

Variable Selection with Copula Entropy

MA Jian

(Hitachi (China) Research & Development Corporation, Beijing, 100084, China)

Abstract: Variable selection is of significant importance for classification and regression tasks in machine learning and statistical applications where both predictability and explainability are needed. In this paper, a Copula Entropy (CE) based method for variable selection which use CE based ranks to select variables is proposed. The method is both model-free and tuning-free. Comparison experiments between the proposed method and traditional variable selection methods, such as distance correlation, Hilbert-Schmidt independence criterion, stepwise selection, regularized generalized linear models and adaptive LASSO, were conducted on the UCI heart disease data. Experimental results show that CE based method can select the ‘right’ variables out more effectively and derive better interpretable results than traditional methods do without sacrificing accuracy performance. It is believed that CE based variable selection can help to build more explainable models.

Keywords: copula entropy; variable selection; distance correlation; Hilbert-Schmidt independence criterion; LASSO; ridge regression; elastic net; adaptive LASSO; AIC; BIC; explainability

2020 Mathematics Subject Classification: 62J99; 62G08; 62H20

Citation: MA J. Variable selection with copula entropy [J]. Chinese J Appl Probab Statist, 2021, 37(4): 405–420.

§1. Introduction

1) Variable Selection Problem

Variable selection is one of the old and widely studied model selection problems in statistics and machine learning^[1,2]. The problem arises when one wants to model the relationship between a variable of interest (response) and a (large) amount of potential explainable variables (predictors) but only a subset of latter explainable variables may be relevant to the former variable. The aim of the problem is to select a subset of variables under certain criteria. The criteria of selection are variable’s ability of prediction and interpretation.

E-mail: majian@hitachi.cn.

Received March 23, 2020. Revised June 17, 2020.

2) Existing Methods

(a) Measure based Variable Selection

The most natural way of variable selection is based on statistical association measures between response and individual predictor. Due to its simplicity and interpretability, measure based selection enjoys widely adoption and empirical successes in practice. The most common traditional measure is Pearson correlation coefficient (CC) for linear models. However, the application of CC assumes Gaussianity, which is unrealistic for most non-linear and non-Gaussian cases. For nonlinear dependence measure, some may considered mutual information (MI) in information theory [3], but only a few applications are reported due to notorious difficulty of estimating MI. Besides MI, several other nonlinear dependence measures were proposed.

Distance Correlation (dCor) is a nonlinear generalization of traditional correlation concept proposed by Székely et al. [4,5]. It generalizes bivariate second-order correlation to multivariate nonlinear cases via distance covariance. dCor between random vectors X and Y is defined as

$$\text{dCor}(X, Y) = \frac{\nu^2(X, Y)}{\sqrt{\nu^2(X)\nu^2(Y)}}, \quad (1)$$

where $\nu^2(X, Y)$ is distance covariance defined with characteristic function f as

$$\nu^2(X, Y; w) = \|f_{X,Y}(t, s) - f_X(t)f_Y(s)\|_w^2. \quad (2)$$

Here, $\|\cdot\|_w$ is the norm in the weighted L_2 function space defined with positive weight function $w(\cdot, \cdot)$ [4,5]. dCor characterizes independence: $\text{dCor}(X, Y) \geq 0$, and $\text{dCor}(X, Y) = 0$ if and only if X, Y are independent. dCor has been proposed as a tool for variable selection [6].

Hilbert-Schmidt independence criterion (HSIC) is another widely studied independence measure [7] and it has multivariate version – d -variable HSIC (dHSIC) [8]. dHSIC defines a nonlinear dependence measure in reproducing kernel Hilbert spaces (RKHS) with kernel function, as follows:

$$\begin{aligned} & \text{dHSIC}(P(X_1, X_2, \dots, X_d)) \\ &= \|\Pi(P(X_1) \otimes P(X_2) \otimes \dots \otimes P(X_d)) - \Pi(P(X_1, X_2, \dots, X_d))\|, \end{aligned} \quad (3)$$

where Π is kernel mean embedding function, and \otimes is tensor products of kernels. dHSIC can be considered as the distance in RKHS between the embeddings of joint distribution and margins. dHSIC also characterizes independence: $\text{dHSIC}(P(X_1, X_2, \dots, X_d)) = 0$ if

and only if X_1, X_2, \dots, X_d are independent so it is natural to apply it to variable selection problem^[9].

(b) Stepwise Selection with Information Criteria

Stepwise Selection is a standard approach for variable selection, usually on linear regression models, which sequentially selecting or eliminating the predictors once at a time based on certain criteria. Two of the main criteria are Akaike information criterion (AIC)^[10] and Bayesian information criterion (BIC)^[11]. Let l denotes the log maximum likelihood of the model, p denotes the number of free parameters of the model and N denotes the number of observation, and then AIC is defined as

$$\text{AIC} = -2l + 2p, \quad (4)$$

and BIC is defined as

$$\text{BIC} = -2l + p \ln N. \quad (5)$$

It can be learned from the definitions that both criteria are defined as penalized likelihood criteria that try to achieve a balance between goodness of fit (likelihood) and penalty on overfitting (number of free parameters) for model selection problem.

(c) Regularized Generalized Linear Models

Linear regression (LR) is the commonly used model in most of the researches that study the relationships under the assumption of linearity. Generalized linear models (GLMs) are a group of variants of LR, such as Logistic regression and Poisson regression, which introduce nonlinear response by means of link function. Due to their poor ability of variable selection, LR or GLMs are not applied directed to many cases for model selection, especially to high dimensional problems. To tackle this issue, regularization techniques are introduced to formalize a new learning problem from GLM problems under the assumption of sparsity. Let y denotes response, \mathbf{X} denote predictors, and β denote coefficients to be estimated, and then regularized GLMs solve the following problem:

$$\min_{\beta} \{L(\beta; y, \mathbf{X}) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2\}, \quad (6)$$

where $L(\cdot)$ denotes likelihood function, $\|\beta\|_i$ denote i th norm of coefficients β , and λ_i denote tuning parameters.

Three main variants of this regularized problem are defined by tuning the parameters λ_1, λ_2 . The problem for the case where $\lambda_1 = 0, \lambda_2 > 0$ is called ridge regression^[12]. The

least absolute shrinkage and selection operator (LASSO) problem corresponds to the cases where $\lambda_1 > 0$, $\lambda_2 = 0$ ^[13]. The problem for $\lambda_1, \lambda_2 > 0$ is called elastic net^[14]. For more variants of regularized GLMs, please refer to [15, 16] and references therein.

Regularized GLMs select a subset of variable by means of shrinking the non-zero coefficients of variables. Suppose the ‘true’ model has a group of sparse coefficients, an estimator is said to has oracle property if it can estimate this coefficients effectively and asymptotically^[17]. Zou^[18] has shown that the LASSO estimator has no oracle property. To address this issue, he proposed the Adaptive LASSO with the adaptive weights technique and demonstrated its oracle property under regularity conditions^[18].

3) Limitations of the Existing Methods

When using Stepwise Selection with different criteria, one may confuse which criterion is fitful to the given problem since different criteria derive different learning models. Comparison between the definition of AIC (4) and BIC (5) shows that BIC penalize the number of free parameter much more than AIC does. Therefore, it is generally believed that AIC tends to overfitting while BIC tends to underfitting^[19].

The existing regularized GLMs have their limitations for variable selection as well. As pointed out in [14], Ridge regression cannot produce parsimonious models. When there are a group of correlated predictors, LASSO tends to select only one from the group^[14], and to include many false positive variable into models while elastic net tends to select the variable group in or out together^[15]. The oracle property of Adaptive LASSO is only attached to GLMs under certain restrictive regularity condition.

The ultimate evaluation criterion for variable selection methods is whether they can discover the ‘right’ variables for the response and hence derive an interpretable model for the given problem with good prediction performance. In many domains, the interpretability of models is a much desired merit than the predictability of models. In this sense, the above existing methods cannot meet the requirement.

Copula entropy (CE) is a recently introduced multivariate statistical independence measure^[20] (more details in Section 2). It is defined rigorously with copula function and was proved to be equivalent to MI. However, it has not been applied to variable selection problem.

In (a) of the previous subsection, two measures for statistical independence (dCor and dHSIC) are introduced. CE, dCor and dHSIC are all promising tools for variable selection. Recently, Böttcher defined copula versions of dCor and dHSIC^[21]. However, no

empirical comparison between these three measures has been done yet.

In this paper, a variable selection method based on CE^[20], which can select the ‘right’ variables for explainable models without sacrificing prediction ability, is proposed. Due to CE, the proposed method is theoretically sound and computationally efficient. The proposed method is evaluated on a biomedical dataset, and compared with the existing methods on variable selection.

§2. Copula Entropy

1) Theory

Copula theory is about the representation of multivariate dependence with copula function^[22,23]. At the core of copula theory is Sklar theorem^[24] which states that multivariate probability density function can be represented as a product of its marginals and copula density function which represents dependence structure among random variables. Such representation separates dependence structure, i.e., copula function, with the properties of individual variables – marginals, which make it possible to deal with dependence structure only regardless of joint distribution and marginal distribution. This section is to define an statistical independence measure with copula. For clarity, please refer to [20] for notations.

With copula density, Copula entropy is define as follows^[20]:

Definition 1 (Copula entropy) Let \mathbf{X} be random variables with marginal distributions \mathbf{u} and copula density $c(\mathbf{u})$. CE of \mathbf{X} is defined as

$$H_c(\mathbf{X}) = - \int_{\mathbf{u}} c(\mathbf{u}) \ln c(\mathbf{u}) d\mathbf{u}. \quad (7)$$

In information theory, MI and entropy are two different concepts^[3]. In [20], Ma and Sun proved that they are essentially same – MI is also a kind of entropy, negative CE, which is stated as follows:

Theorem 2 MI of random variables is equivalent to negative CE:

$$I(\mathbf{X}) = -H_c(\mathbf{X}). \quad (8)$$

The proof of Theorem 2 is simple^[20]. There is also an instant corollary (Corollary 3) on the relationship between information of joint probability density function, marginal density function and copula density function.

Corollary 3

$$H(\mathbf{X}) = \sum_i H(X_i) + H_c(\mathbf{X}). \quad (9)$$

The above results cast insight into the relationship between entropy, MI, and copula through CE, and therefore build a bridge between information theory and copula theory. CE itself provides a mathematical theory of statistical independence measure.

2) Estimation

It has been widely considered that estimating MI is notoriously difficult. Under the blessing of Theorem 2, Ma and Sun^[20] proposed a simple and elegant non-parametric method for estimating CE (MI) from data which comprises of only two steps*:

1. Estimating empirical Copula density (ECD);
2. Estimating CE.

For Step 1, if given data samples $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$ i.i.d. generated from random variables $\mathbf{X} = \{x_1, x_2, \dots, x_N\}^\top$, one can easily estimate ECD as follows:

$$F_i(x_i) = \frac{1}{T} \sum_{t=1}^T \chi(\mathbf{x}_t^i \leq x_i), \quad (10)$$

where $i = 1, 2, \dots, N$ and χ represents for indicator function. Let $\mathbf{u} = (F_1, F_2, \dots, F_N)$, and then one can derives a new samples set $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_T\}$ as data from ECD $c(\mathbf{u})$. In practice, Step 1 can be easily implemented non-parametrically with rank statistics.

Once ECD is estimated, Step 2 is essentially a problem of entropy estimation which has been contributed with many existing methods. Among them, the kNN method^[25] was suggested in [20]. With rank statistic and the kNN method, one can derive a non-parametric method of estimating CE, which can be applied to any situation without any assumptions on the underlying system.

§3. CE based Variable Selection

In this section, we propose a new variable selection method based on CE. It is a new kind of association measure based method. The idea is simple: the CE between response and predictors are estimated from data, and then predictors are selected according to the

*The R package copent for estimating CE is available on CRAN and also on GitHub at <https://github.com/majianthu/copent>.

value of CE. With the selected variables, a model for the prediction problem is built. In the method, CE is estimated nonparametrically with the method in 2) of Section 2.

Since CE has many advantages over traditional association measure CC, it is quite obvious that the new method is superior to CC based method. The new method closely related to previously proposed MI based method but CE has more clear mathematical meaning for all the multivariate cases and is estimated in a new non-parametric way which makes the method efficient, stable and universally applicable. Since CE is defined as a distribution-free measure, the proposed method based on it is therefore model-free.

Another advantage of the new method is that the variable such selected can be attached with biological and physical meaning since CE, as a type of entropy, measures not only statistical dependence between variable and response, but also information transmission or energy exchange in the underlying systems. If a group of variables are selected by the proposed method, it is supposed to correspond to physical, biological or social meanings in the given system. As contrast, traditional methods have no such merit.

§4. Experiments and Results

1) Data

The heart disease dataset in the UCI machine learning repository^[26] is used in our experiments, which contains 4 databases about heart disease diagnosis collected from four locations. The dataset includes 920 samples totally, of which only 899 samples without missing values are used in the experiments. All the dataset have the same instance format with 76 raw attributes, of which the attribute ‘num’ is the diagnosis of patients’ disease. In the past research, only 14 attributes are recommended by researchers for clinical use, as listed in Table 1^[27].

Table 1 The recommended attributes in the UCI heart disease dataset

ID	3	4	9	10	12	16	19
Name	age	sex	cp	trestbps	chol	fbs	restecg
ID	32	38	40	41	44	51	58
Name	thalach	exang	oldpeak	slope	ca	thal	num

2) Evaluation Criteria

To evaluate the variable selection methods, there are two criteria: predictability and interpretability. In our experiments, on one side, we suggest to test the prediction accuracy

of the prediction models building on the selected variables. On the other side, we will try to check the explainability of the selected variable with reference to the established domain knowledge.

3) Experiments

We conducted 10 experiments on heart disease dataset to compare CE based method with other related methods. The goal of the experiments is to predict the diagnosis from other variables. In all experiments, both training data and test data were the whole dataset since we only investigate the variable selection ability of the methods and do not want to verify the generalizability of the models. The first experiment provides the baseline, in which 13 recommended variables were used and a SVM^[28] classifier was trained on these variables. In the following three experiment, a group of variables are selected with the CE, dCor, and dHSIC based method, and then a SVM classifier is trained on such selected variables. The next two experiments are on stepwise selection on GLM with AIC and BIC. Since the response (the ‘num’ attribute of heart disease dataset) has 5 levels of value, the GLMs in all the experiments are set as Poisson regression model with the ‘log’ link function. Stepwise selection is set as a ‘backward’ one under the guidance of AIC and BIC. The next 3 experiments are on regularized GLMs, including LASSO, ridge regression, and elastic net with $\lambda_1, \lambda_2 = 0.5$. For each regularized GLM, the best amount of shrinkage is determined with 10-fold cross validation. The last experiment is on adaptive LASSO with 10-fold cross validation to test its oracle property. The R packages ‘copent’^[29], ‘energy’^[6], ‘dHSIC’^[9], ‘e1071’^[30], ‘glmnet’^[31], and ‘parcor’^[18,32] were used for CE, dCor, dHSIC, SVM, regularized GLMs and adaptive LASSO in the experiments respectively. The default values of the parameters are used for the functions of CE, dCor, and dHSIC in the corresponding packages.

4) Results

The prediction accuracy of the models are listed in Table 2. It can be learned that the SVM classifier with variables selected by CE presents the best result (762 out of 899), better (5 more correct prediction, more than 0.56% improvement) than SVM with the recommended variables. The other two dependence measure based method present comparable results. The remaining methods show only moderate prediction accuracy. Clearly, CE base variable selection improves the performance of prediction.

Regularized GLMs select variables by non-zero coefficients. The coefficients of regu-

Table 2 Prediction accuracy of the models

Model	Accuracy (%)
SVM (recommended variables)	84.20
SVM (CE)	84.76
SVM (dCor)	82.76
SVM (dHSIC)	84.54
Stepwise GLM (AIC)	51.8
Stepwise GLM (BIC)	49.1
LASSO	79.2
Ridge Regression	63.0
Elastic Net	75.9
Adaptive LASSO	35.7

larized GLMs which are used in the prediction task are shown in Figure 1. For multinomial logistic regression, there are 5 groups of coefficients for 5 levels of response. We take the mean of 5 groups of coefficients as the overall coefficients. They indicate the relative importance of the variables in each model. The variables corresponding to non-zero coefficients are considered as selected by the GLMs.

The variables selected by the three dependence measures based method are shown in Figure 2. The dependence strength of ‘fbs’ (fasting blood sugar, #16) is used as the threshold for selection of all the three methods. The variables selected based on ranks of CE, dCor, and dHSIC, are listed in Table 3.

To check the interpretability of the selected variables of different methods, the recommended variables are taken as a golden rule for heart disease diagnosis since they are recommended by professional researchers as clinical relevant^[27]. The variables selected by different methods are summarized in Table 3. We compare the selected variables in each experiment with the recommended set. It can be learned from Table 3 that CE based method selects 11 out of 13 recommended variables, dCor and dHSIC select 9 and 10 out of 13 variables respectively, and meanwhile CE selects a smaller variable set with fewer false positive variables than the other two measures do which means CE has higher selection accuracy. As contrast, experimental results show that regularized GLMs fail to select the recommended variables out of others and adaptive LASSO selects 4 out of 13 variables. Meanwhile, stepwise GLMs with AIC and BIC select 8 and 5 out of 13 variables respectively. It should be mentioned that and the variables (#59–68) corresponding to the properties of vessels are also selected together, all with high CE value, which is meaningful and deserves further investigation in clinical practice. However, stepwise GLM does

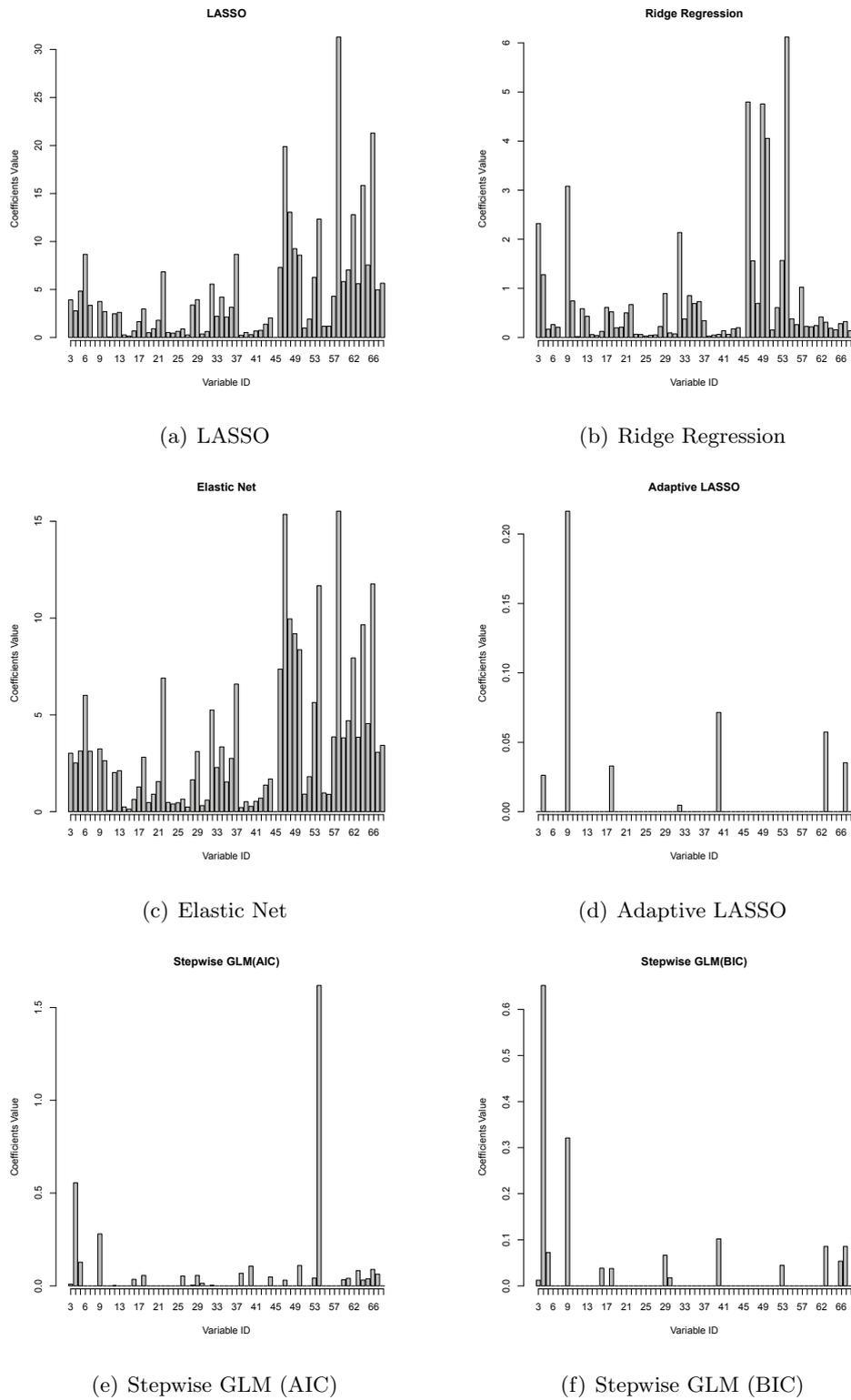
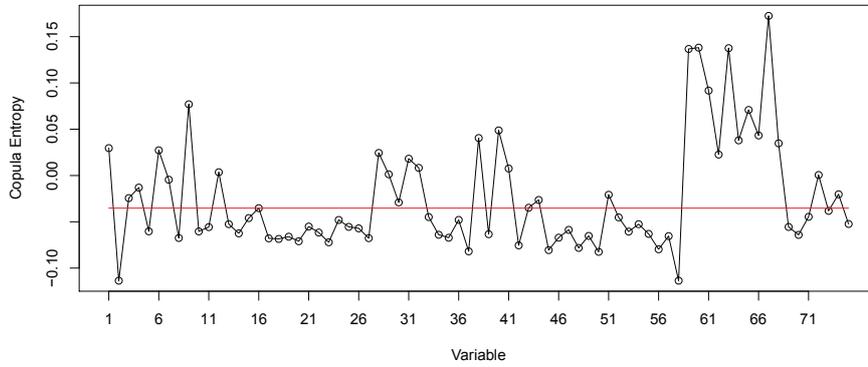
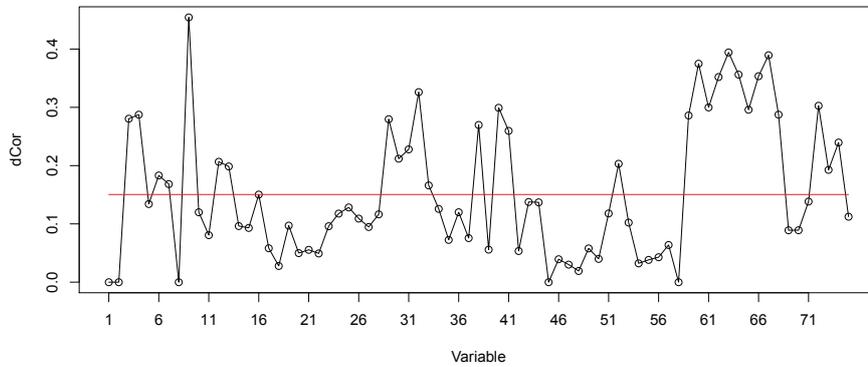


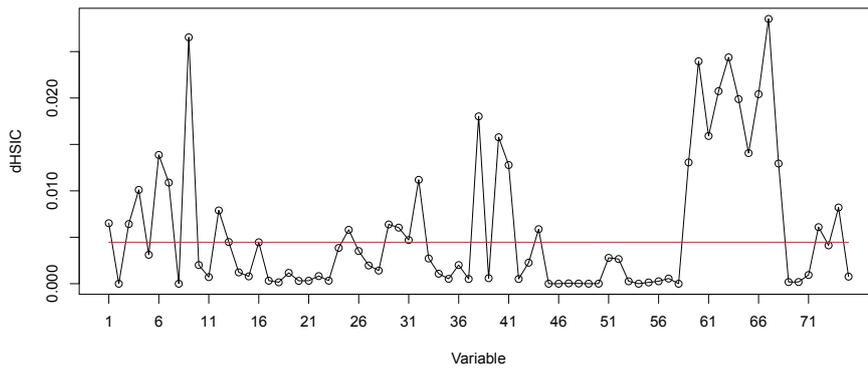
Figure 1 Coefficients of the GLMs



(g) CE



(h) dCor



(i) dHSIC

Figure 2 Variables selected by the three dependence measures

not select them all out. The results demonstrate strong ability of CE based method on selecting meaningful variables against its competitors.

Table 3 Selected variables by different methods

Method	Variable ID
Recommandation	3, 4, 9, 10, 12, 16, 19, 32, 38, 40, 41, 44, 51
LASSO	all except 8, 45
Ridge Regression	all except 8, 45
Elastic Net	all except 8, 45
Adaptive LASSO	4, 6, 9, 18, 32, 40, 63, 67
Stepwise GLM (AIC)	3, 4, 5, 9, 12, 16, 18, 20, 26, 29, 30, 32, 40, 44, 47, 50, 53, 54, 60, 61, 63, 65–67
Stepwise GLM (BIC)	3, 4, 5, 9, 16, 18, 29, 30, 40, 53, 63, 66, 67
dCor	3, 4, 6, 7, 9, 12, 13, 16, 28–33, 38, 40, 41, 52, 59–68
dHSIC	3, 4, 6, 7, 9, 12, 13, 16, 25, 29–32, 38, 40, 41, 44, 59–68
CE	3, 4, 6, 7, 9, 12, 16, 28–32, 38, 40, 41, 44, 51, 59–68

§5. Discussion

In the above experiments, the regularized GLMs select the variables based on the coefficients of the models. To achieve the goal, the methods should set up the models manually first, and then one has to tune both penalty parameter λ_i and shrinkage amount during model training. This means making assumptions of the model of the underlying system, including sparsity of coefficients and specific nonlinearity, which are usually incorrect. Even though the free parameters has been tuned to optimal, experimental results show that regularized GLM failed to select the recommended variables out and meanwhile they presented moderate prediction accuracy. Stepwise GLMs do better than regularized GLMs on variable selection but presented poor prediction performance.

Compared with them, CE based method is both model-free and tuning-free. It does nothing on model setting, tunes no parameter, and yet presents good performance on both prediction accuracy and variable selection. This is because that CE is a distribution-free measure of statistical independence and that its estimation is done in a non-parametric way. When applied, it makes no assumption on the underlying systems. With CE, variable selection is becoming a science with universally applicable theory and efficient method, instead of an art like regularized GLMs and other information criteria.

In the experiments, dCor and dHSIC also present comparable results and are also model-free and almost tuning-free. However, CE presents better results on both prediction accuracy and selection efficiency than dCor and dHSIC do. The good performance of CE over dCor and dHSIC can be explained theoretically. Though all the three measures

characterize multivariate independence, CE has a much rigorous definition as a type of entropy which has many well-known axiomatic properties for statistical dependence. As contrast, dCor and dHSIC are essentially nonlinear generalizations of traditional correlation concept and do not has the axiomatic properties that CE has, such as invariant to monotonic transformation, equivalent to correlation coefficients in Gaussian cases. As a type of entropy, CE also enjoys intrinsic physical meaning as a measure of information or energy exchange in the underlying systems which makes the variables selected with CE interpretable while the physical meaning of dCor and dHSIC are unclear yet.

It can be learned that stepwise GLM selects out less recommended variables than the proposed method does. This is because that the selection criteria of the methods are different. AIC is essentially an approximation of KL divergence^[10] between GLM and the underlying distribution. Since it is under the risks of model misspecification and approximation bias, the result is not so good as CE's. BIC is a criterion which is derived under the Bayesian framework with assumptions on the underlying models and the approximation of Bayesian computation^[11]. It is usually unrealistic for BIC to assume that the true model is within the model family under consideration. As contrast to AIC and BIC, the proposed method is guided by the CE, a well defined and estimated measure instead of a divergence or posteriori, which makes no assumption on the underlying models so there are no above risks of model misspecification or unrealistic assumptions and hence the selection is more advantageous theoretically. Meanwhile, the proposed method is computationally effective and efficient while both AIC and BIC are kind of approximations computed under certain assumptions.

The adaptive LASSO which claims to has oracle property selects only a few 'oracle' variable out. It is because that the adaptive LASSO possesses it oracle property under restrictive model assumptions and regularity conditions. Comparing with it, CE presents very good results because the oracle property of CE is guaranteed by statistical dependence between variables and response measured by CE, which is unconditional and universally applicable.

CE also leads to interpretability of models. CE based variable selection is based on the dependence relationships between variables and response of model. It selects variables based on the dependence strength measured by CE, which is independent of variable's scale and dimension. Such statistical relationships are believed to have real world physical or biological meanings, as has been demonstrated in discovering statistical associations^[33] – a problem closely related to variable selection. The ability of the proposed method to select meaningful variables is also demonstrated in the above experiments where the variables

group selected by CE is very similar to the variable group chosen by the professionals for clinical use and the additional vessels-related variables which are also meaningful but not considered before are also selected. The models built from the variables such selected can be explained with domain knowledge and applied to the cases where explainability matters.

§6. Conclusion

In this paper, we propose a CE based method for variable selection which use CE based ranks to select variables. The proposed method is both model-free and tuning-free. Comparison experiments between the CE based method and traditional variable selection methods, such as dCor, dHSIC, stepwise selection with information criteria, regularized GLMs and adaptive LASSO, are conducted on the UCI heart disease data. Experimental results show that CE based method can select the ‘right’ variables out effectively and derive better interpretable results than traditional methods do without sacrificing predictability. It is believed that CE based method makes variable selection becoming a science instead of an art and can help to build more explainable models which can lead to successful applications where interpretability of model matters. In the future, the proposed method will be applied to more real-world dataset to test its effectiveness and efficiency.

Acknowledgement The author thanks Matsumori Masaki for comments and suggestions.

References

- [1] GUYON I, ELISSEEFF A. An introduction to variable and feature selection [J]. *J Mach Learn Res*, 2003, **3**: 1157–1182.
- [2] GEORGE E I. The variable selection problem [J]. *J Amer Statist Assoc*, 2000, **95(452)**: 1304–1308.
- [3] COVER T M, THOMAS J A. *Elements of Information Theory* [M]. 2nd ed. Hoboken, NJ: Wiley-Interscience, 2006.
- [4] SZÉKELY G J, RIZZO M L, BAKIROV N K. Measuring and testing dependence by correlation of distances [J]. *Ann Statist*, 2007, **35(6)**: 2769–2794.
- [5] SZÉKELY, G J, RIZZO M L. Brownian distance covariance [J]. *Ann Appl Stat*, 2009, **3(4)**: 1236–1265.
- [6] LI R Z, ZHONG W, ZHU L P. Feature screening via distance correlation learning [J]. *J Amer Statist Assoc*, 2012, **107(499)**: 1129–1139.

- [7] GRETTON A, FUKUMIZU K, TEO C H, et al. A kernel statistical test of independence [C] // PLATT J C, KOLLER D, SINGER Y, et al. (eds.) *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, New York: Curran Associates, Inc., 2008: 585–592.
- [8] PFISTER N, BÜHLMANN P, SCHÖLKOPF B, et al. Kernel-based tests for joint independence [J]. *J R Stat Soc Ser B Stat Methodol*, 2018, **80(1)**: 5–31.
- [9] SONG L, SMOLA A, GRETTON A, et al. Feature selection via dependence maximization [J]. *J Mach Learn Res*, 2012, **13**: 1393–1434.
- [10] AKAIKE H. A new look at the statistical model identification [J]. *IEEE Trans Automatic Control*, 1974, **19(6)**: 716–723.
- [11] SCHWARZ G. Estimating the dimension of a model [J]. *Ann Statist*, 1978, **6(2)**: 461–464.
- [12] HOERL A E, KENNARD R W. Ridge regression: biased estimation for nonorthogonal problems [J]. *Technometrics*, 1970, **12(1)**: 55–67.
- [13] TIBSHIRANI R. Regression shrinkage and selection via the lasso [J]. *J Roy Statist Soc Ser B*, 1996, **58(1)**: 267–288.
- [14] ZOU H, HASTIE T. Regularization and variable selection via the elastic net [J]. *J R Stat Soc Ser B Stat Methodol*, 2005, **67(2)**: 301–320.
- [15] FAN J Q, LV J C. A selective overview of variable selection in high dimensional feature space [J]. *Statist Sinica*, 2010, **20(1)**: 101–148.
- [16] WU C, MA S G. A selective review of robust variable selection with applications in bioinformatics [J]. *Brief Bioinform*, 2015, **16(5)**: 873–883.
- [17] FAN J Q, LI R Z. Variable selection via nonconcave penalized likelihood and its oracle properties [J]. *J Amer Statist Assoc*, 2001, **96(456)**: 1348–1360.
- [18] ZOU H. The adaptive lasso and its oracle properties [J]. *J Amer Statist Assoc*, 2006, **101(476)**: 1418–1429.
- [19] DZIAK J J, COFFMAN D L, LANZA S T, et al. Sensitivity and specificity of information criteria [J]. *Brief Bioinform*, 2020, **21(2)**: 553–565.
- [20] MA J, SUN Z Q. Mutual information is copula entropy [J]. *Tsinghua Sci Technol*, 2011, **16(1)**: 51–54.
- [21] BÖTTCHER B. Copula versions of distance multivariate and dHSIC via the distributional transform – a general approach to construct invariant dependence measures [J]. *Statistics*, 2020, **54(3)**: 577–594.
- [22] NELSEN R B. *An Introduction to Copulas* [M]. 2nd ed. New York: Springer, 2006.
- [23] JOE H. *Dependence Modeling with Copulas* [M]. New York: Chapman and Hall/CRC, 2014.
- [24] SKLAR M. Fonctions de répartition à n dimensions et leurs marges [J]. *Publications de l'Institut de Statistique de l'Université de Paris*, 1959, **8**: 229–231.
- [25] KRASKOV A, STÖGBAUER H, GRASSBERGER P. Estimating mutual information [J]. *Phys Rev E Stat Nonlin Soft Matter Phys*, 2004, **69(6)**: 066138 (1–16).
- [26] DUA D, GRAFF C. UCI machine learning repository [DB]. Irvine, CA: University of California, School of Information and Computer Science, 2019.
- [27] NAHAR J, IMAM T, TICKLE K S, et al. Computational intelligence for heart disease diagnosis: a medical knowledge driven approach [J]. *Expert Syst Appl*, 2013, **40(1)**: 96–104.
- [28] SMOLA A J, SCHÖLKOPF B. A tutorial on support vector regression [J]. *Stat Comput*, 2004, **14(3)**: 199–222.

- [29] MA J. copent: estimating Copula entropy and transfer entropy in R [OL]. 2020 [2020-05-27]. <https://arxiv.org/abs/2005.14025>.
- [30] CHANG C C, LIN C J. LIBSVM: a library for support vector machines [J]. *ACM T Intel Syst Tec*, 2011, **2(3)**: 27 (1–27).
- [31] FRIEDMAN J, HASTIE T, TIBSHIRANI R. Regularization paths for generalized linear models via coordinate descent [J]. *J Stat Softw*, 2010, **33(1)**: 1–22.
- [32] KRÄEMER N, SCHÄFER J, BOULESTEIX A L. Regularized estimation of large-scale gene association networks using graphical Gaussian models [J]. *BMC Bioinformatics*, 2009, **10**: 384 (1–24).
- [33] MA J. Discovering association with Copula entropy [OL]. 2019 [2019-07-29]. <https://arxiv.org/abs/1907.12268>.

基于 Copula 熵的变量选择

马 健

(日立(中国)研究开发有限公司, 北京, 100084)

摘 要: 在要求可解释性的机器学习和统计应用中, 变量选择对分类和回归任务十分重要. 本文提出了一种基于 Copula 熵的变量选择方法, 利用 Copula 熵值的阶次选择变量. 本方法既是模型无关的又是参数无关的. 在 UCI 心脏病数据的基础上进行了本方法与传统变量选择方法 (包括距离相关、希尔伯特–施密特独立性准则、逐步选择、正则化广义线性模型和自适应 LASSO) 的对比实验. 实验结果表明, 基于 Copula 熵的方法能够更有效地选择‘正确’的变量, 在不牺牲准确性性能的同时得到比传统方法更具可解释性的模型.

关键词: Copula 熵; 变量选择; 距离相关; 希尔伯特–施密特独立性准则; LASSO; 岭回归; 弹性网络; 自适应 LASSO; AIC; BIC; 可解释性

中图分类号: O212.1; O212.4; O212.7