Categorical Neighbour Correlation Coefficient ($CnCor$) for Detecting Relationships between Categorical Variables

Lifeng Zhang, Shimo Yang, Hongxun Jiang*

School of Information, Renmin University of China
86, Zhongguancun Street, Haidian, Beijing, P.R.China, 100872
l.zhang@ruc.edu.cn, shimo@ruc.edu.cn, jianghx@ruc.edu.cn

Abstract

Categorical data is common and, however, special in that its possible values exist only on a nominal scale so that many statistical operations such as mean, variance, and covariance become not applicable. Following the basic idea of the neighbour correlation coefficient ($nCor$), in this study, we propose a new measure named the categorical $nCor$ ($CnCor$) to examine the association between categorical variables through using indicator functions to reform the distance metric and product-moment correlation coefficient. The proposed measure is easy to compute, and enables a direct test of statistical dependence without the need of converting the qualitative variables to quantitative ones. Compared with previous approaches, it is much more robust and effective in dealing with multi-categorical target variables especially when highly nonlinear relationships occur in the multivariate case. We also applied the $CnCor$ to implementing feature selection by the scheme of backward elimination. Finally, extensive experiments performed on both synthetic and real-world datasets are conducted to demonstrate the outstanding performance of the proposed methods, and draw comparisons with state-of-the-art association measures and feature selection algorithms.

Introduction

Detecting the associations between variables is one of the most important issue in data analysis and machine learning. In the past decades, a number of association measures have been proposed to enable capturing a wide range of complex data relationships. These methods, however, may sometimes exhibit less detection power when dealing with certain types of data. Developing more powerful association detection methods has been a challenging research.

One of the most widely used measures is mutual information (MI), which detects the dependence between variables in the context of information theory. A number of techniques have been proposed to estimate the score of MI, such as kernel density estimation (KDE) (Sohan and Príncipe 2009; Wang, Shen, and Zhang 2005), k-nearest neighbor distances (kNN) (Darbellay and Vajda 1999; Kraskov, Stögbauer, and Grassberger 2004), and bining (partitioning) (Reshef et al. 2011; Heller et al. 2016). Gao et al.(2017) proposed a KSG estimator (MKSG) to provide a better handling of discrete-continuous mixtures. Zeng, Xia, and Tong(2018) proposed a Jackknife version of kernel estimation of MI to free the estimation from bandwidth selection. Distance correlation ($dCor$) (Székely et al. 2007; Székely and Rizzo 2009) has a compact representation analogous to Cor, but enables detecting various nonlinear relationships. Randomized dependence coefficient (RDC) is proposed based on the Hirschfeld-Gebelein-Renyi maximum correlation coefficient (Lopez-Paz, Hennig, and Scholkopf 2013), and can capture a variety of complex associations between random variables of arbitrary dimension. Other approaches include kernel canonical correlation analysis (KCCA) (Bach and Jordan 2003), principal curve based methods (Delicado 2001; Delicado and Smrekar 2009), Hilbert-Schmidt independence criterion (HSIC) (Gretton et al. 2005), and nonlinear spectral correlation (Liu, Sohn, and Jeon 2017). Recently, (Zhang 2020) introduced an order statistics based association detection method called the neighbour correlation coefficient ($nCor$). Since association detection is a basic primitive in machine learning, it is useful in many learning tasks. Various MI estimators and correlation measures have been widely used as selection criteria for constructing filter type feature selection (FS) algorithms (Fleuret 2004; Brown et al. 2012; Shishkin et al. 2016). From the other point of view, assisting in FS has been also commonly considered as a way to evaluate the performance of association measures (Lopez-Paz, Hennig, and Scholkopf 2013; Gao et al. 2017).

To maximize the effectiveness, these measures are usually restricted to certain types of data such as continuous, discrete, or binary. Categorical (qualitative) data, mainly multilevel categorical data, is always difficult to analyze because its possible values are incomparable, and have no significance beyond simply providing a convenient label for a particular value. In such a situation, the data is not appropriate to apply many statistical methods such as mean, variance, and covariance. In many cases, it needs to be converted to quantitative data in order to be able to analyze the data. This conversion can usually be achieved by coding the categorical values into high-dimensional vector spaces, and however inevitably leads to a redundancy of data dimensionality while no additional information is gained. It is clearly that all the product-moment correlation based approaches become inapplicable when dealing with categorical data since they are all

*Corresponding author

composed of mean, variance, and covariance operations. MI estimators may also become less effective in handling categorical data, and probably be significantly influenced by the label configuration that describes the way in which labels are assigned to each category.

To overcome this problem, We propose a new statistical measure named categorical $nCor$ ($CnCor$) to realize a straightforward dependence detection between categorical variables, that is, not only categorical features but also categorical target. It is easy to implement and does not require any data processing or tuning of parameters. Moreover, we reformulate the coefficient of essentialness (COE) with a fast distance matrix updating strategy to measure the importance of one or more features in analysing the target variable with respect to some others. In some sense, the purpose of COE is similar to that of the partial correlation and CMI. Then, we use the COE test to implement FS tasks via the mechanism of backward elimination. Experimental studies are conducted to demonstrate the outstanding performance of the proposed methods in comparison with the existing association measures and FS techniques.

Background of The $nCor$

The $nCor$ has been introduced recently in the study of (Zhang 2020) for measuring the functional relationships between quantitative variables. The basic idea behind the $nCor$ is that if there exists a functional relationship between $(\mathbf{x}, y)$, a pair of data points should be of similar values of y when their values of $\mathbf{x}$ are sufficiently similar. That means, knowing the features determines the value of the target variable, in other words, the value of y is predictable from $\mathbf{x}$.

Consider data $(\mathbf{x}_{(1)}, y_{(1)}), (\mathbf{x}_{(2)}, y_{(2)}), \ldots, (\mathbf{x}_{(N)}, y_{(N)})$ that consists of N independent observations from the joint distribution of $(\mathbf{x}, y)$, where $\mathbf{x}_{(t)} = (x_{i(t)}|1 \leq i \leq M)$. The procedure of computing the $nCor$ can be summarised as follows. First, the sample points need to be reordered by using the permutation $\{n_k|1 \leq k \leq N\}$ that satisfies the criterion of minimizing the total Euclidean distance $\lambda_{n_k,n_{k+1}} = \|\mathbf{x}_{(n_{k+1})} - \mathbf{x}_{(n_k)}\|$ between the neighbouring data points as given in (1).

$$\{n_k|1 \leq k \leq N\} = \underset{\substack{n_k \in \{1,\cdots,N\}, \forall 1 \leq k \leq N \\ n_i \neq n_j, \forall i \neq j}}{\arg\min} \sum_{k=1}^{N-1} \lambda_{n_k,n_{k+1}} \quad (1)$$

In the bivariate case, data reordering can be achieved using order statistics which involves, arranging x in increasing numerical order $x_{(n_1)} \leq x_{(n_2)}, \cdots, \leq x_{(n_N)}$ known as order statistics $\{x_{(k:N)}|1 \leq k \leq N\}$, and then correspondingly reordering $y_{(n_1)}, y_{(n_2)}, \cdots, y_{(n_N)}$ referred to as the concomitants $\{y_{[k:N]}\}$.

In the multivariate case, data reordering can be conducted using the nearest neighbor (NN) algorithm through considering the reordering process as a traveling salesman problem (TSP). The algorithm randomly starts at one data point, then visits the unvisited data point that is nearest to the last visited data point, and repeats this process until all data points have been visited. The obtained route $\{n_1, \cdots, n_N\}$ can be used to generate concomitants $\{y_{[k:N]}\}$. The crucial point is that by the data reordering, both $\{y_{[k:N-1]}\}$ and $\{y_{[k+1:N]}\}$ should obey the same distribution as $\{y_{(t)}\}$ (with only one sample point omitted). Another point is that despite a suboptimal TSP route, NN algorithm can always yield a permutation that is of sufficient quality for conducting the $nCor$ test. That is, the $nCor$ is robust to the exact permutations.

Then, the $nCor$ is defined as below.

$$nCor(\mathbf{x}, y) = \frac{\sum_{k=1}^{N-1} (y_{[k:N]} - \overline{y})(y_{[k+1:N]} - \overline{y})}{\left(\sum_{k=1}^{N-1} (y_{[k:N]} - \overline{y})^2 \sum_{k=1}^{N-1} (y_{[k+1:N]} - \overline{y})^2\right)^{0.5}} \quad (2)$$

where the overbar denotes mean operation.

A key property of the $nCor$ test is that, it is a direct approximation of the R^2 of the underlying relationship $f(\cdot)$ between $(\mathbf{x}, y)$, and $nCor(\mathbf{x}, y) \to R^2 \to \text{var}(f(\mathbf{x}))/\text{var}(y)$ when $N \to \infty$. By this property, three $nCor$ based association measures have been proposed in (Zhang 2020) to characterize the intra and inter structures of the associations from the aspects of nonlinearity, interaction effect, and variable redundancy respectively. One of the measures is called the coefficient of essentialness (COE) which represents an estimate of the association strength between y and $\mathbf{x}_A$ with the effect of $\mathbf{x}_B$ (controlling features, and $\mathbf{x}_A \cap \mathbf{x}_B = \varnothing$) removed. In other words, it measures the dependence between $(\mathbf{x}_A, y)$ given the values of $\mathbf{x}_B$. In this sense, the role and purpose of the COE test is similar to that of the partial correlation coefficient in the linear case, as well as the conditional MI (CMI) in the context of information theory.

The Proposed Methods

As given in (2), the $nCor$ is calculated analogously to the Pearson correlation coefficient, so that it can only be performed on qualitative variables whose possible values must be comparable or relatively comparable whether the data distribution is continuous or discrete. Obviously, it is not applicable to categorical variables since their possible values are incomparable. To overcome this problem, a new association measure is proposed in this study by reforming both the distance measure in feature space and the formulation of correlation coefficient.

Distance Computation with Categorical Features

When computing the $nCor$ for a sample, the sample points need to be rearranged first based on the criterion of minimizing the total distance in $\mathbf{x}$ space between each pair of neighbouring points. For bivariate data, computing the distance is not necessary, since a good permutation can be easily obtained by using order statistics in which case the data points that have same x values would be rearranged together. For multivariate data, the distance becomes essential and, however, cannot be simply computed as the Euclidean norm whenever dealing with categorical features.

Difference measure: The first issue in distance computation is how to calculate the difference between two data points in respect of a multi-level categorical variable without coding the variable into multiple binary ones.

Consider the situation in which a subset of features $\mathbf{c} \subseteq \mathbf{x}$ are categorical, and the rest of the features denoted by $\mathbf{c}^{\mathrm{c}}$ are quantitative. If $x_i \in \mathbf{c}^{\mathrm{c}}$, then the difference is simply derived as $x_{i(p)} - x_{i(q)}$. If x_i is categorical, $x_{i(p)} - x_{i(q)}$ has no significance beyond an indication of whether or not the two data points have the same values of x_i. That is to say, $x_{i(p)} - x_{i(p)} \neq 0$ merely implies that the two values are different whatever nonzero number the subtraction yields. By its very nature, therefore, the subtraction can be replaced by an indicator function $I_{x_{i(p)} \neq x_{i(q)}}$ formulated as follows.

$$I_{x_{i(p)} \neq x_{i(q)}} = \begin{cases} 1, x_{i(p)} \neq x_{i(q)} \\ 0, Otherwise \end{cases} \tag{3}$$

The raw difference between a pair of data points in each x_i can be derived as

$$\Delta x_{i(p,q)} = \begin{cases} I_{x_{i(p)} \neq x_{i(q)}}, & x_i \in \mathbf{c} \\ x_{i(p)} - x_{i(q)}, & x_i \in \mathbf{c}^{\mathrm{c}} \end{cases} \tag{4}$$

Normalization: The second issue is how to normalize the variable (distance) values to adjust the values measured on different scales to a notionally common scale, and thereby make all the variables carry the same weight in distance computation and ensure data reordering equitable.

In this study, before computing $\Delta x_{i(p,q)}$, each $x_i \in \mathbf{c}^{\mathrm{c}}$ needs to be normalized to standard score which has zero mean and unit variance. With normalized x_i, we have

$$\begin{aligned} \sum_{p \neq q} (\Delta x_{i(p,q)})^2 =& N \sum_{q} (x_{i(q)} - 0)^2 + N \sum_{p} (x_{i(p)} - 0)^2 \\ & - 2 \sum_{p \neq q} x_{i(p)} x_{i(q)} \\ =& 2N(N-1)\mathrm{var}(x_i) = 2(N^2 - N) \end{aligned} \tag{5}$$

All the normalization techniques for qualitative variables are obviously not applicable to categorical variables. Actually, shifting and scaling the values of a categorical variable would not change the outcomes of (3). In this study, for each $x_i \in \mathbf{c}$, we normalize the difference matrix rather than the variable sequence. Despite having binary values, the differences obtained from using the indicator function still can be considered as quantitative. Based on this concept, the normalized difference $d_{i(p,q)}$ can be formulated as

$$d_{i(p,q)} = \left(\frac{2\left(N^2 - N\right) \Delta x_{i(p,q)}}{\sum_{n,m=1}^{N} \Delta x_{i(n,m)}} \right)^{0.5} \tag{6}$$

It is clearly that by (6), the sum squared $d_{i(p,q)}$ over the whole matrix for each $x_i \in \mathbf{c}$ is normalized to be equal to that for each $x_i \in \mathbf{c}^{\mathrm{c}}$. For $x_i \in \mathbf{c}^{\mathrm{c}}$, let $d_{i(p,q)} = \Delta x_{i(p,q)}$, and the distance can be computed as follows.

$$\lambda_{p,q} = \left(\sum_{i=1}^{M} d_{i(p,q)}^2 \right)^{0.5} \tag{7}$$

The sample points, then, can be reordered by using $\Lambda = \{\lambda_{p,q}\}^{N \times N}$ and NN algorithm as described in Section 2.

The Categorical $nCor$ ($CnCor$)

Correlation coefficient: The third issue is how to reform the product-moment correlation formula to adapt to categorical target variable without any additional processing of the data.

Consider a categorical target $y \in \{1, 2, \cdots, L\}$. $nCor$ is no longer applicable. Its basic idea, however, is a universal concept regardless of the type of data so that it holds true even for categorical y. The key component of (2) is the covariance operation between $y_{[n]}$ and $y_{[n+1]}$, which is to examine whether the values of y of each pair of neighbouring data points in $\mathbf{x}$ space are close to each other or not. If a functional relationship exists, with sufficiently large N, $(y_{[n]}, y_{[n+1]})$ should be of similar values and thereby exhibit a positive linear correlation. It is obviously that the notion of similar values is inappropriate to categorical variables as their values are just convenient labels. To address this problem, we adapt the above statement to satisfy the property of categorical data. That is to say, if a predictable relationship exists, $(y_{[n]}, y_{[n+1]})$ may very likely have the same value when they are sufficiently close in $\mathbf{x}$ space. Based on this concept, we use an indicator function to measure the relationship between $(y_{[n]}, y_{[n+1]})$ instead of the covariance operation. Likewise, the two variance operations which can be viewed as a special case of covariance are also substituted by indicator functions for scaling the coefficient. Then, (2) can be reformulated as

$$\begin{aligned} & \frac{\sum_{k=1}^{N-1} I_{y_{[k:N]} = y_{[k+1:N]}}}{\left(\sum_{k=1}^{N-1} I_{y_{[k:N]} = y_{[k:N]}} \sum_{k=1}^{N-1} I_{y_{[k+1:N]} = y_{[k+1:N]}} \right)^{0.5}} \\ & = \frac{\sum_{k=1}^{N-1} I_{y_{[k:N]} = y_{[k+1:N]}}}{N-1} \end{aligned} \tag{8}$$

Rescaling: The fourth issue is to adjust the range of (8) to satisfy the needs of association detection.

Clearly, the expectation of (8) is not zero (not even a constant, and varies with the marginal distribution of y) when $(\mathbf{x}, y)$ are independent, and in addition, its maximum value should be less than 1 even if $(\mathbf{x}, y)$ are perfectly associated. This not only is inconsistent with the common view of an association measure, but also leads to a difficulty in intuitively diagnosing the relationship since the threshold of the score yielded by (8) may vary case by case. Hence, (8) needs to be further shifted and rescaled, and the $CnCor$ can be formulated as follows.

Definition 1. *The categorical neighbor correlation coefficient ($CnCor$)*

Let y denote a L-level categorical target variable, and $\{y_{[1:N]}, y_{[2:N]}, \cdots, y_{[N:N]}\}$ denote the concomitants reordered by using the optimal permutation defined in (1). The $CnCor$ can be computed as

$$CnCor(\mathbf{x}, y) = \frac{\sum_{k=1}^{N-1} I_{y_{[k:N]} = y_{[k+1:N]}} - (N-1)\mu}{N - L - (N-1)\mu} \tag{9}$$

where

$$\mu = \sum_{v=1}^{L} \left(\frac{N_{y=v}}{N} \right)^2 \tag{10}$$

and $N_{y=v}$ denotes the number of elements in $\{y_{[k:N]}\}$ that have values of v.

Theorem 1. *The $CnCor$ score, computed as given in Definition 1, should have the following properties.*

(i) $CnCor(\mathbf{x}, y) \leq 1$.

(ii) If $(\mathbf{x}, y)$ are independent, then $\mathbb{E}[CnCor(\mathbf{x}, y)] = 0$, and $\lim_{N\to\infty} CnCor(\mathbf{x}, y) = 0$.

If a predictable relationship exists between $(\mathbf{x}, y)$, then $\mathbb{E}[CnCor(\mathbf{x}, y)] > 0$, and $\lim_{N\to\infty} CnCor(\mathbf{x}, y) > 0$.

(iii) A hypothesis test rejects the null hypothesis of independent if

$$CnCor(\mathbf{x}, y) > \frac{\Phi^{-1}(\alpha) \left((N-1)\mu(1-\mu)\right)^{0.5}}{N - L - (N-1)\mu} \tag{11}$$

where $\Phi^{-1}(\cdot)$ denotes the inverse standard normal cumulative distribution function, and α is the significance level of the $CnCor$ test.

Proof. (i) Suppose a target variable $y \in \{1, 2, \cdots, L\}$ that is reordered perfectly such that, the concomitants are arranged as $\{1, \cdots, 1, 2, \cdots, 2, \cdots, L, \cdots, L\}$. Even in such an ideal situation, there should be $L-1$ out of $N-1$ indicator functions that yield a value of 0, which means $\max\left(\sum_k I_{y_{[k:N]}=y_{[k+1:N]}} | \{n_k\}\right) = N - L$, and therefore,

$$\max\left(CnCor(\mathbf{x}, y) | \mathbf{x} \in \mathbb{R}^M, y \in \{1, \cdots, L\}\right) = 1 \tag{12}$$

(ii) When $(\mathbf{x}, y)$ are independent, $\{y_{[k:N]}\}$ is obviously a randomly reordered sequence. and therefore each $y_{[k:N]}$ is i.i.d that obeys the same distribution as y. That is to say, for all $v, w \in \{1, 2, \cdots, L\}$, $\Pr(y_{[k+1:N]} = v, y_{[k:N]} = w) = \Pr(y_{[k+1:N]} = v)\Pr(y_{[k:N]} = w)$. Then, the expectation of each indicator function can be derived as

$$\begin{aligned}\mathbb{E}[I_{y_{[k:N]}=y_{[k+1:N]}}] =& 1 \sum_{v=1}^{L} \Pr(y_{[k+1:N]} = v, y_{[k:N]} = v) \\ &+ 0 \sum_{v=1}^{L} \Pr(y_{[k+1:N]} \neq v, y_{[k:N]} = v) \\ =& \sum_{v=1}^{L} \Pr(y = v)^2 = \sum_{v=1}^{L} \left(\frac{N_{y=v}}{N} \right)^2 = \mu\end{aligned} \tag{13}$$

$I_{y_{[k:N]}=y_{[k+1:N]}}$ is also i.i.d., by (13), it can be easily obtained that, the expectation of the numerator of (9) is zero so that, $\mathbb{E}[CnCor(\mathbf{x}, y)] = 0$.

When a predictable relationship exists between $(\mathbf{x}, y)$, with a large N such that $\lambda_{n_k, n_{k+1}}$ is sufficiently small, a pair of neighbouring data points which have the same or very similar $\mathbf{x}$ values may be of the same values of y with a higher probability. That is to say,

$$\begin{aligned}&\Pr(y_{[k+1:N]} = v | y_{[k:N]} = v) > \Pr(y_{[k+1:N]} = v) \\ &\Rightarrow \Pr(y_{[k+1:N]} = v, y_{[k:N]} = v) > \\ &\quad \Pr(y_{[k+1:N]} = v)\Pr(y_{[k:N]} = v) \\ &\Rightarrow \mathbb{E}[I_{y_{[k:N]}=y_{[k+1:N]}}] > \sum_{v=1}^{L} \Pr(y = v)^2 \\ &\Rightarrow \mathbb{E}[CnCor(\mathbf{x}, y)] > 0\end{aligned} \tag{14}$$

When $N \to \infty$, by the Kolmogorov's strong law of large numbers, if $(\mathbf{x}, y)$ are independent then $CnCor(\mathbf{x}, y) \to 0$, otherwise $CnCor(\mathbf{x}, y) > 0$, almost surely.

(iii) As discussed above, the $CnCor$ score is expected to be greater than zero when y is predictably dependent on $\mathbf{x}$. A one-tailed test, therefore, can be employed to examine the statistical significance of the score.

Since $I_{y_{[k:N]}=y_{[k+1:N]}}$ obeys the Bernoulli distribution with success probability μ, then, $\sum_k I_{y_{[k:N]}=y_{[k+1:N]}}$ obeys the binomial distribution as

$$\sum_{k=1}^{N-1} I_{y_{[k:N]}=y_{[k+1:N]}} \sim B(N-1, \mu) \tag{15}$$

By the De Moivre–Laplace theorem,

$$\sum_{k=1}^{N-1} I_{y_{[k:N]}=y_{[k+1:N]}} \sim \mathcal{N}((N-1)\mu, (N-1)\mu(1-\mu)) \tag{16}$$

By the definition of $CnCor$, we have

$$\frac{CnCor(\mathbf{x}, y)(N - L - (N-1)\mu)}{\left((N-1)\mu(1-\mu)\right)^{0.5}} \sim \mathcal{N}(0, 1) \tag{17}$$

Then, the confidence limit with a specific significance level α can be easily established as given in (11). □

Remark 1. *For the special case in which the target variable is binary (2-level categorical), the scores obtained from using the $nCor$ and the $CnCor$ tests are very close whether the variables are dependent or not. It is because when $y \in \{0, 1\}$, testing the positive correlationship between $y_{[k:N]}$ and $y_{[k+1:N]}$ is approximate to detecting how many $y_{[k:N]}$ are equal to $y_{[k+1:N]}$. Whenever y is multi-level categorical, the two tests show very different performance. The labelling of the possible values of y may probably cause extensive damage to the detection power of the $nCor$ test and even failure. In contrast, the $CnCor$ test is completely insensitive to the assignment of labels.*

Remark 2. *When dealing with discrete or categorical features, it is often the case that some sample points are of the exact same values of $\mathbf{x}$. Just in case, it is better to randomly rearrange the sample before data reordering to avoid the impact of the original order of sample points on the data reordering, and thereafter the $nCor$ or $CnCor$ estimation, whether for bivariate or multivariate data.*

Fast Distance Computation for COE Test

Suppose we have two subsets of features, $\mathbf{x}_A$ and $\mathbf{x}_B$. COE measures the proportion of the variance of y that can be predicted from $\mathbf{x}_A \cup \mathbf{x}_B$ but cannot be predicted from $\mathbf{x}_B$. This concept can also be applied to categorical data, and we adapt the definition of the COE by using $CnCor$ as

$$\begin{aligned} &COE(\mathbf{x}_A, y|\mathbf{x}_B) \\ &= CnCor(\mathbf{x}_A \cup \mathbf{x}_B, y) - \max\left(0, CnCor(\mathbf{x}_B, y)\right) \end{aligned} \tag{18}$$

where $\mathbf{x}_A, \mathbf{x}_B \neq \varnothing$, $\mathbf{x}_A \cap \mathbf{x}_B = \varnothing$, and $\mathbf{x}_A \cup \mathbf{x}_B \subseteq \mathbf{x}$.

It is noted that whenever calculating a COE score, two distance matrices need to be obtained in advance respectively for executing the two $CnCor$ tests. Suppose that one wants to examine the essentialness of each feature in a dataset by testing $COE(x_1, y|\mathbf{x}/x_1), \cdots, COE(x_M, y|\mathbf{x}/x_M)$, then, totally $M+1$ distance matrices need be calculated, that is obviously computationally expensive. To overcome this problem, here, we proposed a fast distance updating strategy to reduce the computational cost of this kind of consecutive COE test. Consider two feature sets $\mathbf{x}_A$ and $\mathbf{x}_B$, Λ^B can be obtained by modifying $\Lambda^{A\cup B}$ as formulated in (19), and the converse is also valid. Reconsider the above scenario, by the proposed strategy, the computational cost required for obtaining each $\Lambda^{\mathbf{x}/x_i}$ is reduced $|\mathbf{x}/x_i|/|x_i| = M-1$ times.

$$\lambda^B_{p,q} = \left(\left(\lambda^{A\cup B}_{p,q}\right)^2 - \sum_{x_i \in \mathbf{x}_A} d^2_{i(p,q)} \right)^{0.5} \tag{19}$$

Remark 3. *(18) can be interpreted as an assessment of how much improvement can be expected by using $\mathbf{x}_A$ when $\mathbf{x}_B$ is already present in the model. In this sense, it is naturally a convenient tool that can be directly applied to FS. $COE(x_i, y|\mathbf{x}/x_i) > 0$ indicates that x_i is an essential feature that must be involved in model construction. Otherwise, x_i is either irrelevant or redundant, and thus can be removed from $\mathbf{x}$ to reduce the model without loss of prediction accuracy. The more essential x_i, the larger score the COE.*

Experimental Studies

Performance Evaluation of The $CnCor$

Here, we experimentally evaluate the effectiveness of the $CnCor$, and make comparisons with state-of-the-art methods. See Appendix A in supplementary material for details.

The experimental settings are summarized as follows. (i) For each trial, firstly, we generated a set of continuous feature sequences $u_1, \cdots, u_M \sim U(0,1)$ with length of 1000, and produced a continuous target z by $z = f(\mathbf{u})$. Then, we respectively transformed z and each u_i to 5-level and 10-level categorical counterparts y and x_i, by grouping their values into 5 or 10 contiguous bins, each having the same width. (ii) To imitate the easiest, uncertain, and the hardest cases, three types of label configurations were employed. The first one is to successively assign labels from 1 to L to the groups arranged by their values in an ascending order. In this case, y and x_i are ordinal data which is a special type of categorical data, and in a sense can be viewed as quantitative rather than qualitative. The second one is to randomly assign a distinct label from $\{1, \cdots, L\}$ to each group. Thus, the impact of categorization on association detection may vary from trial to trial. The third one is to give each group a fixed label according to an unfavorable configuration that aims to make the detection more difficult. In each trial, only one way out of the three was conducted. (iii) For E4 and E5, a reference dataset was created in every trial to assist in evaluating the detection results. It was generated by randomly rearranging $\{y_{(t)}\}$ such that, y was made entirely independent to $\mathbf{x}$ while remaining its distribution unchanged. The scores of the reference datasets detected by each methods were used to compute the 95th percentile that was thereafter considered as a threshold and compared with the scores of the datasets under examination. (iv) The underlying function of E4 was a superposition of nonlinear main effects, and E5 was a complicated mixture of both main effects and interactions. To increase the difficulty, in E4, we also considered the situations of noise corruption realized by erroneously labeling 20% of $\{y_{(t)}\}$.

We numerically demonstrate the properties of the $CnCor$ especially its robustness against arbitrary label configuration, by contrasting with the $nCor$. Figure 1(a) shows that when (x, y) are independent, the shapes of the cumulative histograms for the $nCor$ and $CnCor$ scores nearly coincide with the theoretical cumulative distribution curve given in Theorem 1 and (Zhang 2020), which confirms the validity of the two statistical tests. Figures 1(b) to 1(d) show that when using random label configurations the $nCor$ scores spread over a wide range in every case, and moreover, in the hardest cases using unfavorable configurations the $nCor$ failed to detect the associations as all the scores were insignificant. That is to say, the effectiveness of the $nCor$ highly depends on the label configuration when dealing with categorical data. In every case, by contrast, the $CnCor$ scores were distributed almost the same way whichever type of label configuration was adopted, and whether $f(\cdot)$ was bivariate or multivariate, main effect or higher-order interaction. Clearly, the $CnCor$ is robust to the label configurations.

We also compared the $CnCor$ with HSIC (Gretton et al. 2005), $dCor$ (Székely and Rizzo 2009), RDC (Lopez-Paz, Hennig, and Scholkopf 2013), QMI (Sohan and Príncipe 2009), MKSG (Gao et al. 2017), and JMI (Zeng, Xia, and Tong 2018). In Figures 1(d) and 1(f), we observed that the threshold values for 5 out of 8 methods noticeably varied with both the distribution of y and the dimensionality of $\mathbf{x}$, which is obviously not desirable. The expectation and dispersion of the score of independent $(\mathbf{x}, y)$ are always preferred to be constant, otherwise it may lead to the misjudgment of independency in some cases. The $CnCor$ performed consistently on the three types of data. In contrast, all the other methods showed sensitivity to the data type to different extent. Compared to on ordinal data, most methods exhibited much less detection power on categorical data, especially on the data that was categorized by unfavorable label configurations. For E5 with specified label configuration and large M, most scores obtained by using these methods fell below the thresholds, which means they all failed in detecting the relationship. On the whole, the $CnCor$ outperformed all the baseline methods.

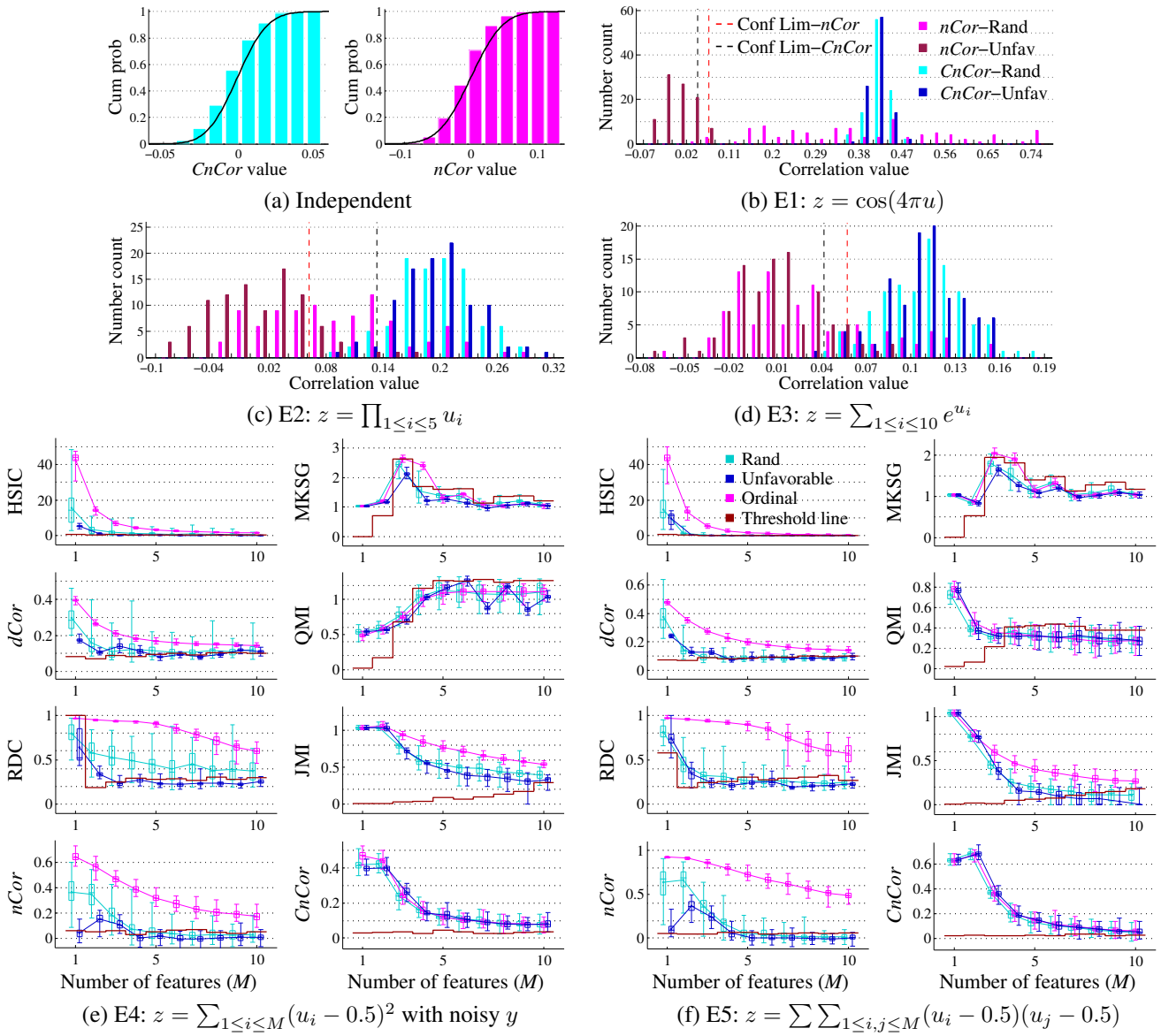

Figure 1: (a): the cumulative histograms over 1000 trials and the theoretical cumulative distributions of the $CnCor(x, y)$ and $nCor(x, y)$ scores for independent (x, y). (b) to (d): the histograms over 100 trials of the $CnCor(\mathbf{x}, y)$ and $nCor(\mathbf{x}, y)$ scores for three complicated relationships with the data categorized by using random and unfavorable label configurations. (e) and (f): the detection results for $(x_1, \cdots, x_M, y)$ with 50 trials in each case obtained from using the eight association measures.

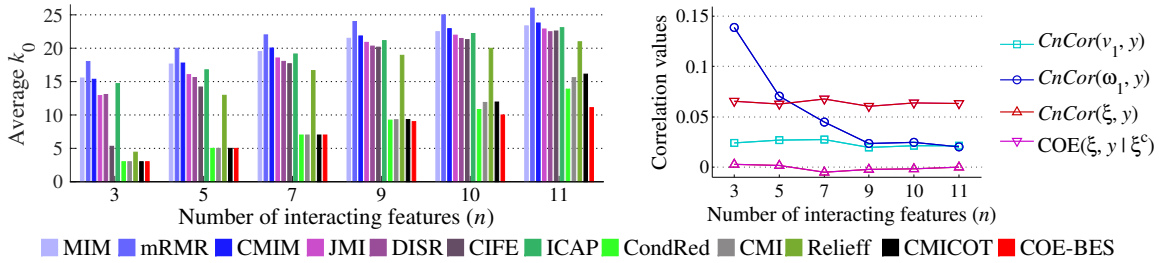

Figure 2: The experimental results of the COE test and FS on 6×100 synthetic datasets. Left: comparisons between $CnCor$-EBS and the baselines in terms of average k_0; Right : the average $CnCor$ and COE scores for ξ, ω_1 and v_1 under different n (the results for ω_2 to ω_n and v_2 to v_{10} are omitted, since they should all represent the same behaviour).

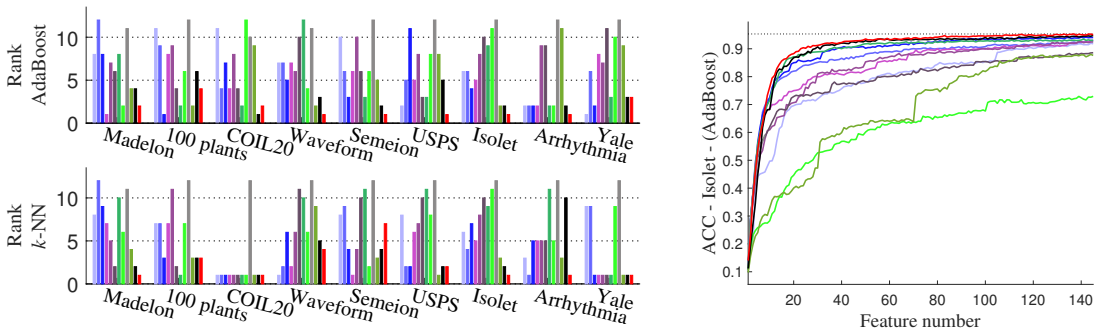

Figure 3: The experimental results of COE-EBS on 9 benchmark datasets. Left: ranking of the highest ACC obtained by using k-NN and AdaBoost on the optimal feature subsets found by the FS methods; Right: an example of the ACC against feature numbers obtained by using AdaBoost classifiers and different FS algorithms

Demonstration of COE and COE-BES

We apply the $CnCor$ based COE test to conducting FS by the mechanism of BES. The procedure of COE-BES involves starting with all candidate features, testing the COE score of each feature, deleting the feature which has the lowest COE score, and repeating this process until there is only one feature left. See Appendix B for details of the algorithm.

To illustrate the effectiveness of the $CnCor$ based COE and COE-BES, we implemented a set of experiments which include 6 synthetic datasets originally introduced by (Shishkin et al. 2016) for assessing the capability of different feature selection techniques to detect high-order feature dependencies. Each dataset consists of a set of binary feature variables $\mathbf{x} = \mathbf{x}_{int} \cup \mathbf{x}_{rel-not-int} \cup \mathbf{x}_{irr}$ and a binary target y. $\mathbf{x}_{int}$, $\mathbf{x}_{rel-not-int}$, and $\mathbf{x}_{irr}$ respectively contain n jointly interacting relevant features, 10 relevant but non-interacting features, and 5 irrelevant features. In the experiments, n was set to be 3, 5, 7, 9, 10, 11. For each case, we randomly generated 100 datasets with sample size of 1000. (Shishkin et al. 2016) also introduced an evaluation metric $k_0 = \min\{k \mid \mathbf{x}_{int} \subseteq S_k\}$ where S_k denotes a feature subset that contains the selected top-k features . The smaller k_0, the more effective the FS method, since it builds the smaller set of features needed to construct the best possible classifier. We also compared COE-BES with 11 well-established FS methods including CMICOT(Shishkin et al. 2016), mRMR(Peng, Long, and Ding 2005), CMIM(Fleuret 2004), MIM(Lewis 1992; Guyon et al. 2006), JMI(Yang and Moody 1999; Bennasar, Hicks, and Setchi 2015), DISR(Meyer, Schretter, and Bontempi 2008), CIFE(Lin and Tang 2006), ICAP(Jakulin 2005), CondRed(Brown et al. 2012), CMI(Brown et al. 2012; Fleuret 2004), and ReliefF(Kononenko 1994; Kira, Rendell et al. 1992). See Appendix C for details and more discussions.

Figure 2 (average k_0) shows that some of the algorithms failed to detect these interacting features appropriately (there is a special feature ξ in $\mathbf{x}_{int}$ that is very easy to omit when using these methods), and most of the others exhibited decreasing detection power as n increased (because with increasing n, $x_i \in \mathbf{x}_{int}$ became less relevant with reference to $x_j \in \mathbf{x}_{rel-not-int}$, especially when $n = 11$). Only COE-BES displayed superior performance and thoroughly captured the associations as k_0 being close to n. Figure 2 (correlation values) confirms that the average $CnCor(\xi, y)$ is almost zero in all the six datasets, which means that ξ is irrelevant when considered alone. Nevertheless, the COE test suggests that ξ is essential and not replaceable even when all the other features have been already invloved in model construction. Moreover, with increasing n the relevance of ω_i decreases, and so does the difficulty of feature selection. In summary, these results are fully consistent with the characters of the data, and further demonstrated the $CnCor$.

Finally, we applied COE-BES and the baseline methods to 9 public benchmark datasets (Guyon et al. 2004; Fanty and Cole 1991; Buscema 1998; Cai et al. 2010; Cai, He, and Han 2011; Breiman et al. 1984; Mallah, Cope, and Orwell 2013; Guvenir et al. 1997). We generated $M - 1$ subsets of top-ranked features by using each method on each dataset. Then, we employed k-NN and AdaBoost models to assess the accuracy (ACC) of classification that can be achieved by each feature subset. Figure 3 (rank numbers) shows that COE-EBS evidently outperformed the other methods on most datasets. Taking Isolet dataset as an example, as shown in the line plot, the ACC line of COE-EBS lies entirely above the other lines, which implies less features but higher accuracy. Actually, COE-EBS showed significant superiority over the baselines on 4 of 9 datasets. For details see Appendix D in supplementary material.

Conclusions

Here, we have proposed a new statistical measure named the $CnCor$, and a $CnCor$ based fast COE test. When dealing with multi-level categorical feature and target variables, the $CnCor$ provides a more impartial distance computation without transforming the qualitative features to quantitative ones, and then effectively measure the association between variables without worrying about whether the underlying label configuration of the targets is unfavorable or not. Despite very concise and easy to implement, the $CnCor$ shows a much better effectiveness and robustness in comparison to previous studies. In addition, we have used the COE test to conduct FS tasks, and numerical studies showed that the new algorithm had a competitive performance on both synthetic and real datasets.

References

Bach, F. R.; and Jordan, M. I. 2003. Kernel Independent Component Analysis. *Journal of Machine Learning Research*, 3(1): 1–48.

Bennasar, M.; Hicks, Y.; and Setchi, R. 2015. Feature selection using joint mutual information maximisation. *Expert Systems with Applications*, 42(22): 8520–8532.

Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1984. Classification and Regression Trees (CART). *Biometrics*, 40(3): 358.

Brown, G.; Pocock, A.; Zhao, M.-J.; and Luján, M. 2012. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The journal of machine learning research*, 13(1): 27–66.

Buscema, M. 1998. MetaNet: The Theory of Independent Judges. *Substance use and misuse*, 33: 439–461.

Cai, D.; He, X.; and Han, J. 2011. Speed up kernel discriminant analysis. *The VLDB Journal*, 20(1): 21–33.

Cai, D.; He, X.; Han, J.; and Huang, T. S. 2010. Graph regularized nonnegative matrix factorization for data representation. *IEEE transactions on pattern analysis and machine intelligence*, 33(8): 1548–1560.

Darbellay, G. A.; and Vajda, I. 1999. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45: 1315–1321.

Delicado, P. 2001. Another Look at Principal Curves and Surfaces. *Journal of Multivariate Analysis*, 77(1): 84–116.

Delicado, P.; and Smrekar, M. 2009. Measuring non-linear dependence for two random variables distributed along a curve. *Statistics and Computing*, 19(3): 255–269.

Fanty, M.; and Cole, R. 1991. Spoken Letter Recognition. In Lippmann, R. P.; Moody, J.; and Touretzky, D., eds., *Advances in Neural Information Processing Systems*, volume 3. Morgan-Kaufmann.

Fleuret, F. 2004. Fast binary feature selection with conditional mutual information. *Journal of Machine learning research*, 5(9): 1531–1555.

Gao, W.; Kannan, S.; Oh, S.; and Viswanath, P. 2017. Estimating Mutual Information for Discrete-Continuous Mixtures. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 5988–5999. Red Hook, NY, USA: Curran Associates Inc.

Gretton, A.; Bousquet, O.; Smola, A.; and Schölkopf, B. 2005. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In *Proceedings of the 16th International Conference on Algorithmic Learning Theory*, 63–78.

Guvenir, H. A.; Açar, B.; Demiroz, G.; and Cekin, A. 1997. A supervised machine learning algorithm for arrhythmia analysis. *Computers in Cardiology 1997*, 433–436.

Guyon, I.; Gunn, S.; Nikravesh, M.; and Zadeh, L. A. 2006. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Berlin, Heidelberg: Springer-Verlag.

Guyon, I.; Gunn, S. R.; Ben-Hur, A.; and Dror, G. 2004. Result Analysis of the NIPS 2003 Feature Selection Challenge. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, volume 4, 545–552.

Heller, R.; Heller, Y.; Kaufman, S.; Brill, B.; and Gorfine, M. 2016. Consistent distribution-free k-sample and independence tests for univariate random variables. *Journal of Machine Learning Research*, 17: 1–54.

Jakulin, A. 2005. *Machine learning based on attribute interactions*. Ph.D. thesis, Univerza v Ljubljani.

Kira, K.; Rendell, L. A.; et al. 1992. The feature selection problem: Traditional methods and a new algorithm. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, 129–134. AAAI Press.

Kononenko, I. 1994. Estimating attributes: Analysis and extensions of RELIEF. In *European conference on machine learning*, 171–182. Springer.

Kraskov, A.; Stögbauer, H.; and Grassberger, P. 2004. Estimating Mutual Information. *Physical review E*, 69: 066138.

Lewis, D. D. 1992. Feature Selection and Feature Extraction for Text Categorization. In *Proceedings of the Workshop on Speech and Natural Language*, 212–217. Association for Computational Linguistics.

Lin, D.; and Tang, X. 2006. Conditional infomax learning: An integrated framework for feature extraction and fusion. In *European conference on computer vision*, 68–82. Springer.

Liu, P.; Sohn, H.; and Jeon, I. 2017. Nonlinear spectral correlation for fatigue crack detection under noisy environments. *Journal of Sound and Vibration*, 400: 305–316.

Lopez-Paz, D.; Hennig, P.; and Scholkopf, B. 2013. The Randomized Dependence Coefficient. In *Advances in neural information processing systems (NIPS) 27*. Vancouver, Canada.

Mallah, C.; Cope, J.; and Orwell, J. 2013. Plant Leaf Classification using Probabilistic Integration of Shape, Texture and Margin Features. *IASTED international conference on Signal Processing, Pattern Recognition, and Applications*, 3842.

Meyer, P. E.; Schretter, C.; and Bontempi, G. 2008. Information-theoretic feature selection in microarray data using variable complementarity. *IEEE Journal of Selected Topics in Signal Processing*, 2(3): 261–274.

Peng, H.; Long, F.; and Ding, C. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8): 1226–1238.

Reshef, D. N.; Reshef, Y. A.; Finucane, H. K.; Grossman, S. R.; McVean, G.; Turnbaugh, P. J.; Lander, E. S.; Mitzenmacher, M.; and Sabeti, P. C. 2011. Detecting Novel Associations in Large Data Sets. *Science*, 334(6062): 1518–1524.

Shishkin, A.; Bezzubtseva, A.; Drutsa, A.; Shishkov, I.; Gladkikh, E.; Gusev, G.; and Serdyukov, P. 2016. Efficient high-order interaction-aware feature selection based on conditional mutual information. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 4644–4652.

Sohan, S.; and Príncipe, J. C. 2009. On speeding up computation in information theoretic learning. In *2009 International Joint Conference on Neural Networks*, 2883–2887.

Székely, G. J.; Rizzo, M. L.; Bakirov, N. K.; et al. 2007. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6): 2769–2794.

Székely, G.; and Rizzo, M. 2009. Brownian Distance Covariance. *The Annals of Applied Statistics*, 3: 1236–1265.

Wang, Q.; Shen, Y.; and Zhang, J. Q. 2005. A nonlinear correlation measure for multivariable data set. *Physica D*, 200: 287–295.

Yang, H. H.; and Moody, J. 1999. Data Visualization and Feature Selection: New Algorithms for Nongaussian Data. In *Advances in Neural Information Processing Systems*, 687–693. Cambridge, MA, USA.

Zeng, X.; Xia, Y.; and Tong, H. 2018. Jackknife approach to the estimation of mutual information. *Proceedings of the National Academy of Sciences*, 115(40): 9956–9961.

Zhang, L. 2020. Systematically Exploring Associations among Multivariate Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 6786–6794.