Statistica Sinica Preprint No: SS-2023-0154	
Title	Detect Complete Dependence via Trace Correlation in the Presence of Matrix-Valued Random Objects
Manuscript ID	SS-2023-0154
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0154
Complete List of Authors	Delin Zhao and Liping Zhu
Corresponding Authors	Liping Zhu
E-mails	zhu.liping@ruc.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

DETECT COMPLETE DEPENDENCE VIA TRACE CORRELATION IN THE PRESENCE OF MATRIX-VALUED RANDOM OBJECTS

Delin Zhao and Liping Zhu

Renmin University of China

Abstract: Various metrics have been developed to test for statistical independence and measure the degree of nonlinear dependence between two random objects. Most of these metrics achieve their lower bound if and only if the two random objects are independent. However, it is often unclear how the two random objects are dependent if they attain their upper bound. Moreover, how to implement these metrics when one of the objects is matrix-valued is rarely touched in the literature. To address these issues, we introduce a new metric called trace correlation, which ranges from zero to one. It equals zero only if the two random objects are independent and attains one only if one random object is functionally dependent on the other. In addition, trace correlation allows one of the random objects to be matrix-valued. We estimate trace correlation using standard U-statistic theory and thoroughly study the asymptotic properties of resultant estimates. Furthermore, we adapt trace correlation in the reproducing kernel Hilbert space. Extensive simulations and an application to the MNIST dataset demonstrate the effectiveness and usefulness of trace correlation.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

Key words and phrases: complete dependence, independence test, matrix-valued object, trace correlation.

1. Introduction

Testing for statistical independence and measuring the degree of nonlinear dependence are fundamental issues in both statistics and machine learning communities. In classification problems, to what extent the features are predictive can be evaluated by quantifying the degree of nonlinear dependence between the features and the class labels. Let us use the MNIST dataset accessible at `http://yann.lecun.com/exdb/mnist/` as an example to illustrate this phenomenon. It comprises of $60,000$ training images and $10,000$ test images, each labeled with an integer between 0 and 9. The images have a resolution of 28×28 pixels and their values are scaled to the range of $[0, 1]$. We hide parts of the image each time and contaminate each image with standard Gaussian white noise. The contaminated pixels are subsequently replaced with 1 if they exceed 1, and set to 0 if they are negative, ensuring that pixel values fall within $[0, 1]$. Our objective is to accurately classify images of handwritten digits that are contaminated and partially unseen into one of the 10 classes. To achieve this, we apply LeNet-5 (Lecun et al., 1998), an efficient convolutional neural network specifically

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

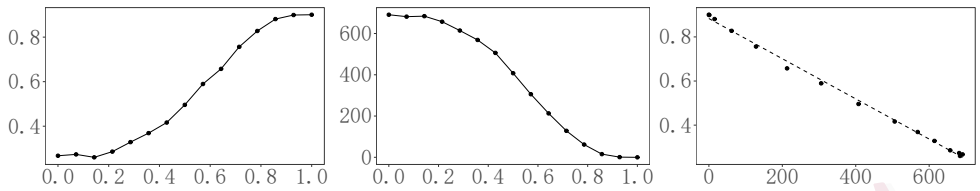
designed for handwritten and machine-printed character recognition. We train the LeNet-5 model on the training set and evaluate its prediction performance on the test set. We use the misclassification rate, the proportion of misclassified observations, to evaluate the prediction power. Additionally, we define the image occlusion ratio as follows:

$$\text{The occlusion ratio} = \frac{\text{The number of hidden pixels}}{\text{The total number of pixels}} \times 100\%.$$

In Figure 1 (a), we demonstrate how misclassification rate varies with the image occlusion ratio, which takes values from $\{0, 1/14, \ldots, 1\} \times 100\%$. As expected, misclassification rate increases with the occlusion ratio. To measure the degree of nonlinear dependence between features and class labels, we introduce trace correlation in this article. Figure 1 (b) shows that the normalized trace correlation decreases as the image occlusion ratio increases. Remarkably, Figure 1 (c) indicates a strong agreement between the normalized trace correlations and the misclassification rates as the image occlusion ratio increases from 0 to 1. This phenomenon suggests that the normalized trace correlation between the features and the class labels is perhaps sufficient to characterize the prediction power without the need to fit complicated nonlinear models.

Many commonly used correlations are designed to test for statistical independence. They achieve the minimum value 0 only if the two random

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154



(a) (b) (c)

Figure 1: (a) Misclassification rates on the vertical axis against occlusion ratios on the horizontal axis; (b) normalized trace correlations against occlusion ratios; and (c) misclassification rates against normalized trace correlations, with a dashed best fitting straight line superimposed.

objects are independent. Examples include distance correlation (Székely et al., 2007; Székely and Rizzo, 2009), Hilbert-Schmidt independence criterion (Gretton et al., 2005, 2007; Sejdinovic et al., 2013), projection correlation (Zhu et al., 2017), ball correlation (Pan et al., 2020), multi-scale graph correlation (Shen et al., 2020), and angle-based correlation (Zhang and Yang, 2024). Székely et al. (2007, Theorem 3) showed that, when the sample distance correlation attains 1, the two random vectors are similar. However, it is unclear that in general situations, if the population distance correlation is 1, how these two random objects are associated. Indeed, for most correlations, it remains unknown how the two random objects are dependent when the correlations reach their maximum value. In contrast,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

Pearson's correlation (Pearson, 1895), Spearman's ρ (Spearman, 1904), and Kendall's τ (Kendall, 1938) are able to measure the strength of linear or monotone relationships. Specifically, the absolute values of these measures are equal to 0 when there is no linear or monotonic dependency between the two objects, and they approach 1 only when one random object appears to be a noiseless linear or monotone function of the other.

Trace correlation possesses a distinctive property: it reaches its upper bound only when one random object is completely dependent on the other (Lancaster, 1963; Kimeldorf and Sampson, 1978). In simpler terms, trace correlation characterizes functional dependence. Most existing correlations quantify deviation from independence based on the discrepancy between joint and marginal distributions. By contrast, trace correlation formulates deviation by comparing conditional and unconditional distributions. It attains the maximum value under functional dependence. Similar idea has been used by Dette et al. (2013), Gamboa et al. (2018), Kong et al. (2019), and Chatterjee (2021) for univariate random variables, and by Yin and Yuan (2020), Ke and Yin (2020), and Deb et al. (2020) for multivariate random vectors. This property makes trace correlation a very useful indicator of prediction performance without the need to fit a nonlinear model, as has been demonstrated in the illustrative example. We further extend

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

the concept of trace correlation into the reproducing kernel Hilbert space, which enables us to characterize nonfunctional dependence.

Another advantage of trace correlation is its ability to handle matrix-valued random objects, which are frequently encountered in the real world but rarely touched in the literature. Matrix-valued random objects arise from the combination of two underlying random variables, such as images constructed from matrices of pixels. Trace correlation allows one random object to be matrix-valued while the other can be either categorical, discrete or continuous, serving as the conditioning variable in trace correlation.

To estimate trace correlation, we use the standard U-statistic theory when the conditioning random object is categorical or discrete, taking a fixed or divergent number of possible values. When the conditioning random object is continuous, we introduce a slicing estimation technique (Li, 1991; Hsing and Carroll, 1992; Zhu and Ng, 1995). In all cases, the resultant asymptotic null distributions do not depend upon the parent distribution of the conditioning variable. In particular, if the conditioning random object takes a divergent number of possible values, the asymptotic null distribution is standard normal. Consequently, there is no need to use bootstrap or random permutations to approximate the asymptotic null distributions, making the implementation of the trace correlation test numerically efficient.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

The asymptotic normality is typically favored by practitioners, particularly by those who are not experts in statistics. We further derive an explicit expression for the asymptotic power against arbitrary fixed alternatives.

This paper is organized as follows. In Section 2, we introduce trace correlation for detecting functional dependence in the presence of matrix-valued random objects. We present comprehensive numerical studies in Section 3 to examine the theoretical properties of trace correlation, followed by a generalization of the concept into the reproducing kernel Hilbert space in Section 4. We conclude this paper with a brief discussion in Section 5. All technical proofs are relegated to an online Supplementary Material.

2. The Trace Correlation

2.1 The rationale

We introduce the rationale of trace correlation first. Define $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{B})$ and $\|\mathbf{A}\|^2 = \langle \mathbf{A}, \mathbf{A} \rangle$, for two generic matrices $\mathbf{A}$ and $\mathbf{B}$. Let $I(E)$ be an indicator function, which equals 1 if the event E is true and 0 otherwise. Let $\mathbf{X} = (X_{k,l}) \in \mathbb{R}^{p \times q}$ be a matrix-valued random object and $Y \in \mathbb{R}^1$ be a univariate one. We allow the conditioning random variable Y to be either categorical, discrete or continuous. We notice that, $\mathbf{X}$ and Y are independent if and only if, for an arbitrary ma-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

trix $\mathbf{B} \in \mathbb{R}^{p\times q}$, $E\{\exp(i\langle \mathbf{B}, \mathbf{X}\rangle) \mid Y\} = E\{\exp(i\langle \mathbf{B}, \mathbf{X}\rangle)\}$ almost surely, or equivalently, $\langle \mathbf{B}, \mathbf{X}\rangle$ and Y are independent. It implies immediately that $E\left\{I(\langle \mathbf{B}, \mathbf{X}\rangle \le x) \mid Y\right\} = E\left\{I(\langle \mathbf{B}, \mathbf{X}\rangle \le x)\right\}$ for all $x \in \mathbb{R}^1$ and $\mathbf{B} \in \mathbb{R}^{p\times q}$. This is equivalent to saying that

$$\mathrm{var}\Big[E\left\{I(\langle \mathbf{B}, \mathbf{X}\rangle \le x) \mid Y\right\}\Big] = 0, \text{ for all } x \in \mathbb{R}^1 \text{ and } \mathbf{B} \in \mathbb{R}^{p\times q}.$$

Unlike most existing correlations that assess deviations from independence through the differences between joint and marginal distributions, the left hand side of the above display quantifies the differences between conditional and unconditional distributions. We advocate using this idea because it simultaneously characterizes statistical independence and complete dependence, on which we shall elaborate shortly. Following Lancaster (1963) and Kimeldorf and Sampson (1978), we define $\mathbf{X}$ to be completely dependent on Y, if there exists a matrix of functions $\mathbf{G} = (G_{k,l}) \in \mathbb{R}^{p\times q}$ such that $\mathrm{pr}\{\mathbf{X} = \mathbf{G}(Y)\} = \mathrm{pr}\{X_{k,l} = G_{k,l}(Y), \text{ for } k = 1, \ldots, p, l = 1, \ldots, q\} = 1$. If $\mathbf{X}$ is completely dependent upon Y, $E\left\{I(\langle \mathbf{B}, \mathbf{X}\rangle \le x) \mid Y\right\} = E\{I(\langle \mathbf{B}, \mathbf{G}(Y)\rangle \le x) \mid Y\} = I(\langle \mathbf{B}, \mathbf{G}(Y)\rangle \le x)$. Accordingly,

$$\mathrm{var}\Big[E\left\{I(\langle \mathbf{B}, \mathbf{X}\rangle \le x) \mid Y\right\}\Big] = \mathrm{var}\Big\{I(\langle \mathbf{B}, \mathbf{X}\rangle \le x)\Big\},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

for all $x \in \mathbb{R}^1$ and $\mathbf{B} \in \mathbb{R}^{p\times q}$. We introduce two positive weight functions, $\omega_1(x)$ and $\omega_2(\mathbf{B})$. The trace correlation is defined as

$$\mathrm{T}_{\mathrm{weight}}(\mathbf{X} \mid Y) = \int_{x\in\mathbb{R}^1}\int_{\mathbf{B}\in\mathbb{R}^{p\times q}} \mathrm{var}\Big[E\left\{I(\langle \mathbf{B}, \mathbf{X}\rangle \le x) \mid Y\right\}\Big] \omega_1(x)\omega_2(\mathbf{B})(d\mathbf{B})(dx) \Big/ \int_{x\in\mathbb{R}^1}\int_{\mathbf{B}\in\mathbb{R}^{p\times q}} \mathrm{var}\Big\{I(\langle \mathbf{B}, \mathbf{X}\rangle \le x)\Big\}\omega_1(x)\omega_2(\mathbf{B})(d\mathbf{B})(dx). \tag{2.1}$$

It follows immediately that $0 \le \mathrm{T}_{\mathrm{weight}}(\mathbf{X} \mid Y) \le 1$, $\mathrm{T}_{\mathrm{weight}}(\mathbf{X} \mid Y) = 0$ if and only if $\mathbf{X}$ and Y are independent because the numerator is zero, and $\mathrm{T}_{\mathrm{weight}}(\mathbf{X} \mid Y) = 1$ if and only if $\mathbf{X}$ is completely dependent upon Y.

To simplify the expression of $\mathrm{T}_{\mathrm{weight}}(\mathbf{X} \mid Y)$ in (2.1), we introduce the following notations. Let $(\mathbf{X}_1, Y_1)$ and $(\mathbf{X}_2, Y_2)$ be two independent copies of $(\mathbf{X}, Y)$. We define

$$d_{\mathrm{weight}}(\mathbf{X}_1, \mathbf{X}_2) = \int_x\int_{\mathbf{B}} \Big\{I(\langle \mathbf{B}, \mathbf{X}_1\rangle \le x) - I(\langle \mathbf{B}, \mathbf{X}_2\rangle \le x)\Big\}^2 \omega_1(x)\omega_2(\mathbf{B})(d\mathbf{B})(dx).$$

If we specify both $\omega_1(x)$ and $\omega_2(\mathbf{B})$ to be standard normal densities (Gupta, 1963; Li and Zhang, 2020; Zhang and Zhu, 2023b), where the standard matrix normal density (Gupta and Nagar, 1999, chapter 2) is defined as

$$\omega_2(\mathbf{B}) = (2\pi)^{-pq/2}\exp(-\|\mathbf{B}\|^2/2),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

the above display reduces to

$$d_{\text{normal}}(\mathbf{X}_1, \mathbf{X}_2) = \arccos\left\{(1 + \langle \mathbf{X}_1, \mathbf{X}_2 \rangle)(1 + \|\mathbf{X}_1\|^2)^{-1/2}(1 + \|\mathbf{X}_2\|^2)^{-1/2}\right\}.$$

We further define $\widetilde{d}_{\text{weight}}(y) = E\{d_{\text{weight}}(\mathbf{X}_1, \mathbf{X}_2) \mid Y_1 = y, Y_2 = y\}$, and $\widetilde{d}_{\text{normal}}(y) = E\{d_{\text{normal}}(\mathbf{X}_1, \mathbf{X}_2) \mid Y_1 = y, Y_2 = y\}$. With these notations, $\mathrm{T}_{\text{weight}}(\mathbf{X} \mid Y) = 1 - E\{\widetilde{d}_{\text{weight}}(Y)\}/E\{d_{\text{weight}}(\mathbf{X}_1, \mathbf{X}_2)\}$ and $\mathrm{T}_{\text{normal}}(\mathbf{X} \mid Y) = 1 - E\{\widetilde{d}_{\text{normal}}(Y)\}/E\{d_{\text{normal}}(\mathbf{X}_1, \mathbf{X}_2)\}$. We summarize the above results in Proposition 1.

Proposition 1. *In general,* $0 \leq \mathrm{T}_{\text{weight}}(\mathbf{X} \mid Y) \leq 1$, $\mathrm{T}_{\text{weight}}(\mathbf{X} \mid Y) = 0$ *if and only if* $\mathbf{X}$ *and* Y *are independent, and* $\mathrm{T}_{\text{weight}}(\mathbf{X} \mid Y) = 1$ *if and only if* $\mathbf{X}$ *is completely dependent upon* Y*. In addition,* $\mathrm{T}_{\text{weight}}(\mathbf{X} \mid Y) = 1 - E\{\widetilde{d}_{\text{weight}}(Y)\}/E\{d_{\text{weight}}(\mathbf{X}_1, \mathbf{X}_2)\}$*. In particular, if we choose both* $\omega_1(x)$ *and* $\omega_2(\mathbf{B})$ *to be standard normal densities, then* $\mathrm{T}_{\text{normal}}(\mathbf{X} \mid Y) = 1 - E\{\widetilde{d}_{\text{normal}}(Y)\}/E\{d_{\text{normal}}(\mathbf{X}_1, \mathbf{X}_2)\}$.

We choose both $\omega_1(x)$ and $\omega_2(\mathbf{B})$ to be standard normal densities to yield a closed expression $\mathrm{T}_{\text{normal}}(\mathbf{X} \mid Y)$. An additional advantage is that the resultant trace correlation is robust to the presence of outliers or extreme values because it does not impose any moment conditions on $\mathbf{X}$ or Y. Indeed as long as both $\omega_1(x)$ and $\omega_2(\mathbf{B})$ are probability density functions, $d_{\text{weight}}(\mathbf{X}_1, \mathbf{X}_2)$, and hence $E\{d_{\text{weight}}(\mathbf{X}_1, \mathbf{X}_2)\}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

are always bounded, regardless of the distribution of $\mathbf{X}$. While there are many other choices of probability densities that yield a closed form of $\mathrm{T}_{\text{weight}}(\mathbf{X} \mid Y)$, we choose standard normal densities for simplicity. By contrast, if we choose $\omega_1(x) = \omega_2(\mathbf{B}) = 1$, $d_{\text{weight}}(\mathbf{X}_1, \mathbf{X}_2) = \|\mathbf{X}_1 - \mathbf{X}_2\|$, and if $q = 1$, $\mathrm{T}_{\text{weight}}(\mathbf{X} \mid Y)$ boils down to the squared expectation of conditional difference (Yin and Yuan, 2020; Xu and Zhu, 2022). To ensure $E\big\{d_{\text{weight}}(\mathbf{X}_1, \mathbf{X}_2)\big\} = E(\|\mathbf{X}_1 - \mathbf{X}_2\|)$ to be finite, it is often assumed that $E\|\mathbf{X}\| < \infty$, which imposes a stringent moment condition on $\mathbf{X}$. In Table 3 we shall show that, if this moment condition is violated, the squared expectation of conditional difference is less efficient in detecting nonlinear dependence.

Subsequently, we work with $\mathrm{T}_{\text{normal}}(\mathbf{X} \mid Y)$, which use $d_{\text{normal}}(\mathbf{X}_1, \mathbf{X}_2)$ and $\widetilde{d}_{\text{normal}}(Y)$ and does not have to use integral in high-dimensional space. For the sake of notational clarity, we suppress the subscript "normal" and refer to as $\mathrm{T}(\mathbf{X} \mid Y)$, $d(\mathbf{X}_1, \mathbf{X}_2)$ and $\widetilde{d}(Y)$ henceforth.

2.2 The estimation

Estimating $\mathrm{T}(\mathbf{X} \mid Y)$ amounts to estimating the denominator, $E\big\{d(\mathbf{X}_1, \mathbf{X}_2)\big\}$, and the numerator, $E\big\{\widetilde{d}(Y)\big\}$, respectively. Suppose a random sample of size n, $\{(\mathbf{X}_i, Y_i), i = 1, \ldots, n\}$, is available. We propose to estimate the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

denominator, $T_1 = E\{d(\mathbf{X}_1, \mathbf{X}_2)\}$, with the standard U-statistic theory. Specifically, we estimate T_1 with

$$\widehat{T}_1 = \{n(n-1)/2\}^{-1} \sum_{1 \le i < j \le n} d(\mathbf{X}_i, \mathbf{X}_j).$$

Estimating the numerator, $T_2 = E\{\widetilde{d}(Y)\}$, is nontrivial, especially when Y is continuous. We first consider the easier case where Y is categorical or discrete, then extend the method to continuous Y by introducing a slicing estimation procedure.

Let us begin by assuming Y is categorical with class labels $\{1, \ldots, H\}$, without loss of generality. We denote the number of observations in the h-th class by n_h, and define $p_h = \text{pr}(Y = h)$ and $\widehat{p}_h = n_h/n$. To facilitate subsequent illustration, we sort the observations $\{(\mathbf{X}_i, Y_i), i = 1, \ldots, n\}$ according to their class labels, such that the sorted observations $\{(\mathbf{X}_{(i)}, Y_{(i)}), i = 1, \ldots, n\}$ satisfy $Y_{(i)} = h$ if $n_1 + \cdots n_{h-1} < i \le n_1 + \cdots n_h$, and $\mathbf{X}_{(i)}$ is the respective concomitant of $Y_{(i)}$. We further introduce the double-subscript notation system (h, i), where the first subscript h represents the class label, and the second subscript i stands for the order number of an observation in the h-th class. It follows that $\mathbf{X}_{(h,i)} = \mathbf{X}_{(n_1 + \cdots + n_{h-1} + i)}$ for $i = 1, \ldots, n_h$ and $h = 1, \ldots, H$. By definition,

$$T_2 = E\{\widetilde{d}(Y)\} = \sum_{h=1}^{H} \widetilde{d}(h) p_h.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

This motivates us to define the estimator

$\widehat{T}_2 = \sum_{h=1}^{H} \widehat{d}(h)\widehat{p}_h$, where $\widehat{d}(h) = \{n_h(n_h-1)/2\}^{-1} \sum_{1\le i<j\le n_h} d(\mathbf{X}_{(h,i)}, \mathbf{X}_{(h,j)})$ is a U-statistic within each class. Accordingly, $\widehat{\mathrm{T}}(\mathbf{X} \mid Y) = 1 - \widehat{T}_2 \,/\, \widehat{T}_1$.

When Y is discrete and takes values in $\{1, \ldots, H\}$, estimating $\mathrm{T}(\mathbf{X} \mid Y)$ follows the same process as in the categorical case. When Y is continuous, we divide the sorted observations $\{(\mathbf{X}_{(i)}, Y_{(i)}), i = 1, \ldots, n\}$ into H slices based on the values of $\{Y_{(1)}, \ldots, Y_{(n)}\}$, such that the h-th slice contains n_h observations. To be specific, we partition the entire range of the conditioning variable Y into H disjoint intervals, $I_1, \ldots, I_H$, and assign $(\mathbf{X}_{(i)}, Y_{(i)})$ satisfying $Y_{(i)} \in I_h$ into the h-th slice. This allocation scheme ensures that all n observations fall into one of the H slices. The idea of slicing was originated from Li (1991), Hsing and Carroll (1992) and Zhu and Ng (1995). Each slice in the continuous case corresponds to a category in the categorical case, making subsequent estimation after slicing similar in spirit. We omit details on estimation for both discrete and continuous cases.

The computational complexity of the sorting algorithm is $O(n \log n)$, and that of calculating $d(\mathbf{X}_i, \mathbf{X}_j)$ is $O(pq)$. This leads to complexities of

$$O\big\{n(n-1)pq\big\} \text{ and } O\left\{\sum_{h=1}^{H} n_h(n_h-1)pq + n\log n\right\}$$

for calculating $\widehat{T}_1$ and $\widehat{T}_2$, respectively. If all n_hs are approximately the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

same order, say, $n_h = O(n/H)$, then the complexity of calculating $\widehat{T}_2$ is

$$O\left\{\sum_{h=1}^{H}(n/H)(n/H-1)pq + n\log n\right\} = O(n^2pq/H + n\log n),$$

and the total complexity of calculating $\widehat{\mathrm{T}}(\mathbf{X} \mid Y)$ is dominated by computing $\widehat{T}_1$, which has the order of $O(n^2pq)$. In the massive data scenario where the computational efficiency is a primary concern, we follow Zhang and Zhu (2023a) and propose a block-wise estimation procedure to alleviate the complexity of calculating $\widehat{T}_1$. To be precise, we divide the whole sample of size n into B blocks, each of size $n_b^B = O(n/B)$. We define the block-wise estimation of T_1 by

$$\widehat{T}_{1,B} = B^{-1}\sum_{b=1}^{B}\left[\{n_b^B(n_b^B-1)/2\}^{-1}\sum_{1\le i<j\le n_b^B} d(\mathbf{X}_{b,i},\mathbf{X}_{b,j})\right],$$

where $\mathbf{X}_{b,i}$ represents the i-th observation in the b-th block. We define the block-wise estimation of $\mathrm{T}(\mathbf{X} \mid Y)$ as $\widehat{\mathrm{T}}_B(\mathbf{X} \mid Y) = 1 - \widehat{T}_2/\widehat{T}_{1,B}$, which has the complexity of $O\{n^2pq(1/B+1/H)\}$. This is a substantial improvement over calculating $\widehat{\mathrm{T}}(\mathbf{X} \mid Y)$ when both B and H are divergent.

2.3 The asymptotic behaviors

Before stating the asymptotic properties of the above estimates, we introduce the following notations. We assume for now that the conditioning variable Y is categorical with class labels $\{1, 2, \ldots, H\}$ or discrete

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

taking the same values, where H can be either fixed or divergent. We assume $i \neq j$, $i \neq k$ and $j \neq k$ unless stated otherwise. By definition, $T_1 = E\big\{d(\mathbf{X}_i, \mathbf{X}_j)\big\}$ and $\widetilde{d}(h) = E\left\{d(\mathbf{X}_i, \mathbf{X}_j) \mid Y_i = Y_j = h\right\}$. Define $\sigma^2 = \mathrm{var}\{d(\mathbf{X}_i, \mathbf{X}_j) - d_1(\mathbf{X}_i) - d_1(\mathbf{X}_j)\}$ for $d_1(\mathbf{X}_i) = E\left\{d(\mathbf{X}_i, \mathbf{X}_j) \mid \mathbf{X}_i\right\}$. Let $\varepsilon_{i,j,h} = d(\mathbf{X}_i, \mathbf{X}_j) - \widetilde{d}(h)$, $r_{i,h} = d_1(\mathbf{X}_i) - E\{d_1(\mathbf{X}_i) \mid Y_i = h\}$, $V_1(h) = \mathrm{cov}(\varepsilon_{i,j,h}, r_{i,h} \mid Y_i = Y_j = h)$ and $V_2(h) = \mathrm{cov}(\varepsilon_{i,j,h}, \varepsilon_{i,k,h} \mid Y_i = Y_j = Y_k = h)$. Let $\tau_1 = \mathrm{var}\Big[\widetilde{d}(Y) - 2\{1 - \mathrm{T}(\mathbf{X} \mid Y)\}E\{d_1(\mathbf{X}) \mid Y\}\Big]$ and $\tau_2 = E\{V_2(Y)\} - 2\{1 - \mathrm{T}(\mathbf{X} \mid Y)\}E\{V_1(Y)\} + \{1 - \mathrm{T}(\mathbf{X} \mid Y)\}^2 E[\mathrm{var}\{d_1(\mathbf{X}) \mid Y\}]$. In addition, $\tau^2 = \tau_1 + 4\tau_2$. We remark here that, σ^2 depends on the distribution of $\mathbf{X}$ only, while τ^2 depends on the joint distribution of $(\mathbf{X}, Y)$.

Theorem 1. *Suppose Y is categorical with class labels $\{1, 2, \ldots, H\}$ or discrete taking the same values. The following convergence results assume implicitly that $n \to \infty$.*

1. *$\widehat{\mathrm{T}}(\mathbf{X} \mid Y)$ converges in probability to $\mathrm{T}(\mathbf{X} \mid Y)$.*

2. *Assume $\mathbf{X}$ is independent of Y. If H remains fixed as $n \to \infty$, then var$\big\{n\big(\widehat{T}_1 - \widehat{T}_2\big)\big\} \to 2(H-1)\sigma^2$, and*

$$n\widehat{\mathrm{T}}(\mathbf{X} \mid Y) \xrightarrow{d} (H-1)\sum_{k=1}^{\infty} \lambda_k (X_k^2 - 1),$$

where all X_ks are independent standard normal, and all λ_ks are positive constants depending on the distribution of $\mathbf{X}$ only. The summa-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

tion of all λ_ks is one.

If $H \to \infty$ as $n \to \infty$, allowing for $H = O(n)$, then var$\{(nc_n)^{1/2}(\widehat{T}_1 - \widehat{T}_2)\} \to 2\sigma^2$, and $(nc_n)^{1/2}\widehat{\mathrm{T}}(\mathbf{X} \mid Y) \xrightarrow{d} \mathcal{N}(0, 2\sigma^2/T_1^2)$, where

$$c_n^{-1} = \sum_{h=1}^{H} n_h/\{n(n_h - 1)\}.$$

3. *Assume $\mathbf{X}$ is dependent but not completely dependent upon Y, H is either fixed, or divergent with $H = o(n)$ and $\max_h p_h = O(n^{-\alpha})$ for some $0 < \alpha < 1$. Then $n^{1/2}\{\widehat{\mathrm{T}}(\mathbf{X} \mid Y) - \mathrm{T}(\mathbf{X} \mid Y)\}/(\tau/T_1) \xrightarrow{d} \mathcal{N}(0, 1)$.*

4. *Assume $\mathbf{X}$ is completely dependent upon Y. Then $pr\{\widehat{\mathrm{T}}(\mathbf{X} \mid Y) = 1\} = 1$.*

Next we describe the asymptotic properties when the conditioning variable Y is continuous. We shall use the following notations. Define $m(Y_i, Y_j) = E\{d(\mathbf{X}_i, \mathbf{X}_j) \mid Y_i, Y_j\}$, $\varepsilon_{i,j} = d(\mathbf{X}_i, \mathbf{X}_j) - m(Y_i, Y_j)$, $r_i = d_1(\mathbf{X}_i) - E\{d_1(\mathbf{X}_i) \mid Y_i\}$, $V_3(Y_i, Y_j) = \mathrm{cov}(\varepsilon_{i,j}, r_i \mid Y_i, Y_j)$, $V_4(Y_i, Y_j, Y_k) = \mathrm{cov}(\varepsilon_{i,j}, \varepsilon_{i,k} \mid Y_i, Y_j, Y_k)$. We further define $\tau_3 = \mathrm{var}\Big[m(Y, Y) - 2\{1 - \mathrm{T}(\mathbf{X} \mid Y)\}E\{d_1(\mathbf{X}) \mid Y\}\Big]$ and $\tau_4 = E\{V_4(Y, Y, Y)\} - 2\{1 - \mathrm{T}(\mathbf{X} \mid Y)\}E\{V_3(Y, Y)\} + \{1 - \mathrm{T}(\mathbf{X} \mid Y)\}^2 E[\mathrm{var}\{d_1(\mathbf{X}) \mid Y\}]$. We remark here that $m(Y, Y)$, $V_3(Y, Y)$ and $V_4(Y, Y, Y)$ correspond to $\widetilde{d}(Y)$, $V_1(Y)$ and $V_2(Y)$, respectively. In addition, we define $\tau_s^2 = \tau_3 + 4\tau_4$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

We define a family of functions $\{f(\mathbf{x}, \cdot)\}$ to be uniformly non-expansive in the metric of $M(\cdot)$ if, for any two points $\mathbf{y}_j$ and $\mathbf{y}_k$,

$$\sup_{\mathbf{x}} |f(\mathbf{x}, \mathbf{y}_j) - f(\mathbf{x}, \mathbf{y}_k)| \leq |M(\mathbf{y}_j) - M(\mathbf{y}_k)| .$$

We assume the following conditions on the smoothness.

(A1) The family of functions, $\{m(y_i, \cdot)\}$ is uniformly non-expansive in the metric of $M(\cdot)$, where $M(\cdot)$ has a total variation of order γ and is non-decreasing over $(-\infty, -B_0]$ and $[B_0, \infty)$. In addition, $\max\{|M(y_i)|^{1/\gamma}, |M(-y_i)|^{1/\gamma}\} \operatorname{pr}(|Y_i| > y_i) \to 0$ as $y_i \to \infty$.

(A2) The families of functions, $\{V_3(y_i, \cdot)\}$ and $\{V_4(y_i, y_j, \cdot)\}$ are uniformly non-expansive in the metric of $V(\cdot)$, where $V(\cdot)$ has a total variation of order ξ and is non-decreasing over $(-\infty, -B_0']$ and $[B_0', \infty)$; In addition, $\max\{|V(y_i)|^{1/\xi}, |V(-y_i)|^{1/\xi}\} \operatorname{pr}(|Y_i| > y_i) \to 0$ as $y_i \to \infty$.

(A3) $\max_h n_h = O(n^\alpha)$, where $\alpha = \min(1/2 - \gamma, 1 - \xi) > 0$.

The conditions of total variation and non-expansiveness are widely used in the literature. See, for example, Hsing and Carroll (1992), Zhu and Ng (1995), Zhu et al. (2006), Li and Zhu (2007) and Lin et al. (2018). We extend the concept of non-expansiveness to uniform non-expansiveness to accommodate multivariate functions.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

Theorem 2. *Suppose Y is continuous. The following convergence results assume implicitly that $H \to \infty$ as $n \to \infty$.*

1. *Assume $\mathbf{X}$ is independent of Y. We allow for $H = O(n)$, then*

$$(nc_n)^{1/2}\widehat{\mathrm{T}}(\mathbf{X} \mid Y) \xrightarrow{d} \mathcal{N}(0, 2\sigma^2/T_1^2).$$

2. *Assume $\mathbf{X}$ is dependent but not completely dependent upon Y, and $H = o(n)$. Under Conditions (A1)–(A3),*

$$n^{1/2}\{\widehat{\mathrm{T}}(\mathbf{X} \mid Y) - \mathrm{T}(\mathbf{X} \mid Y)\}/(\tau_s/T_1) \xrightarrow{d} \mathcal{N}(0, 1).$$

3. *Assume $\mathbf{X}$ is completely dependent upon Y. Under Condition (A1),*

$$\widehat{\mathrm{T}}(\mathbf{X} \mid Y) - 1 = o_p\left(\max_h n_h \Big/ n^{1-\gamma}\right).$$

The number of slices, H, is a crucial factor that affects the asymptotic properties discussed above. Theorem 2 implies that, as long as H is not too small, the slicing procedure always provides a consistent estimation when the conditioning variable Y is continuous. If $\mathbf{X}$ and Y are dependent but not completely dependent, the asymptotic variance of slicing estimation remains unaffected by H, which appears to be a very surprising phenomenon. However, if $\mathbf{X}$ completely dependent upon Y, H does affect the convergence rate of $\widehat{\mathrm{T}}(\mathbf{X} \mid Y)$ indirectly through the quantity $\max_h n_h$. It is worth noting

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

that H affects the power performance of the trace correlation test. Suppose we aim to test whether $\mathbf{X}$ and Y are independent. By Theorem 2, we reject the null hypothesis H_0 if $(nc_n/2)^{1/2}\widehat{\mathrm{T}}(\mathbf{X} \mid Y) \geq z_{1-\alpha}(\sigma/T_1)$, where $z_{1-\alpha}$ represents the $(1-\alpha)\times 100\%$-th quantile of standard normal distribution. Let $\Phi(\cdot)$ stand for the cumulative function of standard normal. The power of the trace correlation test is

$$\Phi\Big[\tau_s^{-1}\{n^{1/2}(T_1 - T_2) - (c_n/2)^{-1/2}(\sigma z_{1-\alpha})\}\Big],$$

which decreases as H increases. It is thus recommended to use a relatively small H to enhance power of the trace correlation test.

To implement the trace correlation test in practice, we let $\widehat{\sigma}^2$ be a consistent estimate of σ^2, and define the normalized trace correlation by $\widehat{\mathrm{T}}_N(\mathbf{X} \mid Y) = (nc_n)^{1/2}\widehat{\mathrm{T}}(\mathbf{X} \mid Y)\Big/\big(2\widehat{\sigma}^2/\widehat{T}_1^2\big)^{1/2}$, which serves as the test statistic. By Theorems 1 and 2, $\widehat{\mathrm{T}}_N(\mathbf{X} \mid Y)$ is asymptotically standard normal under the null hypothesis H_0 when H is divergent. We reject H_0 if $\widehat{\mathrm{T}}_N(\mathbf{X} \mid Y) \geq z_{1-\alpha}$ at the significance level α.

3. Numerical Studies

We demonstrate the theoretical properties and the usefulness of trace correlation through three synthetic examples.

Example 1. Let Y be a univariate categorical variable with H categories,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

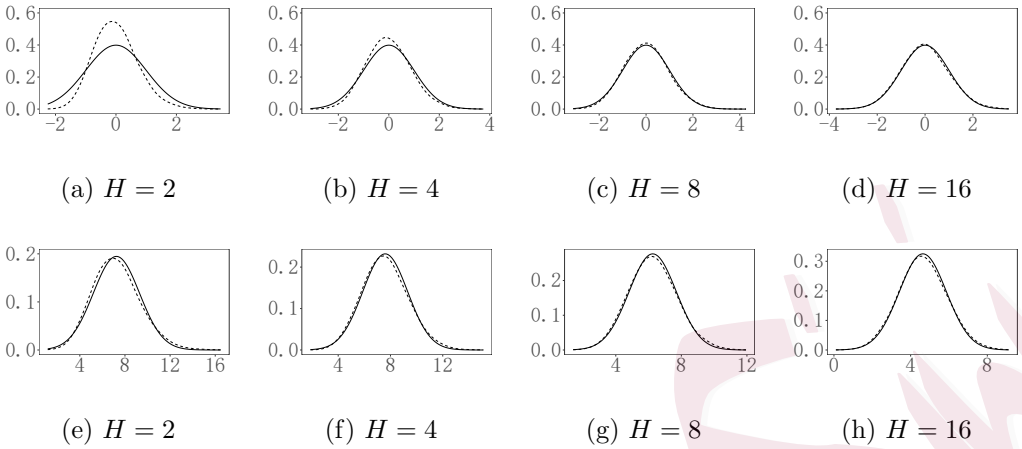
where each category has a probability of occurrence $\text{pr}(Y = h) = 1/H$, for $h = 1, \ldots, H$. We set H to $\{2, 4, 8, 16\}$ and fix the dimensions of matrices to $p = q = 10$. The centering matrices, $\mathbf{C}_h \in \mathbb{R}^{p\times q}$, for $h = 1, \ldots, H$, are generated from a matrix normal distribution $\mathcal{N}_{p,q}(\mathbf{M}, \mathbf{U}, \mathbf{V})$ with mean $\mathbf{M} = \mathbf{0}_{p\times q}$, row covariances $\mathbf{U} = (U_{k,l})_{p\times p}$, where $U_{k,l} = 0.5^{|k-l|}$, for $k, l = 1, \ldots, p$ and column covariances $\mathbf{V} = \mathbf{I}_{q\times q}$. We draw the error matrix $\mathbf{E}$ from a matrix t-distribution $\mathcal{M}_{p,q}(\nu, \mathbf{M}, \mathbf{U}, \mathbf{V})$ using $\nu = 1$ degree of freedom, the same mean, row and column covariances as the matrix normal distribution. We consider two scenarios for generating $\mathbf{X} \in \mathbb{R}^{p\times q}$:

(1) Let $\text{pr}(\mathbf{X} = \mathbf{C}_h + 5\mathbf{E}) = 1/H$, for $h = 1, \ldots, H$.

(2) Let $\mathbf{X} = \mathbf{C}_h + 5\mathbf{E}$ if $Y = h$, for $h = 1, \ldots, H$.

We remark here that $\mathbf{X}$ and Y are independent in the first scenario and dependent in the second. We set the sample size $n = 256$ and repeat this data generating process for $10,000$ times. The normalized trace correlations are shown in Figure 2. We observe that if $\mathbf{X}$ and Y are independent, the normalized trace correlations are normally distributed, as long as H is not too small, say, $H \geq 8$. If X and Y are dependent, the normalized trace correlations are asymptotically normal for a wide range of H. Such findings are exactly in line with our observations in Theorem 1.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154



(a) $H = 2$ (b) $H = 4$ (c) $H = 8$ (d) $H = 16$

(e) $H = 2$ (f) $H = 4$ (g) $H = 8$ (h) $H = 16$

Figure 2: The dashed lines stand for the kernel densities of normalized trace correlations obtained from 10,000 repetitions, and the solid lines stand for the normal densities, which serve as reference curves. $\mathbf{X}$ and Y are independent in (a)–(d) and dependent in (e)–(h). Therefore, the normal densities have mean zero only in (a)–(d).

Next we let Y be a continuous random variable drawn from the uniform distribution on $[-1, 1]$. We generate $\mathbf{E} = (E_{k,l})_{p\times q}$ from the matrix normal distribution $\mathcal{N}_{p,q}(\mathbf{M}, \mathbf{U}, \mathbf{V})$. Let $\mathbf{X} = (X_{k,l})_{p\times q}$, where $X_{k,l} = E_{k,l}$ for $k = 1, \ldots, p$ and $l = m + 1, \ldots, q$. Moreover, for the first m columns of $\mathbf{X}$, i.e., $X_{k,l}$ for $k = 1, \ldots, p$ and $l = 1, \ldots, m$, we generate from one of the following three models:

(3) $X_{k,l} = 0.5Y^2 + \delta E_{k,l}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

(4) $X_{k,l} = 1.2\big\{|Y + 0.5|I(Y < 0) + |Y - 0.5|I(Y \geq 0)\big\} + \delta E_{k,l}$.

(5) $X_{k,l} = 0.25\cos(4\pi Y) + \delta E_{k,l}$.

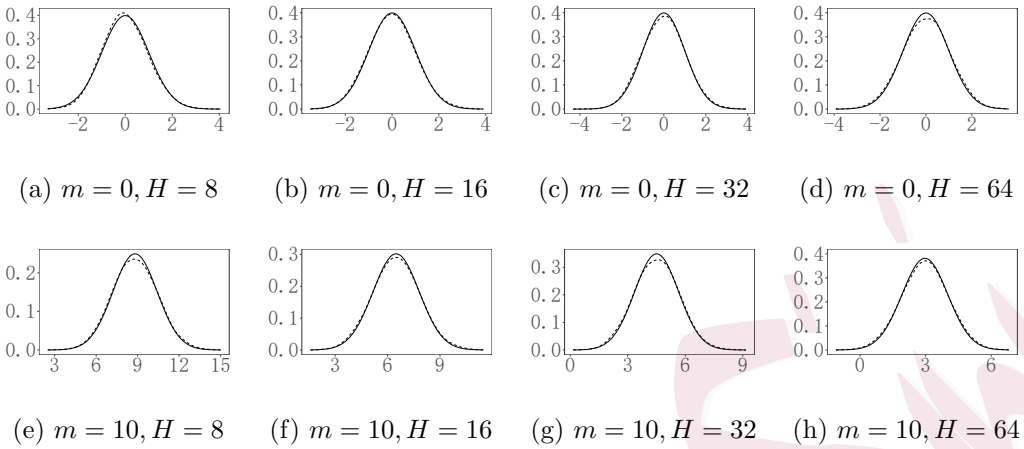
The above models have also been used by Heller et al. (2013), Kong et al. (2019) and Chatterjee (2021). We fix $p = q = 10$ and vary m from 0 to 10. Specifically, when $m = 0$, $\mathbf{X}$ and Y are independent, whereas when $m = 10$ and $\delta = 0$, $\mathbf{X}$ is functionally dependent on Y.

We set $n = 256$, and divide Y uniformly into H slices with $H = \{8, 16, 32, 64\}$. The normalized trace correlations of Model (3) with $\delta = 1$ are depicted in Figure 3. The results indicate that the normalized trace correlations follow normal distribution for all choices of H, regardless of how $\mathbf{X}$ and Y are dependent. These findings align with Theorem 2.

To investigate the relationship between the power of the trace correlation test and H, we consider Models (3)–(5) with $\delta = 1$. We vary m over $\{1, 2, 5, 10\}$ to increase the strength of dependence between $\mathbf{X}$ and Y. Let $H = \{16, 32, 64\}$ in the trace correlation test. The empirical powers of the trace correlation test based on $10,000$ repetitions are summarized in Table 1. It can be seen that the power of the trace correlation test increases as H decreases. This is in line with our theoretical observations in Theorem 2.

Finally we demonstrate trace correlation's ability to measure the strength of nonlinear dependence. In Models (3)–(5) with $m = 10$, we vary δ in

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154



(a) $m = 0, H = 8$ (b) $m = 0, H = 16$ (c) $m = 0, H = 32$ (d) $m = 0, H = 64$

(e) $m = 10, H = 8$ (f) $m = 10, H = 16$ (g) $m = 10, H = 32$ (h) $m = 10, H = 64$

Figure 3: The dashed lines stand for the kernel densities of normalized trace correlations obtained from 10,000 repetitions when H varies over $\{8, 16, 32, 64\}$, and the solid lines stand for the normal densities. All results are based on Model (3) only. In addition, $\mathbf{X}$ and Y are independent in (a)–(d) and dependent in (e)–(h).

$\{1.00, 0.50, 0.25, 0.00\}$ to increase the degree of functional dependence. For trace correlation, we vary H over $\{16, 32, 64\}$. The averages (and standard deviations) of trace correlations based on 10,000 repetitions for Models (3)–(5) are displayed in Table 2. The results show that decreasing δ leads to an increase in the trace correlation. Moreover, when δ is close to zero, the trace correlations are almost equal to one with very small variability. We also note that when $\delta = 0$, increasing H reduces the gap between trace

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

Table 1: The empirical powers of the trace correlation test based on 10,000 repetitions for Models (3)–(5) with $m = \{1, 2, 5, 10\}$ and $H = \{16, 32, 64\}$.

Model	H	$m = 1$	$m = 2$	$m = 5$	$m = 10$
	16	19.0	42.8	94.7	100.0
(3)	32	13.5	26.3	76.0	99.4
	64	10.1	16.9	47.1	89.2
	16	25.2	56.9	98.7	100.0
(4)	32	17.0	37.6	90.9	100.0
	64	12.4	23.1	65.2	97.5
	16	21.5	49.9	97.8	100.0
(5)	32	17.5	38.2	93.0	100.0
	64	12.5	24.3	71.9	99.5

correlations and one, which echoes Theorem 2.

Example 2. In this example, we compare the power performance of several independence tests. We refer to the trace correlation test as TC. In addition, we include the following popular competitors into comparison:

1. The distance correlation test (Székely et al., 2007; Székely and Rizzo, 2009, DC for short), implemented with the `dcor.test` function in the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

Table 2: The averages (and standard deviations) of trace correlations based on 10,000 repetitions for Models (3)–(5) with $\delta = \{1.00, 0.50, 0.25, 0.00\}$ and $H = \{16, 32, 64\}$.

Model	H	$\delta = 1.00$	$\delta = 0.50$	$\delta = 0.25$	$\delta = 0.00$
(3)	16	0.012(0.002)	0.036(0.004)	0.086(0.005)	0.888(0.005)
	32	0.012(0.003)	0.037(0.004)	0.087(0.006)	0.941(0.003)
	64	0.012(0.004)	0.037(0.005)	0.087(0.006)	0.967(0.002)
(4)	16	0.014(0.003)	0.039(0.004)	0.080(0.006)	0.691(0.026)
	32	0.015(0.003)	0.041(0.004)	0.086(0.006)	0.825(0.013)
	64	0.015(0.004)	0.041(0.005)	0.087(0.007)	0.899(0.008)
(5)	16	0.015(0.003)	0.052(0.004)	0.147(0.008)	0.619(0.024)
	32	0.017(0.003)	0.062(0.005)	0.176(0.008)	0.794(0.010)
	64	0.018(0.005)	0.065(0.006)	0.185(0.008)	0.885(0.006)

`energy` package.

2. The Hilbert–Schmidt independence criterion test (Gretton et al., 2005, 2007, HSIC for short), implemented with the `dhsic.test` function in the `dHSIC` package.

3. The Heller-Heller-Gorfine test (Heller et al., 2013, HHG for short)

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

based on ranks of distances, implemented with the `hhg.test` function in the `HHG` package.

4. The expectation of the conditional difference test (Yin and Yuan, 2020, ECD for short).

5. The expected conditional characteristic function-based independence criterion test (Ke and Yin, 2020, ECCFIC for short).

To implement these tests for Models (1)–(2) in Example 1, we vectorize the matrix $\mathbf{X} \in \mathbb{R}^{p\times q}$ into a vector $\mathbf{x} \in \mathbb{R}^{pq\times 1}$. When Y falls into the h-th category, for $h = 1, \ldots, H$, we define $\mathbf{y} \in \mathbb{R}^{H\times 1}$ as a standard unit vector with its h-th element being 1 and all other elements being 0 for the first three tests. We follow Gretton et al. (2007) and Ke and Yin (2020) and implement the HSIC and ECCFIC tests using the Gaussian kernel. We set the sample size $n = 256$, the row dimension $p = 10$, and the significance level $\alpha = 0.05$. We repeat each scenario 1,000 times, and conduct all comparisons in the `R` language. We run 500 random permutations to approximate the asymptotic null distributions. The sizes of all tests for Model (1) are close to 0.05; hence, we report only the powers for Model (2) in what follows.

We vary the column dimension of $\mathbf{X}$, q, in $\{1, 2, 5, 10, 20, 50\}$, and the number of categories, H, in $\{8, 16\}$. The empirical powers are summarized

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

Table 3: The empirical powers of all six independence tests based on Model (2) in Example 1, where the number of categories $H = \{8, 16\}$ and the column dimension $q = \{1, 2, 5, 10, 20, 50\}$.

H	Test	$q = 1$	$q = 2$	$q = 5$	$q = 10$	$q = 20$	$q = 50$
8	TC	85.7	91.7	99.1	100.0	100.0	100.0
	DC	12.3	9.6	9.4	9.9	9.2	10.0
	HSIC	34.3	26.0	23.2	20.1	20.1	23.7
	HHG	21.7	17.1	15.5	13.6	13.2	13.9
	ECD	11.1	8.4	8.8	8.9	8.7	9.0
	ECCFIC	33.8	26.6	22.6	21.0	19.3	22.7
16	TC	68.4	78.9	94.7	99.1	100.0	100.0
	DC	9.7	10.3	10.0	9.2	8.2	8.2
	HSIC	23.1	18.6	17.2	16.9	14.5	15.5
	HHG	12.6	12.1	9.7	8.9	8.5	9.9
	ECD	7.6	6.5	6.4	7.3	7.0	6.4
	ECCFIC	21.9	17.5	17.2	16.1	13.9	15.1

in Table 3. The DC and HSIC tests are known to lose power as the dimension increases (Zhu et al., 2020; Zhang and Zhu, 2023b). It is thus not

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

surprising to see that these tests are not very powerful when the matrix-valued object $\mathbf{X}$ is vectorized. The power performances of the ECD and ECCFIC tests are similar to those of the DC and HSIC tests, respectively. The trace correlation test outperforms these competitors across all scenarios, and its superiority is particularly evident when the number of categories is relatively smaller.

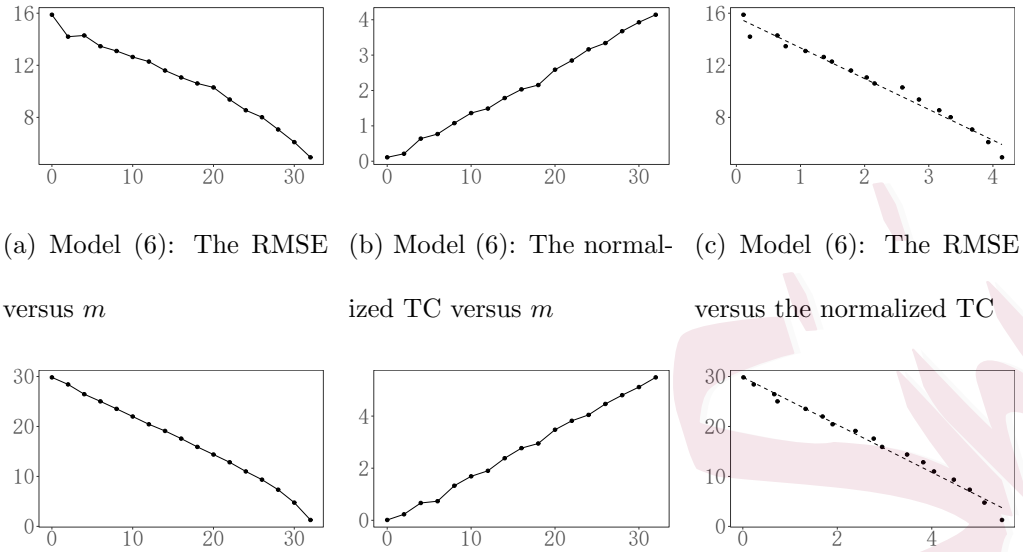
Example 3. We demonstrate the utility of trace correlation between $\mathbf{X} = (X_{k,l}) \in \mathbb{R}^{p\times q}$ and Y in quantifying the prediction power of $\mathbf{X}$ on Y. We consider two models of trace regression that link Y to $\mathbf{X}$ through $\mathbf{Z} = (Z_{k,l}) \in \mathbb{R}^{p\times q}$:

(6) Poisson trace regression: $(Y \mid \mathbf{Z})$ follows Poisson distribution with mean function $\exp(1 + \langle \mathbf{B}, \mathbf{Z}\rangle)$.

(7) Normal trace regression: $(Y \mid \mathbf{Z})$ follows normal distribution with mean function $(1 + \langle \mathbf{B}, \mathbf{Z}\rangle)$ and unit variance.

In both models, $\mathbf{B} = \mathbf{B}_1^{\mathrm{T}}\mathbf{B}_2$, where the rank of $\mathbf{B}$ is r, $\mathbf{B}_1 \in \mathbb{R}^{r\times p}$ and $\mathbf{B}_2 \in \mathbb{R}^{r\times q}$. All entries of $\mathbf{B}_1$ and $\mathbf{B}_2$ are independent and uniformly distributed on $[-1, 1]$. Let $\mathbf{E} = (E_{k,l}) \in \mathbb{R}^{p\times q}$ and $\mathbf{Z} = (Z_{k,l}) \in \mathbb{R}^{p\times q}$. All entries of $\mathbf{E}$ and $\mathbf{Z}$ are independent standard normal. For $k = 1, \ldots, p$, we define $X_{k,l} = Z_{k,l}$, if $l = 1, \ldots, m$, and $X_{k,l} = E_{k,l}$, if $l = m + 1, \ldots, q$. In other words, we

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154



(a) Model (6): The RMSE versus m

(b) Model (6): The normalized TC versus m

(c) Model (6): The RMSE versus the normalized TC

(d) Model (7): The RMSE versus m

(e) Model (7): The normalized TC versus m

(f) Model (7): The RMSE versus the normalized TC

Figure 4: The root mean square errors and trace correlations are abbreviated to RMSE and TC. The best fitting straight lines (dashed) are superimposed in (c) and (f).

contaminate the last $(p-m)$ columns of $\mathbf{Z}$ to form $\mathbf{X} = (X_{k,l}) \in \mathbb{R}^{p\times q}$. We set $p = q = 32$ and $r = 2$, and vary $m = \{0, 2, 4, \ldots, 32\}$. The size of both the training and test sets is $n = 1024$.

Because the observations are generated from low-rank structures, we fit trace regressions penalized by nuclear norms (Zhou and Li, 2014) on the training set. We evaluate the prediction power of $\mathbf{X}$ on Y through the root

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

mean square errors. We calculate both the root mean square errors and the trace correlation between $\mathbf{X}$ and Y on the test set. For Model (7), we fix $H = 64$ and set all n_hs equal to 16 to calculate the trace correlation. The relationships between the normalized trace correlations and the root mean square errors are presented in Figure 4.

The results of Model (6) and (7) are summarized in Figure 4 (a)–(c) and (d)–(f), respectively. In both models, as m varies from 0 to 32, the root mean square errors decrease in Figure 4 (a) and (d), and the normalized trace correlations increase in Figure 4 (b) and (e). Figures 4 (c) and (f) are particularly noteworthy as they demonstrate a strong agreement between the normalized trace correlations and root mean square errors. This, once again, echoes the observation we have made through the illustrative example in Section 1. In Figure 1 (c), Figure 4 (c) and (f), the trace correlations match the prediction powers pretty well. This interesting observation indicates that we can evaluate the prediction power through trace correlation. To be precise, if the trace correlation is large enough, we can conclude that $\mathbf{X}$ is very predictive for Y without fitting a complex predictive model. However, if the trace correlation is relatively small, we may have to collect additional covariates to enhance the prediction power.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

Table 4: The averages (and standard deviations) of trace correlations based on 10,000 repetitions for Example 4 with $\delta = \{1.00, 0.50, 0.25, 0.00\}$.

TC	$\delta = 1.00$	$\delta = 0.50$	$\delta = 0.25$	$\delta = 0.00$
$\widehat{\mathrm{T}}(X \mid Y)$	0.002(0.015)	0.016(0.016)	0.040(0.019)	0.089(0.021)
$\widehat{\mathrm{T}}(X^2 \mid Y)$	0.012(0.018)	0.079(0.028)	0.247(0.035)	0.859(0.008)

4. Detecting Nonfunctional Dependence

Trace correlation is designed to detect functional dependence. In this section, we adapt trace correlation in the reproducing kernel Hilbert space to detect nonfunctional dependence. Let us start from an illustrative example.

Example 4. We generate the random variables (X, Y) through an intermediate variables Z. Specifically, we generate (Z, Y) from a uniform distribution on the unit circle, such that $Z^2 + Y^2 = 1$. In addition, we set $X = Z + \delta\varepsilon$, where ε is standard normal and $\delta = \{0.00, 0.25, 0.50, 1.00\}$. We remark here that, X is not functionally dependent upon Y even when $\delta = 0$, but the transformed random variable, $\Psi(X) = X^2$, is.

We set $n = 256$ and $H = 16$, and calculate $\widehat{\mathrm{T}}(X \mid Y)$ and $\widehat{\mathrm{T}}(X^2 \mid Y)$ at different values of δ. The resultant averages (and standard deviations) of trace correlations based on 10,000 repetitions are displayed in Table 4.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

These results indicate that $\widehat{\mathrm{T}}(X \mid Y)$ remains very small, with a value of just 0.089, even when δ decreases to 0.00. However, by replacing X with X^2, the resulting value of $\widehat{\mathrm{T}}(X^2 \mid Y)$ increases significantly to 0.859. This example motivates us to generalize the concept of trace correlation by introducing some transformations of the random object $\mathbf{X}$.

Let us define the trace correlation in the reproducing kernel Hilbert space. To be precise, let $\mathbf{\Psi}(\mathbf{X})$ denote the feature maps, which are potentially infinite-dimensional transformations of the random object $\mathbf{X}$. Instead of working with $\mathbf{X}$, we use $\mathbf{\Psi}(\mathbf{X})$ and define the trace correlation by

$$\mathrm{T}\{\mathbf{\Psi}(\mathbf{X}) \mid Y\} = 1 - E\{\widetilde{d}_{\mathbf{\Psi}}(Y)\} \Big/ E\Big[d\{\mathbf{\Psi}(\mathbf{X}_1), \mathbf{\Psi}(\mathbf{X}_2)\}\Big],$$

where $\widetilde{d}_{\mathbf{\Psi}}(y) = E\Big[d\{\mathbf{\Psi}(\mathbf{X}_1), \mathbf{\Psi}(\mathbf{X}_2)\} \;\Big|\; Y_1 = y, Y_2 = y\Big]$ and

$$d\{\mathbf{\Psi}(\mathbf{X}_1), \mathbf{\Psi}(\mathbf{X}_2)\} = \arccos\Big(\{1 + \langle \mathbf{\Psi}(\mathbf{X}_1), \mathbf{\Psi}(\mathbf{X}_2)\rangle\} \Big[\{1 + \langle \mathbf{\Psi}(\mathbf{X}_1), \mathbf{\Psi}(\mathbf{X}_1)\rangle\}\{1 + \langle \mathbf{\Psi}(\mathbf{X}_2), \mathbf{\Psi}(\mathbf{X}_2)\rangle\}\Big]^{-1/2}\Big).$$

These quantities merely involve inner products of the form $\langle \mathbf{\Psi}(\mathbf{X}_1), \mathbf{\Psi}(\mathbf{X}_2)\rangle$. Suppose that there exists a reproducing kernel such that $K(\mathbf{X}_1, \mathbf{X}_2) = \langle \mathbf{\Psi}(\mathbf{X}_1), \mathbf{\Psi}(\mathbf{X}_2)\rangle$, it follows that

$$d\{\mathbf{\Psi}(\mathbf{X}_1), \mathbf{\Psi}(\mathbf{X}_2)\} = \arccos\Big(\{1 + K(\mathbf{X}_1, \mathbf{X}_2)\} \Big[\{1 + K(\mathbf{X}_1, \mathbf{X}_1)\}\{1 + K(\mathbf{X}_2, \mathbf{X}_2)\}\Big]^{-1/2}\Big).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

The kernel trace correlation allows us to work with K, instead of $\mathbf{\Psi}$.

We remark that, even when $\mathrm{T}\{\mathbf{\Psi}(\mathbf{X}) \mid Y\} = 0$, $\mathbf{X}$ and Y are not necessarily independent. To ensure $\mathrm{T}\{\mathbf{\Psi}(\mathbf{X}) \mid Y\}$ inherits this desirable property from trace correlation, we must restrict $\mathbf{\Psi}(\mathbf{X})$ into the family of continuous and invertible mappings. This is rigorously formulated in the following proposition.

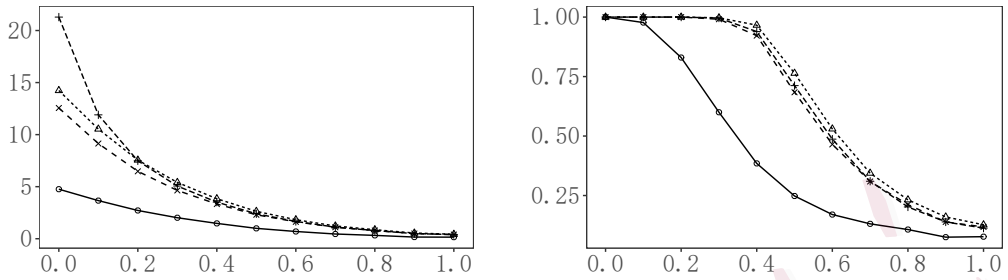
Proposition 2. *In general,* $0 \leq \mathrm{T}\{\mathbf{\Psi}(\mathbf{X}) \mid Y\} \leq 1$. *In particular, suppose the feature mappings* $\mathbf{\Psi}$ *are continuous and injective, or alternatively, the continuous positive definite kernel function* K *is universal,* $\mathrm{T}\{\mathbf{\Psi}(\mathbf{X}) \mid Y\} = 0$ *if and only if* $\mathbf{X}$ *and* Y *are independent, and* $\mathrm{T}\{\mathbf{\Psi}(\mathbf{X}) \mid Y\} = 1$ *if and only if* $\mathbf{X}$ *is completely dependent upon* Y.

The family of universal kernels (Micchelli et al., 2006) includes the most popular choices in the machine learning literature, such as the inverse multiquadric kernel $K(X_1, X_2) = \{1+\gamma(X_1-X_2)^2\}^{-1/2}$, the Gaussian kernel $K(X_1, X_2) = \exp\{-\gamma(X_1 - X_2)^2\}$ and the Laplacian kernel $K(X_1, X_2) = \exp(-\gamma|X_1 - X_2|)$.

We demonstrate the effectiveness of kernel trace correlation in Example 4, by applying the aforementioned universal kernels. We set $\delta = \{0, 0.1, \ldots, 1\}$ and $H = 16$. The results are summarized in Figure 5.

The kernel trick is observed to significantly increase trace correlations

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154



(a) (b)

Figure 5: The horizontal axis stands for $\delta \in \{0, 0.1, \ldots, 1\}$, and the vertical axes correspond to the normalized trace correlations in (a) and the empirical powers in (b), respectively. We compare $\widehat{\mathrm{T}}(X \mid Y)$ ($\circ$) with $\widehat{\mathrm{T}}\{\boldsymbol{\Psi}(X) \mid Y\}$ using Gaussian kernel ($\triangle$), Laplacian kernel ($+$), and inverse multiquadric kernel ($\times$).

and the empirical powers in tests for independence. One possible explanation for this phenomenon is that the feature maps underlying these kernels contain components that are functionally dependent upon Y, which assists in the detection of nonfunctional dependence. For instance, Gaussian kernel has an infinite-dimensional feature map $\boldsymbol{\Psi}(X) = \exp(-X^2/2)\big(1, X, X^2/2^{1/2}, X^3/6^{1/2}, \ldots\big)^{\mathrm{T}}$ by Taylor expansion. In Example 4, all X^d, for even d, are functionally dependent upon Y.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

5. Discussion

This paper introduces trace correlation, a novel method for testing independence and measuring the strength of association. The trace correlation characterizes both independence and complete dependence. This distinctive property makes trace correlation an efficient tool for evaluating the prediction performance before fitting any complicated nonlinear models. Trace correlation allows one random object to be matrix-valued. In the present context, we assume the conditioning random object to be univariate. If it is multivariate, we can simply divide the observations into several blocks using clustering methods like K-means. If it is also matrix-valued, specific data-driven approaches for clustering may be required. We also extend trace correlation for general complete dependence by adapting it in the reproducing kernel Hilbert space.

By definition, trace correlation is asymmetric. If symmetry is desired, an alternative approach is to introduce another trace correlation by switching the roles of the two random objects and taking the maximum. The concept of trace correlation has the potential to be generalized for testing conditional independence, which deserves further investigation.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

Supplementary Material

The online Supplementary Material includes all proofs and technical details.

Acknowledgments

This research is supported by grants from Renmin University of China (22XNA026) and National Natural Science Foundation of China (12225113 and 12171477).

References

Chatterjee, S. (2021). A new coefficient of correlation. *Journal of the American Statistical Association 116*(536), 2009–2022.

Deb, N., P. Ghosal, and B. Sen (2020). Measuring association on topological spaces using kernels and geometric graphs.

Dette, H., K. F. Siburg, and P. A. Stoimenov (2013). A copula-based non-parametric measure of regression dependence. *Scandinavian Journal of Statistics 40*(1), 21–41.

Gamboa, F., T. Klein, and A. Lagnoux (2018). Sensitivity analysis based on Cramér–von Mises distance. *SIAM/ASA Journal on Uncertainty Quantification 6*(2), 522–548.

Gretton, A., O. Bousquet, A. Smola, and B. Schölkopf (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In S. Jain, H. U. Simon, and E. Tomita (Eds.), *Algorithmic*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

Learning Theory, pp. 63–77. Springer Berlin Heidelberg.

Gretton, A., K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola (2007). A kernel statistical test of independence. In J. Platt, D. Koller, Y. Singer, and S. Roweis (Eds.), *Advances in Neural Information Processing Systems*, Volume 20, pp. 585–592. Curran Associates, Inc.

Gupta, A. and D. Nagar (1999). *Matrix Variate Distributions*. Monographs and Surveys in Pure and Applied Mathematics. Taylor & Francis.

Gupta, S. S. (1963). Probability integrals of multivariate normal and multivariate t^1. *The Annals of Mathematical Statistics 34*(3), 792–828.

Heller, R., Y. Heller, and M. Gorfine (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika 100*(2), 503–510.

Hsing, T. and R. J. Carroll (1992). An asymptotic theory for sliced inverse regression. *The Annals of Statistics 20*(2), 1040–1061.

Ke, C. and X. Yin (2020). Expected conditional characteristic function-based measures for testing independence. *Journal of the American Statistical Association 115*(530), 985–996.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika 30*(1–2), 81–93.

Kimeldorf, G. and A. R. Sampson (1978). Monotone dependence. *The Annals of Statistics 6*(4), 895–903.

Kong, E., Y. Xia, and W. Zhong (2019). Composite coefficient of determination and its ap-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

plication in ultrahigh dimensional variable screening. *Journal of the American Statistical Association 114*(528), 1740–1751.

Lancaster, H. (1963). Correlation and complete dependence of random variables. *The Annals of Mathematical Statistics 34*(4), 1315–1321.

Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*(11), 2278–2324.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association 86*(414), 316–327.

Li, Y. and L.-X. Zhu (2007). Asymptotics for sliced average variance estimation. *The Annals of Statistics 35*(1), 41–69.

Li, Z. and Y. Zhang (2020). On a projective ensemble approach to two sample test for equality of distributions. In H. D. III and A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning*, Volume 119 of *Proceedings of Machine Learning Research*, pp. 6020–6027. PMLR.

Lin, Q., Z. Zhao, and J. S. Liu (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics 46*(2), 580–610.

Micchelli, C. A., Y. Xu, and H. Zhang (2006). Universal kernels. *Journal of Machine Learning Research 7*(12), 2651–2667.

Pan, W., X. Wang, H. Zhang, H. Zhu, and J. Zhu (2020). Ball covariance: A generic measure

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

of dependence in banach space. *Journal of the American Statistical Association 115*(529), 307–317.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London 58*(347–352), 240–242.

Sejdinovic, D., B. Sriperumbudur, A. Gretton, and K. Fukumizu (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics 41*(5), 2263–2291.

Shen, C., C. E. Priebe, and J. T. Vogelstein (2020). From distance correlation to multiscale graph correlation. *Journal of the American Statistical Association 115*(529), 280–291.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology 15*(1), 72–101.

Székely, G. J. and M. L. Rizzo (2009). Brownian distance covariance. *The Annals of Applied Statistics 3*(4), 1236–1265.

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

Xu, K. and L. Zhu (2022). Power analysis of projection-pursuit independence tests. *Statistica Sinica 32*(1), 417–433.

Yin, X. and Q. Yuan (2020). A new class of measures for testing independence. *Statistica Sinica 30*(4), 2131–2154.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

Zhang, Y. and S. Yang (2024). Kernel angle dependence measures in metric spaces. *Journal of Computational and Graphical Statistics 0*(0), 1–18.

Zhang, Y. and L. Zhu (2023a). Projection divergence in the reproducing kernel hilbert space: Asymptotic normality, block-wise and slicing estimation, and computational efficiency. *Journal of Multivariate Analysis 197*, 105204.

Zhang, Y. and L. Zhu (2023b). Projective independence tests in high dimensions: the curses and the cures. *Biometrika 111*(3), 1013–1027.

Zhou, H. and L. Li (2014). Regularized matrix regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology 76*(2), 463–483.

Zhu, C., X. Zhang, S. Yao, and X. Shao (2020). Distance-based and RKHS-based dependence metrics in high dimension. *The Annals of Statistics 48*(6), 3366–3394.

Zhu, L., B. Miao, and H. Peng (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association 101*(474), 630–643.

Zhu, L. and K. Ng (1995). Asymptotics of sliced inverse regression. *Statistica Sinica 5*(2), 727–736.

Zhu, L., K. Xu, R. Li, and W. Zhong (2017). Projection correlation between two random vectors. *Biometrika 104*(4), 829–843.

Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China

E-mail: delin1997@ruc.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0154

Center for Applied Statistics and Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China

E-mail: zhu.liping@ruc.edu.cn