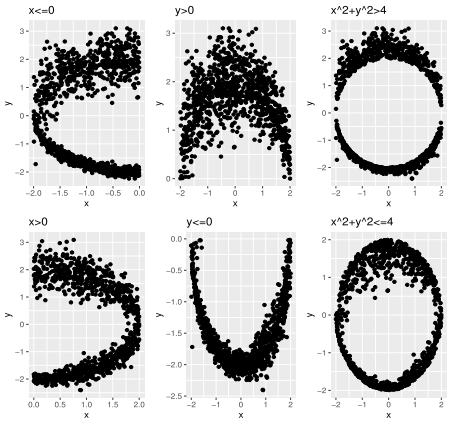


FIGURE 4. Scatterplot of subsets split by the three attributes.

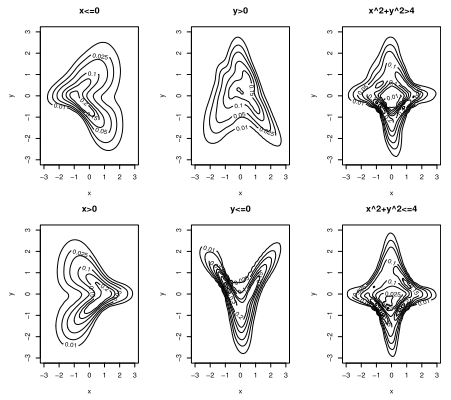


FIGURE 5. Copulas for each subset.

quadrant. Scatterplot of subsets split by the three factors are shown in Figure 4, their corresponding copulas are in Figure 5.

As we can see from Figure 5, different ways of splitting generate subsets with different copulas, in other words, splitting changes relationship, not only in terms of structure but also in strength. To quantify the difference between copulas, we will use copula entropy.

Definition 1: Let $\boldsymbol{x} \in R^N$ be random variables with marginal functions $(F_1, \cdots, F_N)$ and copula density $c(\boldsymbol{u})$. Copula entropy (CE) of $\boldsymbol{c}$ is defined as:

$$\mathrm{CE}(c) = -\int_u c(\boldsymbol{u}) \log c(\boldsymbol{u}) d\boldsymbol{u}.$$

Entropy measures uncertainty of a random variable. For a discrete probability distribution p, the entropy of p is non-negative, and it equals 0 if and only if the random variable is deterministic, but this property generally not hold for continuous distributions. So how about copula entropy? It has been proved that copula entropy is negative mutual information [25]. Therefore, copula entropy inherits the same properties of mutual information. Copula entropy is considered to be the most general measure of the statistical dependence between two random variables [24]. The range of copula entropy is given in the following theorem.

Theorem 1: $\mathrm{CE}(c) \leqslant 0$ with equality if and only if $c(u, v) = 1$.

The proof of the theorem is a direct consequence of the theorem 1 in [25]. Similar to mutual information, we can think of CE(c) as a KL distance from any given copula $c(u, v)$ to the independent copula.

To use copula entropy as a splitting criterion, an estimator from sample data is needed. Since copula entropy and mutual information are essentially the same, we will briefly introduce the estimation of mutual information. One popular class of estimators of mutual information are based on the k-nearest neighbor (k-NN) [26]–[29]. These approaches require number of samples scales exponentially with the value of mutual information [30]. [31] gave a comprehensive review of entropy, mutual information and their estimations. Our mutual information estimator is based on a recently developed nonparametric copula density estimating method. The usual nonparametric smoothing methods suffered from boundary problems not only because the method itself but also because the transformation from joint distribution to copulas will magnify small fluctuations on the boundaries. The remedy proposed by [32], [33] is called probit transformation which transforms the support of a copula into unbounded space, such that usual smoothing methods can be applied without causing any boundary issues. [14] compared the performance of different estimators in different scenarios and found that the transformation local polynomial estimator using nearest-neighbor bandwidths outperform other estimators in mean integrated absolute error. [32] claimed this estimator is accurate and robust to changes of the marginal distributions. So, we will use this estimator as a splitting criterion to construct decision trees. Two splitting criteria serving different purposes will be provided in the next section.

III. TWO TYPES OF TREES

Splitting the data set in different ways may generate subsets with different strength of dependence. The number of subsets will increase exponentially as the number of attributes increases. Consider each possible combination of subsets will dramatically increase calculation burden. We introduce two algorithms used to generate combinations of subsets serving different purposes.

A. REGRESSION TREES

1) METHODOLOGY

Copula entropy can measure the strength of both linear and nonlinear relationship. We will use this property to find the regions where the variables have a strong linear or non-linear relationship. More specifically, the goal is to find regions $R_1, \cdots, R_J$ that minimize the RSS, given by

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2,$$

where $\hat{y}_{R_j}$ is the predicted value of non-parametric regression for the training observations within the jth region.

The two main differences between the problem addressed here and a typical regression tree are: 1. For a regression tree, $\hat{y}_{R_j}$ is the mean response, which is only determined by the region R_j. In our case, different regression methods will produce different $\hat{y}_{R_j}$, therefore we stick on local polynomial regression (LOESS) in this article. 2. The attributes, which are used to divide the predictor space, are given, so that we only need to find the "best" combinations of those attributes. Nevertheless, as the number of feature increases, it is still computationally expensive. So, we take a greedy approach that is known as recursive binary splitting.

For a given set S, attribute A, with two classes a_1 and a_2, splits S into S_1 and S_2. The corresponding copulas associated with sets are denoted by c, c_1 and c_2, respectively. We define copula entropy for the level split by attribute A to be $\sum \frac{|S_i|}{|S|} \mathrm{CE}(c_i)$, where $\mathrm{CE}(c_i)$ is the entropy of copula c_i associated with set S_i. The information gain (InfGain) is defined as:

$$\mathrm{InfGain}(A, S) = \mathrm{CE}(c) - \sum_i \frac{|S_i|}{|S|} \mathrm{CE}(c_i).$$

A higher InfGain indicates a higher reduction in copula entropy, so the attribute with highest InfGain will be chosen at every splitting. By doing this recursively, we get a combination of attributes which have stronger dependent structure than the original data set.

To control the tree size, we could set a threshold for InfGain. The tree will stop growing when the InfGain reaches the threshold. To avoid short-sighted splitting, tree pruning methods can also be adopted by choosing the best λ in the following sequence of trees

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 + \lambda |\,\mathrm{nodes}\,|,$$

where λ is a nonnegative tuning parameter, $|\,\mathrm{nodes}\,|$ is the number of terminal nodes of the tree. The difference between this method and the usual cost complexity pruning is that the region R_j is a region divided by combinations of attributes. Therefore, these regions can be of any shape, not just rectangular.

2) SIMULATION RESULTS

To apply the method to the circular data set, we first calculate InfGain for each attribute. The values of the three attributes A, B and C are -0.07, 0.14 and 0.2 respectively. Attribute C was chosen to split the data since it has the biggest InfGain. Then calculate the InfGain of the remaining two attributes for the subsets where the attribute C is true and false, and select the largest one as the split attribute of the subset. By doing this recursively, we end up with a tree in Figure 6. Attribute A did not participate in splitting on each leaf because its InfGain is negligible or even negative. When an attribute has a positive InfGain, it means that using this attribute to split the data set will make the correlation stronger after the split. Therefore, the correlation of the data set on the leaves is stronger than the original set.

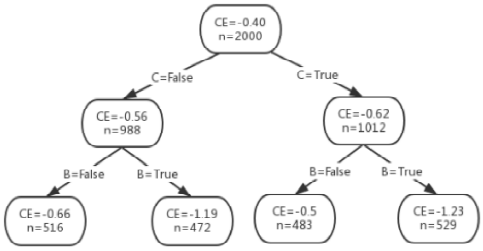


FIGURE 6. The regression tree for the circular data set.

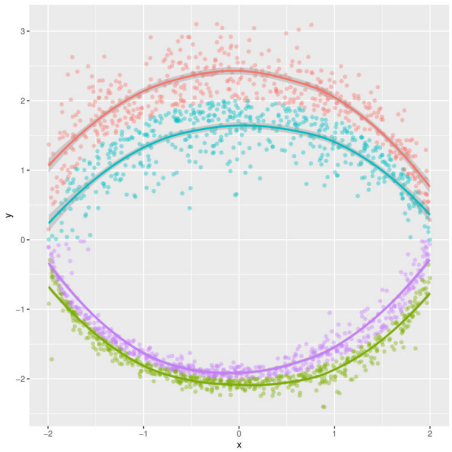


FIGURE 7. The four-region partition of the data set with LOESS smoothing from the regression tree illustrated in Figure 6.

The leaves of the tree in Figure 6 stratify the data set into four regions marked with different colors in Figure 7. For each region, a local polynomial regression (LOESS) was fitted into the data. As a comparison, the LOESS model and the random forest (RF) regression were constructed, and the results are shown in Figure 8 and 9. As can be seen, the models in Figure 7 fit the data much better than the other two models for the whole data set. We conducted a simulation to evaluate the performance of different models for different sample sizes. The MSE of each model was estimated using 10-fold cross-validation in Table 3, where the four regions were represented by numbers from 1 to 4. In each region, we used both LOESS and RF regression to fit the data, which are represented by "Tree + LOESS" and "Tree + RF regression", respectively. The model "tree + LOESS" has the best performance. Notice that RF regression takes categorical variables into account, so it outperforms LOESS for the whole data set, but dividing the data into several regions can still improve the accuracy. If we take the attributes as auxiliary variables, then in the situations where the interest is to predict y values, we can get more accurate estimation by dividing the regions according to the leaves of the tree.

One question raised is whether it will help improve the accuracy of the estimation if we use all possible attribute combinations to divide the entire area to the smallest extent. Table 4 showed the estimated MSE of both models by using

TABLE 3. Estimated MSE using 10-fold cross validation for different models.

		MSE		
Model	Region	$n = 2000$	$n = 5000$	$n = 10000$
Tree + LOESS	1	0.087	0.079	0.080
	2	0.077	0.070	0.066
	3	0.014	0.015	0.015
	4	0.022	0.019	0.016
	Overall	0.05	0.046	0.044
LOESS		2.808	2.730	2.734
Tree + RF regression	1	0.231	0.205	0.223
	2	0.171	0.186	0.180
	3	0.133	0.133	0.140
	4	0.196	0.184	0.177
	Overall	0.183	0.177	0.18
RF regression		0.581	0.560	0.577

TABLE 4. Estimated MSE of regression models on the smallest sets.

	Tree + LOESS	Tree + RF regression
Region 1	0.097	0.207
Region 2	0.053	0.189
Region 3	0.089	0.193
Region 4	0.047	0.161
Region 5	0.005	0.141
Region 6	0.021	0.117
Region 7	0.016	0.165
Region 8	0.018	0.178
overall	0.043	0.169

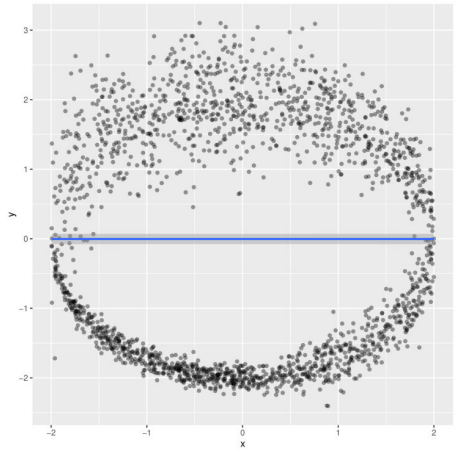


FIGURE 8. LOESS regression for all points.

10-fold cross-validation. As can be seen, the improvement is negligible.

B. CLUSTERING TREES

In this section, we introduce another splitting criterion which choose most ''suitable'' combination of subsets and rank them in an order from strongest relationship to weakest relationship. Recall that in the example of introduction, people are interested in which attribute plays more important role in changing the relationship of two incomes. For example, if the dependence structure between two incomes of the families which think one particular policy is not helpful is significantly different from the dependence structure of the families who think the policy is helpful. Then we will think the policy is effective. So, the purpose of this section will

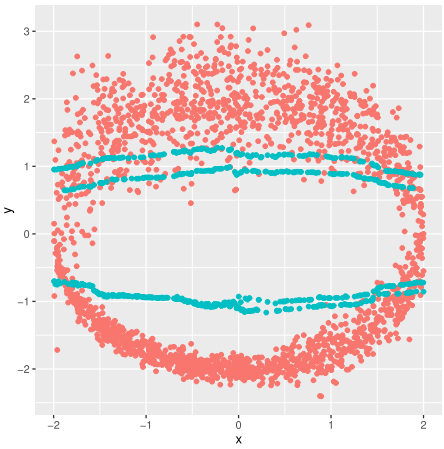


FIGURE 9. Random forest regression (blue dots).

Algorithm 1 A Splitting Criterion for Regression Trees

Input Data: $Attribute_1, \cdots, Attribute_N, X, Y$
Output: Best feature
for $i = 1$ to N **do**
  Split the data set into $subset_{i1}$ and $subset_{i2}$ according to $Attribute_i$
  Estimate copula: $copula_i$,$copula_{i1}$ and $copula_{i2}$
  Estimate copula entropy: CE_i,CE_{i1} and CE_{i2}
  Compute $InfGain_i$ for $Attribute_i$
Best feature $\leftarrow \arg\max_i InfGain_i$

be quantifying the ''changes'' of the relationship caused by attributes and rank them. There are several measures can be used for quantifying the difference between copula densities, i.e. integrated absolute deviation, integrated squared deviation or KL divergence. Since we are not only interesting measuring the similarity of copula densities but also like to rank the copulas according to their strength of dependence, we will construct a quantity based on copula entropy.

By looking at the shape of copulas in figure 5, we noticed that the pair of copulas for attribute B has the ''biggest'' difference among all three pairs. To quantify the difference between each pair of copulas, we would like to use the following transformation to ''normalize'' copula entropy (NCE):

$$\mathrm{NCE}(x) = 1 - e^x,$$

in which x is a given copula entropy. After standardization, NCE has range $[0, 1]$. Then for a given attribute A with categories A_1 and A_2, the quantity measures the difference between the dependence structure of subsets associated with attribute A is defined by

$$\mathrm{DNCE}(A) = |\,\mathrm{NCE}(\mathrm{CE}(A_1)) - \mathrm{NCE}(\mathrm{CE}(A_2))|.$$

Table 5 shows the NCE and DNCE for each subset.

As expected, the pair of copulas split by B has biggest difference in NCE among all three factors. So, in the procedure of constructing a decision tree, attribute B will be the root

TABLE 5. NCE and DNCE for each attribute.

Attribute	A_1	A_2	B_1	B_2	C_1	C_2
NCE	0.25	0.30	0.24	0.54	0.43	0.46
DNCE	0.05		0.3		0.03	

node of the tree. Then calculate the DNCE of the remaining two attributes for the subsets where the attribute B is true and false, and select the largest one as the split attribute of the subset. By recursively repeating this strategy, we end up with a decision tree in Figure 10. There are two ways to interpret the tree:

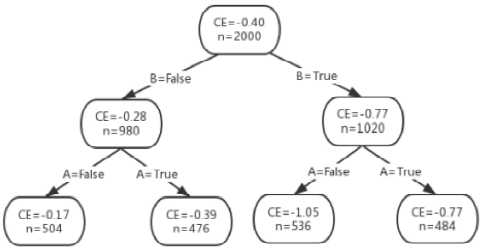


FIGURE 10. The clustering tree for the circular data.

1. The more obvious discriminations between subsets are done first, and the more subtle distinctions are postponed to lower levels. As we can see from Figure 3, the upper semicircle has larger variance than the lower half, which makes the difference between this pair larger than the other two pairs. This indicates attribute B plays the most important role in changing the relationship. Attribute C split the data into inner circle and outer circle, which didn't change much in either the shape or the variance. This explains why it caused the least difference. To determine the children node for the lower semicircle, attribute A causes much larger difference in NCE than attribute C, because when splitting by circle (attribute C), the change of variance is not much different, while left half has constant variance, which is smaller than the variance of right half. So, the lower semicircle should by split by attribute A. For the upper semicircle, the differences are not obvious, so we leave it to DNCE to decide.

2. Each branch of the tree suggests a combination of the attributes, which divide the data into four subsets. The subsets and their corresponding copula functions are shown in Figure 11. This splitting method suggest a way to divide the data set into regions with significant difference in dependence. Further, we can rank them in order.

Note that whether the tree can be constructed depends strongly on the properties of the attributes. For example, if the data set contains only two factors A and C, then $|\,\mathrm{NCE}_{A_1} - \mathrm{NCE}_{A_2}\,| = 0.05$ and $|\,\mathrm{NCE}_{C_1} - \mathrm{NCE}_{C_2}\,| = 0.03$, then neither will be selected for splitting. If it is so, we can conclude that splitting will not change the relationship.

For the stopping criterion, we should take into account both a threshold of $|\,\mathrm{NCE}_{C_1} - \mathrm{NCE}_{C_2}\,|$ and minimum sample size in each node. Because too few points will lead to inaccurate estimation of copula density. According our experience, 200 is the minimum number of points needed for a steady

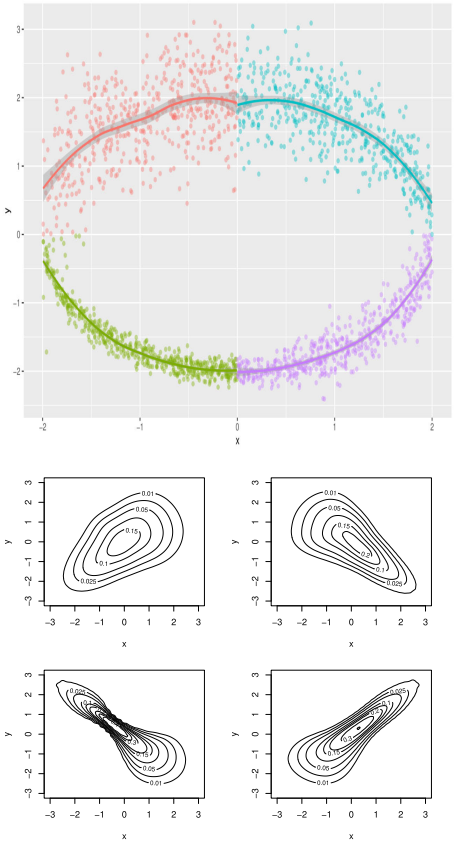


FIGURE 11. A partition suggested by leaves and their corresponding copulas.

Algorithm 2 A Splitting Criterion for Clustering Trees

Input Data: $Attribute_1, \cdots, Attribute_N, X, Y$
Output: Best feature
for $i = 1$ to N **do**
  Split the data set into $subset_{i1}$ and $subset_{i2}$ according to $Attribute_i$
  Estimate copula: $copula_{i1}$ and $copula_{i2}$
  Estimate copula entropy: CE_{i1} and CE_{i2}
  Normalize copula entropy: NCE_{i1} and NCE_{i2}
  $DNCE_i \leftarrow |NCE_{i1} - NCE_{i2}|$
The best splitting feature $\leftarrow \arg\max_i DNCE_i$

estimation. Note that if we grow a full tree with n attributes, which will generate 2^n subsets, then prune it back. For each subset, there corresponds an NCE value, then grouping these NCE values is essentially hierarchical clustering, so some typical clustering algorithms, e.g. k-mean clustering, can be used to prune the tree.

One main difference between the two types of trees is that a regression tree will split only when the weighted sum of copula entropy increases, so it will find the attributes with stronger relationship in a general manner. While a clustering tree is clustering the regions according to the strength of relationship. So, the set of points on leaves are arranged in order of correlation from strong to weak.

C. REAL DATA EXAMPLE

Let's go back to the Sichuan data. The data contains some questions about whether a specific policy is helpful. The answer is either "yes" or "no". As we have discussed earlier, even policies targeting one source of income may have an impact on another source of income. Therefore, these policies are likely to have an impact on the correlation or related structure between the two sources of income. Next, we will divide the population into subgroups according to people's answer to each question, then apply the above two methods.

We are interested in the effectiveness of policies, so the tree size is controlled so that only policies concerning income improving will be included. The questions, which will be used later, are represented by letters in Table 3.

TABLE 6. Questions represented by letters.

CE	Whether to join a co-operative
CF	Whether to take a stake in the cooperative
CK	Is industrial poverty alleviation helping
CO	Whether to develop industries under the leadership of enterprises, cooperatives and large households
EA1	Where is the largest part of the new income? 1. Operating, wages and other non-transfer income; 2. Transfer income such as government subsidies and money from relatives and friends
DA	Is it helpful to increase the income of the family after obtaining employment assistance measures
DD	Whether have loan

The purpose of the first method is to find the question which can "improve" the strength of relationship between two incomes. The InfGain for some attributes are listed on Table 7. After calculating InfGain for all attributes, none of them are high enough, there is no need to proceed to the next level. Therefore, we can conclude that none attributes or any combinations of them have strong relationship between two sources of incomes.

TABLE 7. InfGain for some of attributes.

Attributes	CE	CF	CK	CO	EA1	DA	DD
InfGain	-0.02	-0.02	-0.02	0	0.02	0	0.03

Applying the second method to SiChuan data, we got the following decision tree in Figure 12.

Notice that among all the attributes, the three policies of poverty alleviation through employment, the largest source of new income and whether having mortgage loan have a greater impact on the structure of family income, of which the effect of employment assistance measures is the most significant. The leaves are labeled as leaf one to leaf five from left to right. Besides comparing various policies, leaves of tree also depict the income structure of several types of families. Of the five leaves, leaf three and leaf five have higher CE values of -0.49 and -0.86 respectively. Let's look at what kind of family these two leaves depict. For families in leaf five, employment assistance has no effect on them, and the largest part of new income in the family is transfer income. For families in leaf

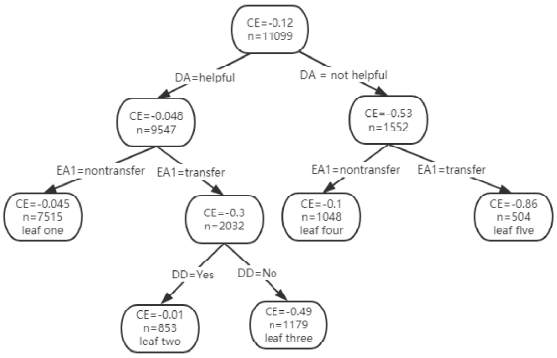


FIGURE 12. Clustering tree for Sichuan data.

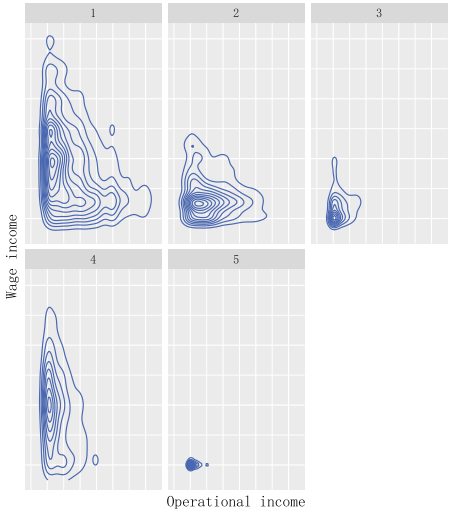


FIGURE 13. Contour plot of joint density of two incomes of the families in each leaf.

three, employment assistance has an effect on them, but this effect seems to be limited, because the largest part of the new income in the family is still transfer income, and the family does not have loans (no loans are not necessarily a good thing, according a survey in 2018 80.8% of the loan families use loans for business activities). To compare joint distributions of two incomes of different leaves, we make nonparametric estimation of the joint density of operational income and wage income of the families in each leaf in Figure 13. Notice that sample sizes of the leaves have no big difference except leaf one ($n = 7515$), but their distributions are significantly different, especially leaf five, which has highly concentrated points.

Next, we perform a hypothesis test for the difference of the joint distributions. The usual KS test is only suitable for single variable. Since no pre knowledge should be assumed for the joint distributions, we adopt the nonparametric testing method for joint density proposed by [34]. Let f_1 and f_2 be two d-dimensional densities. The hypothesis are $H_0 : f_1 = f_2$ vs $H_1 : f_1 \neq f_2$. The test result is shown in Figure 8, the difference between joint distributions are significant.

The decision tree ranks the variables in top-down order according to their importance. This gives us a way to compare

TABLE 8. P-values for pair wise comparison of leaves of Sichuan data.

leaf		P value	leaf		P value
leaf1	leaf2	0	leaf2	leaf3	0
leaf1	leaf3	0	leaf2	leaf4	3.4×10^{-18}
leaf1	leaf4	1.9×10^{-10}	leaf2	leaf5	0
leaf1	leaf5	0	leaf3	leaf4	0
leaf3	leaf5	0	leaf4	leaf5	0

policies. For example, as far as this decision-making tree is concerned, the distinction of employment poverty alleviation policies is clearer than whether they are transfer income or not. Even in all cases, employment poverty alleviation policies are the most effective in changing household income relationships.

The several policy combinations formed by this decision-making tree show significant differences in household income and income structure. This on one hand gives us a way to evaluate the effectiveness of the policy combination. On the other hand, it tells us about some of the common features of extremely poor families. The decision-making tree finds that the most effective way to distinguish the extremely poor families is those who answered that the maximum increase income is transfer income and employment poverty alleviation does not help. This is a reflection of the importance of these two policies for extremely poor families.

IV. CONCLUSION

Compared with the traditional decision tree with single variable as its analysis object, the analysis object in this article is the relationship between variables. By using copula entropy as a measure for the strength of relationship, we constructed two decision trees to serve different purposes. The regression tree is designed to select the attributes or their combinations of which the variables have strong relationship. Note that the relationship is ''strong'' in a general manner, not for any particular subsets. It is possible that a subset with variables having strong relationship may not show its strength after weighted averaging. The clustering tree selects attributes based on the degree of change of each attribute to relationship, thus the space of targeting variables is divided into several regions according to the strength of relationship.

Applying the regression tree method to the analysis of Sichuan Province data, we found that no matter the data is divided according to what kind of attributes or combinations of attributes, there is no strong correlation between operational income and wage income. Applying the clustering tree method to the data in Sichuan Province, we evaluate the effects of multiple policies. This evaluation is not achieved by comparing household incomes with their historical data, but by comparing different policies with each other. This is an effective way to measure policy effectiveness in the absence of historical data. Further, policy combinations with significant difference in distributions are formed by constructing decision trees. On the other hand, the significant difference in distributions also shows that policies are effective which will be more effective when it comes to policy combination.

In this article, we only considered the relationship between two random variables. This idea can be easily extended to three or more random variables. For more than three random variables, the estimation of copula entropy will suffer from the curse of dimensionality, then some other estimation methods e.g. [35]–[37] should be considered. One thing worth to mention is that using copula entropy to construct a splitting criterion is just one of many possible choices, because copula entropy is a general measure of dependence, but not the only choice. In fact, copula entropy also has its limitations, for example, it can only detect the strength of relationship not types, so different types of relationship may have the same copula entropy. Some important information may be ignored when using copula entropy to summarize the copula function. Decision trees for dependence based on other quantities or even multiple quantities will be a possible future direction.

REFERENCES

[1] P. Deheuvels, ''A Kolmogorov–Smirnov type test for independence and multivariate samples,'' *Rev. Roum. Math. Pures Appl.*, vol. 26, no. 2, pp. 213–226, 1981.

[2] J.-D. Fermanian, D. Radulovic, and M. Wegkamp, ''Weak convergence of empirical copula processes,'' *Bernoulli*, vol. 10, no. 5, pp. 847–860, Oct. 2004.

[3] O. Scaillet and J.-D. Fermanian, ''Nonparametric estimation of copulas for time series,'' FAME Res. Paper no. 57, 2002.

[4] I. Gijbels and J. Mielniczuk, ''Estimating the density of a copula function,'' *Commun. Statist.-Theory Methods*, vol. 19, no. 2, pp. 445–464, Jan. 1990.

[5] S. X. Chen and T.-M. Huang, ''Nonparametric estimation of copula functions for dependence modelling,'' *Can. J. Statist.*, vol. 35, no. 2, pp. 265–282, Jun. 2007.

[6] A. Charpentier, J.-D. Fermanian, and O. Scaillet, ''The estimation of copulas: Theory and practice,'' in *Copulas: From theory to Application in Finance*. New York, NY, USA: Springer, 2007, pp. 35–60.

[7] P. A. Morettin, C. M. Toloi, C. Chiann, and J. C. de Miranda, ''Wavelet-smoothed empirical copula estimators,'' *Revista Brasileira de Finanças*, vol. 8, no. 3, pp. 263–281, 2010.

[8] C. Genest, E. Masiello, and K. Tribouley, ''Estimating copula densities through wavelets,'' *Insurance, Math. Econ.*, vol. 44, no. 2, pp. 170–181, Apr. 2009.

[9] A. Sancetta and S. Satchell, ''The bernstein copula and its applications to modeling and approximations of multivariate distributions,'' *Econ. Theory*, vol. 20, no. 3, pp. 535–562, Jun. 2004.

[10] L. Qu, Y. Qian, and H. Xie, ''Copula density estimation by total variation penalized likelihood,'' *Commun. Statist.-Simul. Comput.*, vol. 38, no. 9, pp. 1891–1908, Oct. 2009.

[11] D. Pfeifer, D. Strassburger, and J. Philipps, ''Modelling and simulation of dependence structures in nonlife insurance with Bernstein copulas,'' in *Proc. Int. ASTIN Colloq.*, Helsinki, Finland, Jun. 2009.

[12] X. Shen, Y. Zhu, and L. Song, ''Linear B-spline copulas with applications to nonparametric estimation of copulas,'' *Comput. Statist. Data Anal.*, vol. 52, no. 7, pp. 3806–3819, Mar. 2008.

[13] G. Kauermann, C. Schellhase, and D. Ruppert, ''Flexible copula density estimation with penalized hierarchical B-splines: Flexible copula density estimation,'' *Scandin. J. Statist.*, vol. 40, no. 4, pp. 685–705, Dec. 2013.

[14] T. Nagler, ''Kdecopula: An r package for the kernel estimation of bivariate copula densities,'' 2016, *arXiv:1603.04229*. [Online]. Available: http://arxiv.org/abs/1603.04229

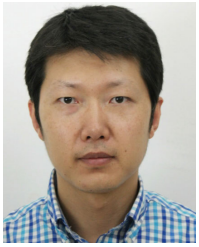
[15] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Amsterdam, The Netherlands: Elsevier, 2014.

[16] S. Moral-García, C. J. Mantas, J. G. Castellano, and J. Abellán, ''Ensemble of classifier chains and credal C4.5 for solving multi-label classification,'' *Prog. Artif. Intell.*, vol. 8, no. 2, pp. 195–213, Jun. 2019.

[17] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.

[18] Y. Mu, X. Liu, and L. Wang, ''A Pearson's correlation coefficient based decision tree and its parallel implementation,'' *Inf. Sci.*, vol. 435, pp. 40–58, Apr. 2018.

[19] M. Li, H. Xu, and Y. Deng, "Evidential decision tree based on belief entropy," *Entropy*, vol. 21, no. 9, p. 897, Sep. 2019.
[20] C. B. Bell, "Mutual information and maximal correlation as measures of dependence," *Ann. Math. Statist.*, vol. 33, no. 2, pp. 587–595, Jun. 1962.
[21] A. Dionisio, R. Menezes, and D. A. Mendes, "Mutual information: A measure of dependency for nonlinear time series," *Phys. A, Stat. Mech. Appl.*, vol. 344, nos. 1–2, pp. 326–329, Dec. 2004.
[22] J. B. Kinney and G. S. Atwal, "Equitability, mutual information, and the maximal information coefficient," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 9, pp. 3354–3359, Mar. 2014.
[23] R. Devon Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," 2018, *arXiv:1808.06670*. [Online]. Available: http://arxiv.org/abs/1808.06670
[24] P. E. Latham and Y. Roudi, "Mutual information," *Scholarpedia*, vol. 4, no. 1, p. 1658, 2009.
[25] J. Ma and Z. Sun, "Mutual information is copula entropy," *Tsinghua Sci. Technol.*, vol. 16, no. 1, pp. 51–54, Feb. 2011.
[26] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 6, Jun. 2004, Art. no. 066138.
[27] J. D. Victor, "Approaches to information-theoretic analysis of neural activity," *Biol. Theory*, vol. 1, no. 3, pp. 302–316, Sep. 2006.
[28] D. Pál, B. Póczos, and C. Szepesvári, "Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1849–1857.
[29] L. Faivishevsky and J. Goldberger, "ICA based on a smooth estimation of the differential entropy," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 433–440.
[30] S. Gao, G. Ver Steeg, and A. Galstyan, "Efficient estimation of mutual information for strongly dependent variables," 2015, *arXiv:1411.2003v3*. [Online]. Available: https://arxiv.org/abs/1411.2003
[31] R. A. A. Ince, B. L. Giordano, C. Kayser, G. A. Rousselet, J. Gross, and P. G. Schyns, "A statistical framework for neuroimaging data analysis based on mutual information estimated via a Gaussian copula," *Hum. Brain Mapping*, vol. 38, no. 3, pp. 1541–1573, Mar. 2017.
[32] H. Safaai, A. Onken, C. D. Harvey, and S. Panzeri, "Information estimation using nonparametric copulas," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 98, no. 5, Nov. 2018, Art. no. 053302.
[33] G. Geenens, "Probit transformation for kernel density estimation on the unit interval," *J. Amer. Stat. Assoc.*, vol. 109, no. 505, pp. 346–358, Jan. 2014.
[34] T. Duong, B. Goud, and K. Schauer, "Closed-form density-based framework for automatic detection of cellular morphology changes," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 22, pp. 8382–8387, May 2012.
[35] M. Ishmael Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. Devon Hjelm, "MINE: Mutual information neural estimation," 2018, *arXiv:1801.04062*. [Online]. Available: http://arxiv.org/abs/1801.04062
[36] M. Gabrié, A. Manoel, C. Luneau, N. Macris, F. Krzakala, and L. Zdeborová, "Entropy and mutual information in models of deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1821–1831.
[37] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 531–540.



QINGSONG SHAN received the B.S. degree in applied mathematics from the China University of Mining and Technology, Xuzhou, China, in 2002, and the M.S. and Ph.D. degrees in mathematical statistics from New Mexico State University, Las Cruces, NM, USA, in 2010 and 2015, respectively.

From 2015 to 2017, he was a Visiting Assistant Professor with the Department of Statistics, Indiana University, Bloomington, IN, USA. Since 2017, he has been an Assistant Professor with the Department of Statistics, Jiangxi University of Finance and Economics, Nanchang, China. His research interests include non-parametric methods, statistical learning methods, and copulas.

QIANNING LIU received the B.S. degree in applied mathematics from the China University of Mining and Technology, Xuzhou, China, in 2002, and the M.S. and Ph.D. degrees in mathematical statistics from New Mexico State University, Las Cruces, NM, USA, in 2014, and 2018, respectively.

Since 2018, she has been an Assistant Professor with the Department of Statistics, Jiangxi University of Finance and Economics, Nanchang, China. Her research interests include machine learning, high-dimensional data analysis, nonlinear differential equations, and nonlinear stochastic differential equations that arise in biology and finance.

Dr. Liu was a recipient of the NMSU Anna Schrufer Kist Endowed Scholarship, from 2015 to 2017, and the National Science Foundation of USA, travel grant to attend the Seminar on Stochastic Processes, in 2017.